Deep Multi-Task Learning for Energy Consumption Forecasting of Household Water Heater Usage

1st Junichiro Niimi
Faculty of Business Management
Meijo University
Aichi, Japan
0000-0002-4618-6272

2nd Takahiro Tsukamoto
School of Economics
Chukyo University
Aichi, Japan
tsukamoto@mecl.chukyo-u.ac.jp

3rd Makito Takeuchi
Graduate School of Economics
Nagoya City University
Aichi, Japan
0000-0002-2041-6770

4th Atomu Shibata
Innovation Center
Rinnai Corp.
Aichi, Japan
atomushibata@rinnai.co.jp

***Abstract*—With electricity market deregulation and the diversification of electricity consumption patterns, precise forecasting of electric demand is crucial for optimal power generation planning and load balancing. However, the increasing complexity of the usage patterns makes the accurate appliance-level prediction more difficult. Water heaters, for example, exhibit distinctive consumption patterns: high electricity usage during active heating periods alternates with complete inactivity (0W), resulting in zero-inflated and high-variance distributions, which are difficult to handle by conventional prediction methods. This study therefore proposes a deep multi-task learning (MTL) framework that simultaneously predicts binary classification (consumed/not consumed) and regression (the amount of consumption) to hierarchically correct the prediction errors for household water heater electricity prediction. We adopt Uncertainty Weighting (UW) which dynamically optimizes heterogeneous loss functions. Using IoT logs with 180,805 samples collected from water heaters in 787 households from January 2022 to May 2024, we demonstrate significant performance improvements over major baselines in RMSE reduction: -22.492% (single-task MLP), -3.962% (Transformer), and -4.223% (LSTM). In addition, the proposed model showed strong generalizability compared to time-series Transformer (PatchTST).**

***Index Terms*—deep learning, multitask learning, electric consumption, energy consumption.**

I. Introduction

In recent years, electricity market deregulation with the unbundling of generation and transmission systems has proceeded in Japan, which increases the importance of electricity demand forecasting from the perspective of load balancing. Additionally, the diversification of household appliances, such as hybrid water heaters that utilize both gas and electricity for water and space heating, further complicates demand prediction, necessitating the development of more precise prediction models.

With the recent advancement in machine learning, particularly deep learning [1], various methods have been proposed in energy prediction [2], [3], [4], [5]. While deep learning methods achieve higher accuracy with nonlinear activations and hierarchical structure compared to traditional models, there remains challenges in applicability to high-frequency Internet-of-Things (IoT) data. Thus, further model development is necessary for industrial application. In particular, for 30-minute intervals, the electric consumption of water heater often becomes zero-inflated distribution. Preliminary analysis in this study revealed that the model struggles to accurately predict consumption of 0W even when no power is being consumed.

Therefore, in this study, with the IoT log which is collected by the hybrid water heater of gas and electricity, we propose a deep learning model for zero-inflated time-series data to predict both the classification whether the electricity is consumed and the regression how much the electricity is consumed at the same time and corrects the prediction. The remainder of the paper is organized as follows: Section 2 summarizes related research, Section 3 describes the proposed model structure, Section 4 summarizes the analysis and results, and Section 5 concludes with the findings and challenges of this study.

This research is supported by JSPS KAKENHI (Grant No. JP23K12481, JP24K16472, and JP25K00678).

II. Related Study

A. Energy Forecasting

Time-series prediction has been addressed through three major paradigms. First, statistical modeling methods, including autoregression (AR), ARIMA, and ARIMAX [6], provide classical approaches grounded in stochastic processes. Second, machine learning methods, such as support vector regression (SVR) [7], random forest (RF) [8], LightGBM [9], and XGBoost [10], leverage non-linear mappings and ensemble learning. Third, deep learning architectures, including multi-layer perceptron (MLP) [11], Long Short-Term Memory (LSTM) networks [12], and Transformer [13], have become increasingly prominent. Among these, Transformer-related architectures have demonstrated remarkable effectiveness in time-series forecasting [14]. Representative variants include Informer [15], Autoformer [16], and PatchTST [17].

