Article

Research on Joint Forecasting Technology of Cold, Heat, and Electricity Loads Based on Multi-Task Learning

Ruicong Han [1], He Jiang [1,*] , Mofan Wei [2] and Rui Guo [1]

1   School of Renewable Energy, Shenyang Institute of Engineering, Shenyang 110136, China; hrc916393409@163.com (R.H.); gr089@sina.com (R.G.)
2   School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870, China; mofan_scholar@163.com
*   Correspondence: jianghescholar@163.com

Abstract: The cooperative optimization and dispatch operation of the integrated energy system (IES) depends on accurate load forecasts. A multivariate load, joint prediction model, based on the combination of multi-task learning (MTL) and dynamic time warping (DTW), is proposed to address the issue of the prediction model's limited accuracy caused by the fragmentation of the multivariate load coupling relationship and the absence of future time series information. Firstly, the MTL model, based on the bidirectional long short-term memory (BiLSTM) neural network, extracts the coupling information among the multivariate loads and performs the preliminary prediction; secondly, the DTW algorithm clusters and splices the load data that are similar to the target value as the input features of the model; finally, the BiLSTM-attention model is used for secondary prediction, and the improved Bayesian optimization algorithm is applied for adaptive selection of optimal hyperparameters. Based on the game-theoretic view of Shapley's additive interpretation (SHAP), a model interpretation technique is introduced to determine the validity of the liquidity indicator and the asynchronous relationship between the significance of the indicator and its actual contribution. The prediction results show that the joint prediction model proposed in this paper has higher training speed and prediction accuracy than the traditional single-load prediction model.

Keywords: multi-task learning; load prediction; DTW; BiLSTM; attention mechanism

1. Introduction

Energy transition is a critical trend in the development of the energy industry, aiming primarily to enhance energy utilization efficiency [1]. Traditional energy systems, such as electricity, natural gas, and heating systems, tend to operate independently and are managed by different energy companies, which results in insufficiently tight coupling between energy systems [2]. As an important aspect of integrated energy system (IES) demand-side energy forecasting, IES load forecasting has become a primary prerequisite for the collaborative IES operation, dispatch, and planning [3]. Accurate load forecasting can better control the balance between production and demand and increase energy use efficiency.

Extensive studies have been carried out in the field, developing various statistical methods [4]. Before the advent of machine learning, linear regression techniques were the dominant approach in forecasting [5]. The autoregressive integrated moving average (ARIMA) method is particularly notable as an advanced autoregressive technique [6]. However, a drawback of linear regression approaches is that they tend to ignore extrinsic elements, like temperature and humidity, in favor of concentrating exclusively on linear connections within load data. A nonlinear function-to-function model proposed in the literature [7], which captures complex time series patterns through a multilayer neural network structure, shows higher prediction accuracy in practical applications. Power

demand forecasting now uses machine learning approaches thanks to the development of artificial intelligence. Techniques such as random forest (RF) [8], artificial neural networks (ANNs) [9–12], and support vector machines (SVMs) [13–16] have become increasingly popular. In SVM applications, selecting parameters related to the hyperplane and kernel function is crucial. It can be difficult to reduce the problem to a quadratic programming (QP) problem since so many variables are involved, particularly when working with large amounts of data. The advent of hybrid forecasting models, which provide more optimal answers, results from deep learning's growth [17]. Recurrent neural networks (RNNs), specifically, the long short-term memory (LSTM) variant, show significant potential in power load forecasting. LSTM, which inherits RNNs' capability for long-term memory, introduces a "forget gate" to address issues related to long-term dependencies. Currently, numerous studies are focused on enhancing LSTM networks. For instance, a combined multivariate linear regression and LSTM approach has been suggested [18], showing promise on huge datasets spanning western China, Uzbekistan, and portions of the United States. It was confirmed that LSTM could reliably extract load characteristics more accurately than support vector regression (SVR) using the power load of Estonia as a case study [19].

With the rapid progress of big data and artificial intelligence technologies, the input features of forecasting models have become increasingly diversified, multidimensional, and complex. In practical applications, multivariate time series data often contain a large number of redundant and irrelevant feature variables, which may mask key feature information and negatively affect the accuracy of time series forecasting models. Therefore, there is an urgent need for a good feature extraction method to filter the input features. Ref. [20] integrated the dynamic time warping (DTW) algorithm with the LSTM model to address data sparsity issues in short-term load forecasting, demonstrating DTW's effectiveness in compensating for missing future data. The significance of the model's capacity for generalization and training speed is rapidly becoming more apparent as a result of the ongoing advancements in deep learning and machine learning technologies, and the convergence process of the model can be effectively accelerated by introducing suitable optimization algorithms, avoiding falling into local optimal solutions, and enhancing the model's generalization ability. The prediction accuracy significantly improved in Ref. [21]'s study based on Bayesian optimization algorithms to improve the traditional LSTM, convolutional neural networks (CNNs), least squares support vector machine, and additional models. At the same time, compared with other models, the LSTM and Bayesian optimization algorithms have better fitness. Ref. [22] proposed a two-stage adaptive superposition prediction of multivariate loads, significantly improving the model's generative memorization capability and stability through the parallel approach of initial prediction and prediction correction.

Accompanied by the accelerated evolution of integrated energy systems, single-load forecasting has increasingly given way to multivariate load forecasting in the study of load forecasting. At present, multivariate load forecasting mainly focuses on deterministic forecasting, which can be categorized into two types according to the number of model research objects. The first category is the separate prediction of multivariate loads, which uses single-task-learning (STL) models to establish independent prediction models for multiple loads, like electricity, heating, and cooling, respectively, and the models are unrelated. Ref. [23] decomposed the data by aggregated modal decomposition and then used a temporal convolutional network (TCN) for prediction. Ref. [24] incorporated a dual attention mechanism into the Seq2Seq model, improving the algorithm's ability to learn temporal and input features. The methods above improve the prediction accuracy of STL models by introducing the attention mechanism or modal decomposition. However, since the STL models are independent, they cannot fully consider the coupling characteristics between multivariate loads. This independence restricts the model from capturing and synthesizing the interactions between loads, which may affect the overall forecasting effect.

The second method is called joint prediction of multiple loads, which may simultaneously produce prediction results for various loads, like heating, cooling, and electricity. It achieves this by predicting multiple loads as a whole using a multi-task learning model.

Ref. [25] proposed an explanatory load forecasting framework that combines multi-task learning (MTL) and LSTM modeling and constructs the input variables through a coupled feature extraction strategy, which makes it possible for the model to effectively capture the strong coupling interactions between the demands of heating, cooling, and electricity. A complex neural network (ComNN) was proposed in the literature of Ref. [26] for multi-task learning. It extracts the salient features of the shared layer using an attention mechanism that distinguishes between subtasks, and it uses the MTL-ComNN network structure for prediction. The methods above predict multivariate loads through MTL, which can fully consider the common data features among loads and exploit the coupling relationship among them, and their prediction accuracy is usually higher than that of STL models. However, there are differences in the sensitivity of meteorological factors to different loads, i.e., the same meteorological factor affects different loads to different degrees. When forecasting with the MTL model, using only the meteorological features strongly correlated with all the loads as inputs may lead to omitting critical meteorological information. In contrast, using meteorological features strongly correlated with any load as inputs may introduce too much noise at the input level. In addition, there are differences in the coupling characteristics among different loads [27], and it is difficult for the MTL model to adequately account for these differences when the coupling strength is weak. These factors limit the prediction accuracy of the MTL model [28]. A single MTL model cannot use future timing information in prediction to improve prediction accuracy. Future time-series information refers to multivariate loads other than the target load at the forecasting moment. For example, due to the dynamic coupling characteristics of multiple loads, the electric, cooling, and heating loads are coupled and related to each other at time t. Therefore, the cooling load and heat load at time t are used in the MTL model. Thus, using the cooling and heating loads at time t as input features for training, the model can be trained to predict the electrical loads at time t based on the values of these loads, thus improving the prediction accuracy. However, for a single MTL model, the multivariate load values at the time, which are the target of prediction, are unknown. Thus, it is impossible to utilize this time-series information to improve the prediction accuracy further.

A multivariate load short-term forecasting model that utilizes multi-task learning and the DTW algorithm is proposed to address the aforementioned issues. The extraction of multivariate load coupling information was considered, concentrating on the problem of limited prediction accuracy due to the inability to fully use the load information at future moments. To effectively utilize the future time-series information, the multi-task learning model based on bidirectional long short-term memory (BiLSTM) first extracts the coupling information between loads and performs a preliminary prediction; second, the DTW algorithm is used to identify the sequences that are most similar to the prediction target, and these similar sequences are added to the training set; third, BiLSTM is applied to establish the multi-task learning shared layer, fully exploiting the characteristics of coupling between loads of heat, electricity, and cold, and the attention mechanism is used to achieve the differential extraction of important features in the shared layer by sub-tasks to enhance the impact of key information; and finally, the established DTW-BiLSTM-MTL model is used to make predictions. Practical examples are employed to evaluate the effectiveness of the presented model.

