Leveraging Multi-task Learning for Energy Consumption Prediction in IoT-Based Data Center

Gaoxiang Jiang*, Yu Sun*, Bo Cheng*, Yanyi Wang*, and Haibo Zhou*
*School of Electronic Science and Engineering, Nanjing University, Nanjing, China, 210023
Emails: jgxwww@sina.cn, yusun@smail.nju.edu.cn, bocheng@smail.nju.edu.cn, wyynju@foxmail.com, haibozhou@nju.edu.cn.

2023 IEEE 23rd International Conference on Communication Technology (ICCT) | 979-8-3503-2595-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCT59356.2023.10419493

***Abstract*—In the context of the rapidly developing Internet era with advancements in software and hardware, the number of data centers, which serve as crucial infrastructures for data computation, processing, and storage, has grown at an astonishing rate. Along with this growth comes significant energy consumption, resulting in not only high electricity costs but also posing significant challenges to environmental protection. Various methods for optimizing data center energy efficiency heavily rely on accurate modeling of data center energy consumption. Currently, the widespread use of wireless Internet of Things (IoT) has enabled researchers to obtain large amounts of highly accurate data related to data center energy consumption. There have been research achievements in using machine learning methods to predict data center energy consumption. However, these studies primarily employ black-box models, which lack interpretability and leave room for further improvement in accuracy. In this paper, we propose a model-wise multi-task Transformer network, ensuring both model accuracy and interpretability. We utilize this approach to predict and model the cooling system, the largest contributor to energy consumption in data centers, using real measurement data from IoT devices. Then, this study further conducts importance analysis to reduce the dimensionality of model input features and eliminate redundant features, which has contributed to saving computational resources, reducing time costs of the prediction model and minimizing the number of IoT sensors required with a satisfactory performance.**

***Index Terms*—Internet of Things, data center, energy consumption, neural networks, multi-task learning**

I. Introduction

Data centers—continuously operating computing infrastructures for large-scale critical tasks—drive the rapid growth of the IT industry and even change the socio-economic system [1]. Demand for various applications has increased in the Internet era of rapid software and hardware advancements. Thus, cloud services, telecommunications operators, and banks have increased their data computation, processing, and storage needs. This trend has accelerated the construction of massive data centers [2]. Data centers consume the most energy globally, rising from 3 percent in 2017 to 4.5 percent in 2025. Meanwhile, their energy costs are doubling every five years [1]. Second, data centers' energy use harms the environment: natural gas and coal power electricity generation. Data centers' high energy use causes significant carbon emissions. Data centers emit 0.3±% of global carbon emissions, and this trend is expected to continue over the next decade [3]. Therefore, companies, countries and the global community must monitor and evaluate data center energy consumption, model and optimize energy consumption scenarios to save energy and reduce environmental stress.

Data center predictive models are being developed using machine learning or deep learning in many research studies. Google tested deep learning-based energy models using ANNs. They predicted data center PUE with an "expert system" using machine learning [4]. Hosoz et al. used an ANN model to predict cooling tower performance parameters like heat rejection, water vapor evaporation, and exhaust airflow temperature [5]. Yang and colleagues suggested detailed machine learning methods. Light gradient boosting machines and random forests predicted data center PUE accurately [6]. Using multivariate linear regression, Smpokos, and his team modeled energy consumption and critical weather parameters. Weather-based energy consumption predictions were compelling [7]. The team then suggested using Gated Recurrent Units (GRU), A temporal neural network model, to account for data center temperature and humidity time-series characteristics, and they created a neural network model to predict PUE values [8]. However, these studies primarily employ black-box models, which entail a deficiency in interpretability and present an opportunity for additional enhancements in accuracy. Also, with the rapid development of Internet of Things (IoT) technology, the method of using wireless sensor networks for environmental monitoring in data centers has become common to enhance data center energy efficiency and ensure operational safety [9]. The increased measurement accuracy and reduced cost of wireless sensor nodes, coupled with the increased deployment quantity, render the utilization of time-series data collected by these Internet of Things (IoT) sensors for training neural network models more precise and reliable in modeling data center energy consumption [10].

