Rocznik Ochrona Środowiska

Volume 28 Year 2026 ISSN 2720-7501 pp. 587-600

https://doi.org/10.54740/ros.2026.040 open access

Received: May 2026 Accepted: July 2026 Published: July 2026

Machine Learning for Multi-Scale Carbon Emission Prediction and Environmental Risk Assessment

Ningyao Yu

College of Marxism, Xiangtan University, Xiangtan, Hunan 411105, China

https://orcid.org/0000-0002-6911-6250

**corresponding author's e-mail: yuny19910602@126.com*

Abstract: Accurate prediction of carbon emissions is crucial for addressing climate change and protecting ecological safety, but many existing models lack rigorous validation on independent data. This study develops and compares two machine learning models—random forest (RF) and convolutional neural network (CNN)—for predicting CO_2 emissions at both international (163 countries) and sub-national (284 Chinese cities) scales. Using two distinct datasets, models were trained and optimized via grid search and 10-fold cross-validation, then evaluated on fully independent test sets. For the global dataset, RF achieved superior performance (test set $R^2 = 0.984$, RMSE = 0.427), while for the more complex urban dataset, CNN demonstrated better generalization (test set $R^2 = 0.890$, RMSE = 0.199). SHAP analysis revealed key drivers: trade openness and GDP were dominant at the country level, whereas economic structure, carbon sequestration, and urban form factors played significant roles at the city level. The study highlights the importance of external validation and shows that model performance depends on data scale and feature complexity, providing robust tools for emission forecasting, environmental risk assessment, and policy support.

Keywords: carbon emission prediction, convolutional neural network, machine learning, model validation, random forest, SHAP analysis

1. Introduction

Carbon dioxide (CO_2) is not only a major greenhouse gas but also an environmental pollutant that poses potential risks to ecosystem stability and environmental safety. Accurate prediction and effective mitigation of CO_2 emissions are therefore paramount in the global effort to combat climate change and protect ecological health. The escalating volume of CO_2 emissions, primarily from industrial processes and energy consumption, underscores the urgent need for robust analytical tools to understand driving factors and forecast future trends (Wang et al. 2022; Dharmapriya et al., 2025). Traditionally, decomposition analysis frameworks such as the expanded Kaya identity and the logarithmic mean divisia index (LMDI) have been instrumental in disentangling the contributory effects of factors like technology, affluence, urbanization, and population scale on emissions (Dai et al., 2015; Jiang et al., 2023). Similarly, econometric models like the STochastic Impacts by Regression on Population, Affluence and Technology (STIRPAT) model, often coupled with ridge regression, have been widely employed to analyze the relationship between carbon emissions and socioeconomic drivers (Fan et al., 2020; Z. Yu et al., 2024; Y. Yu et al., 2024).

In recent years, machine learning (ML) has emerged as a powerful complementary paradigm, offering superior capability in modeling complex, nonlinear relationships inherent in science, technology, and industry (Cao et al., 2026; Balducci et al., 2018; Butler et al., 2018; Mekki et al., 2025; Noé et al., 2020; Pugliese et al., 2021; Sun, et al., 2017; Uncuoglu et al., 2022; Wu et al., 2023; Yalçın et al., 2024). A diverse array of ML algorithms has been applied to CO_2 emission forecasting. Studies have utilized models ranging from fundamental linear regression and support vector machines (SVM) to advanced ensemble methods and deep learning architectures. For instance, Kumari and Singh (2023) conducted a comprehensive comparative study on CO_2 emission prediction in India, employing six different time-series modeling approaches: three statistical models (ARIMA, SARIMAX, Holt-Winters), two machine learning models, and one deep learning-based long short-term memory (LSTM) model, while others have compared the performance of ordinary least squares (OLS), SVM, and gradient boosting regression (GBR) for transportation emissions prediction (Li et al., 2022). Ensemble methods, particularly tree-based models, have demonstrated notable success. Random forest (RF), an evolution of bagging designed to reduce model variance by averaging predictions from numerous de-correlated decision trees, is recognized for its high predictive accuracy and flexibility (Xia et al., 2022; Aria et al., 2021). Comparative studies have evaluated a suite of advanced boosting algorithms, including XGBoost, LightGBM, and NGBoost (natural gradient boosting), with SHapley Additive exPlanations (SHAP) used to interpret factor contributions (Jabeur et al., 2022). In the domain of deep learning, convolutional neural networks (CNNs) have been noted for their ability to concentrate on pertinent patterns and reduce the influence

of noise in real-world datasets (Maji et al., 2023), and LSTM networks optimized with algorithms like particle swarm optimization (PSO) have been developed for forecasting emissions in power grids (Aryai & Goldsworthy, 2023; Leerbeck et al., 2020). Other applied techniques include seasonal autoregressive integrated moving-average (SARIMA) models, decision trees, and various hybrid approaches (Zhong and Li, 2023; Bhatt et al., 2023; Begum & Mobin, 2025).

Despite these advancements, a critical gap persists in the rigorous validation of many proposed models. As noted in the literature, a model's strong explanatory power and performance on internal validation metrics (e.g., high R^2, low RMSE) may not reliably indicate its predictive capability for entirely new, unseen data. To the best of our knowledge, among the cited works, only Jabeur et al. (2022) implemented a rigorous evaluation protocol that included cross-validation followed by external testing on a completely independent hold-out dataset. This step is essential for establishing true generalizability and practical reliability, yet it remains underutilized.

To address this gap and advance robust predictive tools for environmental risk assessment, this study proposes a comprehensive machine learning framework for CO_2 emission modeling. We employ and critically compare two potent algorithms: the robust ensemble RF and the deep learning architecture CNN. Moving beyond internal performance metrics, our methodology subjects both models to stringent validation and robustness evaluation. Crucially, this includes external testing on a fully independent dataset, providing a more trustworthy assessment of their real-world predictive performance and ensuring the findings are not merely artifacts of overfitting to the training data. By doing so, this research aims to deliver not only accurate forecasts but also models whose reliability has been rigorously attested, ultimately supporting the evaluation of CO_2-related ecological risks and environmental safety.

2. Methods

2.1. Data and Variables

This study employs two independent datasets to support high-precision CO_2 emission forecasting at different geographical scales—international (country-level) and sub-national (city-level). The first dataset (Data Set I; see Table S1 in the Supporting Information) is sourced from Dharmapriya et al. (2025). It comprises a balanced panel of 163 countries from 2000 to 2019, classified according to World Bank income groups to account for the effects of development stage on emission trajectories. The target variable is carbon emissions per capita (CE). Key predictors include trade openness (TO), energy consumption per capita (EC), and GDP per capita (GDP), as detailed in Table 1.