2. Methodology

2.1. Introduction to the DTW Algorithm Principle

The DTW algorithm is designed to measure similarity between time series by non-linearly aligning sequences, allowing for a more accurate reflection of the intrinsic nature of the data, especially in the presence of temporal delays. Due to the cyclical characteristics expressed in time delays caused by cold, heat, and power load, such a matching method can mitigate the adverse impact of short-term dissimilarities between sequences, thereby capturing meaningful correlative information.

The following are the algorithm's precise steps to evaluate two time series of N length, X and Y. Firstly, an integer sequence path $p = (p_1, p_2, \cdots, p_L)$ needs to be defined, where $p_i = (n_i, m_i) \in [1:N] \times [1:N] (1 \leqslant i \leqslant L)$, and it must satisfy the following conditions:

(1) Monotonicity: Each point on the path must vary monotonically over time, hence n_i and m_i must satisfy $n_1 \leqslant n_i \leqslant n_L, m_1 \leqslant m_i \leqslant m_L$.

(2) Continuity: For the path, any point $p_i = (n_i, m_i)$ and the next point $p_{i+1} = (n_{i+1}, m_{i+1})$ must satisfy $n_{i+1} - n_i \leqslant 1$ and $m_{i+1} - m_i \leqslant 1$.

(3) Boundary Conditions: The boundary conditions start with $p_1 = (1, 1)$ and end with $p_L = (N, N)$.

Under these conditions, the optimization goal of the DTW algorithm is defined as follows:

$$DTW(X, Y) = \min_{\pi} \sum_{(i,j) \in \pi} |x_i - y_j|^2 \tag{1}$$

where π is the path of the match. The selection criteria for the DTW algorithm are given by the following:

$$\begin{cases} DTW(X, Y) = M(x_R, y_C) = dis(x_R, y_C) + \min \begin{cases} M(X_{R-1}, Y_C) \\ M(X_R, Y_{C-1}) \\ M(X_{R-1}, Y_{C-1}) \end{cases} \\ dis(x_1, y_1) = \sqrt{|x_1 - y_1|^2} \end{cases} \tag{2}$$

$M(x_i, y_j)$ is the value taken from the accumulated distance matrix at point (x_i, y_j). R represents the index position of sequence X; C represents the index position of sequence Y.

According to similarity, the relationship between source features and target features is established, forming feature pairs. Figure 1 is a schematic of DTW.
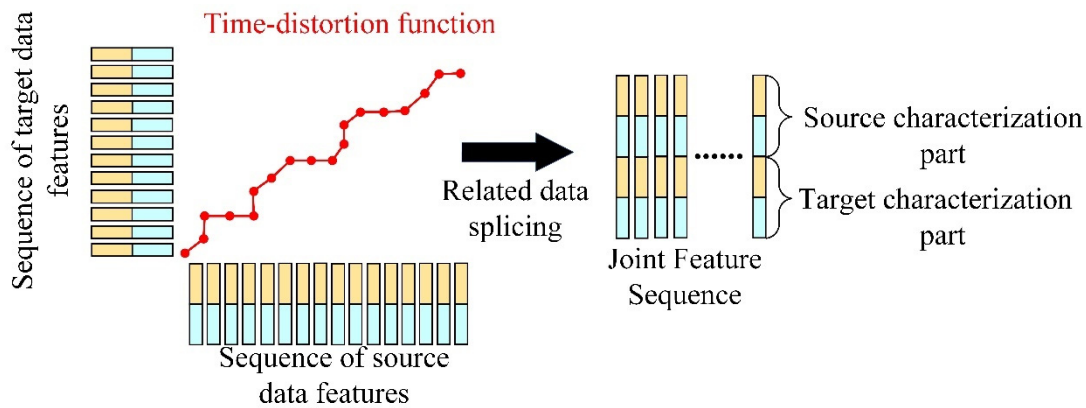

Figure 1. Schematic diagram of DTW.

2.2. Introduction to the BiLSTM Algorithmic Mechanism

LSTM is an improvement of recurrent neural networks (RNNs), aiming to solve the problem of gradient explosion or gradient disappearance of RNNs due to the large span and the influence of the activation function when dealing with long time series data. Figure 2a displays the architecture of the LSTM model. LSTM improves long-term dependent data handling by introducing three gating units—the input gate, forget gate, and output gate—based on RNNs and using sigmoid or tanh activation functions to control the information flow. The input gate is responsible for selectively recording new input information into the memory cell of the previous step; the forgetting gate is responsible for only forgetting and retaining the information in the memory cell state passed during the last step; and the output gate passes the updated cell state to the next time step, as elaborated in Equation (3).

$$\begin{cases} u_t = \sigma(W_{xh}^u x_t + W_{hh}^u h_{t-1}) \\ f_t = \sigma(W_{xh}^f x_t + W_{hh}^f h_{t-1}) \\ o_t = \sigma(W_{xh}^o x_t + W_{hh}^o h_{t-1}) \end{cases} \tag{3}$$
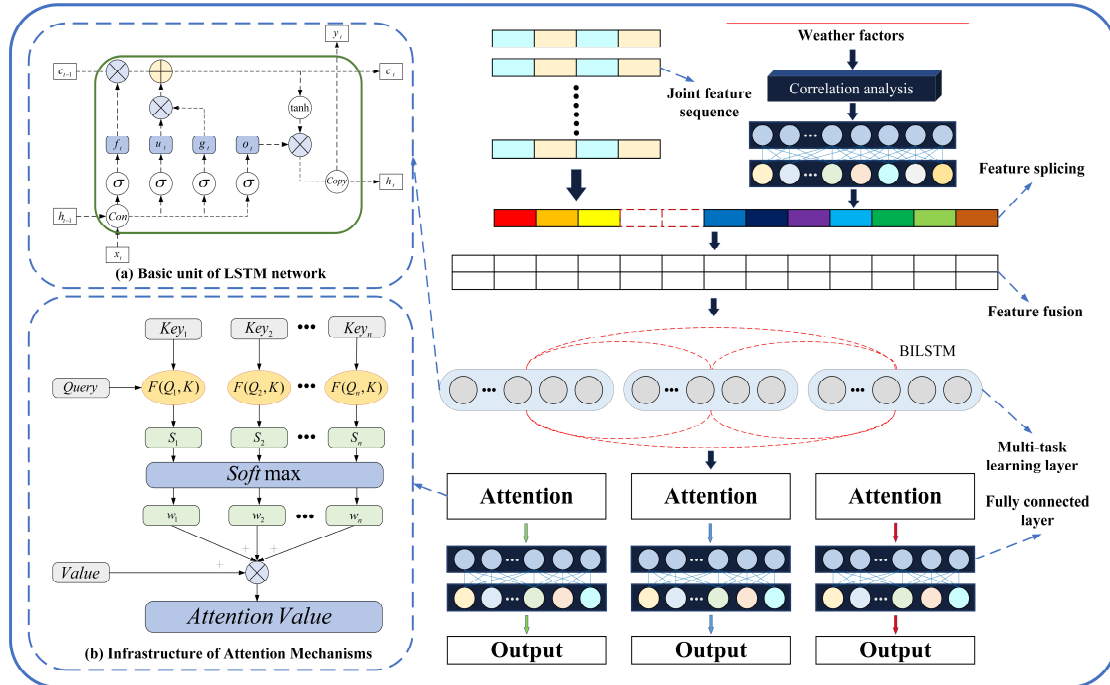

Figure 2. Structure of the proposed model.

Conversely, BiLSTM expands upon the traditional LSTM by utilizing two distinct LSTM networks at every time step: a forward-moving network and a backward-moving network [29]. This approach enables the model to incorporate both preceding and succeeding contexts within the sequence. Although each part functions like a typical LSTM, the input sequence is processed in reverse order by the backward LSTM, and the BiLSTM output is ascertained using the method outlined in Equation (4).

$$y^{(t)} = \sigma(W_{yf} y_f^{(t)} + W_{yb} y_b^{(t)} + b_y) \tag{4}$$

where u_t, f_t, and o_t represent the features of the forgetting, output, and input gates at time t, respectively, and the weights that need to be updated and learned are represented by W_{xh}^u, W_{hh}^u, W_{xh}^f, W_{hh}^f, W_{xh}^o, and W_{hh}^o. The output layer's activation function is indicated by σ, x_t represents the input at moment t, and h_{t-1} denotes the output from the prior time step. The weight matrices required to generate the outputs are W_{yf} and W_{yb}. The output bias matrix is denoted by $y^{(t)}$. The forward and backward LSTM outputs at time step t are represented by $y_f^{(t)}$ and $y_b^{(t)}$, respectively.

