A Mixture of Experts Method with Coordinated Imputation for Multi-Residential Load Forecasting under Missing Data

Zhenyuan Du, *Graduate Student Member*, *IEEE*, Ziming Yan, *Member*, *IEEE*, and Yan Xu, *Senior Member*, *IEEE*
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
Email: e220188@e.ntu.edu.sg; ziming.yan@ntu.edu.sg; xuyan@ntu.edu.sg

***Abstract*—Residential smart-meter data are often incomplete, and training a dedicated forecasting model for each residence is impractical. In this context, this paper proposes a mixture-of-experts (MoE) framework for residential load forecasting under missing data and large user populations. The proposed MoE comprises multiple imputation–forecasting coordinated experts and a gating neural network. For any load series with missing data, the series is fed in parallel to all experts and the gating network, which assigns routing weights based on electricity consumption patterns rather than residence IDs, thereby enabling pattern-level knowledge sharing across users. Each imputation-forecasting coordinate expert consists of a bidirectional recurrent imputation for time series model (BRITS)-based imputer and a long short-term memory neural network (LSTM)-based forecaster, which is optimized through a coordinated experts' loss function considering both interpolation accuracy and the impact of imputation. The forecasting results of each expert will be weighed according to the weight of gating NN and then aggregated to output. The simulation results in a public residential load dataset verify the effectiveness of the proposed method.**

***Index Terms*—Residential load forecasting, missing data, imputation, mixture of experts (MoE).**

I. INTRODUCTION

The large-scale deployment of smart meters has become a fundamental component of modern smart grids, generating massive volumes of high-resolution residential load data [1]. However, real-world smart meter datasets often contain substantial missing values caused by communication disruptions, measurement faults, and data management issues, which severely hinder downstream analytics such as load forecasting [2]. Meanwhile, most existing forecasting practices still follow a one-meter–one-model paradigm [3]. Although this design accommodates household heterogeneity, it becomes impractical at scale: deploying tens of thousands of independent models leads to prohibitive computational overhead and unsustainable model-management complexity. These challenges highlight the need for a scalable forecasting framework that can effectively handle missing data while eliminating the structural limitations of per-household modeling.

In the literature, time-series missingness has been repeatedly shown to deteriorate the performance of downstream forecasting tasks in both PV generation prediction [4] and load forecasting [5], as incomplete sequences disrupt temporal patterns essential for model learning. Consequently, a substantial body of research has focused on imputing missing smart meter readings using statistical methods, spatial–temporal models, or topology-aware estimators to reconstruct continuous load profiles [6]-[8]. While these approaches improve the visual continuity and short-term consistency of reconstructed signals, they predominantly optimize imputation accuracy itself, rather than the suitability of the imputed data for downstream forecasting. Parallel efforts on robust forecasting [9]-[11] aim to tolerate input corruption or uncertainties. However, despite their widespread adoption, existing two-stage pipelines that separately perform missing-data imputation followed by load forecasting suffer from fundamental limitations. First, imputation errors inevitably propagate to the forecasting stage, where the reconstructed values are treated as ground truth and cannot be corrected. Second, optimizing imputation accuracy alone does not necessarily align with forecasting objectives, as imputations that are visually or statistically plausible may still distort temporal dependencies critical for accurate load prediction. This mismatch between reconstruction-oriented and forecasting-oriented objectives leads to suboptimal performance in downstream forecasting tasks. These challenges motivate the need for a coordinated learning framework, in which imputation and forecasting are jointly optimized to ensure compatibility between reconstructed inputs and prediction targets.

Beyond data quality issues, large-scale deployment of smart meters also raises substantial challenges for model management. Existing studies attempt to mitigate scalability concerns through federated learning [12], transfer learning [13], and meta-learning [14] frameworks. While these approaches effectively address privacy preservation, data scarcity, or cross-task generalization, they primarily focus on improving the training efficiency of multiple models rather than reducing the

The work in this paper was supported in part by Ministry of Education (MOE), Republic of Singapore, under grant AcRF TIER-1 RG59/22.

