Short-Term Nodal Voltage Forecasting for Power Distribution Grids: An Ensemble Learning Approach

Yi Wang[a], Leandro Von Krannichfeldt[b], Thierry Zufferey[b], Jean-François Toubeau[c]

[a]*Department of Electrical and Electronic Engineering, The University of Hong Kong, 999077, Hong Kong SAR, China*
[b]*Power Systems Laboratory, ETH Zurich, 8092, Zurich, Switzerland*
[c]*Power Systems and Markets Research (PSMR) Group, University of Mons, 7000, Mons, Belgium*

Abstract

The integration of distributed energy resources (DER) complicates the operation of the power distribution grids, and the nodal voltage may violate frequently. Making accurate predictions of the nodal voltage is fundamental for voltage regulation of the distribution grid. Even though energy forecasting has been widely studied, voltage is still a rarely touched area. This paper enriches the research by proposing an ensemble approach for both deterministic and probabilistic short-term nodal voltage forecasting. Specifically, a new joint model- and data-driven feature selection is first performed to select the most relevant features for distribution grid voltage forecasting. Then, different individual forecasting models are trained using the selected features. On this basis, simple weighted averaging and quantile regression averaging approaches are applied to combine the individual models for deterministic and probabilistic forecasting, respectively. Finally, case studies are conducted on a real-world distribution grid to verify the effectiveness and superiority of the proposed method.

Keywords: nodal voltage forecasting, ensemble learning, quantile regression averaging, distribution grids, situation awareness

1. Introduction

Energy forecasting is an essential part of the decision-making in power systems. Most energy forecasting works focus on electrical load, price, and renewable energy [1]. The integration of distributed energy resources (DER), such as photovoltaics (PV) and electric vehicles (EV), complicates the operation states of the power distribution grids, and the nodal voltage may violate frequently. Thus, there is an incentive to have a real-time situation awareness of the whole distribution grid, i.e., every physical state in the grid, including power injections, power flows, and voltage magnitudes, to observe crucial security constraints [2]. Making accurate predictions of the nodal voltage is a critical part of the distribution grid situation awareness and is fundamental for voltage regulation of the

distribution grid [3]. Even though a massive number of works have been done on energy forecasting, voltage forecasting is still a rarely touched topic.

If all electrical parameters and power injections of the distribution grid are known, the nodal voltages can be indirectly computed. Thus, the nodal voltage forecasting can be implemented by first forecasting power injections (e.g., load and renewable energy) and then calculating the power flows. Bracale *et al.* applied this strategy for voltage forecasting in [4]. The Bayesian-based approach was first proposed to predict the power production of wind, PV, and electrical demands at different nodes. In a second step, probabilistic load flow was performed to calculate the nodal voltages of the distribution grid. Actually, there are strong dependencies among the nodal power injections. Modeling these complex dependencies in probabilistic load flow was not considered in that work. Dobbe *et al.* proposed an approach to improve the real-time nodal voltage forecasting performance with a limited set of real-time measurements, where a linear mapping relationship from load predictions to voltage predictions was build based on recently developed linear approximations for unbalanced three-phase power flow [5]. Hayes *et al.* presented three typical services in distribution network energy management systems based on smart meter data, where forecasting voltage profiles in the low voltage network were one of them [6]. The voltage profile prediction was implemented based on power injection forecasting and energy management as well as network optimization. It should be noted that indirect voltage forecasting and state estimation are distinct. The former provides the values for the future, while the latter provides current values.

Indirect forecasting strategies are confronted with two salient challenges. First, all nodal injections, including reactive power forecasts/estimations, should be used for power flow calculation which is data-intensive, i.e., a lot of data has to be gathered for voltage estimation. Second, if the uncertainties of nodal voltages should be considered, modeling the complex dependencies among the nodal power injections is necessary but nontrivial. Compared with indirect forecasting approaches, direct forecasting approaches model the voltage time-series data itself. Dejamkhooy *et al.* applied Grey system theory-based models to predict the nonstationary envelope voltage magnitude signal. In addition, they used a so-called Fourier correction Grey Model (FGM) to model the residuals in order to improve further the forecasting performance in [7]. However, the voltage magnitude data are measured at a frequency of several milliseconds that is more suitable for transient analysis. Hassanzadeh *et al.* examined the spatial and temporal correlation of the nodal voltage angles of the power systems to which a large number of renewable resources and microgrids are connected in [8]. The spatial and temporal correlations were then modeled by vector autoregressive (VAR) processes and used for nodal voltage angle forecasting. Several regression models were applied for voltage forecasting in [9]. On this basis, these regression methods are combined to produce higher accurate forecasts. Zufferey *et al.* applied and compared two quantile regression models (quantile neural network and quantile K-nearest neighbor) for nodal voltage magnitude predictions, which were then applied for voltage regulation [10]. Note that the nodal voltages are influenced by all the nodal injections and electrical param-

eters of the power systems. The input features of the direct forecasting model will be high-dimensional if lag values of all injections are considered. Feature selection should be performed before training the forecasting model to reduce model complexity and avoid overfitting risks.

Ensemble learning is an effective approach to improve forecasting performance, which has been widely applied in energy forecasting. An ensemble learning model for load forecasting that was able to select appropriate input features adaptively was studied in [11], where the parameters of base models were optimized using evolutionary algorithms. An enhanced ensemble structure using wavelet neural networks was proposed in [12] for short-term load forecasting. The impact of renewable energy on price forecasts was first analyzed in [13], and then a bootstrap aggregated-stack generalized architecture was proposed for very short-term electricity price ensemble forecasting. Another ensemble learning model was applied in [14] for probabilistic intraday electricity price forecasting using simulating trajectories. A decomposition-ensemble learning approach was studied in [15] to combine Complete Ensemble Empirical ModeDecomposition (CEEMD) and Stacking-ensemble learning (STACK) for wind power forecasting. In [16], an optimized stochastic ensemble method was proposed for multi-step wind speed forecasting. Further, a hierarchical probabilistic electric vehicle load forecasting approach was provided in [17] by using a penalized linear quantile regression model. The efficient combination of multiple probabilistic forecasts was studied in [18]. Quantile regression averaging (QRA) was proposed in [19] to combine point price forecasts into probabilistic price forecasts. It was then applied for probabilistic load forecasting [20].

To this end, this paper proposes ensemble approaches for short-term nodal voltage magnitude forecasting (which is simplified as voltage forecasting in the following). The proposed approach has agile responsiveness to the time-varying characteristic of the distribution grid from the ensemble learning perspective. It first trains several nodal voltage forecasting models, where each model performs joint model- and data-driven feature selection to improve the forecasting performance. On this basis, two averaging models, i.e., simple weighted averaging and quantile regression averaging, are proposed to combine these individual forecasts in real-time for both deterministic and probabilistic nodal voltage forecasting. In this way, the proposed method can take full advantage of individual forecasting models to further enhance the forecasting performance and produce probabilistic forecasts. It should be noted that the main focus of this paper includes two aspects, i.e., feature selection and ensemble learning. In the energy forecasting area, different advanced machine learning models have been applied. For example, the long short-term memory (LSTM) neural network is a powerful model that has been applied for PV forecasting [21], load forecasting [22], wind power forecasting [23], and price forecasting [24]. These advanced machine learning models will be the base models in our proposed framework.

This paper makes the following three main contributions:

1. This paper enriches the research on voltage forecasting by developing individual nodal voltage forecasting models with a joint model-driven and

data-driven feature selection process.

2. Since ensemble learning has not been well studied for voltage forecasting, this paper studies this problem by proposing simple weighted averaging and quantile regression averaging ensemble approaches to combine the individual models for deterministic and probabilistic voltage forecasting, respectively.

3. This paper conducts comprehensive case studies and comparisons with a significance test to statistically demonstrate the superiority of the proposed method.

The remainder of this paper is organized as follows: Section 2 briefly describes the dataset to be studied in this paper. Section 3 introduces the framework and technical details of the proposed methods for deterministic and probabilistic nodal voltage forecasting. Section 4 conducts case studies and makes comparisons to verify the superiority of our proposed method. Finally, section 5 draws conclusions and gives an outlook on future work.

2. Data Description

The dataset used for nodal voltage forecasting in this paper is introduced here to easily show how our proposed method works. Note that the proposed method is not limited to a certain dataset or distribution system.

The voltage forecasting is conducted on a distribution grid from a Swiss distribution system operator (DSO). The dataset contains operational state information about the distribution grid (including voltage, power, and reactive power injections and weather information) over one year with a quarter-hour resolution. The distribution grid consists of 196 buses with 197 lines. These 196 buses can be divided into two transformer buses, seven cabinet buses, 88 load buses, and 99 no-load buses. Bus #1 is disregarded for voltage forecasting for the reason of being a root bus. Fig. 1 illustrates the voltage evolution of Bus #3 over one year from 20 October 2016 to 19 October 2017. We can discern a fairly stable voltage pattern over the whole dataset with some spikes around 28 August 2017. The one-year dataset is partitioned into three parts in this paper for model training, ensemble learning, and testing with ratios 50%, 25%, and 25%, which is also shown in Fig. 1.

3. Proposed Methodology

This section first provides a framework of the proposed method and then goes over the details, including individual model training and ensemble learning for both deterministic and probabilistic voltage forecasting. The forecasting horizon for all models is chosen as an hour ahead.

