Improving Clustering-Based Forecasting of Aggregated Distribution Transformer Loadings With Gradient Boosting and Feature Selection

Received December 8, 2021, accepted December 16, 2021, date of publication December 23, 2021,
date of current version January 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137870

Improving Clustering-Based Forecasting of Aggregated Distribution Transformer Loadings With Gradient Boosting and Feature Selection

**GEORGE ROUWHORST[1], (Member, IEEE),
EDGAR MAURICIO SALAZAR DUQUE[1], (Member, IEEE),
PHUONG H. NGUYEN[1], (Member, IEEE), AND HAN SLOOTWEG[1,2], (Senior Member, IEEE)**
[1]Department of Electrical Engineering, Eindhoven University of Technology, 5612AP Eindhoven, The Netherlands
[2]Asset Management, Enexis Netbeheer B.V., 5223MB 's-Hertogenbosch, The Netherlands

Corresponding author: George Rouwhorst (g.d.g.rouwhorst@tue.nl)

This work was supported in part by the project 'Using small data and big data: Neighborhood Energy and Data Management Integration System' (S&B NEDMIS) from 'de Nederlandse Organisatie voor Wetenschappelijk Onderzoek' (NWO).

ABSTRACT Load forecasting is more important than ever to enable new monitor and control functionalities of distribution networks aiming to mitigate the impact of the energy transition. Load forecasting at medium voltage (MV) level is becoming more challenging, because these load profiles become more stochastic due to the increasing penetration of photovoltaic (PV) generation in distribution networks. This work combines medium to low voltage (MV/LV) transformer loadings measured with advanced metering infrastructure (AMI) and machine learning (ML) algorithms to propose a new clustering based day-ahead aggregated load forecasting approach. This four-step approach improves the day-ahead load forecast of a city. First, MV/LV transformer loadings are clustered based on the shape of their load pattern. Second, a gradient boosting algorithm is used to forecast the load of each cluster and calculate the related feature importance. Third, feature selection is applied to improve the forecast accuracy of each cluster. Finally, the day-ahead load forecast of all clusters are aggregated. The case study presented uses 519 measured MV/LV transformer loadings in a city to perform 30 day-ahead load forecasts. Compared against the day-ahead aggregated load forecast without clustering, the average normalized root mean squared error (NRMSE) reduced 12.7 %, the average mean absolute percentage error (MAPE) reduced 18.2 % and the average Pearson Correlation Coefficient (PCC) increased 0.37 %. The 95 % confidence interval of the difference between the average NRMSE, MAPE and PCC without clustering and with the proposed method indicates a statistically significant improvement in accuracy.

INDEX TERMS Aggregated load forecast, clustering, day-ahead, distribution network, feature selection, gradient boosting.

I. INTRODUCTION

Due to the ongoing energy transition an increasing number of Renewable Energy Sources (RES), especially Photovoltaic (PV) generation is connected to distribution networks. At the same time, electricity demand is increasing due to the adoption of Electric Vehicles (EV) and Heat Pumps (HP). As a consequence, traditional unidirectional power flows in distribution networks become bidirectional, electricity generation becomes harder to control and the peak demand increases rapidly [1].

To cope with these issues, Distribution System Operators (DSO) are installing equipment to measure loads in distribution networks at medium to low voltage transformers (MV/LV transformers) and LV feeders. In addition, DSOs are also installing smart meters at the end-customer level. The adoption of this Advanced Metering Infrastructure (AMI) improves monitor and control functionalities of distribution networks by DSOs aiming to mitigate the impact of the energy transition [2], [3]. Enriching data collected from AMI has driven the development of Machine Learning (ML)

The associate editor coordinating the review of this manuscript and approving it for publication was Alexander Micallef.

algorithms for load forecasting, which is widely studied in literature recently. Load forecasts have been applied for many applications depending on the studied time horizon and time resolution, such as day-ahead scheduling, network reinforcement decisions and integration of flexibility services [4]. Short-term Load Forecasting (STLF) aims to forecast load profiles to study network operation applications with a typical time horizon up to a week-ahead. The related time resolution varies up to an hour. On the contrary, Long-term Load Forecasting (LTLF) aims to forecast peak loads during a predefined interval to study network planning applications. The typical time horizon varies up to multiple years [5]. Medium-term Load Forecasting (MTLF) aims to forecast load profiles to study applications, which are used to optimize the utilization of the current network capacity. On the one hand, the typical time horizon of MTLF is exceeding STLF, varying from a week up to a year ahead with usually an hourly time resolution [6]. On the other hand, MTLF is not only forecasting peak loads as in the case of LTLF [7].

A variety of ML algorithms have been deployed extensively for load forecasting so far and many reviews have been written about these studies as well. Reference [8] provides an overview of publications focusing on electricity prediction using ML algorithms between 2005-2015. The use of Neural Networks (NN) is one of the most widely studied methods for load forecasting, including variations, such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN) and Deep Belief Networks (DBN). Other frequently encountered algorithms in literature are based on Support Vector Machines (SVM) and Decision Trees (DT). Driven by the increasing computational power and available data, more advanced ML algorithms, such as Deep Learning (DL), Reinforcement Learning (RL) and Transfer Learning (TL) are introduced to the field of load forecasting more recently to further improve the accuracy [2], [9]–[11]. The aim of a ML algorithm is to find the optimal correlation between a measured load and a set of input parameters, so called features, during the training process. All ML algorithms share the same fundamental assumption that the load to be forecasted has a similar correlation with the set of features as the measured load used for the training process. Subsequently, the ML algorithm is applied to forecast the load based on a set of given features. Traditional regression techniques require stationary properties of the time-series to be forecasted. However, the advantage of ML algorithms over traditional regression techniques is to also find correlations between features and non-stationary time-series, such as load profiles. On the contrary, the pitfall of (advanced) ML algorithms is the increasing complexity of models depending on many parameters and the loss of interpretability of many ML algorithms [10], [12].

So far, literature focuses mainly on load forecasting at the high voltage level of electricity networks. However, one of the main challenges from load profiles measured at lower voltage levels in distribution networks, is that they are more stochastic. In general, the stochasticity of a load profile increases if the voltage level at which the load profile is measured decreases. As a consequence, load profiles measured in distribution networks are harder to forecast. In addition, penetration of solar PV into distribution networks is making forecasting of load profiles in the distribution network even more challenging due to its intermittent properties [11].

To deal with large amounts of stochastic load profiles in an efficient and reliable way, recent studies propose to start with a clustering algorithm to identify load profiles with similar patterns. The aim of clustering algorithms is to minimize the difference between patterns of load profiles within in each cluster, while maximizing the difference between patterns of load profiles with other clusters. Load profile clustering has also been researched extensively for customer characterization and design tariff structures [13]–[18]. Reference [19] summarizes the most important clustering methods applied to smart meter data. Subsequently, the typical load profile of each identified cluster is used for forecasting and a large number of load profiles can be analyzed, without training and forecasting every load profile individually. Furthermore, clustering of load profiles improves the accuracy of the forecast, because the stochastic properties of load profiles are reduced which enables the forecasting algorithm to be trained with higher correlated data [3], [20]–[24].

