Deep Learning in Carbon Neutrality Forecasting:
A Study on the SSA-Attention-BIGRU Network

Jiwei Ran, School of Politics and Public Administration, Guangxi Normal University, China

Ganchang Zou, School of Politics and Public Administration, Guangxi Normal University, China

Ying Niu, College of Liberal Arts and Social Sciences, City University of Hong Kong, Hong Kong*

ABSTRACT

With the growing urgency of global climate change, carbon neutrality, as a strategy to reduce greenhouse gas emissions into the atmosphere, is increasingly seen as a critical solution. However, current forecasting models still face significant challenges and limitations in accurately and effectively predicting carbon emissions and their associated effects. These challenges largely stem from the complexity of carbon emission data and the interplay of anthropogenic and natural factors. To overcome these obstacles, the authors introduce an advanced forecasting model, the SSA-Attention-BIGRU network. This model ingeniously integrates an external attention mechanism, bidirectional GRU, and SSA components, allowing it to synthesize various key factors and enhance prediction accuracy when forecasting carbon neutrality trends. Through experiments on multiple datasets, the results demonstrate that, compared to other popular methods, the SSA-Attention-BIGRU network significantly excels in prediction accuracy, robustness, and reliability.

KEYWORDS

Carbon Neutrality, Deep Learning, Global Climate Change, Policy-Making, SSA-Attention-BIGRU Network

1. INTRODUCTION

In the context of today's global warming, achieving carbon neutrality has become a critically discussed topic of paramount importance (Zhao et al., 2022). Carbon neutrality refers to the goal of reaching net-zero emissions by reducing and offsetting carbon dioxide emissions to combat the threat of climate change. However, despite the widespread discussion and interest in the concept of carbon neutrality within the environmental sector, realizing this goal still presents a series of significant challenges and difficulties (Wu et al., 2022). Firstly, the idea of carbon neutrality involves reducing greenhouse gas emissions into the atmosphere, a challenge in and of itself. Industrial production, energy consumption, and transportation activities worldwide lead to substantial carbon dioxide emissions. Therefore, effectively monitoring, managing, and reducing these emissions poses an urgent issue. Secondly,

*Corresponding Author

forecasting these emissions is also an essential aspect of carbon neutrality. Yet, weather patterns, climatic conditions, and various human-induced factors can influence fluctuating emission levels, making accurate predictions extremely challenging. As such, there is a need to seek more precise and reliable methods for forecasting future emission trends. Lastly, reducing emissions requires the formulation and implementation of a series of effective policies and measures, which in turn, touch upon a broad spectrum of social and economic factors. Hence, ensuring sustainable socio-economic development while reducing emissions becomes a complex problem to address (Waheed et al., 2019). To tackle these challenges, many researchers and scientists have turned to deep learning and artificial intelligence technologies. Deep learning, a potent method within machine learning, can handle large-scale and intricate data, facilitating better emission monitoring, future trend predictions, and the proposition of effective emission reduction strategies. Significant advancements in emission monitoring and forecasting through deep learning offer renewed hope for achieving carbon neutrality (Li et al., 2021).

Deep learning has demonstrated immense potential in various aspects of the carbon neutrality domain, offering innovative solutions to achieve carbon-neutral objectives. As a compelling testament to this, one can observe the successful application of deep learning in the organization of sporting events, intelligent scheduling, and resource management (Wang et al., 2021). Sporting events are large-scale activities that attract millions of viewers globally. However, the organization and management of these events often involve significant energy consumption and carbon emissions (Somu et al., 2021). To mitigate the adverse environmental impacts of these events, deep learning technologies have been introduced to optimize resource utilization and carbon-neutral strategies for the events. By analyzing vast amounts of historical data and real-time information, deep learning algorithms can intelligently adjust the planning, scheduling, and energy consumption of these events, minimizing emissions and ensuring the sustainability of the events. These successful instances not only reduce the carbon footprint but also offer insights for other sectors, indicating the broad applicability of deep learning in carbon neutrality efforts (Amasyali & El-Gohary, 2018).

It's noteworthy that a series of researchers have delved into the references related to deep learning in carbon emission reduction prediction and the field of climate change. These studies not only provide invaluable insights but also lay a solid foundation for research on carbon neutrality using deep learning (Berriel et al., 2017). For instance, in the work related to deep reinforcement learning for energy management, researchers employed deep reinforcement learning to optimize energy management systems, aiming to significantly reduce energy consumption and carbon emissions (Liu et al., 2019). They developed a model based on the Deep Q-Network, which formulates the best energy scheduling strategies by learning real-time environmental data. However, this model encounters computational complexity issues when handling large-scale energy networks, especially when considering multiple diverse energy sources. Furthermore, another research explored the application of Convolutional Neural Networks (CNN) in air quality prediction. Researchers utilized vast amounts of meteorological and pollution data to train the CNN model to predict future air quality (Tang & Li, 2022). A limitation of this model is its typical need for a considerable amount of labeled data, and its performance might decline when dealing with complex meteorological conditions and pollution sources. Additionally, some studies focus on using Recurrent Neural Networks (RNN) to predict energy demands, aiding in optimizing the energy supply chain (Yu, 2023). While RNN models can handle time series data, they may face limitations in addressing long-term dependencies and seasonal changes (Oyando et al., 2023). Lastly, certain research endeavors are dedicated to using deep learning techniques to monitor and estimate carbon emissions. These methods often rely on large datasets, like satellite remote sensing data and ground observation data, to train the deep learning models. However, these approaches might be less effective under data scarcity or poor data quality, and the fine-grained classification and monitoring of emission sources remain challenging (Wenya, 2021).

Given the limitations of the aforementioned studies, we introduce the SSA-attention-BIGRU network. Within the context of carbon neutrality and climate change, this network integrates cutting-

edge deep learning technologies with the aim of offering a more precise prediction and analytical tool. Firstly, the external attention component stands as one of the central elements of this network. This component applies weighted processing to multi-variable input sequences, ensuring that essential points within the sequence are emphasized and retained, while relatively less crucial points are selectively overlooked. Such a strategy empowers the model to enhance the capture of key information within the time series, thus improving prediction accuracy. Secondly, the crux of the network revolves around the Bidirectional Gated Recurrent Unit (BIGRU). BIGRU is designed to grasp the non-linear relationships present within time series and is capable of concurrently considering past and future information, granting the model a comprehensive perspective. As for the SSA component, its primary function is to optimize the hyperparameter combinations of BIGRU. Through this mechanism, SSA ensures that BIGRU operates under its optimal configuration, thereby further amplifying the overall performance of the model.

The SSA-Attention-BIGRU is an integrated and efficient solution, particularly well-suited for prediction tasks related to carbon neutrality and climate change. Given the urgency surrounding climate change, we believe this network will provide invaluable tools and insights for the field. Below, we highlight the three key contributions of this paper:

- We introduced a novel SSA-Attention-BIGRU network, which ingeniously integrates the external attention component, the Bidirectional Gated Recurrent Unit (BIGRU), and the SSA component. This architecture ensures that critical points within the time series are given ample emphasis, while also capturing the non-linear relationships inherent in the time series.
- Through the SSA component, the network is able to automatically optimize the hyperparameter combinations for BIGRU. This significantly enhances the training efficiency and prediction accuracy of the model, allowing it to excel across various tasks and datasets.
- In the realm of carbon emission prediction, the SSA-Attention-BIGRU network offers an efficient and accurate prediction tool. It not only aids in a better understanding of carbon emission dynamics but also provides scientific decision-making support for formulating related policies and measures.

In the rest of this paper, we present recent related work in Section 2. Section 3 introduces our proposed methods. Section 4 showcases the experimental part. Section 5 contains the conclusion.

2. RELATED WORK

2.1 Application of External Attention Mechanism in Climate Prediction

In recent years, the external attention mechanism has gained widespread attention in time series forecasting. A study introduced a model named "Env-AttNet", specifically designed for environmental variable data (Chen et al., 2022). This model, by incorporating the attention mechanism into the traditional recurrent neural network, can capture long-term dependencies in time series with greater precision. Moreover, a model named "ClimateCNN", which blends attention mechanism with Convolutional Neural Networks (CNN), has been explored (Sheng et al., 2023). This model aims to obtain better feature representation both spatially and temporally. Furthermore, the "BiRNN-Attention" model, which combines the external attention mechanism with bidirectional recurrent neural networks, offers an efficient solution for predicting carbon emissions and climate change (Y. Liu et al., 2023).