This study aims to establish a prediction model with high accuracy and interpretability for data center cooling systems using the data collected by the IoT sensor networks. Our study introduces a model-wise deep learning network based on multi-task learning. The study also employs importance analysis to select input features most relevant to energy, reducing the dimensionality of the model. In comparison to previous works, this research demonstrates novelty and contributions in the following aspects:

- We develop a prediction model for energy consumption in data center cooling systems by leveraging a model-wise Transformer network based on the real data col-

lected by the IoT sensors, integrated with a multi-task learning approach, thereby achieving enhanced accuracy and interpretability.

- To improve the interpretability of the established model and reduce input feature dimensions, we employ importance analysis techniques. This not only enhances the model's interpretability but also saves computational resources of the prediction model and reduces the required number of IoT sensors.
- Extensive simulations are conducted to validate the proposed prediction model. The results indicate that this model exhibits higher accuracy compared to traditional approaches in predicting data center cooling systems.

The rest of this paper is organized as follows. Section II introduces a system model of the data center based on IoT sensor network. Section III presents the algorithm for energy consumption prediction in data center cooling systems. The experimental setup and performance evaluation of the established model are presented in Section IV. Finally, Section V concludes this paper.

II. System Model

Data centers use energy for IT equipment, air conditioning and cooling, power distribution, and auxiliary lighting. Cooling and IT systems consume the most energy. IT equipment performance depends on cooling equipment power consumption and operation [11]. Thus, data center energy optimization requires cooling system energy optimization.

In this study, the IoT sensor network-based data center system model is shown as Fig. 1. The cooling system of the data center adopts the architecture of a centralized water-cooled chilled water air conditioning system. The data center cooling system model is shown as the lower section of Fig. 1. As shown in the upper section of Fig. 1, IoT sensor nodes will be deployed at various predetermined locations, including computer racks and different components of the cooling system. Through wireless or wired connections, these sensor nodes will collect measured data, aggregate it, and upload it to the IoT cloud. Applications running in the cloud will store the data in a database, which will be utilized for training and predicting data center energy consumption forecasting models [12]. In this framework, the main energy-consuming components include chiller units, chilled water pumps, cooling water pumps, and cooling towers [13]. Therefore, the model relationship between the total energy consumption and the energy consumption of each device can be determined as follows:

$$p_{cool} = p_{CRAC} + p_{CH} + p_{CWP} + p_{CHP} + p_{CT} + p_{others} + p_{loss}, \quad (1)$$

where p_{CRAC} represents the power consumption of the precision air conditioning unit at the end point, p_{CH} represents the power consumption of the chiller unit, p_{CWP} represents the power consumption of the chilled water pump, p_{CHP} represents the power consumption of the cooling water pump, and p_{CT} represents the power consumption of the cooling tower, p_{others} represents the energy consumption of other devices in the cooling system, p_{loss} represents the energy loss due to reasons such as power transmission losses and equipment heat dissipation in the cooling system.

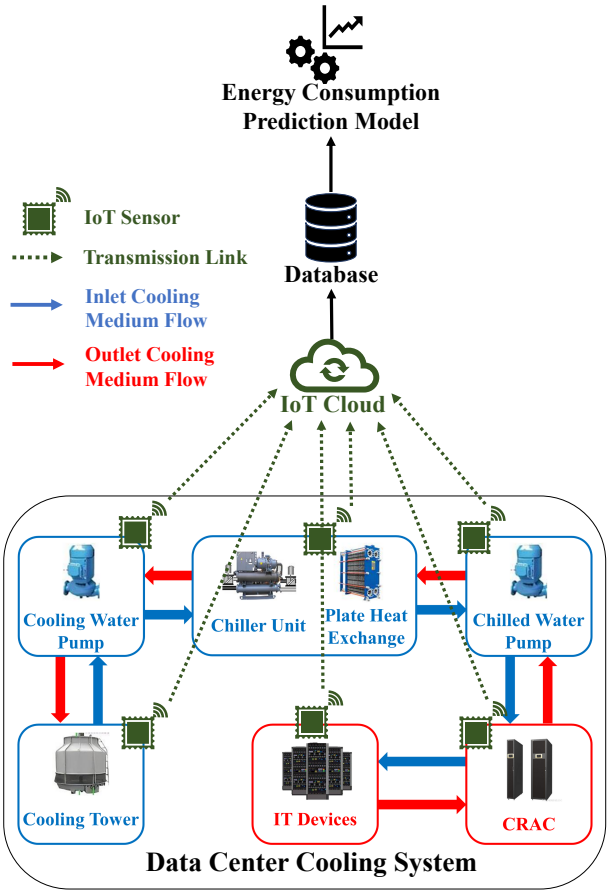

Fig. 1. The IoT enabled monitoring system for data center.