2.3. Improved Bayesian Optimization Algorithm

The Bayesian optimization algorithm works by minimizing or maximizing a black-box objective function. It is based on Bayes' theorem and probabilistic models, such as the Gaussian process (GP). It gradually converges to the global optimal solution by selecting sampling points sequentially in the search space and estimating the optimal value of the function based on the existing sampling data. The expression is shown in Equation (5).

$$h^* = \underset{h \in H}{\mathrm{argmax}} f(h) \tag{5}$$

In this context, $f(h)$ represents the prior distribution model, h^* represents the optimal parameter value under the constraint of $f(h)$, and H represents the candidate set. Compared to grid search and random search, the Bayesian optimization algorithm can achieve satisfactory optimization results with fewer iterations [30].

The Gaussian process is a probabilistic surrogate model that works well when optimizing continuous hyperparameters, like the learning rate, dropout rate, etc. However, it struggles with discrete hyperparameters, like batch size, epoch count, BiLSTM hidden units, and other discrete hyperparameters, like layer count, pooling layer count, and BiLSTM network layer count. It is statistically difficult to perform well in complicated network models if the activation function type, convolution kernel size, or pooling window size cannot be adjusted via the Bayesian optimization procedure.

The topic of this research is the optimization of discrete and continuous hyperparameters in neural network architectures. It makes use of the tree-structured Parzen estimator (TPE) model, which addresses the drawbacks of the conventional GP-based Bayesian optimization models and enables more effective optimization of the neural network architecture and hyperparameters.

The TPE algorithm uses a Gaussian mixture model (GMM) to model the search space. First, according to the Bayesian framework, the conditional probability distribution $P(y|x)$ is decomposed into the product of the likelihood $P(x|y)$ and the prior probability $P(y)$, where $P(y)$ is updated based on new observations. Therefore, the key lies in obtaining $P(x|y)$.

The TPE algorithm employs different strategies based on the search space, replacing the continuous uniform distribution with a truncated Gaussian mixture and the stepwise uniform distribution (discrete) with a truncated categorical mixture, where the weights are adjusted for each category. For different levels of observation $\{x_1, x_2, \cdots, x_n\}$, the algorithm uses different replacements and generates different densities in the configuration space. These two densities are then used to redefine the probability density function $P(x|y)$, obtained by the following Equation (6).

$$p(x \mid y) = \begin{cases} l(x), y < y^* \\ g(x), y > y^* \end{cases} \tag{6}$$

In the equation, $l(x)$ represents the density function formed by the observations $\{x_i\}$ when the evaluation value $y(x_i)$ is less than the threshold y^*, while $g(x)$ represents the density function formed by the observations $\{x_i\}$ when the evaluation value $y(x_i)$ is greater than the threshold y^*.

Therefore, the TPE algorithm constructs two different distributions from the observations, which can be regarded as a distribution of better hyperparameter values and a distribution of worse hyperparameter values.

The choice of the threshold value y^* is made by setting a parameter γ that satisfies the condition $p(y < y^*) = \gamma$. This parameter γ corresponds to a quantile of y. Based on general experience, γ is usually set to 0.25.

Meanwhile, the EI acquisition function with better acquisition performance is used to guide the next set of iterative evaluation points. Therefore, the acquisition function expression is obtained, as shown in Equation (7).

$$\begin{aligned} EI_{y^*}(x) &= \int_{-\infty}^{+\infty} \max(y^* - y, 0) p_M(y \mid x) dy \\ &= \int_{-\infty}^{y^*} (y^* - y) p(y \mid x) dy \\ &= \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(x)} dy \\ &= \int_{-\infty}^{y^*} (y^* - y) \frac{l(x)p(y)}{\gamma l(x) + (1-\gamma)g(x)} dy \\ &= \frac{\int_{-\infty}^{y^*} (y^* - y)p(y)dy}{\gamma + (1-\gamma)\frac{g(x)}{l(x)}} \end{aligned} \tag{7}$$

From the equation above, it follows that $EIy^*(x) \propto (\gamma + (1 - \gamma)g(x)/l(x))^{-1}$, i.e., the value of EI is positively related to the derivative of the denominator of the expression of the acquisition function. When determined, the size of the denominator depends only on the ratio of the two probabilities $l(x)/g(x)$ of x. The physical significance of this ratio lies in the fact that this assessment point is the ratio of the probability of the more effective hyperparameter to the probability of the less effective hyperparameter in the current observed assessment. Therefore, the problem of finding the hyperparameters that maximize the acquisition function translates into finding the hyperparameters that maximize this ratio, and the TPE algorithm is based on this principle to find the next set of hyperparameters.

In summary, the TPE algorithm optimizes the acquisition function by analyzing the ratio of the probability distribution of the hyperparameters, thus effectively determining the optimal combination of hyperparameters.

The improved Bayesian optimization algorithm logically determines which types of hyperparameters different types of hyperparameters belong to and then implements the establishment of probabilistic agent models and corresponding acquisition functions to comprehensively update and fit the hyperparameter combinations with the best diagnostic performance of the model.

2.4. Self-Attention Mechanism

By mimicking how attention resources are distributed in the human brain, the attention mechanism (AM) increases the impact of important information on the model output outcomes [31]. The input information is very complex because load forecasting needs to extract the time series information and characteristic information of uncertain quantities, such as cooling, heating, and electric loads. The self-attention mechanism can adaptably concentrate on various sections of the incoming data and assign greater weights to the important information. Therefore, this paper adopts a neural network centered on the self-attention mechanism to extract the characteristic information of each uncertain quantity and its related variables so as to ensure that the model can focus on the most important characteristic details of each uncertain quantity. The self-attention mechanism is responsible for calculating the attention score, which is used to determine the relevance of the data in question in relation to other data points in the sequence. The configuration of the AM is depicted in Figure 2. In this paper, the dot product is used to calculate the attention score. The specific steps are as follows:

First, the input data x is multiplied with the matrices W_q, W_k, and W_v to obtain the sequence of query vectors q, the sequence of key vectors k, and the sequence of value vectors v, respectively:

$$q = W_\mathrm{q}x \tag{8}$$

$$k = W_\mathrm{k}x \tag{9}$$

$$v = W_v x \tag{10}$$

The value weight coefficients corresponding to each key are then obtained by scaling the dot product, which yields the correlation between the query vector sequence q and all the keys in the key vector sequence k. Then, applying the *soft*max function to normalize the weight coefficients, the value vector sequence v is weighted and summed to produce the attention score sequence a:

$$a = att[(k,v),q] = v^* softmax\left(\frac{k^\mathrm{T}q}{\sqrt{d}}\right) \tag{11}$$

where *att* is the computational mechanism of the attention score; d is the vector sequence's dimension.

The attention mechanism offers two clear benefits: firstly, it allows the model to learn the correlation between distant inputs by lowering the maximum distance between any

two points in the input sequence to 1; secondly, it allows the input data to be processed in parallel, increasing computational efficiency. These two characteristics can be relied upon to ensure the efficacy and rapidity of the prediction model in feature extraction.

2.5. Multi-Task Learning Theory

The multi-task learning approach as a solution to the problem of inadequate training for multidimensional and multivariate data [32]. Let T be the number of target tasks, and let $X \times Y$, $X \in R^n$, $Y \in R$ be the distribution space from which all the data in each task originate. As training samples, each task has m data points, i.e., $\{(x_{11}, y_{11}), (x_{21}, y_{21}), \ldots, (x_{m1}, y_{m1})\}$, sampled from a distribution P_t in the space $X \times Y$, where each task has a different P_t, but they are interrelated. Therefore, the learning of T functions $f_1, f_2, \ldots, f_r$ is the aim of multi-task learning, such that

$$f_t(x_{it}) = y_{it} \tag{12}$$

The knowledge contained in various tasks can be thought of as distinct tasks within multi-task learning during the process. It is important to think about learning many functions simultaneously to balance the limitations and correlations across activities. IES multivariable load forecasting is not simply the sum of predictions for each load category; independent modeling and separate forecasting methods are not applicable. More attention should be paid to the coupling mechanism between multiple loads, exploring more implicit information within, which gives multivariable load forecasting practical significance. Multi-task learning is a method that enhances model representation and generalization capabilities by simultaneously learning multiple tasks, and the regularization of parameters induced by optimizing the learning across tasks demonstrates better generalization performance compared to ordinary training models.

2.6. Shapley Additive Interpretation of Load Forecasting

Like numerous other neural network models, the load forecasting model is opaque. It is necessary to assess the model to ascertain its efficacy and explore the mechanisms by which the characteristics function. Shapley additive interpretation (SHAP) is employed to visually represent the genuine influence of different metrics linked to load on the DTW-BiLSTM-MTL model's output in this work [33]. Effective predictors are connected to the model interpretation by applying Shapley values in the SHAP approach. The Shapley value is the primary parameter of the SHAP method. Since many factors affect power load, the Shapley value can be employed to quantify the contribution of each team member to the overall benefit, whether it rises or reduces. This is feasible if every feature connected to power is thought of as a team member and the overall benefit is taken into account when estimating load.