These innovative approaches showcase the ongoing development in the field of time series forecasting models, emphasizing the increasingly crucial role of integrating external attention mechanisms. The "Env-AttNet" model, with its focus on environmental variable data, has paved the way for more precise predictions in the domain of environmental forecasting (Kaixu Han 2023). By incorporating the attention mechanism into the traditional recurrent neural network, it addresses the challenge of capturing nuanced patterns in time series data related to environmental variables.

Furthermore, the "BiRNN-Attention" model efficiently tackles the prediction of carbon emissions and climate change by combining the external attention mechanism with bidirectional recurrent neural networks (Guipeng Ding, 2023). This combination allows the model to better capture key information in time series, thereby enhancing prediction accuracy and robustness. These studies provide new methods and tools for understanding environmental and climate change, laying a solid foundation for future research and practical applications (Ning et al., 2024).

2.2 Exploration of Bidirectional Recurrent Networks in Meteorology

Bidirectional recurrent networks, especially BIGRU, have been utilized in many time series forecasting tasks (Yang et al., 2022). A model named "MeteorNet", based on bidirectional recurrent networks, has been designed to forecast future meteorological conditions, processing both past and future data simultaneously. Additionally, a model combining bidirectional recurrent networks with external memory units, named "WeatherMNet", aims to provide more accurate predictions for long-term meteorology (Lv et al., 2023). Further research shows that combining bidirectional recurrent networks with attention mechanisms can further enhance the predictive performance of the model.

2.3 Deep Reinforcement Learning in Energy Management

In the field of energy management, deep reinforcement learning has made significant progress. For instance, a model named "EnergyDQN" combines deep Q networks with real-time energy data to offer real-time optimization decisions for energy systems. A model named "Attention-EnergyAgent", which fuses deep reinforcement learning with the attention mechanism, has been developed (Zou et al., 2022). This model can automatically adjust the energy supply chain to minimize carbon emissions to the greatest extent. The "LinearReinforceNet" model, which combines deep reinforcement learning with traditional linear regression, excels in predicting energy consumption and emissions (Zhang et al., 2022).
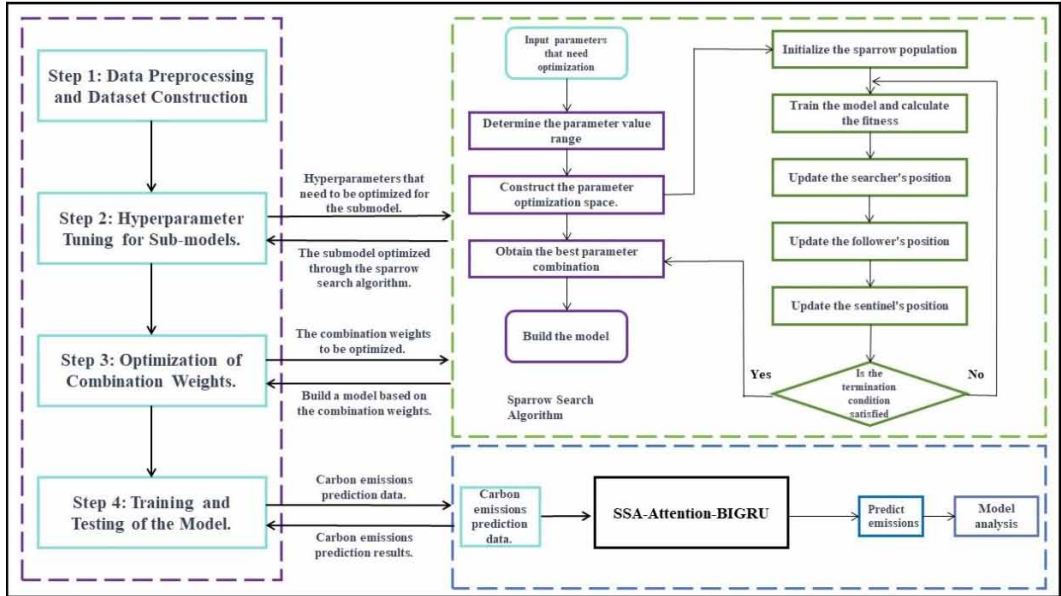
3. METHOD

The SSA-Attention-BIGRU network is a meticulously designed and highly integrated framework aimed at addressing the challenges posed by carbon neutrality and climate change. Its structure ingeniously combines three core components to enhance the precision and reliability of predictions.

Firstly, the external attention mechanism processes multivariate input sequences, automatically identifying and weighting key points in the time series while selectively disregarding those with a lesser impact on predictions(Ding et al., 2020). This balancing strategy significantly boosts the network's ability to capture critical changes within the data.

Secondly, at the core of the network, the BIGRU is responsible for capturing both short-term and long-term dependencies within the time series (Yang et al., 2022). Its bidirectional structure ensures the consideration of both past and future information from the time series, providing a more comprehensive contextual view and resulting in more accurate predictions. Simultaneously, the SSA component seeks and applies optimal hyperparameter combinations, ensuring the BIGRU operates at its prime state, greatly enhancing the model's training efficiency and predictive accuracy.

The construction of this network adheres to strict design principles and spans multiple stages. The process commences with data preprocessing, including handling missing values and normalization. Subsequently, the preprocessed data is input into the external attention mechanism for preliminary processing. Following this, the processed data is fed into the BIGRU to extract time series features. At the conclusion of each training cycle, the SSA component assesses and optimizes hyperparameters for the BIGRU (Sun et al., 2022). Ultimately, the network outputs prediction results, which can further aid in subsequent decision-making or other related applications, as illustrated in Figure 1, depicting the overall network flow.

Figure 1. Overall flow chart of the model

The SSA-Attention-BIGRU network not only provides an efficient solution to the problem but also assists researchers, policymakers, and relevant organizations in gaining a better understanding of the dynamics and patterns of carbon emissions. Moreover, the precise predictions furnished by this model offer robust scientific support for the formulation of related carbon reduction policies and measures.
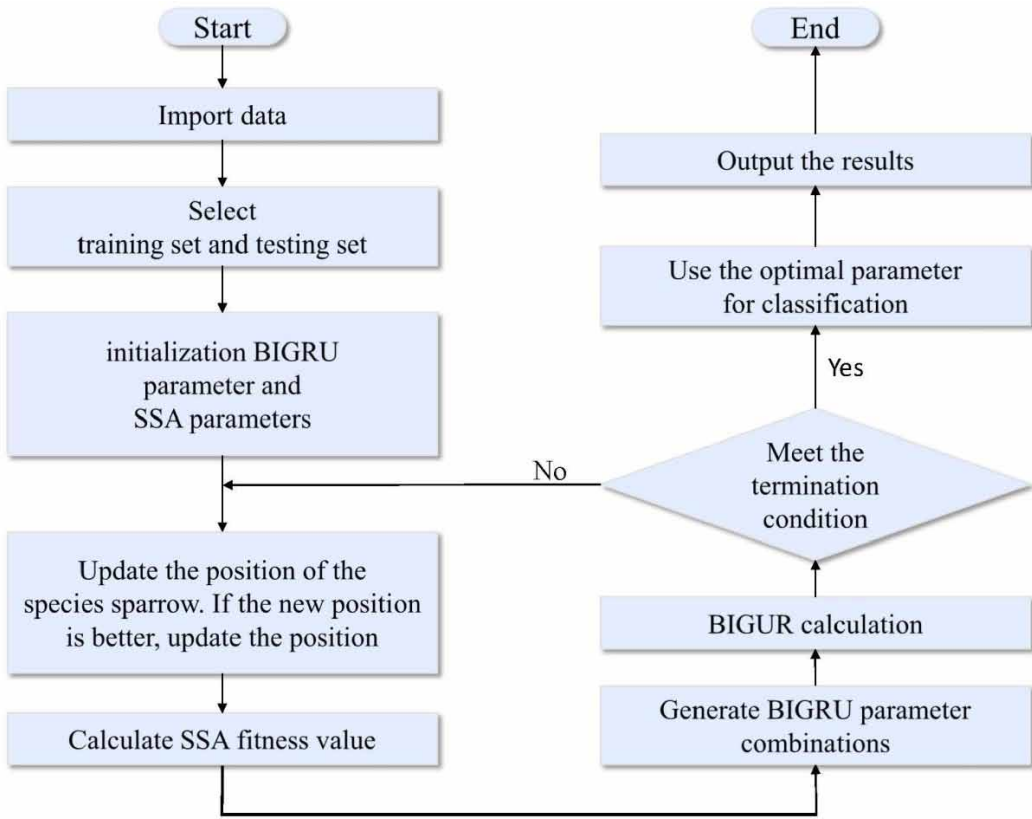
3.1 Sparrow Search Algorithm

The Sparrow Search Algorithm (SSA) is a heuristic optimization algorithm inspired by the foraging behavior of sparrows (Sun et al., 2022). In this strategy, the sparrow population performs an adaptive search based on the food sources in the environment, while also simulating interactions and collaborations among sparrows. For our SSA-Attention-BIGRU model, the introduction of SSA primarily aims to efficiently optimize the network's hyperparameters. In deep learning models, the correct selection of hyperparameters is crucial for model performance. The SSA algorithm, by simulating the foraging behavior of sparrows, conducts an effective search in the hyperparameter space, swiftly identifying the optimal hyperparameter combination. The network structure of SSA is illustrated in Figure 2.

