Probabilistic and Explainable Machine Learning for Tabular Power Grid Data

Alexandra Nikoltchovska
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
alexandra.nikoltchovska@kit.edu

Sebastian Pütz
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
sebastian.puetz@kit.edu

Xiao Li
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
xiao.li@kit.edu

Veit Hagenmeyer
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
veit.hagenmeyer@kit.edu

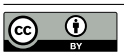
Benjamin Schäfer
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
benjamin.schaefer@kit.edu

Abstract

Modeling power grid frequency stability is becoming increasingly challenging due to the integration of renewable energy sources. Machine learning approaches, such as gradient-boosted trees, have shown promise in analyzing the complex characteristics of power systems. However, these models are inherently deterministic, providing only point estimates. Meanwhile, the task of capturing the underlying uncertainty, particularly through (deep) probabilistic models, is still underexplored, despite its potential to better account for the stochastic nature of power grid dynamics. In this paper, we first compare the performance of TabNet, a deep learning architecture designed for tabular data, to XGBoost for modeling power grid frequency stability. We then present TabNetProba: a probabilistic extension of TabNet, that enables uncertainty-aware estimates comparable to NGBoost. Using these (trained) models, we leverage explainable artificial intelligence (XAI) to analyze the drivers influencing grid stability and identify sources of uncertainty in two major European synchronous areas: Continental Europe and the Nordic region. Our results demonstrate that TabNetProba achieves competitive performance with state-of-the-art methods while providing reliable uncertainty estimates. We find that load and conventional generation ramps, as well as forecast errors, are the key quantities for modeling and explaining mean stability indicators in both synchronous areas. In Continental Europe, renewable generation emerges as a key factor in explaining model uncertainty, while in the Nordic region, load and generation features dominate uncertainty estimation, allowing for more reliable and interpretable stability estimates for modern power systems.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Applied computing** → *Engineering*.



E-ENERGY '25, Rotterdam, Netherlands

ACM ISBN 979-8-4007-1125-1/25/06
https://doi.org/10.1145/3679240.3734623

Keywords

power grid frequency stability, probabilistic machine learning, explainable artificial intelligence, tabular data, deep learning, TabNetProba

ACM Reference Format:
Alexandra Nikoltchovska, Sebastian Pütz, Xiao Li, Veit Hagenmeyer, and Benjamin Schäfer. 2025. Probabilistic and Explainable Machine Learning for Tabular Power Grid Data. In *The 16th ACM International Conference on Future and Sustainable Energy Systems (E-ENERGY '25), June 17–20, 2025, Rotterdam, Netherlands.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3679240.3734623

1 Introduction

The power grid is fundamental to modern society, enabling the reliable delivery of electrical energy to homes, industries, and critical infrastructures. A key challenge in operating this complex system is maintaining stability - keeping the system within operational limits and ensuring it returns to a steady state after disturbances. While stability encompasses many aspects, including voltage levels and phase angles, frequency stability has emerged as a key indicator of overall system health [22]. The integration of renewable energy sources and decentralized generation has transformed modern power systems, introducing new complexities in grid frequency regulation due to their intermittent and less predictable nature. Within this context, accurate modeling of power grid frequency stability has gained importance for grid operators to ensure reliable dispatch decisions and efficient grid operations [17].

Tabular data serves as the primary format for organizing the diverse information needed for stability analysis, including generation types, load patterns, ramp rates, etc. This representation aligns naturally with the discrete hourly intervals of dispatch operations and the temporal definition of grid stability. Traditional modeling approaches, such as physical models and statistical methods, have been employed to analyze this data [8, 38]. However, these conventional methods often prove to be inadequate when confronted with the growing complexity of modern power systems [6].

Machine learning (ML) techniques have been recognized as a promising solution, offering data-driven methods capable of capturing complex relationships and high-dimensional dependencies in tabular data. In particular, gradient-boosted trees (GBTs) have demonstrated success in modeling power grid stability [19].

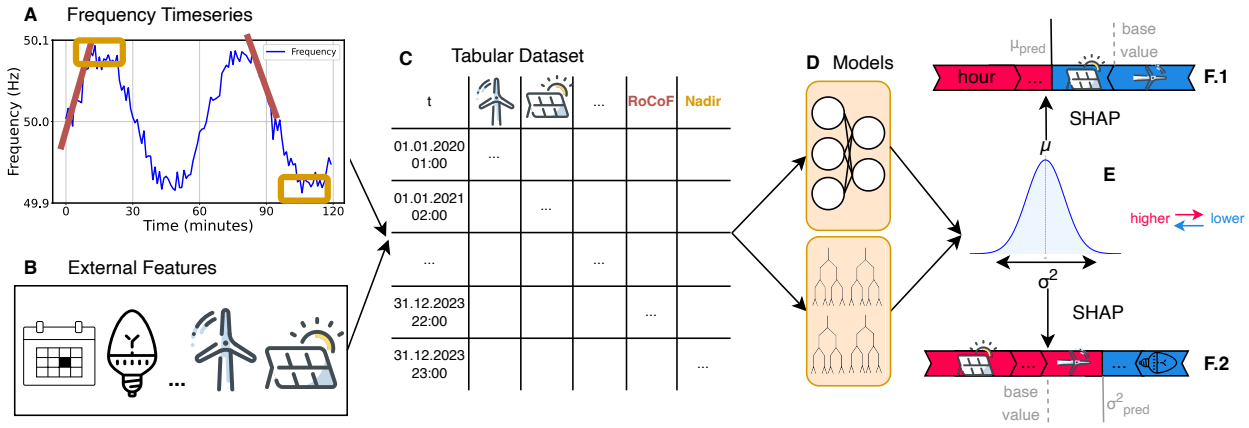


Figure 1: Overview of our probabilistic modeling approach: Starting with high-resolution frequency measurements from the Continental Europe and Nordic power grids (A) and operational data from the ENTSO-E transparency platform (B), we construct a comprehensive tabular dataset with four stability indicators as targets (C). We then employ two probabilistic models - our proposed TabNetProba extension and NGBoost (D) - to estimate the parameters of a Gaussian distribution, capturing both the expected values and uncertainties of power grid stability (E). Finally, we use SHAP to identify key features influencing grid frequency stability (F.1) and drivers of uncertainty (F.2).

The increasing integration of renewable energy introduces complex relationships between generation profiles, meteorological conditions, and grid stability. These relationships, combined with market-driven dynamics, create temporal patterns across multiple timescales (hourly, daily and seasonal), making stability analysis increasingly challenging [4, 36]. Concurrent with this development, deep learning methods have gained prominence in related domains. Deep learning methods offer unique advantages, particularly their capacity to model complex feature interactions and adapt to evolving conditions without extensive feature engineering. While their application in the power grid sector is growing, significant opportunities remain to explore their full potential in addressing grid stability and operational challenges [25, 34].

In recent years, tree-based methods, particularly gradient boosted trees like XGBoost, have established themselves as the standard for tabular data analysis, consistently demonstrating superior performance across various domains [13]. While tree-based approaches remain dominant, there has also been growing interest in adapting deep neural networks for tabular data, leading to architectures like NODE [33], SAINT [40], and FT-Transformer [12]. Among these, TabNet [3] represents one of the first notable attempts to incorporate neural network-based methods with specific inductive biases designed for tabular data. Despite not outperforming tree-based methods in general, TabNet offers certain characteristics that make it interesting for power grid applications: its interpretable feature selection mechanism, ability to capture complex interactions through sequential attention, and potential for extension to probabilistic regression through its neural network foundation.

The inherent stochastic nature of power systems, caused by fluctuating demand, renewable generation, and operational uncertainties, indicates the need for methods that go beyond deterministic estimation. Hence, probabilistic approaches can be a suitable option for capturing the inherent uncertainty of the grid by providing both point estimates and a quantification of uncertainty. While probabilistic extensions of GBTs, such as NGBoost [9], exist, they have not been applied in the power grid context. Moreover, deep learning approaches like TabNet have traditionally focused on deterministic outputs. We therefore propose a probabilistic extension of TabNet to address this limitation, enabling uncertainty quantification in deep learning-based modeling of power grid stability.

Beyond accurate modeling, understanding how models make their decisions is crucial in the context of critical infrastructures like power systems. For grid operators, understanding both the point estimates and their associated uncertainty enables more informed operational decisions - from adjusting dispatch schedules and reserve capacities to planning maintenance windows and emergency responses. While previous work has used SHAP (SHapley Additive Explanations) [21] to interpret the outputs of deterministic models such as gradient-boosted trees in [19], the explainability of uncertainty estimates remains unexplored.

In the present paper, we demonstrate that:

(1) Deep learning approaches for tabular data like TabNet can achieve comparable performance to GBTs in modeling certain aspects of grid stability, though they do not consistently outperform traditional methods across all stability indicators.
(2) Probabilistic methods can effectively model power grid stability while providing valuable uncertainty estimates, advancing beyond traditional deterministic approaches. We illustrate this by utilizing both tree-based methods like NGBoost and extending deep learning architectures through our probabilistic TabNet variant, which achieves comparable uncertainty estimation performance.

(3) By applying the SHAP framework to both deterministic and probabilistic model outputs, we reveal key drivers of power grid frequency stability and sources of uncertainty across different model types, enabling systematic comparison of how different models interpret power grid frequency.

Figure 1 illustrates our methodology, from data preprocessing through model development to explainability analysis.

2 Background and Problem Setting

In this section, we first introduce the concept of power grid frequency stability and the key indicators used to measure it. We then describe the data sources and preprocessing steps used in our analysis. Together, these elements form the foundation for our investigation into probabilistic modeling and explanation of grid stability.

2.1 Power Grid Frequency Stability

Maintaining a stable power grid relies on a balance of supply and demand. This balance is monitored and maintained by keeping the power grid frequency close to its nominal value (e.g. 50 Hz in Europe, 60 Hz in the US), requiring continuous monitoring and control [37]. The power grid frequency is particularly suitable for stability assessment because frequency disturbances propagate rapidly through the grid, making any local measurement effectively a global indication of the power grid stability [32]. A single metric cannot capture all relevant aspects of frequency stability - from rapid fluctuations that may trigger immediate protection measures to sustained deviations that increase operational costs [27]. Therefore, we analyze four complementary indicators that together provide a comprehensive view of grid stability:

(1) The **Rate of Change of Frequency (RoCoF)** describes the steepest slope of the frequency trajectory and measures how rapidly the frequency f changes in short time intervals Δt. It is obtained from the derivative of the frequency time series $\frac{df}{dt}(t)$. High (positive and negative) values are particularly critical as they can trigger protective disconnections before control systems have time to respond [42].
(2) **Nadir** represents the largest positive or negative frequency deviation from the nominal frequency (50 Hz in Europe) within each hour. Exceeding certain Nadir thresholds may require emergency responses such as load shedding to prevent system collapse [37].
(3) The **Mean Squared Deviation (MSD)** quantifies the frequency variability by averaging the squared deviations from the nominal frequency. Higher MSD values indicate that more control effort and resources are needed to maintain grid stability, since larger or more frequent deviations require more active balancing interventions [41].
(4) The **Integral** measures accumulated frequency deviations. It is proportional to the mean deviation within the hour. Large values for this indicator suggest systematic imbalances between generation and demand.

All indicators are calculated using frequency measurements sampled at 1-second intervals following the same methodology as in [19]. See Appendix B.1 for details on the calculation.