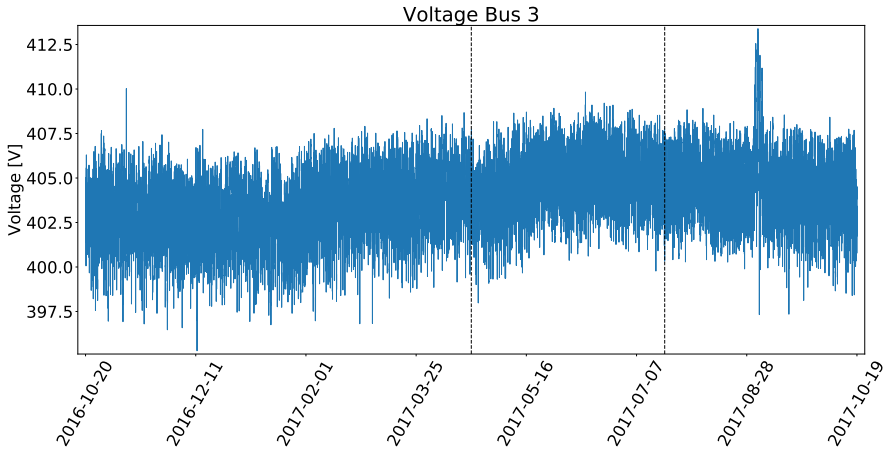

Figure 1: Voltage of Bus #3 over one year and data partitions into model training, ensemble learning, and testing sets.

3.1. Framework

The framework of the proposed methodology can be explained on the partitioned datasets ($\mathbf{D}_{\text{Train}}$, $\mathbf{D}_{\text{Ensemble}}$, $\mathbf{D}_{\text{Test}}$) as shown in Fig. 2. It contains two main stages. The first stage (①&②) consists of training individual models on the dataset $\mathbf{D}_{\text{Train}}$, which will be tested on dataset $\mathbf{D}_{\text{Ensemble}} \cup \mathbf{D}_{\text{Test}}$. In the second stage (③&④), the ensemble combines the base models by training on dataset $\mathbf{D}_{\text{Ensemble}}$, which will be tested on dataset $\mathbf{D}_{\text{Test}}$.

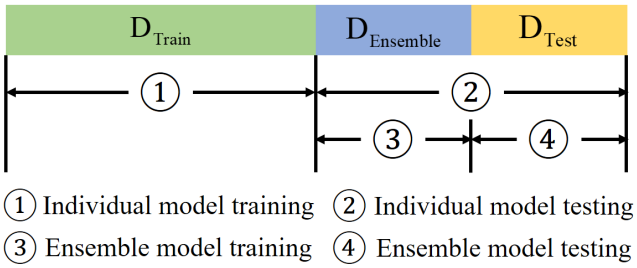

Figure 2: Framework of the proposed method.

3.2. Individual Forecasting Models

In the first stage, individual forecasting models are trained. This means that input features must be optimally selected and that different hyperparameters for each individual model must be chosen.

3.2.1. Feature Selection

The input features of the individual nodal voltage forecasting model are comprised of calendar variables and time-lagged physical distribution grid information written as:

$$\mathbf{X}_t = [D, H, M, \mathbf{x}_{t-h}, \mathbf{x}_{t-h-1}, \mathbf{x}_{t-2h}, \mathbf{x}_{t-2h-1}, \mathbf{x}_{t-3h}, \mathbf{x}_{t-24h}] \tag{1}$$

where $\mathbf{x}_i$ is a feature vector of physical distribution grid information, $D \in \{1, ..7\}$ denotes the weekday, $H \in \{1, ..., 24\}$ the hour and $M \in \{1, ..., 4\}$ an hourly cycle. The variable $h = 4$ denotes the number of time periods in one hour, and $k = t - (\cdot)$ is the lag. In this context, the voltage predictions are made for time t. The lagged features are selected based on the normalized autocorrelation function. As shown in Fig. 3, the most correlated values have very short lag but are also centered around a lag of one day. Furthermore, we observe a significant correlation of the lagged values since they display a higher autocorrelation value than the random noise marked as the blue shaded area. The nodal voltage is influenced by all the injections of the distribution grid. For the voltage magnitude (V), active power (P), and reactive power (Q), all the lagged values from (1) are considered. In addition, weather variables in terms of solar irradiance (S) and temperature (T) are considered, which were measured at a nearby weather station for every time step. For these two quantities, only the value for the most recent lag $t - h$ is used as a feature.

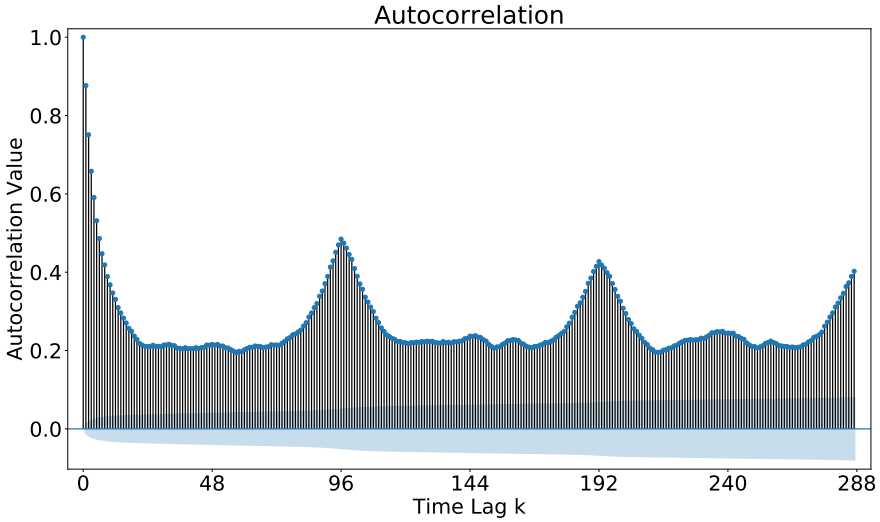

Figure 3: Autocorrelation function for voltage of bus 3 on the batch data $\mathbf{D}_{ens} \cup \mathbf{D}_{test}$ over three days

If the V, P, and Q of all 88 load buses are considered in the features set, the length of the input feature vector $\mathbf{X}_t$ would be $1589 = 3 + 2 + 6 \times 3 \times 88$. Since

there are a massive number of features for the voltage forecasting model, feature selection should be performed before training the regression model. Table 1 summarizes the grid features and includes a flag indicating whether lagged values are included or not.

Table 1: Information about features

Quantity	**Acronym**	**Lagged**
Voltage	V	yes
Active Power	P	yes
Reactive Power	Q	yes
Solar Irradiance	S	no
Temperature	T	no

The feature selection has two phases. The first phase is a model-driven selection of the nearest load buses by examining physical grid interdependences. The three distances which were adopted in [25] are used in the following.

The first distance is defined on the basis of the bus impedance matrix $\mathbf{Z}_{bus}$. With the help of the Thevenin impedance, we can find the Z-distance between two buses i and j as

$$\left|Z_{i,j}^{Thev}\right| = \left|Z_{i,i} + Z_{j,j} - Z_{i,j} - Z_{j,i}\right| \tag{2}$$

where $Z_{i,j}$ is the entry of impedance matrix at position (i, j). The advantage of this metric is its independence of system loadings.

The other two distance measures are related to the grid Jacobian matrix $\mathbf{J}$. The Jacobian matrix is obtained as the solution of the load flow problem and is connected to power flow sensitivities by

$$\begin{bmatrix} \Delta\mathbf{P} \\ \Delta\mathbf{Q} \end{bmatrix} = \mathbf{J} \begin{bmatrix} \Delta\boldsymbol{\theta} \\ \Delta\mathbf{V} \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J}_{P\theta} & \mathbf{J}_{PV} \\ \mathbf{J}_{Q\theta} & \mathbf{J}_{QV} \end{bmatrix} \tag{3}$$

where $\boldsymbol{\theta}$ represents the matrix of voltage angles. Eq. (3) describes the relation between a power injection at bus i and voltage magnitudes as well as angles at bus j. The Jacobian matrix itself can be divided into four sub-matrices, each corresponding to the sensitivity between the quantities in the subscript. Finally, the employed P-distance and Q-distance for voltage forecasting can be written as

$$\frac{\Delta V_{i,i}}{\Delta P_{i,j}} = (J_{PV}^{-1})_{i,i} + (J_{PV}^{-1})_{i,j} - (J_{PV}^{-1})_{j,i} - (J_{PV}^{-1})_{j,j} \tag{4}$$

$$\frac{\Delta V_{i,i}}{\Delta Q_{i,j}} = (J_{QV}^{-1})_{i,i} + (J_{QV}^{-1})_{i,j} - (J_{QV}^{-1})_{j,i} - (J_{QV}^{-1})_{j,j} \tag{5}$$

Eqs. (4) and (5) represent the voltage sensitivity at bus i to an active power injection and reactive power injection at bus i with subsequent withdrawal at bus j, respectively.

For all distance scenarios, including Z-distance, P-distance as well as Q-distance, the lagged V, P, and Q features of the 20 nearest load buses (nearest neighbors, NN) are included in the input feature vector $\mathbf{X}_t$. The nearest load buses selected according to the Z-, P-, and Q-distances are denoted as ZNN, PNN, and QNN methods, respectively. Fig. 4 shows the load buses and the 20 nearest load buses for each forecasting bus according to the distance metrics. In this way, the length of the input feature vector $\mathbf{X}_t$ is reduced to $365 = 3 + 2 + 6 \times 3 \times 20$. The nearest buses are different according to the type of distance measure.

