

Article

Short-Term Heavy Overload Forecasting of Public Transformers Based on Combined LSTM-XGBoost Model

Hao Ma ^{1,*} , Peng Yang ^{2,3}, Fei Wang ⁴ , Xiaotian Wang ¹, Di Yang ¹ and Bo Feng ¹¹ State Grid Hebei Marketing Service Center, Shijiazhuang 050021, China² School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China³ State Grid Hebei Electric Power Co., Ltd., Shijiazhuang 050021, China⁴ Department of Electrical Engineering, North China Electric Power University, Baoding 071003, China

* Correspondence: mahao3313@sina.com; Tel.: +86-133-1596-7750

Abstract: In order to effectively carry out the heavy overload monitoring and maintenance of public transformers in the distribution network, ensure the reliability of the distribution network power supply, and improve customer satisfaction with electricity consumption, this paper presents a short-term heavy overload forecasting method for public transformers based on the LSTM-XGBOOST combined model. The model extracts heavy overload feature variables from four dimensions, including basic parameter information, weather, time, and recent load, and constructs a short-term second highest load prediction model based on the LSTM algorithm to obtain the predicted value of the second highest load rate. After aggregating the heavy overload feature variables and the predicted second highest load rate, the XGboost algorithm is employed to construct a short-term heavy overload prediction model for public transformers to judge whether the public transformers display heavy overload. The test results show that this method has high accuracy in short-term heavy overload forecasting, and can effectively assist in the key monitoring and control of heavy overload in public transformers.

Keywords: network distribution transformer; heavy overload; LSTM; load forecasting; XGBoost



Citation: Ma, H.; Yang, P.; Wang, F.; Wang, X.; Yang, D.; Feng, B. Short-Term Heavy Overload Forecasting of Public Transformers Based on Combined LSTM-XGBoost Model. *Energies* **2023**, *16*, 1507. <https://doi.org/10.3390/en16031507>

Academic Editors: Vitor Monteiro and Abu-Siada Ahmed

Received: 15 November 2022

Revised: 16 January 2023

Accepted: 30 January 2023

Published: 3 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The load status of network distribution transformers (hereafter referred to as transformers) determines a power supply's quality and reliability to a large extent. With rapid economic development, the continuous improvement of people's living standards, and the rapid increase in power consumption related to industrial and commercial production and people's living, increasing loads have been imposed on transformers, causing heavy overloads, aging the equipment, and decreasing their service life [1]. Moreover, the heavy overloading of transformers quickly causes problems such as voltage instability and power outages, causing them to not meet power quality and reliability requirements [2]. Hence, developing a heavy overload prediction method is mandatory in order to effectively improve the power grid's emergency operation and ensure a transformer's power supply reliability. This prediction method will assist in load regulation, reduce voltage instability, and minimize power outages caused by heavy overloads. Furthermore, the heavy overload prediction method can also avoid the passive maintenance and management of transformers [3], improving power supply quality.

With the popularization of smart meters and all-event data collection, electricity consumption information systems already collect a massive amount of data, including the operation data of distribution transformers [4,5], providing a solid basis for big data analysis of such transformers. Considering the transformers' heavy overload prediction, most studies have considered load prediction as the starting point [6–9], and have used historical load data to mine the load variation patterns or employed machine learning methods to build load forecasting models [10–12]. For example, Shen et al. [13] predicted

the daily power load from the perspective of consumer habits. By dividing overall electricity consumption habits into several categories, similar daily load curves were identified, and a kernel function smoothing process weighted the similar daily load curves. After weighing, the per-unit prediction curve was obtained, and the base value was predicted by the multiplication smoothing method. Then, the base value was multiplied by the per-unit prediction to obtain the final prediction curve. To predict the heavy overload during the Spring Festival, Shi et al. [14] proposed a heavy overload prediction method based on a Back Propagation (BP) neural network and a gray model. This method predicted the transformers' load changes before and after the Spring Festival to determine the heavy overload condition. It is worth noting that load prediction is only one key factor in the overload prediction of distribution transformers, with the overload status defined as the load exceeding 80% of the capacity for at least two consecutive hours. Due to low load prediction accuracy, predictions solely involving load forecasting cannot meet the actual requirement because the weather, time, and operating features must also be considered.

Therefore, some studies have explored influencing factors of heavy overload, with He et al. [15,16] analyzing the relationship between the distribution transformers' overload and transformer, meteorological, and customer-related factors. The authors then developed a heavy overload prediction model based on random forest theory, considering the temperature, weekday, season, and other data as feature variables. Zhang et al. [17] examined the relationship between various heavy overload events, including equipment and user attributes, the natural environment, and short-term load features. Then, the heavy overload feature variables were used to establish a short-term heavy overload prediction model. Nevertheless, these methods only focused on the influencing factors of heavy overload and did not consider the fundamental load trend factor that causes heavy overload.

Spurred on by the deficiencies of the abovementioned methods, we propose a short-term heavy overload prediction method based on the combined Long Short-Term Memory (LSTM) -XGBoost model. Our technique first utilizes the LSTM algorithm to build a sub-model for load prediction that predicts the second highest load for the next day, which is a crucial feature variable for heavy overload prediction. Second, considering the historical operation data and meteorological conditions, twenty-two heavy overload feature variables are extracted from the historical data, including time, weather, historical load status, and recent load. Finally, the XGBoost algorithm is utilized to enhance the next day's short-term heavy overload prediction accuracy. The developed method comprehensively considers the influencing factors of heavy overload and the advantages of power load forecasting. Compared with current algorithms, LSTM affords a more accurate load prediction, while XGBoost achieves a more accurate prediction of heavy overload prediction.

2. Basic Theories

2.1. LSTM

A Recurrent Neural Network (RNN) is an enhanced multi-layer neural network comprising the input, hidden, and output layers [18]. The Long Short-term Memory (LSTM) network improves the RNN by solving the gradient disappearance problem. By setting the forget, input, and output gates, the LSTM combines the short-term and long-term memories [19,20], and learns the short-term and long-term data features. LSTM has demonstrated an appealing performance in several time-series forecasting problems [21,22]. Figure 1 illustrates the internal structure of the LSTM network.

Where h_t is the output vector at time t , and c_t is the long-term state vector of the network at time t . W_f , W_i , W_c , and W_o are the input weight matrix of each state, f_t , i_t , and o_t represent the forget gate, input gate, and output gate in the network. c_{1t} represents the current input state, s and $\tanh$ are the activation functions of the network, s is the sigmoid function, and $\tanh$ is the hyperbolic tangent function. The two functions introduce the nonlinear transformation into the neural network such that the network has robust nonlinear expressivity.