Regarding energy prediction, previous studies have primarily focused on area-level [18], household-level [2], [3], [4], and commercial building [19] electricity consumptions. Many studies adopt machine learning and deep learning architectures [20], including LSTM-based [2], CNN-based [21], and Transformer-based [18] approaches.

However, an appliance-level electricity consumption in household, which has yet to be well explored, exhibits fundamentally different characteristics. For instance, some studies [22], [23] using MLP and LSTM pointed out that

there remains challenges to capture the indirect relationship between customer-specific behaviors and exogenous features (e.g., temperatures, energy price). Thus, unlike household-level consumption that maintains relatively continuous power draw, appliance-level prediction is more extraneous. In case of water heater, electricity usage follows a highly intermittent pattern: periods of high consumption during boiling and 0W consumption during inactive state, resulting in zero-inflated and high-variance distributions (cf. Dataset Overview).

The actual zero-inflated distribution can be expressed as the mixture distribution:

$$P(Y = y) = \pi \mathbb{I}(y = 0) + (1 - \pi) P(Y = y | Y > 0) \mathbb{I}(y > 0) \tag{1}$$

where π indicates $P(Y = 0)$. In this study, the parameter $\pi = 0.731$ in a training set, whose nature poses significant challenges for conventional prediction methods. While previous studies have addressed variability and non-linearity in energy consumption, to the best of our knowledge, zero-inflated regression models (e.g., zero-inflated Gamma distribution) have yet to be applied to appliance-level electricity prediction. Thus, this study addresses a distinct challenge in dealing with such complex distributions.

B. Multi-Task Learning

In recent years, multi-task learning (MTL) [24] which simultaneously optimizes the multiple tasks has been developed. In MTL, learning shared representations for multiple tasks can improve the expressiveness of the model and contributes on the model robustness [25]. However, in many cases, since the loss distributions between tasks differ significantly, the design of a loss function across the entire model becomes a significant challenge for the conventional techniques, such as weighted averaging [24] and online regularization [26].

To address these challenges, uncertainty weighting (UW) [25] has been proposed, which learns the variance of loss functions across multiple tasks as uncertainty within the model. In the original UW paper [25], authors conducted an empirical study by internally dividing an image recognition into semantic, instance, and depth tasks, optimizing each loss function, and then integrating their task-level outputs to determine the final output and demonstrated that the model outperformed the aaaaa bbb acccc.

III. Proposed Model

A. Basic Structure

As shown in Fig 1, the proposed model adopts MLP, the typical form of feed-forward neural network with L number of hidden layers. In addition, to address the prediction of exact 0W usage, we construct the hierarchical model which predicts both the probability and amount of energy consumption. More specifically, we set up Task I (classification task) and Task II (regression task) inside the hidden layer (task layer) and calculate the final output value by multiplying the two task-level outputs (output layer).

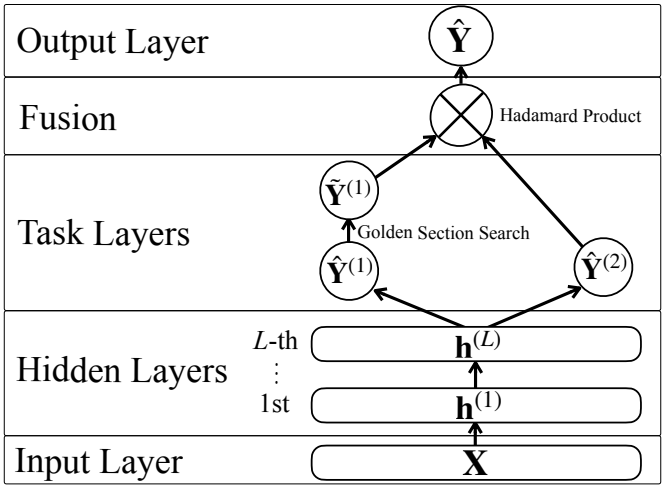


Fig. 1. **Model Architecture**. The proposed model has a MLP architecture with two tasks: Y1 (whether the electricity is consumed) and Y2 (how much electricity is consumed).