III. Model-Wise Multi-Task Transformer Network for Energy Consumption Prediction

In this section, we first introduce the dataset and the data preprocessing methods used in this study, followed by the description of the model-wise multi-task Transformer network model for data center cooling systems that we designed to ensure high prediction accuracy.

A. Dataset

The raw data for this study was collected from a data center located in Jinan, Shandong Province, China, operated by China Unicom. Since this study primarily focuses on modeling and predicting data center energy consumption of the cooling system, our main attention was on the data related to the chiller plant, computer room air conditioner(CRAC), computer rooms, and some other environmental parameters.

The data was collected in real-time using IoT devices and uploaded to a database for storage and real-time updates through the data center's infrastructure management system.

The data covers the period from May 31, 2022, to December 8, 2022, during which the data center was in normal operation. After initial data processing, the dataset contains over 700 data measurement points.

B. Data Cleaning and Prepocessing

1) Data cleaning: Due to the varying sampling intervals of different IoT devices and to ensure an ample amount of data for subsequent modeling, a standard time interval of 1 minute was adopted. For features with intervals exceeding 1 minute, linear interpolation was applied. For features with intervals less than 1 minute, a smoothing average sampling was performed using a 1-minute time window. Regarding the handling of outliers, a series of exceptional values were initially removed based on expert opinions and related work. Subsequently, the box plot method combined with Tukey's test was employed.

2) Data Prepocessing: The data preprocessing steps in this study include data integration, data reduction, and data transformation. In the data integration phase, variables with substantially the same meaning but different names were merged. Next, data reduction was performed by removing variables that were deemed irrelevant to the prediction of data center cooling system energy consumption based on expert knowledge and analysis.

Lastly, to scale data with different dimensions and scales into a consistent range, this study employed the widely used Z-score normalization method from deep learning. The formula for Z-score normalization is as follows:

$$Z = \frac{X - \mu}{\sigma}, \tag{2}$$

where: Z is the standardized value, X represents the original value of the variable, μ is the mean of the variable, σ denotes the standard deviation of the variable.

After applying the aforementioned data processing steps, the original dataset, which initially had a size of 240,000 rows and 700 columns, has been transformed into a usable dataset with dimensions of 188,747 rows and 190 columns. Since the subsequent network model utilizes time series as inputs, the dataset needs to be processed using a sliding window approach. The sliding window is set to a time interval of 8 units. Therefore, the size of the dataset we used is $240,000 \times 700 \times 8$. This processed dataset will be utilized for the subsequent development of prediction models.

C. Multi-Task Transformer Network Model

1) Design and implementation of the model structure: To forecast energy consumption in data center cooling systems, we developed a multi-layered time series regression prediction model based on multi-task learning. This model comprises two layers: subnetwork modules for predicting energy consumption of individual devices and a merging network module for predicting the total energy consumption of the cooling system.

For the subnetwork modules responsible for predicting the energy consumption of each device, we consider combining the MLP (Multilayer Perceptron) network structure, with the Transformer structure, which is currently widely employed in deep learning networks [14]. Specifically, we utilize the Encoder structure of the Transformer to extract parameter features [15], and then replace the Decoder layer of the Transformer with fully connected layers from the MLP network. This modification enables the final feature extraction and result fitting. We integrate merging network module, so that the subnetwork modules of multiple devices are consolidated into a unified network, ensuring the alignment of the network structure and functionality with the actual physical system. The merging network can be described as:

$$h_1 = \text{Concat}(x_1, x_2, ..., x_{n-1}), \tag{3}$$

$$h_2 = f_{MLP}(x_1, x_2, ..., x_{n-1}, x_n), \tag{4}$$

$$Y = \text{Concat}(h_1, h_2), \tag{5}$$

where $x_1, x_2, ..., x_{n-1}$ represents all the input parameters of various cooling devices, including chiller units, cooling towers, chilled water pumps, and other similar cooling equipments. x_n represents the environmental parameters as input parameters. f_{MLP} represents the function of the MLP networks. Y represents the final output of the merging network. The utilized MLP network here comprises only one layer, so f_{MLP} can be described as:

