P-FedNIP: A Multi-Layered Personalized Federated Learning Framework

Rahul Nanduri

July 14, 2025

1 Abstract

Energy forecasting presents unique challenges due to temporal dependencies, seasonal variations, and diverse consumption patterns that require specialized federated learning approaches to capture local patterns while preserving data privacy. While Federated Learning (FL) has emerged as a privacy-preserving alternative for training models on decentralized data, standard algorithms like FedAvg often falter under the non-IID conditions inherent in energy consumption data, leading to reduced convergence speed and suboptimal model accuracy. To address these limitations, we propose P-FedNIP, a novel multi-layered personalized federated learning framework. P-FedNIP extends the FedNIP algorithm by introducing a sophisticated architecture that combines (1) EMD-based client clustering to understand the data landscape, (2) intelligent client selection to optimize training, (3) FedProx regularization to prevent local model drift, and (4) adaptive fine-tuning for deep personalization. This approach aims to create both a robust global model and highly accurate personalized models tailored to the unique energy consumption patterns of each participant.

Keywords: Personalized Federated Learning, Client Selection, Non-IID Data, Energy Forecasting, Time Series Analysis

2 Introduction

2.1 Energy Forecasting in Federated Learning

Energy forecasting is well-suited for federated learning due to its data heterogeneity and privacy requirements. Energy consumption data shows strong temporal patterns, seasonal variations, and location-specific behaviors that differ significantly across households and buildings [2]. This creates non-IID data distributions that challenge standard federated learning algorithms.

Energy consumption data is privacy-sensitive as it can reveal occupancy patterns, lifestyle habits, and economic information. This makes centralized data collection problematic. Federated learning enables collaborative model training while keeping sensitive energy data distributed across participating nodes.

Energy consumption patterns vary due to building characteristics, occupancy patterns, appliance usage, and local climate conditions. This heterogeneity requires personalized approaches that can capture local consumption behaviors while contributing to robust global models.

The remainder of this paper is structured as follows. We begin by reviewing related work in federated learning, personalization, and client selection. We then introduce the proposed P-FedNIP framework in detail, describing its multi-layered architecture. Subsequently, we outline the experimental setup for our evaluation, followed by a presentation and analysis of the comparative results. We conclude with a summary of our findings and potential directions for future research.

3 Background & Related Work

3.1 Federated Learning Fundamentals

Federated Learning (FL) has emerged as a promising approach for training models across decentralized, privacy-sensitive devices without requiring raw data sharing [11]. However, in highly heterogeneous environments such as smart homes, standard FL algorithms like FedAvg struggle due to divergent data distributions across participating households [7]. By aggregating all client model updates uniformly, FedAvg often suffers from slow model convergence and suboptimal performance for individual participants when data is not independent and identically distributed (non-IID).

3.2 Handling Heterogeneity and Personalization

Several strategies have been proposed to mitigate these challenges. FedProx [8] introduces a proximal term to the local loss function, penalizing significant deviations from the global model and thereby stabilizing training. FedDyn [1] dynamically regularizes local updates based on the global objective, providing another powerful approach to handle client heterogeneity.

While these methods improve upon FedAvg, they primarily focus on creating a single, robust global model. The field of Personalized Federated Learning (PFL) goes a step further, aiming to provide each client with a customized model. Techniques range from fine-tuning a global model on local data [15], to model interpolation (e.g., Ditto [9]), and meta-learning approaches like Per-FedAvg [5]. These methods acknowledge that for many real-world applications, a personalized model is more valuable than a generalized one.

3.3 Client Selection in Federated Learning

The performance of FL is also heavily influenced by which clients are chosen to participate in each training round. Random selection is simple but inefficient in non-IID settings. More advanced client selection strategies have been proposed, such as those based on client data quality or contribution to the global model [3, 12]. Our work builds on FedNIP, a framework that uses client clustering and a dynamic ranking system to intelligently select participants, forming the foundation for our personalization layers.

3.4 Research Questions

This paper is guided by two central research questions that build upon the context so far:

1. **RQ1: Adapting FedNIP for Personalization.** How can the FedNIP algorithm, originally designed for efficient global model training, be extended to a personalized framework (P-FedNIP) that creates highly-tuned models for individual smart homes without sacrificing the benefits of its intelligent client selection?

2. **RQ2: Quantifying the Impact of Personalization under High Heterogeneity.** In challenging, non-IID environments what is the specific performance contribution of P-FedNIP's personalization layers when compared to a non-personalized baseline? What about in less heterogeneous environments?

4 The P-FedNIP Framework

This section details the architecture and mechanics of our proposed personalized federated learning framework, P-FedNIP. We begin by formally defining the energy forecasting problem in a federated context and then present the four distinct but integrated layers of our solution.

4.1 Problem Formulation

Consider a federated learning system with a central server and a set of N clients (energy consumers), indexed by $i \in \{1, \ldots, N\}$. Each client i possesses a private local dataset $\mathcal{D}_i = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, where $n_i = |\mathcal{D}_i|$. For energy forecasting tasks, temporal context is crucial—a single sample $\mathbf{x}_{i,j} \in \mathbb{R}^{W \times F}$ represents a window of $W = 60$ time steps with F features capturing energy consumption patterns, while $y_{i,j} \in \mathbb{R}$ represents the target energy consumption value.