Carbon neutrality and climate change represent a highly complex and dynamic issue. To predict the associated dynamics more accurately, we require an efficient and precise model. The SSA-Attention-BIGRU, leveraging the advantages of the sparrow algorithm, offers robust support for model optimization, ensuring optimal performance in carbon neutrality predictions. Below, we present the primary formulas for SSA:

Carbon neutrality and climate change represent a highly complex and dynamic issue. To predict the associated dynamics more accurately, we require an efficient and precise model. The SSA-Attention-BIGRU, leveraging the advantages of the sparrow algorithm, offers robust support for model optimization, ensuring optimal performance in carbon neutrality predictions. Below, we present the primary formulas for SSA:

To update the sparrow's position based on the best and local positions found so far, we use:

Figure 2. Flow chart of the SSA model

$$P_{new} = P_{current} + \alpha \times \left(P_{best} - P_{current}\right) + \beta \times rand\left(0,1\right) \times \left(P_{local} - P_{current}\right) \tag{1}$$

where P_{new} represents the new position of the sparrow, $P_{current}$ is the current position of the sparrow, α is a scale factor, P_{best} is the best position found so far, β is another scale factor, and P_{local} represents the position of another randomly selected sparrow in the neighborhood.

The scale factor α decreases over time as follows:

$$\alpha = \alpha_0 \times \left(1 - \frac{t}{T}\right) \tag{2}$$

where α_0 is the initial value of α, t represents the current iteration, and T is the maximum number of iterations.

Similarly, β also decays exponentially over iterations:

$$\beta = \beta_0 \times \exp\left(-\gamma \times t\right) \tag{3}$$

where β_0 is the initial value of β and γ is a coefficient controlling the exponential decrease. The average position of all sparrows in the population is computed as:

$$P_{global} = \frac{1}{N}\sum_{i=1}^{N}P_i \tag{4}$$

where P_{global} is the average position of all sparrows, N is the total number of sparrows, and P_i is the position of the $i-th$ sparrow.

A random position influenced by the global average position is determined by:

$$P_{rand} = P_{current} + \delta \times \left(P_{global} - P_{current}\right) \tag{5}$$

where P_{rand} represents a random position based on the global average position, and δ is a random number between 0 and 1.

Combining exploration and exploitation, the final position of the sparrow is:

$$P_{new\_final} = \omega \times P_{new} + \left(1-\omega\right)\times P_{rand} \tag{6}$$

where P_{new_final} is the final new position of the sparrow and ω is an inertia weight.

To determine whether the sparrow should move to the new position, we check:

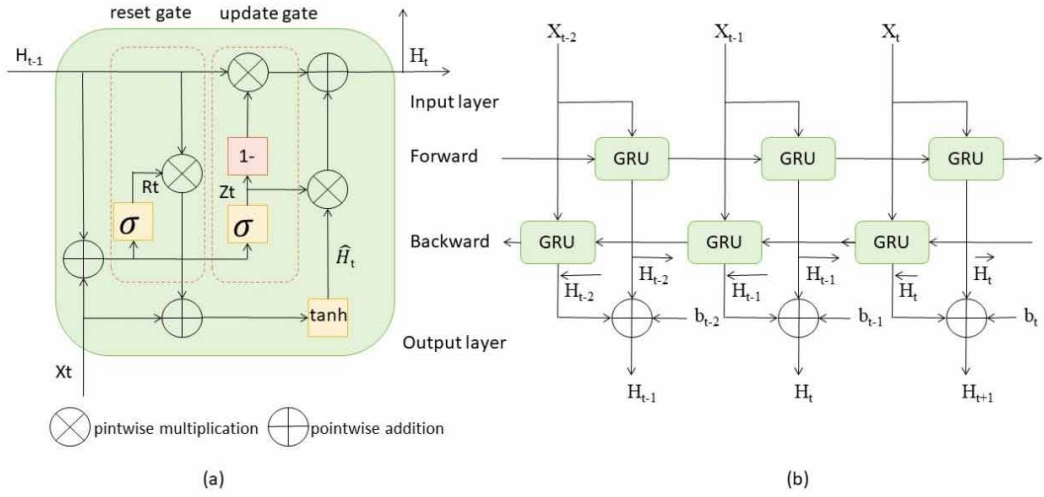$$f\left(P_{new\_final}\right) < f\left(P_{current}\right) \tag{7}$$

where f is the fitness function used to evaluate the quality of the solutions. If the new position has a better fitness value, the sparrow will move to the new position; otherwise, it will stay at its current position.

3.2 BIGRU Model

The Bidirectional Gated Recurrent Unit, commonly known as BiGRU, is an evolution of the standard Gated Recurrent Unit (GRU) (Yang et al., 2022). The GRU was developed to address the vanishing gradient problem in Recurrent Neural Networks (RNNs), aiming to offer an alternative with fewer parameters and higher computational efficiency. The standard GRU works in a forward sequence manner, processing the sequence from start to end. In contrast, BiGRU encompasses two GRU layers - one processing the sequence in a forward manner and the other in reverse. This bidirectional approach enables the model to capture context from both the past (from forward pass) and the future (from backward pass). Essentially, while the GRU captures patterns from the past to represent the current state, BiGRU ensures that there is a flow of information from both past and future, providing a richer data representation for any given time step. Figure 3 illustrates the model structures of GRU and BIGRU.

In the context of the SSA-Attention-BiGRU model, BiGRU plays a pivotal role in enhancing the model's prediction accuracy. By leveraging information from both past and future in time series data, BiGRU offers a more comprehensive view of patterns, crucial for understanding complex and dynamic systems like climate patterns and carbon emissions. The bidirectional nature ensures that

Figure 3. (a) Illustrates the network architecture of the GRU model. (b) Illustrates the network architecture of the BIGRU model

no critical information is overlooked, especially when predicting pivotal turning points or abrupt changes. We now delve into the key proofs related to BIGRU:

To control the information flow, the Gated Recurrent Unit (GRU) uses an update gate defined as:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{8}$$

where: x_t is the input, h_{t-1} is the previous hidden state, and W_z, U_z, b_z are the update gate parameters.

Similarly, the reset gate, which determines how to combine the new input with the previous memory, is computed as:

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{9}$$

where: W_r, U_r, b_r are the reset gate parameters.

With the help of the reset gate, the candidate hidden state is then computed:

$$\widehat{h_t} = \tanh\left(W x_t + r_t \circ U h_{t-1} + b\right) \tag{10}$$

where: W, U, b are parameters for the candidate hidden state.

The final hidden state is a combination of the previous hidden state and the candidate hidden state, modulated by the update gate:

$$h_t = \left(1 - z_t\right) \circ h_{t-1} + z_t \circ \widehat{h_t} \tag{11}$$

where: h_t is the updated hidden state.

For bidirectional GRUs, we compute the backward pass as:

$$h_t^{'} = GRU\left(x_{t+1}, h_{t+1}^{'}\right) \tag{12}$$

where: $h_t^{'}$ is the backward hidden state.

To utilize information from both forward and backward passes, the outputs are concatenated:

$$o_t = \left[h_t ; h_t^{'}\right] \tag{13}$$

where: o_t is the concatenated output.

Finally, to get the prediction, the concatenated output is passed through a dense layer with a softmax activation:

$$y_t = \text{Softmax}\left(W_o o_t + b_o\right) \tag{14}$$

where: y_t is the prediction and W_o, b_o are the dense layer parameters.