Recent studies proposed new methods to improve the accuracy of described clustering based forecasting approaches. Reference [25] described limitations of studies so far due to the independent optimization of clustering and forecasting. Therefore, this study proposes to reduce the forecast error by integrating both optimizations in a closed-loop clustering algorithm. The study described in [26] is based on a novel ensemble forecasting method. Frequently, ensemble forecasting improves the forecast of a load profile by combining multiple forecasts of the same profile with different methods. A weight is assigned to the forecast of each method to reduce the forecast error [10]. However, the study proposed in [26] assigns a weight to the forecast of each cluster to reduce the error of the aggregated load. Reference [27] proposed another ensemble forecasting method to forecast the aggregated load of all clusters. Two forecasting models are applied sequentially and the error of the first model is used by the second model to reduce the error of the forecast.

Until now, many clustering based forecasting approaches proposed in literature focus on clustering of measured smart meters to improve the forecast accuracy of many end-customers. However, this proposed clustering based forecasting approach is focused on clustering measured MV/LV transformer loadings to improve the accuracy of their related aggregated load forecast of the medium voltage (MV) network. To support distribution network planning and operation of cities and neighborhoods and mitigate the impact of the energy transition on distribution networks, focus on MV/LV transformer loadings and their related aggregated loading in MV networks is necessary. On the one hand, forecasting individual MV/LV transformer loadings is hard due to stochastic properties leading to an increasing accumulation error when forecasts of MV/LV transformer loadings are

aggregated [11], [28]. On the other hand, a single model to forecast the aggregated MV/LV transformer loadings at once is troublesome due to the difference in patterns between MV/LV transformers loadings caused by varying behaviours and properties of connected end-customers [3], [24].

Therefore, the proposed model in this paper aims to improve the accuracy of the day-ahead aggregated MV/LV transformer loadings forecast using a four-step approach. First, MV/LV transformer loadings are clustered based on the shape of their load pattern. Second, each cluster is forecasted using a gradient boosting algorithm. Third, the forecast accuracy of each cluster is improved using a feature selection process. Finally, the day-ahead load forecast of all clusters are aggregated to forecast the load profile of the related MV network. Due to the feature selection process for each individual cluster, the accumulation error to forecast the aggregated load is also reduced. The contributions of this paper are

- a novel clustering based day-ahead forecasting approach to identify typical patterns of MV/LV transformers loadings in a city or neighborhood to forecast the specific load profile of each cluster. This information is missing if all MV/LV transformers are aggregated directly without clustering. At the same time, a cumbersome approach of forecasting each individual MV/LV transformer loading is avoided.
- a feature selection process based on calculated specific feature importances related to each cluster of MV/LV transformers to further improve the accuracy of the day-ahead load forecast of each cluster.
- a novel clustering based day-ahead forecasting approach to significantly improve the accuracy of the day-ahead aggregated load forecast representing the MV network related to the studied MV/LV transformers.

The remainder of the paper is organized as follows. Section II formulates the mathematical problem formulation, section III describes the proposed methodology of the model and section IV presents the results and discussion. Finally, in section V the main conclusions of the paper are drawn.

II. PROBLEM FORMULATION

For an area with N MV/LV transformers, the aggregated load profile of all N MV/LV transformers loadings is calculated according to:

$$A(t) = \sum_{i=1}^{N} T_i(t), \tag{1}$$

where $A(t)$ is the aggregated load profile of N MV/LV transformers $[kW]$ and $T_i(t)$ the load profile of transformer i $[kW]$ at timestep t.

Reference [29] proofs that the forecast accuracy of the aggregated load profile $A(t)'$ is higher if N MV/LV transformers are first optimally clustered (K>1) and the forecasts of all clusters are subsequently aggregated than in the case when all N MV/LV transformers are directly aggregated and forecasted ($K = 1$). Therefore, this paper evaluates the accuracy of the day-ahead aggregated load forecast of N MV/LV transformers using a gradient boosting algorithm. All N MV/LV transformers are first clustered in a range from $K = 1, \ldots, K_{max}$ to explore for which number of clusters K the N MV/LV transformers are optimally separated.

When N MV/LV transformers are clustered into K clusters, the aggregated load forecast $A(t)'$ is calculated by the summation of the day-ahead load forecast of each cluster k according to:

$$A(t)' = (\sum_{i=1}^{N_1} T_{i,1}(t))' + (\sum_{i=1}^{N_2} T_{i,2}(t))' + \cdots + (\sum_{i=1}^{N_K} T_{i,K}(t))' \tag{2}$$

with N_1 up to N_K representing the number of MV/LV transformers assigned to each cluster k according to:

$$N = N_1 + N_2 + \cdots + N_K, \tag{3}$$

$T_{i,k}(t)$ is the load of MV/LV transformer i in cluster k $[kW]$ and $A(t)'$ is the day-ahead aggregated load forecast of all K clusters $[kW]$ and thereby of all N MV/LV transformers.

The main goal of the forecasting algorithm is to minimise the difference between the measured aggregated load $A(t)$ and the day-ahead aggregated load forecast $A(t)'$. Three metrics are used to evaluate the accuracy of the day-ahead aggregated load forecast, which are the mean absolute percentage error ($MAPE$), the normalized root mean squared error ($NRMSE$) and the Pearson Correlation Coefficient (PCC). A lower value of the $NRMSE$ and $MAPE$ indicate a lower error of the aggregated load forecast $A(t)'$. These two metrics are calculated according to:

$$NRMSE[\%] = \frac{\sqrt{\frac{1}{P}\sum_{i=1}^{P}(A(t)' - A(t))^2}}{A_{max} - A_{min}} \cdot 100, \tag{4}$$

$$MAPE[\%] = \frac{1}{P}\sum_{i=1}^{P}\left|\frac{A(t)' - A(t)}{A(t)}\right| \cdot 100, \tag{5}$$

where P is the sum of the number of timestamps during the period of the forecast and A_{max} and A_{min} are the maximum and minimum value of the aggregated load $[kW]$ [3], [23].

The PCC is used to quantify the degree of linear dependence between the measured aggregated load $A(t)$ and the day-ahead aggregated load forecast $A(t)'$. The PCC may take any value from [-1,1], where zero indicates not any correlation, while minus one and one indicate a completely negative and positive correlation respectively. Therefore, a PCC value closer to one indicates a more accurate day-ahead aggregated load forecast $A(t)'$ [8]. The PCC is calculated according to:

$$PCC = \frac{cov(A(t)', A(t))}{\sigma_{A(t)'} \cdot \sigma_{A(t)}}, \tag{6}$$

where the *covariance* between the measured aggregated load and the aggregated load forecast is calculated according to:

$$cov(A(t)', A(t)) = \sum_{i=1}^{P}(A(t)' - A_{avg}(t)')(A(t) - A_{avg}(t)) \tag{7}$$

and the standard deviation $\sigma_{A(t)}$ is calculated according to:

$$\sigma_{A(t)} = \sqrt{\sum_{i=1}^{P}(A(t) - A_{avg}(t))^2}. \quad (8)$$

The standard deviation related to $\sigma_{A(t)'}$ is calculated accordingly [8].

