Unlock Data Sharing in Wind Power Forecasting through Privacy-preserving Federated-Learning: Benchmarking of Mixed and Fully Encrypted Frameworks

Lukas Stippel, Simon Camal, Georges Kariniotakis

WESC 2025
24 - 27 JUNE | NANTES, FR

Wind Energy Science Conference
24-27 June 2025
Nantes, France

Unlock Data Sharing in Wind Power Forecasting through Privacy-preserving Federated-Learning: Benchmarking of Mixed and Fully Encrypted Frameworks

Lukas Stippel[a], **Simon Camal**[a], and **Georges Kariniotakis**[a]

[a]Center PERSEE, Mines Paris -PSL, Sophia-Antipolis, France

E-mail: Lukas.stippel@minesparis.psl.eu

Keywords: Wind power forecasting, Data sharing, Federated Learning, Distributed Energy Resources

Introduction:

Accurate forecasts of the power output of wind farms is a prerequisite for their efficient integration into power systems and electricity markets. The state of the art on wind power forecasting is particularly rich. Spatiotemporal models have been established among the main-stream approaches to improve predictability in short-term horizons of a few minutes to a few hours (i.e. 6 hours ahead). Using data from geographically distributed wind farms in an area that can span several tens of kilometers has been shown to produce significant accuracy improvements of up to 20% [1] when predicting the output of the wind farm of interest. However, data from neighboring wind farms cannot always be shared because they are owned by different stakeholders, often competitors, because of concerns about losing their competitive advantage. The objective of this work is to unlock these limitations by developing methods that enable collaboration by exchanging past observations through privacy-preserving methods.

A widely used approach for securely sharing data without revealing critical information is federated learning [2]. Instead of sharing the original raw data, federated learning permits the exchange of by-products of machine learning models, typically gradients, transmitting only encrypted gradients instead of data. Federated learning can be categorized into two types based on its application goals. The first type is horizontal federated learning and addresses data scarcity when individual participants lack sufficient data to train a complex model. Examples in wind energy include wind turbine condition monitoring [3] and wind blade icing detection [4].

Our focus is on the second approach, vertical federated learning, which leverages different types of information, making it particularly suited for spatiotemporal forecasting. Here, each data owner contributes with complementary information, improving forecasting accuracy. Applications include forecasting of wind power [1] and of solar power production [5].

However, applying privacy-preserving frameworks to distributed energy resources presents several challenges. First, most relationships between power output and explanatory input data are inherently non-linear, requiring either approximation techniques (e.g., cubic splines) or models capable of capturing these relationships natively. In addition, sensor or communication malfunctions can lead to missing input features. Ensuring convergence guarantees under such operational constraints is crucial for reliable results. Tree-based models naturally fulfil these requirements, offering robustness and resilience in real-world applications. Studies, including [6] and the HEFTCOM 2024 challenge [7], highlight their superior performance compared to neural networks, which require extensive fine-tuning. In our previous work [1], we proposed a multivariate tree-based model for a resilient, lossless forecasting framework. Our results demonstrated higher efficiency and accuracy than univariate lossless tree-based methods, such as those presented in [8]. An important consideration when sharing features is whether all data require encryption. Certain features, such as the hour of the day or potentially numerical weather prediction (NWP) data, do not require encryption and should remain unencrypted within the framework. A mixed encryption framework, which handles both encrypted and unencrypted data, offers a promising approach to enhance scalability and transparency, ensuring that publicly available features are not unnecessarily treated as sensitive. Furthermore, if features are not revealed to each other, it prevents the same feature from being encrypted twice on accident by different participants.

In this contribution, we evaluate the scalability of our wind power forecasting model, comparing the impact of a fully encrypted and mixed encryption framework. Existing federated frameworks [8, 1] are typically assessed on relatively small datasets, often limiting the number of participants or tree depth. We investigate whether mixed

encryption can mitigate scalability challenges as the number of encrypted features grows more slowly than a fully encrypted model. Using an open-source solar dataset, we also analyze how these frameworks scale when adding new participants with a small number of features versus adding fewer participants with more features.