Table 1. Definitions and units of variables in Data Set I

Variable	Definition	Unit
CE	Carbon emission	Metric ton per capita
TO	Trade openness	Percentage of GDP
EC	Energy consumption	Thousand kWh per capita
GDP	Gross domestic product	Thousands of Dollars per capita

The second dataset (Data Set II; see Table S2) is derived from Wang et al. (2022). It covers 284 prefecture-level cities in China from 2011 to 2017, capturing regional variations in economic structure, policy, and digital development critical for sub-national forecasting. The target variable is the natural logarithm of total CO_2 emissions ($lnCO_2$, in million tonnes).

Core predictors include metrics of digital finance, economic structure, development, and environmental policy. Specifically, these are: the digital financial inclusion index (lnDFI) and its three key dimensions—the coverage breadth index (coverage_breadth), reflecting service accessibility; the usage depth index (lnusg), measuring the intensity of actual use; and the digitization level index (lndig), indicating the degree of facilitation and credit enhancement. Industrial Structure (lnIns) is defined as the ratio of tertiary to secondary industry value-added, along with the separate output of the secondary (second) and tertiary (lnthird) sectors. Economic development is represented by real GDP per capita (lnGDP), technological capacity by the number of granted green patents (green_patent), and policy stringency by environmental regulation (ER), measured as the industrial solid waste utilization rate (Wang et al., 2022). The complete list of variables and their definitions is provided in Table 2.

Table 2. Definitions and units of variables for Data Set II

Descriptor	Definition	Unit
lnCO_2	Natural logarithm of total carbon dioxide emissions.	Logarithm of metric tons
cover-age_breadth	Coverage breadth index for digital financial inclusion (accessibility).	Index score
lnIns	Natural logarithm of Industrial Structure index (Tertiary / Secondary industry value-added).	Logarithm of ratio
lnthird	Natural logarithm of tertiary (service) industry value-added.	Logarithm of 100 million CNY
lnusg	Natural logarithm of the usage depth index for digital financial inclusion.	Logarithm of index score
ER	Environmental regulation, measured by industrial solid waste utilization rate.	Percentage (%)
second	Secondary (industrial) sector value-added.	100 million CNY
lndig	Natural logarithm of the digitization level index for digital financial inclusion.	Logarithm of index score
green_patent	Number of granted green patents.	Count
year	Year of observation.	Calendar year
lnedutech	Natural logarithm of education/science expenditure.	Logarithm of 10 thousand CNY
lnsequ	Natural logarithm of estimated carbon sequestration.	Logarithm of metric tons
sequestration	Carbon sequestration (estimated CO_2 absorption).	Metric tons
lnpop	Natural logarithm of total resident population.	Logarithm of 10 thousand persons
lnprimaryschool	Natural logarithm of the number of primary schools.	Logarithm of count
lnFDI	Natural logarithm of foreign direct investment inflow.	Logarithm of 10 thousand USD
lnpgdp	Natural logarithm of per capita gross domestic product.	Logarithm of CNY per capita
latitude	Geographic latitude coordinate of the city's administrative center.	Degrees (°)
bus	Number of buses.	Vehicles
lnbook	Natural logarithm of public library book collections.	Logarithm of 10 thousand volumes
population	Total resident population of the city.	10 thousand persons
lnroad	Natural logarithm of road length, indicating transportation infrastructure scale.	Logarithm of kilometers (km)
GDP	Gross domestic product, the total market value of goods and services produced.	10 thousand CNY
lnGDP	Natural logarithm of GDP.	None

Each dataset was divided into training and test subsets in a 4:1 ratio using the Kennard–Stone algorithm (Yu, 2020a; Yu, 2025a). Data Set I comprises 3,260 samples in total, with 2,608 (80%; Nos. 1-2608 in Table S1) allocated to the training set and 652 (20%; Nos. 2609-3260) to the test set. Similarly, Data Set II contains 1,988 samples, of which 1,590 (80%; Nos. 1-1590 in Table S2) form the training set, and 398 (20%; Nos. 1591-1988) form the test set. Descriptive statistics for both datasets are provided in Table S3 of the Supporting Information. Model development was conducted through 10-fold cross-validation on the training sets to build RF and CNN models, while the held-out test sets were used for external validation of predictive performance.

2.2. Random Forest Algorithm

The RF algorithm was employed to model the complex nonlinear relationships between socioeconomic drivers and CO_2 emissions. As an ensemble learning method, RF improves predictive accuracy and mitigates overfitting by combining predictions from multiple decision trees, each constructed using bootstrapped samples and a random subset of features for node splitting (Fang et al., 2022; Yu & Zeng, 2022; Yu et al., 2025b).

To adapt the RF model to the differing data characteristics at the two analytical scales, we performed separate hyperparameter tuning for each dataset.

For Data Set I, a grid search was carried out across the following ranges:

mtry (number of features considered at each split): 2 to 4

ntree (number of trees in the forest): 100 to 600 in steps of 100

The minimum node size was fixed at 5 to avoid excessive granularity.

For Data Set II, which has higher feature dimensionality, an expanded search space was defined:

ntree: 100 to 600 (increments of 100)

mtry: 2 to 22

k_features (number of key predictors retained): 10 to 23

The node size was set to 5 again.

Model selection for both datasets was rigorously conducted using 10-fold cross-validation on the training subset. The hyperparameter configuration that achieved the highest cross-validated coefficient of determination (R^2) was selected as the final optimal model. The complete procedure for RF modeling, combined with grid search and cross-validation, for Data Set I is outlined in Algorithm 1.

Algorithm 1: Random forest for Data Set I.

A. Initialization

Set random seed = 42.

Load data:$XX = \mathrm{NUM}(1{:}\,3260, 1{:}\,5)$.

Split data: training set (samples 1–2608), test set (samples 2609–3260).

B. Grid search over hyperparameters

Define search ranges:

$ntree \in \{100,200,300,400,500,600\}$

$mtry \in \{2,3,4\}$

$var\_num \in \{2,3,4\}$

For each combination $(ntree, mtry, var\_num)$:

(1). Select the first var_num features.

(2). Perform 10-fold cross-validation on the training set:

For each fold, train RF with current $ntree$ and $mtry$.

Predict on validation fold and compute metrics (R^2, RMSE, MAE).

(3). Compute average CV metrics.

C. Best model selection

Choose combination $(ntree, mtry, var\_num)$ with highest average CV R^2.

D. Final model training & evaluation

Train RF on full training set using best hyperparameters.

Evaluate on the test set and record final metrics.

Generate visualizations:

E. Results saving

Save best model, hyperparameters, and all performance metrics.

Return best hyperparameters, final RF model, and performance metrics.