979-8-3315-8138-1

number of models deployed. Even lightweight forecasting models still require maintaining one model per household, which becomes impractical when thousands of meters are installed. Moreover, under missing-data conditions, most pipelines must additionally deploy a separate imputation model before forecasting, further increasing computational overhead and system-wide maintenance burden. Ideally, a single universal forecasting model capable of serving a large population of heterogeneous households—even with a moderately larger model size-would substantially simplify system deployment while maintaining high prediction accuracy.

To address missing residential load data and the scalability limitations of per-residence forecasting, this paper proposes a unified imputation–forecasting framework based on a mixture-of-experts (MoE) architecture [15]. Instead of building one model per residence, the proposed method exploits data-level electricity consumption patterns rather than residential identities. Specifically, each incomplete load sequence is fed in parallel to a gating neural network and multiple imputation–forecasting experts, and expert outputs are adaptively weighted according to inferred consumption patterns. This enables a single MoE model to serve a large population of heterogeneous households. Each expert integrates a BRITS-based imputer [16] with an LSTM forecaster and is trained end-to-end using a coordinated objective. Unlike conventional two-stage pipelines, the proposed loss jointly optimizes imputation fidelity and forecasting accuracy, guiding the imputer to generate reconstructions that are both consistent with observed data and beneficial for downstream prediction. This coordinated learning improves both forecasting performance and efficiency.

II. Problem Formulation

A. Milti-Residential Load Forecasting via Electricity Consumption Pattern Sharing

The residential load data which collected from N residences is represented by $D=\left\{\mathbf{X}^{(i)}\right\}_{i=1}^{N}$, and:

$$\mathbf{X}^{(i)}=\left(x_1^{(i)},\cdots,x_T^{(i)}\right)\in\mathbb{R}^T \tag{1}$$

where $\mathbf{X}^{(i)}$ and $x_j^{(i)}$ denotes the load series and jth data point of ith residence, T is the length of forecasting window. Although different residences generally exhibit highly distinct full-day or long-horizon load profiles, we posit that each short-term input segment $\mathbf{X}^{(i)}$ is generated from a finite set of latent consumption patterns. Formally, assume the existence of K electricity consumption patterns $\left\{z_k\right\}_{k=1}^{K}$ such that each sample satisfies

$$\mathbf{X}^{(i)}=f_\theta\left(z_{c^{(i)}}\right)+\varepsilon^{(i)} \tag{2}$$

Where f_θ is a nonlinear mapping shared across all residences, $c^{(i)}\in\left\{1,\cdots,K\right\}$ indicates the electricity consumption pattern that produced the sample, and $\varepsilon^{(i)}$ is random noise.

In our modeling in (2), the patterns $\left\{z_k\right\}_{k=1}^{K}$ are shared across the entire population, but the pattern distribution differs across residence. This means that different residences do not create new patterns but reuse the same $\left\{z_k\right\}_{k=1}^{K}$ with different occurrence probabilities and combination, as long as $\left\{z_k\right\}_{k=1}^{K}$ is complex enough. Therefore, the forecasting task can be reformulated as identifying the latent pattern underlying each input segment and forecasting its corresponding response, rather than building one forecasting model per residence.

B. Unified Formulation for Missing-Data Imputation and Forecasting

In conventional two-stage training, the imputation model and forecasting model are optimized independently with separate objectives, implicitly assuming that minimizing reconstruction error is sufficient for downstream prediction.

Specifically, for an incomplete residential load series $\mathbf{x}$ with the mask of missing data $\mathbf{m}$, an imputer g_θ produces a reconstructed sequence $\mathbf{x}$, and a forecaster h_ϕ forecasts the future load demand y. Traditional two-stage training of the model parameters in imputer and forecaster are formulated by:

$$\theta^*=\arg\min_\theta \mathcal{L}_{imp} \tag{3}$$

$$\phi^*=\arg\min_\phi \mathcal{L}_{forecast} \tag{4}$$

where θ^* and θ are the optimal and original model parameters of imputer, ϕ^* and ϕ denotes the optimal and original model parameters of forecaster, $\mathcal{L}_{imp}$ and $\mathcal{L}_{forecast}$ are the loss function of imputer and forecaster. However, the above training process of imputer in (3) is not necessarily optimized for forecasting. On the other word, the imputed $\mathbf{x}$ may not suitable for the forecasting task in the second stage. Therefore, within the imputation-forecasting process, the imputation model should consider the impact of $\mathbf{x}$ on forecasting performance while completing its imputation task.