In this study, we use three hidden layers ($L = 3$), hidden layer $l = \{0, 1, 2, 3\}$ including input layer $l = 0$. We use Gaussian error unit (GELU) [27] in hidden layers. GELU conducts probabilistic normalization, which is widely adopted in the deep learning models including Transformer. The feature map $\mathbf{h}_l \in \mathbb{R}^{bs \times dim_l}$, denoted bs for batchsize and dim_l for number of dimensions in hidden layer l. The state of the unit can be obtained as:

$$\mathbf{h}^{(l)} = \phi_l(\mathbf{h}^{(l-1)} \mathbf{W}_l + \mathbf{b}_l) \tag{2}$$

where $\mathbf{W}_l \in \mathbb{R}^{dim_{l-1} \times dim_l}$ for weights, $\mathbf{b}_l \in \mathbb{R}^{1 \times dim_l}$ for bias, and ϕ_l for activation function. Input layer $\mathbf{h}_0 = \mathbf{X} \in \mathbb{R}^{bs \times dim_0}$ where dim_0 indicates the number of input variables. In this study, the actual number of dimensions in hidden layers are $[dim_l]_{l=1}^{3} = [512, 256, 128]$.

From the structure above, $\mathbf{h}^{(3)}$ is the shared representation for the each task. Hereinafter, following UW paper [25], task-level outputs are denoted as $\mathbf{Y}^{(1)} \in \{0, 1\}^{bs \times 1}$ and $\mathbf{Y}^{(2)} \in \mathbb{R}^{bs \times 1}$, respectively.

B. Task I: Classification

First, for Task I, we predict whether the household i in time t consume the electricity or not. Combining the shared representation $\mathbf{h}^{(3)}$, task-specific parameters $\theta^{(1)} = (\mathbf{W}^{(1)}, \mathbf{b}^{(1)})$, and sigmoid function $sigm(u) = 1/(1 + e^{-u})$ for the activation, the predicted values are obtained as follows:

$$\hat{\mathbf{Y}}^{(1)} = sigm(\mathbf{h}^{(3)} \mathbf{W}^{(1)} + \mathbf{b}^{(1)}) \in [0, 1]^{bs \times 1} \tag{3}$$

Although the predicted value in this stage is $[0, 1]^{bs \times 1}$ which is continuous values, we need to quantize it with the threshold thr. We adopt golden-section search (GSS) [28], which effectively optimize an 1-dim continuous parameter, compared to grid-search. In the actual exploration, thr is computed with a train set to maximize F1 score of the

classification and apply it to validation and test sets. Therefore, the quantized matrix of $\hat{\mathbf{Y}}^{(1)}$ as

$$\tilde{\mathbf{Y}}^{(1)} = \mathbb{I}\left(\hat{\mathbf{Y}}^{(1)} > thr\right) \in \{0,1\}^{bs\times 1} \tag{4}$$

C. Task II: Regression

Second, for Task II, we predict the amount of consumption $\hat{\mathbf{Y}}^{(2)}$. As well as Task I, we create task-level output using rectified linear unit (ReLU) [29] $relu(u) = \max(0, u)$ for the activation. The predicted value is obtained as follows:

$$\hat{\mathbf{Y}}^{(2)} = relu(\mathbf{h}^{(3)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)}) \in [0, +\infty)^{bs\times 1} \tag{5}$$

D. Output Layer

Finally, we combine two task-level outputs $\tilde{\mathbf{Y}}^{(1)}, \hat{\mathbf{Y}}^{(2)}$. The final output is obtained with the multiplication of each element of the outputs as

$$\hat{\mathbf{Y}} = \tilde{\mathbf{Y}}^{(1)} \odot \hat{\mathbf{Y}}^{(2)} \in [0, +\infty)^{bs\times 1} \tag{6}$$

where $\odot$ indicates Hadamard product. By using this structure, we can strictly predict 0W, compared to the direct prediction of the amount of consumption.

IV. Experiments

A. Dataset Overview

In this study, to construct a time-series prediction model, we divide one day into 48 units and predict electric consumption in the time unit $t \in \{1, 2, \cdots, 48\}$ in day $d \in \{31, 32, \cdots, D_k\}$ for each household $k \in \{1, 2, \cdots, K\}$[1].