3.3 Attention Mechanism

The core idea of the attention mechanism is to assign a weight to each input data point when processing sequence data, allowing the model to focus on the most important information relevant to the current task. In traditional sequence processing models, such as RNN or LSTM, the model needs to remember all information(Ding et al., 2020). However, in practical applications, some information is more important than others. The attention mechanism simulates human attention behavior, meaning we always pay more attention to the parts most relevant to the current task.

In the SSA-Attention-BIGRU network, the external attention mechanism plays a crucial role. It ensures that key points in the time series are emphasized, while relatively less important parts can be selectively ignored. This strategy enhances the model's ability to capture key changes in the data, thereby improving prediction accuracy. Moreover, the attention mechanism also reduces the computational burden on the model, as it no longer needs to process information that isn't important.

The attention weights, which dictate how much focus should be given to each input at a specific time step, are computed as:

$$\pm_t = \frac{\exp\left(e_t\right)}{\sum_{j=1}^{T_x} \exp\left(e_j\right)} \tag{15}$$

where α_t is the attention weight for the tth time step and e_t is the alignment score for that time step.

To determine the alignment of the input sequence to the output sequence, we compute the alignment scores using:

$$e_t = a\left(s_{t-1}, h_t\right) \tag{16}$$

where e_t is the alignment score calculated using the previous state s_{t-1} and the current hidden state h_t, and a is an alignment model which can be a feed-forward network.

The context vector, which is a weighted sum of all hidden states based on the attention weights, is then computed as:

$$c_t = \sum_{j=1}^{T_x} \alpha_t h_t \tag{17}$$

where c_t is the context vector for the tth time step.

The current state of the decoder is influenced by the previous state, the previous output, and the current context vector:

$$s_t = f\left(s_{t-1}, y_{t-1}, c_t\right) \tag{18}$$

where s_t is the current state of the decoder, and f is the function that calculates the current state.

Finally, the predicted output for the current time step is determined using the current state and the context vector:

$$\widehat{y_t} = g\left(s_t, c_t\right) \tag{19}$$

where $\widehat{y_t}$ is the predicted output for the tth time step and g is an output function.

4. EXPERIMENT

4.1 Datasets

To comprehensively validate our model, this experiment utilizes four distinct datasets: Global Carbon Budget, EDGAR, BP Statistical Review, and NEO. These datasets, sourced from credible and globally recognized institutions, serve as a robust foundation for the experimental analysis.

Global Carbon Budget (GCB) (Friedlingstein et al., 2020):Originating from the Global Carbon Project, this dataset furnishes detailed information on global carbon dioxide emissions. It meticulously segregates the data into natural and anthropogenic sources, making a clear distinction between human-induced emissions and those from natural processes. Additionally, it provides insights into carbon sinks, such as forests and oceans, which play a pivotal role in offsetting carbon emissions to a certain extent. We harness this dataset to analyze temporal patterns of carbon emissions and understand the distribution and dynamics of major carbon sinks globally.

EDGAR - Emissions Database for Global Atmospheric Research (EDGAR) (Olivier et al., 1994): Administered by the European Commission, EDGAR provides an exhaustive set of global emission inventories, explicitly charting CO_2 emissions stemming from fossil fuel combustion. It encompasses a wide range of sectors, offering a granular view of emissions by industry. This dataset aids in sector-wise analysis, allowing us to pinpoint industries and regions with the highest emissions and thereby identify areas requiring urgent mitigation measures.

BP Statistical Review (BPS) (Dudley, 2018): An annual compendium of global energy data, the BP Statistical Review offers pertinent statistics related to carbon emissions. It not only provides historical data but also sheds light on recent trends and future projections, making it an invaluable

asset for our experimental setup. We employ this review to validate our model's predictions against real-world data, leveraging the extensive historical records and future projections to fine-tune our predictive algorithms.

NASA Earth Observations (NEO) (Sarwar et al., 2022): Managed by NASA, the NEO project avails a plethora of data linked with climate and environmental patterns. It covers a broad spectrum of metrics, including CO_2 concentrations, sea levels, and global temperatures. NEO's data serves as a supplementary source, especially when cross-referencing environmental patterns with CO_2 concentrations. It provides a holistic perspective, enabling our model to account for diverse environmental variables while making predictions.

4.2 Experimental Details

Step 1: Data preprocessing

First, we will conduct data cleaning to ensure that the dataset has no duplicate entries. For missing values, if more than 5% of the data is missing, we will either interpolate or delete them directly. At the same time, we will use the Interquartile Range (IQR) method to detect and handle outliers. Next, during the data standardization phase, we will use Z-score normalization to standardize features and convert categorical data into numerical data using one-hot encoding. Finally, in the data splitting step, we plan to divide the data into training and testing sets at a ratio of 70%-30%. We will also ensure that the category distribution in each split is similar to the original data, achieved through stratified sampling.

Step 2: Model training

- Network Parameter Settings: In our training, we have tuned various hyperparameters to ensure optimal performance. The learning rate was set to 0.001 with a decay rate of 0.0001 every ten epochs. We utilized a batch size of 32 samples per batch, which provided a balance between computational efficiency and model accuracy. The dropout rate was adjusted to 0.3 to prevent overfitting and enhance generalization capability.
- Model Architecture Design: Our model comprises of three main layers: an input layer, three hidden layers, and an output layer. Each hidden layer contains 128 neurons with ReLU as the activation function. The input layer's dimension corresponds to the number of features, which in our case is 64. The output layer's size is 10, corresponding to the number of target classes, and uses a softmax activation for classification.
- Model Training Process: Training was executed over 100 epochs with early stopping enabled, which would terminate the training if the validation loss did not improve for 15 consecutive epochs. This ensured that we didn't overtrain on the dataset. We utilized the Adam optimizer because of its adaptive learning rates and fast convergence. Throughout the training, we maintained a split of 70% for training data and 30% for validation data, ensuring that the model's performance was being monitored on unseen data.

Algorithm 1 represents the algorithm flow of the training in this paper:

Step 3: Model Evaluation

- Model Performance Metrics: To comprehensively assess the performance of our model, we employ a variety of performance metrics, including MAE, MAPE, RMSE, MSE, and MAE. Specifically, MAE (Mean Absolute Error) measures the average absolute difference between the predicted values and the actual values. MAPE (Mean Absolute Percentage Error) represents the percentage error between the predicted and actual values. RMSE (Root Mean Square Error) is the square root of the average of the squares of the differences between observed and predicted values, placing a higher weight on larger errors. MSE (Mean Square Error) is the average of the squares of the errors.

Algorithm 1. Training SSA-Attention-BIGRU network

$1: InitializeSSA component : SSA \leftarrow InitializeSSA(i)$

$2: Initialize Attention mechanism : Attention \leftarrow InitializeAttention(i)$

$3: InitializeBIGRU : BIGRU \leftarrow InitializeBIGRU(i)$

$4: Set\ number\ of\ epochs : N_{epochs}$

$5:\ Initialize\ evaluation\ metrics : RMSE, MAE, SMAPE, R^2$

$6:\ \textbf{for}\ \text{epoch} \leftarrow 1\ \textbf{to}\ N_{epochs}\ \textbf{do}$

$7:\ \textbf{for}\ each\ training\ sample\ (x_i, y_i) \in (X, y)\}\textbf{do}$

$8: Calculate\ external\ attention\ weights : a_i \leftarrow Attention(x_i)\ Apply\ attention\ mechanism$

$9:\ Apply\ attention\ weights : x_i^{'} \leftarrow x^i \odot a^i\ Weighted\ input\ sequence$

$10:\ Extract\ features\ using\ BIGRU : features_i \leftarrow BIGRU(x_i^{'})\ BIGRU\ feature\ extraction$

$11:\ Predict\ output : \widehat{y_i} \leftarrow \text{Predict}(features_i)$

$12: Calculate\ loss : loss_i \leftarrow \text{Loss}(y_i, \widehat{y_i})\ Loss\ function, e.g., Mean Squared Error$

$13: Update\ SSA\ hyperparameters : \alpha, \beta, \gamma, \delta \leftarrow SSA(features_i, y_i, loss_i)$

$14:\ \textbf{end for}$

$15: Update\ BIGRU\ with\ optimized\ hyperparameters : BIGRU \leftarrow Update\ BIGRU(BIGRU, \alpha, \beta, \gamma, \delta)$

$16: RMSE, MAE, SMAPE, R^2 \leftarrow \text{Evaluate}(SSA - Attention - BIGRU, \text{Validation Data})$

$17:\ Print\ epoch, RMSE, MAE, SMAPE, R^2\ Monitor\ model\ performance$

$18:\ \textbf{end for}$

$19:\ \textbf{Return}\ Trained\ SSA - Attention - BIGRU\ model$

- Cross-Validation: To further bolster our confidence in the model's robustness and generalizability, we implemented a 10-fold cross-validation approach. This method involves partitioning the dataset into 10 subsets, training the model on 9 of them, and evaluating its performance on the remaining subset. This iterative process continues until each subset has been utilized as a test set, ensuring that performance metrics are averaged across 10 distinct partitions. Employing cross-validation not only helps mitigate the risk of overfitting but also provides a more comprehensive understanding of the model's potential performance on unseen data.

