PAPER

Power Peak Load Forecasting Based on Deep Time Series Analysis Method

Ying-Chang HUNG[†a)], ***Member and*** **Duen-Ren LIU**[†], ***Nonmember***

SUMMARY The prediction of peak power load is a critical factor directly impacting the stability of power supply, characterized significantly by its time series nature and intricate ties to the seasonal patterns in electricity usage. Despite its crucial importance, the current landscape of power peak load forecasting remains a multifaceted challenge in the field. This study aims to contribute to this domain by proposing a method that leverages a combination of three primary models - the GRU model, self-attention mechanism, and Transformer mechanism - to forecast peak power load. To contextualize this research within the ongoing discourse, it's essential to consider the evolving methodologies and advancements in power peak load forecasting. By delving into additional references addressing the complexities and current state of the power peak load forecasting problem, this study aims to build upon the existing knowledge base and offer insights into contemporary challenges and strategies adopted within the field. Data preprocessing in this study involves comprehensive cleaning, standardization, and the design of relevant functions to ensure robustness in the predictive modeling process. Additionally, recognizing the necessity to capture temporal changes effectively, this research incorporates features such as "Weekly Moving Average" and "Monthly Moving Average" into the dataset. To evaluate the proposed methodologies comprehensively, this study conducts comparative analyses with established models such as LSTM, Self-attention network, Transformer, ARIMA, and SVR. The outcomes reveal that the models proposed in this study exhibit superior predictive performance compared to these established models, showcasing their effectiveness in accurately forecasting electricity consumption. The significance of this research lies in two primary contributions. Firstly, it introduces an innovative prediction method combining the GRU model, self-attention mechanism, and Transformer mechanism, aligning with the contemporary evolution of predictive modeling techniques in the field. Secondly, it introduces and emphasizes the utility of "Weekly Moving Average" and "Monthly Moving Average" methodologies, crucial in effectively capturing and interpreting seasonal variations within the dataset. By incorporating these features, this study enhances the model's ability to account for seasonal influencing factors, thereby significantly improving the accuracy of peak power load forecasting. This contribution aligns with the ongoing efforts to refine forecasting methodologies and addresses the pertinent challenges within power peak load forecasting.

key words: *time series analysis, self-attention mechanism, transformer mechanism, weekly moving average, monthly moving average*

1. Introduction

This study aims to assess the suitability and effectiveness of GRU [1], Self-attention network [2], and Transformer network [2] in power peak load forecasting, leveraging extensive historical data along with reference data such as "weekly moving average" and "monthly moving average." The experimental dataset is derived from historical data provided by Taiwan Power Company, encompassing daily peak load values spanning from August 1, 2015, to May 31, 2023.

The research encompasses a comparative analysis involving three primary models: the GRU network, GRU + Self-attention network, and GRU + Transformer network, serving as experimental models for ablation experiments. Additionally, two traditional prediction models, ARIMA (Autoregressive Integrated Moving Average model) and SVR (Support Vector Regression), are employed as the foundation for comparative experiments. Furthermore, the study introduces "Weekly Moving Average/Monthly Moving Average" as an additional feature to augment the model's capacity to capture seasonal and temporal variations [3], [4].

1.1 Problem Identification

power peak load forecasting poses a critical and intricate challenge within the energy and power sector. It pertains to predicting the maximum load demand on the power system during specific timeframes. Given the inherent difficulty of storing electrical power, power companies must meticulously allocate resources for generation and consumption based on precise peak load predictions. This practice is essential to maintain a balance between power supply and demand, ensuring grid stability and security. Without accurate forecasts, an excessive demand for electricity can lead to voltage fluctuations, instability, and grid failures, significantly impacting economic activities and daily life. Accurate power peak load forecasting yields several benefits, including improved energy efficiency, reduced waste and carbon emissions, cost savings, risk mitigation, support for renewable energy integration, and enhanced customer satisfaction.

Nonetheless, power peak load forecasting is a formidable task due to its susceptibility to internal and external factors, including weather conditions, holidays, seasons, social events (such as the Covid-19 pandemic), and user behavior. These variables imbue power peak load forecasting with nonlinear, non-stationary, multi-variation, and multi-sequence characteristics. Consequently, addressing these challenges necessitates advanced data analysis and modeling techniques.

In recent years, the advent of machine learning and deep learning has prompted a growing number of researchers to apply these methodologies to power peak load forecasting. Machine learning and deep learning techniques excel in data

Manuscript received September 8, 2023.
Manuscript revised January 22, 2024.
Manuscript publicized March 21, 2024.

[†]The authors are with the Institute of Information Management, National Yang Ming Chiao Tung University, HsinChu, Taiwan, R.O.C.
a) E-mail: Eltonhung@gmail.com

DOI: 10.1587/transinf.2023EDP7187

Copyright © 2024 The Institute of Electronics, Information and Communication Engineers

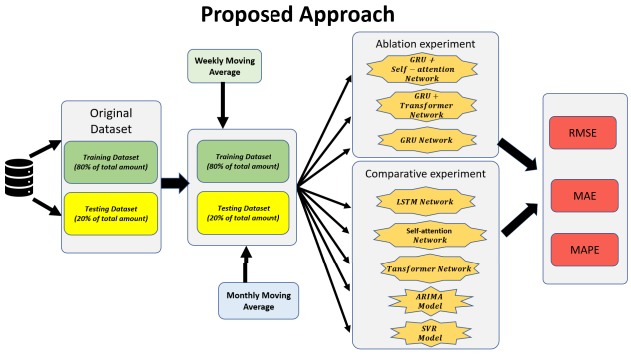

Fig. 1 Proposed methodology

representation, feature extraction, and the discovery of hidden patterns. Prominent among these methods are the Gated Recurrent Unit (GRU), Self-attention network, and Transformer network, all adept at processing time-series data and considering long-term dependencies. While many power peak load forecasting studies predominantly employ time series models like LSTM [5], ARIMA [6] and SVR [7] models, the increasing popularity of Self-attention and Transformer mechanisms in various deep learning domains has introduced new possibilities. To diversify this research, the author utilizes a GRU evolved from LSTM and combines it with these two attention-based deep learning models to explore their viability in power peak load forecasting.

In addition to standard forecasting variables like climate data (temperature, humidity, sunshine hours) and seasonal indices (Seasonal Index), this study introduces "Weekly Moving Average" and "Monthly Moving Average" as supplementary features. These features calculate load data averages over fixed windows of time, such as weekly and monthly intervals, enhancing the model's capacity to capture temporal and seasonal variations.

Figure 1 illustrates the methodology proposed by the author, which encompasses the six predictive networks/models. It incorporates the use of "Weekly Moving Average" and "Monthly Moving Average" as supplementary features to address the crucial issue of peak load electricity forecasting.

