energies

MDPI

Article

Research on Online Temperature Prediction Method for Office Building Interiors Based on Data Mining

Jiale Tang [1,2], Kuixing Liu [2], Weijie You [2], Xinyu Zhang [1,3,*] and Tuomi Zhang [4]

1 State Key Laboratory of Building Safety and Built Environment, Beijing 100013, China
2 School of Architecture, Tianjin University, Tianjin 300072, China
3 China Academy of Building Research, Beijing 100013, China
4 Tianjin International Engineering Institute, Tianjin University, Tianjin 300072, China
* Correspondence: zxyhit@163.com

Abstract: Indoor environmental parameters are closely related to the energy consumption and indoor thermal comfort of office buildings. Predicting these parameters, especially indoor temperature, can contribute to the management of energy consumption and thermal comfort levels in office buildings. An accurate indoor temperature prediction model is the basis for implementing this process. To this end, this paper first discusses the input and output parameters of the model, and then it compares the prediction effects of mainstream prediction model algorithms based on data mining under the same data conditions. The superiority of the XGBoost integrated learning algorithm is verified, and a further XGBoost-based indoor temperature online prediction method is designed. The effectiveness of the method is validated using actual data from a commercial office building in Haidian District, Beijing. Finally, optimization methods for the prediction method are discussed with regard to the scheduler mechanism proposed in this paper. Overall, this work can assist building operators in optimizing HVAC equipment running strategies, thus improving the indoor thermal comfort and energy efficiency of the building.

Keywords: temperature prediction; meteorological parameters; XGBoost; online operation; scheduler; error

check for updates

Citation: Tang, J.; Liu, K.; You, W.; Zhang, X.; Zhang, T. Research on Online Temperature Prediction Method for Office Building Interiors Based on Data Mining. *Energies* **2023**, *16*, 5570. https://doi.org/10.3390/en16145570

Academic Editor: Vincenzo Costanzo

Received: 8 June 2023
Revised: 12 July 2023
Accepted: 20 July 2023
Published: 24 July 2023

1. Introduction

With the improvement of people's living standards, heating, ventilation, and air conditioning (HVAC) systems have been widely applied to meet people's requirements for the comfort of indoor building environments. However, building energy consumption has also been continuously increasing. In 2021, the operation of buildings accounted for 30% of global final energy consumption and 27% of total energy sector emissions [1]. HVAC accounts for 38% of buildings consumption [2]. The HVAC system directly affects indoor thermal comfort and building energy consumption. For commercial office buildings, indoor thermal comfort has an impact on the productivity of indoor personnel [3], and thus, ensuring good indoor thermal comfort is one of the key factors in maintaining its commercial operation. At the same time, the level of building energy consumption is directly related to the evaluation of building owners. Therefore, how to effectively reduce the energy consumption of HVAC equipment while ensuring indoor comfort has become an important research direction, which has promoted the emergence and development of Building energy management systems (BEMS). This system controls and monitors energy use and indoor environmental conditions through the use of appropriate operating strategies to maintain a satisfactory level of indoor building environment and achieve building energy efficiency [4].

A superior operating strategy is the key to the efficient operation of BEMS. Research shows that implementing efficient energy management strategies can lead to 5~15% energy savings in existing buildings [5]. Learning optimal control strategies from optimized

operating datasets is a feasible method for improving the HVAC system efficiency [6]. In the monitoring and management of building energy equipment, predictive results of indoor environmental parameters can serve as an important basis for formulating operating strategies. Among these parameters, air temperature is the most relevant in terms of indoor thermal comfort and it significantly influences building energy consumption [7]. By combining real-time and accurate indoor temperature prediction models, BEMS can effectively control indoor temperature setpoints and implement more efficient energy-saving HVAC control strategies, leading to optimized building energy consumption and improved indoor thermal conditions [8]. Such an indoor temperature prediction model can be of immense value to the building facility manager, and it should be scalable and parameter adjustable for different real-world situations. The prediction algorithm can be integrated with smart sensors and predictive control systems which can be trained for future scenarios [9]. With the widespread application of building automatic systems (BAS), a large amount of building operation data such as indoor environmental parameters and power consumption data can be accumulated, which provides fundamental data support for the development of online prediction models for indoor temperatures.

In recent years, the methods for predicting indoor temperature based on data mining technology have been discussed and experimented with by more and more researchers. However, there are many differences in the selection of model input parameters, design of prediction time length, and selection of model and algorithms. Table 1 summarizes the different methods used by related scholars in predicting room temperature.

Table 1. Different methods for predicting indoor temperature.

Ref	Time Granularity	Predicted Length	Arithmetic	RMSE
[10]	5 min	30 min~3 h	ARX, NNARX	0.126
[11]	5 min	6 days	LSTM	0.5267
[12]	15 min	15/30 min	ARX, ARMAX, BJ, ANN	0.367
[13]	1 min	6 days	ANFIS	1.048
[14]	60 min	non-sequential	ANN	3
[15]	5 min	28 days	NARX	0.288

First, the selection of input variables is an essential part of model development [16]. Using too few parameters will lead to poorly accurate prediction results, and using excess or redundant input parameters creates unnecessary complexity and decreases the computation speed. Mba et al. [10] used only indoor and outdoor temperatures and relative humidity to predict future changes in indoor air temperature and relative humidity. Xu et al. [11] used only historical indoor temperatures as an input to the model to predict future temperature trends. Zakia Afroza et al. [17] used 12 characteristic parameters of the air conditioning equipment system to predict the changes in indoor temperature. The aforementioned situations represent the main types of input parameters commonly used in this kind of research model. However, in practical engineering applications, it is preferable to select parameters that are related to indoor temperature and easy to obtain as the research objects, in order to avoid the increased costs associated with installing a large number of sensors. Secondly, in the face of multiple selectable parameters, several studies have started to show interest in and experiment with the selection of effective parameters [18–20]. Yadav et al. [18] and Talebi et al. [19] used sensitivity analysis techniques to exclude input parameters with low correlation. Guo et al. [20] used the Pearson simple correlation coefficient method to calculate the correlation coefficient between two variables and selected the variables with high correlation to form the feature set of the model. The aforementioned practices are highly recommended for adoption.

The time granularity of indoor temperature prediction is concentrated around 5 min [11,15,21], but the prediction lengths in this field of study are widely distributed, with the shortest prediction lengths being about 15 min [12], longer prediction lengths being

about a week, and the longest prediction lengths being up to 28 days [15]. Xie et al. [13] made use of the adaptive neuro-fuzzy inference system (ANFIS) and the dynamic model constructed based on the energy balance equation (EBE) to predict indoor air temperature for the next 6 days based on 3 months of observed data, respectively. Zakia et al. [22] used a nonlinear autoregressive network (NARX) to achieve temperature prediction for nearly one month by virtue of a 1-year-long training dataset with a total of 12 data features, through the division of the model for different seasons. The prediction accuracy decreased with the growth of the prediction time length, and the short-time indoor temperature prediction had a high practical value from the point of view of guiding the operation and regulation of building air conditioning systems.

In the selection of model algorithms, various types of regression models, especially neural network models, are widely used. Mustafaraj et al. [21] developed and validated two models using external and indoor climate data recorded over three months to predict dry bulb temperature and relative humidity at different time-scales (30 min to 3 h ahead), finding that the neural network-based nonlinear autoregressive model with external inputs (NNARX) model is superior to the linear parametric autoregressive model with external inputs (ARX) model. In similar studies, e.g., Refs. [23–25], the NARX network has the advantages of fast convergence and good generalization ability compared to other networks in dealing with time series datasets. Several previous studies, e.g., Refs. [14,26], compared different data-driven modeling techniques in predicting indoor space temperature or other related parameters and found that nonlinear models, such as neural networks, gave satisfactory results in predicting indoor temperatures. Thomas et al. [12] predicted the indoor temperatures of two buildings for the next 15 to 30 min, finding that a nonlinear ANN model trained using the Levenberg–Marquaurdt algorithm gave more accurate temperature predictions than a linear regression ARX model using the least squares method. Mirzaei et al. [27] found that an advanced model with more input parameters has significantly higher prediction accuracy compared to a simple model, although the computation time is slightly longer, and its advantages greatly outweigh the disadvantages. This modeling technique has become increasingly popular in recent years. There are also researchers in related fields using recurrent neural networks such as long short-term memory (LSTM) to predict temporal changes in temperature and achieving good results [11,28–30], and LSTM is a special type of recurrent neural network (RNN) that is capable of learning long-term dependencies, which is more evident in the prediction of temporal sequences.

It should be noted that currently, most research adopts a single model to solve machine learning problems, which has some obvious drawbacks, such as the inability to leverage the advantages of multiple models, poor generalization ability, high dependence on datasets, etc. As a result, it restricts the overall performance and robustness of machine learning algorithms. In contrast, ensemble learning models can improve the stability and accuracy of models by combining the predictions of multiple models [31]. Ensemble learning can better generalize in various complex data situations and is more tolerant to noise and outliers. Therefore, in future research, more attention should be paid to the application of ensemble learning models. Moreover, current research generally validates the accuracy of models using validation or testing sets partitioned from offline data, with relevant metrics applicable to offline datasets. However, there is less research on methods for online temperature prediction that have greater engineering application value.

