

Prediction of Building Energy Consumption Based on LSTM-SVR-Random Forest Hybrid Model

Zijun Du^{1,a,*}

¹*School of Information and Computing Science, Beijing University of Civil Engineering and Architecture, Beijing, China*

a. 202208010203@stu.bucea.edu.cn

**corresponding author*

Abstract: Accurate building energy prediction can help effectively allocate and utilize energy resources, reducing waste. The dataset used in this study contains data from 539 buildings and includes 12 variables. This study proposes a hybrid model of Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) based on random forest optimization. The model combines the accuracy of SVR with the time series modeling ability of LSTM, which significantly improves the prediction accuracy of building energy consumption. Through an in-depth analysis of historical energy consumption data, this paper identifies key factors influencing energy use and employs the random forest algorithm for feature selection and model optimization. The experimental results show that the accuracy of the LSTM single model is approximately 54%, while the accuracy of the SVR single model is 15%. The hybrid model achieves an accuracy rate of more than 90% on both the training set and the test set, which shows that the SVR-LSTM hybrid model using Random Forest can effectively improve the prediction performance. This research provides important theoretical foundations and practical guidance for energy-efficient building design and smart energy management, while also offering data support for future policy-making, demonstrating broad application prospects.

Keywords: Building Energy Consumption, SVR-LSTM Hybrid Model, Random Forest.

1. Introduction

Accurately predicting building energy consumption is crucial for implementing intelligent building management systems, which in turn are pivotal for achieving energy conservation and reducing emissions. Effective energy management can result in significant cost savings and environmental benefits. Traditional prediction methods often struggle with the inherent complexity and variability of energy consumption data, which poses challenges in terms of both data handling and model accuracy. To address these challenges, advanced techniques such as Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks have been increasingly utilized due to their effectiveness in managing complex and dynamic datasets. LSTM and SVR models are now used in many areas. The application fields of SVR include protein structure prediction, gene expression prediction, etc. The application areas of LSTM include stock price prediction and risk management in the financial field, disease prediction and treatment recommendation in the medical field, etc.

SVM has become a popular tool for energy consumption prediction because of its ability to handle intricate patterns in high-dimensional spaces. Support Vector Machines distinguish different classes by constructing a hyperplane in high-dimensional space or fitting data points to a regression line with minimal error. This method is particularly useful for handling nonlinear relationships and complex data patterns. For instance, Zhang et al. proposed an SVM-LSTM-based method for predicting reservoir turbidity. This study demonstrated the advantages of SVM in dealing with complex and high-dimensional data patterns, showcasing its effectiveness in providing accurate predictions [1]. Similarly, Zhang Qinqin et al. combined LSTM with SVM to study river flow predictions. This approach further validated SVM's capability in managing time series data, underscoring its versatility and effectiveness in various prediction contexts [2].

On the other hand, LSTM networks, a type of Recurrent Neural Network (RNN), are specifically designed to capture and model temporal dependencies in time series data. Unlike traditional RNNs, LSTMs are equipped with memory cells that can retain information over long periods, making them particularly suitable for tasks where historical data plays a crucial role in predicting future values. Gers et al. first highlighted the ability of LSTM to capture long-term dependencies, which is essential for handling sequential data [3,4]. Recent advancements have further showcased LSTM's potential in building energy consumption prediction. For instance, Li et al. proposed a spatial-temporal hierarchical correlation methods of dissemination based on LSTM, which improved the model's interpretability and prediction accuracy by incorporating spatial and temporal correlations [5]. Furthermore, Li et al. introduced Bayesian optimization and spatiotemporal attention mechanisms to refine LSTM's performance in building energy prediction, demonstrating continued advancements in LSTM-based methodologies [6,7].

To further enhance predictive performance, researchers have also explored hybrid and model fusion approaches. Combining multiple predictive models can leverage the strengths of each model to improve overall prediction accuracy. For example, Zhang Lei et al. developed an LSTM-SVM network model for predicting MOOC dropout rates. This model fusion approach combined convolutional neural networks (CNN) for feature extraction, LSTM for temporal modeling, and SVM for classification or regression, showcasing the potential benefits of integrating different models [8]. Similarly, Arkhangelski et al. combined Akima interpolation with deep learning LSTM methods to enhance the accuracy of building energy efficiency predictions. Integrating interpolation techniques with LSTM improved the model's precision in capturing energy consumption patterns [9,10].