For the RF models, the analysis was implemented via the 'randomForest' R package (accessible at https://www.stat.berkeley.edu/~breiman/RandomForests/) within the MATLAB R2024a environment.

2.3. Convolutional Neural Network

In addition to RF, we employed a CNN approach for CO_2 emission prediction. The CNN architecture leverages hierarchical feature learning by treating socioeconomic variables as one-dimensional sequences. Through stacked convolutional layers, the network automatically extracts local patterns and higher-order interactions from the input features without requiring manual feature engineering. Each convolutional block in our implementation consists of convolution, batch normalization, ReLU activation, and dropout operations, enabling the model to learn robust representations while mitigating overfitting (Tinkov et al., 2021; Yu et al., 2026).

Our CNN implementation incorporates several key design elements to optimize performance: (1) automated feature selection through 10-fold cross-validation to identify the most informative predictor subset (10–23 features); (2) a global average pooling layer followed by fully connected layers for regression prediction; and (3) an ensemble strategy employing 10 CNN models with varied architectures for enhanced predictive stability. The models were trained using Adam optimization with piecewise learning rate decay and robust median-IQR normalization, creating an end-to-end learning pipeline for accurate CO_2 emission forecasting. The detailed algorithmic process of Data Set II is as in Algorithm 2.

Algorithm 2: 1D-CNN for Data Set II.
a. Initialization
Set random seed = 42.
Split data: training set (samples 1–1590), test set (samples 1591–1988).
b. Preprocessing
For each feature, calculate IQR bounds and cap outliers in both training and test sets.
c. Feature selection (10-fold CV)
For $k = 10$ to 23:
Select first k features.
Perform 10-fold CV:
Standardize fold data (median & IQR).
Build CNN (2 conv layers, 2 FC layers).
Train for 100 epochs (Adam, LR = 0.001, piecewise decay).
Record validation R^2.
Compute average R^2 across folds.
Choose k^* with max average R^2
d. Final model training
Select first k^* features and standardize.
Build enhanced CNN (3 conv layers, 3 FC layers).
Train for 300 epochs (Adam, LR = 0.001, early-stopping).
e. 10-Model ensemble
Train 10 CNN variants with randomized architectures
Aggregate predictions via averaging
f. Evaluation
Compute R, R^2, RMSE for single model (training/test) and ensemble model (test).
Plot predicted-vs-actual and feature-selection results.
Return k^* trained models and performance metrics.

The CNN architectures were constructed and trained using the Deep Learning Toolbox within MATLAB R2024a.

2.4. Statistical Parameters for Models

The statistical parameters, R, R^2, RMSE, and MAE, were calculated as follows (Roy et al., 2016; 2017; Yu et al., 2026):

$$R = \frac{\sum_{i=1}^{n}\left((y_i-\bar{y})(\tilde{y}_i-\bar{\tilde{y}})\right)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2 \cdot \sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})^2}} \quad (1)$$

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

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(y_i-\tilde{y}_i)^2}{n}} \quad (3)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \quad (4)$$

Here y_i is the observed value of the ith sample; $\tilde{y}_i$ is the predicted value of a model, like an RF or CNN model; $\bar{y}$ is the mean observed value; $\bar{\tilde{y}}$ is the mean predicted value.

Successful models should satisfy the following criteria for the test sets (Roy et al., 2016; 2017; X. Yu et al., 2024; Yu, 2025c):

$$0.85 \le k = \frac{\sum y_i\tilde{y}_i}{\sum {y_i}^2} \le 1.15 \quad (5)$$

$$0.85 \le k' = \frac{\sum y_i\tilde{y}_i}{\sum {\tilde{y}_i}^2} \le 1.15 \quad (6)$$

$$0.5 < q_{ext}^2 = 1 - \frac{\sum(y_i-\tilde{y}_i)^2}{\sum(y_i-\bar{y}_{tra})^2} \quad (7)$$

$$(R^2 - R_0^2)/R^2 < 0.1 \text{ or } (R^2 - {R'}_0^2)/R^2 < 0.1 \quad (8)$$

Here k and k' are the slopes of the regression lines, $\bar{y}_{tra}$ is the average value of the training set, q^2_{ext} is the external validation coefficient, R^2_0 and R'^2_0 are the validation determination coefficients. These were calculated using the following expressions:

$$R^2_0 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2} \tag{9}$$

$$R'^2_0 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \tag{10}$$

Here $\tilde{y}_i^{r_0} = ky$ and $y_i^{r_0} = k'\tilde{y}$.

According to Roy et al. (2016), a QSPR model may be considered to possess good predictive performance provided that it meets the two criteria:

$$\text{MAE} \leq 0.1 \times \text{training set range and MAE} + 3 \times \sigma \leq 0.2 \times \text{ training set range} \tag{11}$$

Here, MAE is mean absolute error (MAE) and σ denotes the standard deviation of the MAE values for the test set data.

In the analysis of applicability domains for QSPR models, the leverage values for all samples were calculated using Eq. (12):

$$h_i = x_i^T (X^T X)^{-1} x_i \tag{12}$$

Where x_i denotes the feature vector of the ith sample, and X is the $n \times p$ data matrix. The warning leverage h^* was calculated with Eq. (13):

$$h^* = 3 \times (p+1)/n \tag{13}$$

Here n is the number of samples and p is the number of features.

3. Results and Discussion

3.1. RF Model I

Based on the CO_2 emission dataset (Data Set I) compiled by Dharmapriya et al. (2025), which comprises 3260 observations from 163 countries spanning 2000 to 2019, a RF regression model was developed to predict per capita carbon emissions (CE). The predictor variables included trade openness (TO), per capita gross domestic product (GDP), per capita energy consumption (EC), and year, yielding four input features.

To identify the optimal model configuration, an exhaustive grid search was performed over three hyperparameters. The search space was defined as follows: the number of features to utilize (N_feature = 2:4), the number of variables randomly sampled at each split (mtry = 2:4), and the number of trees (ntree = 100:100:600). The minimum node size was fixed at 5 in all runs. Respecting the constraint that mtry cannot exceed the number of available features, a total of 36 unique parameter combinations were evaluated. Each combination was assessed using 10-fold cross-validation on the training set comprising 2608 samples, with performance evaluated via the coefficient of determination (R^2).

The optimal model (denoted RF Model I), selected based on the highest cross-validated R^2, was configured with N_feature = 4, ntree = 600, and mtry = 2. Its cross-validation performance on the training set was: $R^2 = 0.945$, RMSE = 1.539, and MAE = 0.777. When retrained on the full training set with these hyperparameters, the final model achieved a training-set performance of R = 0.995, $R^2 = 0.989$, RMSE = 0.684, and MAE = 0.330.