2.2 Data Description

We calculate the four stability indicators (RoCoF, Nadir, MSD and Integral) using the frequency time series data from Continental Europe (CE) and the Nordic region. For Continental Europe, we obtain the frequency recordings from the German Transparency Platform for Energy Generation [15], which provides second-by-second measurements. For the Nordic frequency, we use Fingrid's open data platform [31], which also provides data at one-second resolution. We aggregate these measurements on an hourly basis to obtain hourly target indicators. This is done to match the frequency of the input features. As input features for our analysis we use hourly power system data from the ENTSO-E transparency platform [10] for the Continental Europe and Nordic synchronous areas for the years 2020-2023. The dataset includes generation data by source (nuclear, hydro, wind, solar, etc.), load data and ramp rates, day-ahead forecasts and market data as well as temporal features (hour, weekday, month). The dataset comprises a total of 64 features for CE and 58 features for the Nordic region. For methodological consistency, we use all available features for each synchronous area when training every model type and predicting every target. This approach ensures fair comparisons across both different stability indicators and model architectures. A comprehensive list of these features categorized by type (generation, load, ramps, etc.) along with their regional availability is provided in Table 2 of Appendix A. The data exhibits high correlations between features, particularly among related generation types and their ramp rates (see Appendix B.2). Our preprocessing pipeline includes data cleaning, feature engineering, Yeo-Johnson normalization [44], and correlation analysis, see code at [28].

2.3 Experimental Framework

For each synchronous area (CE and Nordic) and each stability indicator (RoCoF, Nadir, MSD, and Integral), we train four distinct models: two deterministic models (XGBoost and TabNet) and two probabilistic models (NGBoost and TabNetProba). All models use the same feature set and training-validation-test split, allowing for direct comparison of model performance across different architectures. The experimental framework is designed as follows:

- Hourly prediction setting: We focus on hourly prediction due to the power system's natural hourly operational cycles, characteristic frequency patterns at hourly boundaries, and the hourly resolution of available ENTSO-E data [20]. We use features available at the beginning of each hour to predict the aggregated stability indicators for that same hour. This approach provides post hoc insights into how system conditions impact frequency stability within operational time frames.
- Feature consistency: All models for a given synchronous area use the whole identical feature set, regardless of the target indicator being predicted. This ensures that performance differences stem from the model architecture rather than from differences in input information.
- Target-specific models: We train separate models for each stability indicator to capture the unique dynamics and dependencies of each target variable, rather than using a multi-target approach.

3 TabNet vs. GBTs for Modeling Power Grid Frequency Stability

For our deterministic models, we evaluate TabNet [3], a neural architecture for tabular data based on sequential attention, against two alternative approaches: a gradient-boosted trees model from [19] and a daily profile baseline capturing system-specific patterns. In this section, we will describe the model architectures, the experimental setup, and discuss the results across both regions and the four stability indicators.

3.1 Model Architectures

TabNet. TabNet is a neural architecture designed specifically for tabular data that combines elements from decision trees and attention mechanisms. The model processes features through multiple sequential decision steps, where each step selectively focuses on different feature subsets using a sparse sequential attention mechanism (sparsemax). This feature selection is controlled by a prior scaling mechanism that encourages feature diversity across steps. Selected features are then transformed through a shared-weight neural network with batch normalization and residual connections. The outputs from all steps are aggregated and processed through a final layer to generate predictions.

Daily Profile Baseline. Power grid behavior exhibits strong daily patterns driven by human activity, natural cycles and electricity market operations. These patterns manifest in consistent frequency stability indicators, making a daily profile an informative baseline. We compute the profile by averaging each indicator's values across the dataset for each hour of the day: $daily\_profile(h) = \frac{1}{D}\sum_{d=1}^{D} y_{d,h}$, where t is the hour, D is the total number of days in the dataset, and $y_{d,h}$ is the indicator value on day d at hour h.

GBT Model. We reproduce the XGBoost model from [19] on our 2020-2023 dataset, maintaining their original configuration to enable direct comparison. This model serves as our second performance benchmark, representing current state-of-the-art approaches in grid stability prediction.

3.2 Experimental Setup

We randomly split the dataset into training (64%), validation (16%), and test sets (20%). Model hyperparameters were optimized using different strategies based on model complexity. For XGBoost, we performed grid search over parameters including max_depth, learning_rate, subsample, and regularization terms. For TabNet, we used random search to efficiently explore its larger parameter space, sampling the number of decision/attention units (n_d, n_a), decision steps, independent and shared feature transformation layers, and an attention relaxation parameter. A detailed overview of the hyperparameter ranges is provided in Table 4 of Appendix D. For details on the used evaluation metrics, refer to Section E.1 in Appendix E. See also the available code [28] for more details.

3.3 Results

The comparison of model performance (see Figure 2) clearly shows that TabNet and XGBoost outperform the daily profile baseline in modeling frequency stability indicators. The models exhibit complementary strengths across different indicators and regions. In

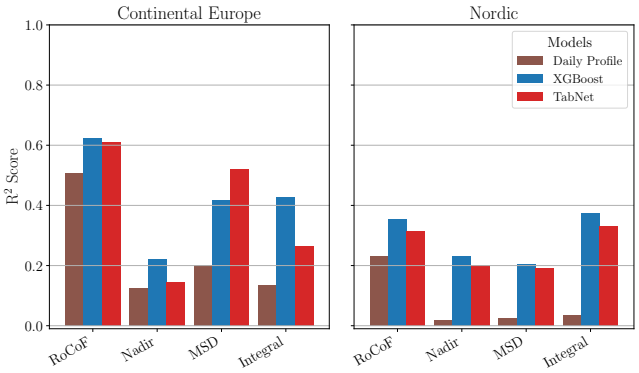


Figure 2: Performance comparison of XGBoost, TabNet, and the daily profile baseline across Continental Europe (left) and Nordic (right) regions. The bars represent the R^2 scores for each stability indicator (RoCoF, Nadir, MSD, and Integral).

Continental Europe, TabNet achieves comparable performance to XGBoost for most indicators. The model particularly excels in MSD predictions where it outperforms XGBoost by a great margin.

The Nordic region, characterized by higher renewable penetration, presents a more challenging prediction environment. XGBoost maintains its lead in RoCoF prediction (R^2 = 0.354), with TabNet still competitive (R^2 = 0.313).

Our analysis highlights three main observations:

(1) TabNet performs competitively with but does not exceed XGBoost in modeling power grid frequency stability. While XGBoost maintains superior performance overall, particularly for RoCoF and Nadir indicators, TabNet shows notable strength in specific cases like MSD predictions in Continental Europe. This indicates that both approaches can effectively predict grid stability indicators, with each method demonstrating strong performance in different scenarios, and neither model consistently outperforming the other by a significant margin. This aligns with broader findings in the literature that gradient boosted trees and deep learning approaches often achieve comparable performance on tabular data tasks, with neither approach demonstrating clear dominance [24, 39]. These findings extend to power grid frequency stability applications, where we observe similar performance patterns between GBTs and deep learning methods.

(2) Both architectures show reduced performance for Nadir predictions across regions, suggesting limitations in capturing the extreme frequency deviations' underlying physics. This performance gap indicates that our current feature set may not adequately represent critical parameters like system inertia and power imbalances that fundamentally drive Nadir behavior.

(3) There is a consistent gap between the performance of Continental Europe and the Nordic regions. While the models in Continental Europe show higher absolute scores, much of this performance is already captured by the daily profile

baseline. In contrast, the Nordic region shows lower absolute performance but higher relative gains over the baseline, suggesting that while modeling grid stability in this region is indeed more challenging, the models are capturing meaningful patterns beyond simple daily cycles.

4 Uncertainty-Aware Grid Stability Prediction

To enable uncertainty-aware predictions of grid stability indicators, we present TabNetProba, our probabilistic extension of TabNet [3] and evaluate it against NGBoost [9], an established natural gradient boosting method, and a probabilistic daily profile baseline that captures system-specific uncertainty patterns. Both NGBoost and TabNetProba estimate the mean (μ) and variance (σ^2) parameters of a Gaussian distribution to approximate prediction uncertainty. For any input x, these models predict $\mu(x)$ and $\sigma^2(x)$, effectively modeling each target y as $y = \mu(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2(x))$. This formulation approximates the distribution of the prediction error ϵ. While mean and variance can fully characterize a Gaussian distribution, they serve here as a computationally tractable approximation that may not capture higher-order moments like skewness or kurtosis.

4.1 Model Architectures

TabNetProba. We adapt TabNet's architecture to incorporate predictive uncertainty estimation. The key modifications to the original architecture enable the model to capture both the expected value and the uncertainty of its predictions. The primary adaptation involves replacing TabNet's standard loss function with the negative log-likelihood (NLL) [29]. Instead of optimizing conventional error metrics like mean squared error, our probabilistic variant maximizes the likelihood of the observed data under the predicted probability distribution. In the final layer of the encoder network, we modify the output structure to generate two parameters: the mean (μ), representing the expected target value (analogous to TabNet's original point estimate), and the variance (σ^2), quantifying predictive uncertainty by estimating the spread of possible outcomes. The core architectural components of TabNet, including its feature selection mechanism and sequential attention layers, remain unchanged from the original implementation (see Figure 11 of Appendix C).

Daily Profile Uncertainty Baseline. To complement the daily profile point prediction baseline, we use the profile's standard deviation to establish a baseline for uncertainty estimation. This approach assumes that both the mean and uncertainty of the system follow consistent daily patterns. For each hour, we calculate the standard deviation of values occurring at that time across all training days, providing a reference for typical daily variability. While this baseline cannot capture dynamic system changes or disturbances, it provides a consistent benchmark for evaluating both the point predictions and uncertainty estimates of like NGBoost and TabNetProba.

NGBoost. Natural Gradient Boosting (NGBoost) [9] extends traditional gradient boosting to predict probability distributions rather than point estimates. Instead of just predicting a single value, NGBoost estimates both the expected value and its uncertainty by learning to predict parameters of a probability distribution. While traditional boosting methods like XGBoost focus solely on improving accuracy, NGBoost optimizes its predictions to also capture how confident it is in those predictions. In our implementation, we configure NGBoost to output Gaussian distributions, providing both a mean prediction for grid stability indicators and an estimate of the prediction uncertainty.

4.2 Experimental Setup

For the probabilistic models, we maintain the same 64:16:20 train-validation-test split and employ random search for hyperparameter optimization. For NGBoost, we focus on base learner configuration and boosting parameters, while for TabNetProba, we explore the same architectural parameters as deterministic TabNet. A detailed overview of the hyperparameter ranges is provided in Table 4 of Appendix D.

For training, we utilize the Negative Log-Likelihood (NLL) loss function, which optimizes the models to predict accurate probability distributions [29]. For evaluation, we compute the Continuous Ranked Probability Score (CRPS) [11, 23], which is a standard metric for judging the quality of the predictive distribution and is widely adopted in probabilistic forecasting of power systems [5, 43]. CRPS considers both the sharpness and calibration of a predictive cumulative distribution function F:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}(z \geq y))^2 \, dz,$$

with the indicator function $\mathbf{1}(z \geq y)$ that is one if $z \geq y$ and otherwise zero with y being the observed value. A lower CRPS score indicates a better prediction. We calculate the CRPS using a sample-based variant using the `properscoring` library. It is important to note that the magnitude of the CRPS values is influenced by the scale of the respective dataset and that a smaller value of CRPS indicates a better performance of the algorithm. Since we scaled our data set, we can still compare the error values of the different models across areas and targets. See Appendix E for more details on the evaluation metrics. See the available code [28] for more details.