Thereafter, it is analyzed if all metrics related to the day-ahead aggregated load forecast have improved statistically significant for the optimal number of clusters K when compared with the day-ahead aggregated load forecast without clustering formulated as:

$$NRMSE_{K=1} > NRMSE_{K_{opt}}, \quad (9)$$

$$MAPE_{K=1} > MAPE_{K_{opt}}, \quad (10)$$

$$PCC_{K_{opt}} > PCC_{K_{K=1}}. \quad (11)$$

III. METHODOLOGY

A schematic overview of the proposed clustering based day-ahead aggregated load forecasting approach is shown in Fig. 1. First of all, an algorithm clusters N MV/LV transformers within an area in a range from $K = 1, \ldots, K_{max}$ clusters. Next, the MV/LV transformers loadings assigned to each cluster k are summed and a day-ahead forecast is performed using a gradient boosting algorithm. Finally, the day-ahead forecast of the aggregated load of all N MV/LV transformer loadings is the sum of the day-ahead forecast of all clusters.

A. CLUSTERING

The electricity consumption of a city is equal to the aggregation of all N MV/LV transformer loadings. The objective is to identify clusters over these MV/LV transformers loadings, which have a similar pattern. The similarity in pattern of MV/LV transformers loadings within the same cluster improves the load forecast of each cluster. One approach to cluster load profiles is to first create one representative load profile (RLP) for each MV/LV transformer as an input for the clustering algorithms. Then, find the optimal number of clusters.

The applied clustering algorithms originate from different paradigms such as spectral-based, hierarchical, density-based and representative-based clustering. Specifically, six methods are used for this study, which are Spectral clustering [30], BIRCH [31], hierarchical Ward [32], Gaussian mixture models [33], k-means and k-medoids [34].

1) REPRESENTATIVE LOAD PROFILES

Let the matrix $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N] \in \mathbb{R}^{P \times N}$ be the MV/LV transformer loadings, where P is the total number of data points recorded during a determined period of time, and N the number of MV/LV transformers. Each vector $\boldsymbol{l}_n = [l_{n,1}, \ldots, l_{n,P}]^\intercal \in \mathbb{R}^P$ is the data of the n^{th} MV/LV transformer for $n = 1, \ldots, N$. The daily MV/LV transformer loading is characterized by averaging every m data points of $\boldsymbol{l}_n$. It is represented as $\bar{\boldsymbol{l}}_{\boldsymbol{n}} = [\bar{l}_{n,1}, \ldots, \bar{l}_{n,m}] \in \mathbb{R}^m$.

The matrix of N representative MV/LV transformer loadings is represented by $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\intercal \in \mathbb{R}^{N \times m}$. The vector $\boldsymbol{x}_n$ is the RLP of a MV/LV transformer, which is the standardized daily load pattern calculated as $\boldsymbol{x}_n = (\bar{\boldsymbol{l}}_n - \mu_{\bar{\boldsymbol{l}}_n})/\sigma_{\boldsymbol{l}_n}$, where $\mu_{\bar{\boldsymbol{l}}_n}$ and $\sigma_{\boldsymbol{l}_n}$ are the mean and standard deviation of the data points which constitutes the vector $\bar{\boldsymbol{l}}_n$.

After processing the dataset into the matrix $\boldsymbol{X}$, the clustering algorithms are applied in the next step to determine the optimal number of clusters. The objective of any clustering algorithm is to find the optimal clustering based on similarities between the patterns of all load profiles. The clustering algorithm assigns each load profile of the dataset to an unique cluster. For load profiles, similar characteristics means load profile patterns during the day.

The set of all N MV/LV transformers, which we define as $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, is clustered in K disjoint clusters Ω_k for $k = 1, \ldots, K$, such that $\boldsymbol{X} = \bigcup_{k=1}^{K} \Omega_k$. Each cluster Ω_k is constituted by a set of RLPs labeled by the clustering algorithm. Each transformer group has a centroid $\boldsymbol{C} = \{c_1, \ldots, c_K\}$ which is computed averaging all the RLPs in the same group, i.e., $c_k = |\Omega_k|^{-1} \sum_{x \in \Omega_k} \boldsymbol{x}$ for $k = 1, \ldots, K$, and $|\Omega_k|$ is the number of MV/LV transformers in cluster k.

The RLPs clustering is an unsupervised learning problem without a ground truth. There is no previous labeling available to define a correct clustering, nor best number of clusters K beforehand. Therefore, the output of the clustering algorithms for a predefined range of cluster numbers should be evaluated using internal indices [35]. The validation indices quantifies for each of the clustering algorithm how well the clusters separate the different RLPs (separation), and how similar are the RLPs within the same (compactness).

2) CLUSTERING QUALITY EVALUATION METRICS

Multiple indices for clustering evaluation have been proposed in the past for load profiling [16], [17], [36]. Four indices have been selected to test the performance of the clustering algorithms, and to aid the selection of the best number of clusters K. All these indices are measured in euclidean distance. There are two primary distances definitions on which the evaluation metrics are built. The *vector-to-set* distance, which measures how distant the vector (RLP) is from the cluster, is expressed as:

$$d(\boldsymbol{x}, \Omega) = \sqrt{\frac{1}{M}\sum_{\boldsymbol{y} \in \Omega} \|\boldsymbol{x} - \boldsymbol{y}\|^2}, \quad (12)$$

which $\|\cdot\|$ denotes the euclidean distance and M is the cardinality of the set Ω. i.e, the number of MV/LV transformers in the set $M = |\Omega|$. The *intra-set* distance, computed by the *vector-to-set* distance for all members of the set Ω, is defined as:

$$\hat{d}(\Omega) = \sqrt{\frac{1}{2M}\sum_{\boldsymbol{x} \in \Omega} d^2(\boldsymbol{x}, \Omega)}. \quad (13)$$

The four indicators evaluated over the clustering algorithm results are:

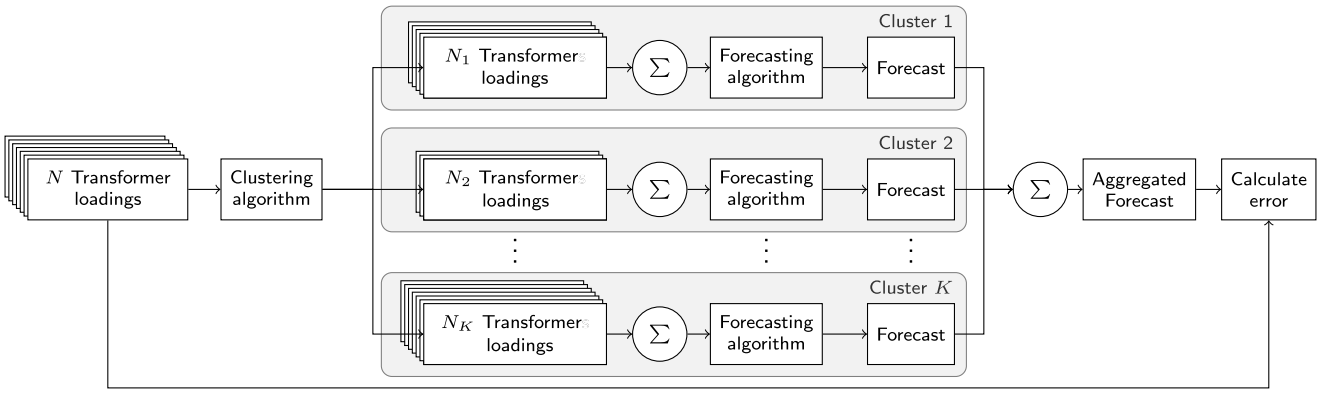

FIGURE 1. Proposed methodology of the clustering based day-ahead forecasting approach of *N* MV/LV transformer loadings.