On the independent test set (652 samples), RF Model I achieved the following performance: R = 0.993, $R^2 = 0.984$, RMSE = 0.427, and MAE = 0.197. The detailed prediction results are provided in Supplementary Table S1 and visualized in Figure 1. The model satisfies all recommended statistical validation criteria for predictive reliability (X. Yu et al., 2024; Yu, 2025c).

$$0.85 \leq k = 0.986 \leq 1.15 \text{ or } 0.85 \leq k' = 1.003 \leq 1.15 \tag{14}$$

$$q^2_{ext} = 0.993 > 0.5 \tag{15}$$

$$R^2 = 0.984 > 0.6 \tag{16}$$

$$\left|R^2 - R^2_0\right| / R^2 = |0.984 - 0.986| / 0.984 = 0.002 < 0.1 \tag{17}$$

Furthermore, with a training-set range of 62.240 (Table S1), the test-set MAE (0.185) and its standard deviation ($\sigma = 0.358$) meet the additional requirements (Roy et al., 2016):

$$\text{MAE} = 0.185 < 6.224\ (0.1 \times 62.240) \tag{18}$$

$$\text{MAE} + 3\sigma = 0.185 + 3\times0.358 = 1.259 < 12.448\ (0.2 \times 62.240) \tag{19}$$

These validation outcomes confirm that RF Model I delivers reliable and accurate predictions for CO_2 emissions.

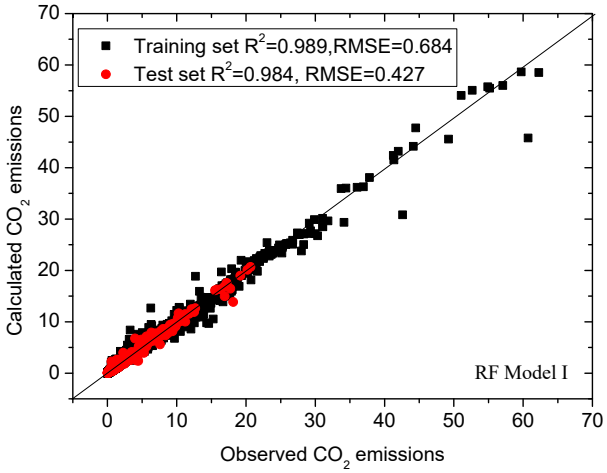

Fig. 1. Plot of observed versus calculated CE values with RF Model I

3.2. RF Model II

A second random forest model (RF Model II) was developed to predict CO_2 emissions based on the 23 socio-economic features identified for 284 Chinese prefecture-level cities (Wang et al., 2022). To identify the optimal configuration, a comprehensive grid search was conducted over 1764 unique hyperparameter combinations (ntree: 100-600, mtry: 2-22, k_features: 10-23). Using 10-fold cross-validation, the best-performing model was selected with ntree = 300, mtry = 16, and 21 retained features, achieving a cross-validated R^2 of 0.912.

When retrained on the full training set, the final model exhibited excellent predictive accuracy. On the training set, the performance metrics were: R = 0.995, R^2 = 0.988, RMSE = 0.081, and MAE = 0.055. On the independent test set, the model attained R = 0.932, R^2 = 0.860, RMSE = 0.225, and MAE = 0.153. These results confirm the model's high accuracy in estimating CO_2 emissions across the prefecture-level cities (Li et al., 2022). The detailed prediction results are listed in Supplementary Table S2 and visually summarized in Figure 2.

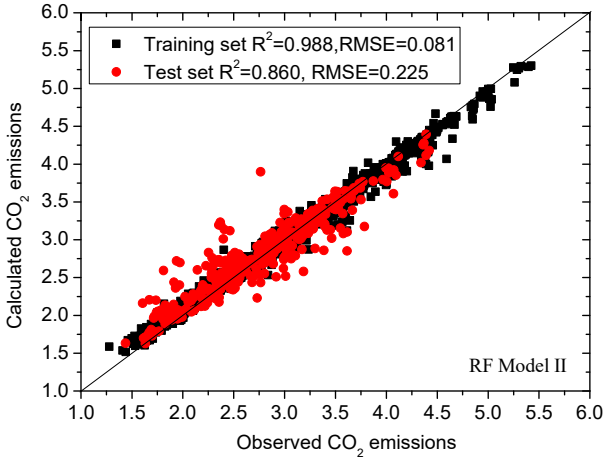

Fig. 2. Plot of observed versus calculated CE values with RF Model II

3.3. CNN Model I

A one-dimensional convolutional neural network (1D-CNN) ensemble model (CNN Model I) was constructed to predict CO_2 emissions using Data Set I. The model employed the four key features identified previously: trade openness (TO), year, per-capita GDP, and per-capita energy consumption (EC). Ten-fold

cross-validation on the training set (2608 samples) confirmed that using all four features yielded the best cross-validated performance ($R = 0.947$, $R^2 = 0.890$, RMSE = 0.271).

The network comprised three convolutional blocks (kernel size = [1, 2]) with filter counts of 128-64, 256-128, and 512-256, each followed by batch normalization, ReLU activation, and dropout (rates 0.1, 0.15, 0.2). After global average pooling, a deep fully-connected network (1024, 512, 256, 128, 64, 32, 16 neurons) was applied, with batch normalization and dropout (0.25, 0.25, 0.2, 0.15) in the initial layers. Training was conducted for 500 epochs using the Adam optimizer (initial learning rate = 0.001, piecewise decay factor = 0.7 every 100 epochs), L_2 regularization ($\lambda = 0.0005$), and a mini-batch size of 32.

To enhance robustness, an ensemble of 10 base CNNs was trained, each on a different 90 % subset of the training data. The ensemble model delivered strong performance: on the training set, $R = 0.974$, $R^2 = 0.946$, RMSE = 1.295, MAE = 0.855; on the independent test set, $R = 0.981$, $R^2 = 0.954$, RMSE = 0.726, MAE = 0.546. Detailed predictions are listed in Table S1 and visualized in Figure 3.

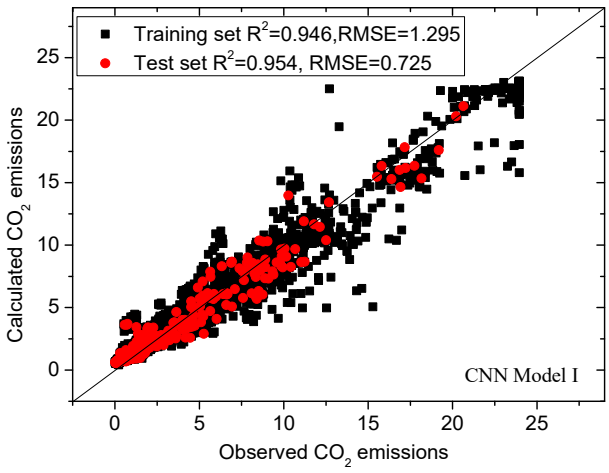

Fig. 3. Plot of observed versus calculated CO_2 emissions with CNN Model I

3.4. CNN Model II

A second 1D-CNN ensemble model (CNN Model II) was developed for CO_2 emission prediction using Data Set II (284 Chinese prefecture-level cities). Through 10-fold cross-validation on the training set (1 590 samples), the optimal number of retained features was determined to be 21, selected from the original 23 candidate variables (cross-validation performance: $R = 0.939$, $R^2 = 0.871$, RMSE = 0.266).