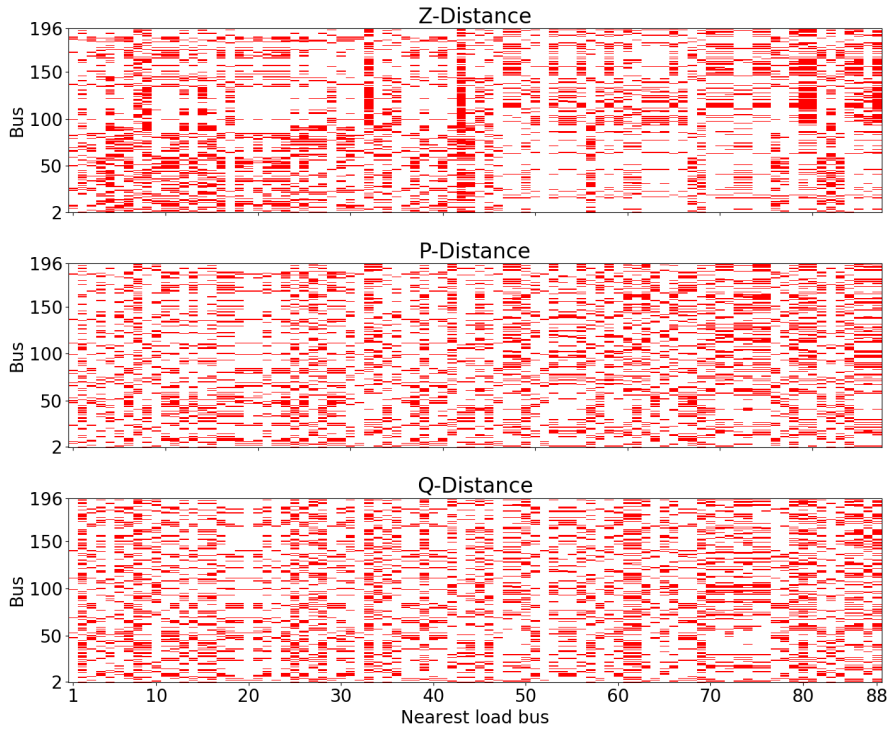

Figure 4: Visualization of the selected load buses.

The second phase of the input feature selection is a data-driven reduction of the number of features by the least absolute shrinkage and selection operator (LASSO). Hereby, each individual input is associated with a coefficient (of the global model) which is an image of its importance in the outcome [26]. Hence, inputs with small coefficients can be safely removed without loss of accuracy. The main idea is to establish a LASSO regression model for voltage forecasting with a certain l_1-regularization value λ. By varying the value of λ, the LASSO regression model with the best performance is chosen through cross-validation. On this basis, the features with a regression coefficient below a certain threshold

are dropped (f.e. a threshold of 10^{-5} is used in this paper). This data-driven feature selection procedure is summarized in Algorithm 1. In this way, joint model- and data-driven feature selection can be implemented to pick the most important features for individual model training.

Algorithm 1: Procedure for LASSO selection

input: Train data $\mathbf{D}_{tr}$

1. Fit LASSO model on $\mathbf{D}_{tr}$ for $\lambda \in [10^{-6}, 10^{-5.5}, ..., 10^{6}]$
2. Cross-validate for the best $\lambda = \hat{\lambda}$ with a 3-fold time series split strategy
3. Fit LASSO model on $\mathbf{D}_{tr}$ with $\hat{\lambda}$
4. Drop all features with a regression coefficient below the threshold 10^{-5}

Table 2 summarizes the basic scenario (in which none of the three distances is considered) and the three additional nearest neighbors scenarios, i.e., ZNN, PNN, and QNN. For the basic scenario, only V, P, and Q of the predicted bus are included. This sums up to $23 = 3 + 2 + 6 \times 3 \times 1$ features being used for nodal voltage forecasting with the basic scenario. By contrast, the mean numbers of features of all forecasting models across the 196 buses for ZNN, PNN, and QNN scenarios are 259, 275, and 280, respectively. The Q-distance displays the highest number of features with around 280 across all buses. In consequence, this might lead to considerably higher computation time.

Table 2: Overview of different feature selection scenarios

Scenario	Distance Measure	Neighbours	LASSO	Mean Feature Number
Basic	No	0	No	23
ZNN	Impedance Matrix	20	Yes	259
PNN	Jacobian Matrix	20	Yes	275
QNN	Jacobian Matrix	20	Yes	280

The selected features through LASSO for each nodal voltage forecasting are illustrated in Fig. 5. Notation-wise, C and W denote the calendar and weather features colored in orange and red, respectively. Other than that, the V, P, and Q features of the bus and of the nearest neighbors (with a "NN"-prefix) are indicated. Even though the selected features differ for the three distance scenarios, the nearest neighbor Q features are of high importance for all distance metrics. Besides, it can be seen that the parts of NN-V are sparser for Z- and P-distance than for Q-distance. That is to say, the bus voltage and nearest neighbor voltage features are less important in the case of Z- and P-distance. This interesting occurrence presents itself even though we are performing voltage forecasting. Nevertheless, the voltage information may be encoded through the AC power laws. Finally, we take note of the dropped bus P and Q features for part of the buses. It can be interpreted that the nearest neighbors also

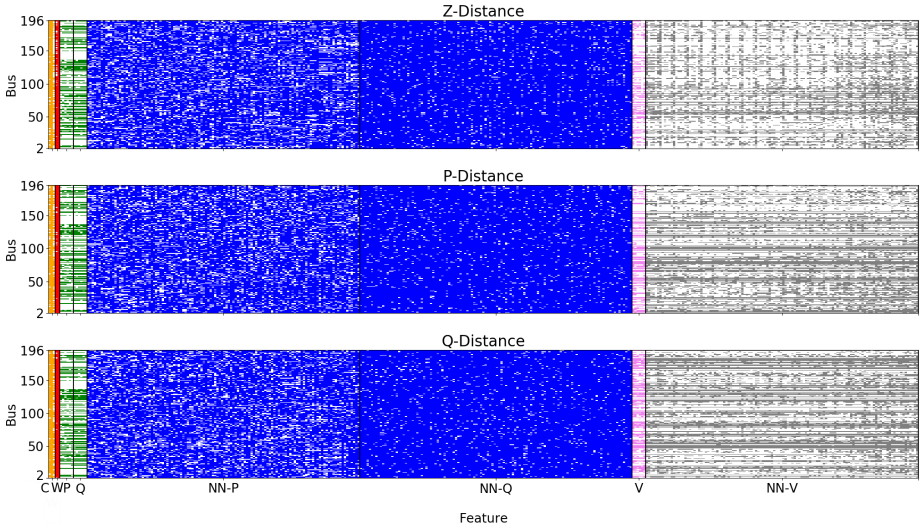

Figure 5: Feature mask for each distance scenario.

hold relevant information. An example would be an injection at a specific bus. Naturally, this would influence the whole bus system. By integrating these neighbors, we may provide useful patterns to other buses. The calendars and weather features hold approximately the same importance across all distance metrics.

The number of selected features is shown according to the bus number in Fig. 6. The maximum number of features is the y-axis upper limit, whereas the mean is drawn in red. On the one hand, the Q-distance displays the highest number of features, i.e., around 280 across all buses. In consequence, this leads to a considerably higher computation time. On the other hand, the Z-distance exhibits a mean of selected features around 260.

3.2.2. Regression Model Training

Based on the selected features, individual short-term voltage forecasting models with a time horizon of 1 hour can be trained for each of the 195 buses. A total of four regression models, including Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Regression Tree (GBRT), and Extremely Randomized Trees (ERT), serve as the base and are individually described hereunder.

Concerning the SVR model, we employ a variation called the ε-SVR that consists of solving a convex optimization problem. The goal is to find a function $f(\mathbf{x})$ that stays within a ε-margin from the labels y_i but also exhibits flatness. The first requirement is met through minimization of the so-called ε-insensitive loss. The second requirement is reflected in the l_2-regularization of the weights. Hence, for a data set $\mathbf{D} = \{(\mathbf{x}_i, y_i), ..., (\mathbf{x}_N, y_N)\}$ the SVR optimization problem

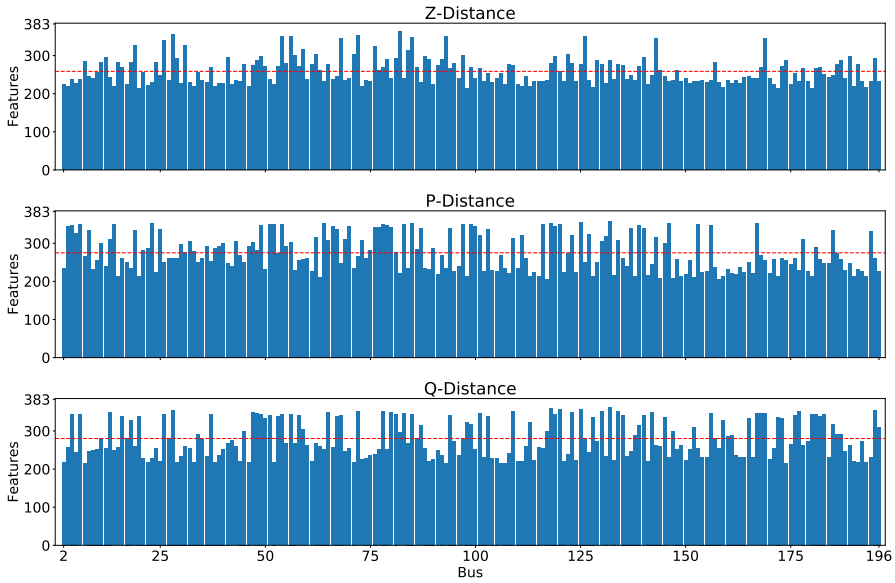

Figure 6: Number of selected features for each distance scenario.

can be formulated as

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & \mathbf{w}^T\phi(\mathbf{x}_i) - y_i \geq \varepsilon + \xi_i \\ & y_i - \mathbf{w}^T\phi(\mathbf{x}_i) \geq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^*, \varepsilon \geq 0 \quad C > 0 \quad i = 1, ..., N \\ & k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \end{aligned} \tag{6}$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is a chosen kernel function and $\phi(\mathbf{x}_i)$ the corresponding feature map. Kernel functions are sometimes also described as efficient inner products. With the help of the feature map, the features are mapped into an inner product space induced by the kernel function. In this product space, the computation happens implicitly and therefore in a more efficient manner. The slack variables ξ_i and ξ_i^* serve the purpose of ensuring feasibility. The constant C is a parameter that balances the importance between the violation of ε precision and flatness of the function.