For a sample $x_i = \{x_i^1, x_i^2, \ldots, x_i^i, \ldots, x_i^n\}$ from past information X with n samples, the Shapley value ϕ_i^j of the j-th feature x_i^j can be defined as follows:

$$\phi_i^j = \sum_{S \subseteq x_i \setminus \{x_i\}} \frac{|S|!(p - |S| - 1)!}{p!} \left(val\left(S \cup \left\{ x_i^j \right\} \right) - val(S) \right) \tag{13}$$

In the formula, S represents the subset of the values of the j-th feature x_i after being selected; $|S|$ represents the amount of features in subset S; and $val(S)$ is the feature function, indicating the degree to which the features in S influence the result of the negative load prediction through the 'cooperative' process. Formula (14) illustrates how to compute its value using the output value of the negative prediction model that makes use of the features that are not part of subset S.

$$val(S) = \int \hat{F}\left(x_i^1, \cdots, x, \cdots, x_i^p \right) dx - \frac{1}{n} \sum_{i=1}^{n} \hat{F}(x_i), x \notin S \tag{14}$$

Equation (14) requires a sum of integrated operation for each load-dependent feature not included in the sample x_i^j, so the SHAP equation can be readily derived from Equation (15).

$$y_t = y_{\text{base}} + f\left(x_i^1\right) + f\left(x_i^2\right) + \ldots + f\left(x_i^n\right) \tag{15}$$

where y_{base} represents the mean amount of the anticipated negative load. The Shapley value of the initial feature within the sample related to the negative load that predicts the negative load outcome y_t is indicated by $f\left(x_i^1\right)$. If $f\left(x_i^1\right) > 0$, it indicates that the initial feature positively influences the predicted negative load result y_t, thereby increasing the relative value of y_t compared to y_{base}; conversely, it decreases it.

SHAP will be employed more effectively to comprehend the direction and significance of each load-related indicator after training the DTW-BiLSTM-MTL model.

3. Framework of Multi-Task Learning-Based Joint Prediction Model for Multiple Loads

The sharing mechanism of MTL can account for the coupling relationship between multivariate loads by sharing the learning layer so tasks can interact. In cases with a high correlation between prediction targets, the shared learning layer does not cause much loss and enhances parameter sharing between models. Models with multiple targets can be trained jointly, reducing the parameter size of the model and preventing model overfitting. For IES, the complex coupling relationship between multiple loads with many uncertainties, the large number of parameters, and the close connection between the sub-tasks of cooling, heating, and electricity make it more suitable to use MTL to construct the model, which has a better fitting effect and better generalization ability. Therefore, the IES multivariate load forecasting model is constructed using MTL in this paper.

In addition, the process of selecting features is a pivotal aspect of model construction, and good or undesirable features play a decisive role in the prediction effect. Traditional methods usually rely on manual experience in feature engineering construction. A combination of manual experience and correlation analysis is used for feature selection in this study. Specifically, the features are scored using Spearman correlation analysis and Pearson correlation analysis. Pearson correlation analysis is mainly used to measure the linear correlation between two variables, while Spearman correlation analysis measures the ranked connection between two variables. By combining these two methods, the importance of each feature can be effectively assessed to help validate and assist decision-making, allowing the modeling process to retain important features and remove redundant features, thus improving the predictive capacity of the model.

The training process for the multivariate load forecasting method based on multi-task learning is as follows:

(1) Data preprocessing. Abnormal data are eliminated, and missing data for cold, heat, and electricity loads and related influencing factors are filled in.

(2) Correlation analysis. Comprehensive analysis of the relevant influencing factors of multivariate loads is conducted by selecting Pearson and Spearman correlation coefficients, deleting the weakly correlated features therein, reducing the impact of the weakly correlated features on the model's prediction accuracy and training speed, and assisting in feature selection.

(3) Construction of an input layer. Preliminary prediction of target data sequences is carried out by a BiLSTM-based MTL model, then historical sequences similar to the target sequences are selected through DTW, and similar data are spliced with the source data to optimize the model's input attributes to enhance the model's capacity for generalization and training.

(4) Shared layer construction. The MTL shared layer is constructed using BiLSTM to extract the coupling information between the loads of electricity, heat, and cold, and the best choice of BiLSTM hyperparameters is achieved through the application of the Bayesian optimization technique.

(5)   Output layer construction. To realize the differential selection of significant features in the shared layer by various subtasks, the attention layer is added before the output layer of each task, which will help the model make better use of the data. The attention mechanism prioritizes the critical characteristics pertaining to heat, cold, and electricity loads while assigning varying probability weights to the BiLSTM's hidden states. Figure 3 depicts the DTW-BiLSTM-MTL model that is suggested in this paper.
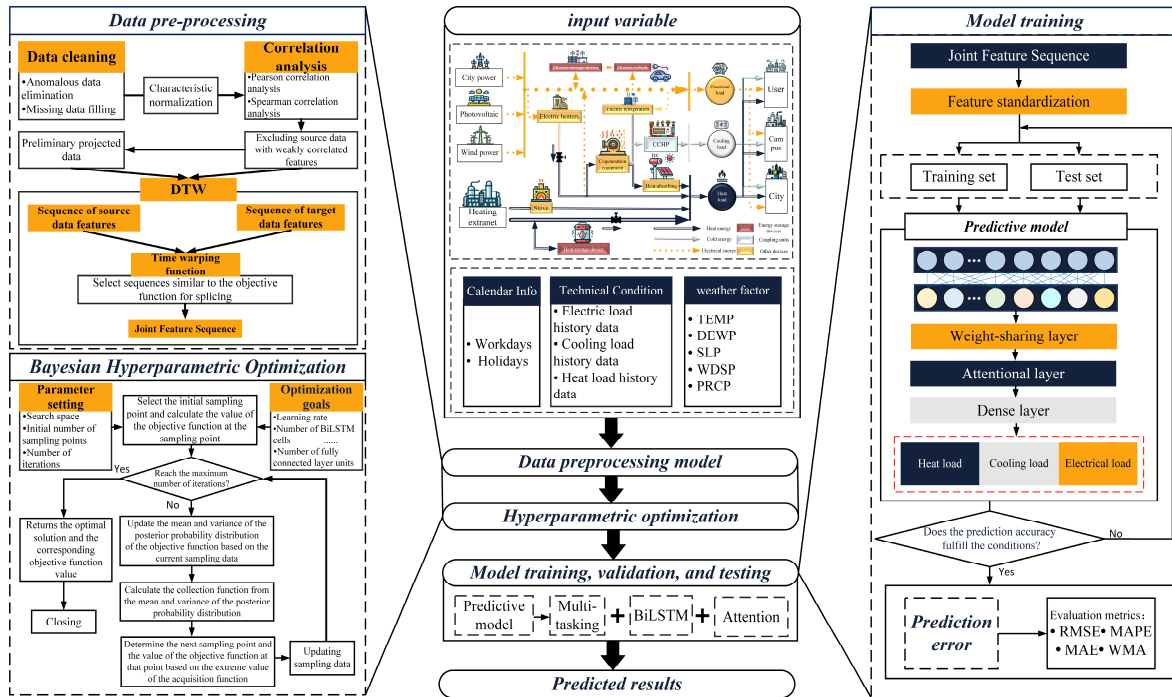

Figure 3. Flowchart of multivariate joint load forecasting based on multi-task learning.

4. Performance Evaluation

This case utilizes multivariate load data from the University of Arizona, Tempe Campus Energy Information System. The dataset contains cooling, heating, and electric load data between January 2023 and December 2023, and the sampling rate of the data is 1 h, i.e., it contains 24 sets of data per day. The meteorological data were obtained from the National Oceanic and Atmospheric Administration (NOAA), and the meteorological stations corresponded to the geographic location of the energy station. At the same 1 h time interval, the meteorological data contained six influencing factors: mean dew point, mean temperature, mean barometric pressure, mean wind speed, maximal wind speed, and precipitation.

4.1. Data Preprocessing

The distance Z from the data points to the overall mean was first calculated using the Z-Score formula, as follows:

$$Z = \frac{x - \mu}{v} \tag{16}$$

where x is the value of the data point, μ is the mean of the overall data, and v is the standard deviation of the overall data. The points where Z is too large are excluded, and the missing data in the dataset are filled in using linear interpolation. Meanwhile, to help the model capture the periodicity of load changes, the original date and time information is converted into a feature vector $[0/1, h]$, which represents whether the day is a working day (yes is 1, no is 0) and whether the current time point is the h-th hour of the day.

Considering the influence of the magnitude problem on other input features, loads of cooling, electricity, and heat, in addition to all the climate data, are normalized to facilitate the input. The calculation formula is as follows:

$$Z = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}} \tag{17}$$

where x_i is the data before normalization of each input eigenvalue; $x_{i,\min}$ and $x_{i,\max}$ are the minimum and maximum values corresponding to x_i, respectively.

