Effect of Clustering in Federated Learning on Non-IID Electricity Consumption Prediction

James S. Nightingale*, Yingjie Wang†, Fairouz Zobiri† and Mustafa A. Mustafa*‡
*Department of Computer Science, The University of Manchester, UK
†Dept. of Electrical Engineering (ESAT-ELECTA) & EnergyVille, KU Leuven, Belgium
‡imec-COSIC, KU Leuven, Belgium
Email: {james.nightingale, mustafa.mustafa}@manchester.ac.uk, {tony.wang, fairouz.zobiri}@kuleuven.be

***Abstract*—When applied to short-term energy consumption forecasting, the federated learning framework allows for the creation of a predictive model without sharing raw data. There is a limit to the accuracy achieved by standard federated learning due to the heterogeneity of the individual clients' data, especially in the case of electricity data, where prediction of peak demand is a challenge. A set of clustering techniques has been explored in the literature to improve prediction quality while maintaining user privacy. These studies have mainly been conducted using sets of clients with similar attributes that may not reflect real-world consumer diversity. This paper explores, implements and compares these clustering techniques for privacy-preserving load forecasting on a representative electricity consumption dataset. The experimental results demonstrate the effects of electricity consumption heterogeneity on federated forecasting and a non-representative sample's impact on load forecasting.**

***Index Terms*—Load Forecasting, Federated Learning, Clustering Methods.**

I. Introduction

Accurate load forecasting is necessary for the massive deployment of renewable energy sources. It is essential to balance renewable power generation and optimise power storage within the grid, and for advanced smart grid applications such as a peer-to-peer energy trading. Residential consumption currently represents 23% of such energy demand [1], but is the most heterogeneous. The data streams provided by smart meters will allow for more accurate usage reporting and consumption forecasting.

There is currently some resistance to the adoption of smart meters, with data privacy and security cited as one of the main reasons for the lack of uptake [2]. Hence, methods for forecasting short-term individual households' energy consumption should treat metering data confidentially and securely. The methods should not expose raw energy consumption data to any system that is not controlled by the client.

However, the datasets held by each household are not large enough to create a well-trained model for the household without data from other households. Centralised methods to create more general models have to be used while maintaining privacy. Federated learning (FL) is a method of training a global model without sharing the clients' raw data with a central entity. FL for short term load forecasting (less than 1 day) has mainly been performed on clustered datasets, either by geographical location or other properties. Experiments performed on these datasets produce results specific to these scenarios but may not generalise to a more diverse population.

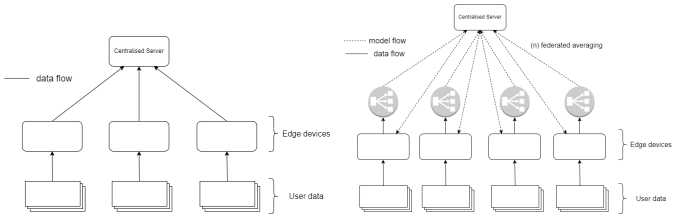
To evaluate the effect of clustering used in energy consumption prediction models based on FL, this work has made the following two main contributions:

- It implements and evaluates a long short-term memory (LSTM) model in predicting power consumption in a centralised (using raw data) and a FL framework.
- It implements and evaluates the effect of different clustering methods on improving the accuracy of the prediction models.

A range of methods have been used for forecasting energy consumption, from simple moving averages [3] to deep learning approaches, such as LSTM models [4], with the latter achieving better results. Centralised learning (CL) is usually used when applying an LSTM model to data from multiple clients. However model performance decreases when the clients' data is not independent and identically distributed (non-IID). To solve this problem, clustering methods that group similar clients together have been proposed [5]. CL also introduces privacy issues. FL addresses these issues and has been applied to several applications from text prediction [6] to handwriting recognition [7], producing comparable results to CL while protecting clients' raw data. It has also been applied to energy forecasting in a range of works such as [8] for a block of offices, or [9] for 200 detached homes in Texas. Clustering has also been introduced in an FL setting [10], [11]. FL clustering been applied to a dataset of homes in London with LSTM [12], and produced better results than individual models. In addition to clustering on the weight updates, other methods of privacy-preserving clustering have been applied, including clustering on hyperparameters of individual models after hyperparameter optimisation [13] and house properties with OPTICS [14].