1.2 Contributions

The contributions of this paper can be summarized as follows:

- Comprehensive Ablation Experiment: This research conducts a thorough ablation experiment, individually and in combination, using deep learning networks such as GRU, Self-Attention, and Transformer mechanisms, along with the incorporation of traditional forecasting models like ARIMA/SVR. The performance of these models is critically evaluated through comparative experiments, providing valuable insights into their predictive capabilities.
- Introducing "Weekly Moving Averages" and "Monthly Moving Averages" proves effective in capturing the impact of seasonal factors on peak load forecasting in power systems. These two moving average methods aid in data smoothing, highlighting both short-term and long-term trends within time series data, particularly demonstrating significant efficacy in handling seasonal variations. While the weekly moving average captures cyclic changes within a week, the monthly moving average is more adept at capturing monthly seasonal variations. Through the incorporation of these features, the model can better account for the influence of seasonal factors on peak load, thereby enhancing predictive accuracy.
- Combining the features of "Weekly Moving Averages" and "Monthly Moving Averages" within a deep learning network holds significant implications for predicting power load. These features capture trends at different time scales: "Weekly Moving Averages" may reflect weekly cyclic patterns, while "Monthly Moving Averages" better represent longer-term seasonal changes. By amalgamating these features, we provide a more comprehensive understanding of load behavior. The deep learning network can leverage this combined information to better comprehend these trends and patterns. Such integration enhances predictive performance, enabling more accurate forecasts of future power demand.
- Practical Verification Using Taiwan Power Peak Load Data: To validate the methodology's practicality and performance, this study utilizes historical power peak load data from Taiwan. Detailed data experiments are conducted, offering insights into the applicability and effectiveness of the proposed approach.

2. Related Works

power peak load forecasting plays a pivotal role in the power system, offering valuable insights to enhance operational efficiency and energy management for both power companies and energy providers. Its primary objective is to predict the maximum demand on the power system within a specified future timeframe, enabling the formulation of appropriate strategies. Power companies rely on these forecasts to optimize power generation and distribution in alignment with user needs. Dispatch centers, known as Independent System Operators, utilize peak load predictions to make well-informed decisions, such as activating or deactivating generating units, adjusting transmission capacity, and managing energy reserves.

Traditional power peak load forecasting methods often hinge on statistical models, including time-series analysis and regression. However, these methods may fall short in capturing non-linear and seasonal load variations [8]–[10]. Self-attention and Transformer-based models, originally prominent in natural language processing, have found their way into the energy sector, particularly in predicting electricity consumption. Recent literature has witnessed an influx of self-attention and Transformer-based models, primarily concentrating on Power Supply Prediction. These models have been notably applied in research areas like wind power generation and solar power generation. Despite exploring actual power peak load forecasting based on

historical data remains a pertinent subject. Thus, this article delves into the application of these models, specifically Self-attention and Transformer-based, in predicting power peak loads in Taiwan.

In this section, the researcher reviews various methodologies employing deep learning models and explores pertinent literature concerning the self-attention mechanism and Transformer for electricity demand forecasting.

Within the realm of power demand prediction, several commonly employed methods include time-series analysis, regression models, artificial neural networks, decision tree methods, support vector machines, and more. The following offers a concise introduction to these methods:

1. Time-Series Analysis is a statistical method that utilizes historical data to analyze temporal trends and predict future data based on these patterns. Its advantage lies in its ability to handle non-linear and non-stationary data, capturing factors like seasonality, periodicity, and randomness. However, it requires a substantial amount of data to build a model and struggles with emergencies or outliers. Common techniques include the ARIMA model [8], [9] and exponential smoothing.
2. The regression model is a supervised learning approach that establishes a function to describe the relationship between independent and dependent variables, enabling the prediction of unknown dependent variables. Its strength lies in handling multivariate data and assessing the influence of each independent variable on the dependent variable. However, it assumes data conformity to a specific distribution and is less effective with nonlinear or highly correlated data. Common regression models encompass linear regression, multivariate regression, etc [10].
3. Artificial Neural Networks (ANNs) mimic the human brain's operations, connecting neurons through learned weight adjustments to map complex nonlinear relationships within large, unstructured datasets. ANNs excel in processing high-dimensional data, extracting hidden features, and uncovering data patterns. Nevertheless, they demand substantial computational resources and training time, making them susceptible to issues like overfitting and local optima [11].
4. The Decision Tree method employs graphical binary judgment rules to categorize data into various values, identifying optimal judgment rules by exploring potential paths. Decision trees effectively handle both categorical and numerical data, offering an intuitive visualization of feature impact on outcomes. However, they are sensitive to noise and outliers, occasionally generating overly complex or oversimplified rules [12].
5. The Support Vector Machine (SVM) relies on statistical theory to attain optimal classification or regression by identifying a hyperplane that maximizes the margin between different categories or values. SVMs excel in handling highly nonlinear and correlated data, and their flexibility and generalization capabilities can be enhanced through Kernel Functions [13].

However, these models exhibit certain limitations in dealing with complex and nonlinear patterns in the data. Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, have proven to yield superior results when tackling data complexity and non-linearity [14], [15].

Since Vaswani et al. introduced the Transformer model for natural language processing in 2017 [2], it has become a key methodology for time series and sequence-to-sequence predictions. In recent years, emerging neural network models like self-attention and Transformer have also made inroads into power peak load forecasting. Self-attention is a machine learning attention mechanism capable of learning correlations between different time steps. The Transformer, an encoder-decoder architecture based on self-attention, has demonstrated remarkable results in natural language processing and image processing.

In the realm of power peak load forecasting, some researchers have begun to explore the application of emerging neural network models such as self-attention and Transformer. For instance, Hu et al. proposed a method that combines self-attention and LSTM for predicting electric vehicle charging demands. They utilized CRPS (Continuous Ranked Probability Score) as the evaluation loss function to assess the recommendations based on the self-attention method for forecasting electric vehicle charging. This approach was chosen because the self-attention method can balance historical patterns and current trends, alleviate long-term forgetting, and enable superior predictions using CRPS. It demonstrated promising outcomes in forecasting electric vehicle charging demands [16]. Jun Wei Chan introduced the Transformer model and compared it to state-of-the-art approaches such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). In the end, he recommended using the "GLEU" activation function to emphasize its speed and impressive accuracy [17].

In addition to the consideration and design of Self-attention and Transformer models, "Weekly Moving Average" and "Monthly Moving Average" represent another set of commonly used techniques in time-series prediction. These techniques primarily aim to capture seasonal and long-term trends present in time series data. "Weekly Moving Average" entails computing the average based on load data from the previous week, while "Monthly Moving Average" calculates the average based on the load data from the preceding month. While these moving averages have found widespread application in various time series forecasting problems, such as stock market and weather forecasting, their utilization in power peak load forecasting remains relatively limited.

"Weekly Moving Average" effectively captures weekly seasonal trends. By calculating the average load data from the past week, a corresponding weekly moving average value is derived and incorporated as a feature in the forecasting model. This enhancement allows the model to better ac-

count for week-to-week patterns, subsequently improving forecast accuracy. Similarly, the "Monthly Moving Average" adeptly captures monthly seasonal trends, such as variations in electricity consumption between summer and winter. By computing the mean of the load data over the past month, a one-month moving average value is obtained as a forecasting model feature. This further augments the model's ability to grasp monthly patterns of change, contributing to improved forecast accuracy [18].

Furthermore, this study conducted a series of related ablation studies and comparative experiments. These studies primarily investigated the impact of different model designs and parameters on the accuracy of power peak load forecasting. Among them, certain studies have indicated that models based on Self-attention and Transformer outperform traditional time series models in power peak load forecasting [19]–[22]. Nevertheless, others have emphasized that the effectiveness of power peak load forecasting is significantly influenced by model design and parameter adjustments. In summary, existing literature suggests that deep learning models, particularly Transformers, effectively manage complexity and non-linearity in power demand data. Concurrently, self-attention mechanisms enhance these models' capacity to capture long-term dependencies. Building upon these findings, this study establishes the use of self-attention mechanisms and Transformers for peak load prediction in Taiwan.