The objective of standard federated learning, such as FedAvg, is to train a single global model with parameters $\mathbf{w}$ that minimizes the aggregate loss across all clients:

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \frac{n_i}{n} \mathcal{L}_i(\mathbf{w})$$

where $n = \sum_{i=1}^{N} n_i$ is the total number of samples and $\mathcal{L}_i(\mathbf{w})$ is the local loss function (in our case, Mean Squared Error) for client i on its local data $\mathcal{D}_i$.

Energy forecasting differs from general time-series prediction due to inherent seasonality, daily cycles, and consumer-specific consumption behaviors that create significant statistical heterogeneity across participants. Due to this data heterogeneity (non-IID) across energy consumers, this single global model $\mathbf{w}$ often fails to perform optimally for individual clients. The goal of our work is to move beyond this one-size-fits-all approach. We aim to develop a set of personalized models $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ that are specifically adapted to each client's local data distribution, thereby achieving superior performance at the local level compared to the global model. P-FedNIP is designed to achieve this by intelligently managing the federated training process to produce both a strong global model and highly effective personalized models.

4.2 P-FedNIP: A Multi-Layered Framework

To address the dual challenges of statistical heterogeneity and the need for personalization, we propose **P-FedNIP**, a novel four-layered federated learning framework. Each layer builds upon the last, creating a comprehensive system that first understands the client landscape, then intelligently selects participants, stabilizes training, and finally, personalizes models for individual users.

4.2.1 Layer 1: EMD-based Client Clustering

The first challenge in a heterogeneous FL environment is to understand the structure of the data disparity. P-FedNIP begins with a profiling step where the server quantifies the dissimilarity between client data distributions.

1. **Local Histogram Generation:** Each client i first analyzes its local dataset $\mathcal{D}_i$. It computes a normalized histogram, $\mathbf{h}_i$, of its target variable values (i.e., the energy consumption data). This histogram serves as a compact, privacy-preserving summary of its local data distribution.

2. **Dissimilarity Calculation with Earth Mover's Distance (EMD):** The server collects these histograms from all clients. To compare them, it computes a pairwise distance matrix $\mathbf{D}$ where each element $\mathbf{D}_{ij}$ is the Earth Mover's Distance (EMD) between the histograms of client i and client j. EMD is chosen over simpler metrics like KL-divergence as it excels at comparing distributions that may not have overlapping support, which is common in non-IID settings.

3. **Client Clustering:** With the dissimilarity matrix $\mathbf{D}$ established, the server uses K-Means clustering to group clients with similar data distributions. To find the optimal number of clusters, K_{opt}, the server computes the silhouette score for a range of K values. The original FedNIP implementation [16] uses the Elbow method, but this can be subjective. We therefore use the silhouette score, which provides a more quantitative measure

for cluster quality [13]. The value of K that maximizes the average silhouette score is selected as the optimal number of clusters.

The output of this layer is a set of K_{opt} clusters, $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_{opt}}\}$, which provides a topological map of the client data landscape, fundamental for the intelligent selection in Layer 2.

4.2.2 Layer 2: FedNIP-based Intelligent Client Selection

Following the initial clustering, this layer aims to dynamically identify and train only the most promising clients. The selection process begins with a mandatory warm-up round where all clients participate. After this round, each client evaluates its newly trained model on its own local validation set and reports the performance score (RMSE) back to the server. The server then uses these self-reported scores to create a ranked list, $\mathcal{R}$, where clients with lower RMSE are assigned a higher rank.

In all subsequent rounds, instead of training all clients, the server selects a mix of clients: the top $k\%$ from the ranked list and a random $r\%$ chosen from the remaining clients. These selected clients receive the global model, train on their local data, and send their updates for aggregation. To ensure the ranked list $\mathcal{R}$ accurately updates over time, a periodic re-evaluation mechanism is triggered every $T_{\text{re-eval}}$ rounds, where all clients report new scores and the list is re-sorted. This allows for "rank-swapping," enabling low-performing clients to improve and high-ranking clients that overfit to be deselected.

4.2.3 Layer 3: Continuous Personalization via Regularization

The core of P-FedNIP's personalization strategy lies in managing client drift while cultivating specialized local models. This is achieved by integrating the FedProx algorithm [8] directly into the training loop. Instead of only minimizing the standard local loss, each participating client minimizes a modified objective function that includes a proximal term:

$$L_i'(w) = L_i(w) + \frac{\mu}{2}\|w - w^t\|^2$$

Here, $L_i(w)$ is the standard Mean Squared Error, w^t represents the global model parameters from round t, and μ controls the regularization strength. This encourages local models to stay close to the global consensus while giving them freedom to diverge in ways beneficial for their local data.

4.2.4 Layer 4: Final Personalization through Fine-Tuning

After the final communication round, the server distributes the global model one last time. Each client then performs a few additional epochs of local training on its own dataset. This fine-tuning is conducted with a significantly reduced learning rate, allowing the model to make subtle adjustments to capture specific local patterns without catastrophically forgetting the generalized knowledge learned. This two-stage process results in a final set of models $\{w_1, \ldots, w_N\}$ highly optimized for individual client performance.

