Data-Efficient Electricity Consumption Forecasting with a Tabular Foundation Model

Seeun Kim
Dept. Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea
sese2121@seoultech.ac.kr

Jungmin Lim
Dept. Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea
ijm0521@seoultech.ac.kr

Jiyoon Byun
Dept. Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea
jiyoon8676@seoultech.ac.kr

Jeongyeon Kim
Dept. Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea
jy_kim@seoultech.ac.kr

Hanul Kim
Dept. Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul, Republic of Korea
hukim@seoultech.ac.kr

***Abstract*—We address building-level electricity consumption forecasting under limited data. We introduce a simple pipeline that couples feature engineering with an electricity consumption predictor comprising a tree-based model and a tabular foundation model. A lightweight meta-learner aggregates their outputs to exploit complementary strengths of these models. We assess the effectiveness of the proposed approach on the KEA-2025 dataset. Experimental results on varying training-history lengths demonstrate that the foundation model excels in low-data regimes, the tree-based model improves as data grow, and the ensemble consistently achieves the best performance.**

***Index Terms*—Electricity consumption prediction, Tabular foundation model, gradient boosting decision tree.**

I. Introduction

Electricity demand has continued to rise [1]. Accurate consumption forecasting becomes important to optimize energy resource management [2]. Because electricity use exhibits persistent temporal structure, many methods aim to model these time-series patterns. Autoregressive integrated moving average (ARIMA) [3] is a prototypical example [4]. However, such statistical models struggle with the complex, nonlinear dynamics of electricity consumption.

To address these limitations, many studies [5]–[7] have explored machine learning (ML) models. Tree-based models including random forest [8] and XGBoost [9] effectively capture nonlinear, multivariate relationships between input variables and consumption, leading to promising performance over statistical baselines. However, these methods typically treat observations as independent, overlooking the sequential structure and long-range dependencies essential for accurate forecasting.

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Goverment under Grants RS-2023-00221365 and RS-2024-00352566.

With advances in deep learning (DL), numerous methods [10]–[16] have leveraged neural architectures, such as temporal convolutional networks (TCNs) [15] and long short-term memory (LSTM) [16], to model long-range temporal dependencies in forecasting applications. However, DL-based methods often require substantial training data and careful hyperparameter tuning to avoid overfitting, which can limit their practicality in data-scarce settings.

Recently, the success of foundation models in natural language processing (NLP) [17], [18] and computer vision (CV) [19], [20] has highlighted strong in-context learning, enabling adaptation to new tasks at inference time without parameter updates. Building on this progress, recent studies have developed foundation models tailored to tabular [21] and time-series [22]–[24] data, demonstrating robust forecasting in data-scarce settings. However, despite the potential of foundation models, their application to electricity consumption forecasting remains underexplored.

In this work, we address building-level electricity consumption prediction in data-scarce regimes. First, we perform feature engineering to extract derived features that are closely related to electricity consumption. Second, we present a simple, data-efficient pipeline that combines a gradient-boosted tree (XGBoost) with a tabular foundation model (TabPFN). We then train a shallow meta-learner to fuse their outputs. The approach requires no fine-tuning of the foundation model for data-scare scenario. Experiments on the Korea Energy Agency's 2025 Electricity Consumption Prediction (KEA-2025) dataset confirm the effectiveness of our approach. Empirically, TabPFN is strongest with short histories, XGBoost improves as history length grows, and the meta-learner's predictions achieve the lowest error across diverse data conditions.