III. Proposed Method

A. The Proposed Mixture of Experts Structure

The proposed MoE structure is shown in Fig. 1. In Fig. 1, the residential load data (with missing data) is first cut into samples according to a given forecasting window. In this stage, the load data is mixed and not identified according to its household identity. Then, the load data sample $\mathbf{x}=\left\{x_1,\cdots x_W\right\}$ with length of window W is input in parallel to the gating NN and all imputation-forecasting coordinate experts (which is introduced in detail in Section III.B), which are formulated by:

$$w_k=\frac{\exp\left(G_k\left(\mathbf{x}\right)\right)}{\sum_{j=1}^{K}\exp\left(G_k\left(\mathbf{x}\right)\right)},k=1,\cdots,K \tag{5}$$

$$y_k=Expert_k\left(\mathbf{x}\right) \tag{6}$$

where $G_k\left(\mathbf{x}\right)$ is the gating NN, which computes the routing weights w_k of kth expert according to the electricity consumption pattern of $\mathbf{x}$, K is the number of experts in MoE, y_k is the forecasting value of kth expert in MoE.

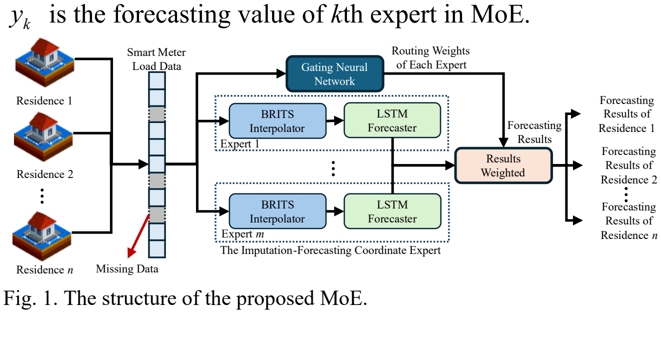


Fig. 1. The structure of the proposed MoE.

979-8-3315-8138-1

Then, all experts' forecasting results will be weighted according to w_k and then output. The final forecasting result of MoE is:

$$y = \sum_{k=1}^{K} w_k y_k \tag{7}$$

where y is the forecasting results of MoE.

To ensure that the gating network assigns appropriate routing weights and that each expert specializes in the electricity consumption patterns it is most competent at forecasting, the gating NN and all experts in the MoE are trained jointly in an end-to-end fashion. The joint training process is represented by:

$$\left(\theta_G^*, \theta_{E_1}^*, \cdots, \theta_{E_K}^*\right) = \underset{\left(\theta_G, \theta_{E_1}, \cdots, \theta_{E_K}\right)}{\arg\min} \mathcal{L}_{fore}\left(D_{Training}\right) \tag{8}$$

where θ_G^* and $\theta_{E_i}^*$ are the optimal model parameters of gating NN and ith expert, θ_G and θ_{E_i} are the initial model parameters of gating NN and ith expert, $\mathcal{L}$ denotes the loss function (means absolute error in this paper), $D_{Training}$ is the training load datasets (with missing data) which collected from multiple residences' smart meter.

In implementation, the gating network takes the same incomplete load window as the expert inputs, together with the corresponding binary missing-data mask. Missing entries are represented via zero-masking, allowing the gating network to infer electricity consumption patterns directly from observed temporal structures rather than residence identities. The gating network outputs a soft assignment over experts via a softmax layer, enabling each input sequence to be processed by multiple experts with different weights.

Note that in this paper, gating NN is a simple multilayer perceptron (MLP), which can also be replaced by more advanced gating NN structure.

B. The Imputation-Forecasting Coordinate Expert

1) The Bidirectional Recurrent Imputation for Time Series Model-based Imputer

BRITS is a recurrent imputation framework that estimates missing values by performing forward and backward temporal inference. In this paper, we adopt a simplified BRITS-based imputer, retaining the core history-based recurrent estimation mechanism while omitting the temporal decay and feature-correlation modules, which are unnecessary for our regularly sampled 1-d smart-meter data.