4.3 Results

Our evaluation reveals that TabNetProba successfully extends TabNet's capabilities to probabilistic prediction while maintaining competitive performance with established methods. We analyze both the accuracy of point estimates and the quality of uncertainty quantification across multiple stability indicators in two major European power grids.

Both models demonstrate substantial improvements over the daily profile baseline across all stability indicators, as shown by their CRPS scores (Table 1). In Continental Europe, NGBoost is particularly good at predicting RoCoF, while TabNetProba is best at MSD prediction. For Nadir prediction (typically the most challenging indicator), TabNetProba provides a marginally better uncertainty quantification than NGBoost. Figure 3 illustrates this performance difference for RoCoF predictions in Continental Europe for both models. While the mean predictions of NGBoost and TabNetProba closely track the daily profile baseline values, their prediction intervals are substantially narrower while still capturing the true values. This improved uncertainty quantification is reflected in the

Area	Target	Daily Profile	NGBoost (Improvement [%])	TabNetProba (Improvement [%])
CE	RoCoF	0.388	**0.336 (13)**	0.340 (12)
	Nadir	0.535	0.513 (4)	**0.507 (5)**
	MSD	0.473	0.388 (18)	**0.385 (19)**
	Integral	0.509	**0.470 (8)**	0.478 (6)
Nordic Area	RoCoF	0.499	**0.454 (9)**	0.464 (7)
	Nadir	0.583	**0.513 (12)**	0.519 (11)
	MSD	0.562	**0.514 (9)**	0.530 (6)
	Integral	0.542	0.453 (16)	**0.446 (18)**

Table 1: Comparison of CRPS scores and improvement over daily profile (in %) for NGBoost and TabNetProba for both synchronous areas and all four stability indicators. The CRPS scores were rounded to the third decimal place, and percentage improvements were calculated based on these rounded scores, with percentages rounded to whole numbers.

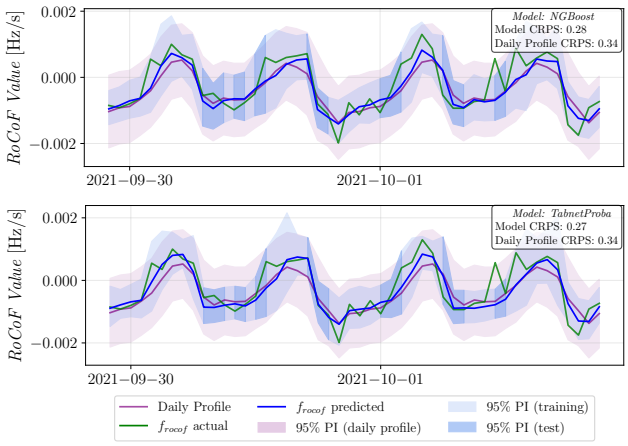


Figure 3: Actual and predicted RoCoF in Continental Europe (CE) for two randomly chosen days using NGBoost (upper) and TabNetProba (lower). Green lines show observed values, blue lines show model predictions, and purple lines represent the daily profile baseline. Shaded regions depict 95% prediction intervals: light purple for the daily profile and blue for model predictions (light indicates training data, dark - test data). CRPS scores (top-right) quantify probabilistic forecast accuracy, with lower values indicating better performance.

CRPS scores, where both models achieve values around 0.27-0.28 compared to the baseline's 0.34.

The Nordic region presents a similar pattern, with both models outperforming the baseline despite the increased prediction difficulty in this region. NGBoost shows stronger performance for RoCoF and Nadir predictions, while TabNetProba outperforms the baseline and NGBoost in Integral predictions. This pattern indicates that the methods achieve similar results through different approaches to capturing prediction uncertainty.

Analysis of uncertainty estimation quality through calibration plots (see Figures 13 and 14, Appendix E) validates the suitability of the Gaussian approximation for modeling predictive uncertainty. Both models demonstrate strong calibration for RoCoF and MSD across both regions, with predicted confidence intervals closely matching observed frequencies of target values. For Nadir predictions, while the models show some deviation at lower confidence levels, they maintain reliable calibration at operational confidence levels (80% and above). The consistency in calibration patterns between NGBoost and TabNetProba, despite their different architectures, suggests that the Gaussian approximation successfully captures the essential characteristics of uncertainty in grid stability predictions.

5 Explaining Model Decisions

In this section, we analyze the drivers of grid stability and their associated uncertainties, focusing on RoCoF in Continental Europe and Nadir in the Nordic region. These indicators were selected to provide complementary insights: RoCoF in CE represents a highly predictable case with clear physical relationships, while Nordic Nadir is a good example of the challenges of predicting complex grid behaviors.

5.1 Challenges in Explaining Power System Models

While SHAP (SHapley Additive exPlanations) [21] values have proven effective for analyzing drivers of grid stability [19], extending these explanations to probabilistic predictions presents new challenges. A limitation of SHAP lies in its assumption of uncorrelated features so that traditional implementations produce misleading attributions when dealing with correlated data [1]. Indeed, power grid data often exhibits complex correlation structures arising from physical constraints, market dynamics, and operational relationships (see the correlation heatmap in Figure 10 of Appendix B.2). For instance, the amount of synchronous generation correlates strongly with the load in the Nordic region, or the hard coal generation is strongly correlated with the lignite generation in Continental Europe.

5.2 Towards Reliable Model Explanations

We address these challenges by following the approach described in [26]. We use the Partition SHAP method, which extends traditional SHAP by computing Owen values [30]. These account for feature interactions within predefined groups while maintaining independence between groups. This approach, combined with correlation-based clustering, enables systematic comparison of explanations across different model types while preserving meaningful feature relationships.

Correlation-Based Feature Clustering. While various approaches exist for grouping correlated features, we implement the hierarchical clustering approach proposed by [26], which groups features based on their Pearson correlation distance through the following steps: First, we compute the correlation distance between two features i and j, defined as $distance(i, j) = 1 - |correlation(i, j)|$. Then, we apply hierarchical clustering using the complete linkage method, which preserves feature relationships while still separating distinct feature groups. Finally, we select correlation distance cutoffs (0.8 for CE and 0.85 for the Nordic region) to form final feature clusters. The cutoffs are used only for better interpretation

of the results and don't change the final importance calculations of the individual features. We choose these thresholds to balance granularity with meaningful feature relationships - for example, grouping solar generation with related solar ramp features while keeping them distinct from wind generation clusters.

The resulting cluster hierarchies are used as inputs to the Partition SHAP explainer, which addresses feature correlations by preserving interactions within clusters while maintaining independence assumptions between clusters. See Figures 15 and 16 in Appendix F for the resulting clustering hierarchies of both synchronous areas.

Quantifying Feature Importance. In order to allow for a systematic comparison of SHAP values, the relative importance of each feature for each model is computed as in [35]. The relative feature contribution is defined as the proportion of the total SHAP contributions attributed to a specific feature. This metric is calculated as follows:

First, we compute the mean absolute SHAP value for each feature across all instances, representing the average contribution of that feature to the model's predictions. Next, the relative feature contribution is obtained by dividing the mean absolute SHAP value of a feature by the sum of the mean absolute SHAP values across all features. Formally, this can be expressed as:

$$\text{Relative Feature Importance}_j = \frac{\frac{1}{M}\sum_{k=1}^{M} |\text{SHAP Value}_{j,k}|}{\sum_{i=1}^{N} \frac{1}{M}\sum_{k=1}^{M} |\text{SHAP Value}_{i,k}|},$$

where j indexes the feature of interest, M is the total number of instances, N is the total number of features. The denominator ensures that the contributions sum to 1, providing a normalized measure of feature importance.

5.3 Core Drivers of Power Grid Frequency Stability

Following the described approach, we obtain the relative feature contributions for the mean predictions of the RoCoF stability indicator in Continental Europe and Nadir in the Nordic region.

Figures 4 and 5 illustrate the relative SHAP feature importance across the four trained models for both synchronous areas. The results show that four feature clusters consistently drive RoCoF predictions across all models, with varying relative importance:

- **Load and generation ramps:** This cluster, containing features like *load ramp*, *load ramp day-ahead*, *total generation ramp*, and *generation ramp day-ahead*, shows the highest overall importance. TabNet assigns it more than a third of the total importance, substantially higher than XGBoost's attribution of about a quarter. For both probabilistic models, this cluster ranks as the second most important, accounting for 20% of the feature importance.
- **Gas, hydro and hard coal ramps:** This cluster consists of ramps of fast-responding (*gas*, *hydro*) and conventional (*hard coal*) generation sources. Its importance varies between models. For NGBoost and TabNetProba, it is the cluster of highest importance, accounting for about one-fifth of the total feature importance, highlighting the significance of these conventional generation sources in modeling the RoCoF stability indicator.
- **Biomass, pumped hydro and price day-ahead ramps:** This cluster demonstrates consistently high importance across models but with varying rankings. For NGBoost and XGBoost, it ranks as the second most important cluster, whereas both TabNet and TabNetProba assign it a lower importance.
- The **solar ramps cluster** plays an important for TabNet's prediction, whereas the other models assign it a considerably lower importance.

For the Nadir mean prediction of the Nordic region, the models demonstrate a higher degree of agreement in feature importance rankings, particularly for the top two clusters (see Figure 5):

- **Load, generation and hydro ramps:** This comprehensive cluster includes a range of key operational parameters: *load ramp*, *load ramp day-ahead*, *reservoir hydro ramp*, *run-off-river hydro ramp*, *total generation ramp*, and *generation ramp day-ahead*. For XGBoost, NGBoost, and TabNet, this cluster stands out as the most influential predictor. However, TabNet assigns it relatively less importance, suggesting a different interpretation of how these fundamental grid parameters affect extreme frequency deviations.
- **Forecast error generation ramp:** This single-feature cluster consistently ranks among the top predictors across all models, particularly for TabNet where it serves as the dominant feature.

Our analysis reveals no consistent pattern in feature importance between models of similar types. For instance, there is no consistent alignment between the probabilistic models (NGBoost and TabNetProba), between the TabNet variants, or among tree-based models.

5.4 Sources of Prediction Uncertainty

In this subsection, we analyze how the uncertainty estimates of the probabilistic models can be explained, and investigate how these explanations differ from the ones derived from mean predictions. Figures 6 and 8 show the relative SHAP feature importance for uncertainty predictions in Continental Europe and the Nordic region respectively, while Figures 7 and 9 provide detailed beeswarm plots illustrating how specific feature values influence uncertainty estimates.

When analyzing uncertainty drivers for RoCoF predictions in Continental Europe, our results reveal distinct patterns that differ notably from those observed when explaining mean predictions. While mean predictions are primarily driven by four ramp-related clusters, uncertainty attributions show a broader and more nuanced distribution across features.

The cluster containing **other renewable** and **biomass generation** emerges as the dominant driver of uncertainty, with markedly different attribution levels between models. It accounts for one fourth of NGBoost's and 17% of TabNetProba's uncertainty predictions, making it the single most influential factor in uncertainty estimation. This finding is particularly interesting given that this generation type plays a relatively minor role in mean predictions and represents only a small fraction of total generation capacity.