Building on these advancements, this paper proposes a novel hybrid model combining SVM, LSTM, and stacking methods—designated as SVM-LSTM-Stacking. The purpose of this paper is to a hybrid model to improve the prediction effect of energy in smart buildings. This model aims to address the complex challenges associated with building energy consumption prediction by leveraging the strengths of SVM in managing nonlinear patterns, the temporal modeling capabilities of LSTM, and the optimization potential of stacking techniques. The SVM component handles complex, high-dimensional data patterns; LSTM captures the temporal dynamics of energy consumption; and stacking techniques aggregate predictions from different models to enhance overall accuracy. This integrated approach promises to provide a more robust and effective tool for building energy management, offering improved prediction accuracy and better insights into energy consumption patterns. In the following paper, the data are collated and normalized for convenience, and then the results predicted by the mixed model are compared with the SVR and LSTM models. Finally, conclusions are drawn.

2. Methods

2.1. Data Source

This paper uses the listing of building energy performance data from 2017 to 2020. The original data is saved in CSV format.

The data used in this study contains 539 instances and 12 variables, with no significant outliers or missing values. Table 1 shows the names, data types, and descriptions of the 12 variables used in this study. The last variable is predicted based on the data from the first 11 variables.

Table 1: Different variables

Full Name	Data type	Explanation
Building size	INT	Building size
year_topcsc	INT	Years obtained topics
Gross_floor_area	INT	The total area of the building
Percent_air_conditioned	INT	Air conditioner floor area percentage
Ave_m_b_occupancyrate	INT	Average monthly building occupancy rate
Age_of_chiller	INT	Age of chiller
Centralized_air_efficiency	INT	Centralizedair conditioning plant efficiency Usage of
Usage_of_led	INT	led
2017year	INT	2017 consumption
2018year	INT	2018 consumption
2019year	INT	2019 consumption
2020year	INT	2020 consumption

2.2. Method Introduction

RNNs are difficult to train, which makes handling long-range dependencies in practical applications even more challenging. LSTM is a type of recurrent neural network that successfully addresses the shortcomings of traditional recurrent neural networks.

The structure of the LSTM model primarily includes the following five components: Input Gate, Forget Gate, Output Gate, Cell State, and Candidate Values. The overall workflow is as follows: first, the input information passes through the Forget Gate, which determines how much of the old information to retain. Then, the Input Gate combines the new information with the Candidate Values to update the Cell State. Finally, the output gate determines the final output based on the current cell state. The LSTM network structure is shown in Figure 1 [3].

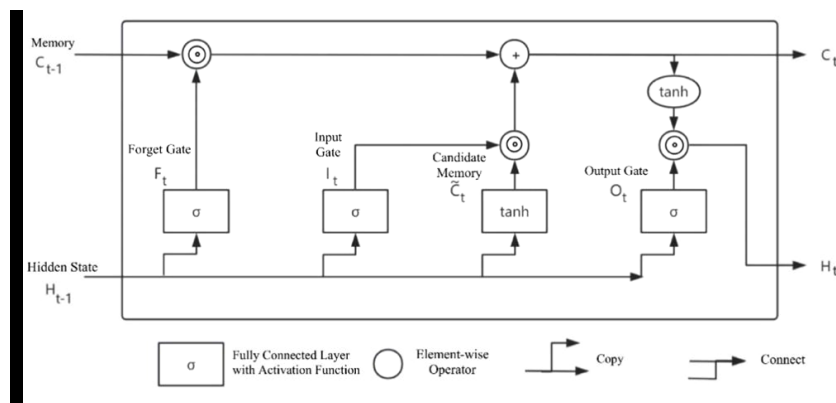


Figure 1: The LSTM network structure [3]

The core computational formulas used in the LSTM model in this paper are:

$$f(x) = \sum_{i=1}^n (\alpha_i + \alpha_i^*) K(x_i, x) + b \quad (1)$$

Where x represents the sample input data, and x_i is the input feature vector of i sample in the training set.