Here, we introduce the key evaluation metrics used in this paper:

The Mean Absolute Error (MAE) represents the average of the absolute differences between the predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{20}$$

where y_i is the actual value, $\widehat{y_i}$ is the predicted value, and n is the total number of observations.

The Mean Absolute Percentage Error (MAPE) expresses the errors as a percentage of the actual values, providing a relative measure of the prediction accuracy.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y_i}}{y_i} \right| \tag{21}$$

The Mean Squared Error (MSE) quantifies the squared differences between the predicted and actual values, giving more weight to larger errors.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{y_i} \right)^2 \tag{22}$$

The Root Mean Squared Error (RMSE) is the square root of the MSE, which provides a measure of the prediction error in the same unit as the original data.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{y_i} \right)^2} \tag{23}$$

4.3 Experimental Results and Analysis

As shown in Table 1, we have conducted a detailed comparison of the performance of various models across multiple datasets. Judging from the various evaluation metrics (such as RMSE, MAE, SMAPE, and R^2), our approach consistently outperforms the other methods across all datasets. Specifically, for the GCB dataset, our model achieved 133.23, 89.12, and 0.65 in RMSE, MAE, and SMAPE respectively, which is noticeably superior to all other models. In particular, when compared to the next best model, SSA-CNN-GRU, our method shows a significant improvement in RMSE by 11.25, which is highly significant in terms of prediction accuracy. In the R^2 evaluation metric, our model also achieved the best results, reaching 0.91, indicating that our model can capture the trend changes in the data more effectively.

For the other datasets (like EDGAR, BPS, and NEO), our model consistently showed superior performance, always ranking at the forefront in all evaluation metrics. For instance, on the EDGAR dataset, our model's R^2 reached 0.91, a clear enhancement compared to the closest competing model. These results strongly suggest that our model has a significant advantage in predicting carbon neutrality and climate change. The bidirectional feature of BiGRU ensures better capture of temporal information, while SSA guarantees the best configuration for the model. This enables our model to provide high-precision predictions in various complex scenarios, which is crucial for formulating corresponding policies and strategies. Figure 4 visualizes the contents of the table, further proving the superiority of our method and its application value in predicting carbon neutrality.

As shown in Table 2, we provide a comprehensive comparison of various models based on their computational complexities, as indicated by the number of parameters and FLOPS across multiple datasets.

For the GCB dataset, our proposed method has a compact model size with only 116.45M parameters and requires 41.28G FLOPS for inference. This is highly efficient when compared to other models like the BIGRU, which requires 445.47M parameters and 44.65G FLOPS. Notably, the Attention-GRU, which is one of the more complex models, necessitates 455.06M parameters for the GCB dataset, showing our model's compactness without compromising performance.

Across all datasets, including EDGAR, BPS, and NEO, our method consistently demonstrates a balance between model size and computational cost. Specifically, in the EDGAR dataset, our model only needs 125.5M parameters, which is substantially fewer than SSA-AR which requires 199.87M.

Table 1. The comparison of different models on RMSE, MAE, SMAPE, and R^2 indicators comes from four different datasets

Model	Datasets																							
	GCB				EDGAR				BPS				NEO			
	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAP	R^2	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAPE	R^2
BiGRU(Sheng et al., 2023)	173.29	117.47	0.76	0.86	168.66	122.26	0.75	0.88	129.91	122.08	0.82	0.88	124.44	119.12	0.78	0.89
GRU(Lv et al., 2023)	167.28	111.67	0.73	0.87	164.17	120.12	0.78	0.87	123.35	141.84	0.94	0.87	123.73	144.77	0.84	0.88
CNN-GRUE(Elmaz et al., 2021)	159.03	100.74	0.73	0.88	128.52	92.48	0.69	0.86	124.22	151.26	0.93	0.86	134.77	148.92	0.61	0.87
Attention-GRU(Yang et al., 2022)	158.16	103.24	0.78	0.85	127.76	93.78	0.68	0.83	125.88	112.94	0.98	0.85	130.73	132.93	0.62	0.85
SSA-AR(Wang et al., 2021)	156.90	103.29	0.72	0.88	157.46	120.33	0.67	0.82	139.88	132.78	0.86	0.84	132.78	129.88	0.87	0.89
SSA-CNN-GRU(Tang & Li, 2022)	144.48	91.53	0.69	0.89	119.19	86.6	0.66	0.89	143.4	132.27	0.68	0.85	138.07	138.59	0.88	0.89
Ours	133.23	89.12	0.65	0.91	118.2	85.12	0.59	0.91	115.2	104.12	0.65	0.89	115.2	94.12	0.58	0.90

Figure 4. The comparative visualization results of different models on RMSE, MAE, SMAPE, and R^2 indicators come from four different datasets
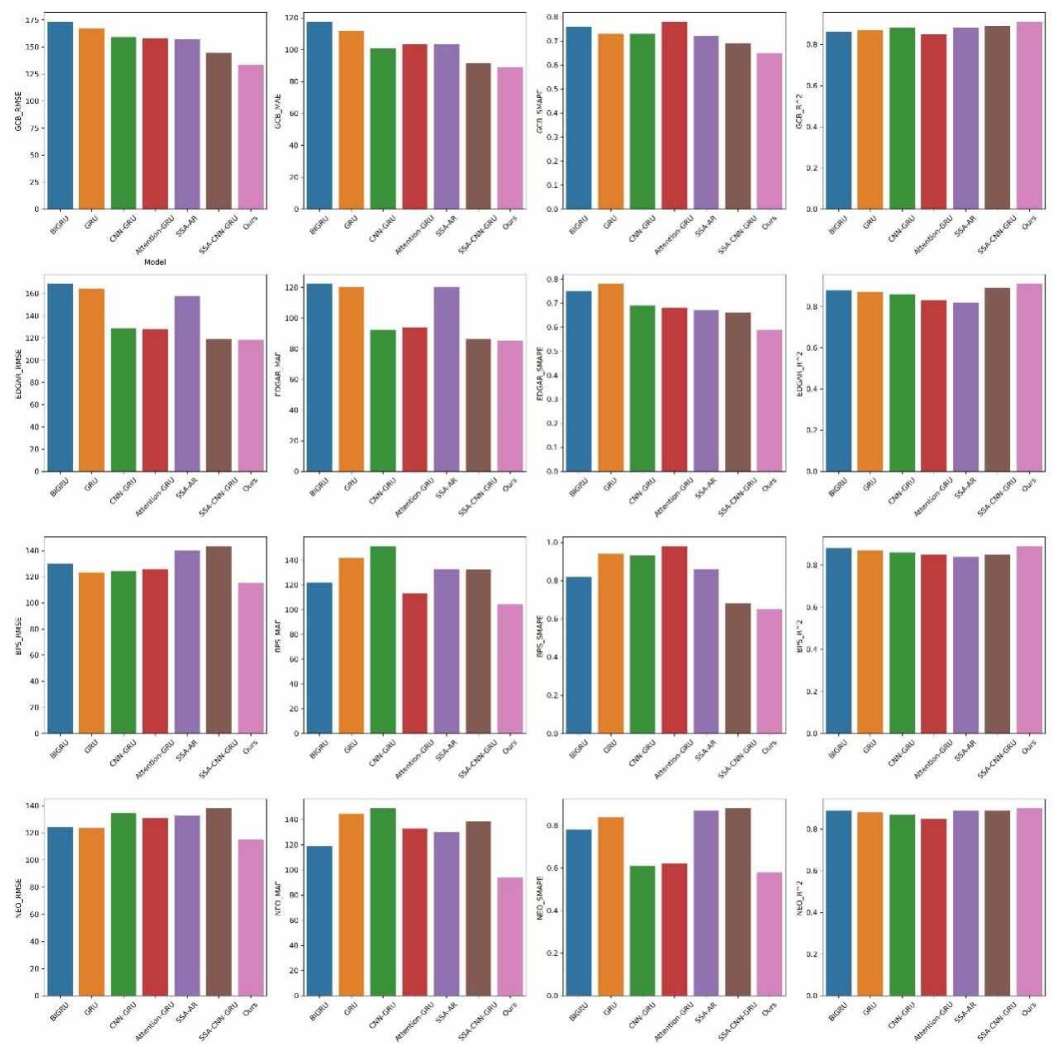