3. Methodology

In this section, the primary focus is to delve into the process of conducting ablation and comparative experiments concerning power peak load forecasting. It involves elucidating fundamental machine learning concepts and the methodology proposed by the author. The aim is to ascertain whether the author's proposed method aligns with the capabilities and significance of power peak load forecasting.

The initial step in this undertaking involves data collection and preparation. This research relies on the 'historical power supply and demand data' obtained from and purchased through the Taiwan Power Company. The sample dataset for short term provided by Taiwan power company is consistently sourced from the Government opendata website (www.data.gov.tw) [23]. Through official channels, data pertaining to 'date,' 'Net peak power supply capacity' (MegaWatt), 'Peak load' (MegaWatt), 'Backup capacity' (MegaWatt), 'Backup capacity ratio', and the power generation data of each generator unit is obtained. And the time period is from 2015/08/01 to 2023/05/31. This study will perform data experiments on "peak load" (MegaWatt), augmenting it with "Weekly Moving Average" and "Monthly Moving Average."

3.1 Data Collection and Preprocessing

The historical power supply and demand data from Taiwan Power Company were collected and split into training and test datasets. Preprocessing steps, including data cleaning, feature selection, and normalization, were executed as follows:

3.1.1 Data Collection

Collect historical electrical load data, encompassing timestamps and corresponding peak load values. This data can be acquired from open data platforms provided by power companies and government sources.

3.1.2 Data Preprocessing

Process the collected data, which may involve tasks such as removing outliers, handling missing values, and smoothing data to reduce noise.

3.1.3 Feature Engineering

Organize and compute the "Weekly Moving Average" and "Monthly Moving Average" from the dataset. These features enhance the model's ability to capture patterns influenced by seasonal changes, a crucial aspect in addressing electricity demand.

3.1.4 Dataset Splitting

Divide the data into training and test datasets, with the first 80% allocated to training and the remaining 20% to testing.

3.2 Model Construction

This section provides a detailed description of the structure and mathematical functions of each model. The researcher will introduce the following models in succession: the GRU network, the GRU + Self-attention network, the GRU + Transformer network, as well as the ARIMA and SVR models.

3.2.1 GRU Network

This is a traditional GRU model that only uses GRU for electricity consumption prediction [1].

Suppose there is a time series of peak load power data, expressed as $X = [x1, x2, \ldots, xn]$, where n represents the peak load power at time stamp t.

Then use GRU to transform this time series into a feature sequence, denoted as $H = [h_1, h_2, \ldots, h_n]$. The recursive formula of GRU is as follows:

$$r_t = \sigma(W_r \odot x_t + U_r \odot h_{t-1} + b_r)$$
$$z_t = \sigma(W_z \odot x_t + U_z \odot h_{t-1} + b_z)$$
$$\tilde{h}_t = \tanh(W \odot x_t + U \odot (r_t \odot h_{t-1}) + b)$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t$$

Where r_t and z_t are update gate and forget gate, $\sigma(x)$

represents the sigmoid function, $\tanh(x)$ represents the hyperbolic tangent function, $\odot$ represents the element-wise product, W_r, U_r, b_r, W_z, U_z, b_z, W, U, and b are addressed as weights and biases of models.

Then, by using a linear layer to perform electricity consumption prediction on the feature sequence H. Assuming that the predicted power consumption sequence is $Y = [y_1, y_2, \ldots, y_n]$, the formula of the forecast model is:

$$y_t = W_y \odot h_t + b_y$$

Where Wy and by are addressed as weight and bias of models

The process of the entire GRU network is as follows: Use GRU to transform the power consumption X of the time series into a feature sequence H. Use the linear layer to predict the electricity consumption of the feature sequence H, and obtain the predicted electricity consumption sequence Y.

The GRU network's relative simplicity, efficient training dynamics, and strong performance with short to medium-length sequences position it as a robust tool for modeling and forecasting sequential data. In our ongoing exploration of power peak load forecasting, it will investigate how the GRU network can be harnessed to enhance predictions and contribute to more effective power management strategies. The mechanism of the GRU network is shown as Model 1 in Fig. 2.

3.2.2 GRU + Self-Attention Network

The model combines the strengths of GRU and the Self-Attention network. GRU focuses on extracting sequence features, while the Self-Attention network adds deep learning-based importance weighting.

In power peak load forecasting, the GRU + Self-Attention network leverages both these architectures to capture temporal patterns and extended dependencies in the data. This hybrid approach merges Gated Recurrent Units (GRUs) and Self-Attention mechanisms, aiming to enhance predictive accuracy in power peak load forecasting by adeptly capturing short-term and long-term temporal patterns in historical power supply and demand data.

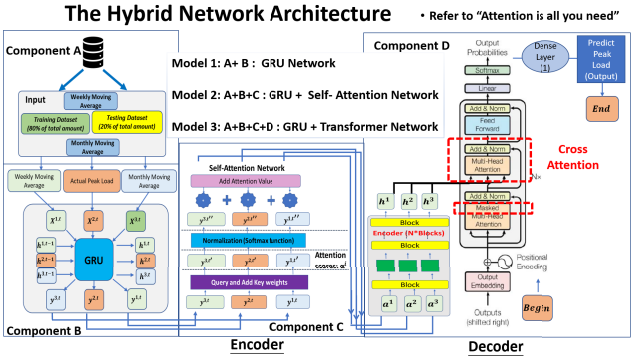

Fig. 2 The hybrid network architecture

Self-Attention Mechanism: The Self-Attention mechanism complements GRU capabilities by capturing distant dependencies and global context. It assigns importance to different time steps based on their relevance to the current prediction. This proves valuable in understanding how past power consumption patterns influence future peak loads, even over extended intervals.

To predict using this hybrid model:

Data Preparation: Historical power supply and demand data, along with relevant features like "weekly moving average" and "monthly moving average," are preprocessed for training and testing.

Model Architecture: The GRU + Self-Attention network design optimally integrates both components. GRU processes sequences, while Self-Attention captures contextual relationships across time steps.

By harnessing the strengths of GRU and Self-Attention, the GRU + Self-Attention network offers a comprehensive power peak load forecasting approach. It adeptly captures various temporal patterns and dependencies, enhancing accuracy for informed decisions in effective power management and supply stability. The mechanism of the GRU + Self-Attention network is shown as Model 2 in Fig. 2.

3.2.3 GRU + Transformer Network

This model combines the advantages of GRU and the Transformer network. GRU handles sequence feature extraction, while the Transformer network manages feature interaction and prediction.

The GRU + Transformer network is a hybrid model that merges Gated Recurrent Units (GRUs) with the Transformer architecture. This fusion enhances the accuracy of power peak load forecasting by leveraging the strengths of both components. It effectively captures short-term dynamics and provides a global context within historical power supply and demand data.

Transformer Architecture: The Transformer introduces a self-attention mechanism critical for capturing long-range dependencies and contextual relationships. This mechanism evaluates the importance of various time steps concerning current predictions, particularly valuable for understanding how past power consumption patterns impact future peak loads across extended intervals.

By combining GRUs and the Transformer architecture, the GRU + Transformer network benefits from both:

Short-Term Patterns: GRU captures immediate consumption trends, revealing short-term variations and patterns.

Global Context: The Transformer's self-attention mechanism widens the context, identifying long-term dependencies and relationships. This is essential for predicting peak loads influenced by seasonal or significant changes.

To forecast power peak load using this hybrid model:

Data Preparation: Historical power supply and demand data, along with features like "weekly moving average" and "monthly moving average," undergo preprocessing for train-

ing and testing.