In summary, the current research lacks discussion of the adjustment of the practical application model or technical solution, including the selection of the input and output data characteristics of the model, and the selection of the model and the application mode, and thus, there is a certain research and development space. To fill the above gaps, the main contributions in this study are as follows.

(1) In this study, the categories of relevant factors affecting room temperature and the detectability of their data were discussed, and the reasonable input parameters of the temperature prediction model were selected via correlation analysis.

(2) In this study, the prediction performance of multiple popular algorithms under the same dataset is compared horizontally, the superiority of the ensemble learning algorithm is verified, and the temperature prediction model is further designed based on such algorithms.
(3) The temperature change rate is studied and selected as the output parameter of the model to improve the prediction effect. The scheduler mechanism is designed to make the model run online and improve the accuracy and stability of the prediction system.
(4) Accurate and effective temperature online prediction can not only reduce building energy consumption and improve indoor comfort, but it can also provide new methods and ideas for the operation of the building energy system, enhancing its flexibility.

The structure of this paper is as follows: Section 1 reviews related work and explains the aims and main contributions of this paper. Section 2 provides a detailed introduction to the research methods, including model parameter selection, algorithm comparison, and online operation methods. Section 3 describes the experimental results obtained from the tests carried out in a case building. Based on the above content, the conclusion and limitations of this paper are presented in Section 4.

2. Methodology

This article proposes an online temperature prediction method for office buildings, which decouples indoor temperature prediction into three stages: model training, model prediction, and online operation. Decoupling the model training phase from the model prediction phase is due to the fact that during the model training phase, the completeness and accuracy of the data can be ensured through data preprocessing. However, during the model prediction phase, the data quality cannot be guaranteed, which often results in prediction biases and program errors. At the same time, considering that indoor temperature may be affected by factors that are not monitored by the building data system, resulting in theoretically unpredictable temperature changes, the model prediction stage is decoupled from the online operation stage, and a scheduler mechanism is introduced to realize the online operation of the model and ensure the accuracy of temperature prediction. The system framework is shown in Figure 1. Before designing the online operation method for the model, this article discusses the selection of model input parameters and designs experimental comparisons that include several mainstream algorithms such as single models and ensemble learning algorithms to select the best-performing model as the core prediction model in this article.

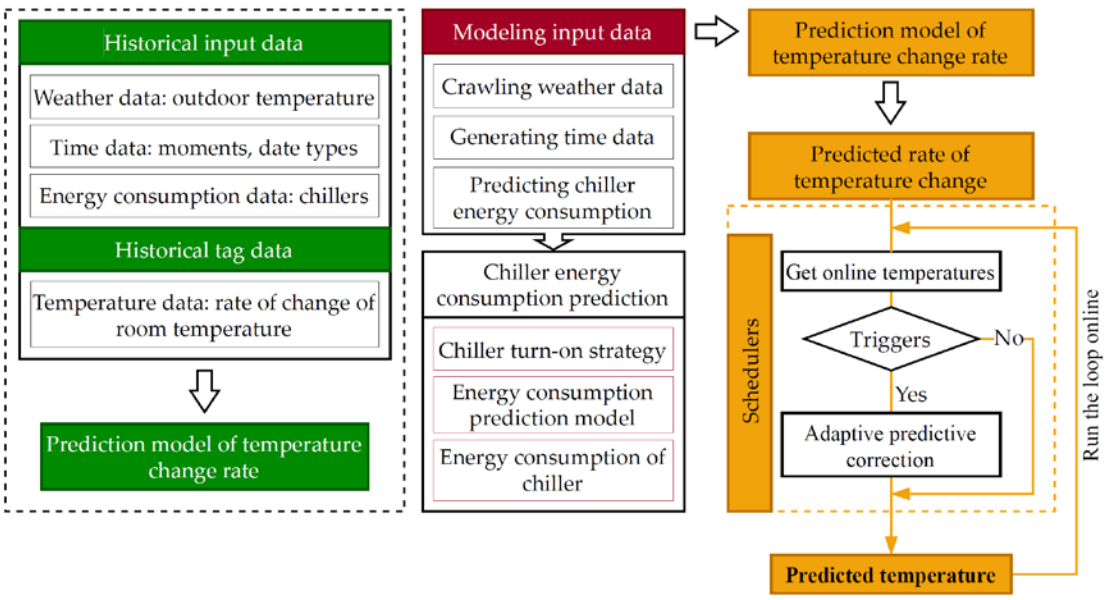


Figure 1. Online indoor temperature prediction framework.

2.1. Indoor Temperature Collection

Firstly, it is necessary to collect temperature data from indoor office buildings for model training and online prediction. An indoor temperature online monitoring system based on Internet of Things technology and temperature sensors is used to achieve 24 h monitoring of indoor temperature. The indoor temperature online monitoring system is based on LoRa technology and combines Wi-Fi to cover remote areas where LoRa signals cannot reach. The upper-level application layer protocol uses the MQTT protocol based on TCP/UDP. The system monitors at a time granularity of 5 min and uploads the temperature data to a web server for storage and record keeping.

2.2. Select Model Input and Output Parameters

2.2.1. Selection of Input Parameters

Insufficient input parameters in a model can lead to low prediction accuracy, while an excessive number of parameters can increase the difficulty of prediction and reduce computational efficiency. This article analyzes the selection of model input parameters by focusing on the factors that affect indoor temperature. Indoor temperature is mainly influenced by four factors: "outdoor weather", "enclosure structure", "personnel and equipment", and "HVAC systems". The theoretical analysis of relevant parameters that affect indoor temperature in this article focuses on their detectability. This is because a necessary condition for model input parameters is that the data can be monitored, collected, and transferred to the model for processing. The specific analysis of relevant parameters that affect indoor temperature is described in the following text.

Firstly, meteorological parameters are among the monitorable parameters. For office buildings, we can easily obtain real-time and future meteorological parameters from various weather stations. These parameters are closely related to indoor temperature. Secondly, the thermal performance of traditional building envelope structures is often difficult to monitor in real time. However, once the envelope structure of most buildings is built, its basic performance will remain relatively stable or change slowly. Therefore, when analyzing the indoor environment, the performance parameters of the envelope structure can be considered as a stable influencing factor. Then, although some parameters of personnel and equipment can theoretically be monitored (e.g., the number of personnel can be recorded through punching cards, AI image recognition, and other techniques, and the power status of equipment can be uploaded to the cloud for recording and analysis through IoT technologies such as Zigbee, NB-IoT, and LoRa), in practice, comprehensive monitoring often requires high equipment costs. Moreover, considering the increase in people's privacy protection awareness, setting up a series of sensors inside commercial office buildings may be resisted by tenants and even cause disputes related to commercial secrets. In addition, in terms of personnel activities inside buildings, the feasibility of recording daily activities through relevant equipment or technologies is low, so this type of parameter is excluded. Finally, the system form of the HVAC (heating, ventilation, and air conditioning) of a building can be considered as a fixed influencing factor along with the building envelope. Although the operational status of the HVAC system can be monitored (e.g., the operation status of chillers, air conditioning units, and fresh air units can be monitored through additional sensors), their parameters are complex, mostly featuring switch status parameters, and their power consumption weights are different. Therefore, calculating the correlation results with the changing indoor temperature may often result in deviation. In addition, the collectible energy consumption data are often the overall energy consumption data of the HVAC system or subsystem, rather than the energy consumption data of the temperature prediction area. Therefore, HVAC system-related data are not considered. At the same time, considering that the type of date (whether a certain day is a working day) will also affect indoor temperature, this study eventually selects outdoor meteorological parameters and date type as input parameters for the model.

2.2.2. Correlation Analysis of Outdoor Meteorological Parameters

Meteorological parameters can be obtained from the meteorological departments of various countries or regions. Due to the numerous types of parameters, they can be combined with the indoor temperature online monitoring system proposed in Section 2.1 to collect room temperature data for correlation analysis. Parameters with a strong correlation to room temperature changes can be selected as input parameters for the model. There are three commonly used correlation coefficients: Pearson linear correlation coefficient (PLCC), Spearman rank-order correlation coefficient (SRCC), and Kendall rank-order correlation coefficient (KROCC) [32]. When the data distribution does not meet normal distribution, the analysis results of PLCC may be biased. At the same time, because the PLCC is calculated based on the variance and covariance of the original data, it is relatively sensitive to outliers, and it measures linear correlation. Therefore, even if the PLCC is 0, it can only indicate that there is no linear correlation between variables, but it is still possible to have nonlinear correlation. KROCC is used to reflect the index of correlation between categorical variables, and is suitable for situations where both variables are ordered categories. It is not suitable for correlation analysis in this paper. SRCC uses the rank order of two variables for linear correlation analysis and does not require the original variable distribution, making it a non-parametric statistical method. Even if outliers appear in the data, the impact on SRCC coefficient is very small because the rank of outliers is usually not significantly different. Therefore, the correlation coefficient used in this article is the SRCC, which is calculated by the following formula:

$$r_{\text{Spearman}} = 1 - \frac{6\sum(U_{\text{i}} - V_{\text{i}})^2}{n(n^2 - 1)} \tag{1}$$

In the formula, U_{i} and V_{i} correspond to the relative position or order of the specific values of X and Y. n denotes the number of singular data samples. Substituting position for absolute value of data improves the robustness of Spearman rank-order correlation coefficient (SRCC) to the distribution of data to a certain extent, making it more suitable for correlation analysis of data that does not meet normal distribution or linear correlation. The relationship between meteorological data and indoor temperature in this article may not be linearly correlated, and the distribution of data in meteorological data is not all normal distribution. Therefore, the SRCC is used. The closer the correlation coefficient is to -1 or 1, the stronger the correlation between variables; the closer the correlation coefficient is to 0, the weaker the correlation between variables.