The network comprised three convolutional blocks with filter counts of 128, 64, and 32 (kernel size = [1, min(3, 21)]), each followed by batch normalization, ReLU activation, and dropout (rates 0.3, 0.25, 0.2). After global average pooling, a fully-connected network (64, 32, 16 neurons) was applied, with batch normalization and dropout (0.15, 0.1). The model was trained for up to 300 epochs using the Adam optimizer (initial learning rate = 0.001, piecewise decay factor = 0.7 every 40 epochs), L_2 regularization ($\lambda = 0.001$), and a mini-batch size of 32, with early stopping (patience = 15).

To enhance robustness, an ensemble of 10 base CNNs was constructed, each with randomized filter counts (±20 % variation) and dropout rates. The single CNN model achieved $R = 0.989$, $R^2 = 0.967$, RMSE = 0.136 on the training set, and $R = 0.938$, $R^2 = 0.861$, RMSE = 0.223 on the test set (398 samples). The ensemble model improved these results to $R = 0.992$, $R^2 = 0.975$, RMSE = 0.118 (training) and $R = 0.952$, $R^2 = 0.890$, RMSE = 0.199 (test). Detailed prediction results are provided in Table S2 and illustrated in Figure 4.

The model satisfies all recommended validation criteria (X. Yu et al., 2024; Yu, 2025c):

$$0.85 \leq k = 0.985 \leq 1.15 \text{ or } 0.85 \leq k' = 1.010 \leq 1.15 \tag{20}$$

$$q_{ext}^2 = 0.929 > 0.5 \tag{21}$$

$$R^2 = 0.954 > 0.6 \tag{22}$$

$$\left|R^2 - R_0'^2\right|/R^2 = (0.954 - 0.881)/0.954 = 0.077 < 0.1 \tag{23}$$

Furthermore, with a training-set range of 4.147, the test-set MAE (0.148) and its standard deviation ($\sigma = 0.133$) meet (Roy et al., 2016):

$$\text{MAE} = 0.148 \leq 0.1 \times 4.147\ (0.415) \tag{24}$$

$$MAE + 3 \times \sigma = 0.148 + 3 \times 0.133 = 0.547 \leq 0.2 \times 4.147\ (0.829) \quad (25)$$

These results confirm that CNN Model II delivers reliable and accurate predictions for CO_2 emissions.

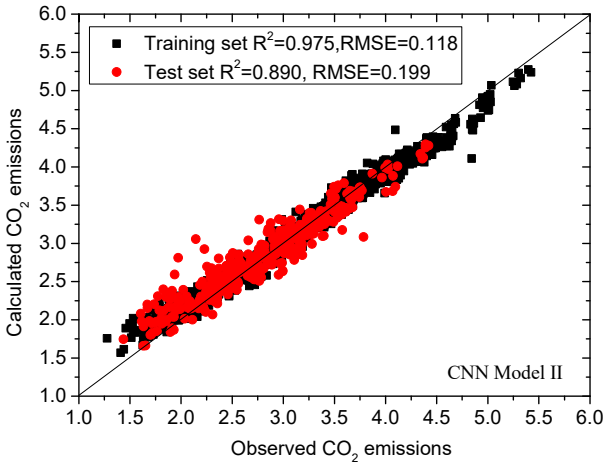

Fig. 4. Plot of observed versus calculated CO_2 emissions with CNN Model II

3.5. Application Domain of Models

The applicability domain of RF Model I was analyzed using the Williams plot (see Figure 5) (Yu et al., 2025b; Yu, 2023). The leverage values for all samples were calculated using Eq. (12), and the warning leverage h* was determined as 0.00575 by Eq. (13) (with p=4 features and n=2608 training samples). In the training set, 52 samples were identified as outliers with |sr| > 3, indicating that the model accurately fitted 98.0% of the training data. For the test set, only two samples (No. 2724 and No. 2904) had |sr| > 3, corresponding to a high prediction accuracy of 99.7%. Additionally, 101 samples (with |sr|<3) exhibited high leverage (h>h*), suggesting that these points are structurally distant from the training set (Yu, 2020b; Yu, 2021). The accurate predictions for these high-leverage points demonstrate the robust extrapolation capability of RF Model I beyond its training domain.

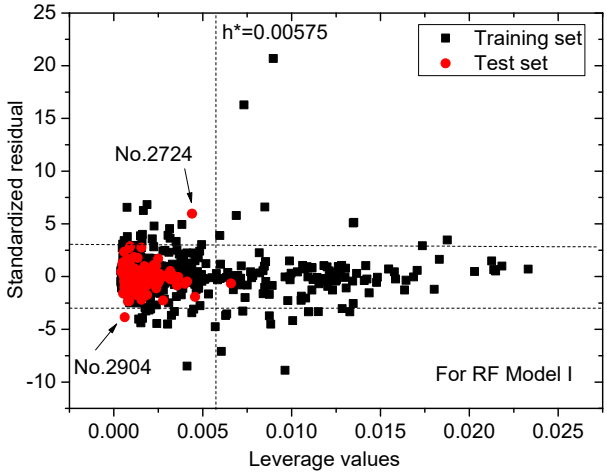

Fig. 5. Williams plot for standardized residual and leverage values for RF Model I

The applicability domain of CNN Model II was similarly evaluated using the Williams plot approach (see Figure 6). In the training set, 19 samples were identified as outliers with |sr|>3, indicating that the model accurately predicted 98.8% of the training samples. For the test set, 30 samples (e.g., Nos. 1853, 1913, 1962, and 1963) exhibited |sr|> 3, corresponding to a prediction accuracy of 92.5%. Furthermore, 39 samples displayed high leverage (h>h*=0.045). Among these high-leverage points, only three samples (No. 9, 1675, and 1827) showed |sr|>3, while the remaining 36 samples had acceptable residuals (|sr|<3). These results demonstrate that CNN Model II also possesses robust extrapolation capability beyond its immediate training domain.