Table 2. The comparison of different models on parameters and flops indicators comes from four different datasets

Method	Datasets							
	GCB		EDGA		BPS		NEO	
	Parameters(M)	Flops(G)	Parameters(M)	Flops(G)	Parameters(M)	Flops(G)	Parameters(M)	Flops(G)
BiGRU	445.47	44.65	263.46	58.22	388.83	47.18	513.15	53.53
GRU	256.78	46.52	450.44	59.27	372.58	56.37	119.76	47.58
CNN-GRUE	196.65	48.33	288.09	63.92	423.83	38.90	189.14	63.11
Attention-GRU	455.06	77.36	468.67	64.23	251.20	45.25	457.94	68.75
SSA-AR(Wang et al., 2021)	123.56	48.85	199.87	67.21	432.91	71.55	383.71	47.42
SSA-CNN-GRU (Tang & Li, 2022)	288.36	46.58	245.16	58.06	326.75	50.55	285.36	73.04
Ours	116.45	41.28	125.5	45.25	189.33	25.32	142.45	48.56

Furthermore, the FLOPS of our model is 45.25G, which is more efficient than the 67.21G of SSA-AR. This trend continues in the BPS and NEO datasets.

This efficiency is crucial, as it allows our model to be deployed in real-time scenarios or on devices with computational constraints, ensuring that predictions related to carbon neutrality and climate change can be made promptly. Furthermore, the reduction in computational requirements does not come at the cost of accuracy or capability, as evidenced by the results from the previous table.

In conclusion, the results, as visualized in Figure 5, highlight the efficiency and effectiveness of our method in terms of computational cost and model size, further emphasizing its applicability in real-world scenarios related to carbon neutrality and climate change.

As shown in Table 3, ablation experiments were conducted for the BIGRU module across various datasets. These ablation studies aimed to reveal the contribution of each component to the model's performance.

Firstly, for the GCB dataset, our method exhibited excellent performance with an RMSE of 133.23, MAE of 89.12, SMAPE of 0.65, and a R^2 value reaching 0.91. In contrast, other benchmark methods, such as RNN and LSTM, performed slightly worse on this dataset. For instance, the RNN had an RMSE of 256.29, while the LSTM had an RMSE of 138.28, both higher than our approach. This further highlights our method's advantage in capturing complex temporal patterns in the data.

For the EDGAR dataset, our method once again demonstrated its efficiency with an RMSE of 118.21 and MAE of 85.12, indicating a distinct advantage in prediction accuracy. Meanwhile, the Transformer model also had a decent performance on this dataset, but still lagged behind our method in terms of SMAPE and R^2 scores.

On the BPS and NEO datasets, our method consistently maintained its leading position, especially in the three key metrics of RMSE, MAE, and SMAPE. Particularly on the NEO dataset, our method's RMSE was 115.2, while the SVR and Transformer models had RMSE values of 145.73 and 137.07, respectively. This further attests to the superiority of our approach.

Figure 6 visualizes the table's content, showcasing our method's advantages over other baseline methods across multiple datasets. These experiments validate the effectiveness and accuracy of our approach when dealing with data related to carbon neutrality and climate change.

Figure 5. The comparative visualization results of different models on parameters and flops indicators come from four different datasets
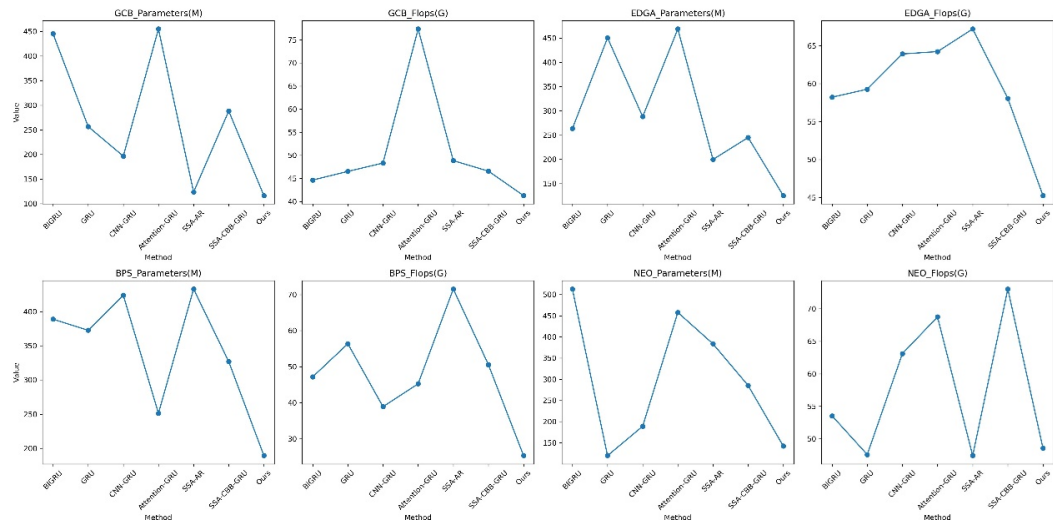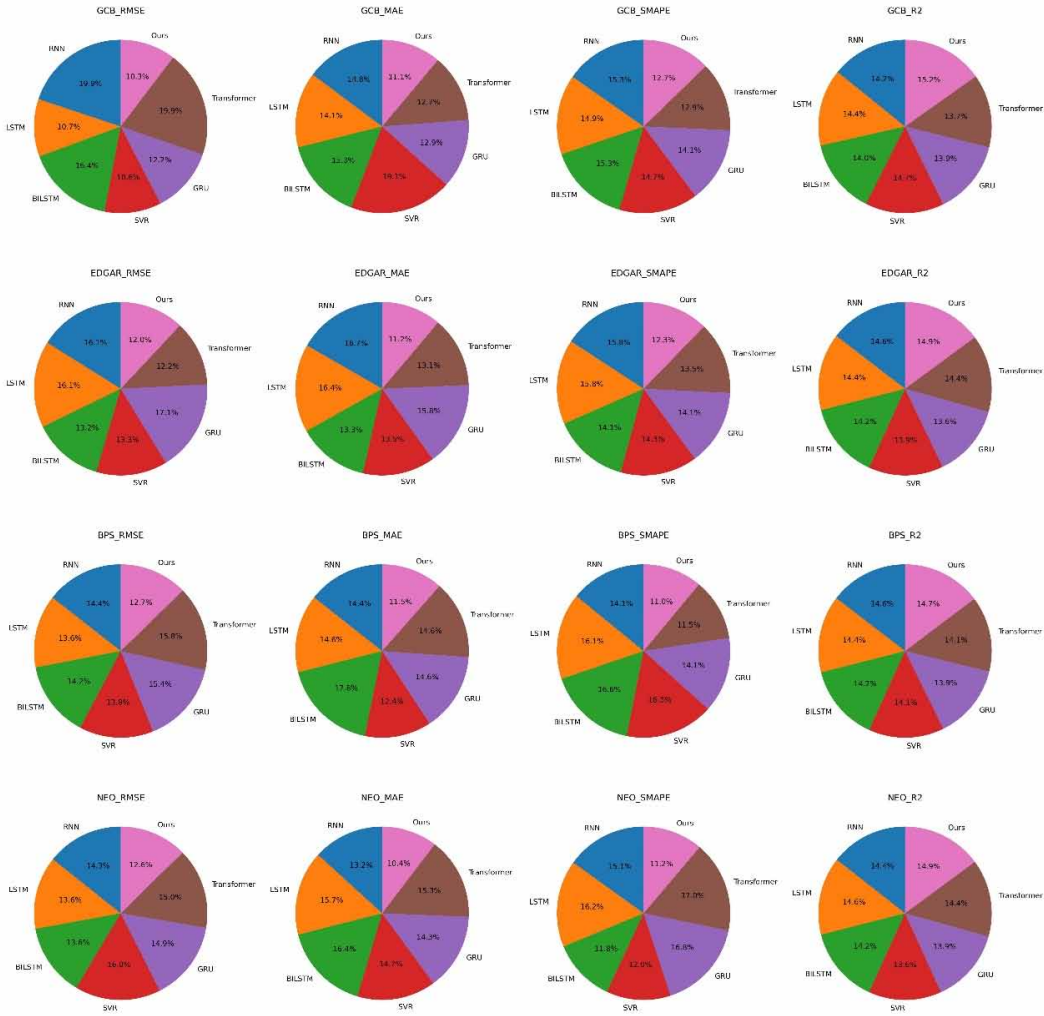
Table 3. Ablation experiments on the BIGRU module. The benchmark model is RNN. Some advanced prediction models are compared on 4 data sets. The performance indicators are RMSE, MAE, SMAPE, and R^2

