Machine Learning-Based Prediction and Benefit Analysis of Photovoltaic Power Generation

Ziyi Qing [1, *], Jiangshan Li [2], Man Wang [1], Ruihuan Wang [2], Zixun Qing [3]

[1] College of Electrical Engineering, Xinjiang University,Urumqi, China, 830049

[2] College of Intelligent Science and Technology, Xinjiang University, Urumqi, China, 830049

[3] College of Humanities and Law, Hebei University of Technology, Tianjin, China,300131

* Corresponding Author Email: qzy2196595135@163.com

Abstract. The increasing global demand for energy and the environmental impact of fossil fuels have propelled the advancement of solar photovoltaic (PV) technology. This paper emphasizes the importance of accurately predicting the efficiency and cost-effectiveness of solar PV systems for informed market and investment decisions. We adopted a machine learning approach due to the inadequacy of traditional models, utilizing historical data from household PV panels. We evaluated three models: Gradient Boosted Regression Tree (GBRT), Bi-Directional Gated Recurrent Unit (BiGRU), and Random Forest (RF). The results indicated that the BiGRU model excels in prediction accuracy, while the RF model demonstrates superior stability and robustness to outliers. Furthermore, the study assesses the economic and environmental returns of household PV systems, concluding that their installation offers significant environmental and financial benefits under a fully grid-connected operation.

Keywords: Domestic Photovoltaic Panels, Machine Learning, Prediction, Economic Benefits.

1. Introduction

As global climate change intensifies and energy consumption continues to grow, the environmental concerns of traditional fossil fuels and the risk of resource depletion have forced countries around the world to seek more sustainable energy solutions [1]. In this context, solar photovoltaic (PV) technology has become a key component of the global energy transition due to its clean and renewable nature [2]. The rapid growth of installed PV capacity worldwide reflects the widespread acceptance and adoption of this technology, while technological advances and continued cost reductions have further fuelled the expansion of the PV industry. However, despite the increasing economics of PV systems, accurate prediction of their power generation efficiency and cost-effectiveness remains a key factor in market expansion and technology investment decisions [3].

With the development of information technology, machine learning has been widely used in several fields, including natural language processing, finance and insurance, image recognition and processing, healthcare, and bioscience. In this study, to solve the problem of predicting the power generation of domestic photovoltaic panels, three machine learning models, namely, Gradient Boosted Regression Tree (GBRT), Bi-directional Gated Recurrent Unit (BIGRU), and Random Forest (RF), are used. These models were first proposed by CHO; LEO BREIMAN; FRIEDMAN et al. The GBRT model reduces the prediction error by gradually adding new decision trees, outperforms the traditional regression model, and is particularly suitable for dealing with nonlinear and complex data relationships [4]. The BIGRU model, as an improved recurrent neural network, can efficiently capture the forward and backward dependencies in time series data, improving the accuracy and stability of prediction [5]. Random forest, on the other hand, enhances the generalization ability of the model and robustness to outliers by constructing multiple decision trees and combining their predictions.

In this study, these models were used to analyze the power generation data from domestic PV panels in Jize County, Handan City, Hebei Province, to predict the power generation. The results obtained not only confirm the feasibility of employing these advanced machine learning techniques

in PV panel power generation prediction but also provide households and businesses with important information about installed capacity and potential economic returns.

2. Data Sources and Methodology

The data in this paper comes from a construction engineering company in Hebei, and the monthly power generation data and corresponding installed capacity of each power station are collected through the software Tianfutong.

Gradient boosted regression tree (GBRT) corrects the prediction error by adding decision trees step by step, similar to gradient descent [6,7]. Each step trains a new model on the data residuals to achieve the purpose of reducing the residuals, combining to form a comprehensive model, and effectively reducing the prediction error. GBRT is good at dealing with complex nonlinear relationships, and the appropriate learning rate helps to control the fitting speed and prevent overfitting.

Bidirectional gated recurrent unit (BiGRU) is a variant of recurrent neural network designed to process sequence data [8]. It runs GRU layers independently in both directions of the sequence, merging information to capture forward and backward dependencies. GRU introduces update gates and reset gates to help solve the gradient problem for long sequence training and improve the accuracy of sequence prediction [9].