In order to guarantee that the predicted values accurately reflect the actual physical significance of each type of load in the IES, the predicted values are ultimately inverted and normalized in order to restore their magnitudes. The formula for the back-normalization is as follows:

$$x_i = Z(x_{i,\max} - x_{i,\min}) + x_{i,\min} \tag{18}$$

4.2. Analysis of the Input Feature Contribution

Considering that multivariate load forecasting involves a complex relationship between several variables, including the supply of different energy sources, weather factors, and so on, taking all the factors as feature inputs will increase noise interference, making the model complex and the training process too long, and overfitting may occur. Through correlation analysis, the degree of interaction between these variables can be determined, helping to identify important influencing factors, thus optimizing the prediction model and improving prediction accuracy.

Since the factors that influence the integrated energy system are complex, there are complicated distribution relationships between different features. The two sets of variables have a linear connection, which is necessary for the Pearson correlation coefficient to be calculated. In contrast, although it can deal with the nonlinear relationship between the variables, the Spearman correlation coefficient is not sensitive to outliers, so the Pearson and Spearman correlation coefficients were chosen simultaneously for the correlation analysis.

The formula for the Pearson correlation coefficient is as follows:

$$r_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}} \tag{19}$$

where r_{xy} is the Pearson correlation coefficient between variables x and y, where the observed values in the sample data are denoted by x_i and y_i. The sample size is denoted by n, and the sample means of x and y are, respectively, $\overline{x}$ and $\overline{y}$.

The calculation formula for the Spearman correlation coefficient is as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{20}$$

where ρ is the Spearman rank correlation coefficient, d_i is the difference between the two variables in the ranking, and n represents the sample size.

The correlation analyses of Pearson and Spearman are utilized to obtain the contribution of cold, heat, and electricity loads and the six meteorological conditions to each other.

(1) Observing the values of each correlation coefficient in Figure 4, it can be observed that the data values of cold, heat, and electricity loads show a significant correlation. The absolute value of Pearson's coefficient is above 0.4. The absolute value of Spearman's coefficient is above 0.7, confirming a close coupling relationship among cold, heat, and electricity loads in this system.

(2) The Pearson's coefficient value of the average temperature is above 0.7, the Spearman's coefficient value is above 0.4, the absolute value of the average air pressure is above

0.6, the absolute value of the Spearman's coefficient value fluctuates above and below 0.5, and the correlation with the load data value is strong. The Pearson's coefficient value between the average dew point and the cold and hot loads is above 0.5, the Spearman's coefficient value is above 0.65, the Pearson's correlation coefficient value with the electric loads is below 0.1, and the Spearman's correlation coefficient value is 0.6, which indicates that the average dew point is strongly correlated with the cold and hot loads and weakly correlated with the electric loads.

(3) The correlation between the average wind speed and maximum wind speed and the load data values is small. The absolute value of Pearson's correlation coefficient is below 0.4; the Spearman's correlation coefficient's absolute value varies from 0.24 to 0.41, which is weakly correlated with the load data values; and the correlation coefficient between the precipitation and the load data values has an absolute value below 0.2, which is very weakly correlated with the load data values.
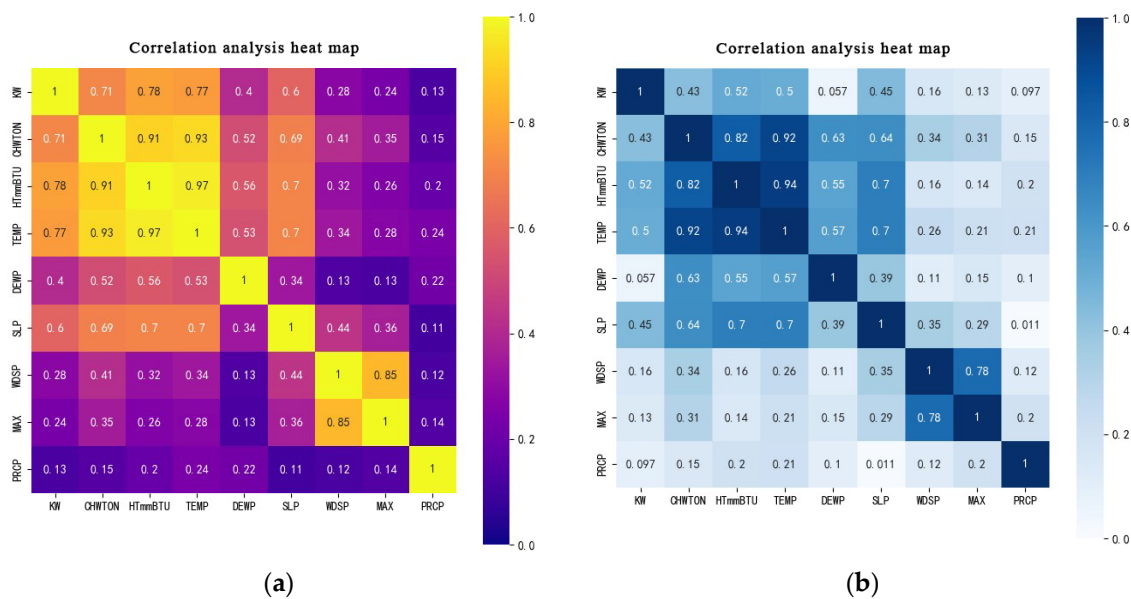

(a)   (b)

Figure 4. Heat maps for correlation analysis. (**a**) Pearson correlation analysis; (**b**) Spearman correlation analysis.

The correlation is strong since the correlation coefficient values of mean temperature, mean air pressure and mean dew point with load data values are large. Therefore, these three factors are considered the main factors influencing the load data values to be input into the model in this paper.

4.3. Evaluation Criteria

A few commonly used evaluation criteria were applied to the IES joint prediction model. Three indexes are used to evaluate computational statistics: weighted mean accuracy (WMA), root mean square error (RMSE), and mean absolute percentage error (MAPE), which are calculated by the following formulas:

(1) MAPE is the mean square error between the actual output value and the predicted output value when the integrated energy system makes a joint prediction of each type of energy, and is calculated by the following formula:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{x(t) - y(t)}{x(t)} \right| \times 100\% \tag{21}$$

where n is the number of samples, $x(t)$ is the actual output value at time t, and $y(t)$ is the forecast output value at time t. *MAPE* is the average absolute error of the load forecast value in the training set in any 24 h day.

(2) When predicting different types of energy simultaneously, the RMSE is the arithmetic square root of the mean square error between the actual and expected output integrative energy system values, calculated by the following formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x(t) - \hat{y}(t))^2} \qquad (22)$$

In the formula, the interpretation of the remaining variables is the same for *MAPE*; $\hat{y}(t)$ is the standard deviation of the expected output value at time t.

(3) The integrated energy system's WMA reflects the distribution of errors of different energy subsystems.

The calculation formula is as follows:

$$\text{WMA} = \frac{\sum_{i=1}^{n} |y_{\text{true},i} - y_{\text{pred},i}| \cdot w_i}{\sum_{i=1}^{n} w_i} \qquad (23)$$

where $y_{\text{true},i}$ and $y_{\text{pred},i}$ represent the actual and predicted values of the i-th sample. w_i represents the weight of this sample, which is based on its influence or priority corresponding to the subsystem.

4.4. Comparative Analysis of Multivariate Load Forecasting Results

All the experiments involved in this study are based on building multivariate load prediction models under a Python 3.7 compilation environment with TensorFlow and Keras frameworks. The model hyperparameters involved in the experiments are shown in Table 1. The data were partitioned for the deep modeling experiments, with 63% of the data designated as the training set, 30% designated as the test set, and the remaining 7% as the validation set. The time step was set to 24 h. To validate the superiority of the DTW-BiLSTM-MTL model multivariate load prediction proposed in this paper, the prediction effects of the model's extreme gradient boosting and multi-task learning (XGBOOST-MTL), CNN-BiLSTM-MTL, and the model proposed in this study were compared and analyzed, respectively. Meanwhile, to test the enhancement effect of DTW and MTL on BiLSTM prediction models, the DTW-BiLSTM-STL and BiLSTM-MTL models were set up for comparison as ablation experiments.

Table 1. Model parameter settings.