Model Architecture: The GRU + Transformer network architecture optimizes both components. GRU focuses on sequential patterns, while the Transformer's self-attention captures contextual interdependencies.

The GRU + Transformer network offers a comprehensive approach to power peak load forecasting by combining different strengths. It captures a wide range of temporal patterns and dependencies, leading to improved forecasting accuracy. This approach supports informed decision-making for effective power management and supply stability, contributing to better resource allocation and energy planning. The mechanism of the GRU + Transformer network is shown as Model 3 in Fig. 2.

3.2.4 LSTM Network

Exploring the Long Short-Term Memory (LSTM) model involves clarifying its fundamental concepts, architecture, and relevance within your research or application. In conjunction with the proposed GRU, Self-Attention, and Transformer models, the LSTM model stands as another prominent technique in time series forecasting, enabling a comparative experiment. LSTM, a subtype of recurrent neural network (RNN), excels in capturing and learning intricate dependencies within sequential data, making it well-suited for tasks such as time series analysis and prediction.

The integration of LSTM into our experimental framework aims to assess its predictive capabilities and determine whether it aligns with or distinctively enhances the capture of temporal dependencies and the prediction of power peak loads.

3.2.5 Self-Attention Network

Self-attention networks are a type of neural architecture that processes sequential data by assessing relationships between each element in the sequence. Unlike traditional models limited by local or sequential processing, self-attention allows for a comprehensive analysis of interdependencies across the entire sequence simultaneously.

What makes self-attention networks powerful is their capability to capture long-range dependencies effectively. They excel in modeling intricate relationships within data, making them highly valuable in various natural language processing tasks like machine translation, language understanding, and text generation. The parallel computation ability of self-attention networks enables efficient processing of extensive sequential data, contributing to their effectiveness in handling large-scale datasets. Their prominence in state-of-the-art models, like the Transformer, signifies their crucial role in modern deep learning architectures for sequence processing.

3.2.6 Transformer Network

Understanding the Transformer model involves exploring its fundamental concepts, architectural intricacies, and its potential application within your research or field. Alongside proposed models like GRU and Self-Attention, the Transformer holds significance in time series forecasting. Unlike LSTM, a recurrent neural network variant, the Transformer specializes in capturing extensive dependencies across sequential data, offering a compelling approach for tasks such as time series analysis and prediction.

Integrating the Transformer into our experimental framework aims to assess its predictive prowess, determining if it aligns with or distinctly surpasses in capturing temporal dependencies and forecasting power peak loads compared to other established models.

3.2.7 ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model is a time series forecasting method designed to capture trends and seasonality within the data. It is commonly employed for predicting future data points based on historical patterns.

Once trained, the ARIMA model can be applied to make predictions on forthcoming data. It is particularly effective for forecasting time-dependent data, such as electricity consumption, as it can discern trends and seasonality in the dataset.

3.2.8 SVR Model

SVR (Support Vector Regression) is an extension of SVM (Support Vector Machine) designed to tackle regression problems. The fundamental concept behind SVR is to transform regression tasks into optimization problems that minimize prediction errors, utilizing the principles and techniques of SVM.

In summary, the SVR model provides a data-driven approach to power peak load forecasting by leveraging historical data and uncovering underlying relationships. Its capacity to handle non-linear associations and incorporate diverse features renders it a valuable tool for precise power load predictions. This, in turn, supports improved resource planning, enhanced energy management, and ensures a stable power supply.

3.3 Add the "Weekly Moving Average" and "Monthly Moving Average" into the Dataset

In the realm of power peak load forecasting, climatic factors exert significant influence. Typically, climate variables such as temperature, rainfall, and relative humidity, along with the incorporation of a "Seasonal Index," are integral components to strengthen forecasting models. These factors collectively constitute pivotal determinants of electricity demand.

However, in this study, the aim is to enhance prediction performance by capturing the influence of climate factors on power peak load. As previously introduced, the "Weekly Moving Average" and "Monthly Moving Average" methods

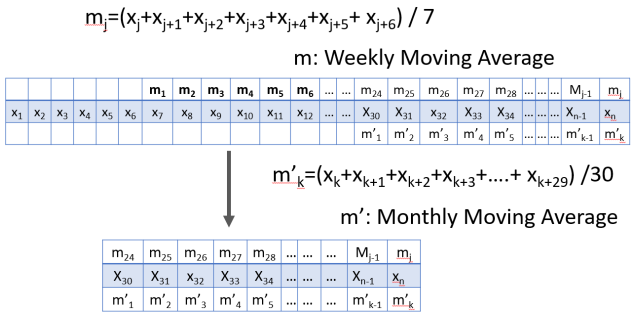

Fig. 3 Weekly/Monthly moving average concept

are employed. In addition to utilizing a "weekly" unit, i.e., averaging data over seven days with a one-day shift to create the weekly moving average, a similar approach is applied to derive the "monthly moving average" based on a "monthly" unit, aggregating data over 30 days with a daily shift to form the monthly moving average. By incorporating these two supplementary datasets, predictions based solely on historical data are substantially reinforced, leading to improved forecasting accuracy-an essential facet of this research.

The utilization of "Weekly Moving Average" and "Monthly Moving Average" to enhance prediction precision extends beyond datasets related exclusively to climate and seasons. It applies broadly to any time series historical dataset, enriching prediction outcomes through the inclusion of these additional features.

The subsequent section provides the mathematical function expressions for adjusting "Weekly Moving Average" and "Monthly Moving Average".

Weekly Moving Average M, $M = [m_1, m_2, \ldots, m_j]$. Where $t = n - 6$, indicating the number of times the sliding window length is 7 days, represented by $t + 7$,

$$m_j = \frac{x_j + x_{j+1} + x_{j+2} + x_{j+3} + x_{j+4} + x_{j+5} + x_{j+6}}{7}$$

Monthly Moving Average M', $M' = [m'_1, m'_2, \ldots, m'_k]$ where the value range of k is from 1 to n, indicating the starting position of each sliding window, and the length of each sliding window is 30 days,

$$m'_k = \frac{x_k + x_{k+1} + x_{k+2} + x_{k+3} + \ldots + x_{k+29}}{30}$$

When forecasting, the adjusted power peak load data (Y1', Y2', . . . , Y12') are used for modeling and forecasting. These adjusted figures will better reflect long-term trends. The mechanism of this concept is shown in Fig. 3.

4. Experiment and Discussion

4.1 Experiment Setting

This section outlines the experiment setup, encompassing dataset preparation, evaluation metrics, and experimental parameter configuration.

4.1.1 Dataset Preparation

The experimental dataset is derived from historical supply and demand data provided by the Taiwan Power Company. Specifically, the dataset covers the timeframe from August 1, 2015, to May 31, 2023, and is divided into two segments. 80% of the total data constitutes the training dataset, while the remaining 20% forms the test dataset. This dataset is subsequently employed for predicting power requirements during power peak load instances.

For each time step, the target variable is the electricity consumption value, with other relevant features (such as date and time) serving as input features.

In the dataset preparation phase, it is imperative to conduct data preprocessing tasks, including feature scaling, handling missing values, and performing feature engineering. These steps ensure that the data is both accessible and suitable for input into the model.

4.1.2 Evaluation Metrics

To assess the performance of various models, the following evaluation metrics will be utilized:

- Root Mean Square Error (RMSE): RMSE represents the square root of the mean squared error between the predicted and actual values. The calculation formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean Absolute Error (MAE): MAE signifies the mean absolute difference between the predicted and actual values. The calculation formula is as follows:

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

- Mean Absolute Percentage Error (MAPE): MAPE signifies the mean absolute Percentage difference between the predicted and actual values. The calculation formula is as follows:

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

A: denotes the actual value, F: denotes the Predict value
These three metrics offer a comprehensive evaluation of the model's predictive performance and depict the extent of dissimilarity between predicted and actual values.