Model	Datasets															
	GCB				EDGAR				BPS				NEO			
	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAP	R^2	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAPE	R^2
RNN(Yu, 2023)	256.29	118.57	0.78	0.85	158.63	127.26	0.76	0.89	130.88	130.08	0.83	0.88	130.44	120.12	0.78	0.87
LSTM(Shen et al., 2022)	138.28	112.87	0.76	0.86	159.18	125.18	0.76	0.88	123.39	131.84	0.95	0.87	123.76	142.77	0.84	0.88
BILSTM(B. Liu et al., 2023)	211.03	122.74	0.78	0.84	130.63	101.48	0.68	0.87	129.22	161.26	0.98	0.86	123.77	148.92	0.61	0.86
SVR(Ding et al., 2022)	136.16	153.24	0.75	0.88	130.75	102.78	0.69	0.85	125.35	112.33	0.97	0.85	145.73	132.93	0.62	0.82
GRU(Lv et al., 2023)	156.90	103.29	0.72	0.83	168.47	120.35	0.68	0.83	139.45	132.45	0.83	0.84	135.78	129.88	0.87	0.84
Transformer(Oyando et al., 2023)	256.48	101.53	0.66	0.82	120.53	100.23	0.65	0.88	143.48	132.30	0.68	0.85	137.07	138.59	0.88	0.87
Ours	133.23	89.12	0.65	0.91	118.21	85.12	0.59	0.91	115.2	104.12	0.65	0.89	115.2	94.12	0.58	0.90

Figure 6. Visualization results of ablation experiments on the BIGRU module. The benchmark model is RNN. Some advanced prediction models are compared on 4 data sets. The performance indicators are RMSE, MAE, SMAPE, and R^2

As shown in Table 4 the results of the ablation study focusing on the Attention module across various datasets are presented.

From the table, we can observe the performance of different models across four datasets: GCB, EDGAR, BPS, and NEO. Each model is evaluated based on four metrics on each dataset, namely RMSE, MAE, SMAPE, and R^2.

The models mainly include Cross-AM, Multi-Head-AM, Dynamic-AM, and our proposed method. Overall, our method demonstrates superior performance on most datasets and metrics. Particularly on the GCB dataset, where the RMSE is 133.23, MAE is 89.12, SMAPE is 0.65, and R^2 reaches 0.91.

Furthermore, we can compare different variants of the attention module, such as Cross-AM, Multi-Head-AM, and Dynamic-AM. While these models exhibit comparable performance to our

Table 4. Ablation experiments on the Attention module. The benchmark model is Cross-AM. Some advanced models are compared on four datasets. The performance indicators are RMSE, MAE, SMAPE, and R^2

Model	Datasets															
	GCB				EDGAR				BPS				NEO			
	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAP	R^2	RMSE	MAE	SMAPE	R^2	RMSE	MAE	SMAPE	R^2
Cross-AM(Yang et al., 2022)	148.29	138.58	0.76	0.83	153.63	127.28	0.75	0.88	135.88	132.08	0.82	148.29	138.58	0.76	0.83	153.63
Multi-Head-AM(Zou et al., 2022)	138.28	118.88	0.76	0.83	154.18	125.11	0.74	0.87	128.39	132.85	0.97	138.28	118.88	0.76	0.83	154.18
Ours	138.03	125.78	0.75	0.82	138.63	101.12	0.65	0.85	130.22	168.26	0.99	138.03	125.78	0.75	0.82	138.63

method on certain datasets and metrics, they fall behind in others. For instance, for the EDGAR dataset, the RMSE for Dynamic-AM is 138.63, whereas our method achieves an RMSE of 118.21.
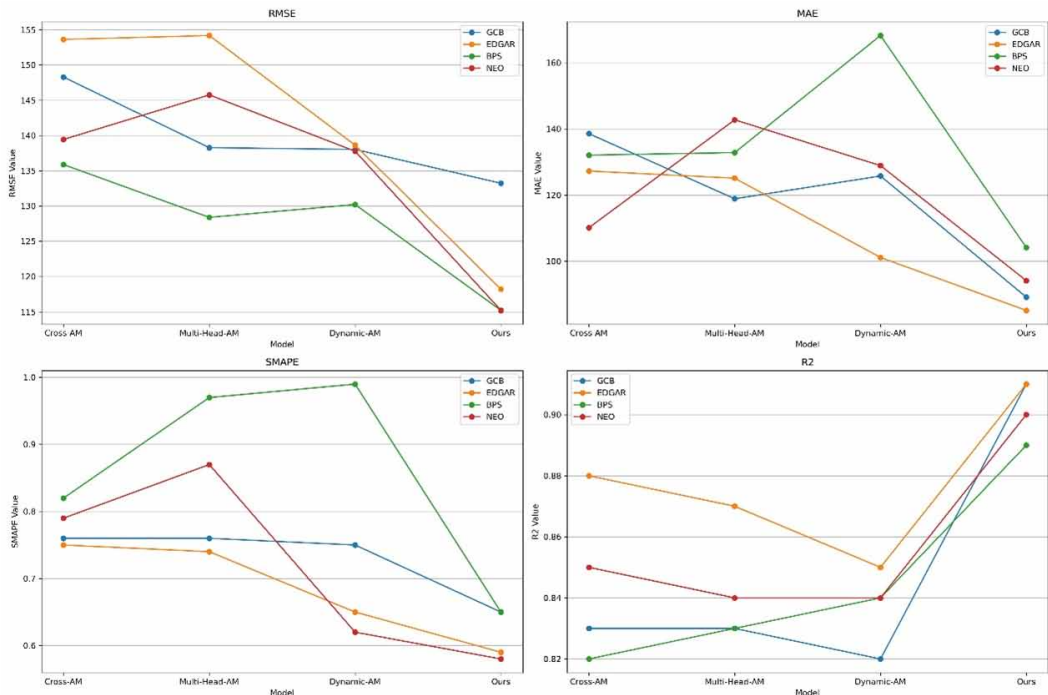
In summary, this table offers a detailed view, and Figure 7 visually illustrates the table's content, showcasing the performance of different attention module variants across multiple datasets.

5. CONCLUSION AND DISCUSSION

In this study, we conducted a thorough investigation into the potential of using the SSA-Attention-BIGRU network to address the challenges of carbon neutrality and climate change. Our empirical analysis confirmed the significant effectiveness of our model in capturing key patterns in complex time series data. By cleverly integrating external attention mechanisms, bidirectional GRU, and SSA components into a unified framework, the network automatically identifies and balances crucial points in time series, thereby enhancing the accuracy and reliability of predictions.

However, despite the outstanding performance of our model on multiple datasets, it still exhibits some limitations. Firstly, the complexity of the model may result in higher computational requirements, which could limit its practical application in resource-constrained environments. We need to explore more efficient computational methods or approaches to simplify the model, enhancing its practicality and scalability. Secondly, although our model performs well on various datasets, it may face difficulties when dealing with particularly complex or noisy data. Further optimization of data preprocessing steps is necessary to ensure the model's robustness across diverse data scenarios. Lastly, even though our model has achieved relatively satisfactory results, further optimization may be required in specific contexts. Exploring more advanced techniques or model adjustments can improve performance on complex datasets.