For a load series with length of W window $\mathbf{x} = \left(x_1, \cdots, x_W\right)$ and a mask of missing data point $\mathbf{m} \in \{0,1\}^W$, the BRITS-based imputer performs forward and backward recurrent imputations to rationally consider information from both temporal directions.

In the forward direction, at each time step t, a recurrent state $h_t^{forward}$ is updated based on the previously imputed value:

$$h_t^f = GRU_f\left(\left(m_t \odot x_t + \left(1 - m_t\right) \odot g_f\left(h_{t-1}^f\right)\right), h_{t-1}^f\right) \tag{9}$$

where GRU_f and g_f denotes the gated recurrent unit neural network and a reconstruction NN layer in forward imputation, $m_t \in \mathbf{m}$ is the mask value of tth datapoint, $\odot$ denotes element-wise multiplication. Thus, a reconstructed sequence from the forward direction $\mathbf{x}^f$ is represented by:

$$\mathbf{x}^f = \left(g_f\left(h_0^f\right), \cdots, g_f\left(h_{W-1}^f\right)\right) \tag{10}$$

Similarly, in the backward direction, another imputation is processed, which is formulated by:

$$h_t^b = GRU_b\left(\left(m_t \odot x_t + \left(1 - m_t\right) \odot g_b\left(h_{t+1}^b\right)\right), h_{t+1}^b\right) \tag{11}$$

where GRU_b and g_b denotes the gated recurrent unit neural network and a reconstruction NN layer in backward imputation. Thus, a reconstructed sequence from backward direction $\mathbf{x}^b$ is represented by:

$$\mathbf{x}^b = \left(g_b\left(h_1^b\right), \cdots, g_f\left(h_W^b\right)\right) \tag{12}$$

Then, the final imputed value of tth data point is taken as the average of $\mathbf{x}^f$ and $\mathbf{x}^b$:

$$x_t = \frac{1}{2}\left(g_b\left(h_t^b\right) + g_f\left(h_t^f\right)\right) \tag{13}$$

2) The LSTM Forecaster

In this paper, the deployed forecaster is a traditional LSTM model, which is a classic NN model for time series forecasting. Note that the forecaster in the proposed method is model-agnostic, which can be modular replaced by another advanced gradient descent-based model.

3) The Optimization Objective of The Proposed Imputation-Forecasting Coordinate Expert

Each coordinate expert in the proposed framework consists of a BRITS-based imputer followed by a LSTM-based forecaster. From the system-level perspective, the forecasting loss (MAE) used in the MoE in (8) already reflects the prediction performance of each expert. However, the primary role of the BRITS-based imputer is still to reconstruct missing load values in a temporally coherent manner, and relying solely on the forecasting loss to update the BRITS-based imputer would be insufficient. On the other hand, optimizing only a pure reconstruction loss is also inadequate, because it ignores how the imputed sequence affects the downstream forecastor. Therefore, the optimization objective of each expert is designed to combine imputation-oriented and forecast-oriented signals in a unified manner.

Therefore, following (13), for an imputed load series $\mathbf{x} = \left(x_1, \cdots, x_W\right)$ with length of window W, the masked imputation loss of imputer is designed as:

$$\mathcal{L}_{imp} = \left\| \mathbf{m} \odot \left(\mathbf{x} - \mathbf{x}\right) \right\| \tag{14}$$

where $\mathcal{L}_{imp}$ is the masked imputation loss, $\mathbf{m}$ denotes the mask series of original load series $\mathbf{x}$, $\odot$ denotes element-wise multiplication. This term ensures that the imputer faithfully reconstructs available measurements without being constrained at originally missing positions. And (14) only penalizes errors on observed entries.

Furthermore, to exploit temporal information from both directions, the expert also enforces forward-backward consistency loss between $\mathbf{x}^f$ and $\mathbf{x}^b$:

$$\mathcal{L}_{cons} = \left\| \mathbf{x}^f - \mathbf{x}^b \right\| \tag{15}$$

(15) encourages the two recurrent inference paths to converge to a coherent reconstruction trajectory, improving the stability of the imputed sequence.