4.1.3 Setting of Experimental Parameters

For each model, there are corresponding parameter settings. For example:

- Setting the hidden state dimension of the GRU layer for

the GRU network model.
- Setting the hidden state dimension of the GRU layer, the number of attention mechanism heads, and the dimension of the linear layer for the GRU + Self-attention network model, and so on.
- Setting the number of layers for the GRU layer and transformer encoder layer, the number of attention mechanism heads, dimensions, and regularization parameters for the GRU + Transformer network model.

These parameters are configured based on empirical knowledge and real-world requirements. Adjusting these parameter values allows for the analysis and comparison of model performance. Additionally, it is essential to prevent overfitting and ensure model stability through techniques like cross-validation.

The dataset is divided into training and testing sets using an 80-20 split, where the initial 80% of the data is allocated for training, and the remaining 20% for testing. The experiments are conducted by implementing the proposed model in the Keras deep learning framework. Training takes place in a cloud environment (Python 3 Google Compute Engine) equipped with an NVIDIA T4 GPU and 15 GB of memory, including 12.7 GB of system memory. For experimental optimization, the model is trained using the "Adam" optimizer with a learning rate of 0.001 and a batch size of 32. The model is trained for 200 epochs.

Subsequently, the validity of the experimental results is confirmed through hypothesis testing. Finally, these five models are compared to each other to assess the strengths and weaknesses of the research findings, leading to the final conclusions.

4.2 Experimental Results

In the ablation experiments, the researcher compared the predictive performance of the GRU network, GRU + Self-attention network, and GRU + Transformer network. Furthermore, comparative experiments were conducted between ARIMA and SVR when compared to the GRU network and GRU + Self-attention network. The results indicated that the GRU + Self-attention network outperformed the others, highlighting the advantage of the self-attention mechanism in extracting sequence features and enhancing prediction accuracy.

4.2.1 Analysis about Training History/ Prediction Trends

Additionally, the following plots depict the convergence during training and the predictions on both training and testing data obtained from experiments for each network or model. Only one example plot for each will be presented.

- For the GRU Network, the training history trend is depicted in Fig. 4.
  - Observing the illustration, it becomes evident that the GRU Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 50th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.

Fig. 4 Training history trend for GRU network

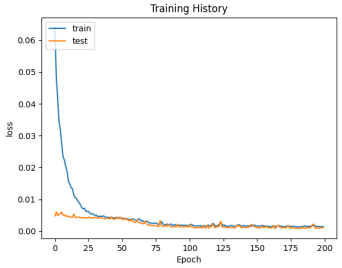

Fig. 5 Training history trend for GRU+ self-attention network

- For the GRU+ Self-Attention Network, the training history trend is depicted in Fig. 5.
  - Observing the illustration, it becomes evident that the GRU+ Self-Attention Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 30th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.
- For the GRU+ Transformer Network, the training history trend is depicted in Fig. 6.
  - Observing the illustration, it becomes evident that the GRU+ Transformer Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 50th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.

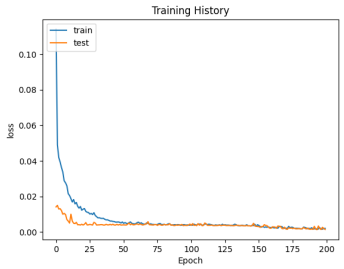

Fig. 6 Training history trend for GRU+ transformer network

Fig. 7 Training history trend for LSTM network

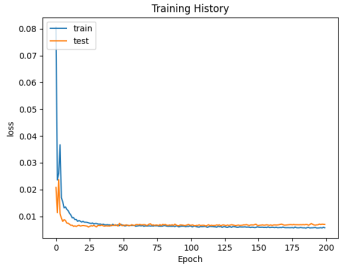

Fig. 8 Training history trend for self-attention network network

- For the LSTM Network, the training history trend is depicted in Fig. 7.
  - Observing the illustration, it becomes evident that the LSTM Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 55th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.
- For the Self-attention network, the training history trend is depicted in Fig. 8.
  - Observing the illustration, it becomes evident that the Self-attention Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 50th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.

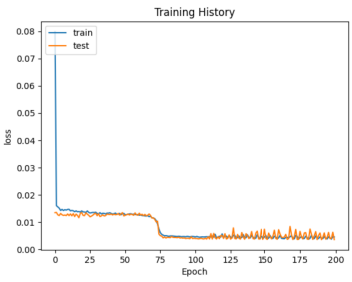

Fig. 9 Training history trend for transformer network

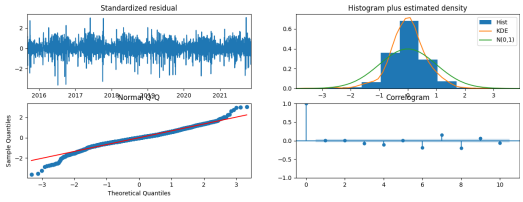

Fig. 10 Training history trend for ARIMA model

- For the Transformer Network, the training history trend is depicted in Fig. 9.
  - Observing the illustration, it becomes evident that the Transformer Network exhibits significant convergence after 200 epochs of computational experiments utilizing both training and test datasets. Notably, starting from approximately the 50th epoch, a gradual convergence between the trends of the test and training datasets becomes apparent. This convergence signifies a close alignment in behavior or performance observed between the test and training data, almost overlapping as the epochs progress.
- For the ARIMA Model, the training history trend is depicted in Fig. 10 and the Actual v.s. Predict comparison is depicted in Fig. 11.
  - From the illustration above, it can "NOT" be observed that the ARIMA Model demonstrates significant convergence after computation experiments using training and test datasets.
- For the SVR Model, the training history trend is depicted in Fig. 12 and the Actual v.s. Predict comparison is depicted in Fig. 13.
  - From the illustration above, it can "NOT" be observed that the SVR Model demonstrates significant convergence after computation experiments using training and test datasets.

4.3 Statistical Significance Test

In this study, seven predictive networks/models have been

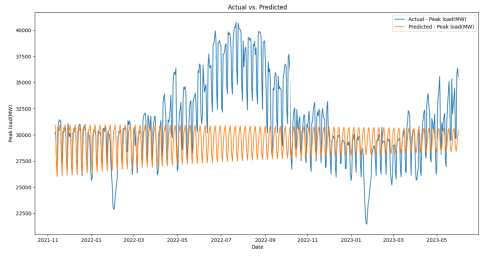

Fig. 11 Actual v.s. predict for ARIMA model

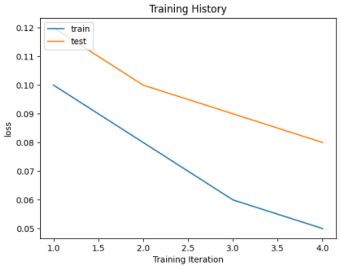

Fig. 12 Training history trend for SVR model

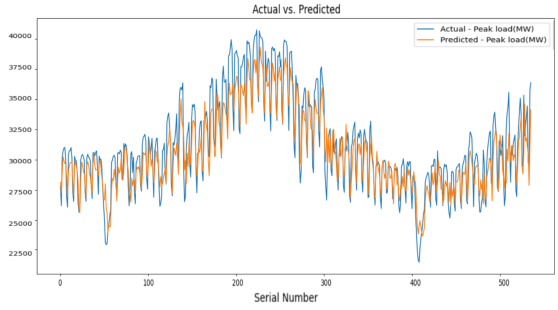

Fig. 13 Actual v.s. predict comparison for SVR model

proposed to address the crucial issue of peak load electricity forecasting. An in-depth exploration of the performance of these models will be conducted, and significant conclusions will be derived through extensive experimentation and statistical analysis.