1) The clustering dispersion indicator (CDI) [16] defined as the ratio between the intraset distance of the RLPs in the same cluster k, and the distance between the K clusters represented by the centroid set C:

$$\text{CDI} = \frac{1}{\hat{d}(\boldsymbol{C})}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\hat{d}^2(\Omega_k)}. \tag{14}$$

2) The modified Dunn index (MDI) [37] computes the distance ratio between the clusters with maximum dispersion, and the closest centroids of the K clusters:

$$\text{MDI} = \max_{1\le k\le K}\{\hat{d}(\Omega_k)\}/\min_{i\neq j}\{\|c_i, c_j\|\}. \tag{15}$$

3) The Davies-Bouldin index (DBI) [38] quantifies the maximum similarity between the K clusters divided by the separation of the sets:

$$\text{DBI} = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq i}\left\{\frac{\hat{d}(c_k, \Omega_k) + \hat{d}(c_i, \Omega_i)}{\|c_k - c_i\|}\right\}, \tag{16}$$

for $i = 1, \ldots, K$. This index can be used to set the "best" choice of clusters, assuming that the clusters are convex. The optimal K minimizes the average similarity between the K clusters (DBI index).

4) The Caliński-Habarasz index (CHI) [39] computes a ratio that compares the dispersion within a cluster k, and the dispersion of the centroids that represent each of the clusters:

$$\text{CHI} = \frac{\text{Tr}(\boldsymbol{S}_B)}{\text{Tr}(\boldsymbol{S}_W)}\cdot\frac{N-k}{k-1}, \tag{17}$$

where $\boldsymbol{S}_W$ is the *intra-set* scatter matrix and $\boldsymbol{S}_B$ is the *inter-set* scatter matrix, defined as:

$$S_B = \sum_{k=1}^{K}|\Omega_k|(c_k - \bar{\boldsymbol{X}})(c_k - \bar{\boldsymbol{X}})^\intercal,$$

$$S_W = \sum_{k=1}^{K}\sum_{\boldsymbol{x}\in\Omega_k}(\boldsymbol{x} - c_k)(\boldsymbol{x} - c_k)^\intercal.$$

The variable $\bar{\boldsymbol{X}}$ is the average of all RLPs in the set $\boldsymbol{X}$, and $\text{Tr}(\cdot)$ is the trace operator. This index penalizes the use of a large number of clusters K, using the second term of (17).

For the DBI, CDI and MDI evaluation indices, lower values indicate a better clustering result; for CHI the opposite applies. To determine the best algorithm and number of clusters for the dataset, an holistic assessment of indexes is performed in the studied case.

B. FORECASTING

Fig. 2 provides an overview of the forecasting model applied to the aggregated load of each cluster, which is described in detail in [7]. However, instead of the described month-ahead forecasting (MTLF) approach, this study involves a day-ahead forecasting (STLF) approach. Therefore, the time interval of the measured load is 15 minutes instead of 60 minutes and the minimum ratio between the training set and forecast is not smaller as described in [7]. The applied features include the same PV-related features, weather-related features and time-related features as described in [7].

The model first decomposes the load profile to be forecasted in stationary profiles and a residual profile, which improves the forecasting accuracy as described in [40]. These stationary profiles are forecasted using an autoregression algorithm.

The applied supervised learning model to forecast the residual profile is a gradient boosting algorithm using DT as the base learner. DT have the advantage of interpreting the learned correlation between the features and forecast. The features are chosen as split points in a DT aiming to minimize the forecast error. Therefore, analyzing the split of features on the reduction of the forecast error indicates the relative importance of features [41]. However, individual DT are usually weak learners leading to inaccurate forecasts. Ensemble learning combines the forecast of many individual weak learners, such as DT, as its base learner to improve the forecast accuracy based on the concept that it is easier to

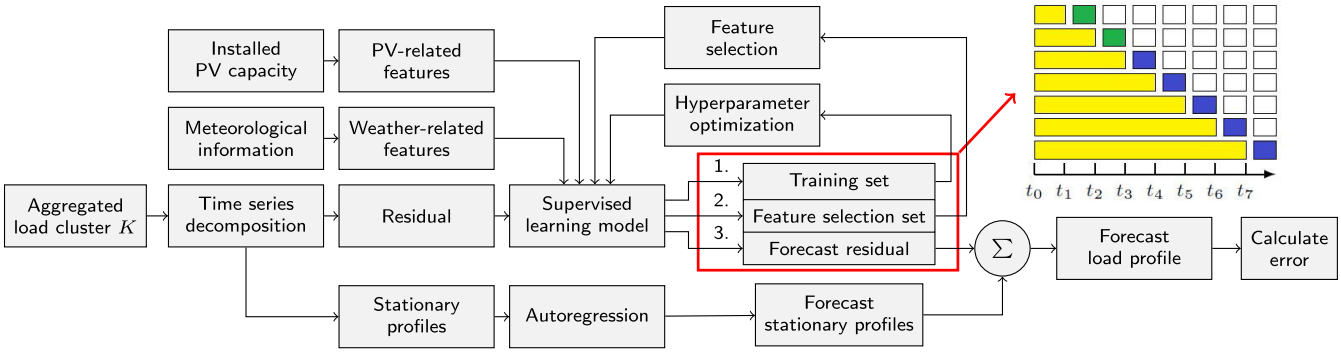

FIGURE 2. Proposed algorithm to forecast the day-ahead aggregated load of cluster *K* with a schematic overview of the applied time series cross validation.

improve the accuracy using the average forecast of many base learners than to find one highly accurate algorithm. Ensemble learning can be performed by parallel (bagging, stacking) -or sequential (boosting) ensembling [42], [43]. The difference is the order of ensembling the forecasts of individual base learners. Parallel ensemble learning methods combine the forecast of all individual base learners to improve the forecast accuracy by assigning weights to the forecast of each base learner. However, sequential ensembling methods, such as (gradient) boosting add a next base learner sequentially. The forecast of the previous base learner is used as input to reduce the forecast error, by assigning different weights based on the accuracy of the forecast of the previous base learner. Unlike regular boosting algorithms, a gradient boosting algorithm is using a gradient descent algorithm to improve the forecast of the previous base learner [43], [44].

To improve the accuracy of the residual forecast, the hyperparameters of the gradient boosting algorithm are first optimized using a Bayesian optimization search as shown in Fig. 2 (1). Secondly, a feature selection process using calculated feature importances is applied to improve the accuracy of the residual forecast further as described in section III-C (2). Finally, the residual forecast and the stationary profiles are summed to forecast the load profile of each cluster (3).

Fig. 2 also shows an example of the applied time series cross validation to every cluster for one week indicating the training set (yellow), the feature selection set (green) and the residual day-ahead load forecasts (blue). As shown, the first two day-ahead load forecasts of every week are used to calculate the average feature importances of these two days. Subsequently, the feature selection process is carried as described in section III-C. Based on the determined feature selection set, the day-ahead load profiles of the next five consecutive days in the week are forecasted.

