Journal of High School Science

Advanced Machine Learning approach for CO_2 emission forecasting: leveraging global patterns to overcome data insufficiency and improve model accuracy

Hung G

Submitted: August 16, 2024, Revised: version 1, November 8, 2024, version 2, December 28, 2024
Accepted: January 8, 2025

Abstract

Carbon dioxide (CO_2) emissions are among the most significant contributing factors to climate change. The increase in industrial activities, transportation, and forest exploitation are direct causes of this ongoing crisis. Due to its rapid development, Taiwan faces the dual challenge of sustaining economic growth while mitigating environmental impacts. Building on previous studies, this paper presents a novel approach that considers the interconnectedness of global CO_2 patterns. This study trains seven machine learning (ML) models using the optimal training ratio to forecast Taiwan's CO_2 trends. The three top-performing models were identified as Gradient Boosting Regressor, FeedForward Neural Network (FFNN), and Random Forest Regressor. After optimization, Gradient Boosting achieved an R^2 score of 0.997, FFNN scored 0.996, and Random Forest reached 0.995. These models demonstrated high accuracy with no signs of overfitting. This paper aims to provide policymakers in Taiwan with insights to effectively address CO_2 emissions.

Keywords

CO_2 emissions, Artificial intelligence, Machine learning, Data modeling, Forecasting emissions, Model optimization, Gradient Boosting Regressor, FeedForward Neural Network, Random Forest Regressor

Gordon Hung, Hsinchu County American School, No. 189, Gaotie 2nd Rd, Zhubei - Hsinchu, Taiwan.
gordonh741@gmail.com

1 Introduction

The escalating concerns revolving around climate change have necessitated profound strategies to combat intense CO_2 emissions worldwide. Significant emissions have caused an increase in atmospheric temperature, the rise of sea levels, and the disruption of natural habitats. The latest data from NASA Climate Science shows that human activities have raised the atmosphere's CO_2 content by over 50% in less than 200 years (1). According to research from Stanford University, in 2023, there were over 40 billion metric tons (MTs) of carbon dioxide in the atmosphere (2). Daily, approximately 456 liters of carbon dioxide are inhaled, which is around 152 times as much water as is consumed each day (3). This significant problem has imposed both physical and mental burdens, with many individuals developing diseases such as Respiratory Acidosis and Chronic Obstructive Pulmonary Disease (COPD) due to prolonged exposure to high levels of CO_2 (4). Furthermore, exposure to air pollutants can also lead to severe mental issues, including impaired cognitive function, psychoses, and dementia.

Globally, the continued increase in CO_2 emissions is intricately connected to various socio-economic factors. The burning of fossil fuels, increasing industrial expansion, and rising demands for heating are all significant contributors to the deteriorating atmosphere (5). While some countries may be larger emitters than others, alleviating CO_2 emissions is not a national task. The interconnectedness of CO_2 emissions between countries stems from the inseparable global supply chain. Manufacturing processes that emit large amounts of CO_2 may take place in one country, while the products are consumed in another. Each stage of production, transportation, and even consumption contributes to CO_2 emissions, making the emissions of countries all heavily correlated. Another prominent factor that contributes to the interconnectedness of the world's CO_2 emissions is international agreements and policies. For example, countries like the United States, China, and Japan endorsed the 2023 International Maritime Organization (IMO) Strategy for the reduction of greenhouse gas (GHG) emissions from ships. This resulted in a reduction of CO_2 emissions worldwide as nations took measures based on the agreement (6). In addition to the amount of CO_2 emission between countries being heavily connected, the detrimental impacts are also co-experienced. When a country emits CO_2, it disperses through the atmosphere, contributing to global greenhouse gas concentrations. This means that emissions from one country can easily affect the air quality in other regions.

Nationally, Taiwan is characterized by its rapid industrialization and technological advancements; this creates a predicament, caught between economic growth and environmental preservation. In response to the significant consequences brought about by excessive CO_2 emissions, Taiwan's government has resolved to establish net-zero carbon emissions policies, first achieving low carbon and then progressing towards zero carbon (7). The government has initiated collaboration with the Chinese National Federation of Industries (CNFI) and related

industries to create a framework for net-zero emissions (8).

To aid policymakers in making more informed decisions, this paper presents a novel approach to forecasting CO_2 emissions, recommending three accurate and robust models that could be implemented to successfully predict future CO_2 trends in Taiwan. The remainder of this paper is organized as follows. Section 2 describes past work and how the study builds upon previous successes. Section 3 discusses data preprocessing and feature extraction techniques, provides a detailed explanation of relevant ML theory, and presents the three top ML models for forecasting CO_2 emissions in Taiwan. Section 4 showcases the results of the three top ML models. Finally, the paper concludes in Section 5.

2 Related Work

Several recent works have investigated using ML to analyze and predict CO_2 emissions. Kumari and Singh used univariate time-series data from 1980 to 2019 to build statistical models: the autoregressive-integrated moving average (ARIMA) model, seasonal autoregressive-integrated moving average with exogenous factors (SARIMAX) model, and the Holt-Winters model (9). Additionally, they utilized ML models, namely linear regression and random forest, along with a long short-term memory (LSTM) model. LSTM was identified as the best model for CO_2 prediction, achieving a 3.101% MAPE, 60.635 RMSE, and 28.898 MedAE. Birjandi et al. focussed on modeling CO_2 in Southeast Asian countries like Malaysia and Singapore (10). The artificial neural network (ANN) employed in the research, utilizing normalized radial basis and tansig transfer functions, demonstrated high precision. The optimal ANN architecture with normalized radial basis and 11 neurons in the hidden layer achieved the highest precision with an R^2 of 0.9997, while the tansig function resulted in a model with an R^2 of 0.9996. Liu employed regression, neural networks, and SVM models to understand their respective performance in the field of CO_2 emissions predictions. It was concluded that each of the models had its own trade-offs but exact means of evaluation were not applied to obtain the most optimal models (11).