To begin with, the following table, depicting numerical data displaying the average RMSE, MAE and MAPE values obtained after 30 rounds of experiments, each comprising 200 epochs, for each network/model, is presented in Table 1.

In the experiments, actual peak load electricity datasets were utilized, and the models' accuracy was assessed using performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). To determine whether performance differences were statistically significant, a series of statistical significance tests, specifically the T-test, was conducted to compare the models' performances.

This study investigated the following aspects concerning the new approach: (a) whether there are significant differences in performance among the proposed combined models in ablation experiments, (b) demonstrating whether the proposed model outperforms both the comparison single model

Table 1 Average RMSE, MAE and MAPE values of experiments

	GRU Network (MegaWatt)	GRU+Self-attention Network (MegaWatt)	GRU+Transformer Network (MegaWatt)	LSTM Network (MegaWatt)	Self-attention Network (MegaWatt)	Transformer Network (MegaWatt)	ARIMA Model (MegaWatt)	SVR Model (MegaWatt)
Average RMSE	1183.440894	1224.916505	1441.561216	1212.160388	3221.234232346685	2463.489007	3954.003889	1684.07
Average MAE	858.6744534	877.7279418	1087.058836	889.3366349	2588.843328256	1962.677212	2758.369277	1369.24
Average MAPE	2.820066107	2.895812276	3.512165202	2.893211719	8.445977231769 78	6.545823848	8.513368586	4.340461516

and traditional models in contrast experiments, and (c) exploring factors and situations influencing the experimental performance and potential areas for improvement.

In this section, statistical significance tests were employed to examine and compare the predictive performance between each experimental group and the control group. Initially, an F-test was applied to evaluate the overall variance in financial performance among the models, enabling the correct formulation for subsequent T-tests. The results of the statistical significance tests comparing the proposed models with the control group are presented in Table 2, with the significance level as 5%. The null hypothesis is formulated as follows:

$$H_0 = \mu_b \geq \mu_a$$

In the ablation experiments section, the results indicated significant differences in performance among the GRU network, GRU + Self-Attention network, and GRU + Transformer network. Specifically, the GRU network outperformed the GRU + Self-Attention network, while the GRU + Self-Attention network exhibited better performance than the GRU + Transformer network.

In the comparative experiments section, the author compared each proposed model with single networks or traditional models such as the LSTM network, Self-Attention network, Transformer Network, ARIMA model, and SVR model. The results indicated that, in most cases, the GRU network and the GRU + Self-Attention network outperformed the other models, including LSTM, Self-attention network, Transformer network, ARIMA, and SVR. Deep learning networks, especially the GRU network and GRU + Self-Attention network, demonstrated superior performance compared to traditional time series models. However, the results showed that there were no significant difference in performance between the the GRU + Self-Attention network and LSTM Network.

Additionally, regarding the comparative experiments between the ARIMA model with SVR model, and the result is the SVR model exhibited better performance than the ARIMA model in certain scenarios.

Several observations arise from the above statements:

- Model Complexity and Overfitting: Regarding self-attention mechanism may increase the model's overall complexity. While self-attention is proficient at capturing long-range dependencies, it also introduces numerous parameters. This complexity can lead to overfitting, particularly when the dataset is not sufficiently large or diverse. In contrast, a standalone GRU network may strike a better balance between model complexity and dataset size, enhancing generalization to test data.
- Data Representation and Feature Extraction: Self-attention mechanisms are designed for sequential data,

Table 2 T-test results for RMSE, MAE and MAPE value

T-Test for RMSE with 30 samples

Model1	Model2	T-Statistic	P-Value	Result	
GRU	GRU+Self-attention	-2.332752184	0.023151983	The significant difference in means.	Reject H0
GRU	GRU+Transformer	-14.75130552	2.75E-21	The significant difference in means.	Reject H0
GRU	LSTM	-1.839105494	0.071020314	The significant difference in means.	Reject H0
GRU	Self-attention	-161.0497198	1.34E-78	The significant difference in means.	Reject H0
GRU	Transformer	-56.53119339	2.01E-52	The significant difference in means.	Reject H0
GRU	ARIMA	-236.8531624	2.66E-88	The significant difference in means.	Reject H0
GRU	SVR	-42.79837256	1.41E-45	The significant difference in means.	Reject H0
GRU+Self-attention	GRU+Transformer	-11.60263577	9.46E-17	The significant difference in means.	Reject H0
GRU+Self-attention	LSTM	0.753863889	0.453980632	No the significant difference in means.	Do not Reject H0
GRU+Self-attention	Self-attention	-140.2650477	3.96E-75	The significant difference in means.	Reject H0
GRU+Self-attention	Transformer	-52.56618212	1.26E-50	The significant difference in means.	Reject H0
GRU+Self-attention	ARIMA	-203.8172915	1.60E-84	The significant difference in means.	Reject H0
GRU+Self-attention	SVR	-34.29110444	3.32E-60	The significant difference in means.	Reject H0
LSTM	GRU+Transformer	-13.79863788	5.67E-20	The significant difference in means.	Reject H0
GRU+Transformer	Self-attention	-128.2265342	7.10E-73	The significant difference in means.	Reject H0
GRU+Transformer	Transformer	-43.76484963	4.01E-46	The significant difference in means.	Reject H0
GRU+Transformer	ARIMA	-193.0616736	3.70E-83	The significant difference in means.	Reject H0
GRU+Transformer	SVR	-18.63491353	3.68E-26	The significant difference in means.	Reject H0
LSTM	Self-attention	-176.0007175	7.85E-81	The significant difference in means.	Reject H0
LSTM	Transformer	-56.94228443	1.33E-52	The significant difference in means.	Reject H0
LSTM	ARIMA	-265.0255188	3.95E-91	The significant difference in means.	Reject H0
LSTM	SVR	-45.61459789	3.88E-47	The significant difference in means.	Reject H0
Transformer	Self-attention	-37.92706213	1.23E-42	The significant difference in means.	Reject H0
Transformer	ARIMA	-76.87904558	4.57E-60	The significant difference in means.	Reject H0
SVR	Self-attention	-318.6254229	9.13E-96	The significant difference in means.	Reject H0
SVR	Transformer	-40.2015368	4.73E-44	The significant difference in means.	Reject H0
Self-attention	ARIMA	-151.88946	3.96E-77	The significant difference in means.	Reject H0
SVR	ARIMA	-1.34E+16	0	The significant difference in means.	Reject H0