Model Name	Hyper Parameterization
DTW-BiLSTM-STL	Number of clusters is 5; number of BiLSTM units: 51; number of fully connected layer units: 30; learning rate: 0.0007; number of training rounds: 50
DTW-BiLSTM-MTL	Number of clusters is 5; number of BiLSTM units: 51; number of fully connected layer units: 30; learning rate: 0.0007; number of training rounds: 50
CNN-BiLSTM-MTL	Number of convolutional layer filters: 64; convolutional kernel size: 5; number of BiLSTM units: 51; number of fully connected layer units: 30; learning rate: 0.0007; number of training rounds: 50
BiLSTM-MTL	Number of BiLSTM units: 51; number of fully connected layer units: 30; learning rate: 0.0007; number of training rounds: 50
XGBOOST-MTL	Maximum depth: 3; learning rate: 0.08; number of trees: 50

In order to guarantee the impartiality of the experiment, each model employed a Bayesian optimization algorithm to identify the optimal hyperparameters. The prediction results of each model for a particular week are shown in Figure 5, and the evaluation metrics on the test set are displayed in Table 2.
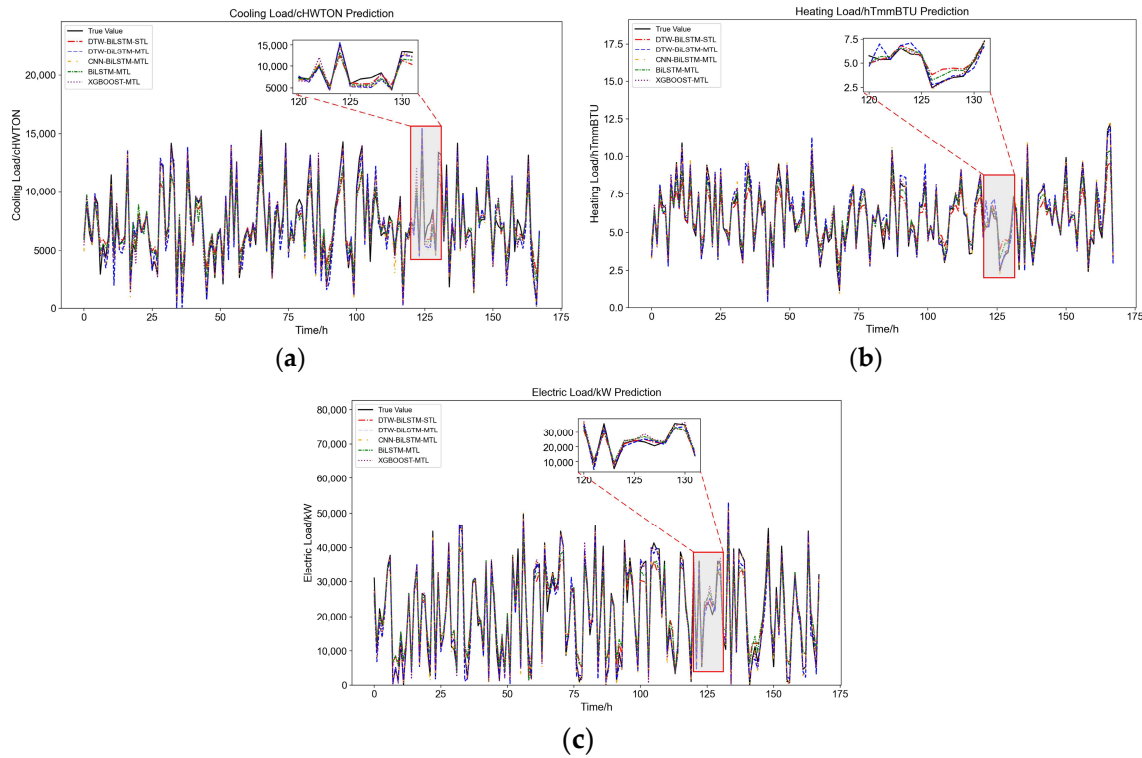

Figure 5. Multivariate load forecasting results. (**a**) Cool load; (**b**) heat load; (**c**) electric load.

Table 2. Prediction performance of test sets based on different prediction models.

Model	Type of Load	RMSE	MAPE (%)	WMA	Time (s)
DTW-BiLSTM-STL	Cold load	0.3266	0.4917	0.3647	577
	Heat load	0.3770	0.5879	0.4907	
	Electric load	0.4166	0.6718	0.6252	
DTW-BiLSTM-MTL	Cold load	0.0888	0.2624	0.0792	542
	Heat load	0.1604	0.4578	0.0977	
	Electric load	0.0723	0.2899	0.0547	
CNN-BiLSTM-MTL	Cold load	0.0964	0.2827	0.0772	558
	Heat load	0.1762	0.5024	0.1099	
	Electric load	0.0903	0.3564	0.0684	
BiLSTM-MTL	Cold load	0.2541	0.4317	0.2762	502
	Heat load	0.3329	0.6373	0.2579	
	Electric load	0.2981	0.6782	0.2279	
XGBOOST-MTL	Cold load	0.0986	0.2868	0.0941	475
	Heat load	0.1701	0.4617	0.1048	
	Electric load	0.1018	0.4499	0.0825	

Figure 5 shows that the prediction error on weekdays is smaller than that on holidays because the energy demand on holidays has greater uncertainty and a lack of regularity data. Regarding prediction accuracy, it can be seen from Table 2 and Figure 6 that when different models are used to predict the loads for heating, cooling, and electricity, the DTW-BiLSTM-STL model has the highest RMSE, MAPE, and WMA, the worst prediction effect, and the longest training time. As an example, when using the DTW-BiLSTM-MTL

model for electrical load prediction, the RMSE and MAPE were reduced by 75.75%, 57.25%, and 62.08% compared to the BiLSTM-MTL model, and the RMSE, MAPE, and WMA were decreased by 82.65%, 56.84%, and 91.25% compared to the DTW-BiLSTM-STL model, verifying the improvement in prediction accuracy of the BiLSTM model through multi-task learning and the dynamic time warping algorithm.
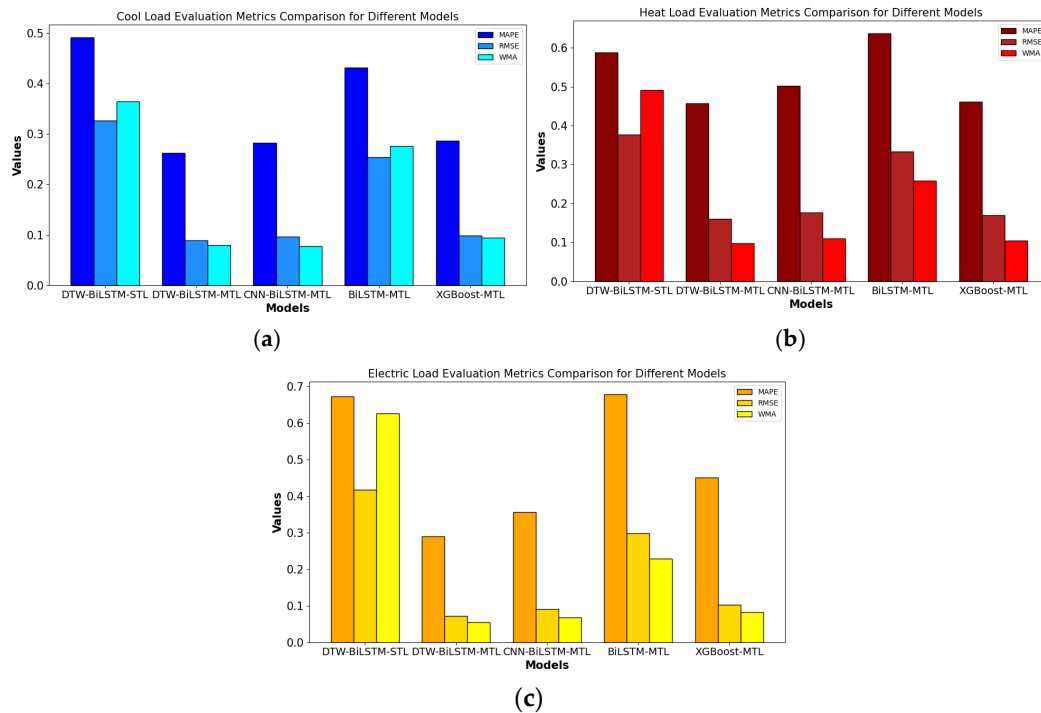

Figure 6. Comparison of multivariate load forecasting accuracy based on different models. (**a**) Cool load; (**b**) heat load; (**c**) electric load.

Compared with the other two comparison models, the DTW-BiLSTM-MTL model has the lowest RMSE, MAPE, and WMA, which indicates that its overall prediction effect is the most accurate. The DTW-BiLSTM-MTL model horizontally realizes the information sharing among different types of loads by establishing the MTL sharing layer based on BiLSTM, learning the characteristics of the coupling between the heating, cooling, and electric load data, and horizontally realizing the information sharing. It can also effectively utilize the auxiliary coupling characteristics to reduce the prediction error when a single load fluctuates greatly. DTW clustering, on the other hand, splices the input data by measuring the similarity of time series and vertically realizes the extraction of future time series information, which leads to a greater enhancement of the prediction effect.

4.5. Model Interpretation

The plot of SHAP of the DTW-BiLSTM-MTL model is shown in Figure 7.

