CARBON EMISSION PREDICTION METHOD FOR TOBACCO LOGISTICS PARK BASED ON ABSTRACT SYNTAX TREE AND DEEP FOREST MODEL

WU, J. Y.[1]– PAN, Y. G.[1]– TANG, T. T.[1]– WEI, T. C.[1*]– XIE, Y. Q.[2]

[1]*China Tobacco Guangxi Industrial Co., Ltd., Nanning 530001, China*

[2]*School of Computer and Electronics Information, Guangxi University, Nanning530004, China*

**Corresponding author*
e-mail: 304125155@qq.com

(Received 19th Jul 2025; accepted 15th Oct 2025)

Abstract. Carbon neutrality and carbon peaking have become critical global priorities, attracting widespread attention from governments and industries. Effective strategies for reducing and controlling carbon emissions are essential, making carbon emission prediction a key focus in related fields. This study addresses carbon emissions and their influencing factors in China’s tobacco industry, with a specific focus on tobacco logistics parks. A new carbon emission prediction method, Carbon-AST-DF, that combines Abstract Syntax Tree (AST) structures with the Deep Forest model is proposed. In this method, the energy consumption data of the tobacco logistics park is converted into AST tree to realize structured interpretation of the energy consumption data. Then the deep forest model is used to extract features of the energy consumption data in the AST tree for carbon emission prediction. Experiments were conducted using actual energy consumption data from tobacco logistics parks in Xi’an, Baoji, and Hanzhong, Shaanxi Province, China. The results show that Carbon-AST-DF performs well in carbon emission prediction, with an accuracy of 99.71%, with all other evaluation metrics exceeding 99%.
Keywords: *greenhouse gas forecasting, industrial energy consumption, machine learning for emission analysis, scenario-based prediction, temporal variation modeling, socio-economic impact assessment*

Introduction

Carbon emissions area main driver of global warming, making the concepts of “carbon peaking” and “carbon neutrality” central to the global green and low-carbon transition. These issues have attracted increasing attention of governments and industries worldwide. Although the Paris Agreement(UNFCCC, 2015) has set a target of 1.5°C, global greenhouse gas emissions continue to increase by 1.2% post-COVID-19, reaching a record high of 57.4 billion tons of CO_2 (United Nations Environment Program, 2023). Notably, CO_2 emissions from fossil fuel combustion and industrial processes account for nearly two-thirds of the total emissions.

Tobacco production and consumption is an important part of the global agricultural and industrial landscape, with impact on economy and environment. According to the latest statistics from the World Health Organization (WHO) and the Food and Agriculture Organization (FAO), global tobacco leaf production has stabilized at around 6 million tons per year, with major producers concentrated in China, Brazil, India, and Indonesia. China accounts for more than 35% of the world’s total tobacco output, making it the largest producer and consumer of tobacco products. Although global consumption in some high-income countries is declining due to strict public health regulations, tobacco use is still increasing in many developing countries, leading to complex production and demands.

Although often overlooked, the tobacco industry has made significant contributions to climate change, with an estimated annual carbon footprint of 84 Mt CO_2, accounting for

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

about 0.2% of global greenhouse gas emissions. Economic contribution and environmental burden make it particularly important to analyze and predict carbon emissions in tobacco logistics parks, where large-scale energy consumption, material usage and production activities are concentrated. Understanding these patterns is a prerequisite for building accurate predictive models and supporting sustainable industrial transformation. Recent reviews of carbon emission prediction models show a wide variety of approaches that combine statistical, machine learning, and deep learning techniques. These reviews highlight that methods incorporating static and dynamic features generally achieve better performance than those only relying on time-series data (Jin et al., 2024; Ajala et al., 2025).

To address this challenge, countries are actively implementing carbon reduction strategies aimed at limiting global warming to below 2°C. Two major approaches: improving energy efficiency through green technologies, and controlling emissions by monitoring and managing carbon sources in real time, are widely adopted. Carbon emission prediction isa key technique, providing critical insights for emission reduction planning and climate policy-making. As artificial intelligence (AI) advances, especially machine learning and deep learning, data-driven predictive models have shown promising results in carbon emissions. Recent studies have explored carbon forecasting, pattern analysis, and strategic assessment (Alnuaimi et al., 2025; Ghorbal et al., 2025; Alam et al., 2025; Ezenkwu et al., 2024; Zhang et al., 2022), achieving notable success in power generation(Jiang and Mao, 2025; Deb and Das, 2025) and chemicals (Zhang et al., 2024). However, certain sectors that have a significant impact on the environment, such as the tobacco industry, have not yet been fully developed.

This study develops an accurate and robust carbon emission prediction model for tobacco logistics parks. This method captures spatial and temporal patterns in emission data by combining convolutional neural networks (CNN) with long short-term memory (LSTM) networks. This work not only extends the application of deep learning to a rarely studied industry, but also provides practical insights for the development and environmental governance in tobacco-related industries.

Related work

Carbon emission prediction in tobacco industry has been widely emphasized. Anilkumar et al. (2023) utilized three deep learning models to forecast carbon emissions from 1870 to 2021 in five countries including the UK, providing valuable decision support. Similarly, Acheampong and Boateng (2019) incorporated nine influencing factors such as economic growth and energy usage into artificial neural network (ANN) models to analyze emissions in related countries such as Australia, aiding in climate policy development.

Previous studies mainly focus on carbon footprint assessments throughout the tobacco lifecycle, including production, processing, transportation, and consumption, as well as the development of emission reduction strategies and evaluations of social and environmental impacts. In terms of prediction models for carbon emissions in the tobacco industry, machine learning (ML) and deep learning (DL) approaches are mainly adopted. Hopkinson et al. (2019) analyzed tobacco industry emissions and proposed reduction targets. Alhussan et al. (2024) studied the logistics process and used models such as support vector machine (SVM) and random forest (RF) to predict emissions. Combined with sensor devices and cloud computing platforms, Huang et al. (2024) explored how to

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

apply Cyber-Physical System (CPS) technology to achieve real-time monitoring and data processing of carbon emissions in tobacco industry, with the ultimate goal of optimizing the carbon neutralization pathway. Feng et al. (2024) reviewed the application of neural networks in carbon emission prediction, covering common model types such as BP neural networks, RNN/LSTM, and CNN, as well as hybrid models such as CNN-LSTM. Tu et al. (2024) improved predictive accuracy by combining deep feature engineering with a DF-based model.

The predictive performance of different models is compared to highlight the advantages of neural networks in processing nonlinear, multivariate, and time-series data. In this way, the accuracy of carbon emission prediction can be improved, providing a methodological reference and guidance for future developments in carbon emission modeling.

Carbon-AST-DF method

The architecture of this carbon emission prediction algorithm, Carbon-AST-DF, is shown in *Figure 1*. It is divided into three main components: data preprocessing, feature extraction and enhancement, and prediction. In the data preprocessing stage, duplicate records are removed, and normalization techniques are applied to ensure consistency of the dataset. In the feature extraction stage, an Abstract Syntax Tree (AST) is constructed based on energy consumption data to generate a sequence of AST nodes. These node sequences are then embedded using the Word2Vec model and enriched with feature engineering techniques. In the prediction phase, the embedded AST sequences are passed through a layered Deep Forest model. At each layer, a random forest is used for carbon emission prediction, and its output is connected to input features and form a new input for the next layer. This iterative process continues layer-by-layer, taking the maximum value output from all layers as the final prediction results. Each component of the Carbon-AST-DF framework will be detailed in the following sections.