α_i, α_i^* is a positive and negative Lagrange multiplier.

$K(x, x_i)$ is a Kernel Function used to calculate the similarity between input x and support vector x_i . b is a Bias Term.

The LSTM network structure, as shown in Figure 1 by Visio.

The following Figure 2 is a schematic diagram of the SVR model created using Word.

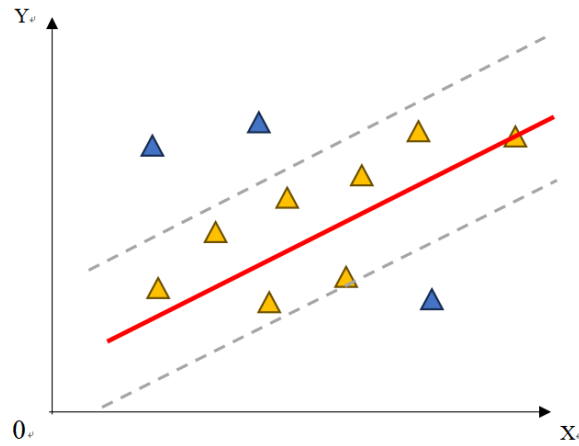


Figure 2: SVR diagram

Random Forest belongs to the Bagging method in ensemble learning. Random Forest is composed of many decision trees, which are not related to each other. When this study performs classification tasks, each decision tree in the forest will make judgments and classifications on new input samples. Each decision tree will produce its classification results. The classification that appears most frequently among the decision trees will be taken as the result of the Random Forest. The following Figure 3 is a flowchart of the Random Forest process created using ProcessOn.

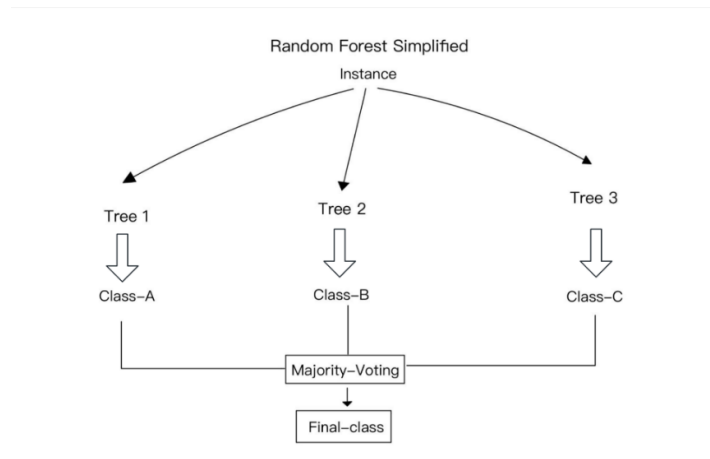


Figure 3: The Random Forest network structure

3. Results and Discussion

3.1. Data Descriptive Analysis

Figure 4 reflects the proportion of various building sizes in the dataset. In the chart, 0 represents small-scale buildings, 0.5 represents medium-scale buildings, and 1 indicates large-scale buildings. As you can see from the figure, the proportion of large buildings is 56%. Illustrates the dominance of large buildings on the dataset. The proportion of small buildings is only 6%.

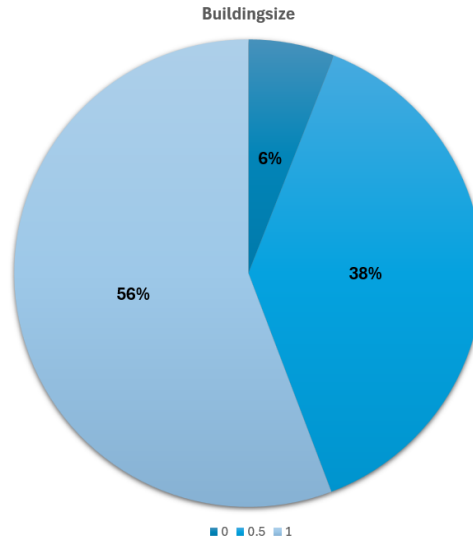


Figure 4: Building size

Figure 5 is the histogram of the distribution of building construction years. In the two periods from 1996 to 2003 and from 2010 to 2017, the highest number of buildings were built, with 108 and 117, respectively. The number of buildings before 1969 is less than 20