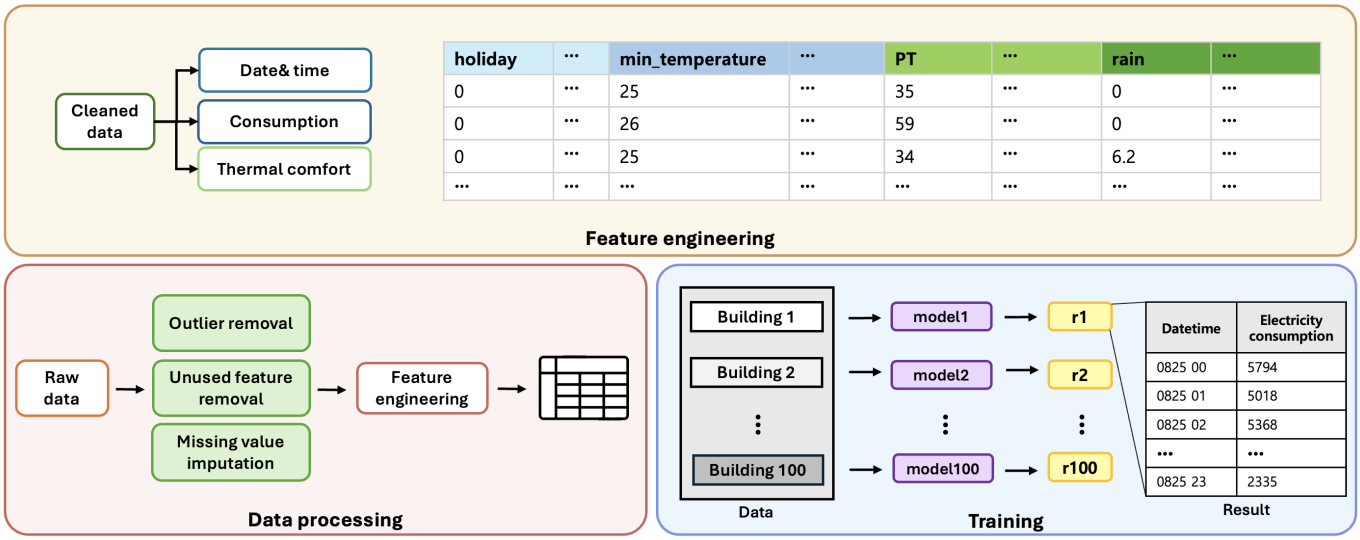

Fig. 1. An overview of the proposed pipeline consists of three stages: (1) data processing, which converts raw building records into training-ready data, (2) feature engineering, which generates derived variables, and (3) an ensemble model of a gradient boosting decision tree (XGBoost) and a tabular foundation model (TabPFN), which fits a per-building predictor for electricity consumption. The final outputs are building-wise forecasts $r_1, \ldots, r_{100}$.

II. Related Work

A. Learning-based Electricity Consumption Forecasting

In [4], ARIMA [3] is combined with clustering to forecast electricity consumption in university buildings. Because ARIMA assumes linearity under stationarity, it struggles to capture the complex patterns common in practice. In [5], XGBoost [25] shows promising results for electricity consumption forecasting in a university building. EA-XGBoost [25], which integrates empirical mode decomposition, ARIMA, and XGBoost, outperforms standalone ARIMA and XGBoost. A review of electricity consumption forecasting studies [26] reported that no single learning-based method outperforms all others across scenarios.

B. Deep Learning Models for Time Series Analysis

Time series analysis is a fundamental problem that includes industrial forecasting [27] and other temporal tasks [28]–[30] across various domains. So, numerous approaches [15], [27], [31]–[39] have been explored. Recurrent neural networks(RNNs) [27], [31], [32] are commonly used to capture long-term dependencies but suffer from limited parallelism and difficulty with very long sequences. Temporal convolutional neural networks [15] and its variants [33]–[36] provide large receptive fields with improved computational efficiency. Inspired by the success of Transformers across modalities [40]–[43], recent studies [37]–[39] adopt attention mechanism to model long-range dependencies in parallel.

C. Tabular Foundation Models

Foundation models such as GPT3 [17] have gained significant attention across domains. Leveraging priors learned from large-scale data, they enable in-context learning that adapts models to new tasks at inference time without parameter updates. In the tabular setting, TabPFN [21] learns a prior-data–fitted Transformer that approximates Bayesian inference over synthetic task distributions, supporting zero-/few-shot transfer. This paradigm has been extended to forecasting via in-context conditioning [21] and to broader time-series foundation models pretrained on heterogeneous temporal corpora [22].

III. Dataset

We use the Korea Energy Agency's 2025 Electricity Consumption Prediction dataset (KEA-2025) to develop our forecasting models. The dataset contains hourly electricity consumption, meteorological variables, and building metadata for 100 buildings collected from June 1 to August 24, 2024. In this work, we use the first 78 days of data for training and the remainder for testing.