Besides investigating the efficacy of different ML models in delivering CO_2 emission prediction, many works have also studied the main contributing factors to the increasing CO_2 emissions. Valeria and Stefano conducted a decomposition analysis for 33 world countries to closely examine the main contributing factors of CO_2 (12). The research concluded that GDP and energy efficiency were the main contributing factors to CO_2 emissions in all considered countries. Therefore, by optimizing these domains, emission levels could be greatly reduced. Yeh et al. utilized an analytic tool of Stochastic Impacts by Regression on Population, Affluence and Technology (STIRPAT) to understand the impact of population and economic growth on carbon emissions in Taiwan (13). This study was based on Taiwan's national data from 1990 to 2014. While the study concluded that population and economic growth were significant contributors to the increasing CO_2 emissions, it also suggested that Taiwan's CO_2 emissions may decrease, despite growth in both

population and economy. Arneth et al. performed a study that analyzed the carbon dioxide emissions caused by land use. Based on dynamic global vegetation model simulations, they suggested that the CO_2 emissions resulting from land-use changes may be severely underestimated (14). This study promoted reforestation and preventive measures against deforestation, as they could significantly reduce CO_2 levels.

3 Materials and Methods

3.1 Objective

Building upon previous studies, this research presents several important goals that will advance the field of CO_2 emission predictions with ML. This paper presents a novel approach that takes into account the interconnectedness of global CO_2 emission trends to train seven ML models: Random Forest, Gradient Boosting, Ada Boosting, Linear Regression, K Neighbors Regressor, Feedforward Neural Network (FFNN), and Long-Short Term Memory (LSTM). The models were trained on CO_2 trends from countries with similar emissions to Taiwan, increasing the sample size and improving the model's ability to generalize. Thereafter, the most optimal models for CO_2 emission predictions were selected by comparing their respective R^2 scores and Mean Absolute Percentage Error (MAPE). Lastly, to enhance the robustness and accuracy of the ML models, hyperparameter searches were performed to optimize the models.

3.2 Data

Data selection and processing is an important task in predictive ML. Figure 1 is a diagram of the steps involved in the dataset preparation. Detailed explanations of each step are given in the following subsections.

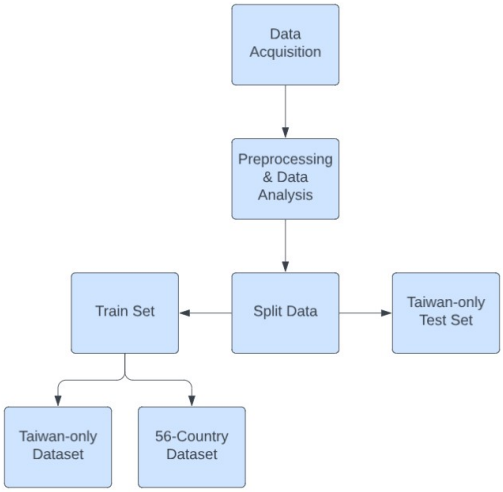

Figure 1. Steps involved in processing and splitting the dataset.

3.2.1 *Acquisition*

This paper utilized a customized dataset that combined four public datasets from 'Our World in Data' (16). The first dataset provided the annual CO_2 emissions of 250 countries. The second dataset provided the number of people per km² of land area – population density. The third dataset provided the primary energy consumption, measured in terawatt-hours, by country. The final dataset provided the total agricultural land, which is a combination of cropland and grazing land, per person.

3.2.2 *Preprocessing*

After acquiring the customized dataset, preprocessing techniques were applied to clean and organize the data for modeling. First, all 'Not Available' (NA) values in the datasets were dropped. Next, all features, except "Entity," were converted to numeric values. As ML models often associate unwanted bias with non-numerical values, one-hot encoding was employed to standardize the country values. A Standard Scaler was then applied to normalize the data for the neural networks. The Standard Scaler transforms the data by removing the mean and scaling it to unit variance, resulting in each feature having a mean of 0 and a standard deviation of 1. Since CO_2 emission values vary significantly among certain countries, countries with CO_2 emissions more than 50% higher or lower than those of Taiwan were filtered out to prevent the models from being overly influenced by extreme values. This approach helped prevent the models from predicting unrealistically high CO_2 emission values after observing trends in countries with such extremes. The objective of this methodology was to surpass the performance of training the models on the traditional Taiwan-only dataset. The final dataset spanned from 1965 to 2022, included 56 countries (including Taiwan), and consisted of a total of 1124 samples.

3.2.3 *Multicollinearity*

Multicollinearity is a highly undesirable condition in regression analysis where independent variables exhibit strong correlations with one another. This can lead to inaccurate evaluations of models and inflated standard errors. To ensure optimal performance, the variance inflation factor (VIF) was utilized to measure the correlation among the features employed. A VIF value between 1 and 5 indicates minimal multicollinearity, whereas a VIF value > 5 suggests significant multicollinearity.

3.2.4 *Correlation*

To gain a more comprehensive understanding of the relationship between individual features, a heatmap representing the average correlation between each feature was generated, where a value of 1 or -1 indicates 100% correlation, and 0 indicates no correlation. The correlation values for each feature across the 56 countries were calculated and subsequently averaged to construct the heatmap.

3.3 Novel approach to CO_2 emission prediction

It is important to ensure the accuracy and robustness of models. Contrary to traditional training methods for CO_2 emission predictions, a novel approach was implemented that combined data from countries with CO_2 trends similar to those of Taiwan in the training set, and the models then used this combined data to

predict CO_2 emissions in Taiwan. This approach alleviated the challenge of data scarcity by enabling the models to understand CO_2 trends in Taiwan by observing similar countries. This strategy highlighted and took advantage of the global nature of CO_2 emissions and leveraged the diverse variations in the world dataset to build more robust and generalized models. It allowed the models to recognize underlying contributing factors rather than overfitting to the characteristics of a specific region.