The rest of this paper is organised as follows. Section II lays out some background on learning, data and clustering. Section III describes our experimental datasets. Section IV performs experiments using different clustering methods and provides a rigorous comparison between them. Section V concludes the work and sets future research directions.

This work was supported in part by the EPSRC through the projects EnnCore EP/T026995/1 and by the Flemish Government through the FWO SBO project SNIPPET S007619. M.A.M. is funded by the DKO Fellowship awarded by The University of Manchester.



(a) Centralised learning overview. (b) Federated learning overview.

Fig. 1: Learning models.

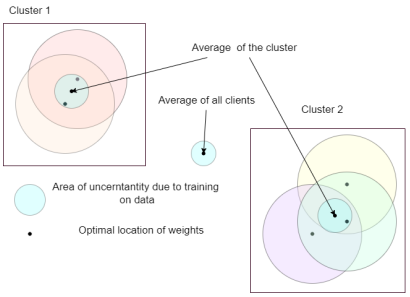


Fig. 2: Federated averaging of non-IID clients in clusters.

II. Background

A. Long Short Term Memory Networks

Long Short Term Memory (LSTM) networks are a type of Recurrent Neural Networks (RNN) developed to address the vanishing gradient problem [15]. They are a type of gated RNN that can regulate what is remembered by the network. They work by replacing the hidden layer in the standard RNN model with an LSTM cell and adding an extra connection between the cell blocks the cell state. The cell state carries extra information along with the hidden state. At each time step, the cell can choose to read from, write to or reset the hidden state using a gating mechanism. This composition of the cell state allows the history information to flow for a more extended period with no risk of vanishing gradients. It can decide what to remember within the cell state and what to forget, instead of holding everything in the network.

B. Learning Styles

1) Centralised Learning (CL): CL is a method to create a centralised model using data held on edge devices such as smart meters (see Fig. 1a). It is performed for two reasons. Firstly, the model has the potential to make predictions on homogeneous systems whose data has not been used within the training process. Secondly, using information from multiple sources allows for better models to be built. This is due to different system data from different scenarios exposing the model to a richer data source. Although CL allows to create better models that have had exposure to a larger number of cases, it raises privacy concerns due to the sharing of raw data with the central server.

2) Federated Learning (FL): FL is a collaborative learning technique that allows the training of a centralised model on data held across multiple edge devices by sharing only gradient updates, as shown in Fig. 1b. Compared to standard machine learning techniques, FL decouples the raw data and the training process. This feature ensures that no other entity can directly extract sensitive information from raw data. It also can reduce substantially the amount of communication needed to train the model in comparison to a standard CL. This decoupling provides security benefits, but in some cases the original data can be reverse-engineered from the weight updates. To mitigate this risk, FL can be combined with other security techniques such as differential privacy [16] and homomorphic encryption [17].

C. Non-Independent and Identically Distributed Data

Within a statistical model for machine learning, there are two samples when working with multiple clients' data. There is $i \sim Q$, the distribution of available clients, and accessing a data point for an individual client $(x, y) \sim P_i(x, y)$ where x is the features and y is the labels. The datasets of two clients i and j are said to be non-IID if for the same set of inputs, the two datasets yield differently distributed outputs.

During training, the difference in the data distributions will cause the model updates to converge to an average location between the optimal weight values. This effect cancels out the accuracy gained by using larger datasets. Fig. 2 represents this effect. It shows that although having more data may improve the model if the data is non-IID, the model created by combining the data may be worse than the individual models.