The hyperparameters C and ε are chosen according to a noise estimation procedure [27]. The penalty parameter C is calculated as

$$C = \max(|\mu_y + 3\sigma_y|, |\mu_y - 3\sigma_y|) \tag{7}$$

where μ_y and σ_y are the mean and standard deviation of the training data target values. In order to achieve a minimal regularization in all cases, a lower

bound $C_{bound} = 0.1$ was set for all $C \leq C_{bound}$. The insensitivity parameter ε is determined with the help of a K-nearest neighbors regression:

$$\varepsilon = 3\hat{\sigma}_{noise}\sqrt{\frac{ln(N)}{N}} \qquad \hat{\sigma}^2_{noise} = \frac{N^{1/5}k}{N^{1/5}k - 1} \cdot \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

where, $\hat{\sigma}^2_{noise}$ is the estimated noise variance, k the number of nearest neighbors, N the number of samples, y_i the true value and $\hat{y}_i$ the estimated regression value. In this respect, we choose $k = 3$ as the number of nearest neighbors. Furthermore, the kernel function is calculated in our case as

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} \qquad \gamma = (d\sigma_x)^{-1} \tag{9}$$

where d is the number of features and σ_x the variance of the training data features [28].

Alternatively to SVR, tree-based methods do not require any feature scaling due to their invariance to monotonic transformations [29]. The number of trees for RF and ERT is taken as a standard value of 100 [28]. The number of boosting stages for GBRT is chosen according to observations in [30]. It is stated that in the investigated problems, the GBRT stabilizes around 1000 stages. These hyperparameters represent a good trade-off between accuracy and computation time. All the remaining hyperparameters are adopted from the standard parameters of the scikit-learn library [28]. A selection of notable parameters of the three tree models are given in Table 3 and 4, respectively.

Table 3: Parameters for RF and ERT

Parameter	**Value**
Number of Trees	100
Min. Samples Split	2
Min. Samples Leaf	1

Table 4: GBRT parameters

Parameter	**Value**
Boosting Stages	1000
Learning Rate	0.1
Min. Samples Split	2
Min. Samples Leaf	1

The performances, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), across all the buses on the test dataset are given in Fig. 7. For the box plot, each error metric is averaged over the four base models. Furthermore, the mean μ and

the median m are indicated as red triangle and orange line inside the boxes, respectively. Their numerical values are displayed at the right box end. The boxes enclose data between the 25%-th quantile and 75%-th quantile, whereas the whiskers extend to the minimum, respectively, maximum of the data points. Two important observations arise from those graphs. First, the inter-quantile range is close for all scenarios. Second, the inclusion of nearest neighbor features leads to performance benefits across all metrics, especially for the P- and Q-distance. Between the P- and Q-distance performances, there are only minor differences.

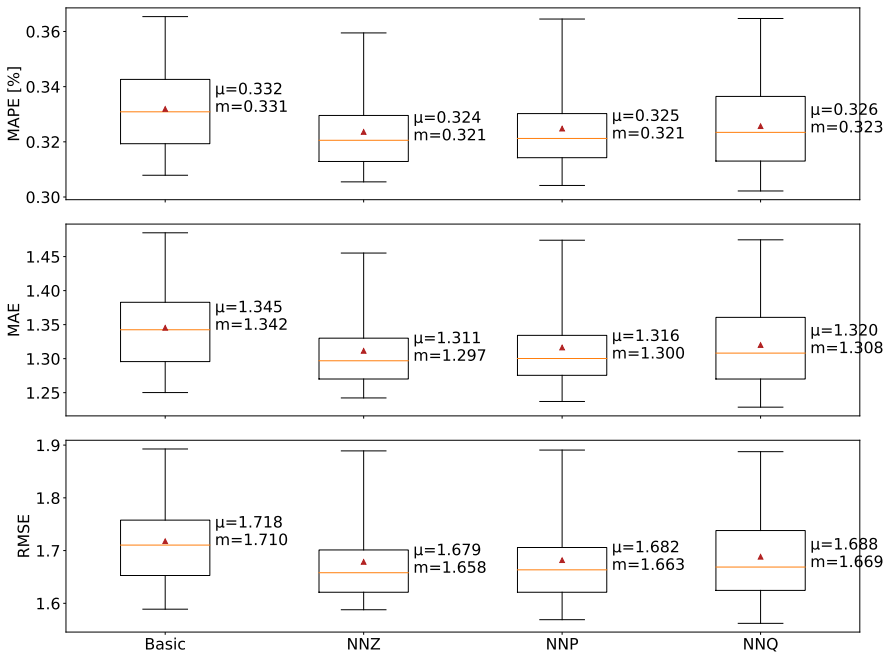

Figure 7: Error boxplots of individual models on $\mathbf{D}_{\text{Test}}$.

The base model with the lowest error metrics on the test dataset is indicated in Fig. 8. If the best model did not coincide for all metrics at one bus, the base model with a higher number of lowest metrics was chosen. Fig. 8 exposes the ERT model as the best performing base model for most of the buses in the test dataset. In contrast, the weakest model is given by the SVR.

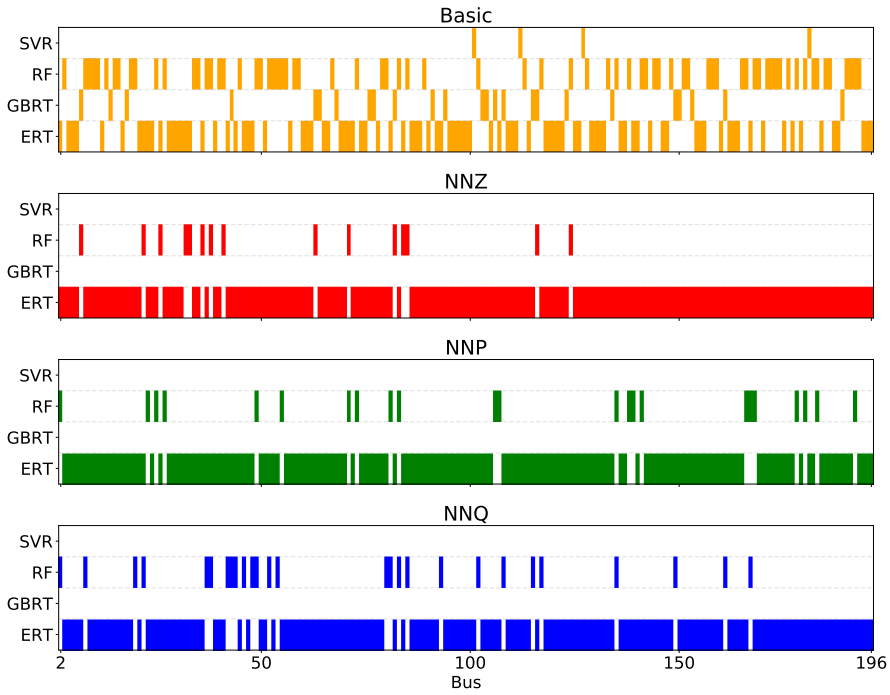

Figure 8: Indication of best individual model on $\mathbf{D}_{\text{Test}}$ for MAPE.

3.3. Ensemble Learning

Ensemble learning is an effective approach to combine different individual models to enhance forecasting performance further. The motivation behind ensemble learning is also rooted in the bias-variance trade-off. Ultimately, ensemble methods aim for a reduction in both bias and variance.

3.3.1. Deterministic Ensemble

The weighted linear combination is a commonly used approach and can be formulated as a convex optimization problem:

$$\begin{aligned} \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \quad & \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \\ \text{subject to} \quad & \hat{y}_i = \sum_{m=1}^{M} w_m \hat{y}_{i,m} \qquad i = \{1, ..., N\} \\ & \sum_{m=1}^{M} w_m = 1 \qquad w_m \geq 0 \end{aligned} \tag{10}$$

with N being the number of data points, M the number of base models, y_i the real value, $\hat{y}_i$ the combined forecast of the base forecasts $\hat{y}_{i,m}$ and w_m the

optimal weights. The batch weights are constrained to sum up to one and to be non-negative.