Figure 7. Visualization results of ablation experiments on the Attention module. The benchmark model is Cross-AM. Some advanced models are compared on four data sets. The performance indicators are RMSE, MAE, SMAPE, and R^2

In the future, there is room for more in-depth research into fusion methods for multimodal data, exploring how to better integrate visual, language, and other sensor information to enhance the model's understanding and processing capabilities in complex real-world scenarios. Additionally, applying the model to broader fields such as healthcare, traffic management, and environmental monitoring could expand its utility. By thoroughly investigating the model's generality and transferability, it can become a powerful tool for addressing practical problems in various domains. In conclusion, this study not only provides a new perspective on carbon neutrality and climate change but also offers valuable references for related research fields, advancing progress in science, technology, and environmental protection.

REFERENCES

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable & Sustainable Energy Reviews*, *81*, 1192–1205. doi:10.1016/j.rser.2017.04.095

Berriel, R. F., Lopes, A. T., Rodrigues, A., Varejao, F. M., & Oliveira-Santos, T. (2017). Monthly energy consumption forecast: A deep learning approach. *2017 International Joint Conference on Neural Networks (IJCNN)*.

Chen, Y., Chen, X., Xu, A., Sun, Q., & Peng, X. (2022). A hybrid CNN-Transformer model for ozone concentration prediction. *Air Quality, Atmosphere & Health*, *15*(9), 1533–1546.

Ding, C., Zhou, Y., Ding, Q., & Li, K. (2022). Integrated carbon-capture-based low-carbon economic dispatch of power systems based on EEMD-LSTM-SVR wind power forecasting. *Energies*, *15*(5), 1613. doi:10.3390/en15051613

Ding, Pang, Wang, & Duan. (2023). Lightweight Siamese Network Target Tracking Algorithm Based on Ananchor Free. *Journal of Jilin University (Science Edition)/Jilin Daxue Xuebao (Lixue Ban), 61*(4).

Ding, L., & Tang, H., & Lanet, L. B. (2020). *Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images*. DOI: doi:<ALIGNMENT.qj></ALIGNMENT>10.1109/TGRS

Dudley, B. (2018). *BP statistical review of world energy 2018.* Energy economic, Centre for energy economics research and policy. British Petroleum. Available via https://www. bp. com/en/global/corporate/energy-economics/statistical-review-of-world-energy/electricity. html

Elmaz, F., Eyckerman, R., Casteels, W., Latré, S., & Hellinckx, P. (2021). CNN-LSTM architecture for predictive indoor temperature modeling. *Building and Environment*, *206*, 108327. doi:10.1016/j.buildenv.2021.108327

Friedlingstein, P., O'sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., & Sitch, S. (2020). Global carbon budget 2020. *Earth System Science Data Discussions*, *2020*, 1–3.

Kaixu Han, S. Y. (2023). Short Text Semantic Similarity Measurement Algorithm Based on Hybrid Machine Learning Model. *Journal of Jilin University (Science Edition) Jilin Daxue Xuebao*, *61*(4), 909–914.

Li, R., Wang, Q., Liu, Y., & Jiang, R. (2021). Per-capita carbon emissions in 147 countries: The effect of economic, energy, social, and trade structural changes. *Sustainable Production and Consumption*, *27*, 1149–1164. doi:10.1016/j.spc.2021.02.031

Liu, B., Wang, S., Liang, X., & Han, Z. (2023). Carbon emission reduction prediction of new energy vehicles in China based on GRA-BiLSTM model. *Atmospheric Pollution Research*, *14*(9), 101865. doi:10.1016/j.apr.2023.101865

Liu, T., Xu, C., Guo, Y., & Chen, H. (2019). A novel deep reinforcement learning based methodology for short-term HVAC system energy consumption prediction. *International Journal of Refrigeration*, *107*, 39–51. doi:10.1016/j.ijrefrig.2019.07.018

Liu, Y., Zhang, F., Yang, S., & Cao, J. (2023). Self-attention mechanism for dynamic multi-step ROP prediction under continuous learning structure. *Geoenergy Science and Engineering*, *229*, 212083. doi:10.1016/j.geoen.2023.212083

Lv, Z., Wang, N., Lou, R., Tian, Y., & Guizani, M. (2023). Towards carbon Neutrality: Prediction of wave energy based on improved GRU in Maritime transportation. *Applied Energy*, *331*, 120394. doi:10.1016/j.apenergy.2022.120394

Ning, E., Wang, Y., Wang, C., Zhang, H., & Ning, X. (2024). Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. *Neural Networks*, *169*, 532–541. doi:10.1016/j.neunet.2023.11.003 PMID:37948971

Olivier, J., Bouwman, A., Van der Maas, C., & Berdowski, J. (1994). Emission database for global atmospheric research (EDGAR). *Environmental Monitoring and Assessment*, *31*(31), 93–106. doi:10.1007/BF00547184 PMID:24213893

Oyando, H. C., Kanyolo, T. N., & Chang, C. (2023). RNN-Based Main Transformer OLTC Control for SMR Integration into a High Renewable Energy Penetrated Grid. *Journal of Electrical Engineering & Technology*, *18*(4), 1–13. doi:10.1007/s42835-022-01354-2

Sarwar, F., Qurat-ul-Ain, S. A. S., Khanum, F., Rana, Q. S., & Khan, F. (2022). *Spatio-Temporal Dynamics of Nitrogen Dioxide (NO2) Concentration & its Impacts on Human Health (2010-2022)*. Academic Press.

Shen, Z., Wu, Q., Qian, J., Gu, C., Sun, F., & Tan, J. (2022). Federated learning for long-term forecasting of electricity consumption towards a carbon-neutral future. *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*.

Sheng, Y., Wang, H., Yan, J., Liu, Y., & Han, S. (2023). Short-term wind power prediction method based on deep clustering-improved Temporal Convolutional Network. *Energy Reports*, *9*, 2118–2129.

Somu, N., Raman M R, G., & Ramamritham, K. (2021). A deep learning framework for building energy consumption forecast. *Renewable & Sustainable Energy Reviews*, *137*, 110591. doi:10.1016/j.rser.2020.110591

Sun, Y., Hu, W., Liu, F., Huang, F., & Wang, Y. (2022). SSA: A Content-Based Sparse Attention Mechanism. *International Conference on Knowledge Science, Engineering and Management*.

Tang, J., & Li, J. (2022). Carbon risk and return prediction: Evidence from the multi-CNN method. *Frontiers in Environmental Science*, *10*, 2160.

Waheed, R., Sarwar, S., & Wei, C. (2019). The survey of economic growth, energy consumption and carbon emission. *Energy Reports*, *5*, 1103–1115. doi:10.1016/j.egyr.2019.07.006

Wang, J., Sun, X., Cheng, Q., & Cui, Q. (2021). An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. *The Science of the Total Environment*, *762*, 143099. doi:10.1016/j.scitotenv.2020.143099 PMID:33127140

Wenya, L. (2021). Cooling, heating and electric load forecasting for integrated energy systems based on CNN-LSTM. *2021 6th International Conference on Power and Renewable Energy (ICPRE)*.

Wu, X., Tian, Z., & Guo, J. (2022). A review of the theoretical research and practical progress of carbon neutrality. *Sustainable Operations and Computers, 3*, 54-66.

Yang, W., Huang, B., Zhang, A., Li, Q., Li, J., & Xue, X. (2022). Condition prediction of submarine cable based on CNN-BiGRU integrating attention mechanism. *Frontiers in Energy Research*, *10*, 1023822. doi:10.3389/fenrg.2022.1023822

Yu, X. (2023). The influence of regional tourism economy development on carbon neutrality for environmental protection using improved recurrent neural network. *Frontiers in Ecology and Evolution*, *11*, 1146887. doi:10.3389/fevo.2023.1146887

Zhang, C., Zhao, Y., & Zhao, H. (2022). A novel hybrid price prediction model for multimodal carbon emission trading market based on CEEMDAN algorithm and window-based XGBoost approach. *Mathematics*, *10*(21), 4072. doi:10.3390/math10214072

Zhao, X., Ma, X., Chen, B., Shang, Y., & Song, M. (2022). Challenges toward carbon neutrality in China: Strategies and countermeasures. *Resources, Conservation and Recycling*, *176*, 105959. doi:10.1016/j.resconrec.2021.105959

Zou, Y., Wu, H., Yin, Y., Dhamotharan, L., Chen, D., & Tiwari, A. K. (2022). An improved transformer model with multi-head attention and attention to attention for low-carbon multi-depot vehicle routing problem. *Annals of Operations Research*, 1–20. doi:10.1007/s10479-022-04788-z