Meteorological variables. For each building, the meteorological record comprises date–time stamp, temperature ($^\circ$C), precipitation (mm), wind speed ($\mathrm{m/s}$), humidity (%), sunshine duration (hr), and solar radiation ($\mathrm{MJ/m^2}$).

Building metadata. Each building record includes seven attributes: building number, building type, floor area ($\mathrm{m^2}$), cooling area ($\mathrm{m^2}$), solar capacity (kW), energy storage system capacity (kWh), and power conversion system capacity (kW). Building types are categorized into ten classes: hotel, commercial, hospital, school, others building, apartment, research institute, department store, IDC (telephone station), and public.

TABLE I

THE SUMMARIZATION OF ENTIRE FEATURE SET FOR BUILDING-LEVEL ELECTRICITY CONSUMPTION FORECASTING, WHICH INCLUDES RAW METEOROLOGICAL AND BUILDING METADATA AND DERIVED VARIABLES THAT ENCODE TEMPORAL REGULARITIES, PER-BUILDING CONSUMPTION STATISTICS, AND THERMAL-COMFORT INDICES (F^{PT}, F^{THI}, F^{THI}_{cls}).

Features	Type	Name	Explanation
Raw	Raw	temperature	Air temperature
		wind speed	Outdoor wind speed
		precipitation, humidity	Rainfall amount and air humidity
		building number	Representation of building id
		building type	Representation of building type(0-9)
		total area, cooling area	Total area of each building and air-conditioned area
Derived	Date-time	month, day, hour	Month, day, and hour extracted from timestamp
		week number	Week of the year
		time sin, time cos	Sine and cosine transformation of hour to capture cyclic nature
		weekend, weekday, holiday	Indicators for weekend, weekday, and holidays
	Statistics	day hour mean	Average electricity consumption by day and hour for each building
		hour mean	Average electricity consumption by hour for each building
		hour std	Standard deviation of electricity consumption by hour for each building
		temperature min	Daily minimum temperature of each building
		temp diff	Difference between current temperature and daily minimum
	Thermal comfort	F^{PT}	Perceived temperature, combining temperature and wind speed
		F^{THI}	Temperature humidity index, reflecting human discomfort
		F^{THI}_{cls}	Categorical variable based on F^{THI} value

IV. METHODOLOGY

A. Feature Engineering

Table I summarizes all raw and derived features used to electricity consumption prediction. First, we encode the date–time stamps into periodic features to capture temporal regularities in electricity demand. Specifically, we derive variables such as day-of-week, weekend/holiday indicators, and sinusoidal encoding.

Next, we compute per-building, hourly summary statistics of electricity consumption on the training set using a leakage-free historical window. These statistics account for building-specific differences in average load and variability. We also aggregate each building's hourly series by day of week to capture weekly cyclic patterns. In addition, we compute, for each building, the daily minimum temperature and the difference between the current temperature and that minimum.

Lastly, we extract features that describe the effects of temperature, humidity, and wind on human thermal sensation, which are factors closely linked to electricity consumption patterns. To this end, we adopt the perceived temperature F^{PT} and the temperature–humidity index F^{THI} as done in [44]. The perceived temperature F^{PT} represents how temperature feels to humans under different wind conditions, given by

$$F^{PT} = 0.6215T + 13.12 + (0.3965T - 11.37)v^{0.16} \quad (1)$$

where v is wind speed in km/h and T is temperature.

The temperature–humidity index F^{THI} quantifies heat stress from temperature and humidity, in which larger values indicate greater thermal discomfort and typically correspond to higher electricity consumption. Specifically, F^{THI} is defined as

$$F^{THI} = 1.8\,T + 32 - 0.55\left(1 - \frac{H}{100}\right)(1.8\,T - 26) \quad (2)$$

where H is humidity. Because human responses to thermal discomfort vary nonlinearly with F^{THI}, we discretize it into four ordinal categories and add their indicators as derived features:

$$F^{THI}_{\text{cls}} = \begin{cases} 0, & \text{if } F^{THI} < 68 \\ 1, & \text{if } 68 \le F^{THI} < 75 \\ 2, & \text{if } 75 \le F^{THI} < 80 \\ 3, & \text{if } 80 \le F^{THI} \end{cases} \quad (3)$$

Here, 0, 1, 2, and 3 correspond to comfortable, mild discomfort, $\sim$50% of people uncomfortable, and strong discomfort for nearly everyone, respectively.