3.4 Establishing baseline models

Model selection is critical to ensuring that the optimal models are employed for a specific task. In this research, a wide range of ML models were studied, including three ensemble methods: Random Forest Regressor, Gradient Boosting Regressor, and AdaBoost Regressor; one linear model: Linear Regressor; one instance-based method: K Neighbors Regressor; and two neural networks: Feedforward Neural Network (FFNN) and Long Short-Term Memory (LSTM). These models were selected because they represent a diverse range of model architectures and have been shown to be effective in forecasting long-term trends. A method workflow diagram is shown in Figure 2. Section A of Figure 2 illustrates the training process of the models, while Section B illustrates the testing process.

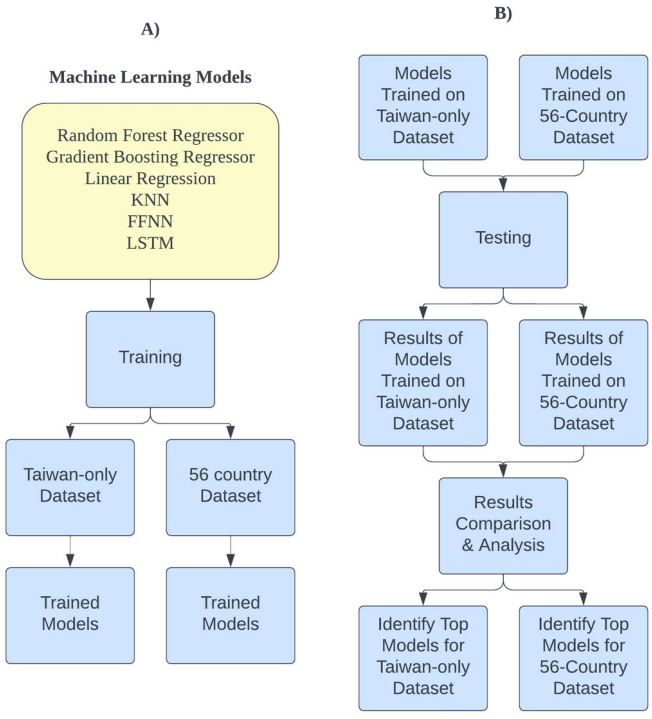

Figure 2. Steps involved in establishing the base models and analyzing the proposed methodology.

The models' performance was evaluated using two methods. The R^2 statistic, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is predictable from the independent variables (18).

$$R^2 = 1 - \frac{SS_{Res}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \quad (1)$$

where SS_{Res} is the sum of squares of residuals (also known as the residual sum of squares), calculated as $\sum_i (y_i - \hat{y}_i)^2$. SS_{TOT} is the total sum of squares, calculated as $\sum_i (y_i - \overline{y}_i)^2$. y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\overline{y}_i$ is the mean of the actual values of the CO_2 emission. R^2 ranges from 0 to 1. A value of 1 indicates that the model explains all the variability in the response variable whereas a value of 0 indicates that the model explains none of the variability in the response variable. While R^2 can successfully evaluate a model's ability to identify the general trend, it does not account for the overall accuracy of the model. On the other hand, MAPE measures the accuracy of the model by calculating the average magnitude of the errors as a percentage of the actual values (19). This offers more insight into the actual accuracy of the models.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (2)$$

where n is the number of observations. A lower MAPE indicates that the predicted value is closer to the actual value. It provides a comprehensive understanding of the relative error in percentage terms. By utilizing the R^2 and MAPE metrics together, this study measured both the models' ability to generalize the data, and the actual accuracy of the predictions.

To ensure the fairness of the evaluation, the models were trained and tested using the same randomly selected data samples. Achieving accurate results requires an optimal ratio for data splitting. Joseph concluded that the optimal training/testing splitting ratio is $\sqrt{p}$:1, where p represents the number of parameters in the given model (15). Since the model in this study used a total of 4 parameters, p = 4. Therefore, the most optimal training/testing ratio was 67:33. The training set consists of 752 samples, 38 of which were from Taiwan and 714 from other countries. For comparison, a Taiwan-only training dataset was also created, consisting of 40 samples exclusively from Taiwan. The testing set included 18 samples of Taiwan data.

3.5 Model optimizations

Model optimization is essential for utilizing the full potential of the selected models. Due to the large number of hyperparameters requiring

extensive optimization and the constraints of limited computational resources, the Random Search algorithm was employed to sample the hyperparameters efficiently.

3.6 Feature importance

In training ML models for predictive tasks, it is important to evaluate the contribution of each feature to understand how the models generate predictions. To achieve this, SHAP values were utilized to assess feature importance. Derived from cooperative game theory, SHAP values illustrate how each individual feature contributes to the final prediction outcome.

4 Results and Discussion

4.1 Data analysis

4.1.1 *Multicollinearity*

As discussed in Section 3, VIF values for each feature were calculated to evaluate potential multicollinearity among the features. Table 1 presents the VIF values for each feature, indicating little to no multicollinearity, which is optimal for training ML models.

Table 1. VIF value of each feature.

Features	VIF Values
Year	1.256
Population Density	1.094
Primary Energy Consumption (TWh)	1.022
Land Use Per Capita	1.136

4.1.2 *Correlation*

In the heatmap in Figure 3, the magnitude of the values represents the strength of the correlation.