5 Experimental Setup

5.1 Dataset & Preprocessing

We use the AMPds2 dataset, a comprehensive energy consumption dataset containing electricity usage data from a Canadian household over two years. This dataset captures the multi-faceted nature of energy consumption with 21 sub-meters at one-minute granularity, representing detailed appliance-level and whole-house consumption patterns typical in energy forecasting applications. We select the main meter ('WHE') as the target variable for our energy forecasting task.

Following standard time-series practices, we partition the data chronologically. The first 70% of the data is used for training, the next 15% for validation, and the final 15% for testing. We start with a standard baseline of 15 clients to understand the behavior of P-FedNIP, but we also perform experiments at N=40 clients to visualize how having multiple actors can influence the results of the federated setup. To simulate a federated environment with N clients, we partition the training data among the clients using a Dirichlet distribution with 3 α parameters at 0.1, 0.3 and 0.6. The smaller the number, the higher the non-IID split of the data distributions. The temporal nature of energy consumption, with its daily cycles, seasonal variations, and consumer-specific patterns, creates the challenging non-IID conditions that make energy forecasting an excellent domain for evaluating personalized federated learning approaches.

5.2 Model Architecture

All clients use an identical LSTM-based model architecture, a common and effective choice for time series energy forecasting [4, 10]. The model consists of a two-layer, bidirectional LSTM with a hidden size of 96 units, a dropout rate of 0.3, and a final linear layer for the forecast.

5.3 Baseline Algorithms

We compare P-FedNIP against two baselines:

- **FedAvg:** The standard algorithm where all clients participate and their updates are uniformly averaged.
- **FedNIP:** The non-personalized version of our algorithm. It uses Layers 1 and 2 but omits FedProx regularization and final fine-tuning, allowing us to isolate the benefits of the personalization layers.

5.4 Hyperparameter Configuration

Our choices were guided by established practices. Our LSTM architecture is consistent with models in similar tasks [4, 10]. The FedProx regularization term, $\mu = 0.01$, aligns with the original implementation [8].

However, through empirical testing for this specific case we find that a $\mu = 0.001$ allows us to see a more dramatized effect of personalisation than higher values. Going even lower than this would severely affect the local fine-tuning we do later on. The choice of 15 and 40 clients aligns with established practices in federated learning for energy forecasting applications. Small-scale studies (10-15 clients) are common in energy forecasting research, as demonstrated by [14] with 13 clients for federated LSTM load forecasting. This scale allows for detailed analysis of personalization benefits while remaining computationally manageable. Research by [6] shows performance gains plateau beyond 15 clients, supporting our smaller configuration for energy forecasting evaluation. Larger studies (30-50 clients) test scalability with increased heterogeneity, such as [2] who used 40 clients across multiple climate zones for energy planning applications. In energy forecasting, this scale captures greater diversity in consumption patterns, seasonal effects, and building characteristics. Our 15-client setup enables detailed analysis of personalization benefits in energy forecasting, while our 40-client configuration evaluates scalability with the increased heterogeneity typical of diverse energy consumption patterns across different consumer types and geographic conditions.

Other parameters were selected through standard empirical tuning, detailed in Table 1.

Table 1: Hyperparameter settings for the federated learning experiments.

Hyperparameter	**Value**
Number of Rounds	400
Number of Clients (N)	15, 40
Dirichlet Alpha (α)	0.1, 0.3, 0.6
Local Epochs	3
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
LSTM Hidden Size	96
LSTM Dropout	0.3
FedProx Mu (μ)	0.001
FedNIP Top-k%	20%
FedNIP Random-r%	20%
FedNIP Re-evaluation ($T_{\text{re-eval}}$)	7 rounds

6 Results and Analysis

The analysis focuses on metrics particularly relevant to energy forecasting applications. RMSE directly relates to prediction accuracy in energy consumption values, which is critical for forecasting applications. The personalization benefits are especially important in energy forecasting because individual consumption patterns can vary dramatically based on building characteristics, occupancy patterns, appliance usage, and seasonal behaviors, making personalized models significantly more valuable than generic global models.

In this section, we present and analyze the empirical results from our simulations. We aim to answer the research questions posed by comparing the performance of P-FedNIP against the FedNIP and FedAvg baselines, focusing on both global model convergence and local model personalization under a high-heterogeneity scenario ($\alpha = 0.1$) and less heterogenous split levels ($\alpha = 0.3, 0.6$).

6.1 Global Model Performance and Convergence

The global model's performance is evaluated across 15 and (in Subsection 6.3) 40 clients under varying degrees of data heterogeneity, controlled by the Dirichlet distribution parameter α, with values of 0.1, 0.3, and 0.6. A lower α indicates higher data heterogeneity, presenting a more challenging learning environment. The analysis focuses on the Root Mean Square Error (RMSE) to assess the performance of FedAvg, FedNIP, and P-FedNIP strategies.

6.1.1 Analysis for $\alpha = 0.1$

With high data heterogeneity ($\alpha = 0.1$), FedAvg achieves its lowest RMSE of 442.8 at Round 228. However, its performance is highly volatile, as evidenced by a final round RMSE of 460.8, representing a 4.1% degradation from its best performance. This instability is characteristic of FedAvg's random client sampling, which struggles to consistently aggregate improvements from highly diverse data distributions. In energy forecasting, this volatility is particularly problematic as it can lead to unreliable predictions that vary significantly between training rounds. The RMSE curve for FedAvg shows significant fluctuations, indicating a struggle to find a stable convergence point.