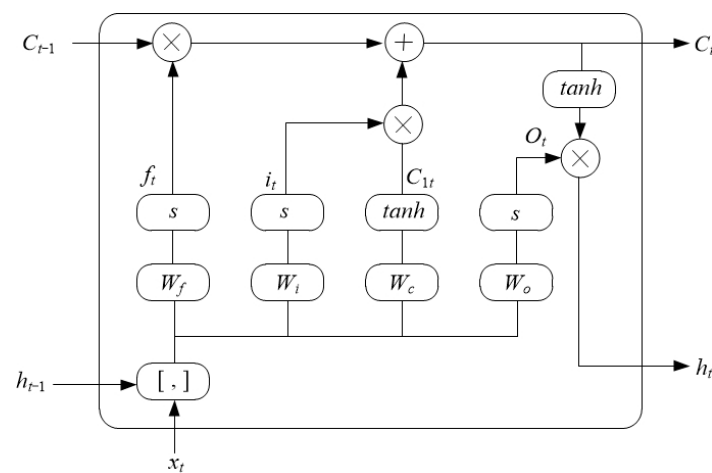


Figure 1. Structure of LSTM.

2.2. XGBoost

Before introducing XGBoost, it is necessary to introduce the ensemble learning and boosting strategies. Ensemble learning comprises multiple individual learners generated from training data through existing learning algorithms, which are combined to obtain a significantly better generalization ability than a single learner.

Boosting is a typical ensemble learning method, the principle of which is that the weight of each sample is the same in the initial training set, and a base learner is obtained through classic training. Then, the weight of each sample in the training set is adjusted according to the base learner result; i.e., the weights of the incorrect samples are increased and of the correct samples are reduced. A new learner is then obtained by training the adjusted samples. This process repeats until the number of basic learners reaches a pre-defined value, and these basic learners are then weighted and combined. Traditional boosting methods include boosting trees, gradient-boosted decision trees, and Adaboost.

The boosting tree is a method that uses a decision tree as the basic classifier. It uses the additive model and forward distribution algorithm to realize optimized learning. When the loss function is the square loss and the exponential loss, the optimization of each step is relatively simple but quite complex for standard loss functions. To solve this, the gradient-boosted decision tree (GBDT) algorithm uses the forward distribution algorithm [23], and the weak learner only uses the Classification and Regression Tree (CART) regression tree algorithm. GBDT combines the gradient algorithm and the regression tree algorithm.

XGBoost stands for extreme gradient boosting, and is an improved version of GBDT that affords high accuracy and efficiency in various classification tasks [24]. According to [25,26], XGBoost can effectively prevent over-fitting and presents an appealing generalization performance.

3. Short-Term Heavy Overload Prediction Model

Heavy overload is defined as the load rate of three or more consecutive sampling points exceeding 80%. A transformer's heavy overload occurrence is related to their condition, the number of users in the area, and the types of electricity consumption. These factors are also affected by weather, period, holidays, and industry. Therefore, we first extract 21 heavy overload feature variables from dimensions and employ the LSTM algorithm to predict the second highest load as the 22nd variable of the short-term heavy overload prediction. Based on the feature variables and the dataset, the XGBoost algorithm is utilized to construct a short-term prediction model for a heavy overload of distribution transformers.

3.1. Heavy Overload Feature Construction

From the basic parameter file information on the transformer, weather, time, recent load conditions, and other dimensions, we explore the feature vectors that affect the heavy

overload of distribution transformers, and we extract 21 feature variables, including region, industry, and capacity.

3.1.1. Features in the Basic Parameter File Dimension

Table 1 reports three feature variables in the basic archives dimension.

Table 1. Feature variables in the archives dimension.

No.	Feature Variables
1	Region
2	Industry
3	Capacity

The above feature variables can be obtained from the archives of public transformers in the electricity consumption information system.

(1) Regional features

The power demand growth rates among regions are uneven, and the heavy overload proportions of the public transformers in different regions substantially vary due to the rapid development of the social economy and other factors. Hence, the regions are divided according to the level of the power supply stations. Statistics regarding the proportion of heavy overload of public transformers in five power supply stations during a certain period (Figure 2) reveal that the public transformers of the JiaJiaKou power supply station suffer from the most frequent heavy overload.

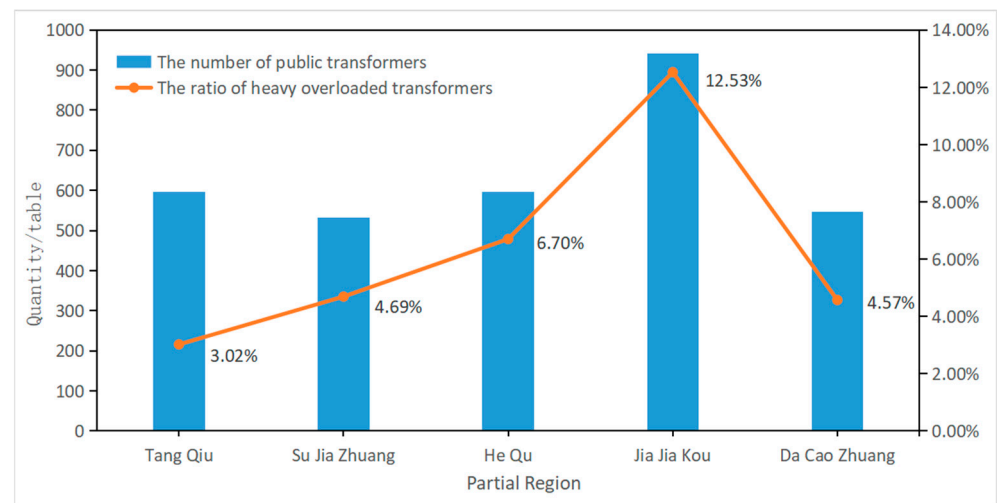


Figure 2. The ratio of the number of heavy overloads of public transformers in five typical power supply substations.

(2) Industrial features

The seasonal features and period laws of the power load of public transformers vary depending on the industry's features, which can be divided into urban residents, rural residents, mixed irrigation and drainage residents, and pure irrigation and drainage users. According to the statistics, the rural resident features present the most frequent heavy overload of public transformers (Figure 3).