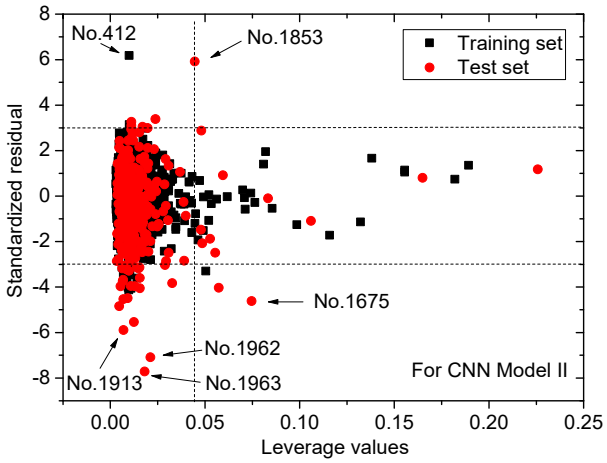

Fig. 6. Williams plot for standardized residual and leverage values for CNN Model II

3.6. Model Comparison

Based on the performance metrics summarized in Table 3, a comparative analysis was conducted between the RF and CNN models for the two distinct datasets. For Data Set I, which contains 3,260 samples from 163 countries, RF Model I demonstrated superior overall predictive performance. On the independent test set (652 samples), RF Model I achieved substantially higher R^2 (0.984) and lower RMSE (0.427) than CNN Model I (R^2 = 0.954, RMSE = 0.726). This indicates that the RF algorithm captured the relationships between the four key global drivers—TO, Year, GDP, and EC—and per capita CO_2 emissions more accurately and with smaller error. The performance difference is also evident in the training set metrics: RF Model I showed a near-perfect fit (R^2 = 0.989, RMSE = 0.684), whereas the ensemble training performance of CNN Model I was lower (R^2 = 0.946, RMSE = 1.295).

For the more complex, high-dimensional Data Set II comprising 284 Chinese cities (with 23 initial features), a different pattern emerged. CNN Model II exhibited better generalization performance on unseen test data. Although both models performed exceptionally well on the training set, CNN Model II attained a higher test-set R^2 (0.890) and a lower test-set RMSE (0.199) than RF Model II (R^2 = 0.860, RMSE = 0.225). This suggests that, for capturing intricate, nonlinear patterns in sub-national urban data—including structural, economic, technological, and policy variables—the hierarchical feature extraction of the CNN architecture offered a slight advantage in generalizing to new cities. It is noteworthy that RF Model II showed a larger drop in performance from training (R^2 = 0.988) to test (R^2 = 0.860) than CNN Model II (training R^2 = 0.975, test R^2 = 0.890), reflecting greater sensitivity to data complexity under these conditions. Furthermore, all models in this study (test set R^2: 0.860–0.984) show accuracy comparable to the result reported by Jabeur et al. (2022) (R^2 = 0.907) on a different dataset (310 samples in the test set).

In conclusion, the choice of optimal model depends on data scale and feature characteristics. For global, macro-level prediction with a small set of strong predictors, the RF model (RF Model I) proved more accurate and reliable. For sub-national, urban-scale prediction involving a larger and more diverse set of features that describe complex socioeconomic systems, the CNN model (CNN Model II) demonstrated marginally stronger generalization performance, delivering results competitive with state-of-the-art benchmarks.

Table 3. Model comparisons in predicting CO_2 emissions

Model	Training set			Test set			Ref.
	n	R^2	RMSE	n	R^2	RMSE	
RF Model I	2608	0.989	0.684	652	0.984	0.427	This work
CNN Model I	2608	0.946	1.295	652	0.954	0.726	This work
RF Model II	1590	0.988	0.081	398	0.860	0.225	This work
CNN Model II	1590	0.975	0.118	398	0.890	0.199	This work
NGBoost	/	/	/	310	0.907	/	Jabeur et al. 2022

3.7. Mechanism Analysis

To quantitatively interpret the models' predictions and assess the contribution of each input feature, SHAP analysis was conducted (Yu et al., 2026). SHAP values, derived from cooperative game theory, provide a consistent and locally accurate measure of a feature's influence on a given prediction (Lundberg and Lee, 2017).

A higher mean absolute SHAP value indicates that, on average, the feature causes a larger change in the model's predicted CO_2 emissions, thus playing a more significant role in the model's decision-making process. The sign of the mean SHAP value indicates the net directional influence of a feature: a positive value suggests that the feature tends to increase the predicted CO_2 emission value, whereas a negative value suggests a tendency to decrease it. The SHAP analysis was performed using the TreeBagger function in MATLAB R2024a for Data Set I and Data Set II. Informed by the theoretical and empirical literature (Dharmapriya et al., 2025; Wang et al., 2022), the mechanisms linking each feature to CO_2 emissions are detailed in Table 4 (RF Model I) and Table 5 (CNN Model II).

Table 4. Features in RF Model I and their mean SHAP values and link to CO_2 emissions

Feature	SHAP	Mechanism linking to CO_2 emissions
TO	0.246	Trade openness affects emissions through economic scale, industrial structure, and technology transfer. Its impact varies by country, making it a key factor explaining emission differences.
year	0.005	The year represents long-term global trends, such as technological progress and climate policies, that gradually affect all countries' emissions.
GDP	0.201	GDP indicates economic scale. Higher economic activity generally requires more energy and production, leading to higher CO_2 emissions.
EC	-0.167	In reality, energy consumption increases CO_2 emissions, but in this model it has a small negative effect. This suggests that, after accounting for trade and wealth, the model uses EC to fine-tune predictions downward rather than treating it as the primary factor behind country differences.

Table 5. Features in Data Set II and their mean SHAP values and link to CO_2 emissions

Feature	SHAP	Mechanism linking to CO_2 emissions
GDP	0.0246	Economic growth is a primary driver of industrial output, energy consumption, and economic activity, which directly increase CO_2 emissions. This aligns with the paper's finding that economic growth is a significant transmission channel for emissions.
lnGDP	0.0227	The log-transformed GDP per capita captures the scale effect of economic development. Higher income levels are consistently associated with greater production and consumption, leading to elevated emissions.
population	-0.0245	While a larger population increases aggregate demand, the negative association in this model may reflect that, after controlling for economic output, factors such as urban form efficiency, shared infrastructure, or the stage of development associated with population metrics are associated with lower emission intensity.
latitude	-0.018	Geographic latitude may serve as a proxy for regional characteristics such as climate, industrial composition, energy mix (e.g., heating needs), and policy enforcement, which collectively correlate with lower-than-baseline emissions in this model's context.
lnroad	-0.0097	Road infrastructure density may indicate the maturity of the transport network. A negative link could suggest associations with efficient logistics, reduced congestion, or correlation with more advanced economic structures that have lower emission footprints.
lnpop	-0.0085	Log population density. High density can enable efficiencies in public services, transportation, and energy use, potentially leading to lower per-capita emissions, which is reflected as a negative contribution to the model's prediction.
lnpGDP	-0.0078	This feature, potentially a per capita or sectoral GDP measure, shows a negative association. It may indicate that, when other factors are controlled, specific aspects of income growth are linked to less emission-intensive development paths in the model.