Figure 7 illustrates the top 20 metrics with the largest impacts on multivariate loads between 1 January 2023 and 31 December 2023. The heat load metrics have the greatest impact, with most metrics impacting both directions. There is a significant impact between different load indicators, which verifies the impact of the tight coupling between different load indicators on the forecast accuracy. Meanwhile, comparing the rankings in Figure 7 to the relevant rankings in Figure 4 can prove the indicator's validity. In addition, when analyzing the metrics' effects at various moments in time, the visualization of multiple time points during the forecasting period is presented in Figure 7, both in the morning and evening. The load significantly affects the forecast accuracy, and the users' electricity consumption patterns during the morning and evening peaks and the lunch breaks tend to

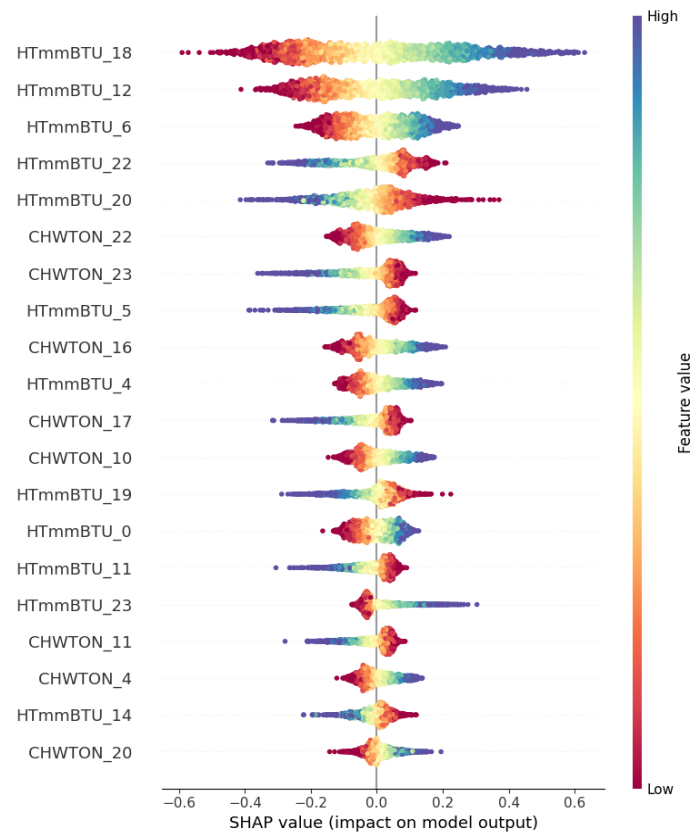increase the load. In contrast, the users' related behaviors during working hours tend to decrease the load.

Figure 7. Distribution of Shapley values for different load-related metrics (backtracked through the proposed DTW-BiLSTM-MTL network).

5. Discussion

This section will present a comprehensive analysis of the practical implications of the proposed model. Additionally, it will address the limitations of the model and offer suggestions for potential improvements.

5.1. Economic Impact

The proposed DTW-BiLSTM-MTL model has been evaluated and found to accurately forecast nonlinear, unstable, static, and periodic fluctuations in cooling, heating, and electricity loads. The accurate forecasting of short-term loads can result in significant economic benefits. For example, a 1% decrease in forecast mistakes can result in cost savings of up to USD ten million [34]. This indicates that the model's high accuracy directly contributes to cost efficiency. Additionally, the model supports energy companies in making informed decisions regarding power generation, distribution, and heating supply, as well as in effectively planning maintenance schedules, thereby significantly reducing operational and maintenance costs.

5.2. Potential Impact on Practical Applications

The inherent instability of loads for electricity, heating, and cooling, coupled with the challenges of energy storage, often threatens the stability of integrated energy systems, possibly resulting in dire circumstances, like energy shortages and waste [35]. The DTW-BiLSTM-MTL multivariate load forecasting model described in this study is capable of accurately identifying load trends, providing reliable information support for energy production and dispatch, and enhancing the operational stability of the system. The DTW algorithm can be employed to adjust load data to ensure the provision of high-quality

inputs to the forecasting model, despite the limitations posed by the availability of limited data. The BiLSTM network is capable of effectively capturing forward and backward loads and is furthermore able to adapt to different time scales. The MTL framework is designed to preserve the coupling relationship between different loads during the prediction process, thereby enhancing the effectiveness of multivariate load forecasting. Furthermore, the DTW-BiLSTM-MTL model demonstrates excellent adaptability, rendering it particularly well-suited for multivariate time series forecasting tasks beyond load forecasting. In scenarios characterized by high load variability or complex weather conditions, the model can provide operators with effective support for energy transmission and dispatch decisions, thereby enhancing the efficiency of resource utilization and reducing energy consumption and potential energy losses. The model's broad applicability makes it an excellent instrument for applications promoting sustainable energy management.

5.3. Differences and Links to Existing Models

The DTW-BiLSTM-MTL model proposed in this study not only retains the advantages of traditional multivariate load forecasting models but also introduces significant innovations. To extract the load coupling information, the model adopts the MTL framework, which is similar to the traditional multivariate load forecasting model, to capture the complex correlations between different loads through the common feature space. The interpretability and correlation analyses in this paper validate the mutual contribution of different loads, and the application of MTL ensures the sharing of information among the load forecasting tasks, thus making full use of the interconnection relationships.

Meanwhile, the DTW algorithm effectively reduces noise and error in the time domain by dynamically adjusting the data, enabling more consistent and higher quality inputs to subsequent models. Such consistency is critical for predictive models that depend on consistency, allowing the model to improve its predictions based on more stable inputs. Compared to traditional simple smoothing or static synchronization methods, DTW significantly improves the consistency of the data load and ensures the reliability of the prediction results. Finally, BiLSTM further enhances the model's capacity to capture time series characteristics. Compared to traditional models, such as LSTM, CNN, and XGBOOST, BiLSTM is more sensitive to cyclical variations in the data, while being able to learn from both historical and future information, with the memory unit ensuring BiLSTM's long-term memory capability and good consistency with inputs from the DTW. In this study, through the organic combination of DTW, BiLSTM, and MTL, a comprehensive modeling of complex interrelationships between multiple loads is achieved. Compared with the existing models, this model not only has advantages in modeling inter-load relationships but also shows significant performance improvement in terms of capturing performance and improving prediction accuracy.

5.4. Limitations and Recommendations

Although the DTW-BiLSTM-MTL model has demonstrated efficacy, it encounters difficulties when forecasting dynamic cooling, heating, and electricity loads. Moreover, the model's forecasting capabilities are constrained in contexts where the energy supply is unstable. Consequently, the selection of a robust foundation model is of paramount importance to boost the forecasting model's applicability. Further investigation into the potential of higher accuracy multi-task learning models as base models or the incorporation of additional models into the base structure may prove beneficial in enhancing the model's applicability. For instance, each household has different appliances, schedules, and usage preferences [36]. When dealing with more complex data, the variational mode decomposition (VMD) approach, which requires less computing work and is unaffected by mode mixing problems, could be utilized for feature selection and signal denoising [37]. It would be beneficial to test the proposed model using larger datasets, shorter time intervals, or data from multiple locations to validate its efficacy. It may be beneficial to consider incorporating

additional variables, such as consumer behavior, seasonal effects, and economic indicators, into future models.

6. Conclusions

The method proposed in this paper fully considers the connection of coupling between multivariate loads and the effective use of future time-series information, uses the DTW algorithm as an optimization method for the input features of prediction models, calculates the distance matrix between different load data sequences, and applies hierarchical clustering methods according to the distance matrix to cluster and splicing, which effectively solves the problem of the traditional prediction model, which relies on historical data and cannot fully utilize the future time-series information. The MTL shared layer is simultaneously established using a BiLSTM neural network, fully exploiting the properties of coupling between electrical, heat, and cold loads and applying the Bayesian optimization approach to determine the prediction model's hyperparameters' global optimal solution.

After finding the optimal parameters, the method has better adaptability and prediction accuracy in load forecasting applications compared to traditional single-task learning and machine learning models. Taking the actual load data of Arizona State University, Tempe Campus as an example and comparing the prediction results of various algorithms under three types of loads, namely, cold, heat, and electricity, it is proven that the DTW-BiLSTM-MTL load forecasting method proposed in this paper can achieve very high prediction accuracy and can adapt to various types of load profiles. The model's training time is short, which allows it to meet the needs of the actual operation of the power system.