To enable the BRITS-based imputer to generate imputation series that aid in forecasting, the MAE loss in MoE in (8) is also incorporated into the imputer's loss function through backpropagation of the LSTM-based forecaster (not directly backpropagation to the imputer). Formally, the gradient calculation from MAE loss to the model parameters of the imputer θ_{imp} is represented by:

$$\frac{\partial \mathcal{L}_{fore}}{\partial \theta_{imp}} = \frac{\partial \mathcal{L}_{fore}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial \theta_{imp}} \tag{16}$$

where $\mathbf{y}$ is the real value set of forecasting task. (16) encourages the imputer to produce imputations that are not only reconstruction-accurate but also beneficial for forecasting.

Combining the above components, the optimization objective of a single imputation-forecasting coordinate expert is defined as

$$\mathcal{L}_{\exp} = \mathcal{L}_{\exp} + \alpha \mathcal{L}_{fore} + \beta \mathcal{L}_{cons} \tag{17}$$

where $\alpha > 0$ and $\beta > 0$ are two trade-off coefficients. In this way, the expert's LSTM-based forecaster is primarily driven by $\mathcal{L}_{fore}$, while the BRITS-based imputer is updated jointly by the reconstruction loss, the forward-backward consistency term, and the forecasting loss backpropagated through the LSTM. This design of loss function ensures that the imputed sequences are both faithful to the observed smart meter data and well aligned with the requirements of downstream load forecasting. In practice, the forecasting loss serves as the primary optimization signal for the LSTM forecaster, while the BRITS-based imputer is updated jointly through reconstruction loss, forward–backward consistency, and forecasting gradients propagated through the forecaster. The trade-off coefficients balance reconstruction fidelity and forecasting utility, ensuring that the imputed sequence remains faithful to observed data while being optimized for downstream prediction.

As for the training of gating NN, all experts are trained jointly in an end-to-end fashion using soft routing, without explicit expert-wise regularization. Expert specialization naturally emerges through competition induced by the gating network, where experts that better fit certain consumption patterns receive higher routing weights and thus stronger gradient updates.

To summarize, (3)-(4) illustrate the decoupled optimization in conventional two-stage pipelines, while (16)-(17) show how the proposed method explicitly couples imputation and forecasting objectives through end-to-end optimization.

IV. Simulation Results

In this paper, a public user-level load dataset [17] with 20 residential buildings is utilized to validate the performance of the proposed method. To evaluate the performance and generalization capability of the proposed method, we adopt a cross-household evaluation protocol on the residential load dataset. Specifically, the dataset consists of 20 residential buildings, and two complementary training-testing splits are considered. In the first setting, the model is trained on the first 10 residences and tested on the remaining unseen 10 residences; in the second setting, the training and test sets are swapped. This two-way evaluation ensures that all residences are evaluated as unseen test data, providing a fair assessment of cross-residential generalization while avoiding data leakage.

The number of experts in the MoE is set to three, which represents a conservative yet sufficient choice to capture diverse electricity consumption patterns without introducing excessive model complexity. To comprehensively assess robustness under missing data, two representative missingness patterns are considered in the experiments: random missing (MCAR) and block missing, with multiple missing rates evaluated for each pattern.

The proposed method is compared with several representative benchmark approaches:

1. LSTM (Mixture Training), where a single LSTM is trained on mixed data from multiple residences without using a BRITS-based imputer or MoE structure;
2. LSTM (Individual Training), where a separate LSTM model is trained independently for each residence;
3. LSTM (MoE), which adopts the same MoE structure and training–testing splits as the proposed method but excludes the BRITS-based imputer and coordinated training mechanism;
4. Imputation/Forecasting Split, a two-stage baseline where imputation and forecasting are trained separately.

The forecasting performance of the proposed method and all benchmarks is summarized in Table I. Overall, the proposed method consistently achieves the lowest average MAE in 20 residences across most missingness settings under both MCAR and block missing patterns. While MoE-based baselines with explicit imputation already belong to the top-performing group, the proposed coordinated framework further improves forecasting accuracy, especially as the missing rate increases.