Table 5. cont.

Feature	SHAP	Mechanism linking to CO_2 emissions
lnprimaryschool	-0.0086	The density of primary schools may proxy for social infrastructure and urban residential development. Its negative link might correlate with community-oriented, potentially less industrial urban forms, which are associated with lower emissions.
lnusg	-0.007	Likely related to urban green space or specific land use. Greater green space is associated with environmental quality and planning, aligning with a negative contribution to emission predictions.
sequestration	-0.0064	Direct carbon sequestration capacity (e.g., forests). Higher sequestration directly reduces net emissions, contributing to lower predicted values.
lnsequ	-0.0063	Log-transformed sequestration. Same mechanism as above, indicating the importance of natural or managed carbon sinks.
lnIns	-0.0033	Industrial structure (tertiary/secondary sector ratio). A higher ratio (more services) is strongly associated with lower CO_2 emissions, as confirmed by the paper's empirical results. This is a key structural factor in reducing emissions.
lnedutech	-0.0032	Investment in education and technology. This fosters human capital and innovation, which can promote energy efficiency and cleaner production methods, reducing emissions.
lnthird	-0.0032	Share of the tertiary (service) sector. Services are typically less energy- and carbon-intensive than heavy industry, leading to lower emissions.
second	-0.0031	Share of the secondary (industrial) sector. Its negative SHAP is notable as industry is emission-intensive. It may reflect that within the model's context, after accounting for total output, a higher industrial share is associated with specific, regulated production clusters that are relatively efficient compared to baseline scenarios.
lnFDI	-0.0047	Foreign Direct Investment. May facilitate technology transfer and introduce higher environmental standards, potentially leading to cleaner production and a negative association with emissions.
bus	-0.0014	Public bus infrastructure. Promotes public transport use, potentially reducing emissions from private vehicles, though its impact appears minor in this model.
ER	-0.0016	Environmental Regulation stringency. Designed to curb pollution, thus exerting downward pressure on emissions, though its direct measured effect here is small.
green_patent	-0.0009	Count of green patents. Represents innovation in environmental technology, which should mitigate emissions, hence the negative contribution.
lndig	-0.0007	Digitization level of DFI. The paper found its direct effect on emissions was statistically insignificant. The negligible SHAP suggests a very weak or complex relationship in this ML model.
lnbook	0.0013	Could proxy for cultural activity or a specific sector. Its minimal positive link might indicate a slight association with consumption patterns that marginally increase emissions.
coverage_breadth	0.0005	Breadth of Digital Financial Inclusion (DFI) coverage. The paper finds DFI increases local emissions. Broader access can stimulate consumption and business activity, potentially raising energy use. Its small SHAP suggests it's a minor direct driver in this specific model.
year	-0.0006	Temporal trend. Captures the net effect of technological progress, policy changes, and efficiency gains over time (2011–2017), applying a slight downward pressure on emission predictions.

The SHAP analysis reveals distinct driver profiles for the two models, reflecting their different analytical scopes. In the global, country-level Data Set I, trade openness (TO) and GDP are the dominant positive drivers, while energy consumption (EC) shows a counter-intuitive negative mean SHAP, acting as a moderating factor

within the model's specific framework. Conversely, in the high-dimensional Data Set II for Chinese cities, economic metrics (GDP, lnGDP) remain key positive drivers. Still, a suite of structural and urban form factors (e.g., industrial structure lnIns, carbon sequestration, population density) show significant negative contributions, highlighting their contextual mitigation roles. In contrast, policy and innovation features exhibit minimal direct explanatory power.

4. Conclusions

This study rigorously compared RF and CNN models for predicting CO_2 emissions across different geographical scales. RF excelled in modeling global, macro-level emissions with a limited set of strong predictors, achieving high accuracy ($R^2 = 0.984$) and robustness. In contrast, CNN demonstrated stronger generalization for complex, high-dimensional urban data, effectively capturing nonlinear interactions among socioeconomic, structural, and policy variables. External validation confirmed the models' reliability beyond training data, addressing a key gap in prior research. SHAP analysis provided interpretable insights into feature contributions, highlighting distinct drivers at country and city levels. These findings underscore the context-dependent performance of machine learning models and offer validated, scalable tools for emission forecasting. Future work may integrate temporal dynamics, expand feature sets, and apply hybrid modeling to further enhance predictive accuracy and policy relevance.

Author contributions

Ningyao Yu: Conceptualization, data curation; methodology, software, writing-original draft.

Declarations

Conflict of interest: The authors have no relevant financial or non-financial interests to disclose.

Data Availability Statement

Additional supporting information can be found online in the Supporting Information section, including Table S1–S3.

Human and Animal Rights

This article does not contain any studies with human or animal subjects.

References

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. *Machine Learning with Applications*, *6*, 100094.

Aryai, V., & Goldsworthy, M. (2023). Day ahead carbon emission forecasting of the regional National Electricity Market using machine learning methods. *Engineering Applications of Artificial Intelligence*, *123*, 106314.

Balducci, F., Impedovo, D., & Pirlo, G. (2018). Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. *Machines*, *6*, 38.

Begum, A.M., & Mobin, M.A. (2025). A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. *Scientific Reports*, *15*, 19469.

Bhatt, H., Davawala, M., Joshi, T., Shah, M., & Unnarkat, A. (2023). Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques. *Cleaner Chemical Engineering*, *5*, 100095.

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine Learning for Molecular and Materials Science. *Nature*, *559*(7715), 547–555.

Cao, X., Zhang, Y., Sun, Z., Yin, H., & Feng, Y. (2026). Machine Learning in Polymer Science: A New Lens for Physical and Chemical Exploration. *Progress in Materials Science*, *156*, 101544.

Dai, X., He, Y., & Zhong, Q. (2015). Analysis of CO_2 emission driving factors in China's agriculture based on expanded Kaya identity. *Journal of University of Chinese Academy of Sciences*, *32*(6), 751–759.

Dharmapriya, N., Gunawardena, V., Methmini, D., Jayathilaka, R., & Rathnayake, N. (2025). Carbon emissions across income groups: exploring the role of trade, energy use, and economic growth. *Discover Sustainability*, *6*, 621.

Fan, Z., Zhao, M., Gong, W., & Wang, C. (2020). Influencing factors and peak values of carbon emission based on STIRPAT model. *Journal of Low Carbon Economy*, *9*(2), 100–110.