We utilize IoT logs collected from water heaters installed in individual households from January 2022 to May 2024. Our prediction task involves forecasting electricity consumption across 48 time slots per day (30-minute intervals) based on the time-series data from the preceding 30 days, requiring a minimum data collection period of 31 days per sample. Given that water heater installation dates vary across households, the duration of available log data differs substantially between households. To mitigate potential overfitting to households with extended data collection periods, we implement a balanced sampling strategy. We randomly extract up to 50 samples for each household and exclude the anomaly samples, including outliers stemming from the defects of the machines and churn in the period.

The dataset[2] finally has 787 households ($K = 787$) with sample size $n = 180,805$. To prevent data leakage, we implement a household-stratified data split, randomly extracting the households and ensuring that all samples from each household were assigned exclusively to either the training (70%), validation (15%), or test (15%) set.

As shown in Table II, both electricity and gas usages have the significantly zero-inflated and high-volume distributions. In addition, even in the same household, the usage pattern drastically changes depending on the seasons. Therefore, to precisely forecast the usage of the electricity, it is crucial to address such zero-inflated and high-variance fitting method with the high generalizability.

Based on the information available in IoT logs, we extract key features from the obtained samples, as shown in Table I. Regarding the temperature-related features, since our logs do not contain the location information of each device, we employ Tokyo meteorological data as a proxy, sourcing previous-day weather forecasts from open datasets [30]. Historical consumption and operational duration features incorporated both same-frame values and 3-frame windowed averages (comprising the target frame and two neighboring frames) across the preceding 30-day period, thereby accommodating potential temporal shifts in routine usage behaviors. The resulting dataset constitutes a multi-household multivariate time series with 657 input dimensions and a temporal depth of 30 days.

B. Optimization

First, we set up two task-level loss functions with binary cross-entropy (BCE) and log mean square error (log-MSE) as follows:

$$\mathcal{L}^{(1)} = -\sum_{i=1}^{bs}\Big[y_i^{(1)}\log(\hat{y}_i^{(1)}) + (1-y_i^{(1)})\log(1-\hat{y}_i^{(1)})\Big] \tag{7}$$

$$\mathcal{L}^{(2)} = \frac{1}{bs}\sum_{i=1}^{bs}\Big[\log(\hat{y}_i^{(2)}) - \log(y_i^{(2)})\Big]^2 \tag{8}$$

where we assume mini-batch learning and calculate the losses for each batch.

In general deep learning, MSE is often used for the evaluation metrics; however, in the power usage of water heater, there is significant variation in consumption, and to take into account cases where sudden large consumption close to outliers occur, we employ log-MSE to calculate the loss for Task II.

As described, since the scales of these task-level loss functions significantly differ and the scale will also change in accordance with the training process, we dynamically weight the change of the scale between tasks and epochs. Therefore, we introduce UW using the variances of each losses (σ_1^2, σ_2^2), the loss function for the overall model is set as follows:

$$\mathcal{L} = \sum_{m=1}^{2}\Big[\frac{1}{2\sigma_m^2}\mathcal{L}^{(m)} + \log(\sigma_m^2)\Big] \tag{9}$$

UW utilizes the uncertainty parameters $(log(\sigma_1^2), log(\sigma_2^2))$ based on the task-level variances and these parameters themselves are also renewed with stochastic gradient descent [31] during the training process. This mechanism effectively optimized the multiple loss functions with different scales simultaneously.

For the optimizer, this study adopts AdaMax [32] which has strength in sparse modeling. The regularization method is not employed. Weights and biases are initialized with the uniform distribution and zero-filling, respectively. As for other hyper-parameters, batch size $bs = 256$, a number of epochs

[1]Since the model predicts the usage amount based on IoT logs obtained in the past 30 days, the day indicator d starts from 31.

[2]For example, if samples were obtained from 3 households for 7 days each with 48 frames per day, the sample size would be $3 \times 7 \times 48 = 1008$ without any outliers.