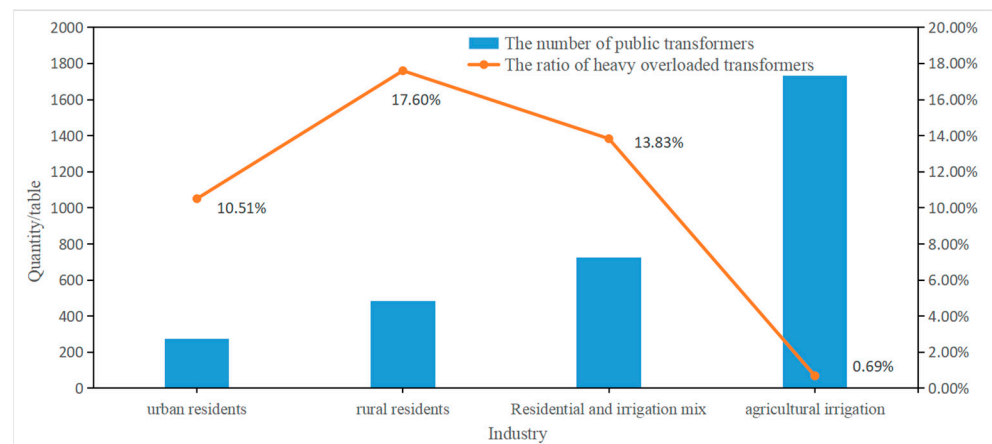


Figure 3. The ratio of the number of heavy overloads of public transformers in a specific period in a region according to the nature of the industry.

(3) Capacity features

The public transformers' loads present different states and change laws due to different capacity levels. Regarding the capacity levels, the public transformers can be divided into 20 kVA, 50 kVA, 100 kVA, and 250 kVA. According to the statistics regarding capacity levels, the 125 kVA-public transformers present the most frequent heavy overload of public transformers (Figure 4).

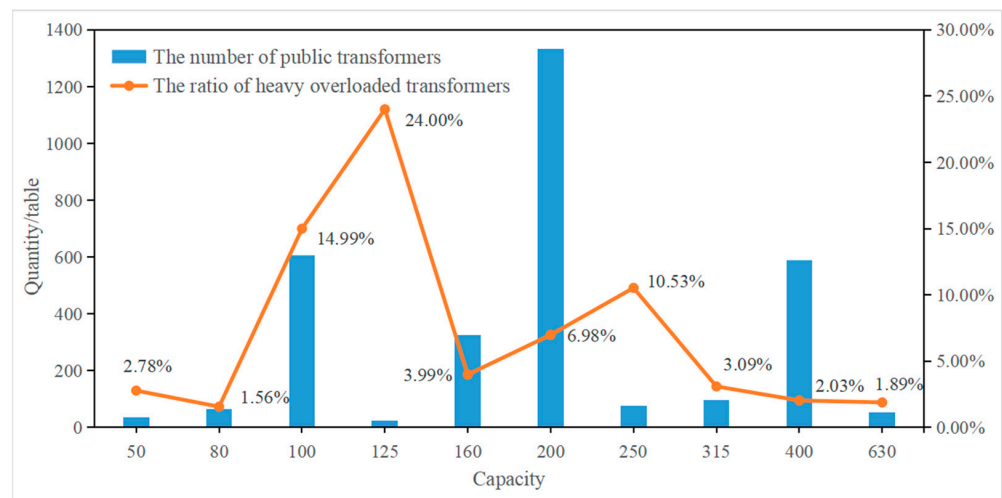


Figure 4. The ratio of the number of heavy overloads of public transformers counted by capacity in a specific period of a region.

3.1.2. Features in the Meteorological Dimension

Moreover, three feature variables in the meteorological dimension are extracted (Table 2).

Table 2. Feature variables in the meteorological dimension.

No.	Feature Variables
1	Daily maximum/minimum temperature
2	Daily maximum humidity
3	Daily precipitation

Meteorological feature variables, such as temperature and maximum humidity, can be obtained from the meteorological service platform. The daily minimum temperature

is selected from 15 November to 15 March of the next year, and the daily maximum temperature is selected for the remaining periods.

(1) Daily maximum/minimum temperature feature variable

During summer, the power load of air conditioners rapidly increases because of the rising temperature, and the probability of a heavy overload of public transformers increases. In winter, the residential electric heating load increases with the continuous temperature decrease, and the heavy overload of some transformers occurs more frequently. Thus, Figure 5 illustrates the load curve of a public transformer, highlighting that the higher the maximum temperature, the more frequent the heavy overload.

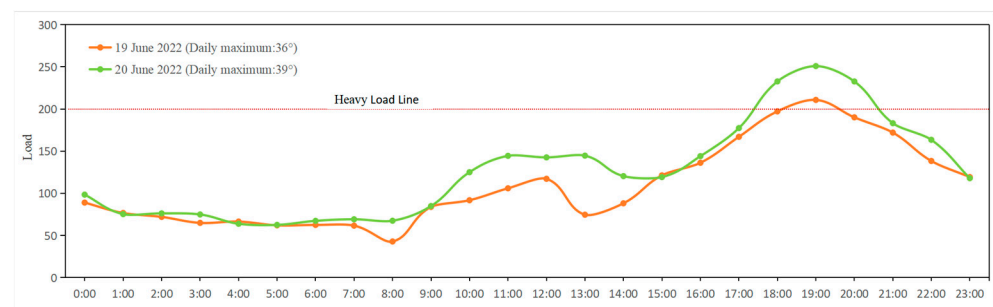


Figure 5. The load curve of a public transformer at different maximum temperatures.

(2) Daily maximum humidity and daily precipitation

Humidity affects the comfort of the human body and indirectly affects temperature regulation and dehumidification load. The greater the relative humidity, the greater the load. Rainfall can effectively reduce temperature, thus reducing the power load, which is evident during rainfall and precipitation on the short-term load of public transformers.

3.1.3. Features in the Time Dimension

Table 3 reports three feature variables in the time dimension.

Table 3. Feature variables in the time dimension.

No.	Feature Variables
1	Month
2	Weekday
3	Holidays

These features can be directly obtained from the electricity consumption information system's historical data and operation data. Holiday features are based on statutory holidays.

(1) Monthly features

The load utilization of public transformers varies depending on the month. For example, during the pure irrigation and drainage period, public transformers suffer a heavy load. In contrast, the heavy overload of public transformers for urban and rural residents occurs more frequently during the air conditioning load period in summer and the electric heating period in winter (Figure 6).

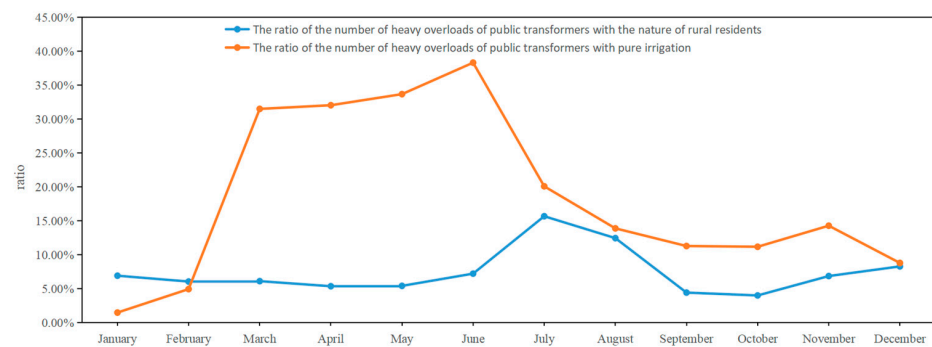


Figure 6. The ratio of the number of heavy overloads of public transformers in a region in different months.