D. Clustering

Clustering is the classification of data points due to certain similarities and is a common unsupervised machine learning task. A standard method for performing clustering is k-means clustering [18]. It defines the number of clusters k, then places and updates a centroid in the centre of each cluster and assigns the data points according to their distance from the centroids.

III. Data Experimentation

A. Dataset

The dataset used is The Smart Metering Customer Behaviour Trials which was collected between 2009 and 2010 with over 5000 Irish homes and businesses participating [19]. It contained 6445 clients: 4225 residential, 285 SME and 1735 other. The clients with a full set of data were divided into:

Clustered dataset: This set contains 30 of the most similar clients in terms of energy consumption clustered by various methods. It is the most homogeneous dataset, and thus a best-case scenario, as the clients have the most similar profiles, hence causing the least problems when used for CL.

Representative dataset: This set contains 30 of the clients that represent the dataset as a whole for training and testing. This was achieved by stratified random sampling from the main dataset, ensuring that the distribution of residential, commercial and other clients was the same between the selected 30 clients and the full dataset. This helps to ensure external validity to the work as the clients are selected from a large geographical location.

TABLE I: Centralised vs individual models.

Training Data	RMSE	Training time per epoch / per house
Single House	0.163467	10.00 / 10.00
30 Houses	0.138887	331.00 / 10.10

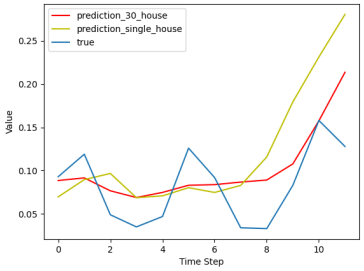


Fig. 3: Example comparison prediction.

TABLE II: RMSE for a single model and per-cluster models.

Framework	RMSE
Single model	0.56440
Model for each cluster	0.52690

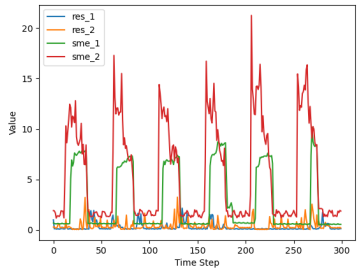


Fig. 4: Example SME vs Residential consumption data.

B. Model Selection and Evaluation Metrics

An LSTM model has been selected due to its common usage for time series forecasting. To compare models, root mean square error (RMSE) and mean absolute percentage error (MAPE) are used. RMSE gives a good measure to compare different models/frameworks that make predictions on data from similar distributions. MAPE gives a better comparison between different models/frameworks, making predictions on different datasets due to the error being scaled to true value.

IV. Evaluation

A. Centralised vs Individual Training

An experiment comparing individual client models and a centralised model was performed, aiming to validate that predictions could benefit from the aggregation of data/models. The data used was the clustered dataset. The clustering step gives an improved selection of clients, minimising the impact of the aggregation on non-IID datasets. An LSTM model was used with hyperparameters selected from an LSTM electrical load forecasting model [20]. The data was split into 80:20 train:test validation. The model predicts the next six hours of consumption using the data from the previous three days.

The results in Table I show an 8% increase in the model's accuracy when training is completed over the clustered dataset compared to the single client model. This comes with longer per epoch but similar per epoch per client training time. The improvement in accuracy, however, only holds if clients have similar power consumption profiles. Such similarity would not be present within a representative dataset, in which case clustering could help. Example predictions are shown in Fig. 3.

B. The Effect of Clustering

1) Clustering on Raw Data: Clients are clustered into similar groups on raw data, which comes with privacy risks. An initial round of k-means clustering was performed using $k = 5$ and data from 6167 clients. The largest cluster (5761 clients) was then clustered too, producing three smaller clusters. In total, data was grouped into seven clusters, with a number of clients within each cluster, ranging from 54 to 2333.