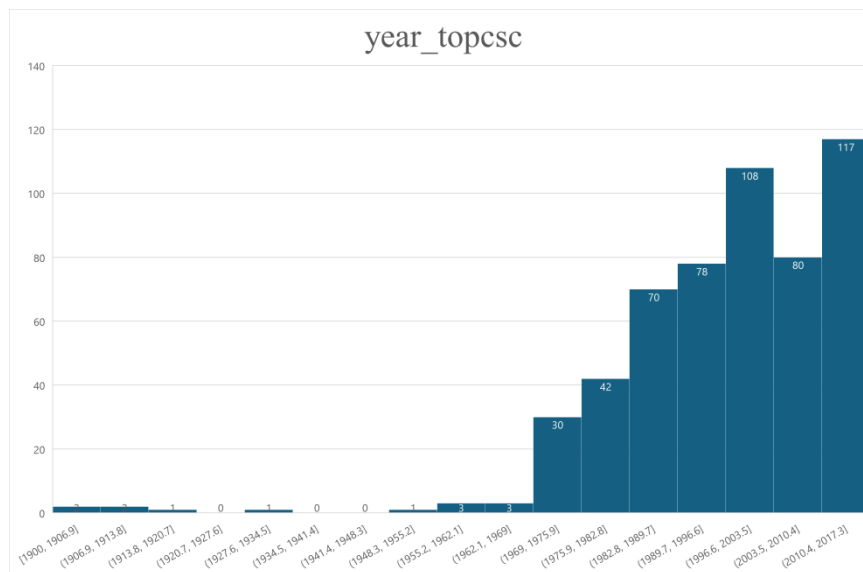


Figure 5: Years building construction

Figure 6 is the histogram of the distribution of the total building area. It can be seen that the largest portion of buildings falls within the area range of 357 to 25,357 square meters.

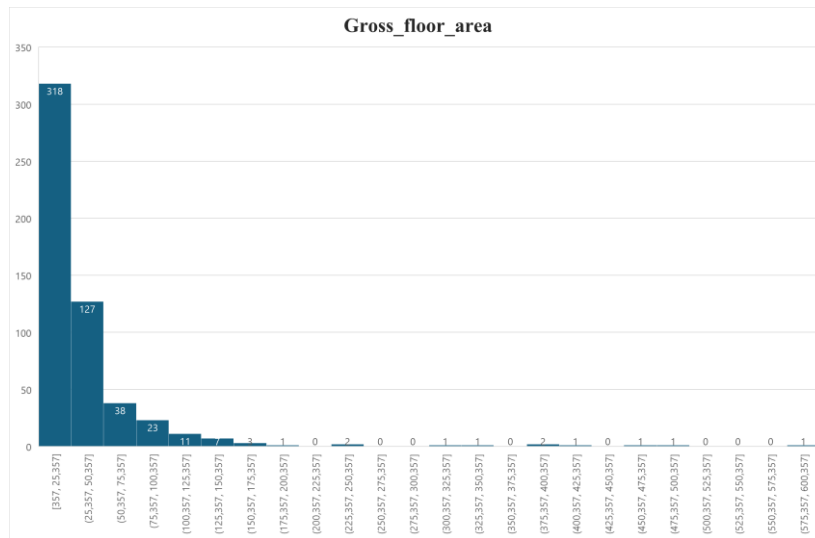


Figure 6: Total building area

Figure 7 is the histogram of the distribution of air conditioner floor area percentages, with the data converted from percentages to decimals. It can be seen that the highest distribution is within the range of 0.768 to 0.96, indicating a high proportion of air-conditioned areas within the building.