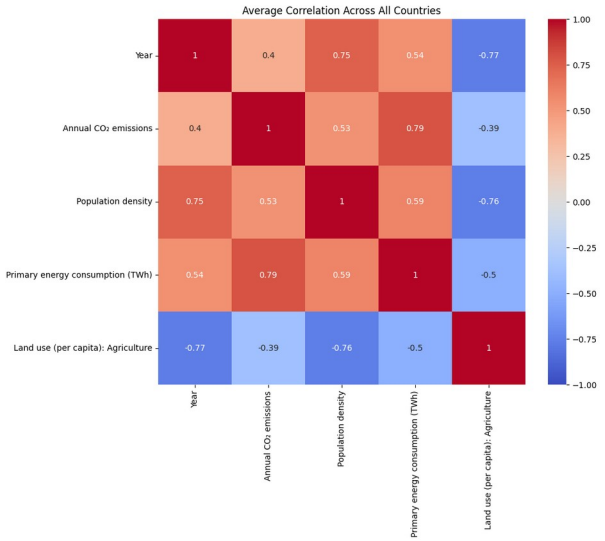

Figure 3. Heatmap that includes all the features.

Values closer to one or a negative one indicate stronger correlations, while values near zero indicate weaker correlations. As illustrated in Figure 3, the target variable, annual CO_2 emissions, is most positively correlated with primary energy consumption (TWh), followed by population density and year. Conversely, land use (per capita) in agriculture exhibits a negative correlation with annual CO2 emissions. The correlation values between all features are below 0.8, showing moderate influence of multicollinearity.

4.2 Performance analysis

Table 2 shows the performance of the models trained on the Taiwan-only dataset, whereas Table 3 shows the performance of the models trained on data from the 56 countries. Models trained on data from the 56 countries generally performed either on par or better compared with those trained on the Taiwan-only dataset. This improvement is most significant in relatively more complex models, including FFNN, LSTM, and Gradient Boosting. Among models trained on the Taiwan-only dataset, Linear Regression performed relatively well, as it was less dependent on sample size, whereas more complex models suffered due to data inadequacy. However, compared to models trained with a greater variety of data samples, Linear Regression performed worse as the dataset became more intricate and complex.

Table 2. Model performance measured by R² and MAPE during testing; models were trained on data from Taiwan only.

Models	R^2	MAPE
Random Forest	0.986	0.078
Gradient Boosting	0.986	0.070
Ada Boosting	0.983	0.115
Linear Regression	0.983	0.087
KNN	0.984	0.121
FFNN	0.975	0.224
LSTM	0.640	1.024

Table 3. Model performance measured by R² and MAPE during testing; models were trained on data from 56 countries.

Models	R^2	MAPE
Random Forest	0.989	0.062
Gradient Boosting	0.995	0.042
Ada Boosting	0.971	0.318
Linear Regression	0.980	0.093
KNN	0.985	0.081
FFNN	0.991	0.059
LSTM	0.987	0.132

Analyzing the evaluation metrics and the graphs of the models during training and testing, there were no apparent signs of significant overfitting within the three top models, even in their most basic state. The relative performance of each model on the dataset with 56 countries is ranked in Table 4. As shown, more complex models demonstrate higher accuracy because they excel at understanding CO_2 trends in other countries and applying that understanding to predict Taiwan's CO_2 trends. The results of the experimentation indicate that while Linear Regression achieves an R^2 score of 0.980, it is still relatively inaccurate compared to the more complex FFNN model. Furthermore, out of all the models tested, Linear Regression ranks 6th out of 7. Although the difference between an R^2 score of 0.980 and 0.990 may not seem significant, it represents the gap between a prediction that is off by millions of tons of CO_2 and one that is off by only a few thousand or even hundreds of tons. Therefore, more complex models successfully minimize this inaccuracy because they can account for finer details in the dataset that are often overlooked by simpler models like Linear Regression. These improvements in the models can offer policymakers the most comprehensive predictions. Based on the results, the three best-performing models were: (1) Gradient Boosting, (2) FFNN, and (3) Random Forest.

Table 4. Model performance ranked from best to worst when trained on the multi-country dataset.

Rankings	Models
1	Gradient Boosting
2	FFNN
3	Random Forest
4	KNN
5	LSTM
6	Linear Regression
7	Ada Boosting

4.3 Top models and optimization

4.3.1 *Gradient Boosting Regressor*

Gradient Boosting Regressor is a sophisticated ensemble model that relies significantly on carefully tuning its hyperparameters. Gradient Boosting starts with a single leaf that represents the initial guess as an overall average for the target. It improves the accuracy of the predictions by focusing on the residuals, or prediction errors, from the initial guess (25). A new regression tree is trained to predict these residuals in each iteration. The newly trained tree's predictions are then added to the existing model with a certain weight determined by a learning rate. This process minimizes the residuals step by step, refining the model's accuracy. Equations 3 through 7 present the steps involved in constructing Gradient Boosting Regressor iteratively.

Initialization

$$F_0(x) = arg\,min_c \sum_{i=1}^{n} L(y_i|, c) \quad (3)$$

Model Construction For m = 1 to M

1. Compute Residuals

$$r_{i,m} = -\left(\frac{\delta(y_i), F_{m-1}(x_i)}{\delta F_{m-1}(x_i)} \right) \quad (4)$$

2. Fit Weak Learner

$$h_m(x) = arg\,min_h \sum_{i=1}^{n} \left(r_{i,m} - h(x_i) \right) \quad (5)$$

3. Update Model

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (6)$$

4. Final Model

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x) \quad (7)$$

where $F_0(x)$ is the initial prediction value. c is the constant that minimizes the loss function in the initial model. $r_{i,m}$ is the residuals in iteration m. $h_m(x)$ is the weak learner fitted to the residuals in iteration m. γ_m is the learning rate that scales the weak learner's contribution. $F_m(x)$ is the updated prediction model after iteration m. $F_M(x)$ is the final model after all iterations.