2.2.3. Selection of Output Parameters

In this paper, the rate of change in the indoor temperature is utilized as the output parameter for the model prediction instead of the actual numerical value of the indoor temperature. The reason for this is that the predicted rate of temperature change, combined with easily obtainable real-time indoor temperature data, can result in a more accurate prediction of temperature. The combination of the two parameters can offer a better precision level. Additionally, the combination of predicted temperature change rate and real-time temperature can respond to abnormal changes in actual indoor temperature, making temperature prediction more flexible.

2.3. Model Algorithm Comparison and Selection

Comparing horizontally the mainstream regression prediction algorithms based on data mining in recent years, we select the best performing algorithm to conduct online prediction experiments. The algorithms compared in this paper include linear regression, support vector regression (SVR), regression tree, artificial neural network (ANN) in the single model algorithms, and Extreme Gradient Boosting (XGBoost) [33] in the ensemble learning algorithms. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree

boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

The input data for these models were meteorological data, date type, and HVAC energy consumption, filtered using the method described in Section 2.2.2, and the output data were the predicted indoor temperature change rate in the future. Historical datasets were divided into training and testing sets in a 7:3 ratio. All five algorithms adopted default parameter settings to ensure relative fairness in the comparison of methods. Specific parameter settings are shown in Table 2. Each model was firstly trained using the training set, and subsequently used the input parameters of the testing set to predict the temperature change rate. Considering the sensitivity of artificial neural network's predictive accuracy to network structure, we need to construct different forms of neural networks for testing to find the optimal neural network structure, so as to avoid significant deviation caused by parameter settings.

Table 2. Model parameter setting.

Algorithm Model	Method	Main Parameters
Linear regression	sklearn.linear_model.LinearRegression	fit_intercept = True, normalize = False, copy_X = True, n_jobs = None
SVR	sklearn.svm.SVR	kernel = 'rbf', degree = 3, gamma = 'scale', coef0 = 0.0, tol = 0.001, C = 1.0, epsilon = 0.1, shrinking = True, cache_size = 200, verbose = False, max_iter = −1
Regression tree	sklearn.tree.DecisionTreeRegressor	criterion = 'mse', splitter = 'best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1
ANN-Default	sklearn.neural_network.MLPRegressor	hidden_layer_sizes = 100, activation = 'relu', solver = 'adam', alpha = 0.0001, batch_size = 'auto'
XGBoost	xgboost.XGBRegressor	base_score = 0.5, booster = 'gbtree', colsample_bylevel = 1, colsample_bynode = 1, colsample_bytree = 1, gamma = 0, importance_type = 'gain'

The detailed prediction errors of each model algorithm are evaluated using the mean absolute error (MAE) and mean squared error (MSE). A smaller MAE and MSE represent better model prediction performance. The specific calculation formulas are as follows:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \tag{3}$$

2.4. Offline Prediction Method of Models

After selecting the optimal model using the method designed in Section 2.3, the model was trained and used to offline predict the rate of temperature change for the following day. It is worth noting that although using larger time span datasets during the training process often results in better predictive performance, in practical engineering applications, most public buildings currently do not have a comprehensive data collection system, which cannot provide larger time span datasets for training models. The collection of larger time span datasets will delay the design of building energy management systems, thereby affecting the construction period of relevant new construction or renovation projects. Therefore, the actual historical dataset used here should not be too large, and it is recommended to choose a period of about one week to match practical engineering application scenarios.

When predicting the rate of temperature change for the next day offline, input data similar to those used in the training phase of the same model are constructed as the input

data for the prediction model in order to achieve a prediction. Future meteorological parameter data can be obtained through meteorological network APIs in various countries or regions; future time data are generated through program-generated timing information and date type; future energy consumption data are an additional part that need to be considered in the model prediction and can be predicted by constructing a prediction model for the running time of the refrigeration unit and the refrigeration energy consumption combined with historical energy consumption data. Based on the above data, the model will predict the rate of future temperature change for the next day, which will serve as the basic data for the model's online operation in the following sections.

2.5. Online Operation Method of Models

This paper designs a scheduling mechanism to realize the online operation of the prediction system and uses online information to improve the accuracy of indoor temperature prediction. The scheduler can convert the predicted temperature change rate outputted by the model prediction phase into a predicted temperature and improve the accuracy and stability of temperature prediction using the indoor online temperature information. The triggers and adaptive prediction methods in the designed scheduler can solve the problems of temperature prediction execution time and execution method (Figure 2).

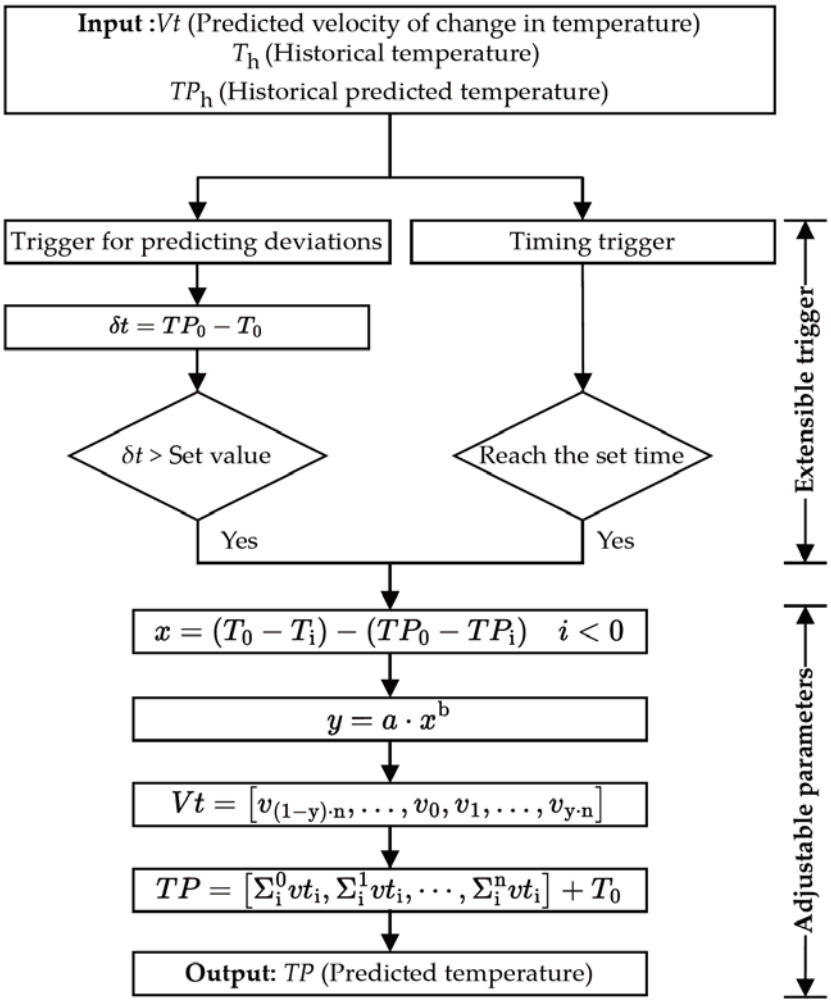


Figure 2. Model online prediction process diagram.

By designing triggers to solve the problem of predicting execution time, the online system only predicts the indoor temperature for the future period when the trigger is activated at a specific time, in order to reduce computational load. Based on this, this paper

designs prediction error triggers and timed operation triggers. When unpredictable factors such as indoor human activity cause prediction errors, the prediction error trigger activates the system to recalculate. The temperature prediction system needs to be regularly re-run to avoid possible temperature prediction errors, so a timed operation trigger is designed to guide the prediction calculation time.

Predictive execution method refers to generating predicted temperature values by combining the predicted temperature change rate with the actual temperature. Based on the current actual temperature, temperature prediction is achieved by accumulating the predicted temperature change rate (Figure 3). Due to the possible occurrence of sudden temperature changes and the problem of low accuracy of the predicted temperature change rate, the direct accumulation of temperature change rate to predict temperature may result in predictions that do not match reality. Therefore, it is necessary to correct the predicted temperature change rate. This article further designs an adaptive prediction correction method, which improves the online temperature prediction effect by adaptively correcting the predicted temperature change rate based on the prediction deviation that occurred in the past.