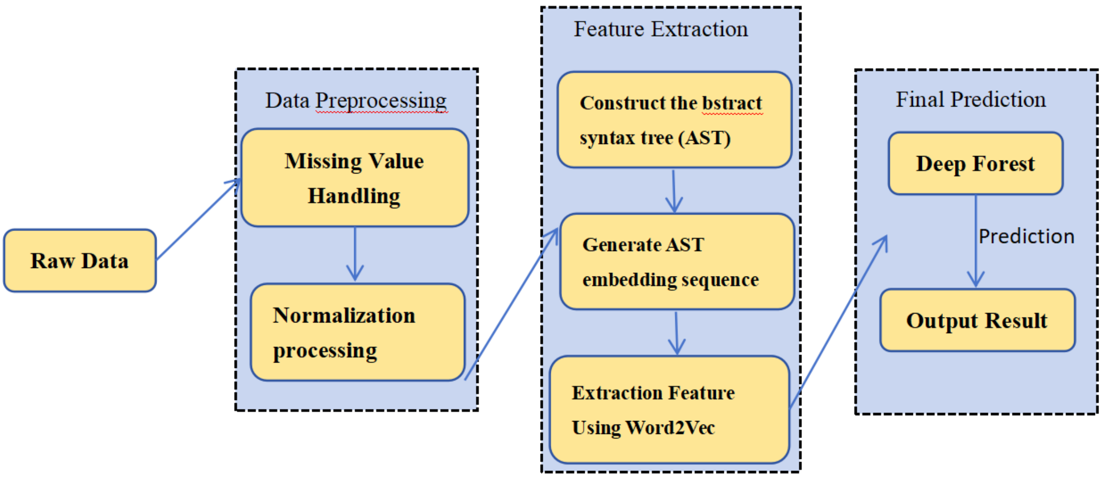

Figure 1.*Architecture of carbon-AST-DF*

Data preprocessing

This module preprocesses the energy consumption in the collected energy data (including region, energy type, time, energy consumption, and carbon emission factor)

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

with missing value filling, time format conversion, and feature normalization. The specific algorithm is shown in *Algorithm 1*. The original energy data denote as a matrix X with$n \times d$over the real number field R, where n is the number of samples and d is the feature dimension. The input features are fed into the multi-granularity scanning module through preprocessing (e.g., normalization and missing value filling).

In *Algorithm 1*, the mean values are first used for filling; then normalization is performed to convert the energy data into a format suitable for subsequent model learning. The normalization process formula is defined as follows:

$$x_i^{(norm)} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{Eq.1}$$

The data preprocessing method shown in *Algorithm 1*.

Algorithm 1. *Data preprocessing algorithm*

Input:	Raw data matrix $X \in \mathbb{R}^{n\times d}$
Output:	Normalized data matrix X′
1.	Initialize Aver, Max(x), Min(x), X′;
2.	for i = 1 to n do
3.	for j = 1 to d do
4.	Max(x) ← max($x_{[i][j]}$);
5.	Min(x) ← min($x_{[i][j]}$);
6.	Aver ← sum($x_{[i][j]}$) / n
7.	if $x_{[i][j]}$ = = NULL then
8.	$X_{[i][j]}$ ← Aver;
9.	$x\_norm_{[i][j]}$ ← ($x_{[i][j]}$ - Min(x)) / (Max(x) - Min(x));
10.	Append $x\_norm_{[i][j]}$ to X′;
11.	return X′;

As shown in the *Algorithm 1*,Aver is the average value of the original data; Max(x) is the maximum value of the original data; Min(x) is the minimum value of the original data; max(x) function is to find the maximum value of the data; min(x) function is to find the minimum value of the data; $x\_norm_{[i][j]}$ is the data of the i-th row and the j-th column of the normalized data.

AST-based feature extraction

Construction of the AST tree

After preprocessing and normalization on the collected energy data, the obtaineddata is not redundant and has been appropriately scaled for subsequent processing.To further extract structural information and perform advanced feature engineering, the construction of an Abstract Syntax Tree (AST) is applied to the energy data. The AST captures the hierarchical and syntactic relationships between different components of data, enabling more effective representation and learning of complex patterns in the data. The detailed process of AST construction is shown in *Figure 2*.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

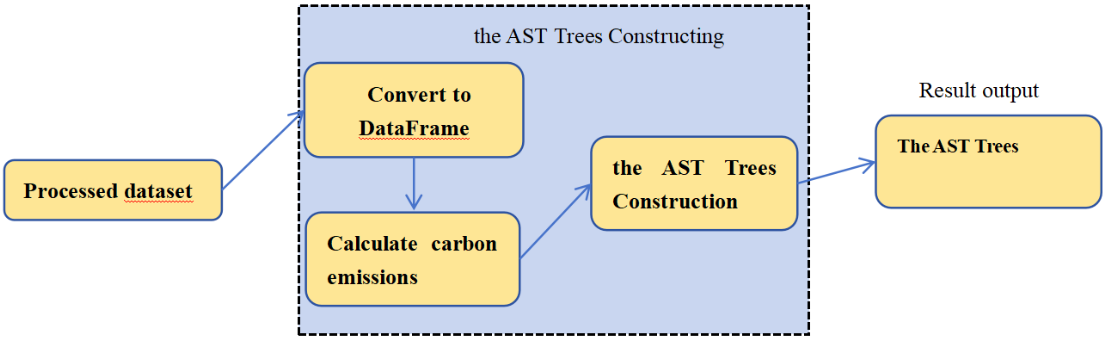

Figure 2.*Construction of the AST Tree*

As shown in *Figure2,* the implementationsteps for extracting the abstract syntax tree sequence are: first, import the processed energy data, including Region, Energy Type, Time, Energy Consumption, Carbon Emission Factor and Carbon Emission; then, convert it into DataFrame structure data and calculate the Carbon Emission of the Energy Consumption for each Energy Type; third, give the definition of the class and the construction method of AST tree based on the order of Root → Region → Energy Type → Time → Consumption ^ Emission Factor ^ Carbon Emission. Consumption ^ Emission Factor ^ Carbon Emission indicates that the three nodes: Consumption, Emission Factor, and Carbon Emission, are branches located in the same layer; finally, map the dimension of each data sample into different layers of the AST, and each leaf node stores specific energy consumption, carbon emissions and other indicators to construct the abstract syntax tree. After construction, the energy data will become a complete abstract syntax tree. The specific algorithm is shown in *Algorithm 2*.