The hyperparameters that contributed most significantly to the robustness and accuracy of the model were selected: number of estimators, learning rate, max features, max depth, min samples split, min samples leaf, and subsample. The number of estimators determines the total number of boosting stages to be built, where each stage corrects the residuals of the previous stages. A higher number of estimators generally improves model performance; however, it also increases computation cost and model complexity. The learning rate determines the significance of each tree to the overall prediction. A smaller learning rate is robust but slow, whereas a greater learning rate is fast but prone to overfitting. Max features

determines the maximum number of features considered for splitting a node during the construction of individual trees; this controls the correlation between each tree, impacting the model's overall generalization ability. Max depth measures the maximum depth of each individual tree: trees that are too deep often suffer from overfitting, whereas trees that are too shallow suffer from underfitting. "Min samples split" determines the minimum number of samples required to split an internal node. A higher value prevents the model from capturing unnecessary noise, while a lower value allows the model to recognize more detailed patterns. Subsample determines the fraction of samples used to fit each individual tree. This hyperparameter adds an element of randomness to the model; lower values reduce the variance of the model by averaging predictions from trees trained on different subsets of the overall data. Table 5 shows the hyperparameter optimization process of the Gradient Boosting Regressor.

Table 5. Hyperparameter Optimization for Gradient Boosting Regressor

Hyperparameters	Values	Results
Number of Estimators	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	800
Learning Rate	0.1, 0.001, 0.0001	0.01
Max Features	auto, sqrt, log2	log2
Max Depth	3, 5, 7, 9, 11, 13, 15, 17, 19, 21, None	9
Min Samples Split	2, 5, 10, 20	20
Min Samples Leaf	1, 2, 4	1
Subsample	0.7, 0.8, 0.9, 1.0	0.7

4.3.2 *Feedforward Neural Network*

A Feed-Forward Neural Network (FFNN) is an artificial neural network architecture that processes input through interconnected layers of neurons, inspired by the perceptron model. It uses a "divide-and-conquer" approach, where the input layer receives data, the hidden layers perform transformations, and the output layer synthesizes the results, with synaptic weights and activation functions optimized through techniques like backpropagation (21-23). Equations (8) and (9) represent the forward pass through a single neuron in the network.

$$u_k = \sum_{k=1}^{n} x_k \times w_k + b_k \tag{8}$$

$$j_k = activation\left(u_k\right) \tag{9}$$

where u_k is the sum of weighted inputs. j_k is the output of the kth neuron. x_k is the k^{th} input fromthe preceding neuron. w_k is the synaptic weight of the k^{th} input. b_k is the bias of the k^{th}

neuron: constant. activation is the activation function.

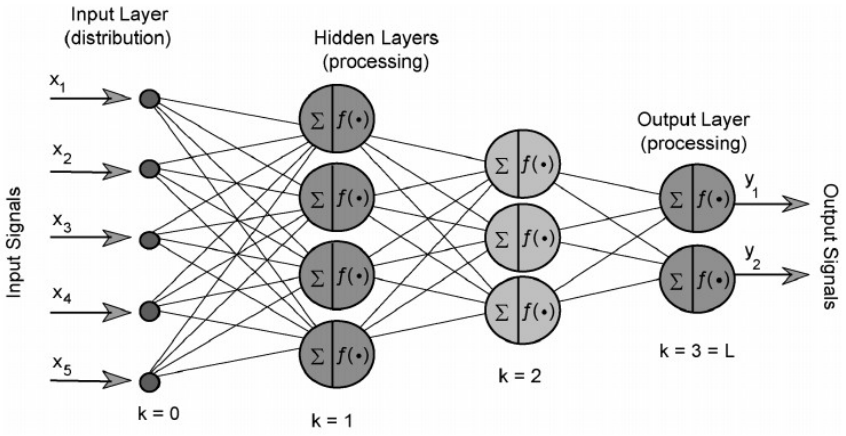

Figure 4. Basic Model Architecture of FFNN that demonstrates the functionalities of the input layer, hidden layers, and output layer.

The number of hidden layers, the number of units in each hidden layer, the activation function, the learning rate, the optimizer, the batch size, the dropout rate, and the number of epochs were tuned. The number of hidden layers determines the number of layers between the input and output layers in the network. A larger number of layers allows the model to understand deeper patterns but also increases the risks of overfitting. The number of units, or neurons, in each hidden layer, influences how well the model captures the general pattern. The activation function introduces non-linearity into the network, allowing the model to recognize more complex behaviors. The optimizer updates the network's weights based on the calculated gradients during backpropagation. Batch size determines the number of training samples used in one forward/backward pass. Smaller batch sizes offer more frequent updates while larger sizes provide smoother gradients. Dropout is a regularization technique where a randomly selected portion of neurons is dropped during training to prevent overfitting. However, it may also disrupt the learning patterns of the model. Lastly, the number of epochs determines the number of times the entire training dataset is passed through the network. Passing the dataset through the model too many times can result in overfitting, whereas too few times would result in underfitting. Table 6 shows the hyperparameter optimization process of FFNN.

Table 6. Hyperparameter Optimization for FFNN

Hyperparameters	Values	Results
Number of hidden layers	1, 2, 3, 4	2
Number of units in hidden layer 1	32, 64, 96, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448, 480, 512	384
Activation Function	relu, tanh, sigmoid	relu
Learning Rate	0.1, 0.01, 0.001, 0.0001	0.001
Units in Hidden Layer 2	32, 64, 96, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448, 480, 512	288
Optimizer	adam, rmsprop, sgd	adam
Batch Size	16, 32, 64, 128	128
Dropout Rate	0.1, 0.2, 0.3, 0.4, 0.5	0.4
Number of Epochs	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	90

4.3.2 *Random Forest Regressor*

A Random Forest Regressor is an ensemble model that builds multiple decision trees and averages their predictions to enhance accuracy and to prevent overfitting. It uses Bootstrap Aggregating to train each tree on a random subset of data, then aggregates their predictions for a final result. Equations 10-13 present the steps involved in constructing the Random Forest Regressor.