B. Electricity Consumption Predictor

Fig. 1 illustrates the overall pipeline to predict per-building electricity consumption. The predictor comprises a tree-based model, a tabular foundation model, and a meta-learner that combines their outputs.

Tree-based model. We adopt XGBoost [9] as the tree-based regressor due to its strong performance on tabular data. XGBoost is trained to minimize mean absolute error (MAE) between predictions and ground-truth electricity consumption. Hyperparameters are tuned with Optuna over 30 trials using a leakage-free, fixed-length sliding-window validation split. After tuning, the model is refit on the training data with the selected settings and evaluated once on the held-out test set.

TABLE II
SMAPE (LOWER IS BETTER) FOR XGBOOST, TABPFN, AND THEIR STACKED ENSEMBLE ACROSS TRAINING-HISTORY LENGTHS (DAYS). THE BEST RESULTS ARE DEPICTED IN BOLD.

Days	XGBoost	TabPFN	Ensemble
1	15.6461	9.3280	**9.0540**
7	7.5434	7.1573	**7.0087**
21	6.8612	6.8973	**6.6412**
50	6.4158	6.7637	**6.2882**

TABLE III
SMAPE (LOWER IS BETTER) FOR XGBOOST, CATBOOST, AND THEIR ENSEMBLE ACROSS TRAINING-HISTORY LENGTHS (DAYS). THE BEST RESULTS ARE DEPICTED IN BOLD.

Days	XGBoost	Catboost	Ensemble
1	15.6461	15.8146	**15.2139**
7	7.5434	9.3465	**7.4607**
21	6.8612	7.0895	**6.7328**
50	6.4158	6.5556	**6.2537**

Tabular foundation model. We employ TabPFN [21], a prior-data-fitted Transformer trained on a large distribution of synthetic tabular tasks to approximate Bayesian inference. At inference time, it performs in-context learning: given a small, leakage-free conditioning set, it predicts targets for new samples without gradient updates. It consumes the same feature set as XGBoost and produces 7-day-ahead forecasts.

Meta-learner. Let $\hat{y}_{\text{xgb}}$ and $\hat{y}_{\text{pfn}}$ denote the predictions from XGBoost and TabPFN, respectively. XGBoost is trained solely on KEA-2025, whereas TabPFN utilizes generalized prior knowledge from large-scale pretraining. Their predictions are therefore complementary. To exploit this, we fit a lightweight stacking head on predictions from both models. The meta-learner's output $\hat{y}$ takes the form

$$\hat{y} = \alpha\, \hat{y}_{\text{xgb}} + (1-\alpha)\, \hat{y}_{\text{pfn}} + \beta, \quad \alpha \in [0,1] \tag{4}$$

where α and β are trainable parameters initialized to 0.5 and 0.0, respectively.

V. EXPERIMENTS

A. Experimental Settings

Training data. We construct training sequences with history lengths from 1 to 50 days and forecast the subsequent 7 days of electricity consumption. The 1-day set contains 24 hourly observations, whereas the 50-day set contains $1{,}200$.

Evaluation metric. We evaluate performance using symmetric mean absolute percentage error (SMAPE) [45]:

$$\text{SMAPE} = \frac{100}{n} \sum_{t=1}^{n} \frac{2\,|\hat{y}_t - y_t|}{|\hat{y}_t| + |y_t|}, \tag{5}$$

where n is the number of observations, y_t and $\hat{y}_t$ are the ground-truth and predicted values at time t, and $|\cdot|$ denotes absolute value.