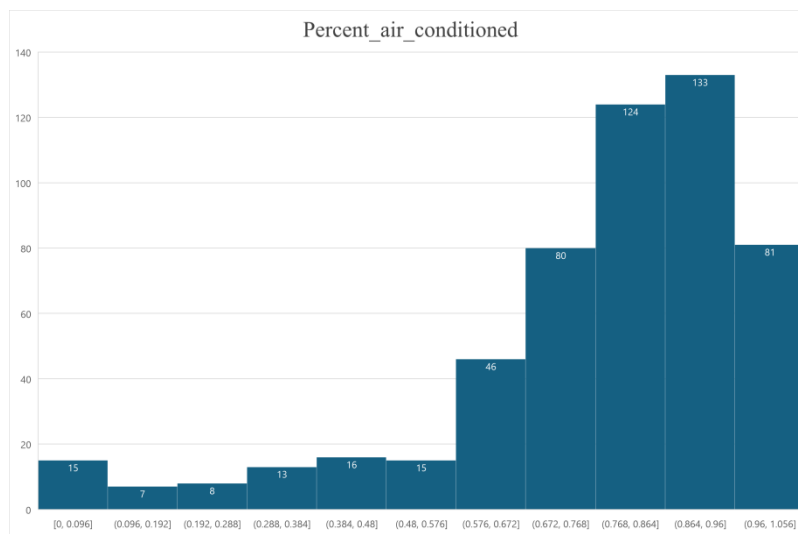


Figure 7: Percent air-conditioned

Figure 8 shows the average monthly building occupancy rate, with the data presented as percentages. It can be observed that the highest occupancy rates fall between 0.902 and 0.973.

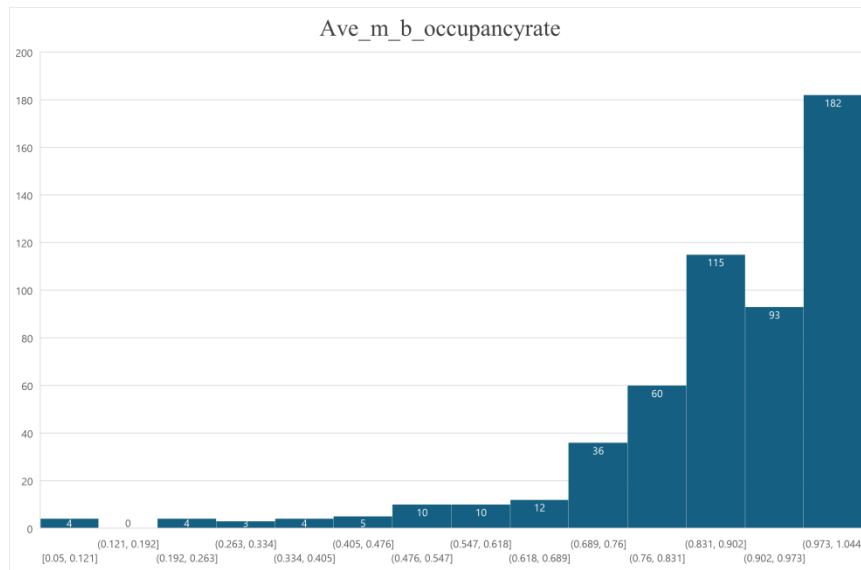


Figure 8: Average monthly building occupancy rate

Figure 9 is the histogram of the distribution of chiller ages. The unit is in years. It can be observed that the highest distribution of chiller ages is between 0 and 8 years, indicating that the age of the chillers is not low, which may affect energy consumption.