$$f_{MLP}(\mathbf{x}) = \text{ReLU}(\text{BN}(\mathbf{W}\mathbf{x} + \mathbf{b})), \tag{6}$$

where we utilize $\text{ReLU}(x)$ as the activation function and the batch normalization operation $\text{ReLU}(x)$ is used to expedite the training process and enhance the network's performance. $\mathbf{W}$ represents the weight matrix and $\mathbf{b}$ is the bias. The study is based on a multi-task learning approach to design the network structure based on physical model, which enhances the interpretability and stability of the network model. We denote this approach as MT-TEMLP. The MT-TEMLP network structure is illustrated as Fig. 2.

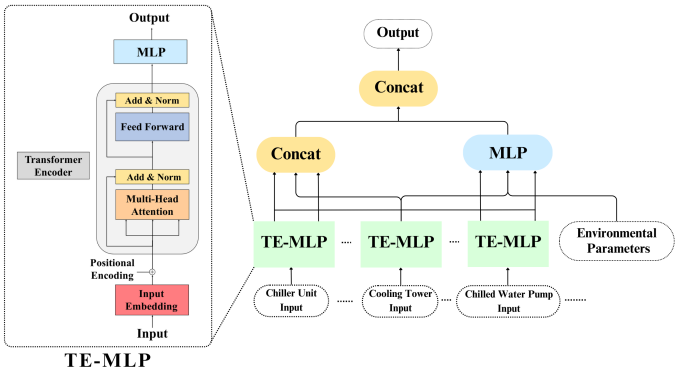

Fig. 2. The network structure of proposed MT-TEMLP.

2) Model training: This study employed a multi-task learning approach for model training. By modifying the loss function, it was set to be a function that is related to both the predictions of individual device energy consumption and the overall energy consumption, which enables the neural

network to simultaneously predict for both device-level energy consumption and total energy consumption [16]. This training method allows the network model to closely align with the network model in real systems, and the loss function was set as follows:

$$L = \sum_i w_i(t) * L_i, \tag{7}$$

where, L represents the loss function for the total task used during model training, L_i denotes the loss function for individual tasks, $w_i(t)$ represents the weight parameters associated with these tasks, and t signifies the training iteration. In this study, we employed a gradient normalization method to dynamically adjust the weight parameters of different tasks, thereby making these weight parameters functions that change with the training iteration t.

Gradient normalization is a crucial method in multi-task learning, aiming to address the issue of disparate convergence rates among different tasks. This method incorporates dynamic adjustment of the learning weights for different tasks based on the changes in their respective loss functions after each learning iteration [17]. The implementation approach and related formulas are presented as follows:

$$G_W^i(t) = \|\nabla_W w_i(t) L_i(t)\|_2, \tag{8}$$

$$\widetilde{L}_i(t) = \frac{L_i(t)}{L_i(0)}, \tag{9}$$

$$r_i(t) = \frac{\widetilde{L}_i(t)}{E_{task}[\widetilde{L}_i(t)]}, \tag{10}$$

$$L_{grad}(t; w_i(t)) = \sum_i |G_W^i(t) - E_{task}[G_W^i(t)] * [r_i(t)]^\alpha|_1, \tag{11}$$

$$w_i(t+1) = w_i(t) + \lambda * \nabla(L_{grad}(t; w_i(t))), \tag{12}$$

where, $G_W^i(t)$ represents the value of gradient normalization for task i, which measures the magnitude of the loss for a specific task. $L_i(0)$ and $L_i(t)$ respectively represent the loss function for task i at step 0 and step t. $\widetilde{L}_i(t)$ is the absolute training speed for task i, while $r_i(t)$ represents the relative training speed between tasks. $L_{grad}(t; w_i(t))$ is a cost function that depends on the weights assigned to different tasks. Ultimately, the cost function, $L_{grad}(t; w_i(t))$, which accounts for the adjustment of task weights, is used for gradient descent to dynamically update the task weights in multi-task learning.