T-Test for MAE with 30 samples

Model1	Model2	T-Statistic	P-Value	Result	
GRU	GRU+Self-attention	-0.968505477	0.336813438	The significant difference in means.	Reject H0
GRU	GRU+Transformer	-11.12025375	5.24E-16	The significant difference in means.	Reject H0
GRU	LSTM	-1.456845237	0.150551922	The significant difference in means.	Reject H0
GRU	Self-attention	-113.3652699	8.75E-70	The significant difference in means.	Reject H0
GRU	Transformer	-60.1514515	5.81E-54	The significant difference in means.	Reject H0
GRU	ARIMA	-127.8333003	8.48E-73	The significant difference in means.	Reject H0
GRU	SVR	-34.35667777	2.99E-40	The significant difference in means.	Reject H0
GRU+Self-attention	GRU+Transformer	-10.9247135	1.06E-15	The significant difference in means.	Reject H0
GRU+Self-attention	LSTM	-0.589048207	0.558116317	No the significant difference in means.	Do not Reject H0
GRU+Self-attention	Self-attention	-128.1524689	7.34E-73	The significant difference in means.	Reject H0
GRU+Self-attention	Transformer	-64.58324473	1.00E-55	The significant difference in means.	Reject H0
GRU+Self-attention	ARIMA	-145.8798032	4.10E-76	The significant difference in means.	Reject H0
GRU+Self-attention	SVR	-38.12613033	9.17E-43	The significant difference in means.	Reject H0
LSTM	GRU+Transformer	-9.612992361	1.31E-13	The significant difference in means.	Reject H0
GRU+Transformer	Self-attention	-102.8928061	2.35E-67	The significant difference in means.	Reject H0
GRU+Transformer	Transformer	-49.18235664	5.49E-49	The significant difference in means.	Reject H0
GRU+Transformer	ARIMA	-117.900461	9.08E-71	The significant difference in means.	Reject H0
GRU+Transformer	SVR	-19.90618939	1.33E-27	The significant difference in means.	Reject H0
LSTM	Self-attention	-111.0565776	2.87E-69	The significant difference in means.	Reject H0
LSTM	Transformer	-58.37192245	3.23E-53	The significant difference in means.	Reject H0
LSTM	ARIMA	-125.4131384	2.56E-72	The significant difference in means.	Reject H0
LSTM	SVR	-32.20180413	1.06E-38	The significant difference in means.	Reject H0
Transformer	Self-attention	-55.32375175	6.86E-52	The significant difference in means.	Reject H0
Transformer	ARIMA	-73.8725842	4.52E-59	The significant difference in means.	Reject H0
SVR	Self-attention	-350.8393793	3.43E-98	The significant difference in means.	Reject H0
SVR	Transformer	-55.0950778	8.69E-52	The significant difference in means.	Reject H0
Self-attention	ARIMA	-48.76719483	8.87E-49	The significant difference in means.	Reject H0
SVR	ARIMA	-1.16E+16	0	The significant difference in means.	Reject H0

T-Test for MAPE with 30 samples

Model1	Model2	T-Statistic	P-Value	Result	
GRU	GRU+Self-attention	-1.218211515	0.228075983	The significant difference in means.	Reject H0
GRU	GRU+Transformer	-11.46603458	1.53E-16	The significant difference in means.	Reject H0
GRU	LSTM	-1.139704954	0.259094185	The significant difference in means.	Reject H0
GRU	Self-attention	-117.7433805	9.81E-71	The significant difference in means.	Reject H0
GRU	Transformer	-65.98756609	2.92E-56	The significant difference in means.	Reject H0
GRU	ARIMA	-119.8995807	3.44E-71	The significant difference in means.	Reject H0
GRU	SVR	-33.1856609	2.03E-39	The significant difference in means.	Reject H0
GRU+Self-attention	GRU+Transformer	-10.71029361	2.30E-15	The significant difference in means.	Reject H0
GRU+Self-attention	LSTM	0.042301885	0.966403335	No the significant difference in means.	Do not Reject H0
GRU+Self-attention	Self-attention	-125.6442175	2.30E-72	The significant difference in means.	Reject H0
GRU+Self-attention	Transformer	-68.29695869	4.07E-57	The significant difference in means.	Reject H0
GRU+Self-attention	ARIMA	-128.8662971	5.33E-73	The significant difference in means.	Reject H0
GRU+Self-attention	SVR	-34.36433005	2.95E-40	The significant difference in means.	Reject H0
LSTM	GRU+Transformer	-10.36741607	8.01E-15	The significant difference in means.	Reject H0
GRU+Transformer	Self-attention	-118.6785975	6.21E-71	The significant difference in means.	Reject H0
GRU+Transformer	Transformer	-59.11892777	1.56E-53	The significant difference in means.	Reject H0
GRU+Transformer	ARIMA	-122.1729429	1.16E-71	The significant difference in means.	Reject H0
GRU+Transformer	SVR	-21.07790104	7.18E-29	The significant difference in means.	Reject H0
LSTM	Self-attention	-118.280269	7.54E-71	The significant difference in means.	Reject H0
LSTM	Transformer	-65.51016276	4.42E-56	The significant difference in means.	Reject H0
LSTM	ARIMA	-120.5978383	2.46E-71	The significant difference in means.	Reject H0
LSTM	SVR	-32.20188592	1.06E-38	The significant difference in means.	Reject H0
Transformer	Self-attention	-53.26324958	5.95E-51	The significant difference in means.	Reject H0
Transformer	ARIMA	-53.56857417	4.30E-51	The significant difference in means.	Reject H0
SVR	Self-attention	-302.7263911	1.77E-94	The significant difference in means.	Reject H0
SVR	Transformer	-66.83456853	1.41E-56	The significant difference in means.	Reject H0
Self-attention	ARIMA	9.775873712	7.14E-14	The significant difference in means.	Reject H0
SVR	ARIMA	-1.20E+16	0	The significant difference in means.	Reject H0

excelling at tasks where capturing global patterns is essential. In the context of electricity peak load Prediction, the sequential nature of GRU networks may better capture temporal patterns. However, the attention mechanism may prioritize global context at the expense of finer temporal details, resulting in relatively poorer performance in this specific task.

- Training Dynamics: There can be significant variations in the training dynamics of GRU network, GRU+ Self-Attention Network, and GRU+ Transformer network. GRU Networks, with their simpler architecture, tend to exhibit more stable training and require less hyperparameter tuning. The more complex architectures of GRU+ Self-Attention Network and GRU+ Transformer network may lead to more challenging optimization during training, potentially converging to suboptimal solutions without achieving superior predictive power.
- Hyperparameter Tuning: Both GRU and Self-Attention mechanisms entail specific sets of hyperparameters that necessitate fine-tuning. Performance disparities may stem from suboptimal hyperparameter configurations in both the GRU+ Self-Attention Network and the GRU+ Transformer network, affecting their predictive ability relative to a standalone GRU network.

In summary, this study offers valuable guidance for selecting models in peak load electricity Prediction. However, it underscores the significance of factors such as model complexity, data characteristics, training dynamics, and hyperparameter tuning in determining performance. In practical applications, choosing the appropriate model requires a comprehensive consideration of these factors and meticulous optimization. This research contributes to the future of electricity demand Prediction and energy management, providing valuable insights in the ever-evolving energy landscape. Accurate peak load Prediction is a critical step in ensuring the reliability and efficiency of power systems in the dynamic energy sector.

5. Conclusion and Future Works

Future research can apply the same deep learning model and incorporate 'Weekly Moving Average' and 'Monthly Moving Average' as auxiliary methods to improve forecasting accuracy. This approach can find applications in various domains characterized by time-series data and large historical datasets. Furthermore, there is room for exploration in terms of additional deep learning models and optimization algorithms. Consideration of other external factors could also enhance the accuracy and applicability of these predictions. This study presents a general approach to data prediction utilizing deep learning models and enhanced dataset techniques. Specific mathematical functions and model details may vary depending on the implementation and deep learning framework used. For more in-depth discussions regarding mathematical functions and model specifics, further details can be provided upon request.