TABLE I
VARIABLE DESCRIPTION

Names	Description
Date, Time	year, month, week, day-of-week, time frame
Temperature	Forecast values from the previous day (min/max temperature, daily temperature range, day-to-day temperature difference)
Past Consumption 1	Electricity/gas consumption at the same time frame over the past 30 days
Past Consumption 2	Average electricity/gas consumption across adjacent 3-frame windows over the past 30 days
Usage duration	Log data both for same time frame and adjacent 3-frame averages (remote control on-time, bathtub filling time, combustion time, boiling time, heat retention time)

TABLE II
SUMMARY STATISTICS FOR ENERGY CONSUMPTION IN 30MIN INTERVAL

Features	Mean	Std	Median
Electricity (W)	34.757	76.987	0.000
Gas ($0.001m^3$)	74.177	32.422	0.000

is 300, and the best model is evaluated with the checkpoint when the highest accuracy is obtained with the validation set. The model performance is calculated with root mean square error (RMSE) for a test set.

C. Baselines

To validate the effectiveness of the proposed model, we set up several baselines. First, we construct a typical single-task MLP which directly predicts $\mathbf{Y}^{(2)}$ as an output. Second, we employ LSTM networks [4] with a 2-layer Stacked LSTM architecture with serial connections, followed by feedforward hidden layers for output generation. Third, we construct Transformer-based model with the Transformer encoder (model dimension $dim_{model} = 128$, attention heads $num_{head} = 8$) followed by two feedforward hidden layers of sizes $[128, 64]$. In addition, we also implement PatchTST for the advanced time-series Transformer baselines. For machine learning-based baselines, XGBoost, LightGBM, and Linear SVR are adopted. Finally, we employ ARIMAX [6] as a statistical time series method, representing an autoregressive model with exogenous variables. Moreover, we implement zero-inflated Gamma GLM, which is the hierarchal model to predict the target variable $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$, and $\mathbf{Y}$ with Logistic regression, Gamma regression, and multiplication, respectively.

D. Results

1) Overall Performance: Table III shows the results, indicating the relative scale in RMSE based on the MLP. First, the proposed model outperforms all the baselines in both validation and test sets.

2) Effect of MTL: Compare between the proposed model and Single-Task MLP, the proposed model significantly reduces the prediction error by -22.492% in the test set. This improvement demonstrates that, even though the model structure becomes more complex, the hierarchical correction mechanism plays a crucial role in handling zero-inflated distributions.

TABLE III
RESULTS (LIFT, VALIDATION-BEST)

Model	Train	Val	Test
Multi-Task MLP (Proposed)	0.606	**0.658**	**0.703**
Transformer	0.604	0.732	0.732
LSTM	0.797	0.699	0.734
GLM (Zero-Inflated Gamma)	0.742	0.810	0.870
LightGBM	0.276	0.912	0.897
Single-Task MLP	1.000	0.907	0.907
PatchTST	0.609	0.957	0.945
XGBoost	**0.072**	0.967	0.960
ARIMAX	0.898	1.204	1.209
SVR	1.604	1.915	1.682

The classification performance of Task I is shown in Table IV. The cross entropy demonstrates that Task I effectively distinguishes between consumption and non-consumption periods, with reasonable generalization performance (test CE: 0.469). The optimal threshold thr calculated from a training set was 0.618, indicating that approximately 62% confidence is required to predict positive consumption.

This classification accuracy directly contributes to the RMSE improvement in two ways: (1) accurate zero-detection prevents false positive predictions during non-consumption periods, and (2) the binary classification task provides regularization effects that improve the regression task's performance. The substantial improvement validates our hypothesis that zero-inflated time-series prediction benefits from explicit modeling of the binary consumption decision process.

TABLE IV
PERFORMANCE OF TASK I (CLASSIFICATION)

Metrics	Train	Val	Test
Cross Entropy	0.358	0.443	0.469

3) Comparison with Transformer and LSTM: Both Transformer and LSTM basically outperform the other baselines on validation and test set, consistent with the previous studies [4], [2]. However, the proposed model achieves further improvements with RMSE reductions of -3.962% and -4.223%, respectively.

This performance relationship suggests that, while attention mechanisms (Transformer) and memory cells (LSTM)

effectively capture temporal dependencies, they struggle with the zero-inflated nature of water heater consumption. In our effective multi-task approach, the classification head addresses the zero-inflated nature while the regression head addresses the high-variance nonlinear nature.