D. Importance Analysis Algorithm

Importance analysis is a crucial method for analyzing and assessing the impact of input parameters on a model. Different models may employ various types of importance analysis methods. In this study, we primarily focused on feature importance based on random permutation, and the pseudocode for this algorithm is outlined in Algorithm I. The algorithm

Algorithm 1 Importance analysis

1: Normalization. Each input data feature X_{ij} and each label data Y_i:

$$X'_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$
$$Y'_i = \frac{Y_i - \mu}{\sigma}$$

where $\mu_j = \sum_i X_{ij}$, $\sigma_j = \sqrt{\sum_i (X_{ij} - \mu_j)^2}$, $\mu = \sum_i Y_i$, $\sigma = \sqrt{\sum_i (Y_i - \mu_j)^2}$. The final dataset obtained: (X', Y')

2: Model training. Based on (X', Y'), a model f is trained. The loss function is g. The sampling quantity is k, and the number of sensitivity cycles is denoted as N. The entire process is repeated for m iterations.

3: **for** each X_j **do**

4: $a = 1$

5: **repeat**

6: The feature X_j is randomly shuffled to become S_j. This process generates a new dataset (S, Y').

7: Calculate the obtained feature importance M_{jn}:
$M_{jn} = L_{jn} - L = g(f(X', Y')) - g(f(S, Y'))$

8: $a = a + 1$

9: **until** $a > m$

10: Calculate the feature-specific difference M_j for the feature X_j:
$M_j = \sum_{n=1}^{N} M_{jn}$

11: **end for**

ultimately yields the output deviation M_j for each feature j. For a pre-established model, the larger the absolute value of the output deviation M_j for feature j, the greater the significance of this feature's impact on the output within the model [18]. Therefore, by employing this importance analysis algorithm and leveraging a predictive model, we can identify input parameters from the dataset that hold greater importance for modeling energy consumption in data centers.

IV. Performance Evaluation

In this section, we first conducted evaluation of the previously established model, comparing it with various models used in prior research. Next, we performed an analysis of the importance of input features based on the developed model. Based on the results of the importance analysis, we reduced the number of input parameters and compared its performance with the previous model.

A. Performance Comparison

In this network, the subnetwork modules of different devices automatically adapt their hyperparameters based on the varying effective input dimensions of each device. Assuming the input feature size of a particular subnetwork module is denoted by N, the corresponding hyperparameters for subnetwork modules are presented in the Table I.

In the dataset employed in this study, the input dimensions (N) for the chiller are 31, the cooling tower is 15, while the cooling water pump and the chilled water pump have

TABLE I
HYPERPARAMETERS OF THE TRANSFORMER ENCODER IN THE NETWORK

Hyperparameters	Values
d_model	N
num_layers	8
num_heads	4
dim_feedforward	$\lceil N/2 \rceil$
Dropout	0.1
Linear1_out	$2N$
Linear2_out	$\lceil N/2 \rceil$

dimensions of 4 each. For the subsequent Dense Layer, the input features consist of the outputs from all preceding network modules, in addition environmental parameters. The hyperparameters for this layer are configured in Table II. The

TABLE II
HYPERPARAMETERS OF THE DENSE LAYER IN THE NETWORK

Hyperparameters	Values
Dropout	0.1
Linear1_in	106
Linear1_out	250
Linear2_out	1

deep learning network employed in this study utilized the Adam optimizer, with the mean squared error (MSE) chosen as the model loss function, which is commonly used for regression problems. The learning rate was set to 3e-4. The dataset was divided using 5-fold cross-validation, and finally the model with the best performance was selected. With the aforementioned settings, the comparison plot between the MT-TEMLP model designed in this study and the actual real-world data results is presented as Fig. 3. The x-axis of the graph

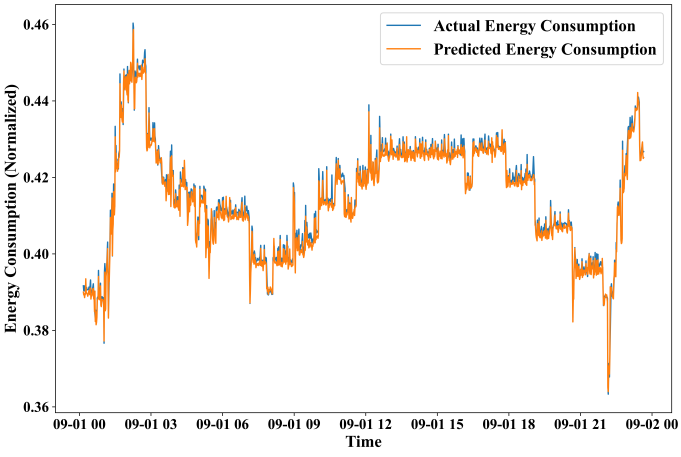

Fig. 3. Comparison diagram of predicted and actual energy consumption of data center cooling systems.