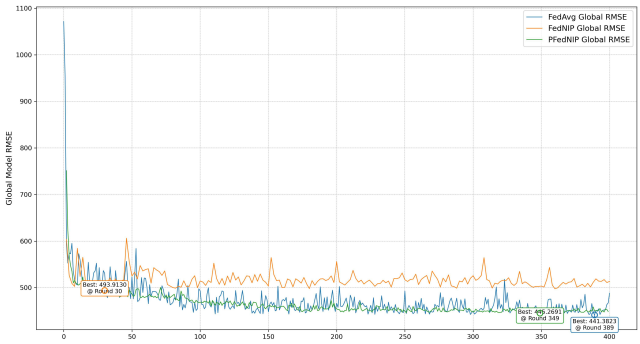

Figure 1: Global Model Convergence Comparison for 15 Clients ($\alpha = 0.1$)

In contrast, FedNIP demonstrates a more stable learning curve, achieving its lowest RMSE of 478.8 at round 59. It concludes with a final RMSE of 501.3, a 4.7% degradation. Although its lowest RMSE is higher than FedAvg's, the smoother curve suggests that the intelligent client selection and swapping mechanism of FedNIP is more effective at mitigating the negative impacts of heterogeneity than FedAvg's random approach.

P-FedNIP further refines this, achieving the lowest RMSE among the three strategies at 446.3 at round 251. It also boasts the highest stability, with a final RMSE

of 454.1, a mere 1.7% degradation from its peak performance. This indicates that the personalization–based clustering in P-FedNIP is highly effective at grouping clients with similar data patterns, leading to more consistent and effective global model updates, even in highly heterogeneous environments.

6.1.2 Analysis for $\alpha = 0.3$

At a moderate level of heterogeneity ($\alpha = 0.3$), FedAvg's performance improves, achieving its lowest RMSE of 444.1 at Round 235 and a final RMSE of 457.8, a 3.1% degradation. The curve remains volatile, but the peaks are less pronounced compared to the $\alpha = 0.1$ case, suggesting that a slight decrease in heterogeneity allows FedAvg to perform more consistently.

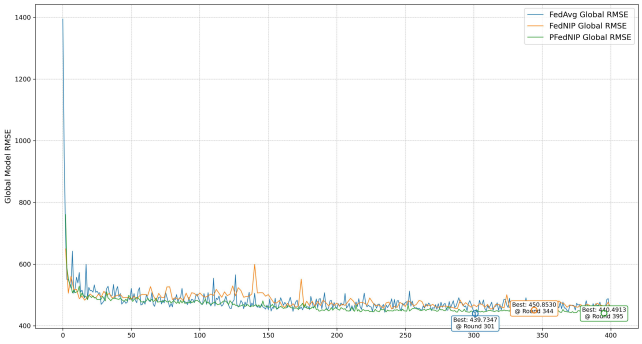

Figure 2: Global Model Convergence Comparison for 15 Clients ($\alpha = 0.3$)

FedNIP reaches its best RMSE of 458.1 at round 336, with a final RMSE of 468.4, showing a 2.2% degradation. P-FedNIP continues to outperform, recording a lowest RMSE of 447.6 at round 208 and a final RMSE of 452.9, a minimal 1.2% degradation. The performance gap between the NIP-based strategies and FedAvg narrows as heterogeneity decreases, which is expected. However, P-FedNIP's consistent stability and low performance degradation highlight its robustness.

6.1.3 Analysis for $\alpha = 0.6$

With low data heterogeneity ($\alpha = 0.6$), the performances of all three strategies converge. FedAvg achieves its lowest RMSE of 442.8 at round 201, with a final RMSE of 460.8, a 4.1% degradation. The curve, while still showing some volatility, is considerably more stable than in the higher heterogeneity scenarios.

FedNIP achieves its lowest RMSE of 458.1 at round 162 with a final RMSE of 469.8 (a 2.5% degradation). P-FedNIP again shows the best performance, with a lowest RMSE of 443.8 at round 246 and a final RMSE of 452.9 (a 2.0% degradation). In this low-heterogeneity environment, the advanced mechanisms of FedNIP and P-FedNIP provide a smaller, yet still noticeable, advantage over FedAvg.

6.1.4 Outlier Analysis and Modifications

Across all α values, FedAvg consistently exhibits the most outliers in the form of sharp spikes in RMSE.

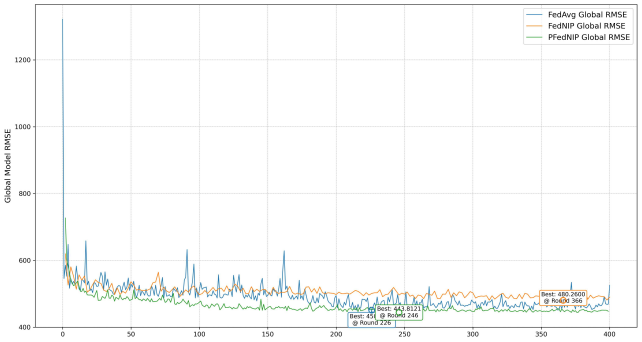

Figure 3: Global Model Convergence Comparison for 15 Clients ($\alpha = 0.6$)

These spikes can be attributed to its random client selection process. In energy forecasting with fluctuating consumption patterns, there is a significant chance of selecting a subset of clients whose data for a given round is anomalous or not representative of typical energy consumption trends, thereby degrading the global model.