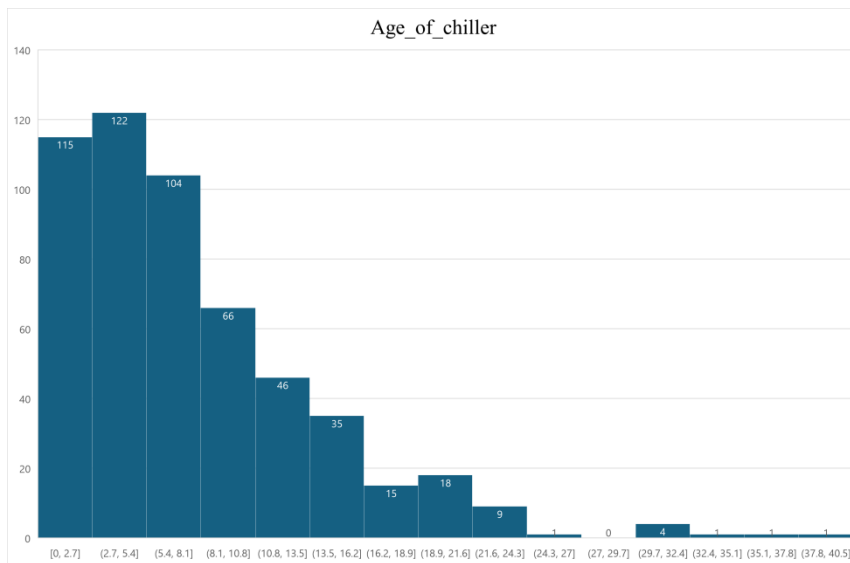


Figure 9: Age of Chiller

Figure 10 is a histogram of LED usage rates. As shown in the figure, a significant proportion of LED usage is concentrated in the range of 0 to 0.16.

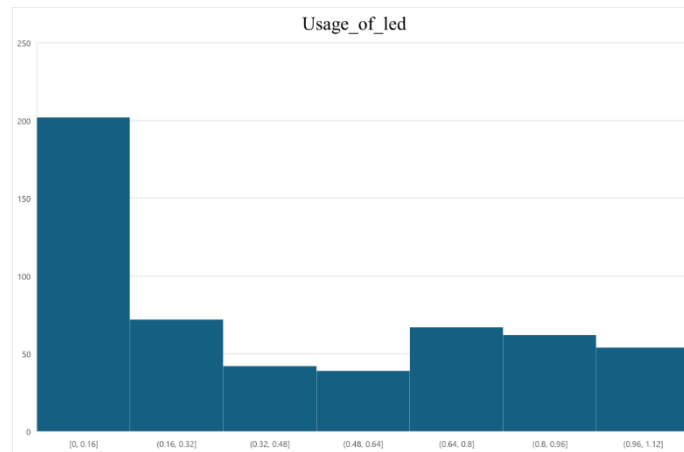


Figure 10: Usage of led

Figure 11 shows the results of the correlation analysis performed in Python using all variables from the dataset. It can be seen that some variables have correlations as high as 0.98, while others have correlations of -0.12. This indicates that some data may not be suitable for training the model, so these variables will be excluded in subsequent calculations.



Figure 11: Correlation Analysis Results

3.2. Single Model Results

The LSTM model's prediction results show a relatively low mean squared error (0.0021), indicating that the model has a small prediction error. As can be seen from Table 2 the R-squared value is 0.542, This indicates that the model can explain approximately 54.2% of the variability, which, while not very high, indicates some explanatory power. As can be seen from Table 3 the differences between actual and predicted values are small, especially for several data points in the table. It is recommended that the model performs well on these data points. However, there is still room for improvement to enhance prediction accuracy and explanatory capability.

Table 2: LSTM Model Evaluation Results

	Result
Mean Squared Error	0.002098
R-squared	0.5423

Table 3: LSTM Model Some Data Predicted

Actual	Predicted
0.199315	0.218120
0.168002	0.162747
0.107246	0.126053
0.048958	0.051842
0.104734	0.088292

In the SVR model, the mean squared error is 0.0039, which is not considered high, but it shows a noticeable deviation compared to previous results. As can be seen from Table 4 The R-squared value is 0.151, indicating that the model can only explain about 15.1% of the variability, demonstrating weak explanatory power and an inability to effectively capture the data trends. As can be seen from Table 5 The differences between actual and predicted values are significant, particularly for certain data points (such as the second and third), indicating that the model performs poorly for these points.

According to Figure 12, it can be seen that most of the data points are concentrated in the region with low values, such as the Y value of about 0.1, indicating that the prediction of the model in the low-value area is relatively concentrated.