Regarding PatchTST, despite comparable performance to the proposed model, it exhibits clear overfitting tendency on validation and test sets. This behavior highlights a critical limitation of complex Transformer architectures when applied to zero-inflated appliance-level data, where the challenge lies not in capturing long-range dependencies but in accurately modeling the intermittent consumption patterns. The superior generalization of our proposed model compared to PatchTST demonstrates that task-specific architectural design (multi-task learning for zero-inflation) is more effective than simply scaling model complexity for this particular prediction problem.

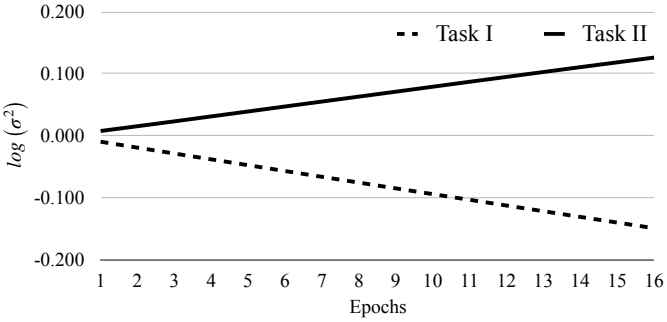
4) Comparison with Other baselines: For other baselines, we adopt XGBoost, LightGBM, Linear SVM, and ARIMAX. Boosting methods, such as XGBoost and LightGBM tends to be high performance in a training set while they underperform in a test set. ARIMAX, as a linear auto-regressive model, naturally underperforms compared to the Single-Task MLP due to its limited capacity to capture non-linear relationships in the data. However, the performance of Linear SVM is significantly lower than even ARIMAX across all metrics. This suggests that temporal dependencies play a crucial role in water heater electricity prediction. While ARIMAX explicitly models autoregressive structure and incorporates external variables (e.g., temperature, day-of-week, seasons), Linear SVM treats each time point independently despite receiving 30 days of historical features. Consequently, Linear SVM fails to capture the inherent temporal correlations and periodic usage patterns (daily, weekly, monthly cycles) that are essential for accurate electricity consumption forecasting.

Interestingly, the statistically principled Zero-Inflated Gamma GLM outperforms several sophisticated deep learning architectures, highlighting the importance of incorporating domain knowledge about the zero-inflated nature of the data rather than relying solely on model complexity.

5) Generalizability: The proposed model has the best performance in both validation and test sets. Even compared with the Transformer, the discrepancy between the train and test performance is smaller in the proposed model. Even though LightGBM and XGBoost has significant high performances in training set, low generalizability of these boosting-based models are compliant with the previous study [33], [34]. Moreover, as described, PatchTST has poor generalization to the hold-out samples, including validation and test. These indicate that the proposed model has a well-balanced generalizability among the baselines.

6) Effect of Uncertainty Weighting: Finally, Fig. 2 shows the change of the uncertainty weights in the training process. As can be seen from these results, the degree of uncertainty in each task clearly changes as learning progresses. In particular, the importance of the regression task increases relatively as learning progresses. These results suggests that uncertainty weighting which dynamically adjusts the importance of the multiple tasks is shown to be effective for the multi-task learning between different loss functions.

Fig. 2. Change of Uncertainty



V. Conclusion

A. Summary

This study implements deep multi-task learning model to predict the appliance-level electricity consumption of the water heater in each household. To address distinct challenges of zero-inflated and high-variance prediction on the time-series data, we simultaneously optimize both classification and regression using Uncertainty Weighting and correct the final output in 30-min intervals. Through the exhaustive analysis, the proposed model which concurrently predicts whether the electricity is consumed and the amount of the consumption shows the best performance, compared with the deep learning architectures, machine learning methods, and time-series statistical models.

B. Theoretical Implications

This study has several theoretical implications. First, as far as authors know, this study is the first application of UW to zero-inflated time-series prediction. In particular, zero-inflated modeling requires both strict binary classification and precise regression, which is suitable task for UW. This approach can be extended to other zero-inflated time-series applications such as intermittent demand forecasting, fault detection, and renewable energy generation. Second, we demonstrate UW effectiveness between BCE and log-MSE which are heterogeneous loss functions. Although the original UW paper [25] optimized three loss functions, it did not handle loss functions with properties as different as those in this study. Also, the model shows the superiority in generalization. In particular, even with Transformer, performance was almost equivalent to the proposed model during training, but performance declined significantly during testing. Most of the sophisticated architectures like PatchTST struggles with the hold-out samples, indicating that to address zero-inflated distribution is critical issue for the model complexity.