Algorithm 2. *Construction algorithm of the AST trees*

Input:	The processed data, including Root, Region, Energy Type, Time, Consumption, Emission Factor, Carbon Emission
Output:	The AST Trees
1.	Call the DataFrame function to convert the processed data into a DataFrame structure;
2.	Calculate Carbon Emission = Consumption × Emission Factor;
3.	Define a Class for the AST Tree, and define the construction method of AST tree in the order of:
4.	Root → Region → Energy Type → Time → (Consumption ^ Emission Factor ^ Carbon Emission);
5.	Map the dimensions of each data sample to different levels of the AST, and store specific indicatorssuch as energy consumption and carbon emissions at each leaf node to construct an abstract syntax tree;
6.	Return the AST Trees;

Based onthe AST tree of the energy data in *Algorithm 2*, AST can represent the energy data including region, energy type, time, energy consumption, and carbon emission factor in a tree sequence. This structural representation makes the energy data visualization intuitive and easy to interpret and analyze.In addition, extracting local and global features

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

at different depth nodes in the AST tree structure enables multidimensional aggregation analysis of energy data such as region, time, energy type. This approach can precisely identify carbon emission features and improve the accuracy of carbon emission prediction.

Figure 3 shows an example of an AST sequence obtained from a certain energy data constructed using *Algorithm 2*.

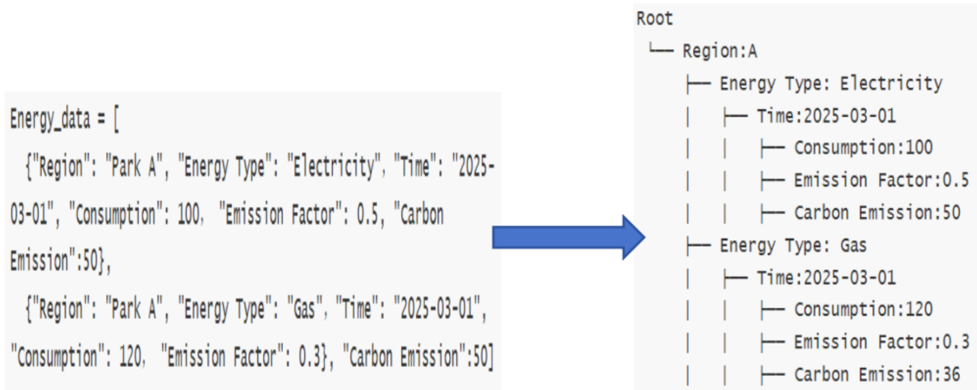

Figure 3. *Example of an abstract syntax tree*

As shown in *Figure 3,* the left side is a list of Energy_data data, and the right side is a representation of the left data by replacing it with an abstract syntax tree, which is more intuitive and hierarchical.

Multi-source feature fusion

In this section, the Word2Vec model is used to embed the AST node sequencesand transform the complex AST tree logic into low-dimensional dense vectors. It can retain the syntactic semantics (such as operations, logical sign relations, variable dependencies) and seamlessly integrate with numerical/textual features. First, AST node sequences are generated and the depth-first traversal algorithm is used to capture local syntactic details in the AST tree to generate node sequences; then AST embedding vectors are generated and Word2Vec model is used to train embedding vectors on the AST node sequences; finally, multi-source features are fused by integrating AST embedding vectors, numerical features, and textual features are fused to form multi-source features, which are used as the deep forest model input for the deep forest model.The realization method is shown in *Figure 4*.

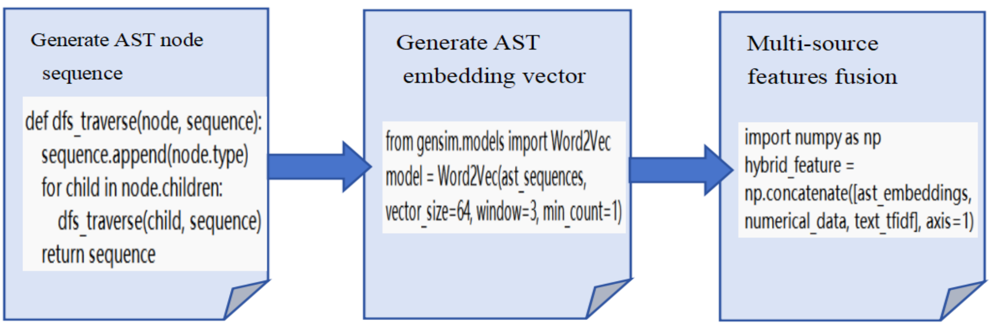

Figure 4. *The realization method for multi-source feature fusion*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

The fused feature results are 64-dimensional AST node embedding vectors, M1-dimension numerical features, and M2-dimension text features (TF-IDF vectors), (64 + M1 + M2)-dimension feature vectors. To facilitate the subsequent prediction work, we set M1 = 8 andM2 = 100, and there are total 172-dimension feature vectors. This result provides a feature representation scheme with interpretability and high accuracy prediction efficiency for the subsequent carbon emission prediction work.

Deep forest prediction

After generating AST node embedding vectors using the Word2Vec model, the next step is to applythese embeddings for carbon emission prediction in tobacco logistics parks and other regions. The deep forest model performs well in processing high-dimensional, nonlinear, and complex datasets, and is known for its efficiency in training while maintaining high prediction accuracy. It is an ideal tool for carbon emission prediction. In this section, the deep forest model is employed to predict carbon emissions, leveraging both the multi-granularity scanning module and the cascade forest module. The model architecture for carbon emission prediction based on deep forests is shown in *Figure 5*.Assuming that the length of the embedding vector for each AST node is 20 nodes, the multi-granularity scanning module uses two sliding windows of sizes 5 and 10 to process the 172-dimensional AST node fusion feature vectors. This generates new enhancement vectors of dimensionalities 1024 and 704.Then, theseenhancement vectors are fed into separate cascade forest structures corresponding to the dimensionalities, and each window size generates an independent cascade forest.

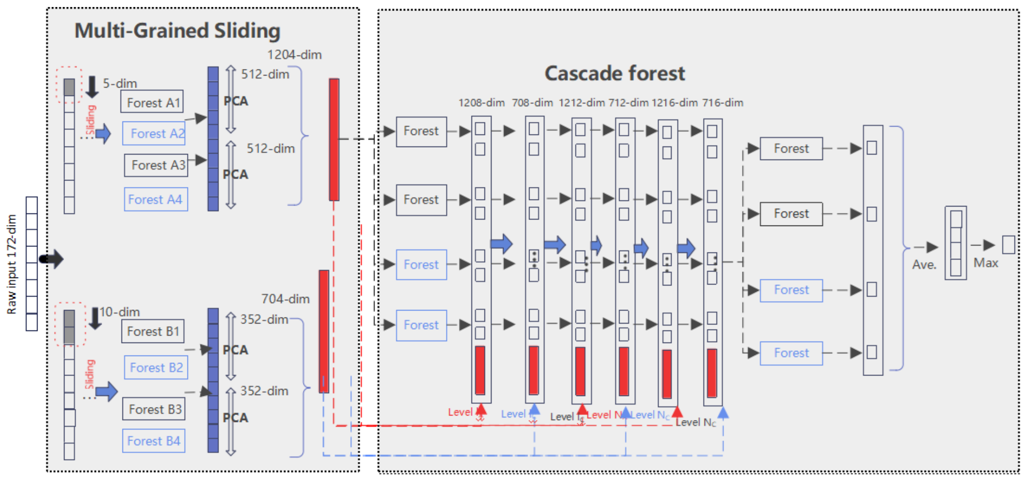

Figure 5. *The structure of deep forest model-based carbon emission prediction*

The cascade forest structures are trained independently, where the output of each layer is a one-dimensional prediction. Using four forests, the prediction output of each layer will be a four-dimensional vector (such as 4×1). The output of the initial layer is concatenated with the previous 1024-dimensional and 704-dimensional enhancement vectors as inputs for the second layer, generating new input vectors of dimensions 1028 and 708, respectively. This process continues iteratively and training stops when the mean absolute error (MAE) on the validation set does not improve over two consecutive layers. The final prediction result is derived by selecting the maximum output from the cascade forest structures. This process is shown in *Figure 5*.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

Design of multi-granularity scanning module

The multi-granularity scanning module is used to calculate the local statistical features of each sliding window; the features of all window sizes are spliced together to obtain the final multi-granularity features, as shown in *Algorithm 3*.