3.3.2. Probabilistic Ensemble

Quantile regression is a regression model to produce a series of quantiles instead of the expected value. For a linear model, we find the quantile weights $\mathbf{w}_q$ through the minimization problem:

$$\begin{aligned} \hat{\mathbf{w}}_q &= \arg\min_{\mathbf{w}_q} \sum_{i=1}^{N} \ell_q(y_i, \mathbf{x}_i \cdot \mathbf{w}_q) \\ \ell_q(y, \hat{y}_q) &= \begin{cases} q(y - \hat{y}_q) & \text{if } y \geq \hat{y}_q \\ (q-1)(y - \hat{y}_q) & \text{if } y < \hat{y}_q \end{cases} \end{aligned} \tag{11}$$

Hereby, ℓ_q is the so-called quantile loss and represents a tilted absolute value function. In probabilistic forecasting, quantiles are used to construct a Prediction Interval (PI) with corresponding Prediction Interval Nominal Confidence (PINC):

$$\begin{aligned} PI &= [y_{q_l}, y_{q_u}] \\ PINC[\%] &= P(y_{q_l} \leq Y \leq y_{q_u}) = 100\% \cdot (1 - \alpha) \\ \alpha &= 1 - (q_u - q_l) \qquad \alpha \in (0, 1) \end{aligned} \tag{12}$$

The quantiles numbers y_{q_l} and y_{q_u} are the lower and upper bound of the PI, while α is the nominal confidence. The PINC describes the likelihood of a new observation to fall into the PI range $[y_{q_l}, y_{q_u}]$

Ensemble learning can also be used in connection with quantile regression. A method following this rationale is known as Quantile Regression Averaging (QRA) [31]. It uses the forecasts of deterministic base models as input regressors [31]. Thus, an input feature of the QRA reads:

$$\mathbf{x}_i = [1, \hat{y}_{1,i}, ..., \hat{y}_{M,i}] \qquad i \in [1, ..., N] \tag{13}$$

where the constant offset represents the intercept and $\hat{y}_{m,i}$ the i-th forecast of the m-th deterministic base model.

For both deterministic and probabilistic forecasting, the optimization problems of (10) and (11) are convex and can be easily solved in an offline fashion on dataset $\mathbf{D}_{\text{Ensemble}}$.

4. Experimental Results

In this section, the benchmarks for comparisons and evaluation metrics are introduced. Afterwards, the results of deterministic and probabilistic nodal voltage forecasting are reported.

4.1. Benchmarks

In order to verify the superiority of the ensemble methods presented in section 3.3, three other approaches for ensemble learning are also tested. In the course of investigations, it crystallized that linear learners were more suitable as ensemble combiner for non-linear base models. Therefore, we decided to take linear stochastic approximation methods as benchmarks. The three benchmark methods are based on the general optimization problem [32]

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \Big[d(\mathbf{w}, \mathbf{w}_t) + \eta_t \ell(y_t, \mathbf{w} \cdot \mathbf{f}_t) \Big] \tag{14}$$

where $d(\cdot)$ is a distance function, $\ell(\cdot)$ a loss function and η_t the learning rate. These two terms have opposing objectives for each stochastic training pass t. On one hand, the distance function tries to prevent information loss by keeping the new weights close to the old ones. On the other hand, the loss function tries to integrate the new sample by having a small loss on it.

1. SGD: Stochastic Gradient Descent (SGD) [33] is a well-known iterative optimization algorithm. By choosing the l_2-distance and an arbitrary loss for (14), the weight update can be written as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}_t} \ell(y_t, \mathbf{w}_t \cdot \mathbf{f}_t) \tag{15}$$

   In deterministic forecasting we use the squared loss function, whereas in probabilistic forecasting we minimize the quantile function. We denote the latter model as Quantile Stochastic Gradient Descent (QSGD).

2. Adam: Adaptive Moment (Adam) [34] is an extension of SGD and is based on adaptive estimates of lower-order moments according to the formula

$$\begin{gathered} \mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{\sqrt{\mathbf{v}_t} + s} \mathbf{m}_t \\ \mathbf{m}_t = \frac{\gamma_1 \mathbf{m}_{t-1} + (1-\gamma_1)\nabla \ell_t}{1-\gamma_1^t} \qquad \mathbf{v}_t = \frac{\gamma_2 \mathbf{v}_{t-1} + (1-\gamma_2)\nabla \ell_t^2}{1-\gamma_2} \end{gathered} \tag{16}$$

   where the terms $\mathbf{m}_t$ and $\mathbf{v}_t$ are bias-corrected mean and variance terms of the gradient, γ_1 as well as γ_2 are decaying constants and s is a smoothing constant.

3. PAR: Passive Aggressive Regression (PAR) [35] is another linear approximation method obtained by choosing the l_2-distance and the ε-insensitive loss for (14). The resulting weight update has the form

$$\mathbf{w}_{t+1} = \mathbf{w}_t + sign(y_t - \mathbf{w}_t \cdot \mathbf{x}_t)\tau_t \mathbf{x}_t \qquad \tau_t = min \left\{ C, \frac{\ell_\varepsilon(y_t, \mathbf{w}_t \cdot \mathbf{x}_t)}{\|\mathbf{x}_t\|_2^2} \right\} \tag{17}$$

   where $sign(\cdot)$ is the signum function, C the aggressiveness parameter and $\ell_\varepsilon(\cdot)$ the ε-insensitive loss.

All benchmarks are trained with a stochastic approximation procedure with early stopping. To this end, a train-test shuffling procedure without replacement is employed with parameters given in Table 5. The hyperparameters for the different models are shown in Table 6. The learning rate for SGD was found through cross-validation with a common decay parameter. For Adam, the recommended parameters by the authors of the algorithm were taken [34]. In the case of PAR, the model furthermore displayed an insensitivity for a large range of hyperparameters. For this reason, we used common default values for ϵ and C. It should be noted that insensitivity to changes in hyper-parameters is a good quality of the model since it does not require strong expert knowledge to fine-tune the model. It may boost acceptability at the industry level.

Table 5: Early stopping parameters

Parameter	**Method**		
	SGD	Adam	PAR
Maximum Epochs	1000	1000	1000
Validation Split	75%/25%	75%/25%	75%/25%
Patience	5	5	5
Tolerance	0.001	0.001	0.1

Table 6: Selected hyperparameters for all case studies

Method	**Hyperparameter**			
SGD	$\eta_0 = 0.01$	$p = 0.25$		
Adam	$\eta_t = 0.001$	$\gamma_1 = 0.9$	$\gamma_2 = 0.999$	$s = 10^{-8}$
PAR	$\varepsilon = 0.1$	$C = 1$		

4.2. Evaluation Metrics

For the purpose of evaluating the deterministic point forecasts $\hat{y}_t$ with respect to the true values y_t for a batch of data, the three common metrics MAE, MAPE, and RMSE are assessed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{18}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{20}$$

A natural evaluation metric of a probabilistic forecasting model is the quantile loss since it is also minimized in quantile regression. We will define a model's quantile loss on a batch of data as Pinball Loss (PBL) with:

$$PBL = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\mathcal{Q}}\sum_{q}\ell_q(y_i,\hat{y}_q)\right) \tag{21}$$

where q is the particular quantile, $\mathcal{Q}$ the total number of quantiles and ℓ_q the quantile loss defined in 11. The summation over the quantiles is calculated for the whole set $q \in \{0.01, ..., 0.99\}$ implying $\mathcal{Q} = 99$.

When assessing the prediction interval derived from a probabilistic forecasting model, the reliability and sharpness of the interval are key properties. The reliability can be examined with the Average Coverage Error (ACE). This metric is specified by the difference between the PI coverage probability and nominal confidence $(1-\alpha)$ reading

$$ACE = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\{y_i\in[L_i,U_i]\}} - (1-\alpha) \tag{22}$$

The ACE can either be positive or negative, and a well-performing model displays an ACE close to zero [36]. The sharpness may be assessed with the Winkler Score (WKS) given by

$$WKS = \frac{1}{N}\sum_{i=1}^{N}W(y_i,L_i,U_i) \tag{23}$$

where, $\delta_i = U_i - L_i$ is the interval width defined by lower and upper bound L_i and U_i of the PI and W denotes the winkler loss which is expressed as follows:

$$W(y_i,L_i,U_i) = \begin{cases} \delta_i, & L_i \le y_i \le U_i \\ \delta_i + 2(L_i - y_i)/\alpha, & y_i < L_i \\ \delta_i + 2(y_i - U_i)/\alpha, & y_i > U_i \end{cases} \tag{24}$$

Intuitively, the WKS penalizes the performance relative to the distance of the constructed PI. Therefore, a high-quality PI has a low WKS [37].

4.3. Deterministic Forecasting

Fig. 9 depicts the performance of the best individual method, i.e., ERT, and our proposed ensemble method (OPT) regarding the different scenarios, respectively. The corresponding mean and standard derivations of different methods on all the 195 nodes are provided in Table 7. It stands out that the proposed OPT method with Q-distance has the best performance. Compared with the best individual method (ERT-NNZ), there is about an average improvement of 1.09% in terms of MAE, MAPE, and RMSE. Besides, the standard deviation

of these three evaluation metrics is also smaller than the best individual methods. This means that the forecasts of our method are more reliable for these three metrics since the forecasting error moves in a smaller range. Note that we are examining ensemble models in the case study. The performance of the ensemble model is strongly dependent on its base models. Since all ensemble combiners have the same base with strong learners such as SVR, which already have high accuracy, an average improvement of 1.09% in terms of MAE, MAPE, and RMSE and an average improvement of 9.48% in terms of their variances for the ensemble combiner are nontrivial.