A comparison was performed between a framework that creates a single model for all clients and one that creates a model for each cluster. Thirty clients were selected randomly from the complete dataset for model training and evaluation. The task was to predict the next 30 minutes of power consumption using the previous 22 half-hour readings. This task was selected as it trains faster than the 6-hour future prediction. For the non-clustered framework, an LSTM model was trained on the data, and for the clustered framework, an individual model was trained for each cluster with the same clients.

RMSE was used for the comparison. To create a combined RMSE for the clusters T_{rmse} the equation below was used:

$$T_{rmse} = \frac{\sum_{n=0}^{N} \frac{C^n_{rmse} \times C^n_{count}}{N_{total}}}{N} \quad (1)$$

where C^n_{rmse} is the RMSE for cluster n, C^n_{count} – number of clients from the sample in cluster n, N_{total} – total number of sample datasets and N – total number of clusters. This gives a balanced weighting to each client in the overall RMSE score reducing the impact of a good score from a cluster with only a few clients compared to a score that finds an average.

The comparison results, shown in Table II, indicate a 7.1% decrease in RMSE between the single model and the clustered framework, from 0.56440 to 0.52690.

2) Clustering on Client Types: Factors about a client, such as type (residential, SME), property type (house, apartment) and number of occupants, play a significant role in its energy consumption, thus making these data an ideal candidate to perform clustering upon. The Irish dataset [19] provides extra information about the clients included in the dataset, from their tariff and property allocation to the number of occupants.

Initially clustering was performed between the groups residential, SME and other. Example consumption data from four clients, two residential and two SME, is shown in Fig. 4. On average, the SME has higher power consumption and a more regular consumption pattern. The heterogeneity between the different groups makes them ideal candidates for clustering.

Fifteen clients from each category (residential, SME, other) were randomly sampled to create a subset for training. Four

TABLE III: Results from clustering based on client types.

(a) All types of clients.

Model	RMSE
Residential	0.5336634
SME	1.5565122
Other	0.8768399
Clusters average	0.9890052
Combined	1.0665991

(b) Residential types of clients.

Model	RMSE
Apartment	0.4350906
Terrace	0.5415721
Detached	0.7925013
Clusters average	0.5897213
Combined	0.6314942

different federated models were made from this sample set of clients, one for each of the client property allocations and one containing all of the clients within the sample. The data in each cluster is used to train an LSTM model to predict the next six hours of data using the last three days of measurements. To ensure fair allocation of training resources per cluster, the number of training rounds for each cluster was proportional to the number of clients within the cluster. Residential, SME and other each performed 100 rounds and the combined 300 rounds. These models were then compared using RMSE.

The accuracy of the clustered model improved by 7.2% compared to the single combined model (see Table IIIa), demonstrating that this clustering method improves forecasting accuracy without sharing users' private data. Further improvements may be possible by performing further clustering on other building features such as type, insulation and number of occupants. The same experiment was performed with the samples being taken from the residential clients and the clusters being apartments, terrace houses and detached houses. There is an improvement of 6.6% in RMSE (see Table IIIb), which is a significant improvement and shows that further clustering can improve the forecasting ability of the models.

3) Clustering on Weight Updates: FL works by an iterative process where the central server broadcasts a model to clients, who update the model by using their raw data for training and return only weight updates, which then the server aggregates to create a new model. Clients can also be clustered based on weight updates. The central server performs the clustering upon receiving the weight updates to determine how the clients are allocated to different clusters. Then this clustering information is used to create cluster-specific models.

Thirty random clients were selected from the testing dataset. LSTM model was used to predict the next six hours from the previous three days of data. Five rounds of training were performed and the client weights were used for k-means clustering with four clusters and 40 rounds of training performed for each cluster. Table IV shows the results of our experiment. Compared to the federated model with no clustering, the model with clustered brings 9.2% improvement. This result may be further improved with parameter optimisation such as number of clusters/rounds of training and different distance metrics.