Table 4: SVR Model Evaluation Results

	Result
Mean Squared Error	0.003891
R-squared	0.151267

Table 5: LSTM Model Some Data Predicted

Actual	Predicted
0.199315	0.217721
0.168002	0.148891
0.107246	0.178378
0.048958	0.102177
0.104734	0.100768

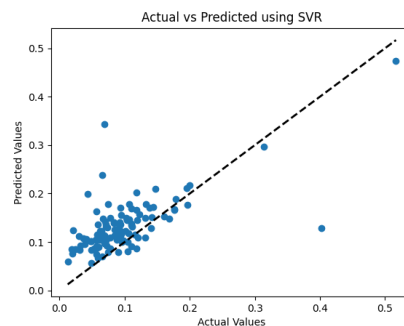


Figure 12: SVR Results

3.3. Random Forest Stacking Results

The prediction results of the Random Forest Stacking model are relatively good. The MSE is 0.00047, indicating that the model's prediction error is very small, demonstrating good accuracy. As can be seen from Table 6 the R-squared value is 0.898, meaning this model can explain about 89.8% of the variability, showing strong explanatory power. As can be seen from Table 7 the differences between actual and predicted values are also small, particularly for the first few data points. This indicates that the model performs well in these areas. Overall, this model exhibits excellent performance in predictions and is worth further application and validation.

Table 6: Stacking Model Evaluation Results

	Result
Mean Squared Error	0.000469
R-squared	0.897615

Table 7: LSTM Model Some Data Predicted

Actual	Predicted
0.199315	0.192892
0.168002	0.162449
0.107246	0.113041
0.048958	0.052509
0.104734	0.100794

4. Conclusion

This paper combines the accuracy of SVR with the time series modeling capabilities of LSTM to study. This research can significantly improve the prediction accuracy of building energy consumption. This experiment demonstrates that the prediction method using Random Forest stacking with LSTM and SVR yields effective results for predicting building energy consumption. The accuracy of both the test set and training set is around 90%. By analyzing the relevant factors affecting energy consumption and collecting historical data, this hybrid model can provide more accurate predictions for future energy consumption. With the development of smart buildings and the Internet of Things, this model can be integrated with smart control systems to enable real-time monitoring and adjustments, optimizing energy management. By accurately predicting energy consumption, it helps improve energy efficiency in building design and operation, promoting the development of sustainable buildings. Future research can provide data support for policy-making, aiding in the formulation of more effective energy management and conservation policies. The application of this model is not limited to the building sector but can also be extended to industrial, transportation, and other fields, achieving broader energy management.

References

- [1] Zhang, H. (2020). A reservoir-type turbidity prediction method based on the SVM-LSTM model. *Ecological Modelling*, 432.
- [2] Zhang, Q., et al. (2023). River flow prediction based on the LSTM-SVM model. *Journal of Tianjin Normal University (Natural Science Edition)*, 43(6), 45-52.
- [3] Gers, F. A., Eck, D., & Schmidhuber, J. (2001). Applying LSTM to time series predictable through time-window approaches. In *International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, 22-25.
- [4] Boasso, E., Manganelli, M., & Martirano, L. (2021). Impact of EV charging on the electric load of smart buildings and energy communities. In *2021 IEEE Industry Applications Society Annual Meeting (IAS)* 234.

- [5] Li, G., et al. (2022). *A spatial-temporal layer-wise relevance propagation method for improving interpretability and prediction accuracy of LSTM building energy prediction*. *Energy and Buildings*, 452.
- [6] Li, G., et al. (2024). *BO-STA-LSTM: Building energy prediction based on a Bayesian optimized spatial-temporal attention enhanced LSTM method*. *Developments in the Built Environment*, 18, 114-117.
- [7] Arkhangelski, J., et al. (2024). *Combined Akima interpolation and deep learning LSTM method-based building energy efficiency prediction*. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*.
- [8] Zhang, L., et al. (2019). *Construction of a CNN-LSTM-SVM network model and MOOC dropout prediction method*. Patent CN110059716A.
- [9] Arkhangelski, J., et al. (2024). *Combined Akima interpolation and deep learning LSTM method-based building energy efficiency prediction*. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*.
- [10] Hu, T. (2020). *Research on building energy consumption prediction method based on EDA-LSTM model*. MS thesis. Shenzhen University.