Random Forest is an integrated learning method that improves prediction by constructing multiple decision trees and averaging or majority voting. Each tree is trained using random samples and features to increase model generalisation and reduce variance. It shows strong robustness and accuracy especially in predicting the nonlinear relationship between installed capacity and power generation.

In this paper, we predict monthly power generation from PV panels using three types of models: neural network, decision tree, and random forest. The correlation coefficient R^2 measures how well the model predictions match actual data variability, while mean-square deviation (MSE) quantifies prediction errors. Given MSE's sensitivity to large residuals, we also apply root-mean-square deviation (RMSE) to mitigate this issue, aiming to determine the optimal model. We include formulas for R^2, MAE, and RMSE.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i| \tag{2}$$

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

3. Performance of Various Models in Photovoltaic Power Prediction

3.1. GBRT model

The following figure shows the fit of the gradient boost regression model for the monthly generation of PV panels (see Figure 1).

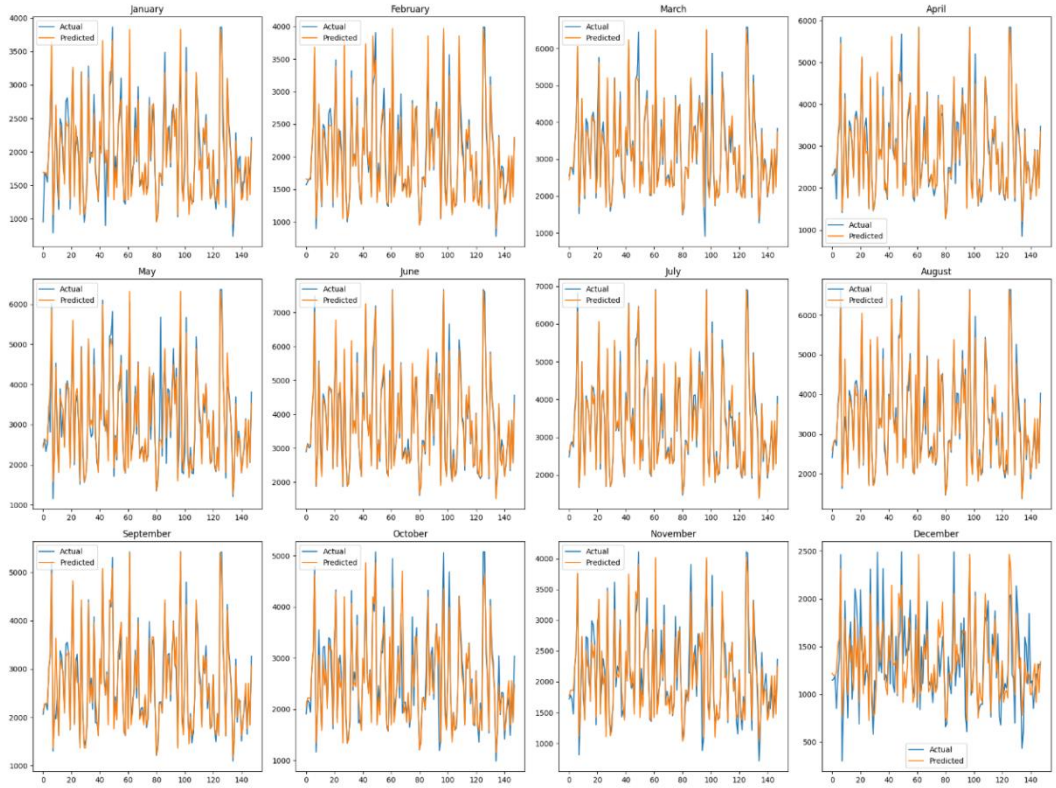

Fig. 1 Gradient boost regression model fit for monthly electricity generation from photovoltaic panels

In the performance of the model, the gradient boosting regression model showed high training and test R^2 values in most of the months, indicating that the model was able to capture the relationship between installed capacity and generation to a considerable extent, and the model showed high predictive accuracy, reflecting the GBRT model's strong data fitting ability and predictive performance. However, the decline in the tested R^2 values in October-December, especially in October when the tested R^2 value dropped to 0.78, shows the challenge of the model's ability to generalize to these months.