References

1. Rifkin, J. The Third Industrial Revolution: How Lateral Power Is Transforming Energy, the Economy, and the World. *Civ. Eng.* **2012**, *82*, 74–75.
2. Mo, Y.; Kim, T.H.; Brancik, K.; Dickinson, D.; Lee, H.; Perrig, A.; Sinopoli, B. Cyber—Physical Security of a Smart Grid Infrastructure. *Proc. IEEE* **2012**, *100*, 195–209.
3. Wang, Y.; Ma, K.; Li, X.; Liang, Y.; Hu, Y.; Li, J.; Liu, H. Multi-type Load Forecasting of IES Based on Load Correlation and Stacked Auto-Encode Extreme Learning Machine. In Proceedings of the 2020 10th International Conference on Power and Energy Systems, Chengdu, China, 25–27 December 2020; pp. 585–589.
4. Panda, S.K.; Ray, P.; Salkuti, S.R. A Review on Short-Term Load Forecasting Using Different Techniques. In Proceedings of the Recent Advances in Power Systems, Berlin, Germany, 14 February 2022; pp. 433–454.
5. Veeramsetty, V.; Mohnot, A.; Singal, G.; Salkuti, S.R. Short Term Active Power Load Prediction on A 33/11 kV Substation Using Regression Models. *Energies* **2021**, *14*, 2981. [CrossRef]
6. Lekshmi, M.; Subramanya, K.N.A. Short-Term Load Forecasting of 400kV Grid Substation Using R-Tool and Study of Influence of Ambient Temperature on the Forecasted Load. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms, Gangtok, India, 25–28 February 2019; pp. 1–5.
7. Wang, Q.; Wang, H.; Gupta, C.; Rao, A.R.; Khorasgani, H. A non-linear function-on-function model for regression with time series data. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 232–239.

8.   Mishra, D.P.; Jena, S.; Senapati, R.; Panigrahi, A.; Salkuti, S.R. Global solar radiation forecast using an ensemble learning approach. *Int. J. Power Electron. Drive Syst.* **2023**, *14*, 496–505. [CrossRef]

9.   Abedinia, O.; Amjady, N. Short-term load forecast of electrical power system by radial basis function neural network and new stochastic search algorithm. *Int. Trans. Electr. Energy Syst.* **2016**, *26*, 1511–1525. [CrossRef]

10.   Veeramsetty, V.; Rakesh Chandra, D.; Salkuti, S.R. Short term active power load forecasting using machine learning with feature selection. In *Next Generation Smart Grids: Modeling, Control and Optimization*; Springer Nature: Singapore, 2022; Volume 824, pp. 85–101.

11.   Singh, S.; Hussain, S.; Bazaz, M.A. Short term load forecasting using artificial neural network. In Proceedings of the 2017 Fourth International Conference on Image Information Processing, Shimla, India, 21–23 December 2017; pp. 1–5.

12.   Dehalwar, V.; Kalam, A.; Kolhe, M.L.; Zayegh, A. Electricity load forecasting for urban area using weather forecast information. In Proceedings of the 2016 IEEE International Conference on Power and Renewable Energy, Shanghai, China, 21–23 October 2016; pp. 355–359.

13.   Duan, M.; Darvishan, A.; Mohammaditab, R.; Wakil, K.; Abedinia, O. A novel hybrid prediction model for aggregated loads of buildings by considering the electric vehicles. *Sustain. Cities Soc.* **2018**, *41*, 205–219. [CrossRef]

14.   Hu, L.; Zhang, L.; Wang, T.; Li, K. Short-term load forecasting based on support vector regression considering cooling load in summer. In Proceedings of the 2020 Chinese Control And Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 5495–5498.

15.   Niu, D.X.; Wanq, Q.; Li, J.C. Short term load forecasting model using support vector machine based on artificial neural network. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; pp. 4260–4265.

16.   Qiuyu, L.; Qiuna, C.; Sijie, L.; Yun, Y.; Binjie, Y.; Yang, W.; Xinsheng, Z. Short-term load forecasting based on load decomposition and numerical weather forecast. In Proceedings of the 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China, 26–28 November 2017; pp. 1–5.

17.   Ahmad, N.; Ghadi, Y.; Adnan, M. Load forecasting techniques for power system: Research challenges and survey. *IEEE Access* **2022**, *10*, 71054–71090. [CrossRef]

18.   Li, J.; Deng, D.; Zhao, J.; Cai, D.; Hu, W.; Zhang, M.; Huang, Q. A novel hybrid short-term load forecasting method of smart grid using MLR and LSTM neural network. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2443–2452. [CrossRef]

19.   Jiang, Q.; Zhu, J.; Li, M.; Qing, H. Electricity power load forecast via long short-term memory recurrent neural networks. In Proceedings of the 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), Wuhan, China, 19–21 April 2018; pp. 265–268.

20.   Gunawan, J.; Huang, C. An extensible framework for short-term holiday load forecasting combining dynamic time warping and LSTM network. *IEEE Access* **2021**, *9*, 106885–106894. [CrossRef]

21.   Munem, M.; Bashar, T.R.; Roni, M.H.; Shahriar, M.; Shawkat, T.B.; Rahaman, H. Electric power load forecasting based on multivariate LSTM neural network using Bayesian optimization. In Proceedings of the 2020 IEEE Electric Power and Energy Conference (EPEC), Edmonton, AB, Canada, 18 January 2020; pp. 1–6.

22.   Ge, L.; Li, Y.; Yan, J.; Zhang, J.; Li, X. Multivariate Two-stage Adaptive-stacking Prediction of Regional Integrated Energy System. *J. Mod. Power Syst. Clean Energy* **2022**, *11*, 1462–1479. [CrossRef]

23.   Li, W.W.; Zhang, P.Y.; Shi, Q.; Feng, C.Y.; Li, D. Load correction forecasting for integrated energy systems based on aggregated hybrid modal decomposition and time-series convolutional neural networks. *Power Syst. Technol.* **2022**, *46*, 3345–3357.

24.   Bai, B.Q.; Liu, J.T.; Wang, X.; Jiang, C.W.; Jiang, T.; Zhang, S.X. Short-term forecasting of urban energy multiple loads based on MRMR and dual attention mechanism. *Autom. Electr. Power Syst.* **2022**, *46*, 44–55.

25.   Wu, K.; Gu, J.; Meng, L.; Wen, H. An explainable framework for load forecasting of a regional integrated energy system based on coupled features and multi-task learning. *Prot. Control Mod. Power Syst.* **2022**, *7*, 24. [CrossRef]

26.   Zhao, P.; Cao, D.; Hu, W.; Huang, Y.; Hao, M.; Huang, Q. Geometric loss-enabled complex neural network for multi-energy load forecasting in integrated energy systems. *IEEE Trans. Power Syst.* **2023**, *39*, 5659–5671. [CrossRef]

27.   Wu, C.; Yao, J.; Xue, G.Y.; Wang, J.; Wu, Y. Load forecasting of an integrated energy system based on MMoE multitask learning and long and short-term memory networks. *Electr. Power Autom. Equip.* **2022**, *42*, 33.

28.   Gilanifar, M.; Wang, H.; Sriram, L.M.K.; Ozguven, E.E.; Arghandeh, R. Multitask Bayesian spatiotemporal Gaussian processes for short-term load forecasting. *IEEE Trans. Ind. Electron.* **2019**, *67*, 5132–5143. [CrossRef]

29.   Li, K.; Mu, Y.; Yang, F.; Wang, H.; Yan, Y. A novel short-term multi-energy load forecasting method for integrated energy system based on feature separation-fusion technology and improved CNN. *Appl. Energy* **2023**, *351*, 121823. [CrossRef]

30.   Biju, G.; Pillai, G. Hyperparameter Optimization of Long Short Term Memory Models for Interpretable Electrical Fault Classification. *IEEE Access* **2023**, *11*, 123688–123704.

31.   Niu, D.; Yu, M.; Sun, L.; Gao, T. Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism. *Appl. Energy* **2022**, *313*, 118801. [CrossRef]

32.   Fujii, K.; Kawahara, Y. Supervised dynamic mode decomposition via multitask learning. *Pattern Recognit. Lett.* **2019**, *122*, 7–13. [CrossRef]

33.   Zhang, Z.; Liu, J.; Pang, S.; Shi, M.; Goh, H.H.; Zhang, Y. General short-term load forecasting based on multi-task temporal convolutional network in COVID-19. *Int. J. Electr. Power Energy Syst.* **2023**, *147*, 108811. [CrossRef]

34. Peng, L.; Lv, S.X.; Wang, L. Effective electricity load forecasting using enhanced double-reservoir echo state network. *Eng. Appl. Artif. Intel.* **2021**, *99*, 104132. [CrossRef]

35. Li, S.; Kong, X.; Yue, L.; Liu, C.; Khan, M.A.; Yang, Z. Short-term electrical load forecasting using hybrid model of manta ray foraging optimization and support vector regression. *J. Clean. Prod.* **2023**, *388*, 135856. [CrossRef]

36. Quilumba, F.L.; Lee, W.J.; Huang, H.; Wang, D.Y. Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Trans. Smart Grid* **2014**, *6*, 911–918. [CrossRef]

37. Ribeiro, M.H.D.M.; Silva, R.G.; Moreno, S.R.; Canton, C. Variational mode decomposition and bagging extreme learning machine with multi-objective optimization for wind power forecasting. *Appl. Intell.* **2024**, *54*, 3119–3134. [CrossRef]