C. FEATURE SELECTION

As explained, (gradient) boosting algorithms based on DT have the advantage over other ML algorithms to interpret the correlation between the applied features and forecasted load based on the calculated relative importance of the features.

Reference [41] explains that the relative importance I of feature j from a DT T is generally calculated according to:

$$I_j^2(T) = \sum_{t=1}^{J-1} i_t^2 1(v_t = j), \tag{18}$$

where t are the non-terminal nodes related to terminal node J, v_t is the splitting variable associated with feature j and i_t^2 is the squared improvement of the DT due to the split with feature J.

The importance of feature j of all M DT used for the gradient boosting algorithm is calculated according to:

$$I_j{}^2 = \frac{1}{M} \sum_{m=1}^{M} I_j^2(T_m), \tag{19}$$

which equals to the average importance of the calculated importance for each DT.

The initially applied features for each cluster are the same time-, weather- and PV-related features. However, the gradient boosting algorithm is trained and optimized separately for each cluster. As a consequence, the gradient boosting algorithm enables to identify relative (un)important features for each cluster depending on typical behaviours and load properties of customers connected to MV/LV transformers. Identifying features with a low importance can be used to further improve the forecast accuracy of each cluster, because including features with a low importance could lead to overfitting. Therefore, a feature selection process is applied for each cluster to improve the forecast of each cluster and thereby the forecast of the aggregated load.

The feature selection process starts by calculating the feature importances for each cluster. Subsequently, the algorithm forecasts the two days from the feature selection set after dropping the feature with the lowest calculated importance until the lowest forecast error is calculated and the optimal set of features for each cluster is determined. This optimal set of features is used to forecast the consecutive five days of the week. The feature selection process is repeated weekly to account for changing correlations between features and load over a longer period of time due to external factors such as changing weather conditions.

D. ERROR EVALUATION

The day-ahead forecast error of each cluster in the range from $k = 1, \ldots, K$ clusters and the related aggregated load are calculated according to (4) and (5) with $K = 1, \ldots, K_{max}$. To analyze if the results confirm that the stated inequalities of (9) and (10) are significant, the day-ahead load forecast is carried out during six consecutive weeks according to the time series cross validation shown in Fig. 2. The first two days of each week (the feature selection set) are not considered in this error evaluation of the day-ahead forecast accuracy. Thus, in total the day-ahead forecast accuracy of 30 days is analyzed. Subsequently, the day-ahead forecast accuracy without feature selection is compared with day-ahead forecast accuracy with feature selection. Finally, it is calculated if the average day-ahead forecast error of the aggregated load for $NRMSE_{K=1}$ and $MAPE_{K=1}$ is outside of the 95 % confidence interval of the calculated average day-ahead forecast error of the aggregated load for $NRMSE_{K_{opt}}$ and $MAPE_{K_{opt}}$.

E. DATASET

For this study, the load profiles of 519 MV/LV transformers from the city of Enschede are measured by the DSO every 15 minutes from the 17th of July until the 7th of November in 2020. First of all, these MV/LV transformers are clustered in the range from K $= 1, \ldots, K_{max}$. Subsequently, the applied initial training set is from the 17th of July until the 26th of September as shown in Fig. 2. To evaluate the day-ahead forecast error as described in section III-D and following the time series cross validation shown in Fig. 2, the period from the 27th of September until the 7th of November is split into six weeks [45], [46]. Fig. 2 indicates the data related to the time window of the first week. The minimum ratio between the training set and forecast is therefore 0.986/0.014.

The proposed clustering based day-ahead forecasting algorithm makes use of centralized ML approaches. When facing privacy issues regarding centralized storage and use of measured load profiles of MV/LV transformers, the clustering based day-ahead forecasting algorithm can also be studied with decentralized ML learning approaches. However, the study of these decentralized ML techniques are not considered in the scope of this research.

IV. RESULTS

A. CLUSTERING

The data processing discussed in Section III-A1 is applied to the dataset of the MV/LV transformers for the month of August. The clustering algorithms are computed in the range from $K = 2, \ldots, K_{max}$ clusters with $K_{max} = 10$ and each cluster k is named $C1, \ldots, C10$. The results of the validation metrics for all clustering algorithms are shown in Fig. 3. Overall, all algorithms show a consistent performance and the algorithms have a continuous decrease for the CDI and CHI indicating algorithm stability. However, from the information gain point of view, the MDI and CHI indexes show a decrease in performance for $K > 4$ clusters, indicating that there is no natural clustering with more clusters. The MDI and CHI are the most clear metrics to define the number of clusters. The adequate balance between cluster compactness and separation was found for $K = 3$ clusters.

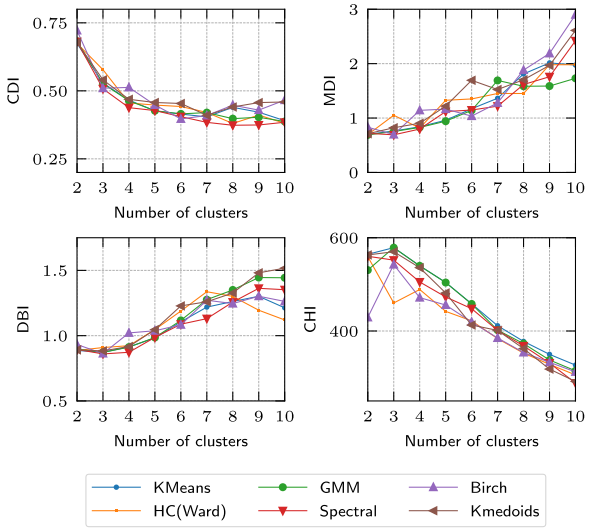

FIGURE 3. Calculated quality evaluation indices for six clustering algorithms in the range from $K = 2, \ldots, 10$ clusters. The minimum value of DBI and the maximum of CHI indicate that the optimal number of clusters is $K = 3$.

TABLE 1. Calculated clustering quality evaluation indices for $K = 3$ clusters. The best two performing algorithms are highlighted in bold. The best performing algorithm is underlined. The adjusted rand scores evaluates the consensus between the results of each algorithm referenced against the Spectral algorithm.

	DBI	MDI	CDI	CHI	Adj. Rand score
Spectral	**0.860**	**0.692**	**0.508**	551.99	1.000
BIRCH	**0.863**	**0.696**	**0.510**	542.40	0.930
GMM	0.872	0.753	0.527	**578.70**	0.858
KMeans	0.873	0.761	0.528	**578.74**	0.849
Kmedoids	0.883	0.821	0.540	570.02	0.759
HC(Ward)	0.909	1.051	0.578	460.11	0.297

The numerical values of the quality evaluation indices for $K = 3$ clusters are shown in Table 1. The best performing clustering algorithm is the spectral clustering and the second best is BIRCH. Both algorithms grouped the set or RLP similarly according to the adjusted rand score metric [47], which is a consensus metric to assess how similar are the clustering results between algorithms. The spectral clustering results are used for the next steps of the framework shown in Fig. 1.