(2) Weekday features

The load features of public transformers are closely related to the residents' daily life rules. According to the statistics related to weekday features, the quantity proportion of heavy overload of public transformers in a certain region within one year (Figure 7) reveals that the heavy overload of public transformers occurs more frequently on Saturdays.

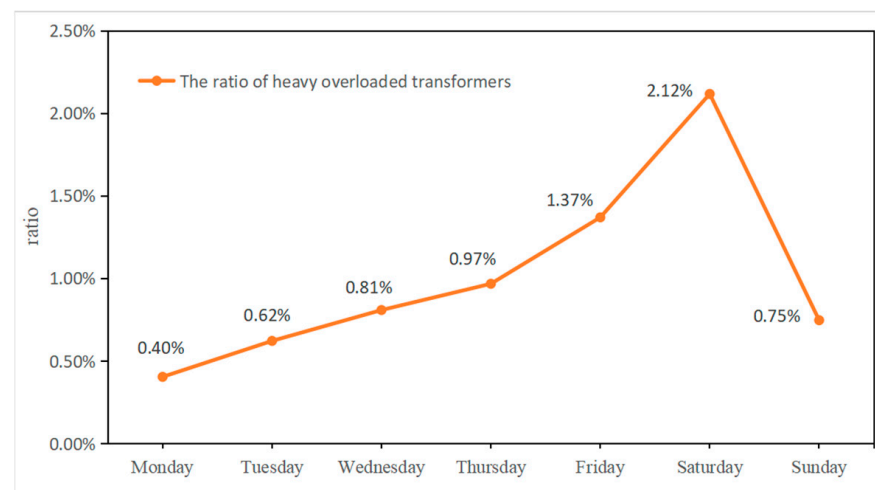


Figure 7. The ratio of heavy overload of public transformers in a certain area according to the weekly characteristics.

(3) Holiday features

Behaviors such as gathering and traveling during holidays (e.g., New Year's Day, Spring Festival, Tomb-Sweeping Day, May Day, Dragon Boat Festival, and National Day) greatly influence regional power load (Figure 8), which easily results in the heavy overload of some public transformers.

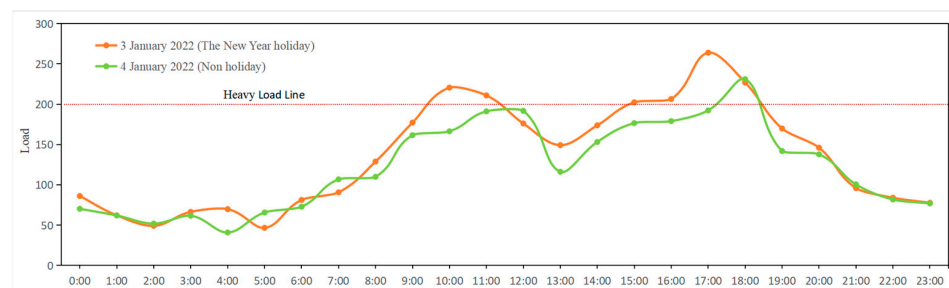


Figure 8. The load curve of a public transformer on New Year's Day holiday and non holiday.

3.1.4. Features in the Recent-Load Dimension

We also extract twelve feature variables in the recent-load dimension (Table 4).

Table 4. Feature variables in the recent load dimension.

No.	Feature Variables
1	Average daily maximum load rate for the previous three days
2	SD of daily maximum load rate for the previous three days
3	Number of days with heavy overload events in the previous three days
4	Number of heavy overload events in the previous three days
5	Number of points with heavy overload events in the previous three days
6	Average non-light or no-load points in the previous three days
7	Average daily maximum load rate for the previous seven days
8	SD of daily maximum load rate for the previous seven days
9	Number of days with heavy overload events in the previous seven days
10	Number of heavy overload events in the previous seven days
11	Number of points with heavy overload events in the previous seven days
12	Average non-light or no-load points in the previous seven days

The 12 feature variables in Table 4 are calculated from the operation data of the electricity consumption information system. The specific features are calculated as follows:

The average daily maximum load rate for the previous three days $r_{\max avg_3}$ is the average value of the maximum load rate per transformer to be forecasted in the previous three days, calculated as:

$$r_{\max avg_3} = \frac{\sum_{i=1}^3 r_{i\max}}{3} \quad (1)$$

where $r_{i\max}$ denotes the daily maximum load rate of the transformer on the i -th day considering the previous three days.

The SD of the daily maximum load rate for the previous three days r_{sd_3} is the standard deviation of the maximum load rate per transformer to be forecasted in the previous three days, calculated as:

$$r_{sd_3} = \sqrt{\frac{\sum_{i=1}^3 (r_{i\max} - r_{\max avg_3})^2}{3}} \quad (2)$$

The number of days with heavy overload events in the previous three days T_{zgz_3} is calculated by:

$$T_{zgz_3} = \sum_{j=1}^3 t_{zgz_j} \quad (3)$$

where t_{zgz_j} indicates whether heavy overload occurs on the j -th day. If heavy overload occurs, $t_{zgz_j} = 1$. Otherwise, $t_{zgz_j} = 0$.

The number of heavy overload events in the previous three days C_{zgz_3} is given by:

$$C_{zgz_3} = \sum_{j=1}^3 c_{zgz_j} \quad (4)$$

where c_{zgz_j} is the number of heavy overload events on the j -th day.

The number of points with heavy overload events in the previous three days D_{zgz_3} is given by:

$$D_{zgz_3} = \sum_{j=1}^3 d_{zgz_j} \quad (5)$$

where d_{zg_j} is the number of points with heavy overload events on the j -th day.

The average non-light or no-load points in the previous three days D_{fqz_3} is calculated as:

$$D_{fqz_3} = \sum_{j=1}^3 d_{fqz_j} \quad (6)$$

where d_{fqz_j} is the sum of the non-light or no-load points on the j -th day. The remaining six variables are the same but were calculated for the previous seven days.

3.2. Short-Term Second Highest Load Rate Prediction Based on LSTM

The operation data of the transformer is collected once a day on the hour; i.e., 24 times per day. The ratio of the second highest load to the capacity is called the second highest load rate.