Initialization

$$D_b \approx Bootstrap(D) \tag{10}$$

Tree Construction

$$T_b = DecisionTree(D_b, m) \tag{11}$$

Prediction Aggregation

$$F(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{12}$$

Final Model

$$F_M(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{13}$$

Where b ranges from 1, 2…, B, and B is the total number of decision trees in the forest. D_b is the bootstrapped sample used to train the b^{th} decision tree. m is the number of features considered at each split in the decision trees. $T_b(x)$ is the prediction made by the b^{th} decision tree for the input x. $F_M(x)$ is the final output of the model after aggregating the predictions from all B trees.

The following hyperparameters were tuned: number of estimators, max features per split, max tree depth, min samples per split, min samples per leaf, and bootstrap. The number of estimators represents the number of trees in the forest. A larger number of estimators can improve the model's robustness but is also more computationally challenging. Max features per split represents the number of features to consider when looking for the optimal split at each node. Max tree depth sets the maximum depth of each tree in the forest; this reduces the risks of overfitting by preventing the trees from being overly influenced by unnecessary noise. Min samples per split specifies the minimum number of samples needed to split a node. A larger value can prevent overfitting as it forces the trees to be more generalized, whereas a smaller value more effectively captures the details of the dataset. Lastly, bootstrap is a boolean parameter that determines whether bootstrapping is applied when constructing the trees. If set to True, each tree is trained on a random subset of the training data; if set to False, the entire dataset will be used to train each tree. Table 7 shows the hyperparameter optimization process of the Random Forest Regressor.

Table 7. Hyperparameter Optimization for Random Forest Regressor

Hyperparameters	Values	Results
Number of estimators	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	900
Max Features per Split	auto, sqrt, log2	sqrt
Max Tree Depth	3, 5, 7, 9, 11, 13, 15, 17, 19, 21, None	17
Min Samples per Split	2, 5, 10, 20	5
Min Samples per Leaf	1, 2, 4	2
Bootstrap	True, False	True

4.4 Optimized models

The optimized models outperformed the base models. Table 8 displays the best results of the optimized models. Figures 5-7 show the graphical view of the models' performances on the test set composed of Taiwan's data. The real data samples are taken directly from the targets of the test set while the predictions are taken directly from the model's outputs. The closer the data points are to the ideal fit, the better the model.

Table 8. Performance of the optimized models measured in R^2 and MAPE during testing.

	FFNN	Gradient Boosting Regressor	Random Forest Regressor
R^2 Score Prior Optimization	0.991	0.995	0.989
MAPE Score Prior Optimization	0.059	0.042	0.062
R^2 Score After Optimization	0.996	0.997	0.995
MAPE Score After Optimization	0.049	0.040	0.058

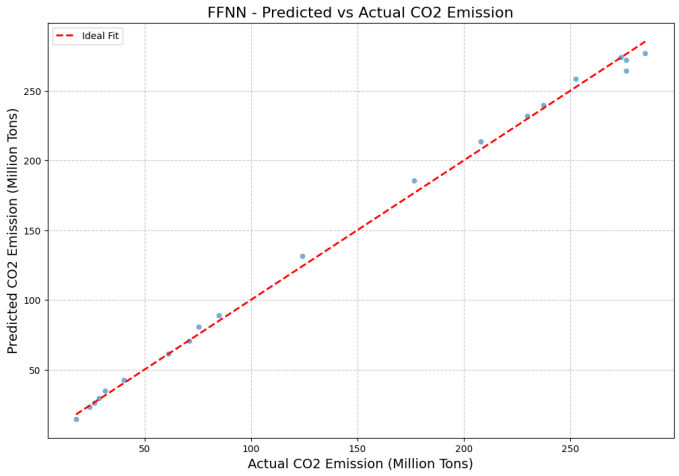

Figure 5. Graph of the predictions of FFNN after optimization

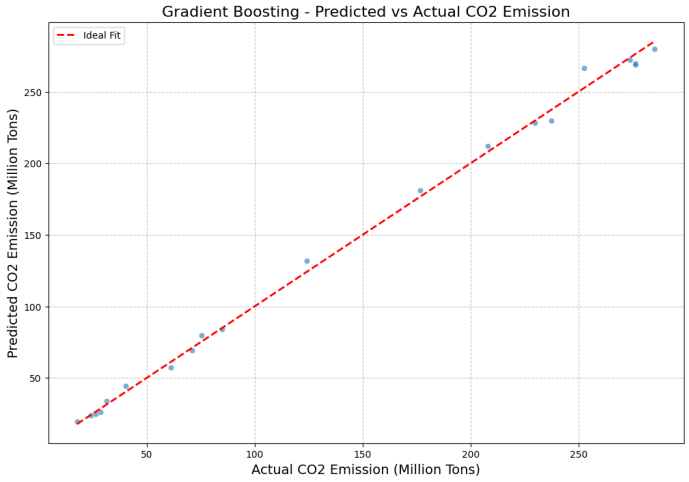

Figure 6. Graph of the predictions of Gradient Boosting Regressor after optimization

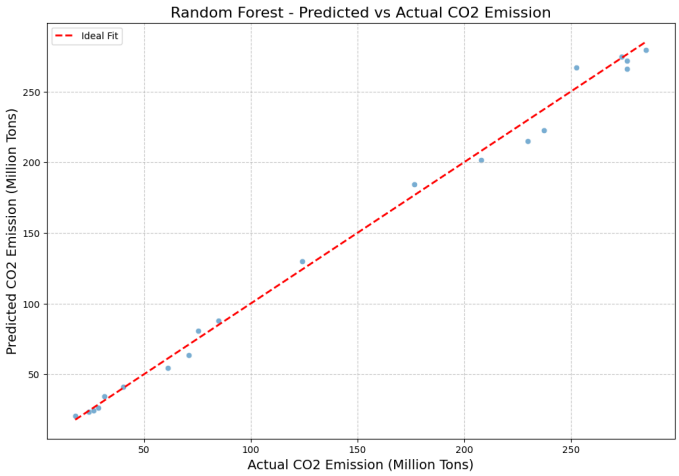

Figure 7. Graph of the predictions of Random Forest Regressor after optimization

The performance of each model improved slightly after extensive optimizations. The final evaluation of the models shows the top three most optimal models: Gradient Boosting, followed by FFNN, and lastly Random Forest. The robustness and accuracy of this novel approach to predicting CO_2 emissions are directly evident in the high performance of the models.