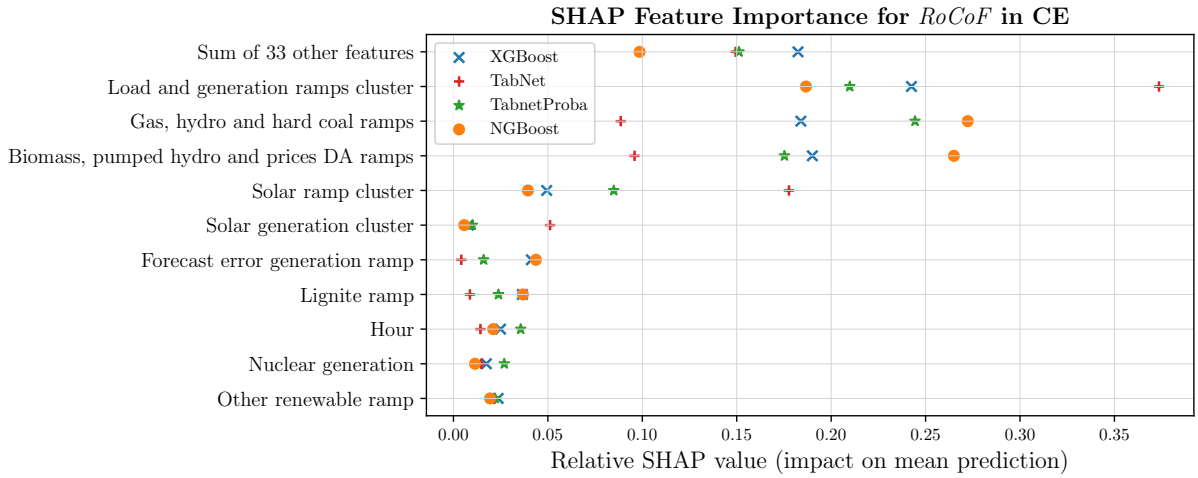


Figure 4: Relative SHAP feature importance for the point prediction of the RoCoF stability indicator in the Continental Europe region.

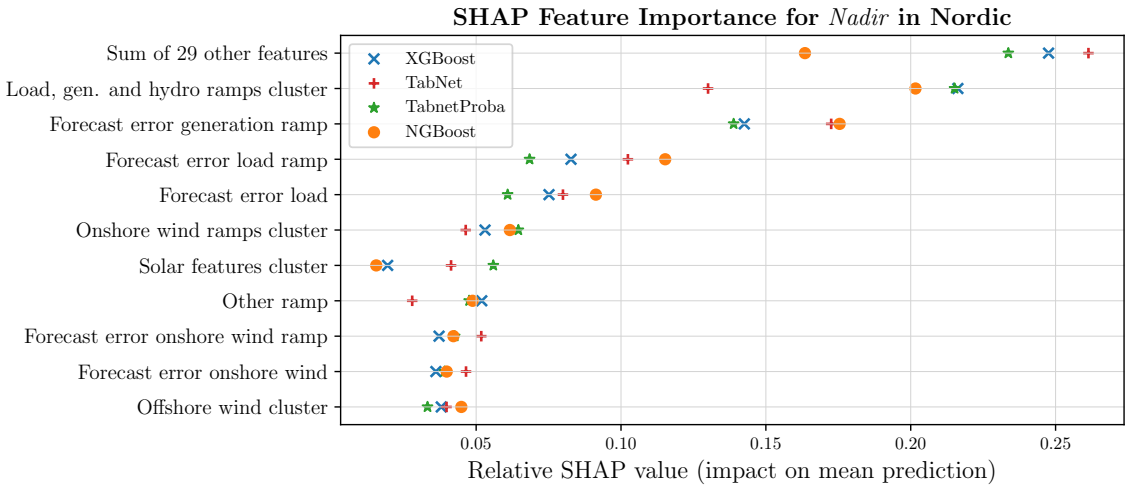


Figure 5: Relative SHAP feature importance for the point prediction of the Nadir stability indicator in in the Nordic region.

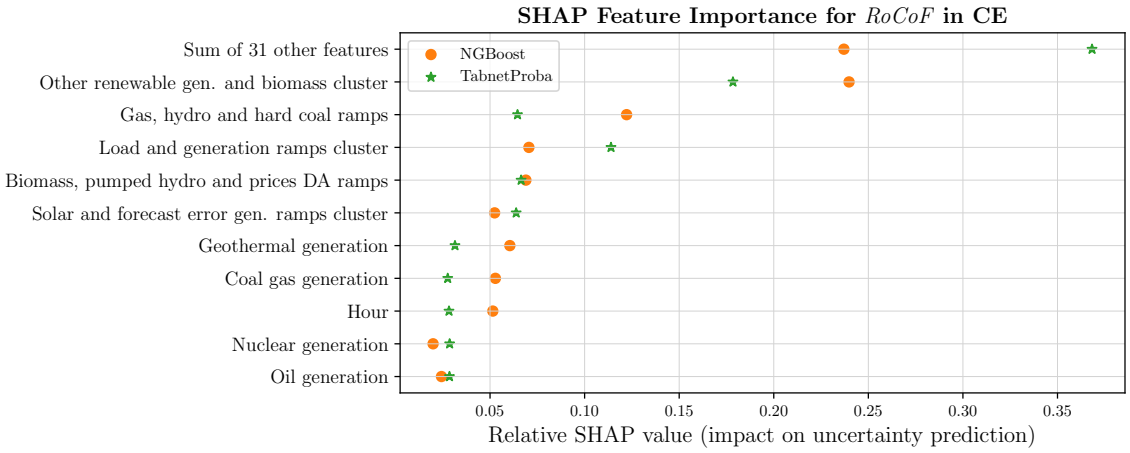


Figure 6: Relative SHAP feature importance for the uncertainty prediction of the RoCoF stability indicator in CE.

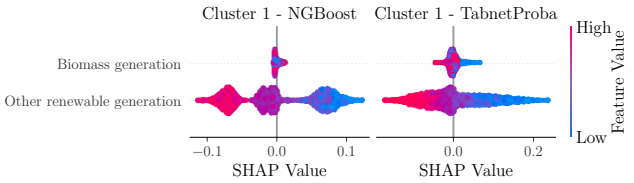


Figure 7: Beeswarm plot illustrating SHAP values for the most influential feature cluster in RoCoF uncertainty prediction for the Continental Europe dataset. Each point represents a SHAP value for a specific data instance, with color indicating the feature value. The plot highlights the contribution of individual feature values to the model's predictions.

Examining detailed effects through the beeswarm analysis in Figure 7 reveals a counterintuitive relationship: higher values of *other renewable generation* consistently correspond to decreased uncertainty in both models' predictions. This suggests that despite its limited direct impact on RoCoF values, other renewable generation may serve as a proxy for broader system conditions that enhance predictability. This stabilizing effect on uncertainty, despite minimal influence on mean predictions, indicates that other renewable generation might capture underlying grid dynamics that are not directly observable but influence system predictability.

The ambiguous nature of *other renewable generation* in our dataset complicates interpretation of these results. The ENTSO-E transparency platform [10] does not offer detailed specifications of what qualifies as *other renewable generation*, and this classification varies across different countries within the Continental European grid. While we can observe and quantify the statistical relationship between this feature and prediction uncertainty, we cannot definitively attribute this effect to specific types of renewable generation.

Similar to RoCoF, no single factor fully explains the uncertainty predictions for the Nadir stability indicator in the Nordic region, underscoring the complex nature of uncertainty in frequency stability modeling.

For NGBoost, the **load, reservoir hydro, and generation cluster** is the most significant contributor, consisting of six features: *load*, *reservoir hydro generation*, *total generation*, *synchronous generation*, *scheduled generation* and *load day-ahead*. This cluster captures the complex interactions between load and generation that are likely to affect the system's response to frequency deviations and is the second most influential one, accounting for over 20% of the uncertainty. This cluster includes features related to load and generation variability (such as load ramp, total generation ramp, and day-ahead load ramp) that introduce variability that NGBoost captures in its uncertainty estimate.

In contrast, TabNetProba's predictions emphasize the **load, generation, and hydro ramps cluster** as the primary uncertainty driver, with other clusters contributing significantly less, each with a relative SHAP value near or below 0.1. This contrast suggests that TabNetProba prioritizes short-term fluctuations in load and generation over broader dynamics when attributing uncertainty.

The beeswarm plots in Figure 9 clearly show that the distribution of SHAP values is random and there is no obvious pattern to indicate what drives uncertainty up or down. The color scheme demonstrates that certain features are associated with both high and low SHAP values, regardless of their value. Examples of this include synchronous generation or load ramp. However, reservoir hydro generation is an exception. Low values here increase the uncertainty in NGBoost's predictions, while higher values reduce it. This is consistent with the intuition that stable reservoir hydro generation provides a buffering effect, reducing uncertainty by increasing system stability during variability.

To put these observations into context, it is useful to revisit the concept of the Nadir, which is defined as the maximum deviation of the frequency from its nominal value, whether that deviation is high or low. Hence, the Nadir represents the situation with the greatest power surplus or deficit within the hour. It is well established that qualitatively different dynamics occur depending on whether this deviation is extreme in a positive or negative direction. Consequently, the analysis suggests that further research could benefit from exploring alternative definitions of the Nadir, such as isolating cases of only high or only low Nadirs, and deriving different explanations for situations with negative and positive Nadirs.

6 Conclusion

Understanding and explaining uncertainty in power grid stability predictions is becoming increasingly critical as power systems evolve to incorporate more renewable sources. Our work demonstrates how modern approaches can not only model power grid frequency stability but also quantify and explain the sources of uncertainty in these predictions. In this work, we adapt TabNet to provide uncertainty estimates by adding probabilistic outputs while maintaining its core architecture for tabular data analysis. While we use this adaptation for power grid stability modeling, the approach generalizes to other domains requiring uncertainty quantification in tabular data analysis.

Transparency of machine learning models for critical systems, such as the power grid, is essential for their deployment and acceptance. Going beyond traditional SHAP, we demonstrate how Partition SHAP makes machine learning models for tabular data transparent while respecting underlying correlations between features. Our analysis reveals consistent patterns across models while highlighting important differences in how they interpret grid behavior. For Continental Europe's RoCoF predictions, ramp features emerge as the most influential across all models, though their exact ranking varies. This is consistent with earlier findings [19]. Interestingly, our analysis reveals that even structurally similar models can be influenced by different features in their decisions. For instance, the deterministic and probabilistic versions of TabNet often differ as much in their important features as they do compared to completely different model types such as tree-based approaches. However, the Nordic region's Nadir predictions show more consistency across models in terms of which features matter and how they affect predictions, suggesting that despite methodological differences, the models capture similar underlying patterns.