FedNIP shows fewer outliers, but is not immune. A potential cause for outliers in FedNIP is the "cold start" problem for new clients that are swapped in. A client that has been out of the training cluster for many rounds might have a significantly diverged local model, and its introduction can temporarily destabilize the global model.

To improve the P-FedNIP algorithm's robustness to these outliers, several tweaks could be considered for future work. A "warm-up" period for newly swapped-in clients could be introduced, where their model updates are weighted less until they have participated in a few rounds of training. This would dampen the shock of introducing a potentially divergent model. Additionally, a momentum term could be added to the global model update rule, which would smooth out the effect of outlier rounds and create a more stable convergence path. This would involve maintaining a velocity vector that accumulates a fraction of the past updates, giving the optimization process inertia and making it less susceptible to drastic changes from a single round.

6.2 Local Performance and Personalization

To further analyze the performance of P-FedNIP, we must look at the performance of the final, personalized models on each client's local validation data. In energy forecasting applications, local performance is crucial as individual consumption patterns can vary significantly based on household characteristics, seasonal behaviors, and appliance usage. In the P-FedNIP framework, the final, personalized layers of the client models are never aggregated into the global model. When a client receives an updated global model from the server, it only overwrites its shared layers, preserving the specialized ones. This architecture ensures that extreme updates from a single client's data distribution—a likely occurrence in a highly non-IID setting—are averaged out in the global model and do not disrupt the locally adapted

layers of other clients. This effect is evidenced by the box plots below, which compare the final local validation RMSE for all clients across the three strategies.

6.2.1 Analysis for $\alpha = 0.1$

Under high data heterogeneity ($\alpha = 0.1$), the benefits of personalization are most pronounced. For P-FedNIP, the box plot shows a median RMSE of 500.6. The key advantage, however, is the tighter interquartile range compared to the other strategies, indicating that most clients receive a model that is consistently adapted to their local energy consumption patterns. Despite its strong general performance, P-FedNIP registers one significant outlier with an RMSE of 812.3, suggesting that for this specific client, the personalization process was not as effective, likely due to a highly unique local data distribution that was not well-represented even within its own cluster.

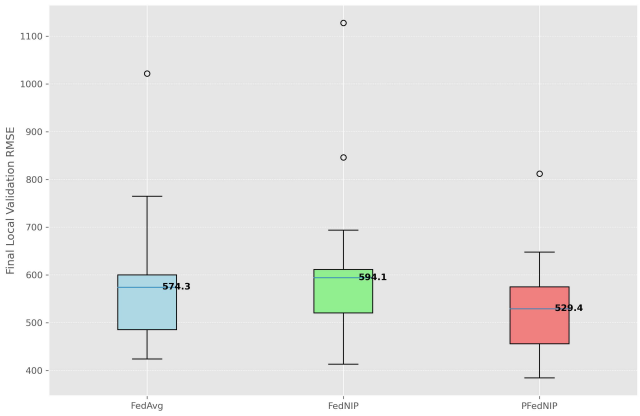

Figure 4: Local Performance Comparison for 15 Clients ($\alpha = 0.1$)

FedNIP achieves a median RMSE of 593.8. While the median is higher than P-FedNIP's, the most notable difference is the significantly larger spread of the data, with RMSE values ranging from 413.2 to 1127.9. This high variance indicates that while the intelligent swapping of FedNIP benefits some clients, it fails to deliver a consistently performing model for all, a direct consequence of applying a single global model to diverse local energy consumption patterns. FedAvg performs similarly, with a median RMSE of 574.3 and a very wide spread, underscoring its struggle to serve all clients effectively in a high-heterogeneity environment.

6.2.2 Analysis for $\alpha = 0.3$

With moderate heterogeneity ($\alpha = 0.3$), P-FedNIP maintains its edge, delivering a median RMSE of 492.7. The interquartile range remains tight, and while one outlier is present at 698.5, the overall performance for the majority of clients is strong and consistent. This demonstrates that the personalization approach is robust even as heterogeneity decreases.

FedNIP's median RMSE is 486.8, which is competitive with P-FedNIP. However, its variance is visibly

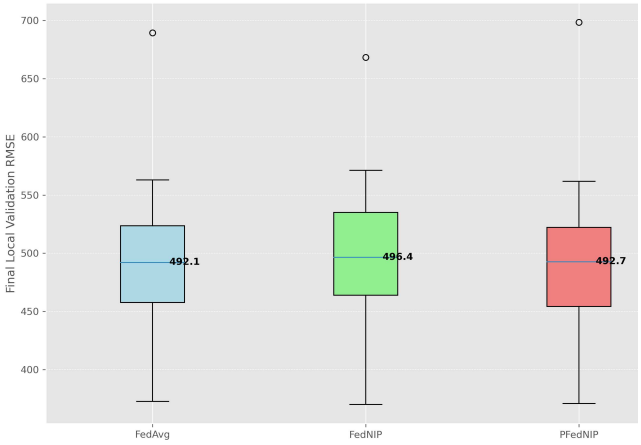

Figure 5: Local Performance Comparison for 15 Clients ($\alpha = 0.3$)

higher, with a wider box and longer whiskers, indicating less consistent performance across clients. FedAvg posts a median RMSE of 492.1, very close to the other two, but with the largest variance of the three, showing that many clients receive a model that is substantially worse than the median.