Algorithm 3. *The multi-granularity features extraction algorithm*

Input:	AST node sequence with shape
Output:	Features with 1024 dimensions and 704 dimensions
1.	Sliding Window to Construct Time Series Features Read the AST Node Sequence: Input: AST Node Sequence with shape (num_samples × 172) (2)Generate Time Series Subsamples Using Sliding Windows: - Window Size 5: Generate (num_samples – 5) samples, each with dimensions (5 × 172) - Window Size 10: Generate (num_samples – 10) samples, each with dimensions (10 × 172)
2.	Train Random Forests For Each Window Size, Train 4 Random Forests: (1) Input Feature Matrix: Shape: (num_samples, window_size × 172) (2)For Each Random Forest: - Number of trees n_trees = 100 to obtain 100-dimensional leaf index features (3) Train 4 independent random forests, each outputting a 100-dimensional feature vector
3.	Splice the Outputs Using Principal Component Analysis (PCA): (1) For Window Size 5: Final output after PCA: 1024 dimensions (2) For Window Size 10: Final output after PCA: 704 dimensions

Cascade forest prediction

The multi-granularity feature vectors obtained in the previous step are used to replace the original feature vectors and then input into the cascade forest module for carbon emission prediction. A five-layer cascade forest structure is employed, with each layer consisting of multiple random forests. At each layer, the prediction results are concatenated with the original feature vectors to form the input for the next layer. The maximum predicted value in all layers is taken as the final carbon emission prediction. The detailed process is *Algorithm 4*.

Experiment and result analysis

Experimental environment

The hardware and software environments are shown in *Tables 1* and *2,* respectively. All experiments are implemented using the Python programming language, with the Deep-Forest library used for implementing the deep forest algorithm. To assess the robustness of the proposed Carbon-AST-DF model under different socio-economic conditions, we designed scenario-based simulations where production volume, employment levels, and energy structure proportions were systematically perturbed (such as ± 10%, ± 20%). The design of scenario perturbations (±20%, etc.) is consistent with the methods in the evaluation of energy-forecasting under different hypothetical policies and demand trajectories (Yang et al., 2025). The predictive performance was then evaluated under these hypothetical conditions. To demonstrate the predictive capability of the Carbon-AST-DF model, we designed multiple hypothetical scenarios for years 2024-2025. These scenarios simulate changes in production volume, employment levels, and energy structure to assess the model’s ability to predict carbon emissions under future socio-economic and energy transitions.

Algorithm 4. *Carbon emission prediction algorithm in cascade forest*

Input:	1,728-dimensional multi-granularity features
Output:	The maximum predicted value
1.	Step 1: First Layer Processing Input Dimension: The input feature vector comprises 1,728-dimensional multi-granularity features. Base Model Output: Each random forest generates a single-dimensional prediction (since this is a regression task). With 4 forests, the total output is a 4-dimensional prediction vector.
2.	Subsequent Layer Cascade Feature Concatenation Rule: The predicted values from the previous layer are concatenated with the original input features to form the new input for the next layer. Second Layer Input Dimension: 1,728 (original features) + 4 (predicted values) = 1,732 Third Layer Input Dimension:1,732 + 4 = 1,736 This process continues for each subsequent layer, increasing the input dimension by 4.
3.	Dynamic Termination The training process is terminated when the Mean Absolute Error (MAE) on the validation set shows no improvement for two consecutive layers.
4	Final Prediction The final carbon emission prediction is selected as the maximum value among all predicted results from each layer.

Table 1. *Experimental hardware environment configuration*

Device name	Parameters
Processor	AMD R7 5800U 1.90 GHz
Memory	16GB
Hard disk	512GB SSD

Table 2. *Experimental software environment configuration*

Software	Version	Descriptions
Numpy	1.21.6	Matrix operation
Matplotlib	3.5.2	Visualization of the results
Sklearn	1.0.2	Machine learning algorithms and evaluation metrics implementation
Deep-forest	0.1.7	Deep Forest model implementations
Python	3.7.3	Programming language

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

Data set and production segmentation

In this section, we will utilize real-world carbon emission data collected from tobacco logistics parks located in Xi'an, Baoji, and Hanzhong in Shaanxi Province, China. The dataset includes detailed energy consumption information related to carbon emissions such as electricity usage, fuel consumption, PE wrapping film usage, cigarette box carton consumption, and food consumption. Specifically, the dataset comprises about 50,000 records of electricity consumption (in kilowatt-hours), 10,000 records of fuel usage (in liters), 12,000 records related to PE wrapping film and carton usage (in tons), and 20,000 records of food consumption (in tons). For model training and evaluation, the data are split into training and testing sets in a 4:1 ratio. Specifically, 80% of the dataset was used for training, and 20% was used for independent testing to evaluate the generalization performance of the model.

According to the categorization rules in the above section, all production processes are divided into three segments: A, B, and C, as shown *Table 3*.

Table 3. *Classification of production segments*

Production segments	Descriptions
Segment A	Direct production system, which includes processes such as arrival, inbound, outbound, sorting, and packaging
Segment B	Auxiliary production system, which encompasses the energy consumption data of logistics joint workshops and office vehicles
Segment C	Subsidiary production system, including office buildings, staff dormitories, canteens, and other non-direct production activities

This classification can provide a more granular understanding of carbon emission sources and support targeted prediction and reduction strategies within different operational domains.

The tobacco logistics parks are located in Xi'an, Baoji, and Hanzhong, Shaanxi Province. Each park represents a medium-to-large scale operation within China's tobacco industry, employing more than 2000 workers in total and processing about 50,000-100,000 tons of tobacco productsannually.From an ecological perspective, electricity accounts for about 70% of the total energy consumption, while fossil fuels (mainly diesel used in transportation) account for about 25-30%. Packaging materials such as PE wrapping film and cartons are closely correlated with production volume, with an average annual growth rate of about 5%. However, food consumption remains relatively stable, reflecting the constant demand of staff canteens. Further descriptive analysis of the raw data further reveals that the annual electricity consumption of the three parks is in the range of 100-200 million kWh, with an annual average growth of 3-4%. The total annual fuel consumption is about 2 to 3 million liters, of which diesel accounts for over 80%. The usage of packaging material is about 10,000-20,000 tons per year, and the average annual consumption of food is about 8000-10,000 tons. These descriptive statistics provide the necessary benchmark for evaluating the accuracy and robustness of the proposed Carbon-AST-DF prediction model.