Our uncertainty analysis reveals distinct patterns across regions and indicators. In Continental Europe's RoCoF predictions, *other renewable generation* is a dominant uncertainty driver, though its impact may be a proxy for unobserved system conditions. The

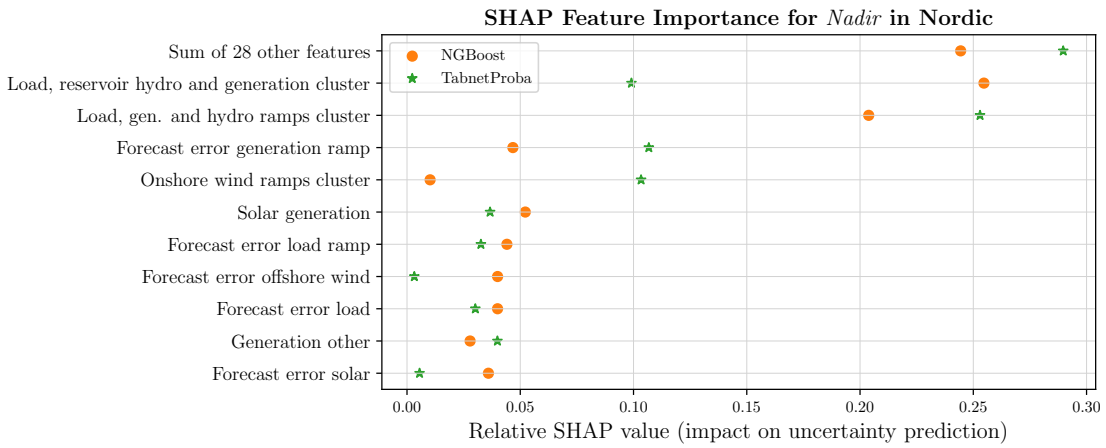


Figure 8: Relative SHAP feature importance for the uncertainty prediction of the Nadir stability indicator in the Nordic region.

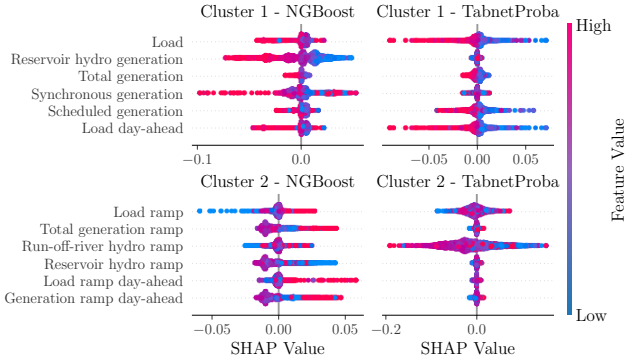


Figure 9: Beeswarm plot illustrating SHAP values for the most influential feature clusters in Nadir uncertainty prediction for the Nordic area.

Nordic region shows more distributed uncertainty drivers for Nadir predictions. The different uncertainty attribution patterns between NGBoost and TabNetProba suggest value in ensemble approaches for more robust uncertainty assessment.

Our work highlights that SHAP explanations require careful interpretation in complex systems with correlated features and indirect relationships. While SHAP reliably identifies important features, distinguishing between direct influences and proxy effects remains challenging. For instance, the importance of *other renewable generation* in uncertainty prediction in CE likely stems from its correlation with broader system conditions rather than direct causation. This suggests the need for both better theoretical frameworks to separate direct and indirect effects in SHAP explanations, and domain-specific guidelines for interpreting SHAP values in power systems.

In summary, our findings provide understanding of the distinct features driving mean estimates versus uncertainty distributions - an important contribution for grid operators to make risk-aware decisions. Furthermore, our application of Partition SHAP to handle the correlated data typical for energy systems represents a methodological advancement for explaining complex relationships in critical infrastructures.

With TabNetProba and Partition SHAP successfully applied to quantify and explain power grid stability, there remain several promising future research questions. First, development of composite features that better capture the directionality and magnitude of grid events, particularly for complex indicators like Nadir, could improve model performance. Second, the integration of conformal prediction techniques [2] could further improve the reliability of uncertainty estimates. Exploring causal extensions to SHAP [14] could help differentiate between direct influences and proxy effects in both mean predictions and uncertainty estimates, particularly for renewable generation's impact on grid stability. Future work will also include a comparison with physics-based and physics-informed machine learning approaches that combine domain knowledge with neural networks [16, 18]. Finally, exploring how local frequency fluctuations and spatial aspects of load fluctuations affect the frequency stability will be considered for future work.

Acknowledgments

This research is funded by the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI under grant no. VH-NG-1727 and the Sino-German (CSC-DAAD) Postdoc Scholarship Program (91870333). We also acknowledge support by the state of Baden-Württemberg through bwHPC.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2020. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. doi:10.48550/arXiv.1903.10464 arXiv:1903.10464 [stat].

[2] Anastasios N. Angelopoulos and Stephen Bates. 2022. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511 [cs.LG] https://arxiv.org/abs/2107.07511

[3] Sercan Arik and Tomas Pfister. 2021. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (05 2021), 6679–6687. doi:10.1609/aaai.v35i8.16826

[4] Sabine Auer and Tim Kittel. 2020. Modeling the dynamics and control of power systems with high share of renewable energies. doi:10.48550/arXiv.2012.05164 arXiv:2012.05164 [physics].

[5] Jonathan Berrisch and Florian Ziel. 2024. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. *International Journal of*

Forecasting 40, 4 (Oct. 2024), 1568–1586. doi:10.1016/j.ijforecast.2024.01.005

[6] Otavio Bertozzi, Harold R. Chamorro, Edgar O. Gomez-Diaz, Michelle S. Chong, and Shehab Ahmed. 2024. Application of data-driven methods in power systems analysis and control. *IET Energy Systems Integration* 6, 3 (2024), 197–212. doi:10.1049/esi2.12122 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/esi2.12122.

[7] Peter J. Bickel and Kjell A. Doksum. 2015. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, New York. doi:10.1201/9781315369266

[8] Youhong Chen. 2024. Power System Modelling and Analysis Techniques. In *Stability Assessment of Power Systems with Multiple Voltage Source Converters: Bifurcation-Theory-Based Methods*, Youhong Chen (Ed.). Springer Nature Switzerland, Cham, 47–86. doi:10.1007/978-3-031-63095-8_2

[9] Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. 2020. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2690–2700. https://proceedings.mlr.press/v119/duan20a.html ISSN: 2640-3498.

[10] ENTSO-E. 2024. ENTSO-E Transparency Platform. https://transparency.entsoe.eu/. Accessed: 2024-10-07.

[11] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102 (March 2007). doi:10.1198/016214506000001437 Publisher: Taylor & Francis.

[12] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. https://openreview.net/forum?id=i_Q1yrOegLY

[13] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on tabular data? doi:10.48550/arXiv.2207.08815 arXiv:2207.08815 [cs].

[14] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 4778–4789. https://proceedings.neurips.cc/paper_files/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[15] Steuerungsgruppe Horizontaler Belastungsausgleich (SG HoBA). 2024. German Transparency Platform for Energy Generation. https://www.netztransparenz.de/de-de/Regelenergie/Daten-Regelreserve/Sek%C3%BCndliche-Daten Accessed: 2024-10-26.

[16] Bin Huang and Jianhui Wang. 2023. Applications of Physics-Informed Neural Networks in Power Systems - A Review. *IEEE Transactions on Power Systems* 38, 1 (Jan. 2023), 572–588. doi:10.1109/TPWRS.2022.3162473

[17] Raja Kandukuri and Yaser M. Banad. 2024. Forecasting Smart Grid Stability Using Machine Learning Models. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*. 214–217. doi:10.1109/AIxSET62544.2024.00041

[18] Johannes Kruse, Eike Cramer, Benjamin Schäfer, and Dirk Witthaut. 2023. Physics-Informed Machine Learning for Power Grid Frequency Modeling. *PRX Energy* 2, 4 (Oct. 2023), 043003. doi:10.1103/PRXEnergy.2.043003 Publisher: American Physical Society.

[19] Johannes Kruse, Benjamin Schäfer, and Dirk Witthaut. 2021. Revealing drivers and risks for power grid frequency stability with explainable AI. *Patterns* 2, 11 (Nov. 2021), 100365. doi:10.1016/j.patter.2021.100365

[20] Jeremy Lin and Fernando H. Magnago. 2017. *Electricity Markets: Theories and Applications: Theories and Applications* (1 ed.). Wiley. doi:10.1002/9781119179382

[21] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[22] Masayoshi Mase, Art B. Owen, and Benjamin Seiler. 2020. Explaining black box decisions by Shapley cohort refinement. arXiv:1911.00467 [cs.LG] https://arxiv.org/abs/1911.00467

[23] James E. Matheson and Robert L. Winkler. 1976. Scoring Rules for Continuous Probability Distributions. *Management Science* 22, 10 (1976), 1087–1096. https://www.jstor.org/stable/2629907 Publisher: INFORMS.

[24] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Benjamin Feuer, Chinmay Hegde, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. 2024. When Do Neural Nets Outperform Boosted Trees on Tabular Data? doi:10.48550/arXiv.2305.02997 arXiv:2305.02997 [cs].

[25] Seyed Mahdi Miraftabzadeh, Andrea Di Martino, Michela Longo, and Dario Zaninelli. 2024. Deep Learning in Power Systems: A Bibliometric Analysis and Future Trends. *IEEE Access* 12 (2024), 163172–163196. doi:10.1109/ACCESS.2024.3491914 Conference Name: IEEE Access.

[26] Chistoph Molnar. 2023. *Interpreting Machine Learning Models With SHAP*. Chistoph Molnar c/o MUCBOOK, Heidi Seibold. https://leanpub.com/shap

[27] Marcel Nedd, Waqquas Bukhsh, Callum MacIver, and Keith Bell. 2021. Metrics for determining the frequency stability limits of a power system: A GB case study. *Electric Power Systems Research* 190 (Jan. 2021), 106553. doi:10.1016/j.epsr.2020.106553

[28] Alexandra Nikoltchovska, Sebastian Pütz, Xiao Li, Veit Hagenmeyer, and Benjamin Schäfer. 2025. prob-xai-power-grid. https://github.com/KIT-IAI-DRACOS/prob-xai-power-grid

[29] D.A. Nix and A.S. Weigend. 1994. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 1. 55–60 vol.1. doi:10.1109/ICNN.1994.374138

[30] Guillermo Owen. 1977. A Value for Games with a Priori Unions. In *Essays in Mathematical Economics and Game Theory*, Robert Henn and Oskar Moeschlin (Eds.). Springer, Berlin, Heidelberg, 76–88.

[31] Fingrid Oyj. 2024. Frequency - historical data. https://data.fingrid.fi/en/datasets/339 Accessed: 2024-10-26.

[32] Laurent Pagnier and Philippe Jacquod. 2019. Inertia location and slow network modes determine disturbance propagation in large-scale power grids. *PLOS ONE* 14, 3 (March 2019), 1–17. doi:10.1371/journal.pone.0213550 Publisher: Public Library of Science.

[33] Sergei Popov, Stanislav Morozov, and Artem Babenko. 2019. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. http://arxiv.org/abs/1909.06312 arXiv:1909.06312 [cs, stat].

[34] Sebastian Pütz, Hadeer El Ashhab, Matthias Hertel, Ralf Mikut, Markus Götz, Veit Hagenmeyer, and Benjamin Schäfer. 2024. Feasibility of Forecasting Highly Resolved Power Grid Frequency Utilizing Temporal Fusion Transformers. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (e-Energy '24)*. Association for Computing Machinery, New York, NY, USA, 447–453. doi:10.1145/3632775.3661963

[35] Sebastian Pütz, Johannes Kruse, Dirk Witthaut, Veit Hagenmeyer, and Benjamin Schäfer. 2023. Regulatory Changes in German and Austrian Power Systems Explored with Explainable Artificial Intelligence. In *Companion Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 26–31. doi:10.1145/3599733.3600247

[36] R Radhika and K C Sindhu Thampatty. 2023. Impacts of Renewable Energy Integration on Power System Stability. In *2023 IEEE Technology & Engineering Management Conference - Asia Pacific (TEMSCON-ASPAC)*. 1–6. doi:10.1109/TEMSCON-ASPAC59527.2023.10531586

[37] Meysam Saeedian, Bahman Eskandari, Shamsodin Taheri, Marko Hinkkanen, and Edris Pouresmaeil. 2021. A Control Technique Based on Distributed Virtual Inertia for High Penetration of Renewable Energies Under Weak Grid Conditions. *IEEE Systems Journal* 15, 2 (June 2021), 1825–1834. doi:10.1109/JSYST.2020.2997392 Conference Name: IEEE Systems Journal.