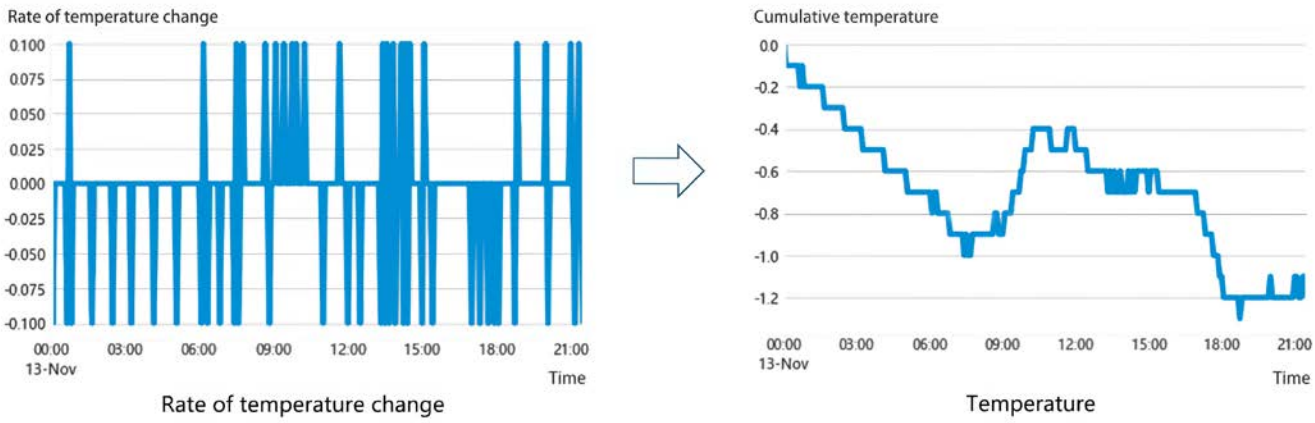


Figure 3. Method for prediction of temperature generation.

The corrected predicted temperature change rate consists of two parts: the historical actual temperature change rate over a certain period of time and the offline-predicted temperature change rate proposed in Section 2.4. The proportion of the actual historical temperature rate is determined by the prediction error over the past period of time. The proportion follows the following principle: the greater the offline prediction error over the past period of time, the more inclined it is to use the historical actual temperature change rate. Guided by this principle, this article proposes the following mathematical formula for automatically adjusting the proportion of historical actual temperature change rates:

$$y = a \cdot x^b \tag{4}$$

$$a = maxerr^{-b} \tag{5}$$

In Equation (4), y represents the actual temperature rate ratio over a past period. The parameter 'a' is related to the maximum allowable cumulative error, *maxerr*, over the same period, and its calculation is given in Equation (5). The specific growth trend of the ratio y is correlated to parameter 'b': when b is greater than 1, the function exhibits exponential growth; when b is equal to 1, the function exhibits linear growth; and when b is between 0 and 1, the function exhibits decreasing growth.

3. Case Results

3.1. Correlation Analysis of Meteorological Parameters

In this paper, an office building in Haidian District, Beijing was used as a case building. The office building has a construction area of 95,000 m^2 and a height of 68.5 m, of which 17 floors are above ground and 2 floors are underground. The working hours in the office area are mostly from 8:00 to 18:00. This building uses a fan coil unit and fresh air system for year-round ventilation, a chilled water unit for cooling during the summer months, and municipal heating during the winter. The original outdoor meteorological parameters were obtained from the China Meteorological Administration for the area, and 32 parameters were covered in total, including indoor temperature and date type, as shown in Table 3. Due to the absence of solar radiation data in the historical meteorological data provided by China Meteorological Administration, and the lack of an API for solar radiation data in that region, it is currently not feasible to include it in data analysis from a data availability perspective. This article collected and organized the meteorological data for Haidian District, Beijing and the indoor temperature data for the case study room from 3 January 2020 to 10 January 2020, obtaining a total of 2016 data points with 5 min intervals for the 32 parameters. The frequency distribution histogram is shown in Figure 4.

Table 3. Meaning of Meteorological Data Identification.

Parameter Name	Meaning	Unit	Parameter Name	Meaning	Unit
Station_Id_C	Station ID (Character)	/	WIN_D_INST_Max	Wind direction of extreme wind speed	/
Year	Year	Year	WIN_S_Max	Maximum wind speed	m/s
Mon	Month	Month	WIN_D_S_Max	Wind direction of maximum wind speed	/
Day	Day	Day	WIN_S_Avg_2mi	2 min average wind speed	m/s
Hour	Time	Hour	WIN_D_Avg_2mi	2 min average wind direction	/
PRS	Pressure	hPa	WEP_Now	Current weather	/
PRS_Sea	Sea level pressure	hPa	WIN_S_INST_Max	Extreme wind speed	m/s
PRS_Max	Maximum pressure	hPa	TEM_App	Apparent temperature	°C
PRS_Min	Minimum pressure	hPa	WIN_Pow	Wind power	/
TEM	Outdoor temperature	°C	VIS	Horizontal visibility (artificial)	m
TEM_Max	Maximum temperature	°C	CLO_Cov	Total cloud cover	%
TEM_Min	Minimum temperature	°C	CLO_Cov_Low	Low cloudiness	%
RHU	Relative humidity	%	CLO_Cov_LM	Cloud cover (low- or mid-level clouds)	%
RHU_Min	Minimum relative humidity	%	CLO_Height_LM	Cloud height (low- or mid-level clouds)	m
VAP	Vapor pressure	hPa	TEM_Indoor	Indoor temperature	°C
PRE_1h	1 h of precipitation	mm	D_Type	Date type	/

Excluding the data items that remain unchanged, such as Year and Month, as well as similar data items such as PRS_Min and PRS_Max, the 11 most typical variables are selected, which are Day, Hour, PRS, WIN_D_Avg_2mi, WIN_S_Avg_2mi, TEM, RHU, VAP, WEA_Now, WIN_Pow, and TEM_App. Referring to Formula (1), the SRCC is calculated by combining the aforementioned variables and indoor temperature, and the final result is shown in Figure 5. The absolute correlation of indoor temperature is ranked from high to low, with the results being room temperature, hour, perceived temperature, outdoor temperature, weather, day, humidity, wind power, wind speed, air pressure, water vapor pressure, and wind direction. It can be seen that the correlation between the time (Hour) and the indoor temperature reaches 0.59, indicating that the indoor temperature changes have a certain time regularity and show high repeatability in the data of one week. There is a strong correlation between outdoor temperature (TEM) and apparent temperature (TEM-App) and indoor temperature, while the absolute correlation of other parameters with indoor temperature is less than 0.24, indicating a weak correlation. For public build-

ings, parameters such as wind direction, wind speed, and water vapor pressure outside the building have low correlation with indoor temperature. Considering that the China Meteorological Administration does not provide future apparent temperature prediction values, in order to avoid constructing another apparent temperature prediction model that may result in a decline in the speed and accuracy of temperature change rate prediction, this paper ultimately chooses the time and outdoor temperature as the meteorological data input for the model.