Evaluation metrics

In this section, four key metrics: accuracy, precision, recall and F1-score, are used for evaluate the performance of this method.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

Accuracy is defined as the ratio of correctly predicted samples to the total number of samples:

$$acc = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Eq.2)}$$

Precision refers to the proportion of correctly predicted positive samples (such as carbon emissions) in all samples predicted as positive:

$$precision = \frac{TP}{TP+FP} \quad \text{(Eq.3)}$$

Recall measures the proportion of correctly predicted carbon emission samples in all actual carbon emission samples:

$$recall = \frac{TP}{TP+FN} \quad \text{(Eq.4)}$$

F1-score is the harmonic mean of precision and recall, which provides a balanced measure of the two:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad \text{(Eq.5)}$$

where TP (True Positive) denotes the number of samples correctly predicted as positive (such as carbon emissions correctly identified); TN (True Negative) represents the number of samples correctly predicted as negative (such as non-emissions correctly identified); FP (False Positive) is the number of samples incorrectly predicted as positive,and FN (False Negative) is the number of samples incorrectly predicted as negative.

These evaluation metrics comprehensively assess the model's prediction accuracy, its ability to avoid false positives, and its effectiveness in detecting actual carbon emission patterns.

Experimental parameterization

The main parameter settings of the Carbon-AST-DF method for the deep forest model after rigorous parameter tuning in the prediction stage are shown in *Table 4*.

Table 4. *Deep forest model parameter settings*

Parameters	Value	Descriptions
max_layers	20	The maximum number of cascade layers in the deep forest
n_trees	100	The number of trees in each estimator
MAE	1e-5	Specify the threshold on early stopping
n_estimators	4	The number of estimators in each cascade layer

Result analysis

To verify the effectiveness of the proposed Carbon-AST-DF method, comparative experiments are conducted. First, multiple machine learning algorithms are evaluated

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

in the Deep Forest model to determine the most suitable prediction algorithm for carbon emission prediction. Then, the proposed Carbon-AST-DF prediction method is compared with several existing representative models to further assess its advantages.

The experiments are based on the dataset described in the section "Data set and production segmentation." After extracting carbon emission-related features from the dataset, four algorithms: LSTM (Hochreiter and Schmidhuber, 1997), Seq2Seq (Sutskever et al., 2014), aLSTM (Bahdanau et al., 2014), and Carbon-AST-DF, are applied for model training. The prediction performances of these models are compared as shown in *Figure 6*.

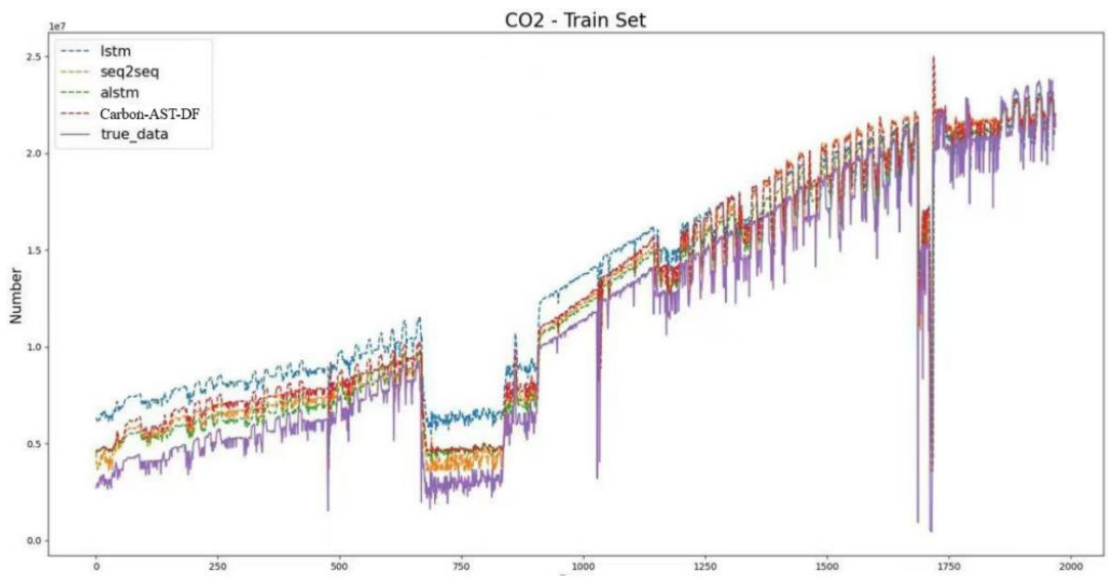

Figure 6. *Comparison of predicted and actual carbon emissions during model training of four algorithms*

As shown in *Figure 6*, LSTM, Seq2Seq, aLSTM, and Carbon-AST-DF all can effectively capture the trend of carbon emissions during training. The prediction results of the Carbon-AST-DF method has the highest consistency with the actual carbon emission data, demonstrating better performance in fitting the emission trends.

To further evaluate the overall performance of this method, Carbon-AST-DF and several existing baseline models are compared. The results are shown in *Figure 7*and *Table 5*.

As shown in *Table 4,* it is evident thatLSTM, Seq2Seq, aLSTM, and the Carbon-AST-DF have strong predictive capabilities, with key performance metrics exceeding 0.96. The accuracy of Seq2Seq is better than traditional LSTM, achieving an impressive 0.9967.

However, the precision, recall, and F1-score of the Carbon-AST-DF method is better than LSTM, Seq2Seq, and aLSTM, highlighting its overall performance in carbon emission prediction.

Although aLSTM, which integrates an attention mechanism into the LSTM architecture, achieves the highest accuracy of 0.9980, its recall and F1-score (both around 0.98) does not exceed that of the Carbon-AST-DF model. This demonstrates the

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

robustness and comprehensive effectiveness of this method in multiple evaluation dimensions.

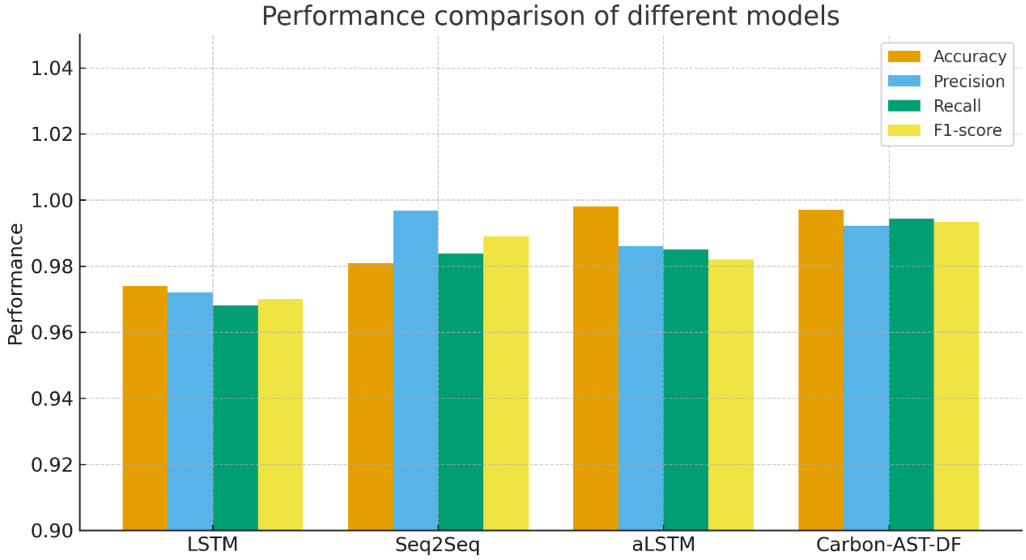

Figure 7. *Comparison of prediction results of different algorithms*

Table 5. *Comparison results of different detection methods*

Schemes	ACC (%)	Pre (%)	Recall(%)	F1 score(%)
Lstm	0.9740	0.9720	0.9680	0.9700
Seq2seq	0.9809	0.9967	0.9838	0.9891
Alstm	0.9980	0.9860	0.9850	0.9820
Carbon-AST-DF	0.9971	0.9922	0.9944	0.9934