The short-term load is primarily affected by weather and holidays. Therefore, when using LSTM to predict the second highest load rate, the model's input data should include the second highest load rate of the previous day and the weather, holidays, and weekday variables on the forecast day (Table 5). This is because the load is significantly affected by temperature, with the maximum summer and minimum winter temperatures strongly affecting the load. Moreover, the holiday variable is introduced to reflect the impact of holidays on the second highest load. If the day evaluated is a holiday, its value is one; otherwise, it is zero. The load variation from Monday to Sunday is reflected through the weekday variable; i.e., Monday = 1 and Sunday = 7.

Table 5. Feature variables of the prediction model for the second highest load rate.

Feature Variables	Content	Value
Weather	Maximum/minimum temperature	Actual number
Holiday	New Year's Day, Qingming, May Day, and Mid-Autumn Festival	0,1
Weekday	Monday, Tuesday, . . . , and Sunday	1 . . . 7

The output data is the predicted second highest load rate on the prediction day, with the corresponding LSTM model illustrated in Figure 9.

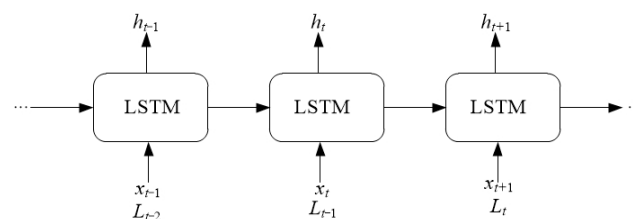


Figure 9. Second highest load rate prediction.

The LSTM's internal structure is depicted in Figure 1, where L_t is the second highest load rate on the t -th day, x_t includes the highest temperature, the lowest temperature, whether it is a holiday, and the weekday on the t -th day. h_t is the model output of the second highest load rate on the t -th day. The predicted second highest load rate is used as a feature variable for the short-term prediction of heavy overload.

3.3. Feature Processing

The extracted feature variables involve discrete string items and continuous numeric items. In order to facilitate machine learning and improve processing efficiency, the feature variables must be processed.

3.3.1. Discrete String Items

The region, industry, weekday, month, and holiday are discrete string variables. For example, the industry features include five values; i.e., urban residents, rural residents, resident irrigation and drainage mixture, pure drainage and irrigation, and unclassified. The machine learning algorithm cannot recognize the characters' meaning; thus, it is necessary to convert the values into numerical values, i.e., 1–5. Similarly, the region, industry, weekday, month, and holiday variables are numerically processed.

3.3.2. Normalizing the Continuous Features

Some feature variables, such as the average maximum load rate and the standard deviation of the maximum load rate in the previous three days, vary. If a feature's value range is quite different from the other features, it affects the sample distance calculated by the model, and the results will be inconsistent with the actual situation. Therefore, it is necessary to normalize the feature values to eliminate the negative impact of the value range on the results and increase the gradient decrease speed to quickly find the optimal solution. In this study, we exploit the linear normalization method:

$$z' = \frac{z - z_{\min}}{z_{\max} - z_{\min}} \quad (7)$$

where z' is the normalized feature value, z is the original feature value, and $z_{\min}$ and $z_{\max}$ are the maximum and minimum values of the variable. After linear normalization, the variables range is between 0 and 1.

3.4. Heavy Overload Prediction Based on XGBoost

The XGBoost algorithm builds a heavy overload prediction method and forecasts whether a heavy overload event will quickly occur. Specifically, the feature dataset is divided into training and test sets, with the XGBoost algorithm, exploited to construct the overload prediction model based on the training set. The prediction model is evaluated on the test set, and the prediction results are compared with the actual operation data to determine the model's accuracy.

3.4.1. Training Set

Typically, the data set is divided based on a specific ratio, e.g., 4:1, or using the n -fold cross-validation method, i.e., the dataset is divided into n equal parts, and one of them is sequentially selected for testing. The remaining $n - 1$ parts are used for model training, and the average of the n -tests determines the model's accuracy. The heavy overload prediction in this study has temporal characteristics. Moreover, in the actual prediction process, the model can only be trained on data collected before the prediction period. Hence, the random test set selection scheme from the feature data set does not conform to the actual application scenario. Therefore, in the case of fixed epochs, a certain period before the test set should be set as the training set. In the modeling process, the training set size affects the algorithm's accuracy and efficiency. Thus, the training set involves data related to the first three months in the test set.

3.4.2. Parameter Adjustment

In the heavy overload prediction model based on XGBoost, the learning rate, maximum depth, and the number of iterations affect the model's accuracy. Thus, adjusting these parameters optimizes the model's accuracy.

3.5. Steps of the Heavy Overload Prediction Process

The flowchart of short-term heavy overload prediction of public transformers illustrated in Figure 10.

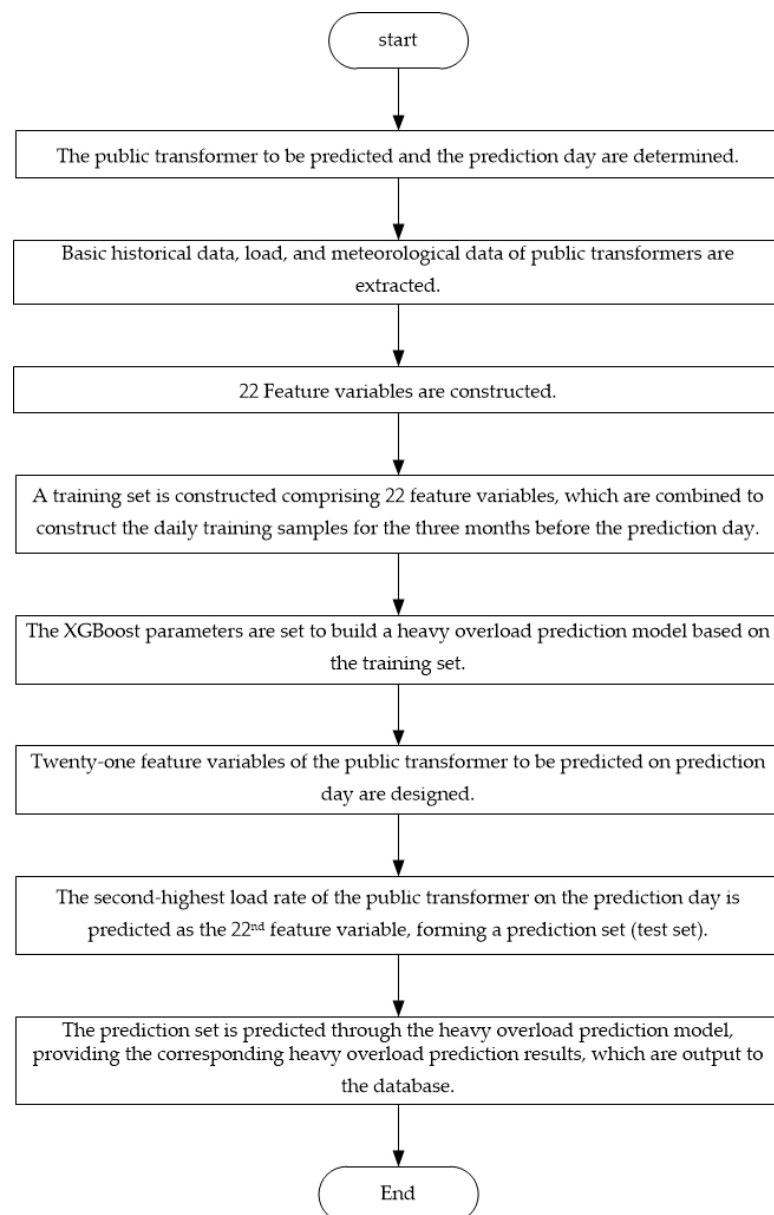


Figure 10. The flowchart of short-term heavy overload prediction of public transformers based on combined LSTM-XGBoost model.