The main reason for the poor performance of the fitting results in winter is affected by seasonal factors. Jize County is located in the north and has a temperate monsoon climate, where the sunlight hours become shorter and the solar radiation relatively decreases in October-December. The snowy autumn and winter seasons, coupled with snow-covered household PV panels lead to a further decline in power generation efficiency, thus making the probability of outliers much higher, leading to differences in model fitting and poorer results. In future research, consider using the Interquartile Range (IQR) method to identify and handle outliers, reducing their impact on model performance.

3.2. BiGRU model

The following figure shows the fitting of the bidirectional gated recirculation unit (BiGRU) model for the monthly electricity production of PV panels (see Figure 2).

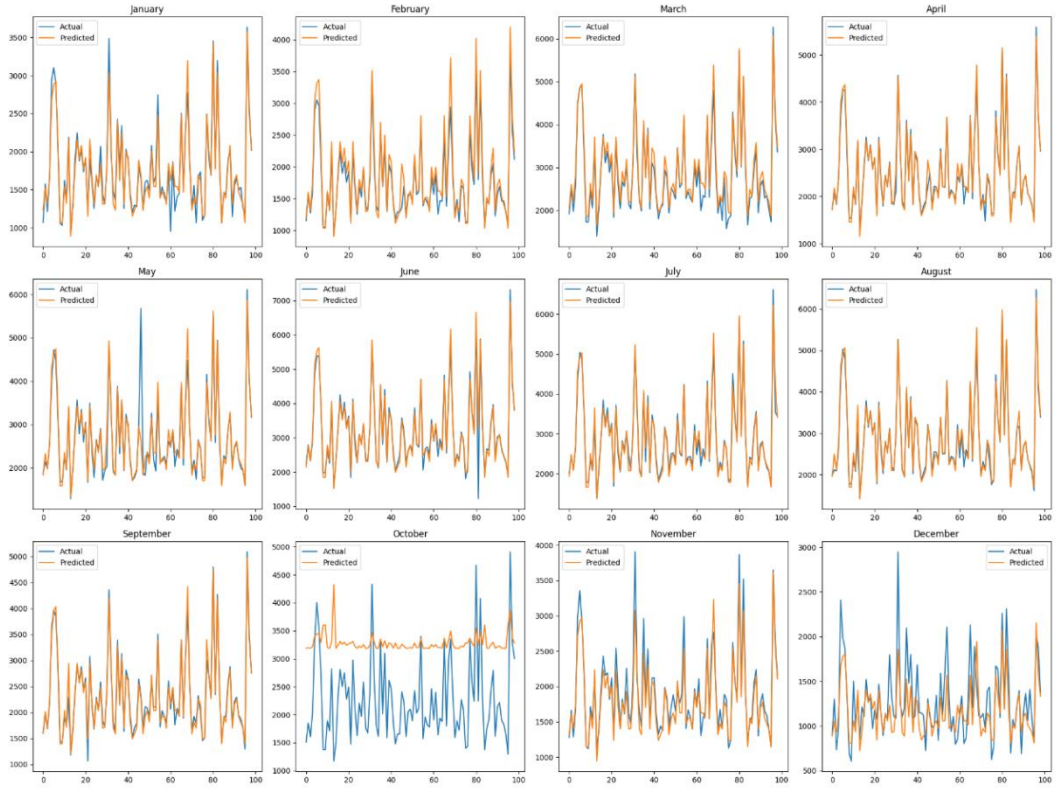

Fig. 2 BiGRU model fit for monthly electricity generation from PV panels

Based on the results obtained from the bi-directional gated recurrent unit (BiGRU) model, the model shows good predictive ability in most months, with the model exhibiting high R^2 values on the test set, indicating that the model can capture the dynamics between installed capacity and power generation more accurately. However, the model's performance drops significantly in October and December, especially in October where the R^2 value is -0.57, showing that the model predictions perform extremely poorly in this month. The very low R^2 value in October indicates that the model is extremely sensitive to outliers or fluctuations in the data. This is due to the model overfitting certain features in the training data, leading to a decrease in prediction performance when faced with volatile data.