[38] Md. Abdus Salam. 2020. Power System Stability Analysis. In *Fundamentals of Electrical Power Systems Analysis*, Md. Abdus Salam (Ed.). Springer, Singapore, 411–460. doi:10.1007/978-981-15-3212-2_9

[39] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (May 2022), 84–90. doi:10.1016/j.inffus.2021.11.011

[40] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. 2021. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. doi:10.48550/arXiv.2106.01342 arXiv:2106.01342 [cs].

[41] Melvyn Tyloo and Philippe Jacquod. 2021. Primary Control Effort Under Fluctuating Power Generation in Realistic High-Voltage Power Networks. *IEEE Control Systems Letters* 5, 3 (July 2021), 929–934. doi:10.1109/LCSYS.2020.3006966 Conference Name: IEEE Control Systems Letters.

[42] Andreas Ulbig, Theodor S. Borsche, and Göran Andersson. 2014. Impact of Low Rotational Inertia on Power System Stability and Operation. *IFAC Proceedings Volumes* 47, 3 (Jan. 2014), 7290–7297. doi:10.3182/20140824-6-ZA-1003.02615

[43] Julian Vossen, Baptiste Feron, and Antonello Monti. 2018. Probabilistic Forecasting of Household Electrical Load Using Artificial Neural Networks. In *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. 1–6. doi:10.1109/PMAPS.2018.8440559

[44] In-Kwon Yeo and Richard A. Johnson. 2000. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* 87, 4 (2000), 954–959. https://www.jstor.org/stable/2673623 Publisher: [Oxford University Press, Biometrika Trust].

A Data Description

Ex-post	Ramps [MW/h]	Load ramp, Total generation ramp, Biomass ramp, Coal gas ramp^{CE}, Fossil peat ramp^{Nordic}, Gas ramp, Geothermal ramp^{CE}, Hard coal ramp, Lignite ramp^{CE}, Nuclear ramp, Offshore wind ramp, Onshore wind ramp, Oil ramp, Other ramp, Other renewables ramp, Pumped hydro ramp^{CE}, Reservoir hydro ramp, Run-off-river hydro ramp, Solar ramp, Waste ramp
	Generation and load [MW]	Load, Total generation, Synchronous generation, Biomass generation, Coal gas generation^{CE}, Fossil peat generation^{Nordic}, Gas generation, Geothermal generation^{CE}, Hard coal generation, Lignite generation^{CE}, Nuclear generation, Oil generation, Other generation, Other renewable generation, Pumped hydro generation^{CE}, Reservoir hydro generation, Run-off-river hydro generation, Solar generation, Waste generation, Wind offshore generation, Wind onshore generation
	Forecast errors of generation and load [MW]	Forecast error load, Forecast error total generation, Forecast error solar, Forecast error offshore wind, Forecast error onshore wind
	Forecast errors of ramps [MW/h]	Forecast error load ramp, Forecast error generation ramp, Forecast error solar ramp, Forecast error offshore wind ramp, Forecast error onshore wind ramp
Day-ahead	Generation and load [MW]	Load day-ahead, Scheduled generation, Solar day-ahead, Offshore wind day-ahead, Onshore wind day-ahead
	Ramps [MW/h]	Load ramp day-ahead, Generation ramp day-ahead, Solar ramp day-ahead, Offshore wind ramp day-ahead, Onshore wind ramp day-ahead
	Other	Price ramp day-ahead [Currency/MWh/h], Prices day-ahead [Currency/MWh], Hour, Weekday, Month

Table 2: All external features in the data set. Features marked with ^{CE} are specific to the Continental Europe (CE) region, and those marked with ^{Nordic} are specific to the Nordic area. The units correspond to those used in our publicly available data set.

Area	Number of features	Number of data points
Continental Europe	64	35040
Nordic	58	35064

Table 3: Properties of the used data sets for both synchronous areas.

B Stability Indicators and Correlation Analysis

B.1 Calculation of Grid Frequency Stability Indicators

For the calculation of the stability indicators we use an approach from previous work [19], which we summarize in this section. The four frequency stability indicators are calculated from frequency time series $f(t)$ with 1-second resolution, centered around the nominal frequency: $f(t) = \tilde{f}(t) - 50$ Hz. For the i-th hour starting at time t_i, let $I_i = \{t_i, t_i + \tau, ..., t_i + \gamma\}$ be the set of hourly time steps with $\gamma = 3600$ seconds. The RoCoF is calculated by finding the steepest slope within a window $W_i = [t_i - T, t_i + T]$ around the beginning of each hour:

$$\mathbf{RoCoF}(t_i) = \frac{df}{dt}\left|\left(\arg\max_{t \in W_i} |\frac{df}{dt}|\right)\right.$$

The derivative $\frac{df}{dt}(t)$ is estimated using a low-pass filter on the frequency increments $\Delta f(t) = f(t) - f(t-\tau)$, implemented as a rectangular rolling window of length L, where $L = T = 60s$ for Continental Europe and $L = T = 30s$ for the Nordic region. The other three indicators are calculated as follows:

$$\mathbf{Nadir}(t_i) = f\big(\arg\max_{t \in I_i} |f(t)|\big), \quad \mathbf{MSD}(t_i) = \frac{1}{\gamma}\sum_{t \in I_i} f^2(t), \quad \mathbf{Integral}(t_i) = \tau \sum_{t \in I_i} f(t).$$

B.2 Correlation Analysis

The correlation heatmap in Figure 10 illustrates the feature relationships in our external features for the Continental Europe region. We only visualize correlations with absolute values ≥ 0.7. This analysis reveals strong correlations between related generation types and their corresponding ramp rates, as well as dependencies between actual values and day-ahead forecasts.

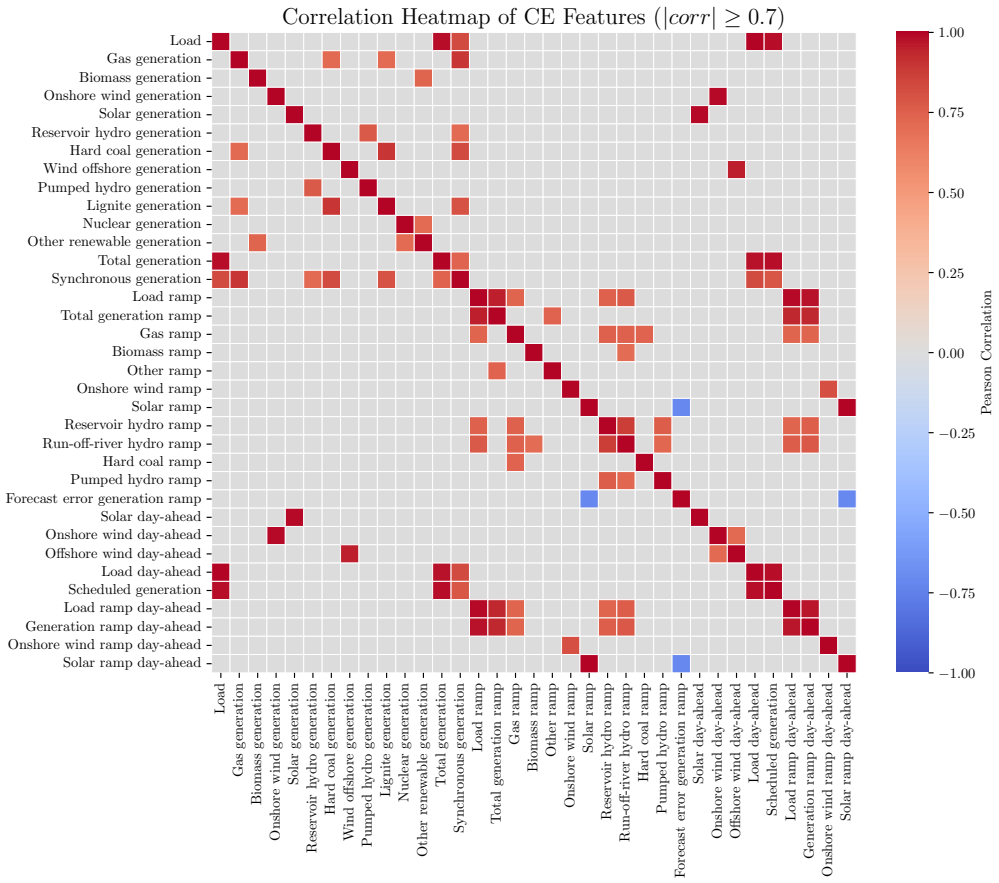


Figure 10: Correlation matrix showing absolute Pearson correlations among features in the Continental Europe data, including only features with absoulte correlation coefficients ≥ 0.7.

C TabNetProba: Modified Architecture

In Figure 11, the key components of TabNet's modified architecture are illustrated. Our key modification involves changing the output layer and loss function to predict probabilistic distributions rather than point estimates. The final layer now outputs both the mean μ and variance σ^2 parameters of a Gaussian distribution. Additionally, we replace the standard MSE loss with negative log-likelihood (NLL) to optimize these distributional parameters. The core architectural elements - feature transformers, attentive transformers, and the prior scaling mechanism - remain unchanged from the original TabNet design.