Fang, Z., Yu, X., & Zeng, Q. (2022). Random forest algorithm-based accurate prediction of chemical toxicity to Tetrahymena pyriformis. *Toxicology*, *480*, 153325.

Jabeur, S.B., Ballouk, H., Arfi, W.B., & Khalfaoui, R. (2022). Machine learning-based modeling of the environmental degradation, institutional quality, and economic growth. *Environmental Modeling & Assessment*, *27*(6), 953–966.

Jiang, P., Gong, X., Yang, Y., Tang, K., Zhao, Y., Liu, S., & Liu, L. (2023). Research on spatial and temporal differences of carbon emissions and influencing factors in eight economic regions of China based on LMDI model. *Scientific Reports*, *13*, 7965.

Kumari, S., & Singh, S.K. (2023). Machine learning-based time series models for effective CO_2 emission prediction in India. *Environmental Science and Pollution Research*, *30*(55), 116601–116616.

Leerbeck, K., Bacher, P., Junker, R., Goranovi´c, G.G., Corradi, O., Ebrahimy, R., Tveit, A., & Madsen, H. (2020). Short-term forecasting of CO_2 emission intensity in power grids by machine learning. *Applied Energy*, *277*, 115527.

Li, X., Ren, A., & Li, Q. (2022). Exploring patterns of transportation-related CO_2 emissions using machine learning methods. *Sustainability*, *14*(8), 4588.

Lundberg, S.M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4768–4777.

Maji, D., Shenoy, P., & Sitaraman, R.K. (2023). Multi-day forecasting of electric grid carbon intensity using machine learning. *ACM SIGEnergy Energy Informatics Review*, *3*(2), 19–33.

Mekki, Y., Moujahdi, C., Assad, N., & Dahbi, A. (2025). Enhancing CO_2 emissions predictions through historical events-aware artificial intelligence models. *International Journal of Environmental Science and Technology*, *22*, 15289–15308.

Noé, F., Tkatchenko, A., Müller, K.R., & Clementi, C. (2020). Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry*, *71*, 361–390.

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, *4*, 19-29.

Roy, K., Das, R.N., Ambure, P., & Aher, R.B. (2016). Be Aware of Error Measures: Further Studies on Validation of Predictive QSAR Models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18–33.

Roy, K., Ambure, P., & Aher, R.B. (2017). How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemometrics and Intelligent Laboratory Systems*, *162*, 44–54.

Sun, W., Wang, C., & Zhang, C. (2017). Factor analysis and forecasting of CO_2 emissions in Hebei, using extreme learning machine based on particle swarm optimization. *Journal of Cleaner Production*, *162*, 1095–1101.

Tinkov, O.V., Grigorev, V.Y., & Grigoreva, L.D. (2021). QSAR analysis of the acute toxicity of avermectins towards Tetrahymena pyriformis. *SAR and QSAR in Environmental Research*, *32*(7), 541–571.

Uncuoglu, E., Citakoglu, H., Latifoglu, L., Bayram, S., Laman, M., Ilkentapar, M., & Oner, A.A. (2022). Comparison of Neural Network, Gaussian Regression, Support Vector Machine, Long Short-Term Memory, Multi-Gene Genetic Programming, and M5 Trees Methods for Solving Civil Engineering Problems. *Applied Soft Computing*, *129*, 109623.

Wang, X., Wang, X., Ren, X., & Wen, F. (2022). Can digital financial inclusion affect CO_2 emissions of China at the prefecture level? Evidence from a spatial econometric approach. *Energy Economics*, *109*, 105966.

Wu, F., Zhang, X., Fang, Z., & Yu, X. (2023). Support Vector Machine-Based Global Classification Model of the Toxicity of Organic Compounds to Vibrio fischeri. *Molecules*, *28*, 2703.

Xia, Y., Jiang, L., Wang, L., Chen, X., Ye, J., Hou, T., Wang, L., Zhang, Y., Li, M., Li, Z., Song, Z., Jiang, Y., Liu, W., Li, P., Rosenfeld, D., Seinfeld, J.H., & Yu, S. (2022). Rapid assessments of light-duty gasoline vehicle emissions using on-road remote sensing and machine learning. *Science of the Total Environment*, *815*, 152771.

Yalçın, G., Bayram, S., & Çıtakoğlu, H. (2024). Evaluation of Earned Value Management-Based Cost Estimation via Machine Learning. *Buildings*, *14*(12), 3772.

Yu, X. (2020a). Prediction of chemical toxicity to Tetrahymena pyriformis with four descriptor models. *Ecotoxicology and Environmental Safety*, *190*, 110146.

Yu, X. (2020b). Quantitative structure-toxicity relationships of organic chemicals against Pseudokirchneriella subcapitata. *Aquatic Toxicology*, *224*, 105496.

Yu, X. (2021). Support vector machine-based model for toxicity of organic compounds against fish. *Regulatory Toxicology and Pharmacology*, *123*, 104942.

Yu, X., & Zeng, Q. (2022). Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes. *Aquatic Toxicology*, *251*, 106265.

Yu, X. (2023). QSPR-based model extrapolation prediction of enthalpy of solvation. *Journal of Molecular Liquids*, *376*, 121455.

Yu, X., Zhang, Z., & Wang, H. (2024). Global classification model for acute toxicity of organic compounds towards Tetrahymena pyriformis. *Process Safety and Environmental Protection*, *192*, 1221–1227.

Yu, Y., Chen, Q., Zhi, J., Yao, X., Li, L., & Shi, C. (2024). Carbon peak prediction in China based on Bagging-integrated GA-BiLSTM model under provincial perspective. *Energy*, *313*, 133519.

Yu, Z., Meng, L., Han, X., & Wang, S. (2024). Research on carbon emission peak prediction in China-Based on STIRPAT model and driving factor decomposition analysis. *Advances in Applied Mathematics*, *13*(7), 3442–3455.

Yu, X. (2025a). Quantitative structure-property relationship of glass transition temperatures for organic compounds. *Molecular Physics*, *123*, e2413005.

Yu, X., Fang, Z., & Wu, F. (2025b). Random Forest based approach for predictions of glass transition temperatures in polymers. *Polymer Engineering & Science*, *65*, 6238–6247.

Yu, X. (2025c). Predicting chemical toxicity towards Raphidocelis subcapitata with quantum chemical descriptors. *Algal Research*, *89*, 104055.

Yu, X., & Deng, J. (2026). Predicting the Flory-Huggins parameter for polymer-solvents from quantum chemical descriptors. *Journal of Polymer Science*, *64*, 724–733.

Zhong, J., & Li, Z. (2023). Research on national carbon emission forecasting based on trend time series. *Operations Research and Fuzziology*, *13*(4), 3870–3881.