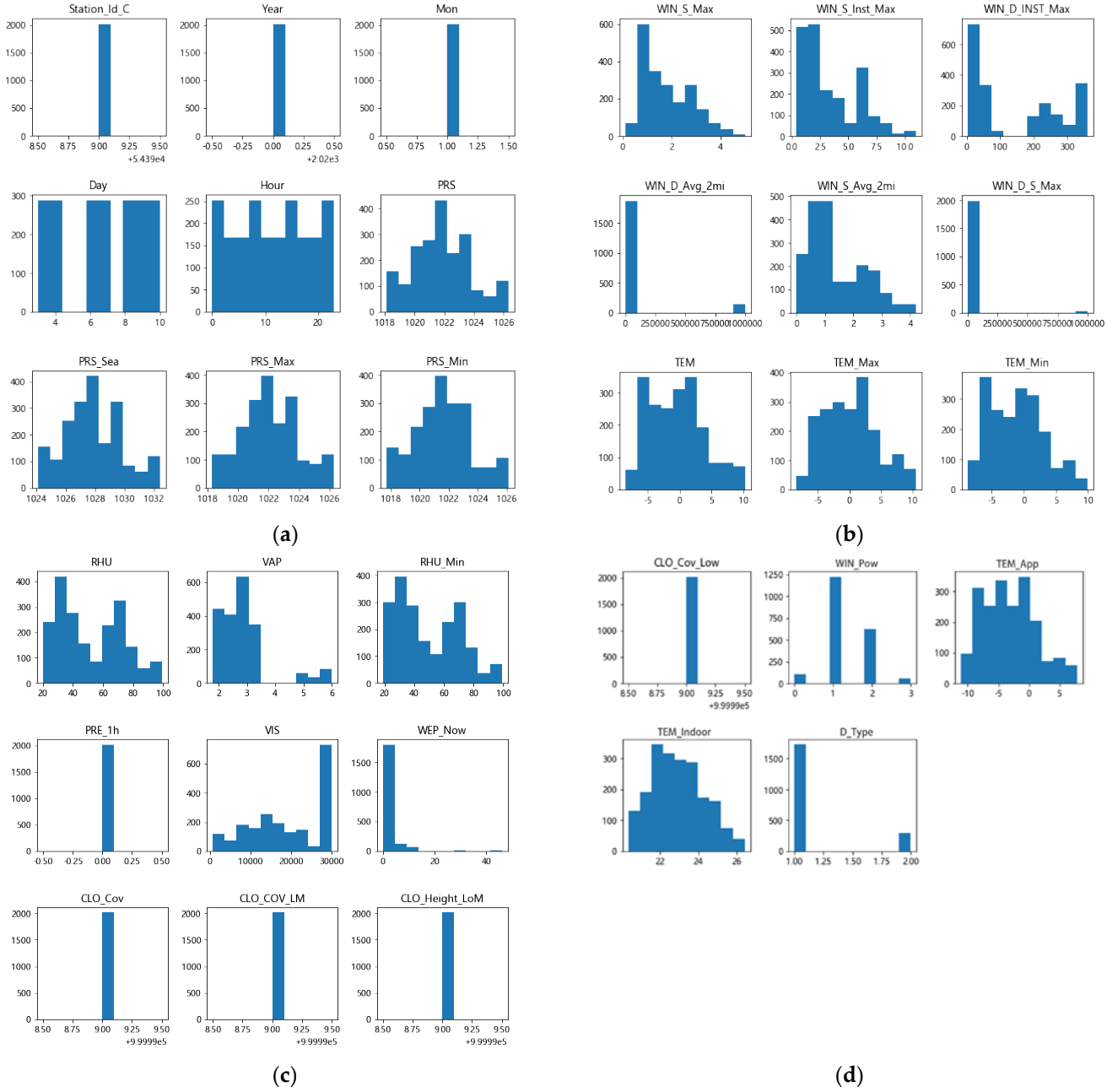


Figure 4. Histogram of frequency distribution of data: (**a**) No. 1~9; (**b**) No. 10~18; (**c**) No. 19~27; (**d**) No. 28~32.

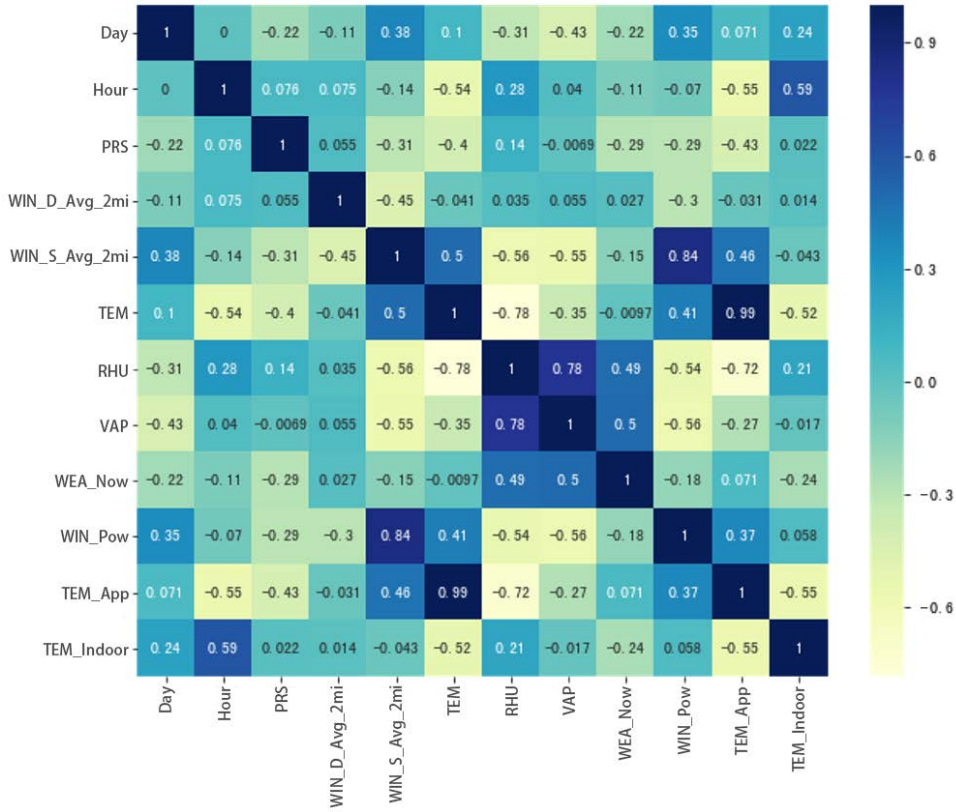


Figure 5. Thermographic map of the correlation between meteorological data and indoor temperature.

3.2. Horizontal Comparison of Model Algorithms

Based on the input data categories selected in the previous section, the relevant dataset from Beijing's Haidian district and case rooms from 3 January 2020 to 10 January 2020 were divided into training and testing sets in a 7:3 ratio. According to the parameters set in Section 2.3, each model was first trained using the training set and then used the input parameters from the testing set to predict the temperature change rate. The prediction effects of the five models are shown in Figure 6.

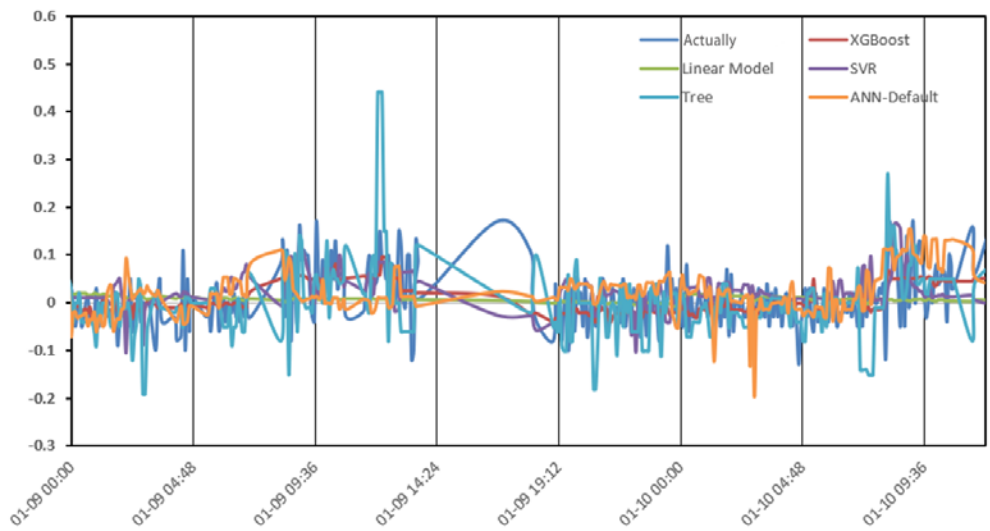


Figure 6. Prediction results of temperature change rate of different model algorithms.

In the ANN models, except for ANN-Default, which uses default parameters, other ANNs have undergone model adjustments, such as ANN-10xN, where 10 represents the

number of neurons in each hidden layer, and N at the end represents the number of hidden layers in the ANN. For example, ANN-10 × 2 represents an artificial neural network with two hidden layers, each with 10 neurons. In the paper, neural networks with 2, 4, 8, 16, 32, and 64 hidden layers were designed, and their predictive performance is shown in Figure 7.

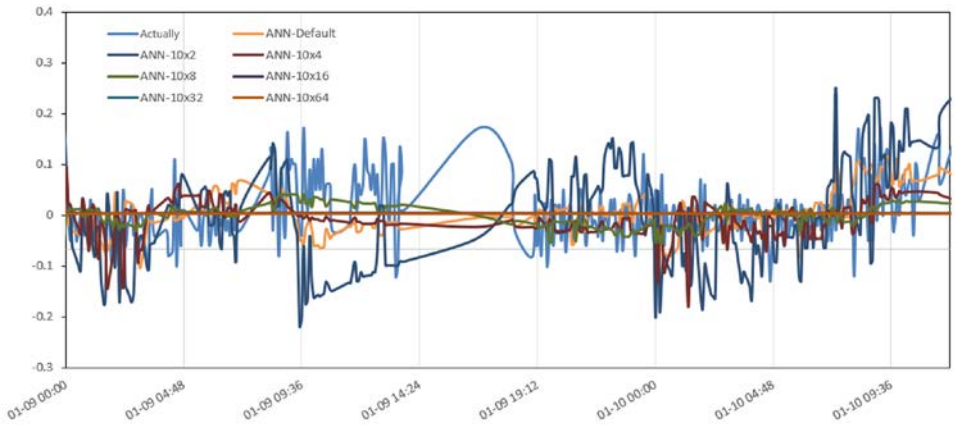


Figure 7. ANN model prediction results for various parameters.