The model demonstrated good predictive ability in most months. This reflects the model's strength in dealing with time-dependent serial data. In most months, the model has a high R^2 value for the test set, implying that the model can explain the variance of the target variable to a considerable extent, demonstrating a good fit.

3.3. RF model

The following figure shows the random forest (RF) model fit for monthly PV panel generation (see Figure 3).

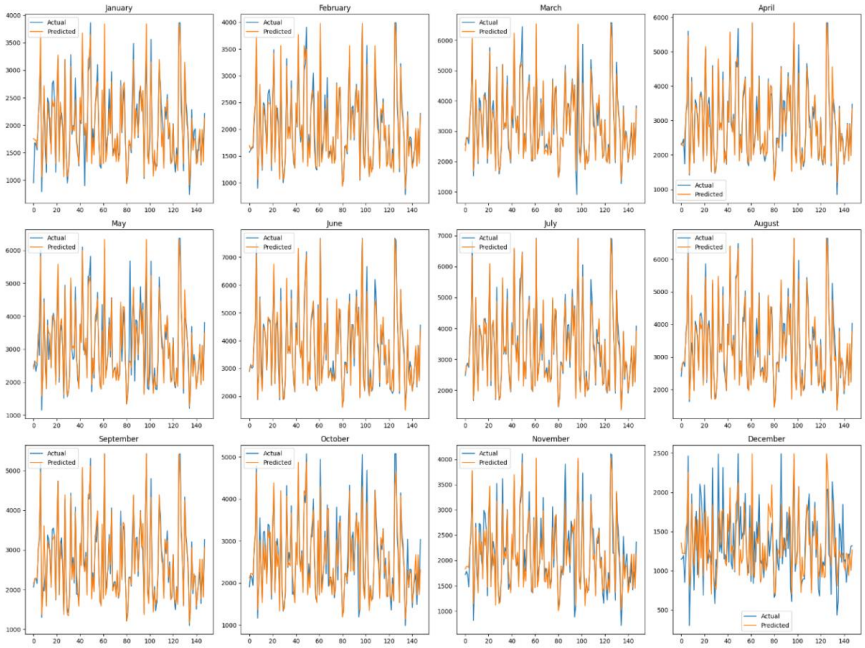

Fig. 3 Random Forest model fit for monthly power generation of PV panels

The results obtained by the model show that the R^2 values of the test set are higher in most months, especially in August-October, indicating that the Random Forest model has a better prediction performance in these months and can reflect the trend of power generation more accurately. The Random Forest model shows better predictive ability in most months, showing its effectiveness in handling this regression task. It has the advantage of being robust to outliers and is usually not prone to overfitting. However, the models do not perform well in some months, so further model tuning, feature engineering, or consideration of seasonal variations inherent in the data is required.

The three models were used to train the models using the same data to obtain their predicted correlation coefficients (R^2), root mean square errors (RMSE), and mean absolute errors (MAE), which are further measures of model fit.

3.4. Comparative analysis of R2 of the three models

Through the training of the models, we get the comparison of the correlation coefficient R^2 of the three models (see Table 1).

Table 1 Comparison of R^2 of correlation coefficient of three models

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
GBRT	0.92	0.93	0.86	0.96	0.94	0.88	0.85	0.85	0.95	0.78	0.76	0.51
BiGRU	0.94	0.95	0.95	0.98	0.87	0.95	0.98	0.98	0.97	-0.57	0.91	0.58
RF	0.90	0.95	0.89	0.85	0.89	0.92	0.90	0.90	0.93	0.92	0.86	0.45

The GBRT showed high R^2 values in most months, especially peaking in April and September, indicating that the model had high prediction accuracy in these months. However, by December, the R^2 value decreases, indicating a significant drop in the model's forecasting accuracy. The BiGRU model shows extremely high stability and forecasting accuracy throughout the year, especially in March, April, July, and August, where the R^2 value is close to or reaches 0.98. This suggests that the BiGRU model is extremely effective in capturing the dynamics of time series data. However, care needs to be taken to note that the negative R^2 value in October indicates extremely poor model prediction performance, which is due to data anomalies or model overfitting. Random Forest maintains a more consistent performance for most of the time and is particularly good in February and October. However, in December, the R^2 values plummeted, similar to the performance of the GBRT model in the same month, showing a decline in prediction accuracy at the end of the year.