The hierarchical visualization of the BIRCH results shows identified load profile characteristics related to $K = 1, \ldots, 5$ clusters, which is shown as dendogram in Fig. 4. The mean and 95% confidence interval computed from the RLPs are highlighted in black and red lines, respectively. Generally, when the number of groups increases, the margins of the 95% confidence interval shrinks, meaning a reduction in variance

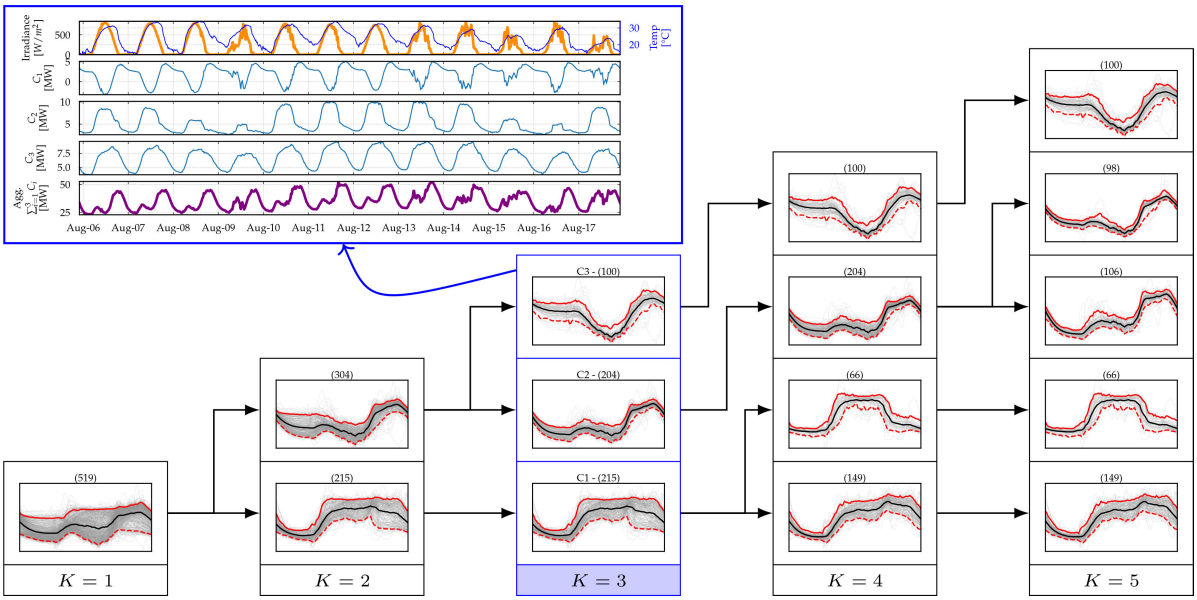

FIGURE 4. The hierarchical clustering results in the range from $K = 1, \ldots, 5$ clusters including the solar irradiance and temperature during 12 days of August.

within the clusters. A clear daily activity separation with defined characteristics is seen for $K = 3$ clusters. The clusters $C1$ and $C2$ for $K = 3$ clusters show residential load profiles. The cluster $C1$ has a pronounced *duck curve* which implies PV generation. Cluster $C3$ has a mixed activity behavior with a mixture of commercial and residential consumption with high activity at night, confirmed by the separation of this cluster when $K = 4$ clusters. The differences between the clusters start to be more evident for $K > 5$ clusters.

The lowest subplot in the upper left corner in Fig. 4 shows the aggregated load profile for $K = 3$ clusters during 12 days of August. The upper subplot shows the solar irradiance and the temperature during the same days, which corresponds to a heat wave period in the Netherlands. The subplots in the middle show the load profiles of the related clusters $C1$, $C2$ and $C3$. Cluster $C2$ shows that the shape of the load profile in day light hours is directly affected by the solar irradiance, confirming that PV generation is the cause of the change of the load profile. Additionally, the increase of temperature caused an increase of load at night times of 20% for cluster $C1$ and 11% for cluster $C2$, indicating that the temperature changed the load profile of each cluster differently.

B. FORECASTING

Based on the clustering results, the average day-ahead forecast accuracy of 30 days in the range from $K = 1, \ldots, 6$ clusters is calculated using (4) and (5) for each cluster in the range from $k = 1, \ldots, K$ and the aggregated load. An overview of the optimized hyperparameters using the Bayesian optimization search are shown in Table 2 in case of $K = 6$ clusters. The objective is *reg* : *squarederror* and the booster is *gbtree* for all clusters using *init_point* $= 25$ and *n_iter* $= 125$.

TABLE 2. The optimized hyperparameters found with the Bayesian optimization search, which are applied to the XGBoost algorithm to forecast the day-ahead residual profiles in case of K = 6 clusters.

Paramaters	C1	C2	C3	C4	C5	C6
max_depth	4	4	5	5	5	4
gamma	1.30	1.72	0.57	0.78	1.69	0.55
colsample_bytree	0.5	0.94	0.95	0.92	0.94	0.87
learning rate	0.233	0.013	0.005	0.160	0.009	0.004
subsample	0.46	0.63	0.44	0.41	0.48	0.49
n_estimators	524	635	618	515	529	802
min_child_weights	3.33	4.52	3.69	5.57	3.07	5.50
reg_alpha	0.28	1.26	1.15	2.01	1.43	0.94
reg_lambda	1.56	0.92	0.68	0.13	1.56	1.73

TABLE 3. The average calculation time in minutes per week of the applied clustering algorithm, Bayesian optimization search, forecasting algorithm and feature selection process.

	Time [Min.]
Clustering	1.5
Bayesian optimization/10 iter./cluster	4.7
Forecast gradient boosting/cluster	3.9
Feature selection/cluster	5.7

The forecasting algorithm shown in Fig. 2 is calculated on a Lenovo notebook with a Intel(R) Core (TM) i7-8750 CPU @ 2.20GHz processor. The average calculation time in minutes per week of the applied clustering algorithm, Bayesian optimization search, forecasting algorithm and feature selection process are shown in Table 3. The total calculation time of this proposed clustering based day-ahead forecasting algorithm is less than 24 hours and therefore applicable for online day-ahead applications.

TABLE 4. Calculated average NRMSE and MAPE for the day-ahead aggregated load forecast and each separate cluster in the range from $K = 1, \ldots, 6$ clusters of 30 days. The related average NRMSE and MAPE during the defined daily peak period between 07:00 and 22:00 are shown in italics.