Referring to Formulas (2) and (3), the prediction errors of different algorithms were calculated, and the results are shown in Table 4 and Figure 8. Under the default parameter settings of the five models, XGBoost outperforms the other four methods with default configurations, and it has advantages and disadvantages compared to the adjusted ANN, such as ANN-10 × 8 and ANN-10 × 16, which have some improvement in MAE but are inferior to XGBoost in MSE. At the same time, considering that XGBoost has good compatibility with input parameter formats and occupies an advantage in the upgrading and adjustment of subsequent models, XGBoost has a good generalization performance. Therefore, this paper selects XGBoost as the core prediction model for further prediction and optimization discussions.

Table 4. Prediction error of different models.

Model	MAE	MSE	Model	MAE	MSE
XGBoost	0.052329	0.008973	ANN-10 × 2	0.093193	0.01829
Linear Model	0.056628	0.00993	ANN-10 × 4	0.074262	0.012442
SVR	0.059398	0.010467	ANN-10 × 8	0.052177	0.00924
Tree	0.073302	0.013473	ANN-10 × 16	0.051239	0.009021
ANN-Default	0.063937	0.01075	ANN-10 × 32	0.057315	0.010021
			ANN-10 × 64	0.056734	0.009971

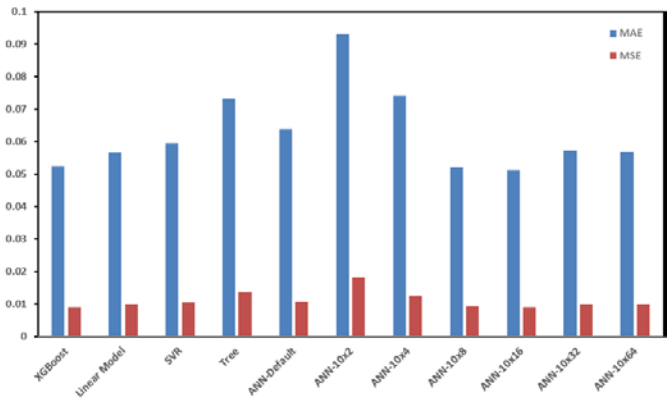


Figure 8. Prediction error of different models.

3.3. Online Prediction of Indoor Temperature

3.3.1. Model Offline Prediction

Continuing with the example of a certain office building in Haidian District, Beijing, as stated in the previous text, relevant data from 3 January 2020 to 9 January 2020, a total of 7 days, were selected for training. The outdoor weather parameter data were obtained from China Meteorological Administration with a data time interval of 1 h, totaling 168 entries. The indoor temperature historical data had a time interval of 5 min, totaling 2016 entries. Temperature fluctuation rate data were obtained by taking the difference between consecutive data points and filling in the first entry with a 0, resulting in 2016 entries. After the aforementioned training, this model was used to predict the indoor temperature fluctuation rate of a specific room in the building on 10 January 2023, providing historical forecasting results for online operation and generating more accurate modified forecast temperature fluctuation rates by combining actual temperature fluctuation rates in the recent past. Future weather data were obtained from China Meteorological Administration through a daily sample of outdoor temperature points with a time interval of 1 h and was subsequently refined and completed into 288 entries with a time interval of 5 min using a spline interpolation. This in turn generated 288 timestamps and date types. Based on the temperature rate forecasting model mentioned above, this paper predicts the temperature rate fluctuation of the tested area in the building on 10 January 2023, as shown in Figure 9.

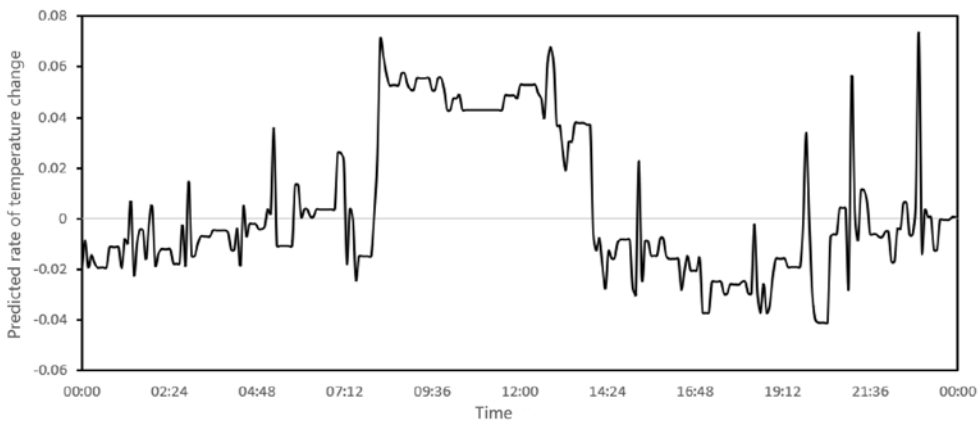


Figure 9. Temperature rate of offline prediction.

3.3.2. Online Running Results

In actual operation, this study only runs from 7:00 to 22:00 every day. During this time period, the majority of tenants and merchants in the case building work and do business. The temperature prediction system is meaningful for guiding the operation of the air conditioning system and improving the indoor environment. The significance of temperature prediction outside of working hours is insufficient and therefore not predicted.

The maximum error threshold setting of the online operation scheduler's prediction deviation trigger is set to 0.7 °C, and the timer trigger is set to 1 h. Referring to Formulas (4) and (5), the *maxerr* setting in the adaptive prediction method of the scheduler is set to 1.2 °C, and *b* is set to 0.7. The temperature prediction system prediction effect is shown in Figure 10.

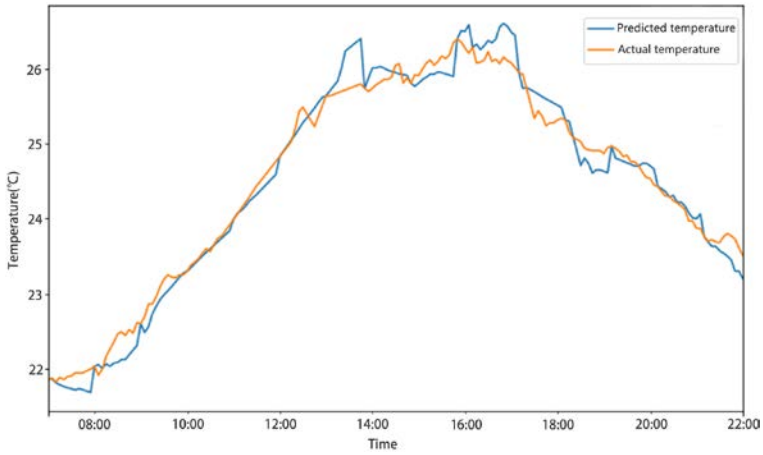


Figure 10. Online prediction run results.

It is clear that the predicted temperature is in line with the actual temperature performance. The maximum error occurs at 13:45 with an error of 0.6 °C and an average of −0.008 °C for the error, which indicates that the prediction error does not have a significant skew. The mean absolute error (MAE) is 0.16 °C, the overall error is low, and the prediction results are valid and practical. The specific error distribution is shown in Figure 11.

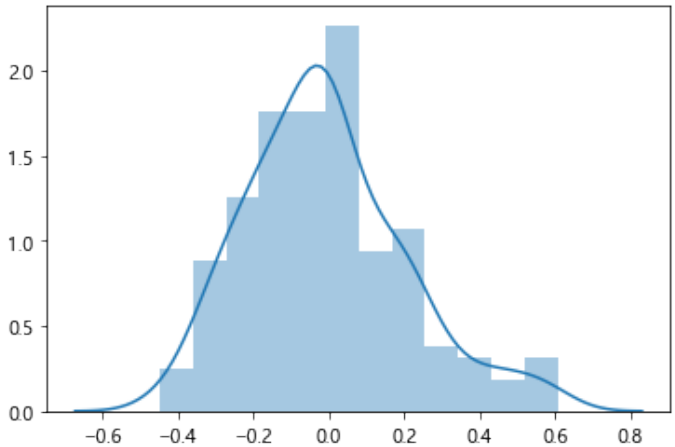


Figure 11. Error distribution histogram.

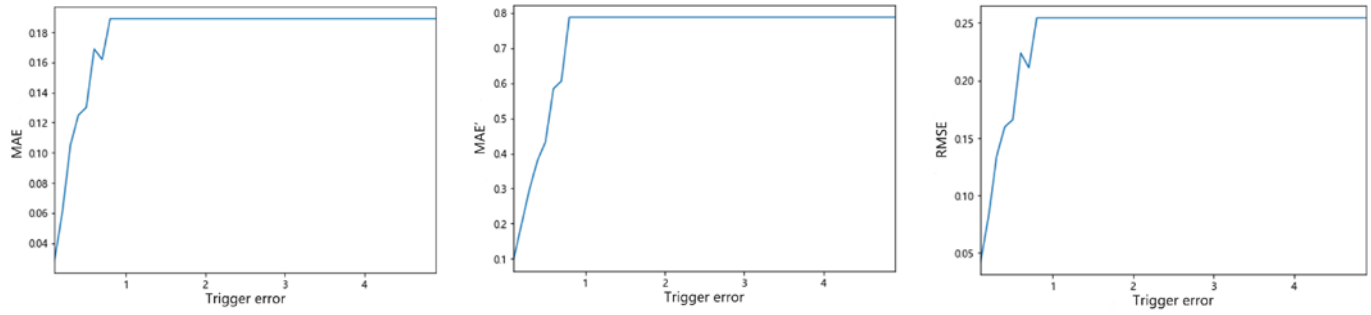
3.4. Optimization of the Scheduler

The prediction error of online operation depends on the parameter settings of both the temperature change rate prediction model and the scheduler. The following discussion will focus on the impact of the parameter settings, particularly those related to Formulas (4) and (5), in the scheduler on the prediction results. The error metrics used to measure accuracy are mean absolute error (MAE), maximum absolute error (MAE′), and root mean square error (RMSE).