depicts time, focusing on data extracted from a particular day within the dataset (September 1, 2022). On the y-axis, we find the normalized energy consumption values of the cooling system within the data center. It becomes evident upon examination that the MT-TEMLP model exhibits a notable prowess in effectively forecasting the energy consumption of the data center's cooling system, thereby substantiating its capacity for accurate prediction. We focused on evaluating the regression prediction models using three commonly used metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 score(R^2). The formulas for these three metrics are expressed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

Additionally, other methods were also explored and compared in this study, such as decision trees, polynomial regression, support vector machines (SVM), recurrent neural networks (RNN), and convolutional neural networks (CNN). The comparative results are listed in Table III. Based on the experiment

TABLE III
PERFORMANCE COMPARISONS OF DIFFERENT MODELS

Models	MAE	MSE	R^2
	Values	Values	Values
MT-TEMLP	**0.00408**	**2.122e-5**	**0.964**
GRU	0.00783	5.95e-5	0.925
MLP	0.00844	6.182e-5	0.928
ResNet	0.0208	9.873e-5	0.833
Random Forest	0.0144	8.277e-5	0.913
SVM	0.0402	6.194e-4	0.826
Polynomial regression	0.0551	0.00122	0.798

results of different models, it can be concluded that the innovative MT-TEMLP model proposed in this study outperforms general deep learning models such as GRU and MLP, which are commonly used for regression prediction problems. Furthermore, the MT-TEMLP model offers improved interpretability to users by incorporating the structural elements of a physical model, distinguishing it as an advantage over other models.

B. Results of Importance Analysis

This study conducted an analysis of the MT-TEMLP model using the previously mentioned importance analysis methods. The features were sorted based on the output deviation M_j , and the top 10 features in terms of importance ranking are shown in Fig. 4. Considering expert knowledge and thermodynamics principles, these features are closely associated with the energy consumption of various components within the cooling system, which reinforces the accuracy of the prediction

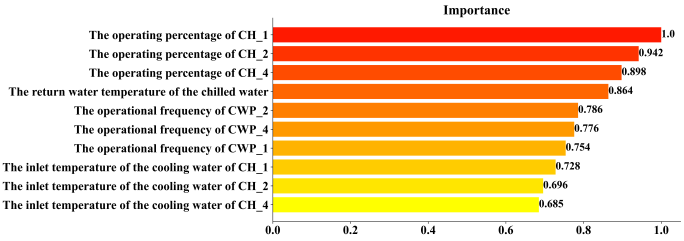

Fig. 4. Illustration of the top 10 features in importance analysis.

model established in this study. Subsequently, a process of re-modeling was undertaken through the application of the MT-TEMLP technique, where the selection encompassed the foremost 45 input parameters, as identified and prioritized through meticulous importance analysis. A comparison was made between the model predictions using the 45 selected input parameters and the model predictions using the initial 190 input parameters. The error comparison is depicted in the figure as Fig. 5. As a result of the reduction in the quantity

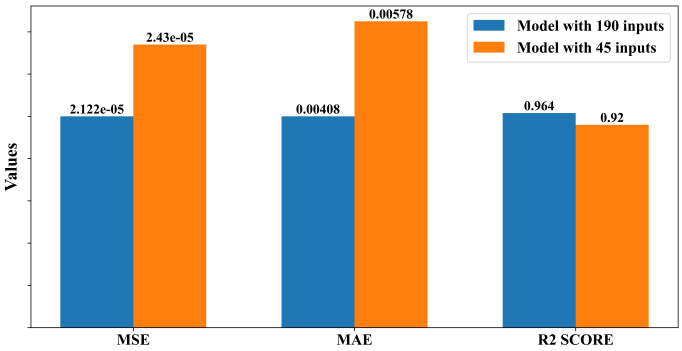

Fig. 5. Comparison between model with 45 inputs and model with 190 inputs.

of input parameters, there is an inherent likelihood of a moderate decrease in the model's precision. Nevertheless, when contrasted with alternative algorithms, this model continues to demonstrate a notably elevated degree of accuracy. Within a permissible margin of performance reduction, it becomes plausible to efficiently curtail the input parameters of the model, thereby streamlining the intricacy of the model and expediting the optimal implementation of IoT sensors. Taking into account the importance of both model accuracy and model size, this study concludes that the selection of different input parameters based on importance analysis results holds certain significance.