Clusters agg.		C1		C2		C3		C4		C5		C6	
NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE
7.67	5.08	7.67	5.08										
9.03	*4.96*	*9.03*	*4.96*										
7.81	5.31	3.41	2.24	8.54	7.58								
9.58	*5.79*	*3.98*	*2.21*	*11.11*	*7.96*								
6.85	4.32	3.50	2.43	6.74	5.34	9.44	**173.40**						
8.76	*4.87*	*4.16*	*2.36*	*8.99*	***5.81***	*12.80*	***249.52***						
7.33	5.02	4.30	2.67	7.12	5.74	10.71	**64.50**	4.18	2.91				
9.30	*5.69*	*4.29*	*2.42*	*9.53*	*3.16*	*14.46*	***60.06***	*5.88*	*6.35*				
7.49	5.22	4.41	2.87	9.67	10.18	9.89	**54.18**	4.12	2.83	4.89	3.49		
9.11	*5.72*	*5.61*	*2.55*	*13.02*	*10.56*	*13.35*	***55.64***	*4.44*	*3.00*	*5.78*	*3.44*		
7.77	5.42	4.45	2.86	9.03	9.22	9.19	**64.02**	4.53	3.04	5.57	4.26	15.34	**24.42**
9.49	*6.00*	*6.08*	*2.52*	*9.03*	*9.77*	*12.25*	***79.68***	*4.40*	*3.21*	*6.75*	*4.37*	*22.77*	***29.70***

Table 4 provides an overview of all average errors of the day-ahead load forecasts expressed in NRMSE and MAPE. Table 4 also indicates the average errors of the day-ahead load forecast in italics during the period between 07:00 and 22:00 when the load is typically high. Table 4 indicates a minimum average error of the day-ahead aggregated load forecast for $K = 3$ clusters expressed in both calculated error metrics. The average error of the day-ahead aggregated load forecast increases when the number of clusters equals $K > 3$. This observation is supported when the calculated PCC is considered, which is shown in Table 5. The PCC is the closest to 1 for $K = 3$ clusters, while the PCC is lower when the number of clusters equals $K > 3$. All metrics therefore indicate that $K = 3$ clusters is the optimal number of clusters, which coincides with the results discussed in section IV-A.

Table 4 also indicates extremely large forecast errors expressed in MAPE for cluster $C3$ and $C6$. The load profiles of these two clusters drop to values close to zero due to PV generation during certain periods. As a consequence, (5) indicates that the calculated MAPE value increases, because the denominator is approaching to zero. Therefore, these MAPE values do not reflect the forecast error accurately, but these extremely large forecast errors are due to inherent properties of MAPE. If the forecast error of these same clusters is expressed in NRMSE, these extremely large forecast errors are not observed in Table 4. However, these extremely large forecast errors expressed in MAPE do confirm the discussion related to Fig. 4, where the profiles indicated that clusters $C3$ and $C6$ are clusters of households with a relative high penetration of PV generation. Due to this PV generation, the difference between PV generation and demand is small during certain periods leading to a load close to zero. As a consequence, a large forecast error is calculated when expressed in MAPE.

Fig. 5 shows the calculated NRMSE, MAPE and PCC from the day-ahead aggregated load forecast of the same 30 days with the proposed model for $K = 1$ (red) and the optimal number of $K = 3$ clusters (black). Fig. 5 also shows the

TABLE 5. Calculated average PCC related to the day-ahead aggregated load forecast and each separate cluster in the range from $K = 1, \ldots, 6$ clusters of 30 days. The related average PCC during the defined daily peak period between 07:00 and 22:00 is shown in italics.

Clusters Agg.	C1	C2	C3	C4	C5	C6
0.978	0.978					
0.962	*0.962*					
0.976	0.996	0.967				
0.957	*0.990*	*0.948*				
0.981	0.996	0.983	0.936			
0.963	*0.990*	*0.976*	*0.891*			
0.979	0.994	0.980	0.910	0.988		
0.961	*0.993*	*0.949*	*0.857*	*0.963*		
0.978	0.995	0.955	0.929	0.988	0.991	
0.961	*0.994*	*0.918*	*0.876*	*0.961*	*0.985*	
0.977	0.995	0.961	0.777	0.987	0.989	0.946
0.959	*0.994*	*0.923*	*0.774*	*0.946*	*0.981*	*0.887*

day-ahead aggregated load forecast with an Autoregressive Integrated Moving Average (ARIMA) model for comparison (blue), which is a traditional and widely applied model for time series forecasting [48], [49]. Fig. 5 indicates that the NRMSE and MAPE from the day-ahead aggregated load forecast is generally lower for the optimal number of $K = 3$ clusters than for $K = 1$. Correspondingly, the PCC is generally closer to one for the optimal number of $K = 3$ clusters than for $K = 1$. All metrics indicate a more accurate day-ahead aggregated load forecast for $K = 3$ clusters than for $K = 1$. All three metrics in Fig. 5 also indicate the improvement in accuracy of the day-ahead aggregated load forecast with the proposed model for $K = 1$ and the optimal number of $K = 3$ clusters compared with the day-ahead aggregated load forecast with the ARIMA model. However, Fig. 5 also indicates that the calculated values of all metrics variate rather strongly between these days. On average, the NRMSE and MAPE indicate a decreasing trend and the PCC indicates an increasing trend over time due to the increasing length of the training set for each consecutive day-ahead

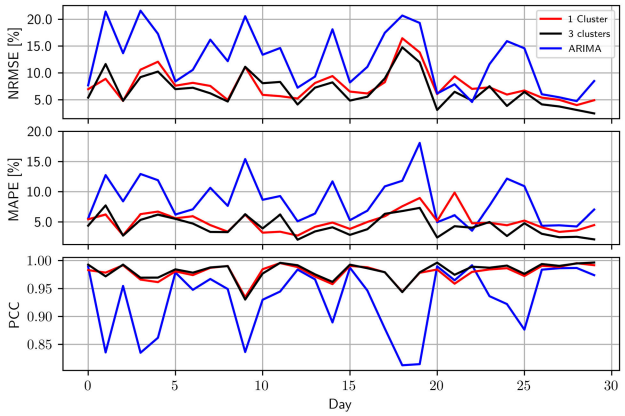

FIGURE 5. Calculated NRMSE, MAPE and PCC of the day-ahead aggregated load forecast of 30 days with the proposed model with $K = 3$ clusters, $K = 1$ (without clustering) and ARIMA model.

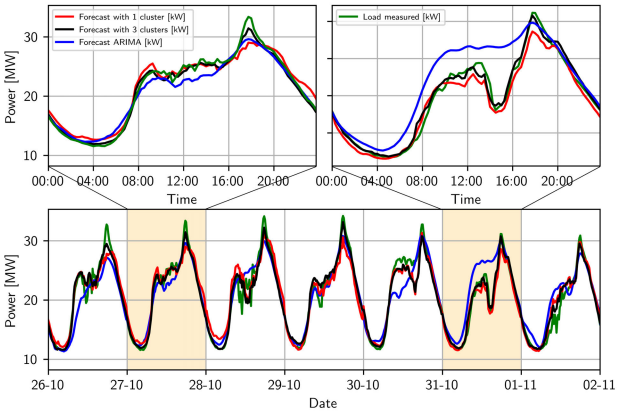

FIGURE 6. Day-ahead aggregated load forecasts from the 26th of October until the 1st of November with the proposed model with $K = 3$ clusters, $K = 1$ (without clustering) and ARIMA model.

aggregated load forecast. Additionally, the NRMSE and MAPE is generally larger and the PCC is generally lower on weekend days than on working days, because only two from the seven days are weekend days. Thus, the algorithm has less days in the training set with a similar pattern to train and forecast.