Federated learning with multivariate trees

This section introduces federated multivariate trees, discusses their functionality, and describes the key factors when analyzing scalability. First, we introduce our setting and security assumptions. For our collaborative participants, we assume the so-called semi-honest but curious setting [9]. We assume that participants try to gather as much information as possible about others by analyzing encrypted messages and shared information. They do not, however, collude against each other or send malicious false data and follow the protocol.

Our setting distinguishes between active parties (data owners) and passive participants (publicfeatures).

Suppose we have d active data owners and one public party κ. The set of active data owners is $\mathcal{O} \setminus \{\kappa\}$. For this set, an active owner $i \in [d]$ posses a sample matrix $X_i \in \mathbb{R}^{n \times \zeta_i}$, and a target vector $Y_i \in \mathbf{R}$ usually consisting of power observations. This means we have n past observations with ζ_i features.

The public party κ only contributes with a sample matrix $X_\kappa \in \mathbb{R}^{n \times \zeta_\kappa}$ and does not contribute to the loss.

Now, we introduce the multivariate tree model. Multivariate trees, as shown in [10], are a natural extension of univariate gradient-boosted trees, but they have a d dimensional leaf instead of a one-dimensional one. Let bold variables now describe the d or higher dimensional corresponding vectors. Hence, suppose we have an input vector $\mathbf{x} = [x_1, x_2, \ldots x_d, x_\kappa]$ then a prediction $\hat{\mathbf{y}} = [y_1, y_2, \ldots, y_d]$, is the sum of all T trees predictions. Hence, let $f_t(\mathbf{x})$ describe the prediction of tree t then $\hat{\mathbf{y}} = \sum_{i=1}^{T} f_i(\mathbf{x})$. This also intuitively shows the optimization process, as gradient-boosted trees optimize additively, meaning that in this example, the $(T+1)-th$ tree would correct the previous prediction. This requires a twice differentiable convex loss function. Then, by applying the Taylor approximation, we obtain:

$$\mathscr{L}^{T+1}(\mathbf{y},\hat{y}) = \sum_{i=1}^{n} l(\mathbf{y}_i,\hat{y}_i + \mathbf{g}_i^T \mathbf{w} + \frac{1}{2}\mathbf{w}^T \mathbf{H}_i \mathbf{w} + \frac{\lambda}{2}||\mathbf{w}||_2^2 + \gamma T. \tag{1}$$

Here, $\gamma, \lambda \geq 0$ describe the regularization parameters, $\mathbf{g}, \mathbf{H}$ the gradients and Hessians respectively, and $\mathbf{w}$ the leaves which we want to optimize for. A key criterion for efficiently solving leaf weights in multivariate trees is whether the Hessian matrix is diagonal. This is the case when the loss function is separable. Separability is also crucial for privacy. With separability, each owner can compute their leaf outputs independently. Thus, final predictions remain private and are computed locally, preventing inference attacks. For forecasting, the common loss function is the Root Mean Square Error (RMSE), the RMSE is inherently additively separable allowing localized gain calculations. A detailed representation of the algorithm is given in [1]. The goal of this work is to evaluate the scalability; hence, the encryption will only be briefly explained. Secret-sharing is a multi-party control operating on a ring structure. This means one can perform multiplication and addition. A value is encrypted by owner i by sending the other $d-1$ owners a random value. The share of $i < x >^i$ is build by building the difference of $< x >^i = x - \sum_{j=1 j \neq i}^{d-1} < x >^j$. This allows us to perform the loss-aggregation encrypting by localizing the gain calculations and employing randomization to calculate an encrypted multiplicative inverse for the common tree gain calculation. In our case studies, we evaluate the impact of different data configurations on the framework's scalability. While theoretical properties have been discussed in [1], we focus on numerical scaling effects introduced by Python. Specifically, we compare the efficiency of adding data by increasing features within an existing party versus distributing data among multiple smaller parties. Finally, as prior evaluations were conducted on a small dataset, we extended our experiments to more diverse conditions to assess generalization performance.

Case-studies and results This subsection presents the two case studies on which the analysis is performed and describes the main results.