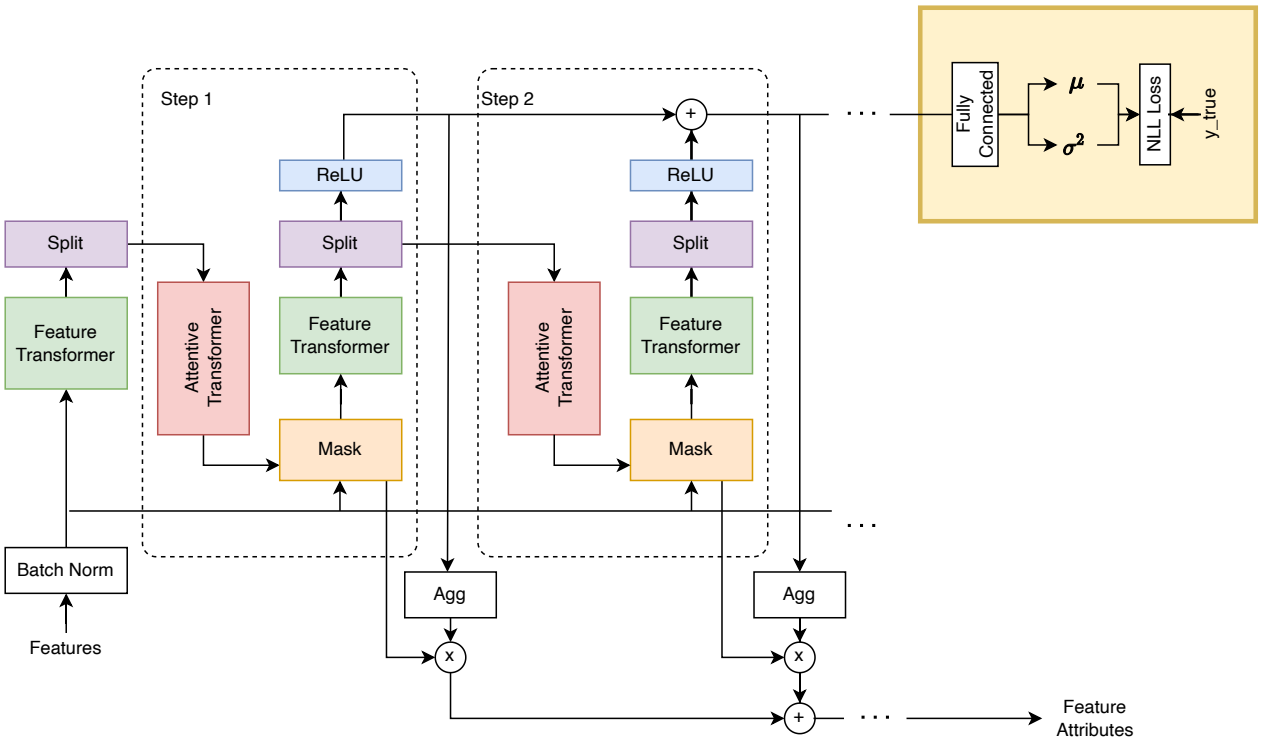


Figure 11: Overview of our probabilistic extension to TabNet (adapted from [3]). While maintaining TabNet's core architecture for feature processing, we modify the output layer (highlighted in yellow) to predict parameters μ and σ^2 of a Gaussian distribution instead of point estimates. The model is trained using negative log-likelihood (NLL) loss computed between the predicted distribution parameters and true values *y_true*.

D Hyperparameter Search Spaces

In this section, we provide a detailed overview of the hyperparameter ranges used for optimizing the machine learning models mentioned in this paper. Table 4 outlines the specific hyperparameters and their respective ranges or options considered during the tuning process for each model.

E Model Evaluation

E.1 Evaluation Metrics

In our work, we employed two metrics at different stages of model development: Mean Squared Error (MSE) during training and the R^2 score (coefficient of determination) for evaluation.

Mean Squared Error (MSE). The Mean Squared Error (MSE), used during model training, calculates the average squared differences between predicted and actual values [7]:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{1}$$

where N is the number of samples, y_i represents the true value, and $\hat{y}_i$ is the predicted value for the i-th observation. As a negatively oriented metric, lower MSE values indicate better model performance.

R^2 Score. For model evaluation, we utilized the R^2 score, which quantifies the proportion of variance explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2}$$

Table 4: Hyperparameter Ranges and Search Types for Model Optimization

Model	Tuning Technique	Parameter	Range/Options
XGBoost	Grid Search	max_depth	{3, 5, 7, 9, 11}
		learning_rate	{0.01, 0.05, 0.1}
		subsample	{0.4, 0.7, 1.0}
		min_child_weight	{1, 5, 10, 30}
		reg_lambda	{0.1, 1, 10}
NGBoost	Random Search	n_estimators	[350, 1000]
		learning_rate	[0.01, 0.8] (uniform)
		minibatch_frac	[0.1, 0.9]
		max_depth	[1, 20]
TabNet/TabNetProba	Random Search	n_d, n_a	[8, 64]
		n_steps	[3, 10]
		n_independent	[3, 10]
		n_shared	[3, 10]
		gamma	[3, 10] (uniform)

where $\bar{y}$ represents the mean of the true values. The R^2 score typically ranges from 0 to 1, with scores closer to 1 indicating better model performance.

For evaluating our probabilistic models, we employed proper scoring rules that assess both the sharpness (precision of predictions) and calibration (reliability of uncertainty estimates) of model outputs. During training, we utilized both the Continuous Ranked Probability Score (CRPS) and Negative Log-Likelihood (NLL), while the CRPS served as our primary metric for final model evaluation.

CRPS Score. The CRPS evaluates the quality of probabilistic predictions by measuring the difference between predicted and observed cumulative distribution functions [11, 23]:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}(z \geq y))^2 \, dz \tag{3}$$

where F is the predicted cumulative distribution function and y is the observed value. The CRPS effectively quantifies how well the predicted probability distribution aligns with actual observations, making it particularly suitable for comparing probabilistic predictions across different models.

NLL. During training, we also used the NLL, which measures how well the model's predicted distributions fit the likelihood of the observed data [29]:

$$\text{NLL} = -\sum_{i=1}^{n} \log p(y_i | \mathbf{x}_i, \theta) \tag{4}$$

For our Gaussian distribution assumptions, this specifically takes the form:

$$\text{NLL} = \frac{1}{2} \sum_{i=1}^{n} \left(\log(2\pi\sigma^2) + \frac{(y_i - \mu)^2}{\sigma^2} \right) \tag{5}$$

where μ is the predicted mean, σ^2 is the predicted variance, y_i represents observed values, n is the number of observations. While both metrics were used during model development, we prioritized CRPS for final evaluation due to its intuitive interpretation and direct assessment of both sharpness and calibration in a single metric.

E.2 Model Performance on Point Estimation

Figure 12 gives an overview of the performance of all trained models for modeling the frequency stability indicators in the Continental Europe and Nordic regions measured using the R^2 score. The probabilistic models were evaluated as if they were deterministic by using the mean estimates to compute the R^2 score, in order to be able to compare their performance to the deterministic ones.

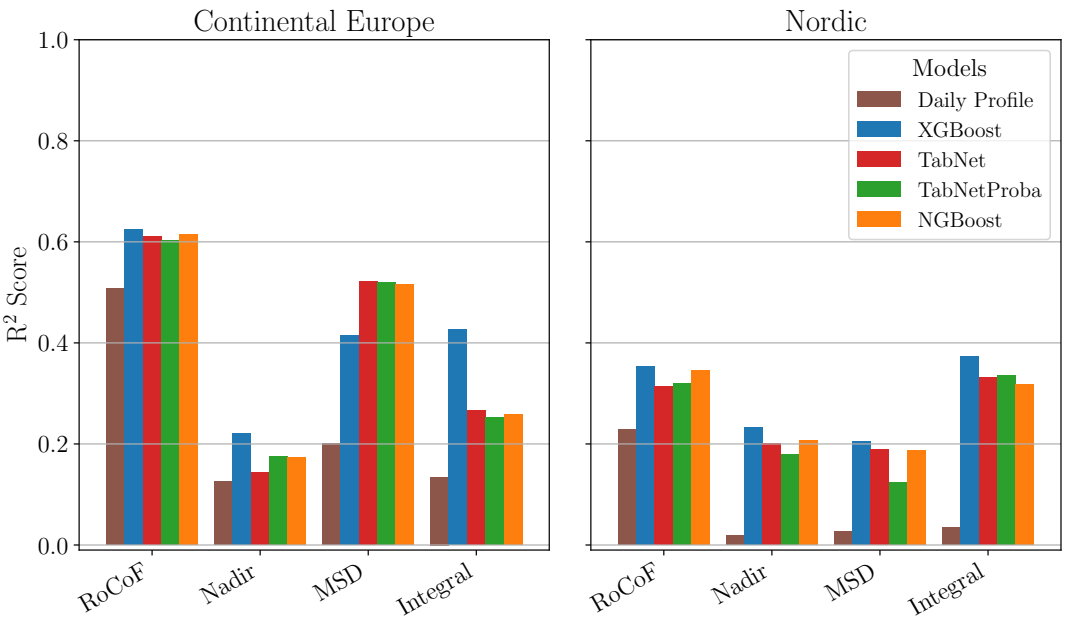


Figure 12: Performance of all trained models.

E.3 Calibration Curves

Calibration analysis for Continental Europe reveals distinct patterns across models and stability indicators (Fig. 13). NGBoost demonstrates strong calibration for RoCoF and Integral, with calibration curves closely tracking the ideal diagonal. While MSD shows good overall calibration, minor deviations appear at lower confidence levels. The Nadir target proves challenging, with NGBoost exhibiting underconfidence at lower confidence levels before improving above 80% confidence. TabNetProba achieves comparable performance for RoCoF and superior calibration for MSD relative to NGBoost. However, it consistently underestimates uncertainty for Integral across all confidence levels, with increasing severity at higher levels. While TabNetProba shows marginally better calibration than NGBoost for Nadir, both models struggle with this target, particularly at lower confidence levels. Both models achieve reliable calibration for RoCoF and MSD, with NGBoost showing superior performance for Integral predictions. The persistent challenge of calibrating Nadir predictions across both models suggests underlying complexities in capturing extreme frequency deviations, though TabNetProba demonstrates modest improvements in this regard.

Figure 14 shows the calibration curves for the Nordic region, which display very similar trends to those observed in CE. NGBoost (left plot) performs well for RoCoF and Integral, with curves closely following the diagonal, while showing slight underconfidence for MSD at lower confidence levels and more pronounced underconfidence for Nadir, especially at lower intervals. TabNetProba (right plot) also performs similar to CE, with good calibration for RoCoF and MSD. However, it is underconfident for Integral at higher confidence levels and struggles with Nadir for lower levels, though does slightly better than NGBoost. In summary, both models perform similarly in the Nordic region as they do in CE, with strong calibration for RoCoF and MSD, while Nadir remains the most challenging target for both models.

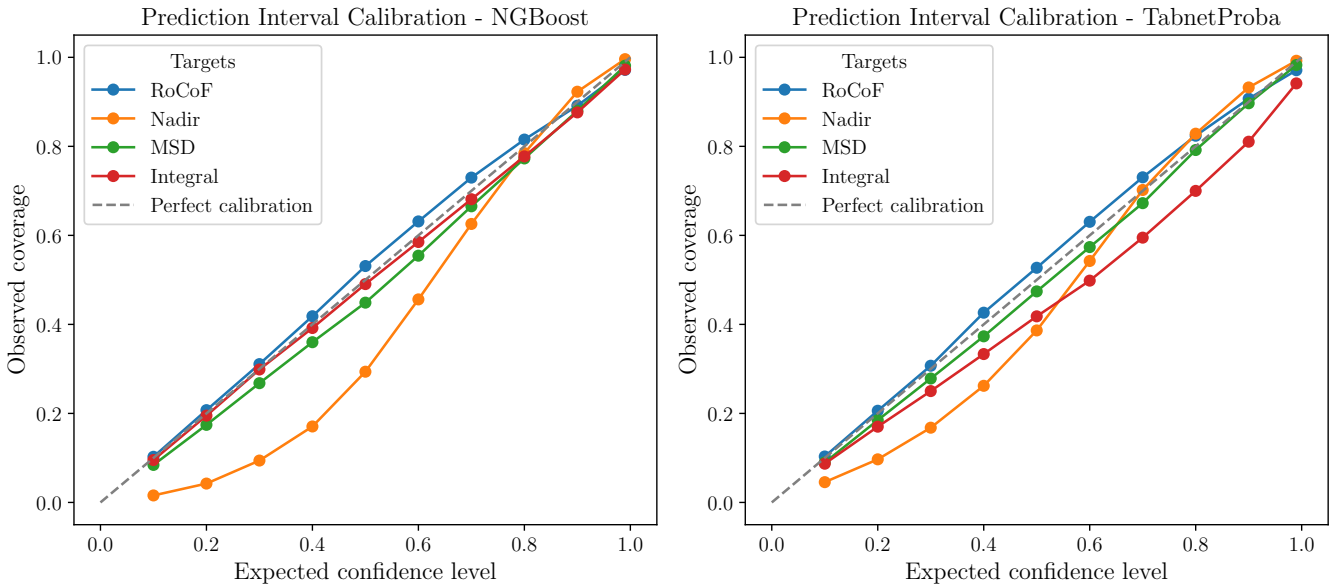


Figure 13: Calibration Plot of Prediction Models: Continental Europe.