In contrast, the remaining benchmarks exhibit significantly higher forecasting errors, indicating limited robustness to missing data and reduced generalization across households. Moreover, the comparison between LSTM (Mixture Training)

TABLE I
Performance of The Proposed Method and Benchmarks in 20 Residential Building Dataset

Missing Pattern	Parameter Setting of Data Missing			Method				
	Missing Rate	Block Rate/Step	Block Duration	The Proposed	LSTM (Mixture Training)	Imputation/ Forecasting Split	LSTM (Individual Training)	LSTM (MoE)
MCAR	10%	-	-	0.0402	0.0416	0.0401	0.0492	0.0404
	15%	-	-	0.0400	0.0410	0.0400	0.0493	0.0409
	20%	-	-	0.0403	0.0417	0.0406	0.0505	0.0408
	25%	-	-	0.0399	0.0427	0.0401	0.0477	0.0417
Block Missing	-	6%	4~10 Steps	0.0411	0.0442	0.0417	0.0496	0.0435
	-	6%	6~12 Steps	0.0417	0.0440	0.0419	0.0514	0.0437
	-	10%	4~10 Steps	0.0414	0.0443	0.0418	0.0491	0.0433
	-	10%	6~12 Steps	0.0419	0.0448	0.0416	0.0482	0.0437

and LSTM (Individual Training) shows that mixed training across multiple residences improves predictive performance, but remains insufficient to support reliable single-model cross-residential forecasting. These results collectively demonstrate that the proposed method not only improves forecasting accuracy under missing data but also generalizes effectively across heterogeneous households.

To explore the forecasting strategy in MoE, the forecasting curves of three experts in the same input load data series under MCAR pattern (25% missing probability) are shown in Fig.2. As shown in Fig. 2, expert 1 has better tracking and forecasting capabilities for regular load data, while expert 2 is better at handling load data with sudden peaks. Notably, expert 3 is better at handling load data with long-term low values. To verify this, we selected the three load samples with the highest routing weights from the three experts across all tests set in 15 residential load datasets, which is shown in Fig. 3.

Fig. 3 shows that the three experts receive the highest routing weights for distinct electricity consumption patterns. Experts 1 and 2 mainly capture load sequences with pronounced fluctuations in the early and later stages, respectively, whereas Expert 3 specializes in long-term low-load patterns. These results confirm that the MoE framework effectively assigns input sequences with different characteristics to appropriate experts. Fig. 4 further illustrates the distribution of routing weights across experts. Expert 2 receives higher average weights, indicating stronger adaptability to common residential load patterns, while Expert 3 is assigned lower weights due to the relatively low occurrence of persistent low-load patterns in typical residential data.

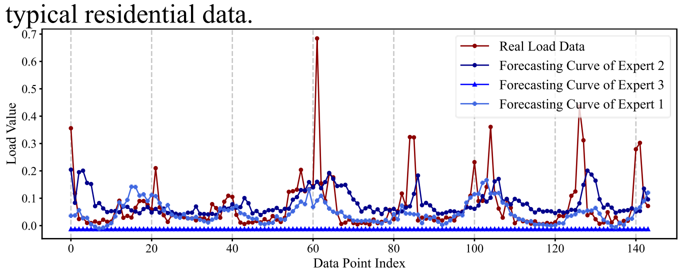
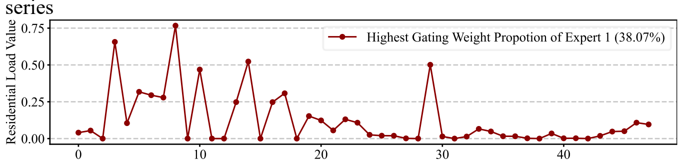
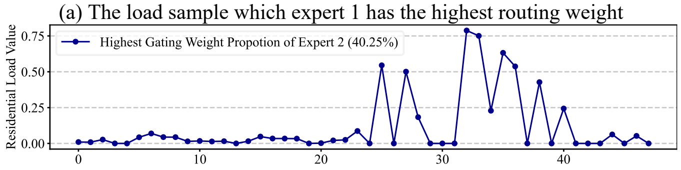


Fig. 2. The forecasting curves of three experts in the same input load data series



(a) The load sample which expert 1 has the highest routing weight



(b) The load sample which expert 2 has the highest routing weight

(c) The load sample which expert 3 has the highest routing weight

Fig. 3. Three load samples with the highest weights from the three experts.