V. Conclusion

In this paper, we have investigated a model-wise multi-task Transformer model for energy consumption modeling prediction in data center cooling systems. Combining the physical model of data center energy consumption, we have proposed a novel MT-TEMLP neural network model, which has higher accuracy and interpretability. We conduct tests on a real dataset collected by IoT sensor networks in data center from China Unicom and compare the results with other methods. The results demonstrate that the proposed model achieves superior performance in terms of MSE, MAE, and R2 score, with values of 2.122e-05, 0.00408, and 0.964, respectively, outperforming other models in this scenario. Furthermore, based on this model, we have performed an importance analysis to decrease the model complexity and optimize deployment of IoT sensors while maintaining accuracy. In future work, we aim to explore intelligent and automated optimization of data center energy efficiency based on this model.

References

[1] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2015.

[2] Z. Cao, X. Zhou, H. Hu, Z. Wang, and Y. Wen, "Towards a systematic survey for carbon neutral data centers," *IEEE Communications Surveys & Tutorials*, 2022.

[3] N. Jones *et al.*, "The information factories," *Nature*, vol. 561, no. 7722, pp. 163–6, 2018.

[4] J. Gao, "Machine learning applications for data center optimization," 2014.

[5] M. Hosoz, H. M. Ertunç, and H. Bulgurcu, "Performance prediction of a cooling tower using artificial neural network," *Energy Conversion and Management*, vol. 48, no. 4, pp. 1349–1359, 2007.

[6] Z. Yang, J. Du, Y. Lin, Z. Du, L. Xia, Q. Zhao, and X. Guan, "Increasing the energy efficiency of a data center based on machine learning," *Journal of Industrial Ecology*, vol. 26, no. 1, pp. 323–335, 2022.

[7] G. Smpokos, M. A. Elshatshat, A. Lioumpas, and I. Iliopoulos, "On the energy consumption forecasting of data centers based on weather conditions: Remote sensing and machine learning approach," in *2018 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*. IEEE, 2018, pp. 1–6.

[8] P. Zhao, L. Yang, Z. Kang, and J. Lin, "On predicting the pue with gated recurrent unit in data centers," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*. IEEE, 2019, pp. 1664–1670.

[9] Y. Xu, H. Zhou, J. Chen, T. Ma, and S. Shen, "Cybertwin assisted wireless asynchronous federated learning mechanism for edge computing," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[10] A. Medina-Santiago, A. D. P. Azucena, J. M. Gómez-Zea, J. A. Jesús-Magaña, M. de la Luz Valdez-Ramos, E. Sosa-Silva, and F. Falcon-Perez, "Adaptive model iot for monitoring in data centers," *IEEE Access*, vol. 8, pp. 5622–5634, 2019.

[11] A. Thakkar, K. Chaudhari, and M. Shah, "A comprehensive survey on energy-efficient power management techniques," *Procedia Computer Science*, vol. 167, pp. 1189–1199, 2020.

[12] B. Qian, H. Zhou, T. Ma, K. Yu, Q. Yu, and X. S. Shen, "Heterogeneous multi-operator spectrum sharing architecture for massive IoT access with NOMA," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 2020, pp. 1–6.

[13] Q. Zhang, Z. Meng, X. Hong, Y. Zhan, J. Liu, J. Dong, T. Bai, J. Niu, and M. J. Deen, "A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization," *Journal of Systems Architecture*, vol. 119, p. 102253, 2021.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] R. Mohammadi Farsani and E. Pazouki, "A transformer self-attention model for time series forecasting," *Journal of Electrical and Computer Engineering Innovations (JECEI)*, vol. 9, no. 1, pp. 1–10, 2020.

[16] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[17] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International conference on machine learning*. PMLR, 2018, pp. 794–803.

[18] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.