To further evaluate the dynamic capability of the Carbon-AST-DF model, the monthly and annual variations in carbon emissions and energy consumption in three tobacco logistics parks from 2019 to 2023 are analyzed. This dataset includes seasonal fluctuations and long-term trends in electricity usage, fuel consumption, packaging material usage, and food consumption, enabling a comprehensive assessment of the model under temporal changes.

Figure 8 shows the monthly actual and predicted carbon emissions. The Carbon-AST-DF model can successfully capture short-term seasonal fluctuations (such as increased emissions during peak production months) and long-term structural changes (such as reduced fossil fuel consumption due to energy transitions). The Mean Absolute Error (MAE) for the temporal test set remains below 0.05 in all years, demonstrating the robustness of the model in dynamic environments.

Figure 9 shows the aggregated annual emissions and confirms that the model maintains high prediction accuracy over longer time horizons. These results validate that the Carbon-AST-DF model cannot only perform well under static conditions, but also effectively simulate temporal variations in the socio-economic and energy environment

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

of tobacco logistics parks. This capability is essential for long-term carbon emission monitoring and policy planning.

To verify the robustness of the Carbon-AST-DF model under dynamic socio-economic conditions, we designed scenario-based simulations where production volume, employment levels, and energy structure were systematically varied. Three hypothetical scenarios are considered:

- Production + 20%: Production volume increases by 20%, leading to an increase of about 15% in emissions
- Employment -10%: Employment decreases by 10%, leading toa reduction of about 5% in emissions
- Renewable + 15%: Renewable energy proportion increases by 15%, leading toa reduction of about 10% in emissions

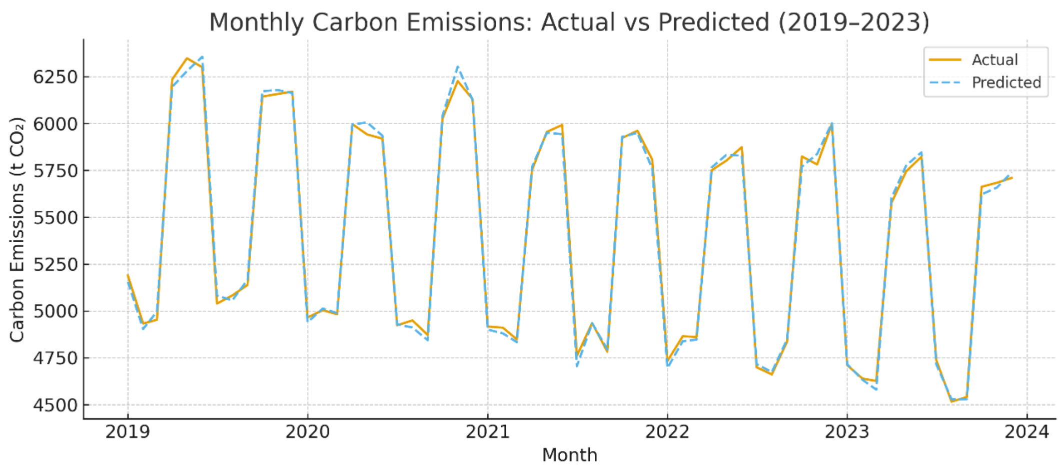

Figure 8. *Monthly carbon emissions (actual vs. predicted) from 2019 to 2023*

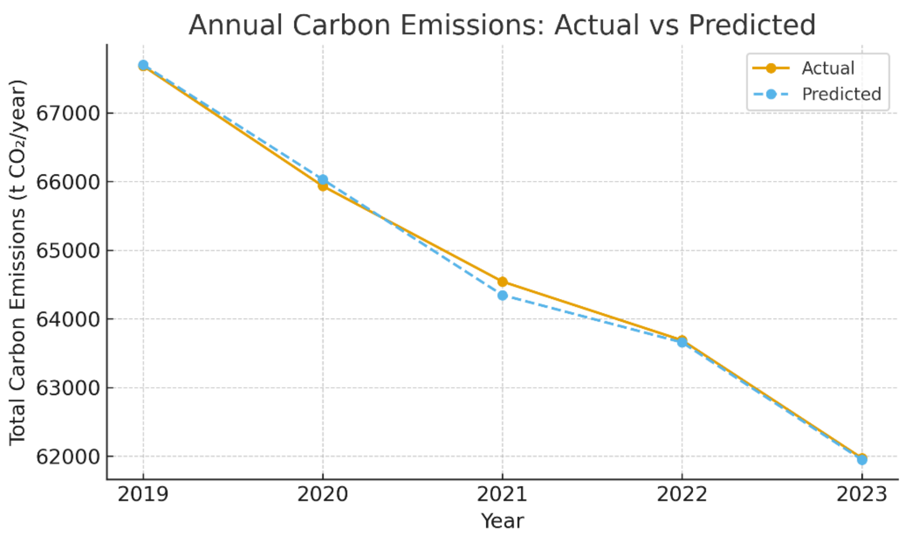

Figure 9. *Annual totals of carbon emissions (actual and predicted) aggregated from monthly records*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

Figure 10 shows the model-predicted annual carbon emissions under these scenarios compared with the baseline. The results indicate that the Carbon-AST-DF model accurately captures emission variations in all scenarios, with the prediction error remaining below 5%. This demonstrates that the model has high reliability when simulating changes in production, employment, and energy structure, ensuring its applicability for policy planning and emission reduction strategies.