C. Model Comparison

Clustering methods have shown that forecasting accuracy of an LSTM model can be improved beyond individual models within a privacy-preserving framework. However, these results

TABLE IV: No clustering vs clustering on weight updates.

Model	RMSE
Federated no clustering	0.736478
Federated clustering on weight updates	0.673984

TABLE V: Results of predictions from different frameworks.

Model	RMSE	MAPE
Individual	0.562207	67.44599
Federated no clustering	0.648050	78.63596
Federated allocation clustering	0.604902	69.92027
Federated weight update clustering	0.599697	75.14829

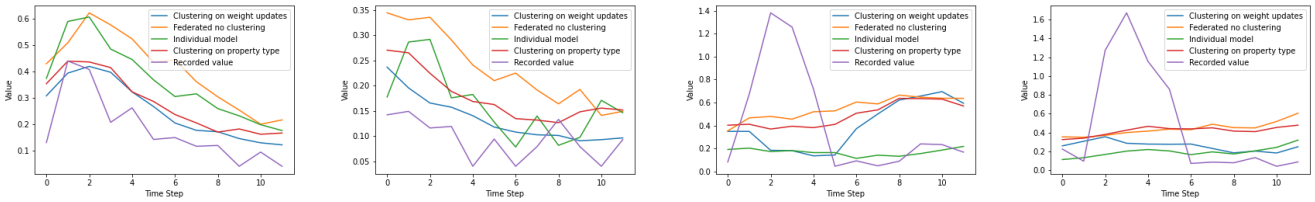
have been produced for clients who have similar features. Next we evaluate the effect of these clustering methods on datasets that are representative of the whole population. We compare four methods of conducting energy forecasting, aiming to find a trade-off between model accuracy and privacy protection.

- Individual: A separate model trained for each client.
- Federated no clustering: A single model for all clients trained using FL.
- Federated allocation clustering: Clients are clustered for Residential/SME/other; each cluster is trained using FL.
- Federated weight update clustering: Clients are clustered on weight updates. Five rounds of training are performed before clients are clustered into five clusters and the rest of the time is spent training using FL.

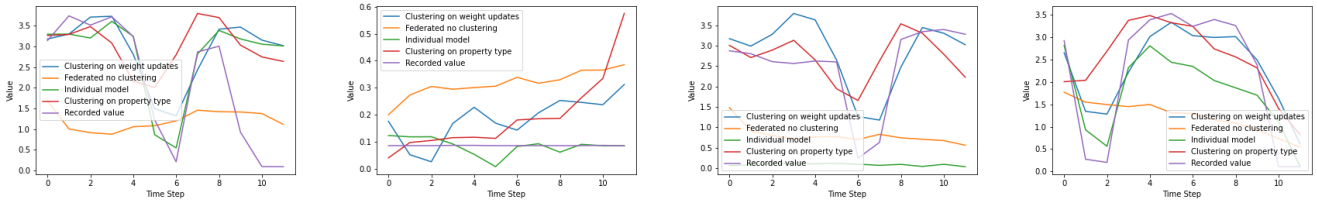
The frameworks are used to produce LSTM models that use the 30-minute data from the last three days to predict the next six hours of data at a 30-minute interval. To ensure fairness, each model is given the same training epochs on the same hardware with RMSE as the loss function. The data is split into an 80:20 test: train split with the models being compared using RMSE and MAPE metrics.

When comparing the models with RMSE, FL with no clustering performs the worst, and the individual model performs the best (see Table V). This shows that the heterogeneity of the different client's data outweighs the benefits from the extra data, contrary to findings from the experiment looking at centralised learning vs individual learning. The clustering methods improve the prediction accuracy by 6.4% for the clustering on client property type allocation and by 7.5% when clustering on the weight updates. These results are similar to those obtained in earlier experiments.