B. Experimental Results

Table II compares single models (XGBoost, TabPFN) and their ensemble across training-history lengths. When training data are scarce, TabPFN outperforms XGBoost: with 1 day of history, SMAPE drops from 15.65 (XGBoost) to 9.33 (TabPFN), and with 7 days from 7.54 to 7.16. This advantage reflects TabPFN's strong prior and in-context learning, which stabilize training with small n. As data increase, XGBoost overtakes TabPFN: at 21 days it is slightly better (6.86 vs. 6.90), and at 50 days the gap widens (6.42 vs. 6.76), indicating stronger task-specific fitting with more samples. The ensemble achieves the best performance at all settings, consistently improving over the stronger base model. These results suggest complementary error profiles between the two models.

To analyze the ensembling effect in Table II, we train an additional CatBoost model and ensemble it with XGBoost under the same Optuna-based tuning and retraining protocol. Table III reports the results. Compared to the XGBoost+TabPFN ensemble, the XGBoost+CatBoost ensemble yields smaller gains: with 1 day of history the error remains high (15.21 vs. 9.05), with 7 days it is higher (7.46 vs. 7.01), and with 21 days it is still worse (6.73 vs. 6.64). This weaker effect is expected because XGBoost and CatBoost are both GBDT models and thus share similar inductive biases, leading to highly correlated errors and limited diversity. Notably, at 50 days the XGBoost+CatBoost ensemble slightly outperforms XGBoost+TabPFN (6.25 vs. 6.29), and CatBoost alone also surpasses TabPFN (6.56 vs. 6.76), suggesting that with sufficient data TabPFN's prior advantage diminishes.

VI. CONCLUSION

In this work, we presented a simple and data-efficient pipeline to predict building-level electricity consumption. First, we performed feature engineering to construct strong features for electricity consumption modeling. We then combined two complementary predictors: a tree-based model (XGBoost) and a tabular foundation model (TabPFN) using a lightweight meta-learner. We validated the proposed method on KEA-2025 dataset. Experimental results demonstrate that our ensemble consistently attains the lowest SMAPE across training-history lengths. Analyses describe that TabPFN excels in low-data regimes through strong priors and in-context learning, whereas XGBoost improves as data grow by capturing task-specific patterns; their diversity drives the ensemble gains.

Limitations and future work. Our approach relies on in-context learning rather than explicit sequence modeling. Future work will incorporate recurrent or transformer-based time-series architectures to directly capture long-range temporal dependencies.

REFERENCES

[1] International Energy Agency, "Electricity 2024," IEA, Paris, Tech. Rep., 2024, licence: CC BY 4.0. [Online]. Available: https://www.iea.org/reports/electricity-2024

[2] G. Hafeez, K. S. Alimgeer, A. B. Qazi, I. Khan, M. Usman, F. A. Khan, and Z. Wadud, "A hybrid approach for energy consumption forecasting with a new feature engineering and optimization framework in smart grid," *IEEE Access*, vol. 8, pp. 96 210–96 226, 2020.

[3] O. Anderson and M. Kendall, "Time-series. 2nd edn." *J. R. Stat. Soc. (Series D)*, 1976.

[4] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and arima model for energy management in buildings," *Japan Architectural Review*, vol. 3, no. 1, pp. 62–76, 2020.

[5] J. Barzola-Monteses, F. Parrales-Bravo, V. Macas-Espinosa, G. Bórquez-Vargas, and M. Espinoza-Andaluz, "Energy consumption of a building using the xgboost algorithm: A forecasting study," in *SCCC*. IEEE, 2024, pp. 1–7.

[6] S.-Y. Shin and H.-G. Woo, "Energy consumption forecasting in korea using machine learning algorithms," *Energies*, vol. 15, no. 13, p. 4880, 2022.

[7] Y. Liu, H. Chen, L. Zhang, and Z. Feng, "Enhancing building energy efficiency using a random forest model: A hybrid prediction approach," *Energy Reports*, vol. 7, pp. 5003–5012, 2021.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *SIGKDD*, 2016, pp. 785–794.

[10] A. M. Elshewey, M. M. Jamjoom, and E. H. Alkhammash, "An enhanced cnn with resnet50 and lstm deep learning forecasting model for climate change decision making," *Scientific Reports*, vol. 15, no. 1, p. 14372, 2025.