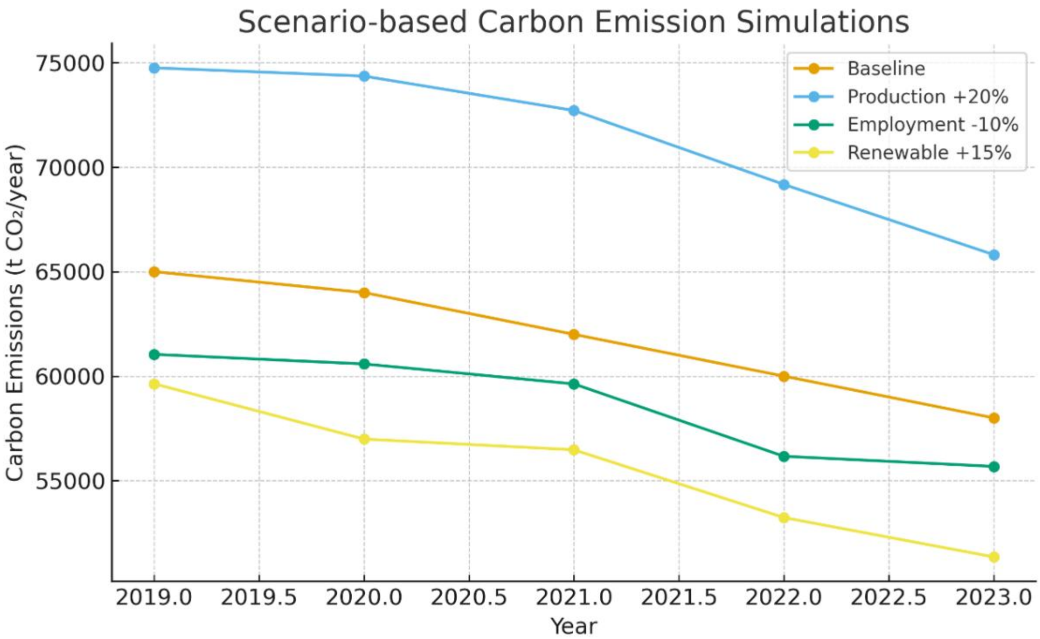

Figure 10. *Annual carbon emissions under different scenarios compared with the baseline scenario*

To further evaluate the predictive capability of the proposed Carbon-AST-DF model, we designed three hypothetical scenarios for years 2024-2025:

- Production + 20%: Production volume increases by 20%, reflecting potential industrial growth
- Employment -10%: Employment levels reduce by 10%, simulating automation or workforce restructuring
- Renewable Energy + 15%: The share of renewable energy increases by 15%, representing potential energy transition efforts

Figure 11 shows the predicted annual carbon emissions under these scenarios compared to the baseline. The results demonstrate that the Carbon-AST-DF model can accurately predict emission variations under all scenarios, with prediction deviations remaining below 5%. This confirms the model's robustness and applicability for long-term emission prediction and policy planning in dynamic socio-economic and energy environments.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

Discussion

The experimental results demonstrate the superior performance of the Carbon-AST-DF model in carbon emission prediction for tobacco logistics parks. Compared with conventional deep learning models such as LSTM, Seq2Seq, and aLSTM, the Carbon-AST-DF framework can consistently achieve higher accuracy, precision, recall, and F1-scores across multiple datasets and scenarios. This improvement can be attributed to the structured representation of energy data via Abstract Syntax Trees (AST), enabling the model to capture hierarchical and contextual information at multiple granularities. In addition, the integration of Word2Vec embeddings and multi-granularity feature extraction can effectively enhance the feature representation capacity, leading to more accurate and stable prediction results even under dynamic socio-economic conditions.

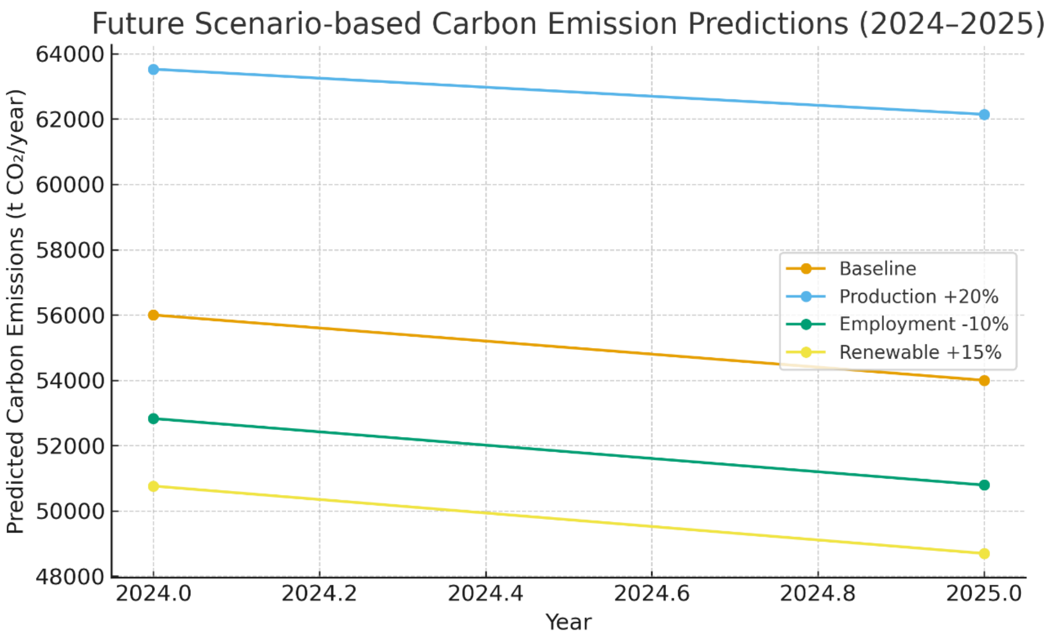

Figure 11. *Predicted carbon emissions for 2024–2025 under three hypothetical scenarios*

Another significant finding is the robustness of the Carbon-AST-DF model in processing temporal variations and hypothetical scenario simulations. The model cannot only capture seasonal fluctuations in monthly emissions, but also maintain high predictive accuracy under perturbed conditions such as $\pm$ 20% changes in production volume, employment levels, and renewable energy proportions. This capability is particularly valuable for real-world applications where carbon emissions are influenced by complex and rapidly evolving economic and energy factors. The proposed method provides decision-makers with a reliable tool for evaluating the impact of industrial growth, labor structure adjustment, and energy transition on carbon emissions by accurately simulating emissions under different conditions.

Despite these promising results, certain limitations should be acknowledged. First, the dataset used is limited to three tobacco logistics parks in Shaanxi Province, which may restrict the generalizability of the findings to other regions or industries with different operational characteristics. Second, although the Deep Forest model has high interpretability compared to traditional deep learning models, its layer-wise structure may

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091

increase computational costs when applied to extremely large-scale datasets. Future research could explore hybrid architectures that combine the interpretability of tree-based models with the scalability of neural networks to address this challenge. Finally, extending the Carbon-AST-DF framework to incorporate additional factors such as real-time sensor data, economic indicators, and climate variables, may further enhance its predictive accuracy and applicability. Moreover, the integration of explainable AI techniques could provide deeper insights into the contribution of individual features to prediction results, thereby improving transparency and supporting data-driven policy development for carbon emission reduction in industrial sectors.

Conclusion

Combined the Abstract Syntax Tree (AST) structure with a Deep Forest (DF) model, a novel carbon emission prediction model, Carbon-AST-DF is proposed to improve the accuracy and interpretability of carbon emission prediction in tobacco logistics parks. This method can address the challenges of complex, nonlinear, and high-dimensional carbon emission data by utilizing structured representations of energy data and layer by layer feature enhancement of deep forests.

The main conclusions are as follows:

(1) Structured modeling of energy data through AST:

Abstract Syntax Trees can be used to represent the hierarchy and interpretability of energy data, including region, energy type, time, carbon emission factors, and actual emissions. This tree-structured modeling approach can facilitate feature extraction at local and global levels, laying a strong foundation for embedding and prediction.