3.4.1. Effect of Trigger Error on Results

Set the trigger error to 0.1, 0.2, 0.3,..., 5 °C, respectively, for a total of 50 trigger errors and observe the operation of the scheduler. The default setting for *maxerr* is 1.2 °C and for *b* is 0.7. From Figure 12, it can be seen that the MAE, MAE′, and RMSE have similar trends. After a certain trigger error, all error indicators remain unchanged, or more precisely, the error indicators for trigger error settings of greater than or equal to 0.8 °C are exactly the same. This is because the trigger of the trigger is set at every 1 h interval for offline temperature prediction, and the error between the predicted temperature and the actual temperature does not exceed 0.8 °C for that day. Therefore, under the setting of larger

trigger errors, the error trigger has never been triggered, resulting in exactly the same performance of the scheduler for trigger errors of greater than or equal to 0.8 °C.



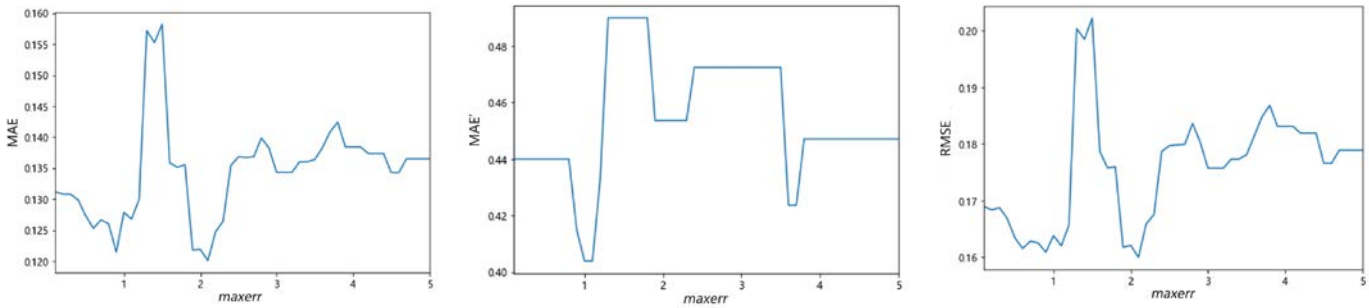
(**a**) (**b**) (**c**)

Figure 12. (**a**) MAE for different triggering errors; (**b**) MAE′ for different triggering errors; (**c**) RMSE for different triggering errors.

The mechanism of error trigger lies in recalculating and correcting larger prediction errors; therefore, the smaller the triggering error is, the better the final error performance. However, setting a smaller trigger error leads to frequent computer calculations, and more importantly, frequent corrections cause temperature prediction to lose its predictive significance. Therefore, the trigger error should be set according to the actual situation, taking into account the fact that too large an error will result in invalid triggers, and also according to the balancing prediction accuracy and calculation frequency. In this paper, a trigger error of 0.3~0.7 °C has shown good performance.

3.4.2. Effect of the Maxerr Parameter of the Adaptive Correction Method on the Results

Set *maxerr* to 0.1, 0.2, 0.3, . . ., 5 °C, respectively, totaling 50 cases. The trigger error is set at 0.5 °C, with the default set to 0.7. From Figure 13, it can be observed that the error indicators for different *maxerr* have similar trends. When *maxerr* varies within the range of 0.1~0.8 °C, the absolute maximum error remains unchanged, while the average absolute error and root mean square error show a decreasing trend. This indicates that the growth of *maxerr* within the range of 0.1~0.8 °C has a significant effect on improving the overall performance of temperature prediction, but it does not have a significant improvement effect for extreme temperature prediction. When *maxerr* varies within the range of 0.1~0.8 °C, all three error evaluation indicators perform well, with low prediction errors when *maxerr* = 1 °C. When *maxerr* varies within the range of 1.2~2.1 °C, the prediction errors are larger. When *maxerr* = 2.1 °C, the root mean square error and average absolute error indicators remain steady compared to when 0.8 °C $\leq$ *maxerr* $\leq$ 1.2 °C, but the maximum absolute error indicator performs poorly. In conclusion, setting *maxerr* within the range of 0.1~1.2 °C yields better results for the scheduler.

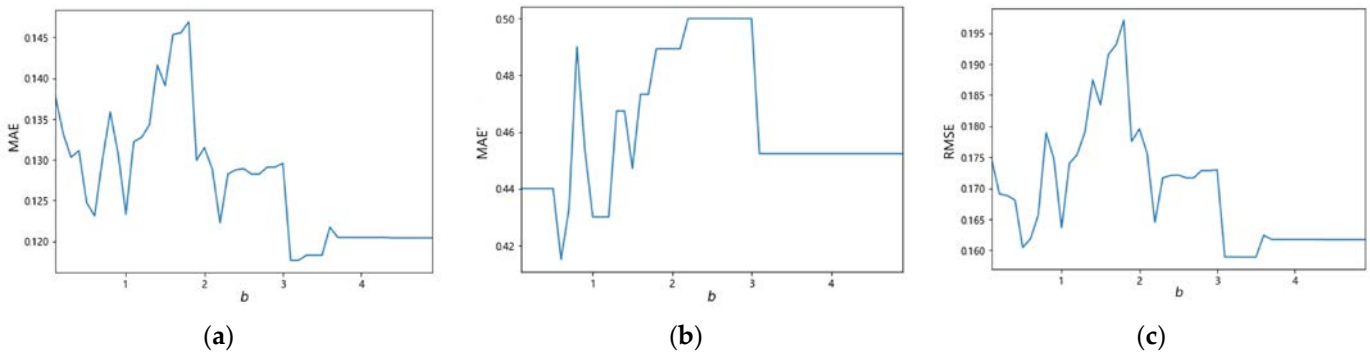


(**a**) (**b**) (**c**)

Figure 13. (**a**) MAE for different *maxerr*; (**b**) MAE′ for different *maxerr*; (**c**) RMSE for different *maxerr*.

3.4.3. Effect of the b-Parameter of the Adaptive Correction Method on the Results

Set b to 0.1, 0.2, 0.3,..., 5, for a total of 50 scenarios. According to the results of Sections 3.4.1 and 3.4.2, set the trigger error to 0.5 °C and *maxerr* to 1.2 °C. The experimental results are shown in Figure 14. It is observed that the relationship between the error indicators and scheduling parameter b exhibits irregular characteristics. When $0.1 \leq b \leq 3$, the three error indicators fluctuate. However, when $b \geq 3$, the performance of all three indicators tends to stabilize. This is because larger b values make the scheduler less sensitive to past errors and more likely to use offline temperature predictions directly. Hence, all schedulers are in a state of insensitivity to past errors, resulting in similar performance. Based on the performance of the three error indicators, a value of b = 0.5 is determined to yield the best performance of the scheduler.



(**a**) (**b**) (**c**)

Figure 14. (**a**) MAE for different b; (**b**) MAE′ for different b; (**c**) RMSE for different b.

4. Conclusions

The predicted values of the indoor temperature in large office buildings can provide important data support for the development of HVAC operation strategies. However, most of the current research focuses on offline prediction using single models, and the associated indicators are only applicable to offline datasets, with relatively little research on the technology of online temperature prediction. In this paper, an online prediction method for indoor temperature in office buildings based on the XGBoost algorithm was designed, and a scheduler was developed with the purpose of ensuring online operation and predicting accuracy. The impact of different parameter configurations of the scheduler on the prediction results was also discussed. The effectiveness and accuracy of the method were validated using actual data from a commercial office building in Haidian District, Beijing. The specific conclusions are as follows:

(1) Through analysis, it was determined that outdoor meteorological parameters and date type can be used as input parameters for the model. Additionally, the suitability of time information and outdoor temperature as input parameters for this experiment was proven through Spearman rank-order correlation coefficient (SRCC) calculation.
(2) Through the designed experiments, a horizontal comparison of several mainstream single model algorithms and ensemble learning algorithms, including the XGBoost algorithm, was carried out to verify the superiority of the ensemble learning algorithm XGBoost. To avoid the influence of the ANN model structure on the test accuracy, six additional ANN models were constructed and added to the comparative experiments to improve the experimental rigor.
(3) The temperature change rate was selected as the output parameter of the core temperature prediction model, and relevant explanations were provided. The proposed dispatching mechanism can improve the accuracy and stability of the prediction system by utilizing online information while realizing the online operation of the model. The test results of online operation showed no obvious bias and had a low overall error.