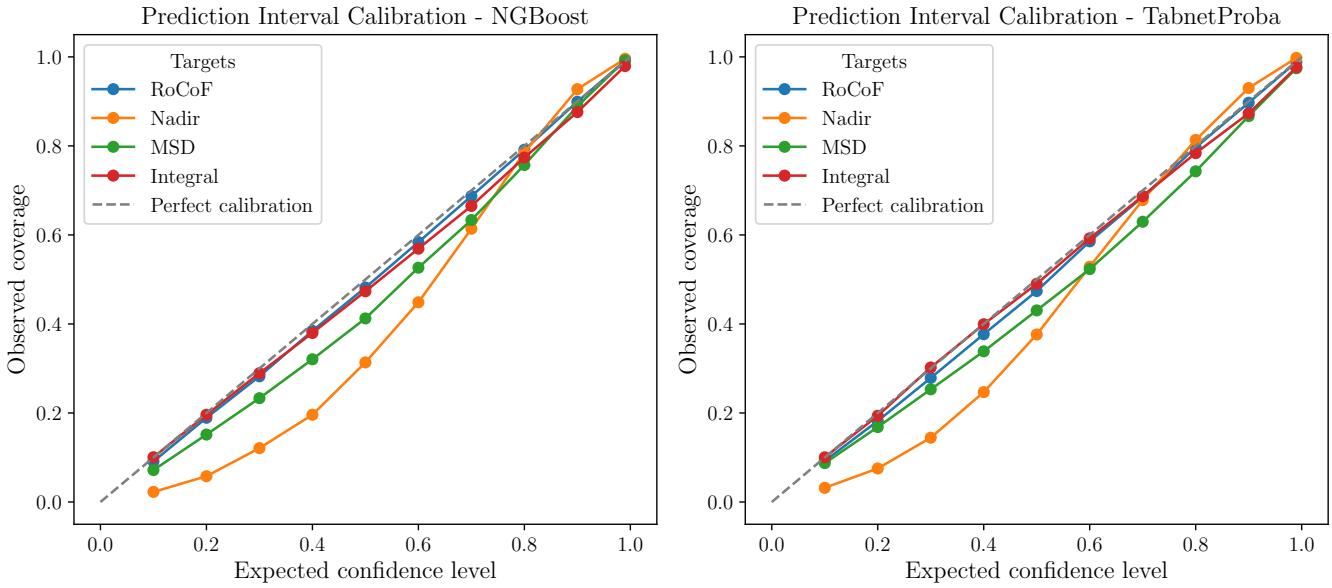


Figure 14: Calibration Plot of Prediction Models: Nordic Region.

F Clustering Dendrograms of Input Features

In this section, we present clustering dendrograms that illustrate the correlation-based hierarchical clustering of input features. Figures 15 and 16 show these dendrograms for Continental Europe (CE) and the Nordic region, respectively.

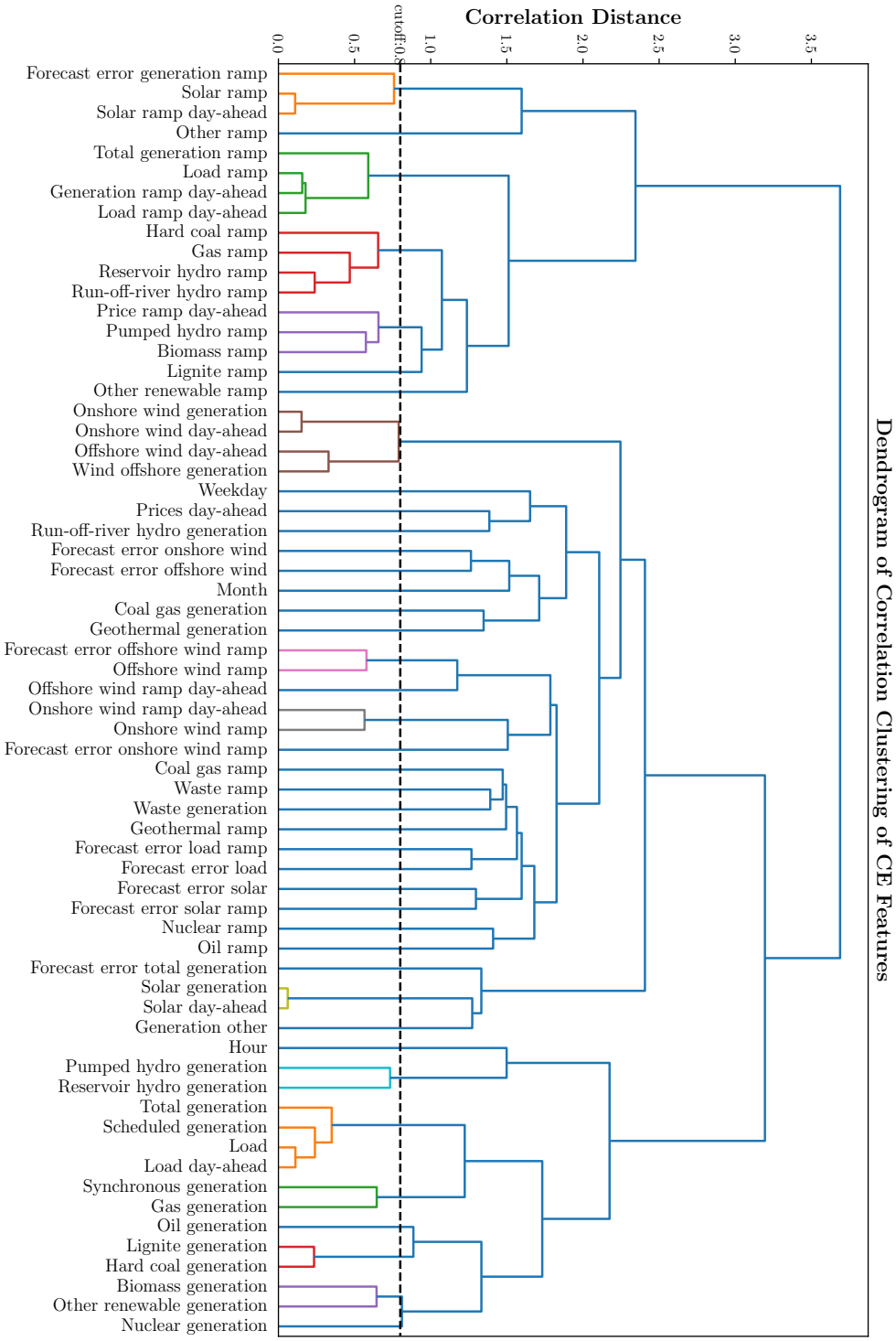


Figure 15: Hierarchically clustered features based on correlation for Continental Europe. The dashed line represents a correlation cutoff threshold of 0.8. Feature groups under this line are grouped together for the interpretation of the results with the Partition SHAP Explainer.

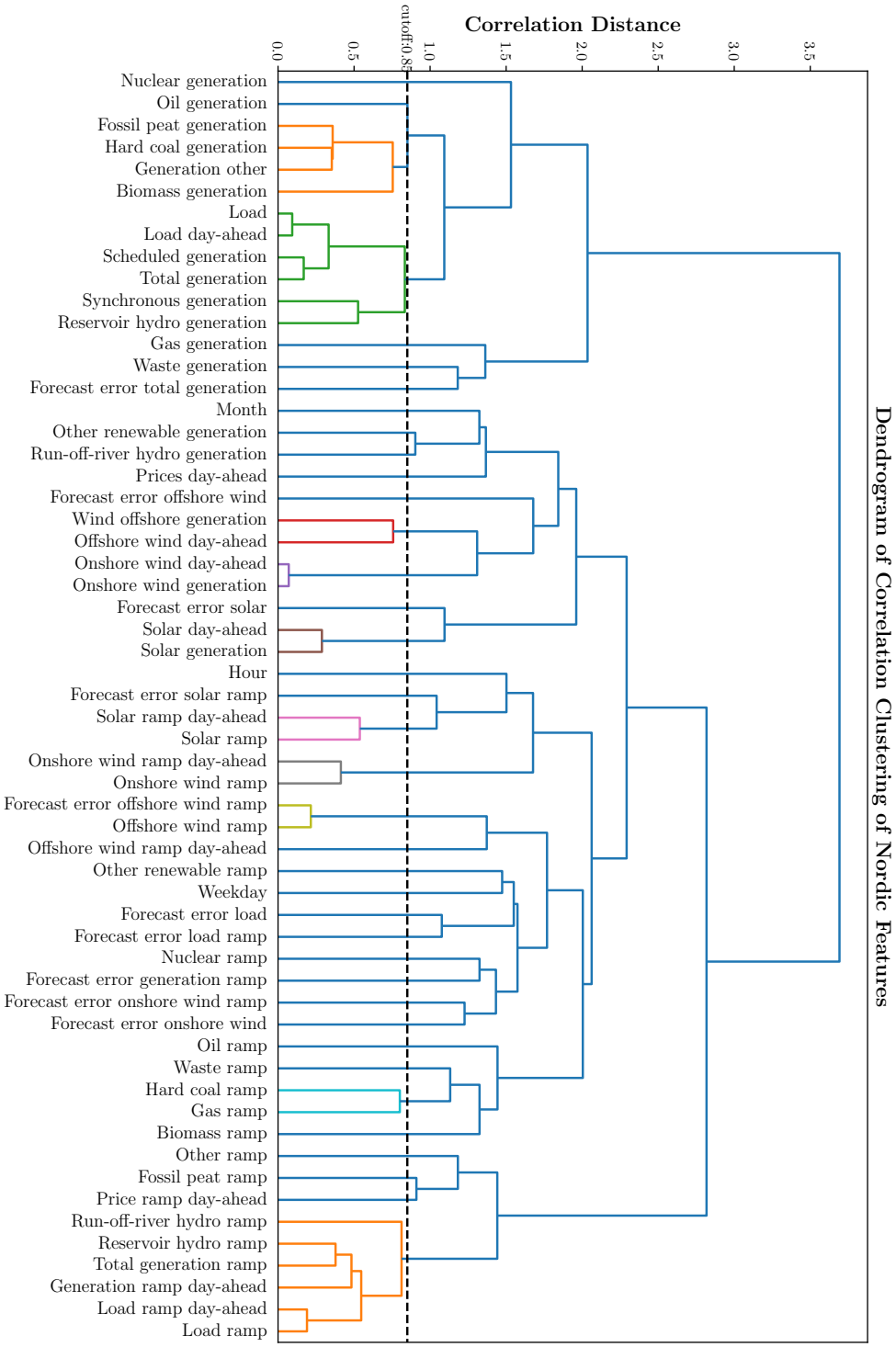


Figure 16: Hierarchically clustered features based on correlation for the Nordic region. The dashed line represents a correlation cutoff threshold of 0.85. Feature groups under this line are grouped together for the interpretation of the results with the Partition SHAP Explainer.