4.5 Feature importance results

Figures 8-10 showcase visualizations of the SHAP values of the top models when trained on the dataset with 56 countries.

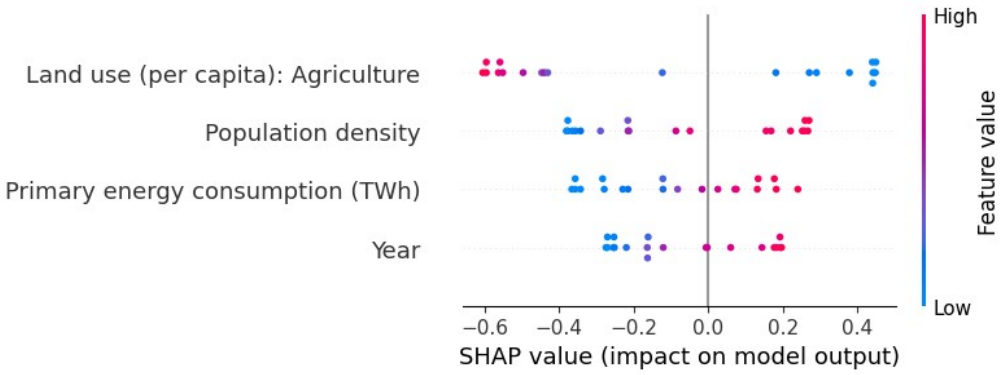

Figure 8. Feature importance of Gradient Boosting Regressor measured with SHAP values when trained on the dataset with 56 countries.

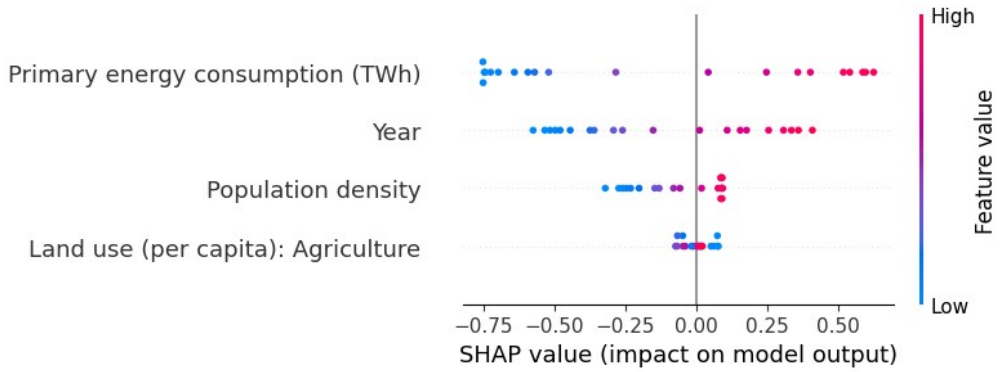

Figure 9. Feature importance of FFNN measured with SHAP values when trained on the dataset with 56 countries.

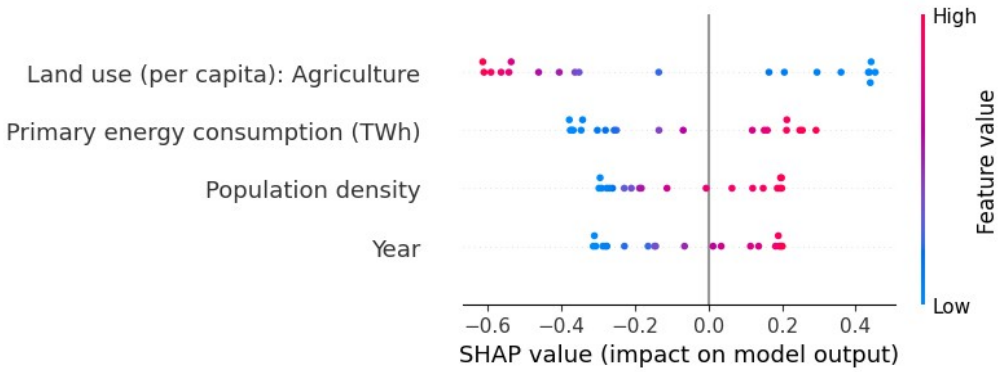

Figure 10. Feature importance of Random Forest Regressor measured with SHAP values when trained on the dataset with 56 countries.

As shown in the figures above, each model has its own feature importance due to their unique architecture. Therefore, the average SHAP value for each feature was calculated to identify the most important and least important features across the dataset. This final evaluation of each feature is shown in Table 9.