Step 1: The public transformer to be predicted and the prediction day are determined.

Step 2: Basic historical data, load, and meteorological data of public transformers are extracted. Specifically, data concerning public transformer archives and load in the three months before the prediction day are extracted from the power consumption information acquisition system. Then, the meteorological service platform extracts meteorological data, such as maximum temperature, minimum temperature, maximum humidity, and daily precipitation, in this region during a three-month period.

Step 3: Feature variables are constructed. First, according to the public transformer historical and load data involving the three months before the prediction day, 18 features such as historical data dimension, time dimension, and recent load dimension are formed. Then, according to meteorological information data, three features of the meteorological dimension are extracted. After that, the second highest load rate of the public transformer on the previous day is extracted as a feature variable, and 22 feature variables are formed.

Step 4: A training set is constructed comprising 22 feature variables, which are combined to construct the daily training samples for the three months before the prediction day.

Step 5: The XGBoost parameters are set to build a heavy overload prediction model based on the training set.

Step 6: Twenty-one feature variables of the public transformer to be predicted on prediction day are designed. Specifically, 18 feature variables involve the historical data dimension, time dimension, and recent load dimension, and three feature data, i.e., maximum/minimum temperature, daily maximum humidity, and daily precipitation, are extracted on the prediction day.

Step 7: The second highest load rate of the public transformer on the prediction day is predicted as the 22nd feature variable, forming a prediction set (test set). The second highest load rate of the prediction day is obtained using an LSTM algorithm that exploits the data of the second highest load rate, the maximum temperature, the minimum temperature, and holidays/weekdays of the public transformer to be predicted for three months before the prediction day. Hence, a prediction set is constructed based on the designed 22 feature variables.

Step 8: The prediction set is predicted through the heavy overload prediction model, providing the corresponding heavy overload prediction results, which are output to the database.

4. Experiment

4.1. Hardware and Software

A big data analysis platform was built for this study, comprising nine servers (one interface server and eight Hadoop CDH cluster servers). The interface server utilizes the Oracle interface database, and the distribution data is extracted to the Hive data warehouse of the Hadoop CDH cluster through the unified interface program. The Hadoop CDH cluster consists of two master nodes and six sub-node servers, and uses the Hive data warehouse to process and analyze the massive data of the distribution variables, build and verify the model through Spark and TensorFlow, and store the results in the Hive data warehouse for a visual demonstration.

4.2. Experimental Data

In order to verify the accuracy of the proposed heavy overload prediction method, the power load data of 66,200 transformers in a specific area are extracted from the electricity consumption information system involving four test periods with typical seasonal features selected as the test set (Table 6). The training set comprises the feature data of the transformer in the previous three months, while the big data analysis platform is used for feature extraction and model construction.

Table 6. Typical test periods.

No.	Typical Test Periods	Test Dataset
1	From 20 March to 26 March	T1
2	From 24 July to 30 July	T2
3	From 30 October to 5 November	T3
4	From 25 December to 31 December	T4

4.3. Evaluation Indices

The main evaluation metrics are accuracy, recall rate, and $F1$, with accuracy P calculated as:

$$P = \frac{TP}{TP + FP} \quad (8)$$

where TP denotes the true positive, and FP is the false positive. The recall rate R is calculated by:

$$R = \frac{TP}{TP + FN} \quad (9)$$

where FN denotes false negative. $F1$ is calculated as:

$$F1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

4.4. Short-Term Second Highest Load Rate Prediction Results Based on LSTM

Taking the typical test period from 24 July to 30 July, the LSTM algorithm, neural network algorithm, and support vector machine regression algorithm are respectively applied to build the sub-high load rate model for 66,200 public transformers. The actual load rate of the transformers is used for verification, with the corresponding results illustrated in Figure 11, where the y -axis is the absolute value of the average error per model. The latter figure highlights that the LSTM method's error in the test period is lower than that of the competitor methods.

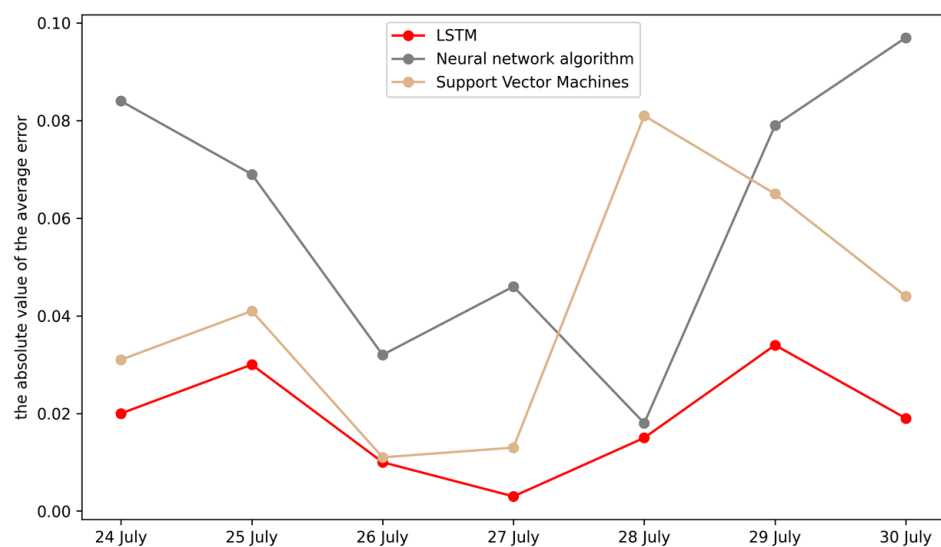


Figure 11. Prediction error in the second highest load rate of each algorithm.

4.5. Results

Based on the 22 heavy overload feature variables, the XGBoost algorithm is used to construct the prediction model, utilizing the parameters presented in Table 7. Table 8 and Figures 12–14 report the accuracy P , recall rate R , and $F1$ value of the test results.