3.5. Comparative analysis of RMSE of the three models

The following table shows the RMSE comparison of the three models (see Table 2).

Table 2 Comparison of RMSE of the three models

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
GBRT	199	195	477	237	309	521	543	503	238	466	377	323
BiGRU	144	129	204	120	352	247	147	138	129	933	186	276
RF	197	145	358	359	347	368	377	374	251	245	249	297

In the table, we observe that the RMSE values of all three models are generally high, which is due to the relatively large range of variation in the historical data of the training model, the average value of the installed capacity is 24603.4Wp, and the average value of the power generation from January to December ranges from 1300 to 4000KW, with a large range of variation in the values, so it leads to a generally large RMSE value. Its relative error is in the range of 4%~10%, so it can be considered that the prediction effect is better. On this basis, the BiGRU model exhibits lower RMSE values in most of the months, which indicates that the BiGRU model has high prediction accuracy in these months. The GBRT model has the highest RMSE values in July and August, which indicates that its prediction performance is relatively poor in summer. The RMSE values of the Random Forest model were more stable in several months, but generally higher than those of the BiGRU model, indicating that its overall prediction accuracy was lower than that of the BiGRU model.

3.6. Comparative analysis of MAE of the three models

The following table shows the MAE comparison of the three models (see Table 3).

Table 3 Comparison of MAE of the three models

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
GBRT	133	104	190	132	195	215	210	181	133	228	198	238
BiGRU	100	95	141	82	151	120	99	97	89	786	131	207
RF	132	87	181	169	189	186	194	194	154	165	180	241

The MAE values are large due to the large magnitude of the training data but fluctuate less relative to the magnitude of the training data. In particular, the BiGRU model shows the lowest MAE values in the majority of months, especially in October, and despite the unusually high MAE values in this month, the performance of the BiGRU model in the rest of the months suggests a better average prediction accuracy compared to the other models. The Random Forest model has advantageous MAE values in February, July to September compared to GBRT and BiGRU models but performs worst in December. The GBRT model has moderate MAE values in most of the months but its error is higher in May, June, and December.

3.7. Stability analysis

Based on the comprehensive performance analysis of the GBRT, BiGRU, and RF models, we obtained the variance of the R^2 values of each model for each month of the year, which were 0.0143 for the GBRT model, 0.1795 for the BiGRU model, and 0.0162 for the Random Forest model.

The GBRT model and the Random Forest model have a lower variance of R^2 values of 0.0143 and 0.0162, respectively, which suggests that these two models maintain a more stable performance within each month of the year. In contrast, the BiGRU model has an R^2 value variance of 0.1795, which is significantly higher than the other two models, indicating that the performance of the BiGRU model fluctuates more from month to month and is less stable. The performance of the BiGRU model is not as stable as that of the GBRT and the Random Forest model when dealing with datasets with large variations or outliers.

In summary, the BiGRU model is relatively the best forecasting model in terms of forecasting accuracy but is unstable compared to the other two models. The RF model demonstrates better robustness with relatively stable performance. The GBRT demonstrates high R^2 values in some months and the model is relatively stable.

4. Economic and Environmental Benefits Analysis of Photovoltaic Systems

After solving the problem of predicting power generation, we can get the unknown power generation according to the prediction model, and then we can further analyze and evaluate the economic and environmental benefits based on the power generation.

Currently, the economic benefits of PV systems are mainly determined by the generation cost and operation mode.

4.1. Economic analysis

When evaluating the economics, the time value of money must be taken into account, i.e. by discounting the costs and benefits. In this paper, a discount rate of 5% is used to convert future costs into present value. The cost of a PV system includes the initial investment, equipment replacement, and O&M costs. The initial investment covers PV modules, inverters, racking, convergence boxes, and insurance, as well as construction costs. Although the warranty period of a PV system is usually 25 years, its service life can exceed this period, and this paper uses 25 years as the evaluation standard. Since the service life of an inverter is usually no more than 10 years, the equipment replacement cost mainly refers to the cost of replacing the inverter. The O&M cost is then calculated based on the unit O&M cost and the potential capacity of the system. By using Equation (4), the specific generation cost of the PV system can be calculated.