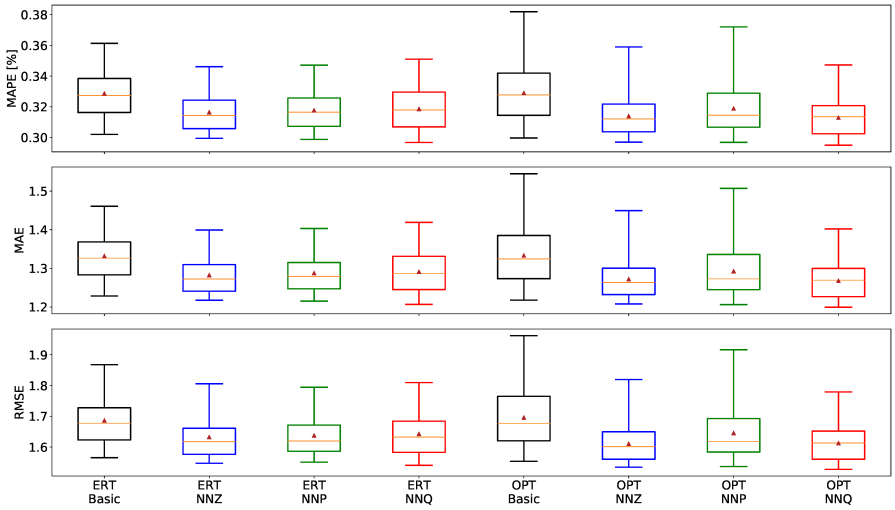

Figure 9: Deterministic forecasts for different distance scenarios.

Table 7: Deterministic forecasting performance with different distance

	MAE	MAE-std	RMSE	RMSE-std	MAPE	MAPE-std
ERT-Basic	1.332	0.057	1.686	0.076	0.329%	0.014%
ERT-NNP	1.288	0.048	1.638	0.065	0.318%	0.012%
ERT-NNQ	1.291	0.052	1.642	0.068	0.319%	0.013%
ERT-NNZ	1.282	0.051	1.633	0.070	0.316%	0.013%
OPT-Basic	1.333	0.070	1.695	0.089	0.329%	0.017%
OPT-NNP	1.292	0.062	1.646	0.081	0.319%	0.015%
OPT-NNQ	**1.268**	**0.045**	**1.613**	**0.056**	**0.313%**	**0.011%**
OPT-NNZ	1.272	0.046	1.611	0.057	0.314%	0.012%
Improvement	1.09%	6.25%	1.22%	13.85%	0.95%	8.33%

It is also interesting to see that not all OPT methods have better performance than the corresponding best individual method. For example, combining four basic forecasts (OPT-Basic) without consideration of nearest neighbor-

hood nodes performs worse than ERT. One possible reason is that the four basic individual methods do not provide a diversity of forecasting. Moreover, the difference between the Z- and Q-distance metrics is marginal. For this reason, we continue the investigation with the Z-distance due to its computational advantages for the nearest neighbor selection.

Fig. 10 illustrates the deterministic forecasting results obtained by different ensemble combining algorithms for the Z-distance. It can be found that compared with SGD, Adam, and PAR algorithms, the OPT method has the best performance with the lowest variances in terms of MAE, MAPE, and RMSE. Fig. 11 shows the deterministic forecasts obtained by our proposed method for Bus 100 over one week. The predicted profile can effectively capture the basic trend of the voltage magnitude changes.

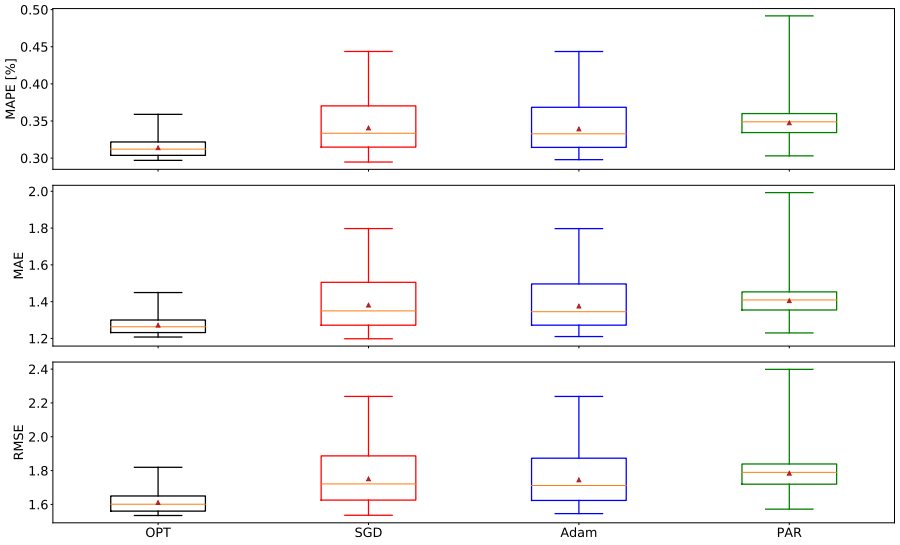

Figure 10: Box plot of deterministic forecasting performances for the Z-distance.

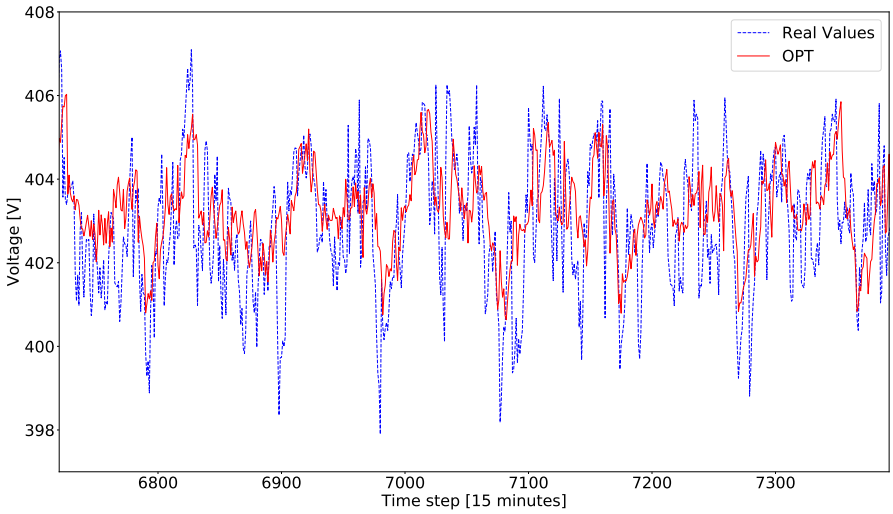

Figure 11: OPT forecast for Bus 100 over one week.

In order to validate the forecasting accuracies, the modified Diebold-Mariano test [38, 39] is performed for the MAE and averaged over all buses. The corresponding results are shown in Fig. 12. The null hypothesis is formulated as the fact that the forecasts of the two models in question have the same accuracy. We see that in the case of a significance level of 8 %, the null hypothesis can be rejected for all tests involving OPT since the p-values is below 0.1. Therefore, OPT does not have the same accuracy as the other models according to the Diebold-Mariano test.

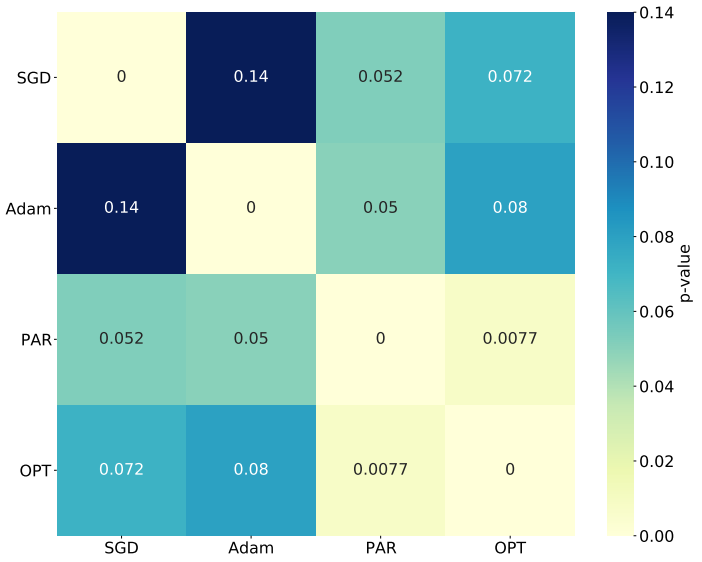

Figure 12: P-values of the Diebold-Mariano test for the MAE criterion averaged over all buses.

Fig. 13 shows the MAPE of the 195 buses and its corresponding Z-distance (Ohm) from bus 2, the nearest bus to the feeder bus on the low voltage. There is a clear trend that with the increase of Z-distance from bus 2, the MAPE increases, i.e., the predictability of the nodal voltage decreases. The main reason is that the voltage of the root bus is fixed, and the further from bus 2, the larger the variance of the nodal voltage.