Fig. 5 shows example predictions of two clients, a residential (Fig. 5a) and an SME (Fig. 5b), which highlight the strengths and weaknesses of the different models. As the residential clients dominated the majority of the dataset, the majority of the models (Fig. 5a) fit the general trend of the data but do not perform well when predicting peaks. Predictions of a client SME (Fig. 5b) show the effects of heterogeneous data and the advantages of clustering. As the client is an SME, it has a vastly different consumption distribution from the rest of the clients. This is reflected in the poor predictive performance of the non clustered federated model. This shows the effect of



(a) Consumption profile/predictions of a client - a residential.



(b) Consumption profile/predictions of a client – an SME.

Fig. 5: Example predictions of client power consumption.

non-IID. The poor performance is due to the model training being dominated by data from different distributions. The models that use clustering and the single model have a better prediction accuracy as the training is performed on data with a similar distribution to the client or has a higher proportion of data from the distribution within the training dataset.

V. Conclusions and Future Work

Our work has shown that for clients with similar power consumption, better forecasting models can be produced by creating a central model. This model can be trained in a privacy-preserving distributed environment using federated learning coupled with privacy-preserving clustering techniques to improve the performance of the models.

Furthermore, a representative dataset of a whole population was compiled to compare the models. On such diverse dataset, the performance of the individual models was better than any of the centralised methods. Nevertheless, clustering the models on both building types and weight updates did improve the performance compared to the non-clustered federated learning framework. Although this reduces the value of federated learning for consumption forecasting, it provides a robust method for new clients to be introduced into the system without requiring a backlog of data from them.

Future research includes deployment of privacy-enhancing techniques to protect client weight updates and use of advanced clustering techniques.

References

[1] I. IEA, "World energy statistics and balances, iea," *France*, 2019.

[2] N. Balta-Ozkan, O. Amerighi, and B. Boteler, "A comparison of consumer perceptions towards smart homes in the UK, Germany and Italy: reflections for policy and future research," *Technology Analysis & Strategic Management*, vol. 26, no. 10, pp. 1176–1195, 2014.

[3] C. Yuan, S. Liu, and Z. Fang, "Comparison of China's primary energy consumption forecasting by using ARIMA model and GM (1, 1) model," *Energy*, vol. 100, pp. 384–390, 2016.

[4] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017.

[5] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert systems with applications*, vol. 140, p. 112896, 2020.

[6] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[7] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[8] J. Li, C. Zhang, Y. Zhao, W. Qiu, Q. Chen, and X. Zhang, "Federated learning-based short-term building energy consumption prediction method for solving the data silos problem," in *Building Simulation*, vol. 15, no. 6. Springer, 2022, pp. 1145–1159.

[9] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, M. D. Mueck, and S. Srikanteswara, "Energy demand prediction with federated learning for electric vehicle networks," in *GLOBECOM*, 2019, pp. 1–6.

[10] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[11] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. on NNLS*, vol. 31, no. 9, pp. 3400–3413, 2019.

[12] C. Briggs, Z. Fan, and P. Andras, "Federated learning for short-term residential energy demand forecasting," *arXiv:2105.13325*, 2021.

[13] N. Gholizadeh and P. Musilek, "Federated learning with hyperparameter-based clustering for electrical load forecasting," *Internet of Things*, vol. 17, p. 100470, 2022.

[14] Y. L. Tun, K. Thar, C. M. Thwal, and C. S. Hong, "Federated learning based energy demand prediction with clustered aggregation," in *Int. Conf. on Big Data and Smart Computing*. IEEE, 2021, pp. 164–167.

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

[16] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[17] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in *USENIX Annual Technical Conf. (USENIX ATC 20)*, 2020, pp. 493–506.

[18] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[19] "Home, irish social science data archive." [Online]. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[20] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Multi-sequence lstm-rnn deep learning and metaheuristics for electric load forecasting," *Energies*, vol. 13, no. 2, p. 391, 2020.