6.2.3 Analysis for $\alpha = 0.6$

In the low heterogeneity setting ($\alpha = 0.6$), the performance of the three strategies becomes more similar, as expected. P-FedNIP achieves a median RMSE of 499.5. The distribution of its performance is the tightest among the three, showcasing the consistent benefits of its personalization, even when client data is more alike.

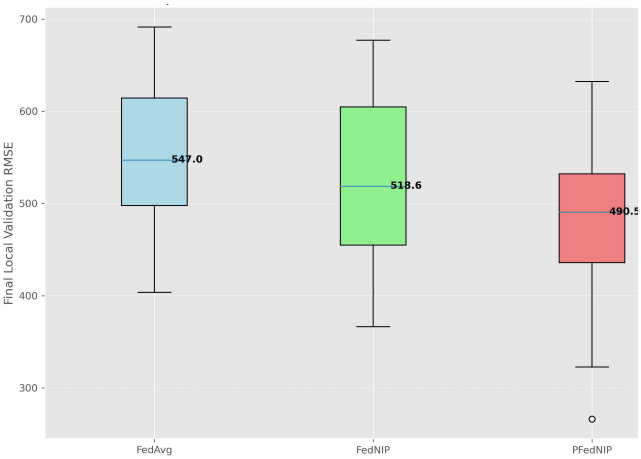

Figure 6: Local Performance Comparison for 15 Clients ($\alpha = 0.6$)

FedNIP has a median RMSE of 518.6, and FedAvg has a median of 547.0. While the medians are comparable, the box plots reveal that both FedNIP and FedAvg have a significantly larger spread in client performance than P-FedNIP. For FedAvg in particular, the high variance means that while the global model is decent on average, its local performance is a lottery; some clients get a well-performing model while others get a

poor one. P-FedNIP, by contrast, provides a more equitable and reliable level of performance for the majority of clients, which is a primary goal of personalization in federated learning.

6.3 Summary of 40-Client Results

We evaluated FedAvg, FedNIP, and P-FedNIP with 40 clients for $\alpha = 0.1$ (high heterogeneity) and $\alpha = 0.6$ (moderate heterogeneity). The main results are:

- **Global Model Performance:**
  - For $\alpha = 0.1$, the lowest global RMSEs were: FedAvg 361.1, FedNIP 384.6, and P-FedNIP 365.6. P-FedNIP reached its best performance much earlier in training.
  - For $\alpha = 0.6$, the lowest global RMSEs were: FedAvg 388.1, FedNIP 391.5, and P-FedNIP 378.7.
- **Local Model Performance:**
  - For $\alpha = 0.1$, the median local RMSEs were: FedAvg 575.2, FedNIP 552.0, and P-FedNIP 484.9.
  - For $\alpha = 0.6$, the median local RMSEs were: FedAvg 469.5, FedNIP 464.0, and P-FedNIP 445.4.
- **Stability and Fairness:**
  - P-FedNIP showed the most stable convergence and the best local fairness (lowest RMSE variance) in both settings.
  - FedAvg was the most volatile, especially for $\alpha = 0.1$.
- **Effect of Scaling the Number of Clients:**
  - Increasing the number of clients from 15 to 40 generally led to lower global and local RMSEs for all methods, indicating improved overall performance at larger scale.
  - The performance gap between P-FedNIP and the other methods became more pronounced with more clients, especially under high heterogeneity, highlighting the scalability and robustness of the personalized approach.

In summary, P-FedNIP provided the best global and local performance, as well as the most consistent and fair results, for both high and moderate data heterogeneity in the 40-client setting. Scaling up the number of clients further amplified the advantages of personalized federated learning. Table 2 in the Appendix and Figures 7, 8, 9, and 10 show the results of conducting 40 client experiments at high and low heterogeneity. Table 2 shows a high-level summary of the results achieved through the experimentation process for both 15 and 40 clients.

6.4 Communication Cost and Overhead

The primary Communication Cost is associated with the the transfer of the aggregated LSTM model's learned biases and weights back to clients at the start of each round. The LSTM, although relatively simple and rudimentary still has 1'288'577 parameters. Every time a client sends its updated model to the server, it incurs a cost of 5.15 MB. To preserve the idea of intelligence within the P-FedNIP framework, we adjusted our P-FedNIP strategy to add two relatively small communication costs. At the very start, each client sends a 20-number summary of its data. This costs just 80 bytes, and is sent in the form of a histogram. After training, the client will send its new RMSE logged score back, costing only 4 bytes. Communication costs were also kept to a minimum because of the inherent logic of the framework - we do not train all 15 or 40 clients throughout the simulations but only a small effective group based on our $k\%$ and $r\%$. These are the only clients that participate in the 5.15 MB model transfer each round. This provides a 60% reduction in our main communication cost, as when compared to training all clients. Furthermore, P-FedNIP learns which clients are the most valuable during proxy evaluation and other phases through the 80-byte histograms and the 4-byte RMSE score. This is the key trade-off of the P-FedNIP framework: accepting a negligible overhead cost in contrast to gaining a massive 60% resources saving on the primary cost.