Table 9. The average importance of features ranked from most important to least important

Rankings	Features
1	Primary Energy Consumption (TWH)
2	Land Use (Per Capita): Agriculture
3	Year
4	Population Density

5. Conclusion

Successful forecasting of CO_2 emissions is critical in the contemporary world. As the demands for energy and economic activities continue to rise, understanding and mitigating the impacts of CO_2 emissions becomes increasingly crucial. Accurate CO_2 emission predictions allow policymakers, governments, and organizations to make informed decisions that benefit the deteriorating environment. This paper presented an ML-based CO_2 emission forecasting model. The study identified the three most optimal ML models—FFNN, Gradient Boosting, and Random Forest—for predicting CO_2 emissions in Taiwan. Furthermore, the research highlighted the most suitable hyperparameters for each of the selected models. The results indicated that the proposed method may be useful in forecasting CO_2 emissions in the absence of sufficient data. Future research can exploit the flexibility of the models to adjust them for other specific countries.

5.1 Limitations

A limitation of this study lies in the selection of countries. The selection process arbitrarily included countries with emissions ranging from 50% more to 50% less than those of Taiwan. Future work can address this limitation by utilizing models specifically designed to process domain transfer, such as Domain-Adversarial Neural Networks (DANN), Transfer Learning, and Generative Adversarial Networks (GANs) for domain adaptation. Furthermore, additional models such as regression support vector machine (RSVM) and Adaptive Neuro-fuzzy Inference System (ANFIS) can be studied to identify more effective models. Lastly, time-series models such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR) can be utilized to capture temporal dependencies and potentially improve prediction accuracy.

6. Acknowledgments

I would like to thank my mentor, Dr. Salinna Abdullah, for supporting me throughout the experimentation and writing processes.

7. References

1. Climate change - nasa science. *NASA*. https://science.nasa.gov/climate-change/

2. Global carbon emissions from fossil fuels reached record high in 2023. Stanford Doerr School of Sustainability (2024). https://sustainability.stanford.edu/news/global-carbon-emissions-fossil-fuels-reached-record-high-2023

3. U.-E. Team, How much pollution do you breathe in every day? what happens when you breathe in. *U* (2022). https://www.u-earth.eu/post/how-much-pollution-breathe-every-day

4. World-first study shows increased atmospheric CO2 levels damage young lungs. Telethon Kids Institute. https://www.telethonkids.org.au/news--events/news-and-events-nav/2021/january/study-shows-increased-co2-levels-damage-lungs/

5. Causes and effects of climate change. United Nations. https://www.un.org/en/climatechange/science/causes-effects-climate-change

6. IMO strategy on reduction of GHG emissions from ships. International Maritime Organization. https://www.imo.org/en/OurWork/Environment/Pages/IMO-Strategy-on-reduction-of-GHG-emissions-from-ships.aspx

7. *National Development Council-Taiwan's pathway to net-zero emissions in 2050*. https://www.ndc.gov.tw/en/Content_List.aspx?n=B154724D802DC488

8. Industrial Development Administration, Ministry of Economic Affairs, The Ministry of Economic Affairs Collaborate with the Chinese National Association of Industries to establish the "Industrial Carbon Neutrality Alliance." *Industrial Development Administration, Ministry of Economic Affairs*. https://www.ida.gov.tw/ctlr?lang=1&PRO=news.rwdNewsView&type=News&id=40744

9. Kumari S, Singh SK. Machine learning-based time series models for effective CO2 emission prediction in India. Environ Sci Pollut Res 30:116601-116616, (2023). https://doi.org/10.1007/s11356-022-21723-8

10. Birjandi AK, Alavi MF, Salem M, Assad MEH, Prabaharan N. Modeling carbon dioxide emission of countries in southeast of Asia by applying artificial neural network. Int J Low Carbon Technol 17:321-326, (2022). https://doi.org/10.1093/ijlct/ctac002

11. Liu X. CO2 emissions prediction based on regression, neural network and SVM. Appl Comput Eng 54:98-103, (2024). https://doi.org/10.54254/2755-2721/54/20241407

12. V. Andreoni, S. Galmarini, Drivers of CO2 emissions variation: A decomposition analysis for 33 world countries. Energy 103:27–37, (2016). https://doi.org/10.1016/j.energy.2016.02.096

13. J.-C. Yeh, C.-H. Liao, Impact of population and economic growth on carbon emissions in Taiwan using an analytic tool STIRPAT. Sustainable Environment Research. 27, 41–48 (2017). https://doi.org/10.1016/j.serj.2016.10.001

14. Arneth, A., Sitch, S., Pongratz, J. et al. Historical carbon dioxide emissions caused by land-use changes are possibly larger than assumed. Nature Geosci 10, 79–84 (2017). https://doi.org/10.1038/ngeo2882

15. V. R. Joseph, Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal. 15, 531–538 (2022). https://doi.org/10.1002/sam.11583

16. Our World in Data, M. Roser, Our world in data. Our World in Data (2024). https://ourworldindata.org/

17. Pearson correlation. Pearson Correlation - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/pearson-correlation

18. J. Fernando, R-squared: Definition, formula, uses, and limitations. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

19. Working with planning. Oracle Help Center (2024). https://docs.oracle.com/en/cloud/saas/planning-budgeting-cloud/pfusu/insights_metrics_MAPE.html

20. Some specific models of artificial neural nets. McCullogh-Pitts and Perceptron Models. https://www.cs.cmu.edu/afs/club/user/cmccabe/ecee.colorado.edu/~ecen4831/lectures/NNet2.html

21. A. K. Ghosh, Divide and conquer algorithm - working, Advantages & Disadvantages. Learn to Code, Prepare for Interviews, and Get Hired. https://www.scholarhat.com/tutorial/datastructures/divide-and-conqure-algorithm

22. Multilayer Perceptron. Multilayer Perceptron - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

23. J. Li, J. Cheng, J. Shi, F. Huang, Brief introduction of back propagation (BP) Neural Network algorithm and its improvement. Advances in Intelligent and Soft Computing, 553–558 (2012). https://doi.org/10.1007/978-3-642-30223-7_87

24. What is linear regression? IBM (2021). https://www.ibm.com/topics/linear-regression

25. Gradient boosting regression. Scikit. https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html