The indoor temperature online prediction method proposed in this paper can assist BEMS in executing more efficient HVAC system group control strategies in a timely fashion based on the predictive results. This can achieve desired objectives such as improving the effectiveness of group control strategies for chillers or heat pumps, thereby reducing building energy consumption and enhancing indoor comfort. In addition, the actual operation of the building will be recorded in the historical indoor temperature monitoring platform and power monitoring platform, providing data support for the building operation strategy library system.

Due to the complexity of the experiment, further investigation is needed to explore whether there is a synergistic effect between different hyperparameters and schedulers. At the same time, due to the limitations of the current experimental conditions, the impact of solar radiation and indoor carbon dioxide on the model prediction and the performance differences in models in different seasons will be the focus of future research in this field of study.

Author Contributions: Conceptualization, J.T. and K.L.; data curation, W.Y. and T.Z.; formal analysis, J.T., W.Y. and T.Z.; funding acquisition, X.Z.; methodology, J.T. and K.L.; project administration, X.Z.; resources, K.L.; software, T.Z.; supervision, X.Z.; validation, K.L. and W.Y.; writing—original draft, J.T.; writing—review and editing, K.L., W.Y. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Opening Funds of State Key Laboratory of Building Safety and Built Environment & National Engineering Research Center of Building Technology, BSBE2022-01, and the Basic Research Program Projects in Qinghai Province, China in 2023, 2023-ZJ-738.

Data Availability Statement: Data are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IEA—International Energy Agency. Available online: https://www.iea.org/data-and-statistics/data-sets (accessed on 28 December 2022).
2. González-Torres, M.; Pérez-Lombard, L.; Coronel, J.F.; Maestre, I.R.; Yan, D. A Review on Buildings Energy Information: Trends, End-Uses, Fuels and Drivers. *Energy Rep.* **2022**, *8*, 626–637. [CrossRef]
3. Lamsal, P.; Bajracharya, S.B.; Rijal, H.B. A Review on Adaptive Thermal Comfort of Office Building for Energy-Saving Building Design. *Energies* **2023**, *16*, 1524. [CrossRef]
4. Antunes, H.C.; Soares, S.; Gomes, Á. An Integrated Building Energy Management System. In *Mediterranean Green Buildings & Renewable Energy*; Sayigh, A., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 191–199. ISBN 978-3-319-30745-9.
5. Brambley, M.R.; Haves, P.; McDonald, S.C.; Torcellini, P.; Hansen, D.; Holmberg, D.R.; Roth, K.W. *Advanced Sensors and Controls for Building Applications: Market Assessment and Potential R&D Pathways*; EERE Publication and Product Library: Washington, DC, USA, 2005.
6. Tian, Z.; Ye, C.; Zhu, J.; Niu, J.; Lu, Y. Accelerating Optimal Control Strategy Generation for HVAC Systems Using a Scenario Reduction Method: A Case Study. *Energies* **2023**, *16*, 2988. [CrossRef]
7. Zhao, H.-X.; Magoulès, F. A Review on the Prediction of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3586–3592. [CrossRef]
8. Magalhães, S.M.; Leal, V.M.; Horta, I.M.L.; Isabel, M. Horta Modelling the Relationship between Heating Energy Use and Indoor Temperatures in Residential Buildings through Artificial Neural Networks Considering Occupant Behavior. *Energy Build.* **2017**, *151*, 332–343. [CrossRef]
9. Deb, C.; Eang, L.S.; Yang, J.; Santamouris, M. Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. *Energy Build.* **2016**, *121*, 284–297. [CrossRef]
10. Mba, L.; Meukam, P.; Kemajou, A. Application of Artificial Neural Network for Predicting Hourly Indoor Air Temperature and Relative Humidity in Modern Building in Humid Region. *Energy Build.* **2016**, *121*, 32–42. [CrossRef]
11. Xu, C.; Chen, H.; Wang, J.; Guo, Y.; Yuan, Y. Improving Prediction Performance for Indoor Temperature in Public Buildings Based on a Novel Deep Learning Method. *Build. Environ.* **2019**, *148*, 128–135. [CrossRef]
12. Thomas, B.; Soleimani-Mohseni, M. Artificial Neural Network Models for Indoor Temperature Prediction: Investigations in Two Buildings. *Neural Comput. Appl.* **2006**, *16*, 81–89. [CrossRef]

13. Xie, Q.; Ni, J.-Q.; Bao, J.; Su, Z. A Thermal Environmental Model for Indoor Air Temperature Prediction and Energy Consumption in Pig Building. *Build. Environ.* **2019**, *161*, 106238. [CrossRef]
14. Ashtiani, A.; Mirzaei, P.A.; Haghighat, F. Indoor Thermal Condition in Urban Heat Island: Comparison of the Artificial Neural Network and Regression Methods Prediction. *Energy Build.* **2014**, *76*, 597–604. [CrossRef]
15. Su, Y.; Chan, L.-C.; Shu, L.; Tsui, K.-L. Real-Time Prediction Models for Output Power and Efficiency of Grid-Connected Solar Photovoltaic Systems. *Appl. Energy* **2012**, *93*, 319–326. [CrossRef]
16. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
17. Afroz, Z.; Urmee, T.; Shafiullah, G.; Higgins, G. Real-time prediction model for indoor temperature in a commercial building. *Appl. Energy* **2018**, *231*, 29–53. [CrossRef]
18. Yadav, A.K.; Malik, H.; Chandel, S. Selection of Most Relevant Input Parameters Using WEKA for Artificial Neural Network Based Solar Radiation Prediction Models. *Renew. Sustain. Energy Rev.* **2014**, *31*, 509–519. [CrossRef]
19. Talebi, B.; Haghighat, F.; Mirzaei, P.A. Simplified Model to Predict the Thermal Demand Profile of Districts. *Energy Build.* **2017**, *145*, 213–225. [CrossRef]
20. Guo, Y.; Wang, J.; Chen, H.; Li, G.; Liu, J.; Xu, C.; Huang, R.; Huang, Y. Machine Learning-Based Thermal Response Time Ahead Energy Demand Prediction for Building Heating Systems. *Appl. Energy* **2018**, *221*, 16–27. [CrossRef]
21. Mustafaraj, G.; Lowry, G.; Chen, J. Prediction of Room Temperature and Relative Humidity by Autoregressive Linear and Nonlinear Neural Network Models for an Open Office. *Energy Build.* **2011**, *43*, 1452–1460. [CrossRef]
22. Afroz, Z.; Shafiullah, G.; Urmee, T.; Higgins, G. Prediction of Indoor Temperature in an Institutional Building. *Energy Procedia* **2017**, *142*, 1860–1866. [CrossRef]
23. Zemouri, R.; Gouriveau, R.; Zerhouni, N. Defining and Applying Prediction Performance Metrics on a Recurrent NARX Time Series Model. *Neurocomputing* **2010**, *73*, 2506–2521. [CrossRef]
24. Lin, T.; Horne, B.G.; Tiiio, P.; Giles, C.L. Learning Long-Term Dependencies in NARX Recurrent Neural Networks. *IEEE Trans. Neural Netw.* **1996**, *7*, 1329–1338. [PubMed]
25. Xie, H.; Tang, H.; Liao, Y.-H. Time Series Prediction Based on NARX Neural Networks: An Advanced Approach. In Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, Baoding, China, 12–15 July 2009; pp. 1275–1279.
26. Lu, T.; Viljanen, M. Prediction of Indoor Temperature and Relative Humidity Using Neural Network Models: Model Comparison. *Neural Comput. Appl.* **2009**, *18*, 345–357. [CrossRef]
27. Mirzaei, P.A.; Haghighat, F.; Nakhaie, A.A.; Yagouti, A.; Giguère, M.; Keusseyan, R.; Coman, A. Indoor Thermal Condition in Urban Heat Island—Development of a Predictive Tool. *Build. Environ.* **2012**, *57*, 7–17. [CrossRef]
28. Jiang, B.; Gong, H.; Qin, H.; Zhu, M. Attention-LSTM Architecture Combined with Bayesian Hyperparameter Optimization for Indoor Temperature Prediction. *Build. Environ.* **2022**, *224*, 109536. [CrossRef]
29. Fang, Z.; Crimier, N.; Scanu, L.; Midelet, A.; Alyafi, A.; Delinchant, B. Multi-Zone Indoor Temperature Prediction with LSTM-Based Sequence to Sequence Model. *Energy Build.* **2021**, *245*, 111053. [CrossRef]
30. Weng, K.; Mourshed, M. RNN-Based Forecasting of Indoor Temperature in a Naturally Ventilated Residential Building. In Proceedings of the 16th IBPSA Conference, Rome, Italy, 2–4 September 2019; pp. 3103–3108.
31. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A Survey on Ensemble Learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
32. Akoglu, H. User's Guide to Correlation Coefficients. *Turk. J. Emerg. Med.* **2018**, *18*, 91–93. [CrossRef]
33. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.