7 Conclusion

In this paper, we addressed the dual challenges of statistical heterogeneity and the need for personalization in federated learning for energy forecasting. We introduced P-FedNIP, a novel multi-layered framework that integrates EMD-based client clustering, intelligent client selection, continuous regularization with FedProx, and a final fine-tuning step to produce highly specialized models for individual clients.

From an energy forecasting perspective, P-FedNIP addresses critical challenges where privacy-preserving collaborative learning is essential, but consumer heterogeneity prevents effective standard federated learning. The framework's ability to maintain both strong global performance and superior personalized performance makes it particularly suitable for energy forecasting applications where accurate local predictions are as important as aggregate forecasting capabilities. The significant improvements in local model performance demonstrate the value of personalization in capturing the diverse energy consumption patterns inherent in real-world energy forecasting scenarios.

Our experimental evaluation in a variety of heterogeneity environments demonstrated that P-FedNIP significantly outperforms both standard FedAvg and the non-personalized FedNIP baseline. The results showed that our framework achieves superior performance not only for the global model but, more critically, for the personalized local models, confirming the effectiveness of its integrated design. Furthermore, the effect of scal-

ing the number of clients did not drastically hurt model performance or outlier quantity at the local and global level, proving that there is potential for the framework to be applied in higher stress environments with varying degrees of heterogeneity.

Future Work

Currently, the P-FedNIP algorithm extends the FedNIP algorithm with personalisation techniques. Exploring different ranking and clustering algorithms may help build the complete profile. Additionally, exploring the impact of different clustering algorithms or more advanced personalization techniques within the P-FedNIP framework could yield further improvements. We have explored the effect of increased clients, but this research can further benefit from testing at more granular levels of heterogeneity, especially for $\alpha = 0.3$ at $N = 40$ and for extreme number of clients as well. Deploying and evaluating the framework on real-world hardware would provide valuable insights into its practical communication and computation trade-offs. Finally, experimenting with different energy forecasting datasets such as UKDALE and REDD may yield different results for the simulations, and help build a better generalizable profile for P-FedNIP across diverse energy consumption scenarios.

8 Use of AI

An initial query using ChatGPT's Deep Research feature was used to gather preliminary sources for the research proposal. Gemini ("Gemini 2.5 Pro") was used to validate LaTex functions for better readability and general structuring of the report within Overleaf. The author has not used any suggested code verbatim when building the code-base, reviewed all suggestions as needed carefully and relevant to the task, and takes full responsibility for the content of this work.

References

[1] Durmus Alp Acar, Yue Zhao, Ramon Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations (ICLR)*, 2021.

[2] M. Ali, R. Wazir, K. Imran, K. Ullah, A. K. Janjua, A. Ulasyar, A. Khattak, and J. M. Guerrero. Comparative analysis of data-driven algorithms for building energy planning via federated learning. *Energies*, 16(18):6517, 2023. doi: 10.3390/en16186517.

[3] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.10106*, 2020.

[4] K. D. Dinh et al. A novel approach for short-term building energy consumption forecasting using a huber-lstm-based model. *Sustainability*, 15 (24):16738, 2023.

[5] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[6] Z. Jiang, W. Wang, B. Li, and B. Li. Pisces: Efficient federated learning via guided asynchronous training. *arXiv preprint arXiv:2206.09264*, 2022. URL `https://arxiv.org/abs/2206.09264`.

[7] Qiang Li, Bingsheng He, and Dawn Song. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37 (3):50–60, 2020.

[8] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

[9] Tian Li et al. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning (ICML)*, 2021.

[10] M. Y. Maarif et al. Design and implementation of household electricity usage forecasting model based on long short-term memory. In *2023 International Conference on Data Science and Its Applications (ICoDSA)*, 2023.

[11] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[12] Takeru Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019.

[13] S. A. P. Raj and Vidyaathulasiraman. Determining optimal number of k for e-learning groups clustered using k-medoid. *International Journal of Advanced Computer Science and Applications*, 12(6), 2021.

[14] M. Wang, R. Xin, M. Xia, Z. Zuo, Y. Ge, P. Zhang, and H. Ye. A federated lstm network for load forecasting using multi-source data with homomorphic encryption. *AIMS Energy*, 13(2):265–289, 2025. doi: 10.3934/energy.2025011.

[15] P. Wang et al. Federated learning with matched averaging. In *ICLR Workshop*, 2019.

[16] S. Zagema. Fednip: A statistical heterogeneity aware dynamic ranking algorithm for federated learning. Master's thesis, University of Twente, 2024.