[11] Y. Li, W. Zhou, Y. Wang, S. Miao, W. Yao, and W. Gao, "Interpretable deep learning framework for hourly solar radiation forecasting based on decomposing multi-scale variations," *Applied Energy*, vol. 377, p. 124409, 2025.

[12] A. Li, J. Li, and Z. Shen, "An efficient modern convolution-based dynamic spatiotemporal deep learning architecture for ozone prediction," *ENSO*, vol. 188, p. 106424, 2025.

[13] H. Pankka, J. Lehtinen, R. J. Ilmoniemi, and T. Roine, "Enhanced eeg forecasting: a probabilistic deep learning approach," *Neural Computation*, vol. 37, no. 4, pp. 793–814, 2025.

[14] S. I. Abir, S. Shoha, M. M. Hossain, N. Sultana, T. R. Saha, M. H. Sarwer, S. I. Saimon, I. Islam, and M. Hasan, "Machine learning and deep learning techniques for eeg-based prediction of psychiatric disorders," *JCSTS*, vol. 7, no. 1, pp. 46–63, 2025.

[15] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *CVPR*, 2017, pp. 156–165.

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

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *CoRR*, 2024.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *CVPR*, 2023, pp. 4015–4026.

[20] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *CVPR*, 2025.

[21] H.-J. Ye, S.-Y. Liu, and W.-L. Chao, "A closer look at tabpfn v2: Strength, limitation, and extension," *arXiv preprint arXiv:2502.17361*, 2025.

[22] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "Moment: A family of open time-series foundation models," *arXiv preprint arXiv:2402.03885*, 2024.

[23] S. Dooley, G. S. Khurana, C. Mohapatra, S. V. Naidu, and C. White, "Forecastpfn: Synthetically-trained zero-shot forecasting," *NeurIPS*, vol. 36, pp. 2403–2426, 2023.

[24] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," in *ICML*, 2024.

[25] W. Yucong and W. Bo, "Research on ea-xgboost hybrid model for building energy prediction," in *Journal of Physics: Conference Series*, vol. 1518, no. 1. IOP Publishing, 2020, p. 012082.

[26] A. Groß, A. Lenders, F. Schwenker, D. A. Braun, and D. Fischer, "Comparison of short-term electrical load forecasting methods for different building types," *Energy Informatics*, vol. 4, no. Suppl 3, p. 13, 2021.

[27] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *IJF*, vol. 36, no. 3, pp. 1181–1191, 2020.

[28] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: Bidirectional recurrent imputation for time series," *NeurIPS*, vol. 31, 2018.

[29] U. Yokkampon, A. Mowshowitz, S. Chumkamon, and E. Hayashi, "Robust unsupervised anomaly detection with variational autoencoder in multivariate time series data," *IEEE Access*, vol. 10, pp. 57 835–57 849, 2022.

[30] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *IJCNN*. IEEE, 2017, pp. 1578–1585.

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NeurIPS*, 2014.

[32] S. Lin, W. Lin, W. Wu, F. Zhao, R. Mo, and H. Zhang, "Segrnn: Segment recurrent neural network for long-term time series forecasting," *arXiv preprint arXiv:2308.11200*, 2023.

[33] D. Luo and X. Wang, "Moderntcn: A modern pure convolution structure for general time series analysis," in *ICLR*, 2024, pp. 1–43.

[34] H. Wang, J. Peng, F. Huang, J. Wang, J. Chen, and Y. Xiao, "Micn: Multi-scale local and global context modeling for long-term series forecasting," in *ICLR*, 2023.

[35] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, and Q. Xu, "Scinet: Time series modeling and forecasting with sample convolution and interaction," *NeurIPS*, vol. 35, pp. 5816–5828, 2022.

[36] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *ICLR*, 2023.

[37] 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 *AAAI*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[38] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *ICML*. PMLR, 2022, pp. 27 268–27 286.

[39] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *ICLR*, 2023.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[42] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28 492–28 518.

[44] J. Moon, S. Rho, and S. W. Baik, "Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with shapley values," *SETA*, vol. 54, p. 102888, 2022.

[45] W. Yan, "Toward automatic time-series forecasting using neural networks," *TNNLS*, vol. 23, no. 7, pp. 1028–1039, 2012.