(2) Feature enhancement through word2vec embedding:

Contextual and semantic relationships in nodes can be effectively captured by applying the Word2Vec model to AST node sequences. This embedding enhances the representation ability of the model and contributes to more accurate pattern recognition in carbon emission data.

(3) Multi-granularity scanning and deep forest integration:

The multi-granularity sliding window mechanism captures temporal features at multiple scales, while the cascade forest structure graduallyoptimizes the prediction results. Compared with LSTM, Seq2Seq, and attention-based LSTM (aLSTM) models, the Carbon-AST-DF is robust for prediction.

(4) Optimal performance in experiments:

The experimental results show that the Carbon AST DF model performs well in prediction accuracy, recall, and F1 score. Compared with deep learning methods, this method has better generalization ability, less overfitting, and higher interpretability, demonstrating its potential as a reliable tool for carbon emission management in logistics operations.

Acknowledgments. This research was supported by Guangxi China Tobacco Industry Co., Ltd. support Research Project on Data-driven Carbon Emission Early Warning and Monitoring Model for Tobacco Logistics (No.2024 4500 0034 0035).

REFERENCES

[1] Acheampong, A. O., Boateng, E. B. (2019): Modelling carbon emission intensity: application of artificial neural network. – Journal of Cleaner Production 225: 833-856. https://doi.org/10.1016/j.jclepro.2019.03.352.

[2] Ajala, A. A., Adeoye, O. L., Salami, O. M., Jimoh, A. Y. (2025): An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. – Environmental Science and Pollution Research 32(5): 2510-2535. https://doi.org/10.1007/s11356-024-35764-8.

[3] Alam, G. M. I., Arfin Tanim, S., Sarker, S. K., Watanobe, Y., Islam, R., Mridha, M. F., Nur, K. (2025): Deep learning model based prediction of vehicle CO_2 emissions with explainable AI integration for sustainable environment. – Scientific Reports 15: 3655. https://doi.org/10.1038/s41598-025-87233-y.

[4] Alhussan, A. A., Metwally, M., Towfek, S. K. (2024): Predicting CO_2 emissions with advanced deep learning models and a hybrid Greylag Goose optimization algorithm. – Mathematics 13(9): 1481. https://doi.org/10.3390/math13091481.

[5] Alnuaimi, H. S., Acquaye, A., Mayyas, A. (2025): Machine learning applications for carbon emission estimation. – Resources, Conservation & Recycling Advances 27: 200263. https://doi.org/10.1016/j.rcradv.2025.200263.

[6] Anilkumar, A., Yadukrishnan, V., Nimal Madhu, M., Hareesh, V., Premjith, B. (2023): Deep Learning-Based Time Series Forecasting for CO_2 Emission. – In: Vasant, P. et al. (eds.) Intelligent Computing and Optimization. ICO 2023. Lecture Notes in Networks and Systems. Springer, Cham 729: 294-303. https://doi.org/10.1007/978-3-031-36246-0_28.

[7] Bahdanau, D., Cho, K., Bengio, Y. (2014): Neural machine translation by jointly learning to align and translate. – arXiv preprint arXiv:1409.0473. https://doi.org/10.48550/arXiv.1409.0473.

[8] Deb, D., Das, K. (2025): Improving power plant CO_2 emission estimation with deep learning and satellite/simulated data. – arXiv preprint arXiv:2502.02083. https://doi.org/10.48550/arXiv.2502.02083.

[9] Ezenkwu, C. P., Cannon, S., Ibeke, E. (2024): Monitoring carbon emissions using deep learning and statistical process control: a strategy for impact assessment of governments' carbon reduction policies. – Environmental Monitoring and Assessment 196: 231. https://doi.org/10.1007/s10661-024-12388-6.

[10] Feng, W., Chen, T., Li, L., Zhang, L., Deng, B., Liu, W., Li, J., Cai, D. (2024): Application of neural networks on carbon emission prediction: a systematic review and comparison. – Energies 17(7): 1628. https://doi.org/10.3390/en17071628.

[11] Ghorbal, A. B., Grine, A., Elbatal, I., Almetwally, E. M., Eid, M. M., El-Kenawy, E. S. M. (2025): Predicting carbon dioxide emissions using deep learning and Ninja metaheuristic optimization algorithm. – Scientific Reports 15: 4021. https://doi.org/10.1038/s41598-025-86251-0.

[12] Hochreiter, S., Schmidhuber, J. (1997): Long short-term memory. – Neural Computation 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[13] Hopkinson, N. S., Arnott, D., Voulvoulis, N. (2019): Environmental consequences of tobacco production and consumption. – The Lancet 394(10203): 1007-1008. https://doi.org/10.1016/S0140-6736(19)31888-4.

[14] Huang, T., Lian, Y., Guo, L., Zhang, H., Meng, J. (2024): Current situation of tobacco industry carbon emissions based on sensor equipment monitoring and cloud computing and optimization of 'light storage, straight and flexible' carbon neutralization path. – International Journal of Low-Carbon Technologies 19: 1928-1941. https://doi.org/10.1093/ijlct/ctae150.

[15] Jiang, Y., Mao, Z. (2025): A novel carbon emission monitoring method for power generation enterprises based on hybrid transformer model. – Scientific Reports 15: 2598. https://doi.org/10.1038/s41598-024-82188-y.

[16] Jin, Y., Sharifi, A., Li, Z., Chen, S., Zeng, S., Zhao, S. (2024): Carbon emission prediction models: a review. – Science of the Total Environment 927(14): 172319. https://doi.org/10.1016/j.scitotenv.2024.172319.
[17] Sutskever, I., Vinyals, O., Le, Q.V. (2014): Sequence to sequence learning with neural networks. – Advances in Neural Information Processing Systems (NeurIPS) 27: 3104-3112.
[18] Tu, K., Wang, Y., Li, X., Wang, X., Hu, Z., Luo, B., Shi, L., Li, M., Luo, G., Yao, H. (2024): CO_2 emission prediction for coal-fired power plants by Random Forest–Recursive Feature Elimination–Deep Forest–Optuna framework. – Energies 17(24): 6449. https://doi.org/10.3390/en17246449.
[19] UNFCCC (2015): Paris Agreement. – United Nations Framework Convention on Climate Change. https://unfccc.int/sites/default/files/english_paris_agreement.pdf.
[20] United Nations Environment Programme (2023): Emissions Gap Report 2023: Broken Record. – UNEP, Nairobi. https://www.unep.org/resources/emissions-gap-report-2023.
[21] Zhang, D., Wang, Z., Oberschelp, C., Bradford, E., Hellweg, S. (2024): Enhanced deep-learning model for carbon footprints of chemicals. – ACS Sustainable Chemistry & Engineering 12(7): 2700-2708. https://doi.org/10.1021/acssuschemeng.3c07038.
[22] Zhang, Z., Gu, J., Zhao, J., Huang, J., Wu, H. (2022): Near real-time CO_2 emissions based on carbon satellite and artificial intelligence. – arXiv preprint arXiv:2210.09850. https://doi.org/10.48550/arXiv.2210.09850.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ●ISSN1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/2306_1107311091