Table 7. Parameter configuration of XGBoost.

No.	Parameter	Value
1	Number of iterations	5
2	Maximum depth	5
3	Weight minimum sum	1
4	Learning rate	0.1
5	L1 regular term coefficient	0.2
6	L2 regular term coefficient	0.2

The performance of LSTM-XGBoost was compared against the BP neural network, random forest, and GBDT. For fairness, the parameters of each algorithm were optimized before the trials.

From the perspective of business, grassroots business units pay more attention to the prediction accuracy of whether a public transformer will be overloaded in the short term.

The high accuracy of heavy overload prediction can better assist in the active repair and load monitoring of public transformers.

Table 8 reveals that the LSTM-XGBoost algorithm attains the highest accuracy of heavy overload prediction accuracy in all four test periods, with an average of about 85.87% and a recall rate of 72.67%. The results indicate that this study's heavy overload prediction method outperforms current methods.

Table 8. Performance of short-term heavy overload prediction of each algorithm.

Indices	Algorithm	T1	T2	T3	T4	Mean
<i>P</i>	BP	60.13%	68.12%	61.35%	72.52%	65.53%
	Random forest	74.32%	76.64%	70.51%	69.88%	72.84%
	GBDT	74.30%	80.48%	72.52%	75.36%	75.67%
	LSTM-XGBoost	86.91%	85.81%	83.99%	86.76%	85.87%
<i>R</i>	BP	59.88%	55.39%	64.65%	68.57%	62.12%
	Random forest	49.20%	68.86%	63.21%	62.21%	60.87%
	GBDT	65.11%	74.61%	69.83%	66.56%	63.64%
	LSTM-XGBoost	74.46%	75.54%	71.13%	69.55%	72.67%
<i>F1</i>	BP	60.00%	61.10%	62.96%	70.49%	63.64%
	Random forest	59.21%	72.54%	66.66%	65.82%	66.06%
	GBDT	69.40%	77.43%	71.15%	70.69%	72.17%
	LSTM-XGBoost	80.20%	80.35%	77.03%	77.21%	78.70%

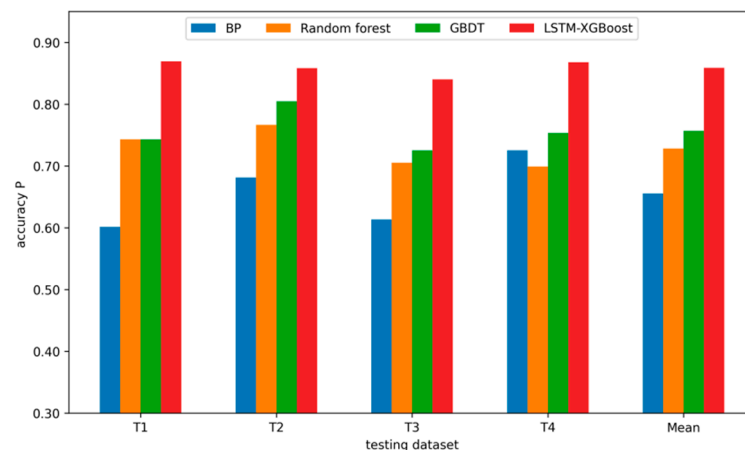


Figure 12. Accuracy P of heavy overload prediction models of public transformer based on various algorithms.

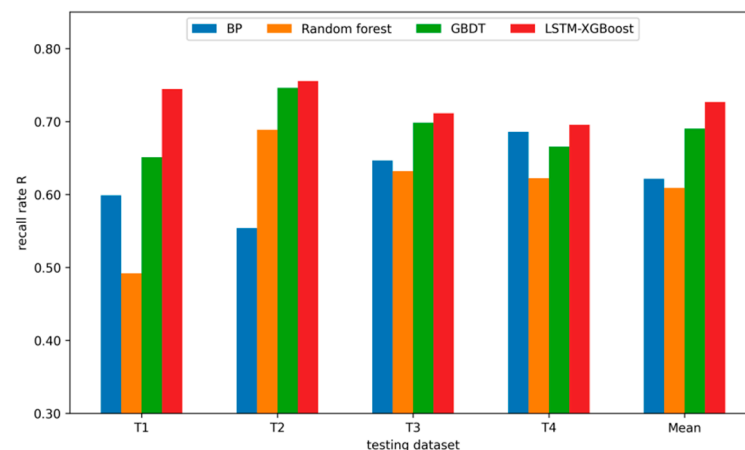


Figure 13. Recall rate R of heavy overload prediction models of public transformer based on each algorithm.

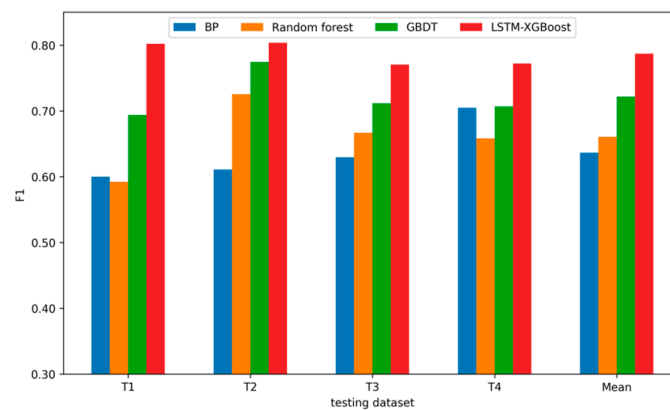


Figure 14. F1 of heavy overload prediction models of public transformer based on each algorithm.

5. Conclusions

This study extracts heavy overload feature variables from four dimensions, including basic parameter information, weather, time, and recent load, and constructs a second highest load rate prediction model based on the LSTM algorithm to obtain the predicted value of the second highest load rate. After aggregating the heavy overload feature variables and the predicted second highest load rate, the XGboost algorithm constructs a short-term heavy overload prediction model for distribution transformers. After experimental verification, the following conclusions are drawn:

- (1) The proposed LSTM-XGBoost model considers the influencing factors of heavy overloads, such as weekday, month, and temperature, and analyzes the impact of load prediction on the heavy overload. The results highlight that the proposed method improves short-term heavy overload prediction accuracy.
- (2) Future research will include the current heavy overload situation as a feature variable. Moreover, the industry types will be subdivided to construct the correlation model between the heavy overload and factors such as weather, period, and industry types. Then, a score will be obtained for each factor combination, and the higher the score, the more likely heavy overload occurs when the combination of factors occurs. The correlation score will be used as a heavy overload prediction feature variable, improving prediction accuracy.
- (3) The developed ultra-short-term heavy overload prediction model can detect and accurately locate heavy overload risks in advance. This is important, especially in periods such as an epidemic, helping frontline workers conduct early maintenance and quickly formulate plans in the presence of heavy overload events, ensuring high power supply efficiency in the epidemic prevention area.