1. Windpower GEFCom 2014 The first dataset utilized for evaluation is the GEFCom 2014 dataset [11]. It has 10 wind farms in Australia for a period of over two years, from January 01, 2012, to November 30, 2013. Previous evaluations for the multivariate tree model were only done on a small cluster. Here, we can clearly analyse the scaling effect and the difference between full and mixed encryption. The dataset possesses hourly observations of the wind farm's power production normalized by the wind farm's nominal capacity, as well NMPs such as the u and v wind speed components at 10m and 100m heights. It is split into one year of training data set for tuning and one year for evaluation via a sliding window. Feature selection revealed that the past 6 observations are the optimal spatio-temporal input. The results illustrate the benefits in forecasting accuracy and are compared to the case of fully available data for spatiotemporal forecasting.

2. PV data set

The second data set is the one utilized in [12]. It provides 44 generation units localized in a microgrid Evora covering from February 1st, 2011 to March 6th, 2013, allowing for a similar approach with training and testing sets as the previous case study. The relevant past observations were analyzed in [12] to be three features per party, allowing us to evaluate many participants with few features versus the impact of few participants with many features. **Results** Initial experiments showed that mixed provides needed efficiency gains compared to the existing vertical federated learning methods. We expect this to hold up in the full evaluations. Furthermore, we demonstrate that mixed encryption and smart utilization of how the participants are designed by either minimizing the number of data owners and maximizing the number of features per data owner or splitting data owners into smaller parties) is a solution to transfer privacy-preserving methods to a more applied setting.

Acknowledgements

Part of this work is carried out in the framework of the AI.NRGY project, funded by France 2030 (PEPR TASE Programme Grant No: ANR-22-PETA-0004). This work is also supported by the sdw (German Business Foundation)

References

[1] Lukas Stippel, Simon Camal, and Georges Kariniotakis. Multivariate federated tree-based forecasting combining resilience and privacy: Application to distributed energy resources. In *PSCC'2024, 23nd Power Systems Computation Conference*, 2024.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[3] Lorin Jenkel, Stefan Jonas, and Angela Meyer. Privacy-preserving fleet-wide learning of wind turbine conditions with federated learning. *Energies*, 16(17):6377, September 2023.

[4] Xu Cheng, Fan Shi, Yongping Liu, Jiehan Zhou, Xiufeng Liu, and Lizhen Huang. A class-imbalanced heterogeneous federated learning model for detecting icing on wind turbine blades. *IEEE Transactions on Industrial Informatics*, 18(12):8487–8497, 2022.

[5] Carla Gonçalves, Ricardo J. Bessa, and Pierre Pinson. Privacy-preserving distributed learning for renewable energy forecasting. *IEEE Transactions on Sustainable Energy*, 12(3):1777–1787, 2021.

[6] Tim Januschowski, Yuyang Wang, Kari Torkkola, Timo Erkkilä, Hilaf Hasson, and Jan Gasthaus. Forecasting with trees. *International Journal of Forecasting*, 38(4):1473–1481, 2022.

[7] Jethro Browell; Sebastian Haglund; Henrik Kälvegren; Edoardo Simioni; Ricardo Bessa; Yi Wang; Dennis van der Meer. Hybrid energy forecasting and trading competition, 2023.

[8] Lunchen Xie, Jiaqi Liu, Songtao Lu, Tsung-Hui Chang, and Qingjiang Shi. An efficient learning framework for federated xgboost using secret sharing and distributed optimization. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5):1–28, 2022.

[9] Andrew Paverd, Andrew Martin, and Ian Brown. Modelling and automatically analysing privacy properties for honest-but-curious adversaries. *Tech. Rep*, 2014.

[10] Leonid Iosipoi and Anton Vakhrushev. Sketchboost: Fast gradient boosted decision tree for multioutput problems. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25422–25435. Curran Associates, Inc., 2022.

[11] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

[12] Carla Gonçalves, Ricardo J. Bessa, and Pierre Pinson. A critical overview of privacy-preserving approaches for collaborative forecasting. *International Journal of Forecasting*, 37(1):322–342, 2021.