The primary contribution of this study lies in introducing a predictive method tailored for time-series data. It leverages 'Weekly Moving Average' and 'Monthly Moving Average' as auxiliary tools to enhance prediction accuracy. The model combines GRU, the Self-attention mechanism, and Transformer in a novel design. Experiments conducted on Taiwan's power peak load data showcase the superiority of the proposed model over comparative models. Nonetheless, there exist certain limitations and avenues for future research in this work. Firstly, the proposed model could benefit from further optimization through exploration of different hyperparameters and training strategies. Secondly, incorporation of external factors like weather conditions and economic indicators may elevate forecast accuracy. Thirdly, assessing the model's generalizability by applying it to other regions and countries warrants attention.

To conclude, this study underscores the effectiveness of the Self-attention mechanism and Transformer in Predicting Taiwan's power peak load. The proposed model has

the potential for practical applications in the power industry's decision-making processes. In the future, expanding the application of these models to diverse fields and exploring methods to enhance their predictive capabilities will be valuable directions for further research.

References

[1] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation" arXiv:1406.1078v3 [cs.CL] 3 Sep 2014 DOI: DOI.ORG/10.48550/arXiv.1406.1078, 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 31st Conference on Neural Information Processing Systems (NIPS 2017), arXiv:1706.03762 DOI: DOI.ORG/10.48550/arXiv.1706.03762, 2017.

[3] A. Guntuboyina, "Statistics 153 (Time Series): Lecture Three," 2012-01-24, Accessed 2024-01-07.

[4] R.J. Hyndman, "Moving averages", 2009-11-08, Accessed 2020-08-20.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," vol.9, no.8, pp.1735–1780, Neural Computation, 15 Nov. 1997, DOI: DOI.ORG/10.1162/neco.1997.9.8.1735

[6] G. Box, "Box and Jenkins: Time Series Analysis, Forecasting and Control," A Very British Affair, pp.161–215, Palgrave Advanced Texts in Econometrics, Palgrave Macmillan, 1970 DOI: DOI.ORG/10.1057/9781137291264 6

[7] M.O. Stitson, J.A.E. Weston, A. Gammerman, V. Vovk, and V. Vapnik, "Theory of support vector machines," Technical Report, CSD-TR-96-17, 31 Dec. 1996.

[8] P.C. Huy, N.Q. Minh, N.D. Tien, and T.T.Q. Anh, "Short-Term Electricity Load Forecasting Based on Temporal Fusion Transformer Model," IEEE Access, vol.10, pp.106296–106304, 2022, DOI: 10.1109/ACCESS.2022.3211941

[9] C. Tarmanini, N. Sarma, C. Gezegin, and O. Ozgonenel, "Short term load forecasting based on ARIMA and ANN approaches," Energy Reports, vol.9, pp.550–557, Supplement 2023, DOI: DOI.ORG/10.1016/j.egyr.2023.01.060

[10] T. Tumiran, S. Sarjiya, L. M. Putranto, E. Nugraha Putra, R. F. Setya Budi, and C. Febri Nugraha, "Long-Term Electricity Demand Forecast Using Multivariate Regression and End-Use Method: A Study Case of Maluku-Papua Electricity System," 2021 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP), pp.258–263, 2021, DOI: DOI.ORG/10.1109/ICT-PEP53949.2021.9601144

[11] G. Sideratos, A. Ikonomopoulos, and N.D. Hatziargyriou, "A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks," Electric Power Systems Research, vol.178, p.106025, 2020.

[12] I.U. Khan, N. Javaid, C.J. Taylor, K.A.A. Gamage, and X. Ma, "Big Data Analytics Based Short Term Load Forecasting Model for Residential Buildings in Smart Grids," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2020, DOI:DOI.ORG/10.1109/INFOCOMWKSHPS50562.2020.9163031

[13] M.K. Azad, S. Uddin, and M. Takruri, "Support vector regression based electricity peak load forecasting," 2018 11th International Symposium on Mechatronics and its Applications (ISMA), 2018 DOI: DOI.ORG/10.1109/ISMA.2018.8330143

[14] T.-Y. Kim and S.-B. Cho, "Predicting Residential Energy Consumption Using CNN-LSTM Neural Networks," Energy, vol.182, pp.72–81, 2019, DOI: DOI.ORG/10.1016/j.energy.2019.05.230

[15] K. Yan, X. Zhou, and J. Chen, "Collaborative Deep Learning Framework on IoT Data with Bidirectional NLSTM Neural Networks for Energy Consumption Forecasting," Journal of Parallel and Distributed Computing, vol.163, pp.248–255, 2022, DOI: DOI.ORG/10.1016/j.jpdc.2022.01.012

[16] T. Hu, H. Ma, H. Liu, H. Sun, and K. Liu, "Self-Attention-Based Machine Theory of Mind for Electric Vehicle Charging Demand Forecast," IEEE Trans. Ind. Informat., vol.18, no.11, pp.8191–8202, 2022, DOI: DOI.ORG/10.1109/TII.2022.3180399

[17] J.-W. Chan and C.-K. Yeo, "Electrical Power Consumption Forecasting with Transformers," 2022 IEEE Electrical Power and Energy Conference (EPEC), pp.255–260, 2023.

[18] T.G. Grandón, J. Schwenzer, T. Steens, and J. Breuing, "Electricity demand forecasting with hybrid statistical and machine learning algorithms: Case study of Ukraine," Applied Energy, arXiv:2304.05174, 2023.

[19] C. Wang, Y. Wang, Z. Ding, T. Zheng, J. Hu, and K. Zhang, "A Transformer-Based Method of Multienergy Load Forecasting in Integrated Energy System," IEEE Trans. Smart Grid, vol.13, no.4, pp.2703–2714, 2022, DOI: DOI.ORG/10.1109/TSG.2022.3166600

[20] S. Chapaloglou, A. Nesiadis, P. Iliadis, K. Atsonios, N. Nikolopoulos, P. Grammelis, C. Yiakopoulos, I. Antoniadis, and E. Kakaras, "Smart energy management algorithm for load smoothing and peak shaving based on load forecasting of an island's power system," Applied Energy, vol.238, pp.627–642, 2019, DOI: DOI.ORG/10.1016/j.apenergy.2019.01.102

[21] D. Syed, H. Abu-Rub, A. Ghrayeb, S.S. Refaat, M. Houchati, O. Bouhali, and S. Banales, "Deep Learning-Based Short-Term Load Forecasting Approach in Smart Grid With Clustering and Consumption Pattern Recognition," IEEE Access, vol.9, pp.54992–55008, 2021, DOI: doi.org/10.1109/ACCESS.2021.3071654

[22] Z. Chang, Y. Zhang, and W. Chen, "Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform," Energy, vol.187, 115804, 2019, DOI: DOI.ORG/10.1016/j.energy.2019.07.134

[23] "Taiwan Electric Power Company for Past electricity supply and demand information," website: https://data.gov.tw/en/datasets/19995.

Ying-Chang Hung is currently a doctoral student at the Institute of Information Management, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, R.O.C. He holds a keen interest in various domains of wireless communication systems, network management systems, power application management systems, system integration, and project management. In addition to his academic pursuits, he has achieved a Project Management Professional (PMP) certificate, further highlighting his dedication and proficiency in managing complex projects.

Duen-Ren Liu is a Professor of the Institute of Information Management, National Yang Ming Chiao Tung University, HsinChu, Taiwan, R.O.C. Professor Liu's major research interests include data mining, knowledge engineering, electronic commerce, and recommender systems.