Author Contributions: Conceptualization, P.Y.; Methodology, H.M.; Software, H.M.; Validation, H.M.; Formal analysis, H.M. and F.W.; Investigation, H.M., P.Y., D.Y. and B.F.; Resources, X.W., D.Y. and B.F.; Data curation, H.M.; Writing—original draft, H.M.; Writing—review and editing, F.W.; Visualization, P.Y. and F.W.; Project administration, P.Y. and X.W.; Funding acquisition, P.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Wang, T.; Guan, L.; Zhang, X.; Zhang, Y. A new method for power system stability assessment based on extended k-Nearest neighbor classifier. *Autom. Electr. Power Syst.* **2008**, *32*, 18–21, 75.
2. Zhang, Y.; Kou, L.; Sheng, W.; Wang, J.; Liang, Y.; Song, Q. Big Data Analytical Method for Operating State Assessment of Distribution Transformer. *Power Syst. Technol.* **2016**, *40*, 768–773.

3. Cai, D.; Wang, W.; Ma, X.; Xu, M.; He, Z.; Tang, Z.; Zhou, C.; Han, N.; Wang, Y. Analysis of Heavy Load and Overload Distribution Transformer in Regional Power Grid. In Proceedings of the 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China, 20–22 October 2018; pp. 1–5.
4. Zhu, E.; Liu, X. Construction and application of electric energy information acquisition system. In Proceedings of the 2011 IEEE 3rd International Conference on Communication Software and Networks, Xi'an, China, 27–29 May 2011; pp. 114–116.
5. Ma, H.; He, C.; Wang, L.; Yang, P.; Shen, H.; Tao, P. Load assessment of public transformer based on power consumption information acquisition and big data. *Electr. Meas. Instrum.* **2020**, *57*, 99–105.
6. Qian, K.; Wang, X.; Yuan, Y. Research on Regional Short-Term Power Load Forecasting Model and Case Analysis. *Processes* **2021**, *9*, 1617. [\[CrossRef\]](#)
7. Fallah, S.N.; Ganjkhani, M.; Shamshirband, S.; Chau, K.-W. Computational intelligence on short-term load forecasting: A methodological overview. *Energies* **2019**, *12*, 393. [\[CrossRef\]](#)
8. Li, M.; Zhou, Q. Distribution transformer mid-term heavy load and overload pre-warning based on logistic regression. In Proceedings of the 2015 IEEE Eindhoven PowerTech, Eindhoven, The Netherlands, 29 June–2 July 2015; pp. 1–5.
9. Torkzadeh, R.; Mirzaei, A.; Mirjalili, M.M.; Anaraki, A.S.; Sehhati, M.R.; Behdad, F. Medium term load forecasting in distribution systems based on multi linear regression & principal component analysis: A novel approach. In Proceedings of the 19th Electrical Power Distribution Conference (EPDC2014), Tehran, Iran, 6–7 May 2014; pp. 66–70.
10. Kwac, J.; Flora, J.; Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430. [\[CrossRef\]](#)
11. Albert, A.; Rajagopal, R. Smart meter driven segmentation: what your consumption says about you. *IEEE Trans. Power Syst.* **2013**, *28*, 4019–4030. [\[CrossRef\]](#)
12. Pei, S.; Qin, H.; Yao, L.; Liu, Y.; Wang, C.; Zhou, J. Multi-Step Ahead Short-Term Load Forecasting Using Hybrid Feature Selection and Improved Long Short-Term Memory Network. *Energies* **2020**, *13*, 4121. [\[CrossRef\]](#)
13. Chen, S.; Qin, J.; Sheng, W.; Fang, H. Study on Short-Term Forecasting of Distribution Transformer Load Using Wavelet and Clustering Method. *Power Syst. Technol.* **2016**, *40*, 521–526.
14. Shi, C.K.; Yan, W.; Zhang, X.; Zhang, B.; Fan, Y.; Tang, W. Heavy overload forecasting of distribution transformer during the spring festival based on BP network and grey model. *J. Electr. Power Sci. Technol.* **2016**, *31*, 140–145.
15. He, J.; Wang, H.; Ji, Z.; Meng, X.; Zhang, T. Analysis of factors affecting distribution transformer overload in smart grid. *Power Syst. Technol.* **2017**, *41*, 279–284.
16. He, J.; Wang, H.; Ji, Z. Heavy overload forecasting of distribution transformers based on random forest theory. *Power Syst. Technol.* **2017**, *41*, 2593–2597.
17. Zhang, G.; Wang, X.; Deng, C. Heavy overload prediction method for distribution network based on association analysis and machine learning. *Big Data* **2018**, *11*, 106–116.
18. Zarai, R.; Kachout, M.; Hazber, M.A.G.; Mahdi, M.A. Recurrent Neural Networks & Deep Neural Networks Based on Intrusion Detection System. *Open Access Libr. J.* **2020**, *7*, e6151. [\[CrossRef\]](#)
19. Wang, Y.; Zhang, N.; Chen, X. A Short-Term Residential Load Forecasting Model Based on LSTM Recurrent Neural Network Considering Weather Features. *Energies* **2021**, *14*, 2737. [\[CrossRef\]](#)
20. Bashir, T.; Haoyong, C.; Tahir, M.F.; Liqiang, Z. Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN. *Energy Rep.* **2022**, *8*, 1678–1686. [\[CrossRef\]](#)
21. Zhang, D.; Tong, H.; Li, F.; Xiang, L.; Ding, X. An Ultra-Short-Term Electrical Load Forecasting Method Based on Temperature-Factor-Weight and LSTM Model. *Energies* **2020**, *13*, 4875. [\[CrossRef\]](#)
22. Stratigakos, A.; Bachoumis, A.; Vita, V.; Zafiropoulos, E. Short-Term Net Load Forecasting with Singular Spectrum Analysis and LSTM Neural Networks. *Energies* **2021**, *14*, 4107. [\[CrossRef\]](#)
23. Ke, G.; Xu, Z.; Zhang, J.; Bian, J.; Liu, T.-Y. DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 3–7 August 2019.
24. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting System. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: Singapore, 2016; pp. 785–794.
25. Li, L.; Situ, R.; Gao, J.; Yang, Z.; Liu, W. A hybrid model combining convolutional neural network with XGBoost for predicting social media popularity. In Proceedings of the 25th ACM International Conference on Multimedia (MM'17), New York, NY, USA, 23–27 October 2017; ACM: Singapore, 2017; pp. 1912–1917.
26. Pan, B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *113*, 012127. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.