$$LCC = \sum_{i=1}^{2.5} \frac{((C_{eq,i} + C_{m,i}) \times capacity)}{1.05^i} \tag{4}$$

In Eq. (4): LCC represents the overall generation cost of the PV system (in ¥), Ceq, i represents the initial investment and replacement cost of each unit of equipment of the PV system in the year i (in ¥), while Cm, i indicates the operation and maintenance cost of each unit of the PV system in the year i (in ¥) [10]. The levelised cost of electricity generation metric, i.e. the life cycle average cost of electricity generation (LCOE), is derived by averaging the total cost of the project over its entire life cycle with the total electricity generation, which is the core parameter for measuring the economic efficiency of the PV system, and it can be calculated by using Eq. (5):

$$LCOE = \frac{LCC}{E_{p25}} = \frac{\sum_{i=1}^{25} \frac{\left\{\left(C_{eq,i} + C_{m,i}\right) \times Capacity\right\}}{1.05^j}}{\sum_{i=1}^{25} E_{Pi}} \tag{5}$$

The operation modes of PV systems can be categorized into three types according to the source of economic benefits: fully grid-connected mode, self-generated electricity for self-consumption and residual electricity for on-grid use, and fully self-consumption mode [11]. In the fully grid-connected mode, all the power generated by the PV system is supplied to the grid for use, and the economic benefits are derived from the grid-connected revenue. The self-consumption and grid-connected mode allows a portion of the power generated by the PV system to be used for self-consumption and the remaining portion to be connected to the grid, and the economic benefits include both the cost savings of substituting traditional energy sources and the grid-connected revenues. In the fully grid-connected mode, the power generated by the PV system is fully connected to the electricity supply bureau. In this paper, we study the revenue of domestic PV panels in the fully grid-connected mode, and the revenue of the PV system in this mode of operation can be calculated by Equation (6):

$$Profit = \sum_{i=1}^{25} \frac{E_{p,i} \times P_k}{1.05^i} \tag{6}$$

In Eq. (6), Profit represents the total return of the PV system (in ¥), Ep, i denotes the amount of electricity produced by the PV system in the ith year (in kWh), and Pk is the unit price of electricity in a specific scenario k (in ¥). The economic performance of a PV system is evaluated through several core metrics, including net present value (NPV), net income (NI), internal rate of return (IRR), and return on investment (ROI). Net Present Value (NPV) is the difference between the present value of all future cash inflows and the present value of all future cash outflows after considering the discount rate. The calculation of this value can be done according to equation (7):

$$NPV_i = \sum_{t=1}^{i} \frac{R_i}{1.05^i} \tag{7}$$

In Eq. (7), NPV_i represents the net present value of the PV system in the year i (in ¥), while R_i represents the net cash inflow of the PV system in that year (in ¥). Net income (NI) is defined as the residual income after deducting the cost of PV power generation, which does not involve the consideration of the time value of money [18]. The exact calculation of net income can follow equation (8):

$$NI = Profit_k - LCC_k \tag{8}$$

In Eq. (8), $Profit_k$ represents the economic return of the PV system under a specific scenario k (in ¥), while LCC_k represents the total cost of PV power generation under the same scenario (in ¥).

4.2. Assessment of economic benefits

Taking Jize County, Handan City, Hebei Province as an example, the PV panels are sold at a grid price of RMB 0.3641 per kWh, and the annual electricity generation from 1KW PV panels is 1,374 kWh. Considering the initial investment cost (see Table 4), it is calculated that the householders can receive net benefits as early as the 8th year until the end of the life of the PV panels. This estimate does not include government subsidies; if subsidies are included, the payback period can be shortened accordingly. For commercial and industrial users, the payback period is estimated to be 4 to 6 years due to the impact of peak tariffs.