A Appendix

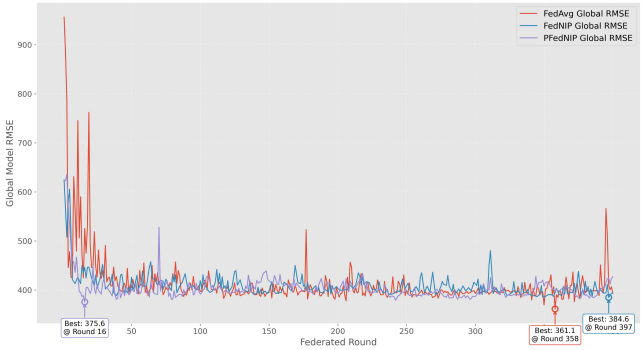

Figure 7: Global Performance Comparison for 40 Clients ($\alpha = 0.1$)

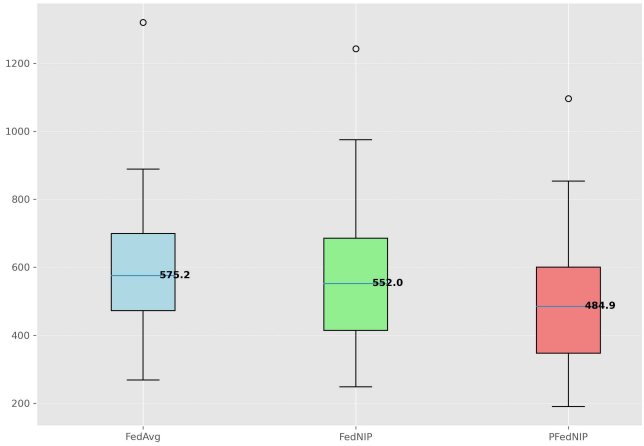

Figure 8: Local Performance Comparison for 40 Clients ($\alpha = 0.1$)

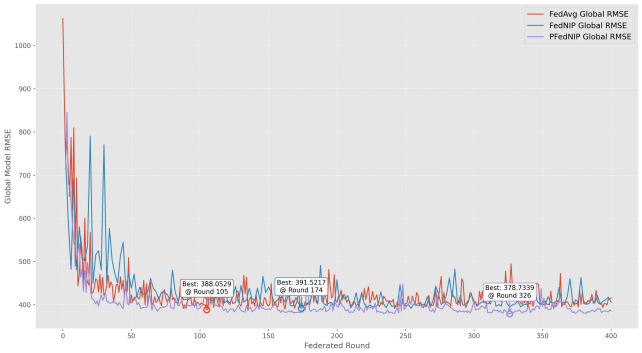

Figure 9: Global Performance Comparison for 40 Clients ($\alpha = 0.6$)

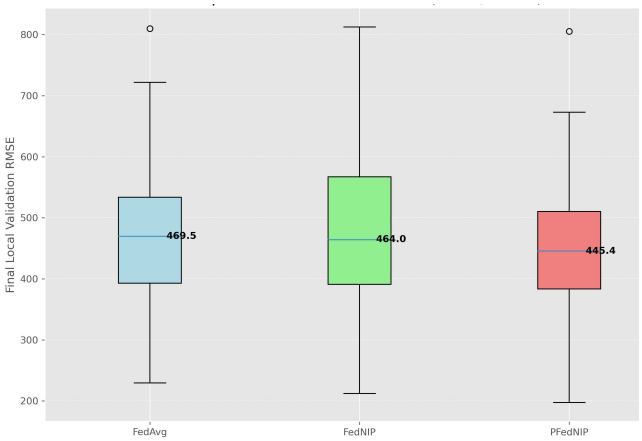

Figure 10: Local Performance Comparison for 40 Clients ($\alpha = 0.6$)

Table 2: Simplified Summary of Performance Findings for 15-Client and 40-Client Experiments

Metric	α	**FedAvg**	**FedNIP**	**P-FedNIP**
Final Global RMSE	0.1	460.8	513.4 **(Worst)**	454.1 **(Best)**
	0.3	457.8	469.8	452.9 **(Best)**
	0.6	489.7 **(Worst)**	455.8	452.9 **(Best)**
	0.1, 40	405.47	393.23	415.58
	0.6, 40	415.57	406.28	384.99
Lowest Global RMSE	0.1	442.8	478.8	446.3
	0.3	444.1	457.2	445.4
	0.6	443.8	447.6	445.6
	0.1, 40	361.1	384.6	365.6
	0.6, 40	388.1	391.5	378.7
Stability (Degradation)	0.1	Volatile (4.1%)	More Stable (7.2%)	Highly Stable (1.8%)
	0.3	Volatile (3.1%)	Stable (2.8%)	Excellent (1.7%)
	0.6	Unstable (10.3%)	Stable (1.8%)	Excellent (1.6%)
	0.1, 40	Volatile	Stable	Highly Stable
	0.6, 40	More Stable	Stable	Stable
Median Local RMSE	0.1	574.3	594.1 **(Highest)**	529.4 **(Lowest)**
	0.3	492.1	496.4	492.7
	0.6	547.0 **(Highest)**	518.6	490.5 **(Lowest)**
	0.1, 40	575.2	552.0	**484.9 (Best)**
	0.6, 40	469.5	464.0	**445.4 (Best)**
Local Fairness (RMSE Variance)	0.1	Poor (High Var.)	Poor (High Var.)	Excellent (Low Var.)
	0.3	Poor (High Var.)	Fair (Mod. Var.)	Excellent (Low Var.)
	0.6	Poor (High Var.)	Fair (Mod. Var.)	Excellent (Low Var.)
	0.1, 40	Poor (High Var.)	Poor (High Var.)	Excellent (Low Var.)
	0.6, 40	Fair (Mod. Var.)	Fair (Mod. Var.)	Excellent (Low Var.)