C. Practical Implications

This study offers several practical implications for industry applications. First, accurate predictions of zero-inflated electricity usage enable utilities to optimize peak load management strategies, leading to improved demand response program effectiveness. Second, unlike the Transformer-based architectures that require extensive embedding preprocessing, the proposed model enables direct integration with the customer database. Finally, the model's capability for high-frequency IoT data extends beyond energy management to broader applications. The real-time analysis of household consumption patterns provides valuable insights for lifestyle-based marketing, such as offering optimal price plan recommendations and sending personalized energy-saving notifications.

D. Limitation

Finally, we organize the limitations of our study. First, although we have collected IoT data over two years, the experiment adopts 48 frames $\times$ 30 days $=$ 1,440 steps for the actual sequences. While PatchTST significantly underperforms the proposed model, Transformer-based time-series models are basically built for extremely long sequences (e.g., 10,000 steps), which may not be optimal for our experimental design. Therefore, future work should investigate the effectiveness of the proposed model with the longer steps. Second, this study does not collect the personalized information, such as the size, constitutions, location information, and contract plans with electric power companies of each household. These information facilitate the personalized predictions.

ACKNOWLEDGMENT

The dataset was provided by Rinnai Corporation. J.N., T.T., and M.T. serve as technical advisors of Rinnai Corporation.

REFERENCES

[1] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[2] J. F. Torres, F. Martínez-Álvarez, and A. Troncoso, "A deep lstm network for the spanish electricity consumption forecasting," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10 533–10 545, 2022.

[3] M. Frikha, K. Taouil, A. Fakhfakh, and F. Derbel, "Predicting power consumption using deep learning with stationary wavelet," *Forecasting*, vol. 6, no. 3, pp. 864–884, 2024.

[4] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[5] C.-H. Liu, J.-C. Gu, and M.-T. Yang, "A simplified lstm neural networks for one day-ahead solar power forecasting," *Ieee Access*, vol. 9, pp. 17 174–17 195, 2021.

[6] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[7] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems (NIPS2017)*, vol. 30, 2017.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation, In parallel distributed processing, explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT Press, 1986, pp. 318–362.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[14] C. Wang, H. Wang, X. Zhang, Q. Liu, M. Liu, and G. Xu, "A transformer-based industrial time series prediction model with multivariate dynamic embedding," *IEEE Transactions on Industrial Informatics*, 2024.

[15] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[16] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[17] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *The Eleventh International Conference on Learning Representations*, 2023.

[18] A. L'Heureux, K. Grolinger, and M. A. Capretz, "Transformer-based model for electrical load forecasting," *Energies*, vol. 15, no. 14, p. 4993, 2022.

[19] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.

[20] I. K. Nti, M. Teimeh, O. Nyarko-Boateng, and A. F. Adekoya, "Electricity load forecasting: a systematic review," *Journal of Electrical Systems and Information Technology*, vol. 7, no. 1, p. 13, 2020.

[21] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho, and S. W. Baik, "Improving electric energy consumption prediction using cnn and bi-lstm," *Applied Sciences*, vol. 9, no. 20, p. 4237, 2019.

[22] S. Welikala, C. Dinesh, M. P. B. Ekanayake, R. I. Godaliyadda, and J. Ekanayake, "Incorporating appliance usage patterns for non-intrusive load monitoring and load forecasting," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 448–461, 2017.

[23] C. Yang, H. Zhou, X. Chen, and J. Huang, "Demand time series prediction of stacked long short-term memory electric vehicle charging stations based on fused attention mechanism," *Energies*, vol. 17, no. 9, p. 2041, 2024.

[24] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[25] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[26] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 109–117.

[27] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[28] J. Kiefer, "Sequential minimax search for a maximum," *Proceedings of the American mathematical society*, vol. 4, no. 3, pp. 502–506, 1953.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[30] J. Niimi, "7days weather forecast in japan (dataset)," 2024, hugging Face Datasets (huggingface.co/datasets/jniimi/weather_forecast_japan).

[31] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[34] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.