Table 4 Initial PV System Investment Costs

Component	Price (¥/W)
Fuse	0.035
Convergence box	0.1
Line	0.2
Inverter	0.3
PV mounting	0.3
Engineering	0.6
PV module	2
Initial investment cost	3.535

4.3. Analysis of Environmental Benefits

The integration of photovoltaic systems not only effectively reduces the temperature of building surfaces, but also helps to reduce cooling demand through power generation, further alleviating the urban heat island phenomenon. In the context of China's pursuit of a "dual-carbon" goal, it is crucial to conduct an in-depth assessment of the environmental benefits of PV systems. Studies have shown that for every 1 kWh of electricity saved, 0.4 kg of standard coal consumption is reduced, with corresponding reductions of 0.272 kg of carbon dust, 0.997 kg of carbon dioxide (CO_2), 0.03 kg of sulfur dioxide (SO_2), and 0.015 kg of nitrogen oxides (NO_X) [12].

5. Conclusions

This paper focuses on the relationship between installed capacity and electricity generation and compares and analyses the fitting effectiveness of three models. The results show that the BiGRU model performs best in fitting the relationship between installed capacity and electricity generation, although the October data shows some sensitivity, which suggests that more features and parameters need to be introduced for training to optimize the model. The Random Forest model can effectively handle high-dimensional features is robust to outliers, and is generally less prone to overfitting. To improve its prediction accuracy, it is recommended to add feature engineering and consider seasonal variations. The gradient-boosting regression model shows strong prediction ability in most months, however, it still faces challenges in generalization ability. This paper also analyses the economic and environmental benefits of generating electricity from domestic photovoltaic panels. It is shown that the initial cost of installing domestic PV panels can be recovered through the economic benefits generated within eight years, even without government subsidies, and the net benefits can be realized within four to six years with government subsidies.

This paper provides a research idea and framework for PV panel power generation forecasting and benefit analysis, and analyses the fitted forecasts of historical power generation data from three models, BiGRU, RF, and GBRT, and finds that the BiGRU model has a 7.6% higher forecasting accuracy than the other two models, which further demonstrates the feasibility of power generation forecasting using the BiGRU model.

References

[1] Yang Xiaozhan, editor-in-chief; Feng Wenlin, and Ran Xiuzhi, associate editors. Introduction to New Energy and Sustainable Development [M]. Chongqing: Chongqing University Press,2019

[2] Wang Huan, Ma Bing, Jia Lingxiao, et al. Role of key minerals in the clean energy transition under the carbon neutrality target, supply and demand analyses and their suggestions[J]. China Geology,2021,48(06):1720-1733.

[3] LAI Bo,CHEN Minyu,ZHONG Haiwang,et al. Overview of long-term planning for new power systems with a high proportion of renewable energy[J]. Chinese Journal of Electrical Engineering,2023,43(02):555-581.

[4] MA Jingwen, LI Shuqing, XIA Mengyao. A research review on the application of machine learning in GDP forecasting[J]. Science and Technology Intelligence Research,2022,4(03):73-94.

[5] ZHANG Chi,GUO ¥,LI Ming. A review of the development and application of artificial neural network models[J]. Computer Engineering and Applications,2021,57(11):57-69.

[6] Park S, Jung S, Lee J, et al. A Short-Term Forecasting of Wind Power Outputs Based on Gradient Boosting Regression Tree Algorithms[J]. Energies, 2023, 16(3): 1132.

[7] Fu Lianlian, Wu Jian. Pig price prediction based on gradient boosting regression model[J]. Computer Simulation,2020,37(01):347-350.

[8] ZHANG Changfan, LIU Jiafeng, HE Jing, et al. Bearing fault diagnosis based on improved convolutional bidirectional gated recurrent network[J]. Journal of Electronic Measurement and Instrumentation,2021,35(11):61-67.

[9] REN Liqiang, JIA Shuyi, WANG Haipeng, et al. A review of deep learning based time series classification research[J/OL]. Journal of Electronics and Information,1-23[2024-04-28].

[10] Zhu Xingyu. Technical and economic analysis of rooftop photovoltaic power generation system for universities based on dual-carbon target [D]. Qilu University of Technology,2023.

[11] Gao Ruiming. Research on the operation mode of the new energy enterprise of CGN under the background of dual-carbon target [D]. Xi'an University of Technology, 2023.

[12] Yang Yuxin. Analysis of environmental benefits of campus rooftop photovoltaic power generation under the background of "dual carbon"[J]. Automation Application,2023,64(01):41-42+70.