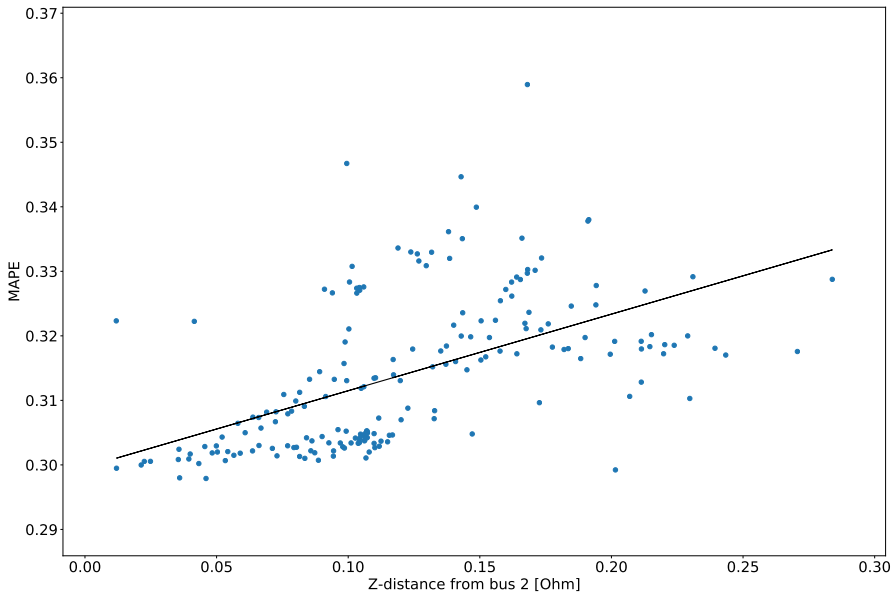

Figure 13: Scatter plot of MAPE and Z-distance from bus 2.

4.4. Probabilistic Forecasting

Fig. 14 depicts the performance of different probabilistic ensemble methods in terms of PBL, WKS, and ACE, respectively. The corresponding mean and standard derivations of different methods on all the 195 nodes are provided in Table 8. It should be noted that the absolute value of ACE is given for ease of comparison.

Compared with QSGD, QAdam and QPAR, the proposed QRA model has the best performance for all three criteria, which means our proposed method has better reliability, sharpness as well as calibration. Especially for ACE, our proposed method largely outperforms others. Fig. 15 shows the probabilistic forecasts obtained by our proposed method for Bus 100 over one week. The predicted interval can cover most of the voltage, and the interval changes over time which is identical to the change of real voltage magnitude. Again, this figure verifies the reliability and calibration of our forecasts.

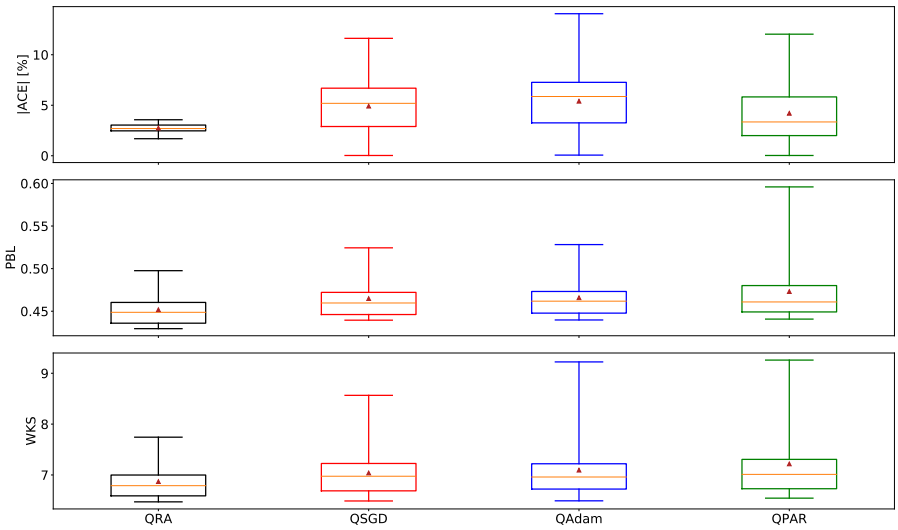

Figure 14: Box plot of probabilistic forecasting performances for the Z-distance

Table 8: Probabilistic forecasting performance with different ensemble methods

	ACE	ACE-std	PBL	PBL-std	WKS	WKS-std
QRA	**2.75%**	**0.37%**	**0.452**	**0.019**	**6.869**	**0.341**
QSGD	4.92%	2.62%	0.465	0.022	7.040	0.445
QAdam	5.40%	2.82%	0.466	0.022	7.095	0.536
QPAR	4.21%	2.94%	0.473	0.033	7.218	0.673

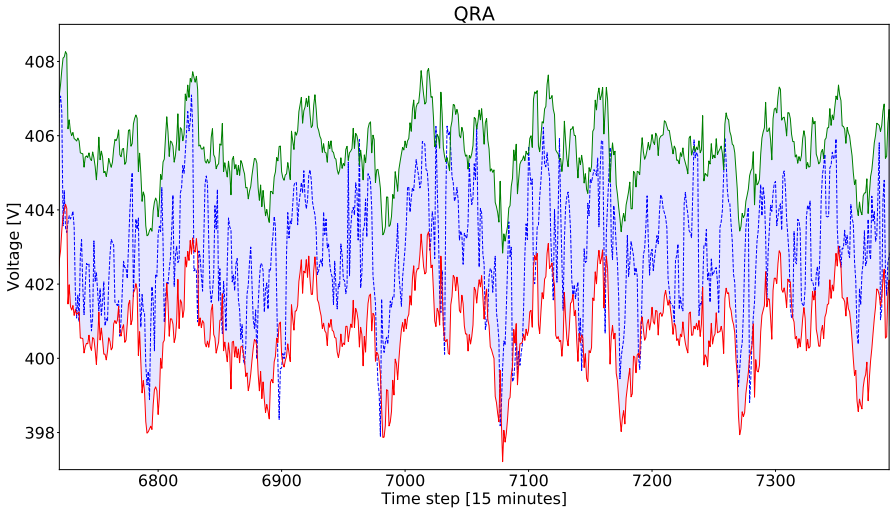

Figure 15: QRA forecast for Bus 100 over one week.

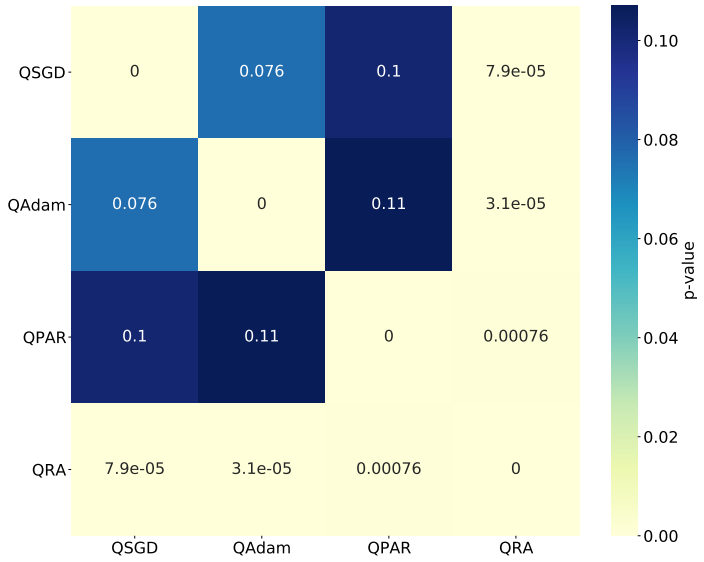

Figure 16: P-values of the Diebold-Mariano test for the PBL criterion averaged over all quantiles and all buses.

The forecasting accuracies for the PBL are again validated with the modified Diebold-Mariano by averaging over all quantiles and buses. The resulting p-values are given in Fig. 16. The null hypothesis is formulated as the fact that the forecasts of the two models in question have the same accuracy. Assuming

again a significance level of 1%, the null hypothesis can be rejected for all tests involving QRA since the p-values are below 0.1. Therefore, QRA does not have the same accuracy as the other models according to the Diebold-Mariano test.

Fig. 17 shows the PBL of 195 buses and its corresponding Z-distance (Ohm) from bus 2. Similar to Fig. 13, the probabilistic forecasting performance gets worse with the increase of Z-distance from bus 2. This again highlights the importance of voltage regulation of the bus that are far from the bus.

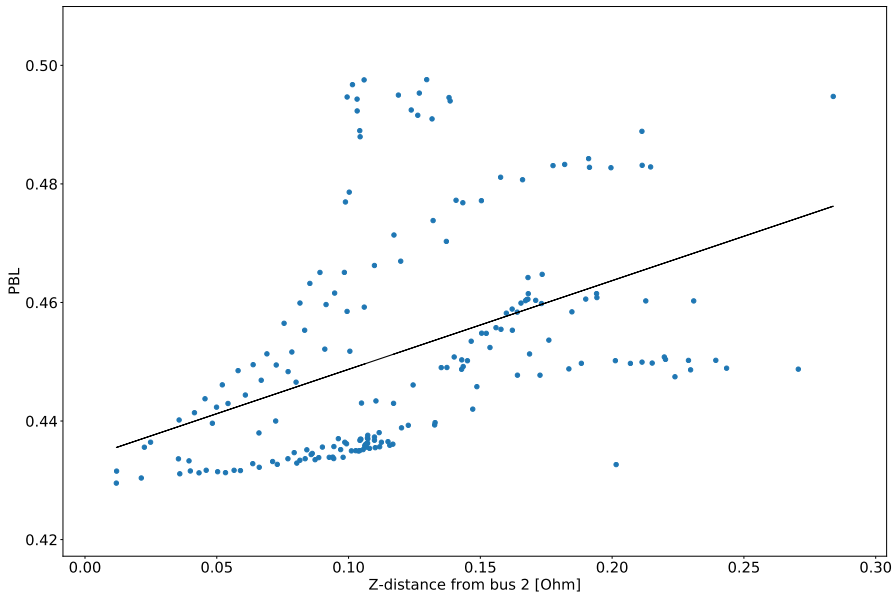

Figure 17: Scatter plot of PBL and Z-distance from bus 2.