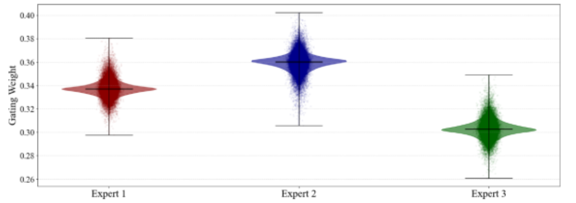


Fig. 4. Routing weight distribution of three experts.

V. Conclusions and Future Work

This paper proposes an imputation-forecasting coordinate MoE method for multi-residential load forecasting with missing data. Interestingly, this paper demonstrates that the MoE-driven residential load forecasting method has the potential to handle a large amount of diverse residential load data using a single model. This provides a potential avenue for developing generalized forecasting models in the field of residential load.

References

[1] J. D. Rhodes, C. R. Upshaw, C. B. Harris, et al., “Experimental and data collection methods for a large-scale smart grid deployment: Methods and first results,” *Energy*, vol. 65, pp. 462–471, 2014.

[2] Y. Wang, Q. Chen, T. Hong, et al., “Review of smart meter data analytics: Applications, methodologies, and challenges,” *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[3] W. Kong, Z. Y. Dong, D. J. Hill, et al., “Short-term residential load forecasting based on LSTM recurrent neural network,” *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[4] T. Kim, W. Ko, and J. Kim, “Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting,” *Appl. Sci.*, vol. 9, no. 1, p. 204, 2019.

[5] K. Park, J. Jeong, D. Kim, et al., “Missing-insensitive short-term load forecasting leveraging autoencoder and LSTM,” *IEEE Access*, vol. 8, pp. 206039–206048, 2020.

[6] S. R. Kuppannagari, Y. Fu, C. M. Chueng, et al., “Spatio-temporal missing data imputation for smart power grids,” in *Proc. 12th ACM Int. Conf. Future Energy Syst.*, 2021, pp. 458–465.

[7] D. Kodaira and S. Han, “Topology-based estimation of missing smart meter readings,” *Energies*, vol. 11, no. 1, p. 224, 2018.

[8] J. Peppanen, X. Zhang, S. Grijalva, et al., “Handling bad or missing smart meter data through advanced data imputation,” in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. (ISGT)*, 2016, pp. 1–5.

[9] Su, Heng-Yi, and Chia-Ching Lai, “Toward Improved Load Forecasting in Smart Grids: A Robust Deep Ensemble Learning Framework,” *IEEE Trans. on Smart Grid,* vol. 15, no. 4, pp. 4292–4296, 2024.

[10] Zhang, Tong, et al., “Short-Term Residential Load Forecasting Framework Based on Spatial–Temporal Fusion Adaptive Gated Graph Convolution Networks,” *IEEE Trans, on Ne. Net. and Lea. Syst.*, 2025 *early access*.

[11] Shabbir, Attia, et al. "Robustness against data integrity attacks in decentralized federated load forecasting." *Electro.* vol. 13, no.23, pp.4803, 2024.

[12] M. A. Husnoo, A. Anwar, N. Hosseinzadeh, et al., “A secure federated learning framework for residential short-term load forecasting,” *IEEE Trans. Smart Grid*, vol. 15, no. 2, pp. 2044–2055, 2023.

[13] Z. Zhang, P. Zhao, P. Wang, et al., “Transfer learning featured short-term combining forecasting model for residential loads with small sample sets,” *IEEE Trans. Ind. Appl.*, vol. 58, no. 4, pp. 4279–4288, 2022.

[14] Y. He, F. Luo, and G. Ranzi, “Transferrable model-agnostic meta-learning for short-term household load forecasting with limited training data,” *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 3177–3180, 2022.

[15] W. Cai, J. Jiang, F. Wang, et al., “A survey on mixture of experts in large language models,” *IEEE Trans. Knowl. Data Eng.*, 2025, early access.

[16] W. Cao, D. Wang, J. Li, et al., “BRITS: Bidirectional recurrent imputation for time series,” *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[17] M. David, S. Lina and S. Vladimir, “REFIT: Electrical Load Measurem ents (Cleaned)”, https://pureportal.strath.ac.uk/en/datasets/refit-electrica l-load-measurem ents-cleaned.