5. Conclusions

This paper proposes two-stage nodal voltage forecasting methods. In the first stage, several individual deterministic forecasting models are trained by integrating model- and data-driven feature selection. In the second stage, weighted averaging and QRA models are proposed to combine individual models to provide final deterministic and probabilistic forecasting. We demonstrated that joint model- and data-driven feature selection as well as ensemble learning could effectively improve the performance of short-term nodal voltage. We also observed that the distance to the root node has a close relationship with the predictability of nodal voltage.

References

[1] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, H. Zareipour, Energy forecasting: A review and outlook, IEEE Open Access Journal of Power and Energy 7 (2020) 376–388. doi:10.1109/OAJPE.2020.3029979.

[2] J. Lin, C. Wan, Y. Song, R. Huang, X. Chen, W. Guo, Y. Zong, Y. Shi, Situation awareness of active distribution network: Roadmap, technologies, and bottlenecks, CSEE Journal of Power and Energy Systems 2 (3) (2016) 35–42.

[3] S.-E. Razavi, E. Rahimi, M. S. Javadi, A. E. Nezhad, M. Lotfi, M. Shafiekhah, J. P. Catalão, Impact of distributed generation on protection and voltage regulation of distribution systems: A review, Renewable and Sustainable Energy Reviews 105 (2019) 157–167.

[4] A. Bracale, P. Caramia, G. Carpinelli, A. R. Di Fazio, P. Varilone, A bayesian-based approach for a short-term steady-state forecast of a smart grid, IEEE Trans. Smart Grid 4 (4) (2013) 1760–1771.

[5] R. Dobbe, W. van Westering, S. Liu, D. Arnold, D. Callaway, C. Tomlin, Linear single-and three-phase voltage forecasting and bayesian state estimation with limited sensing, IEEE Trans. Power Systems 35 (3) (2020) 1674–1683.

[6] B. P. Hayes, M. Prodanovic, State forecasting and operational planning for distribution network energy management systems, IEEE Trans. Smart Grid 7 (2) (2016) 1002–1011.

[7] A. Dejamkhooy, A. Dastfan, A. Ahmadyfard, Modeling and forecasting nonstationary voltage fluctuation based on grey system theory, IEEE Trans. Power Delivery 32 (3) (2017) 1212–1219.

[8] M. Hassanzadeh, C. Y. Evrenosoğlu, L. Mili, A short-term nodal voltage phasor forecasting method using temporal and spatial correlation, IEEE Trans. Power Systems 31 (5) (2015) 3881–3890.

[9] A. F. Bastos, S. Santoso, V. Krishnan, Y. Zhang, Machine learning-based prediction of distribution network voltage and sensor allocation, in: IEEE Power & Energy Society General Meeting (PESGM), IEEE, 2020, pp. 1–5.

[10] T. Zufferey, S. Renggli, G. Hug, Probabilistic state forecasting and optimal voltage control in distribution grids under uncertainty, Electric Power Systems Research 188 (2020) 106562.

[11] J. Tian, K. Li, W. Xue, An adaptive ensemble predictive strategy for multiple scale electrical energy usages forecasting, Sustainable Cities and Society 66 (2021) 102654. doi:https://doi.org/10.1016/j.scs.2020.102654. URL https://www.sciencedirect.com/science/article/pii/S2210670720308702

[12] G. T. Ribeiro, V. C. Mariani, L. dos Santos Coelho, Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting, Engineering Applications of Artificial Intelligence 82 (2019) 272–281. doi:https://doi.org/10.1016/j.engappai.2019.03.012.
URL https://www.sciencedirect.com/science/article/pii/S0952197619300624

[13] K. Bhatia, R. Mittal, J. Varanasi, M. Tripathi, An ensemble approach for electricity price forecasting in markets with renewable energy resources, Utilities Policy 70 (2021) 101185. doi:https://doi.org/10.1016/j.jup.2021.101185.
URL https://www.sciencedirect.com/science/article/pii/S0957178721000199

[14] M. Narajewski, F. Ziel, Ensemble forecasting for intraday electricity prices: Simulating trajectories, Applied Energy 279 (2020) 115801. doi:https://doi.org/10.1016/j.apenergy.2020.115801.

[15] R. G. da Silva, M. H. D. M. Ribeiro, S. R. Moreno, V. C. Mariani, L. dos Santos Coelho, A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting, Energy 216 (2021) 119174. doi:https://doi.org/10.1016/j.energy.2020.119174.
URL https://www.sciencedirect.com/science/article/pii/S0360544220322817

[16] J. Zhao, J. Wang, Z. Guo, Y. Guo, W. Lin, Y. Lin, Multi-step wind speed forecasting based on numerical simulations and an optimized stochastic ensemble method, Applied Energy 255 (2019) 113833. doi:https://doi.org/10.1016/j.apenergy.2019.113833.

[17] L. Buzna, P. De Falco, G. Ferruzzi, S. Khormali, D. Proto, N. Refa, M. Straka, G. van der Poel, An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations, Applied Energy 283 (2021) 116337. doi:https://doi.org/10.1016/j.apenergy.2020.116337.

[18] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen, C. Kang, Combining probabilistic load forecasts, IEEE Transactions on Smart Grid 10 (4) (2018) 3664–3674.

[19] J. Nowotarski, R. Weron, Computing electricity spot price prediction intervals using quantile regression and forecast averaging, Computational Statistics 30 (3) (2015) 791–803.

[20] B. Liu, J. Nowotarski, T. Hong, R. Weron, Probabilistic load forecasting via quantile regression averaging on sister forecasts, IEEE Transactions on Smart Grid 8 (2) (2015) 730–737.

[21] F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, M. Shi, A day-ahead pv power forecasting method based on lstm-rnn model and time correlation modification under partial daily pattern prediction framework, Energy Conversion and Management 212 (2020) 112766. doi:https://doi.org/10.1016/j.enconman.2020.112766.
URL https://www.sciencedirect.com/science/article/pii/S0196890420303046

[22] H. Abbasimehr, M. Shabani, M. Yousefi, An optimized model using lstm network for demand forecasting, Computers & Industrial Engineering 143 (2020) 106435. doi:https://doi.org/10.1016/j.cie.2020.106435.
URL https://www.sciencedirect.com/science/article/pii/S0360835220301698

[23] S. Rodrigues Moreno, R. Gomes da Silva, V. Cocco Mariani, L. dos Santos Coelho, Multi-step wind speed forecasting based on hybrid multi-stage decomposition model and long short-term memory neural network, Energy Conversion and Management 213 (2020) 112869. doi:https://doi.org/10.1016/j.enconman.2020.112869.
URL https://www.sciencedirect.com/science/article/pii/S0196890420304076

[24] Z. Chang, Y. Zhang, W. Chen, Electricity price prediction based on hybrid model of adam optimized lstm neural network and wavelet transform, Energy 187 (2019) 115804. doi:https://doi.org/10.1016/j.energy.2019.07.134.
URL https://www.sciencedirect.com/science/article/pii/S0360544219314768

[25] P. Cuffe, A. Keane, Visualizing the electrical structure of power systems, IEEE Systems Journal 11 (3) (2017) 1810–1821.

[26] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[27] V. Cherkassky, Y. Ma, Practical selection of svm parameters and noise estimation for svm regression, Neural Networks 17 (1) (2004) 113–126.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, G. et. al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[29] M. LeBlanc, J. Crowley, A review of tree-based prognostic models, Springer US, Boston, MA, 1995, pp. 113–124.

[30] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics, 2009.

[31] J. Nowotarski, R. Weron, Computing electricity spot price prediction intervals using quantile regression and forecast averaging, Computational Statistics 30 (3) (2015) 791–803.

[32] J. Kivinen, M. K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors, Information and Computation 132 (1) (1997) 1 – 63.

[33] H. E. Robbins, A stochastic approximation method, Annals of Mathematical Statistics 22 (2007) 400–407.

[34] D. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations (12 2014).

[35] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer, Online passive-aggressive algorithms, Journal of Machine Learning Research 7 (2006) 551–585.

[36] C. Wan, Z. Xu, Y. Wang, Z. Y. Dong, K. P. Wong, A hybrid approach for probabilistic forecasting of electricity price, IEEE Trans. Smart Grid 5 (1) (2014) 463–470.

[37] R. L. Winkler, A decision-theoretic approach to interval estimation, Journal of the American Statistical Association 67 (337) (1972) 187–191.

[38] F. X. Diebold, R. S. Mariano, Comparing predictive accuracy, Journal of Business & Economic Statistics 20 (1) (2002) 134–144. doi:10.1198/073500102753410444.

[39] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, International Journal of Forecasting 13 (2) (1997) 281–291. doi:https://doi.org/10.1016/S0169-2070(96)00719-4.
URL https://www.sciencedirect.com/science/article/pii/S0169207096007194