Fig. 6 shows seven consecutive day-ahead aggregated load forecasts with the proposed model for $K = 1$ (red), for the optimal number of clusters $K = 3$ (black) and with the traditional ARIMA model (blue) compared with the actual measured aggregated load profile (green). Both forecasts with $K = 1$ and $K = 3$ clusters follow the pattern of the measured aggregated load profile rather accurately, but importantly the day-ahead aggregated load forecast during the peak period is generally improved when the number of clusters equals $K = 3$ which is supported by Table 4 and Table 5. The forecast with the ARIMA model indicates that the peak load around 18:00 of the day-ahead aggregated load forecast is rather accurate on some days, but it has most difficulty with forecasting an accurate pattern between 08:00 and 20:00 when the load is relatively high.

To analyze the impact of feature selection on the day-ahead aggregated load forecast of the proposed model, Fig. 7 shows the average NRMSE and MAPE reduction of the day-ahead aggregated load forecast and the average PCC increase of the day-ahead aggregated load forecast due to feature selection in a range from $K = 1, \ldots, 6$ clusters. The impact of the feature selection is the largest when the number of clusters equals $K = 1$ followed by $K = 3$. On the one hand, Fig. 7 indicates that the impact of the feature selection is correlated with the performance of the clustering. On the other hand, Fig. 7 also indicates that the impact of the feature selection is depending on the initial forecast error without feature selection.

Fig. 8 shows the average relative NRMSE and MAPE reduction and average relative PCC increase of the day-ahead aggregated load forecast in a range from $K = 1, \ldots, 6$ clusters when compared with the average NRMSE, MAPE and

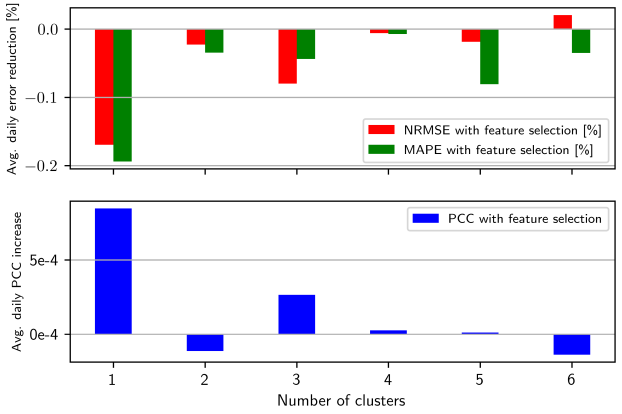

FIGURE 7. Average NRMSE and MAPE reduction and average PCC increase calculated from the average day-ahead aggregated load forecast due to feature selection in the range from $K = 1, \ldots, 6$ clusters over 30 days.

PCC of the day-ahead aggregated load forecast for $K = 1$ without feature selection. Based on the results of Table 4, the largest average relative NRMSE and MAPE reduction is expected for $K = 3$ clusters with feature selection, which is also confirmed by Fig. 8. This average NRMSE reduction is 12.7 % and this average MAPE reduction 18.2 %. Correspondingly, the largest average relative PCC increase is also expected for $K = 3$ clusters with feature selection based on Table 5, which is confirmed by Fig. 8 as well. This average PCC increase is 0.37 %.

Fig. 9 shows the average NRMSE, MAPE and PCC calculated from the day-ahead aggregated load forecast over the same 30 days in a range from $K = 1, \ldots, 6$ clusters including their 95 % confidence interval. On the one hand, Fig. 9 indicates that the average MAPE calculated from the day-ahead aggregated load forecast for $K = 1$ (5.08 %) is just outside the 95 % confidence interval of $K_{opt} = 3$ clusters (4.32 % $\pm$ 0.58 %). On the other hand, Fig. 9 also indicates that the average NRMSE calculated from the

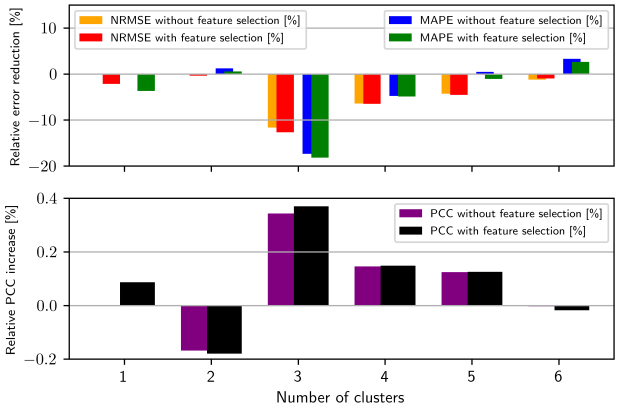

FIGURE 8. Average relative NRMSE and MAPE reduction and average relative PCC increase calculated from the day-ahead aggregated load forecast in the range from $K = 1, \ldots, 6$ clusters and the day-ahead aggregated load forecast for $K = 1$ clusters without feature selection.

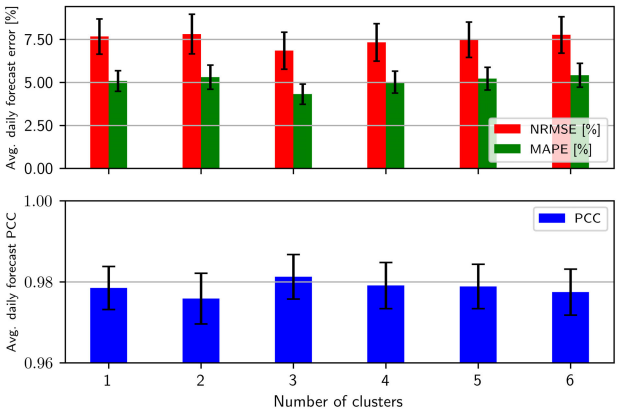

FIGURE 9. Average NRMSE, MAPE and PCC calculated from the day-ahead aggregated load forecast in the range from $K = 1, \ldots, 6$ clusters including their 95 % confidence interval over 30 days.

day-ahead aggregated load forecast for $K = 1$ (7.67 %) is just inside the 95 % confidence interval of $K_{opt} = 3$ clusters (6.85 % ± 1.07 %). Fig. 9 also indicates that the average PCC calculated from the day-ahead aggregated load forecast for $K = 1$ (0.979) is just inside the 95 % confidence intervals of $K_{opt} = 3$ clusters as well (0.981 ± 0.006). Based on this analysis, no statistically significant improvement can be observed yet.

Therefore, further analysis considers the confidence interval of the difference between the average NRMSE, MAPE and PCC for $K = 1$ and $K_{opt} = 3$ clusters. The improvement is statistically significant if zero is outside the 95 % confidence interval of the difference between the two load forecasts. The calculated confidence interval of the difference in NRMSE is -0.82 % ± 0.51 % and in MAPE is -0.77 % ± 0.50 %. The calculated confidence interval of the difference in PCC is 0.0027 ± 0.0018. Altogether, this enables to conclude that the improvement of the day-ahead aggregated load forecast expressed in (9), (10) and (11) are all statistically significant.

V. CONCLUSION

This paper proposes a new four-step approach to significantly improve clustering based day-ahead forecasting of aggregated MV/LV transformer loadings using a gradient boosting algorithm in combination with a feature selection process. First, all MV/LV transformer loadings are clustered in a variety of numbers based on similarity in patterns. Second, a gradient boosting algorithm is used to forecast the load profile of each cluster. Third, a feature selection process is applied to each separate cluster to improve the day-ahead forecast accuracy of each cluster. Finally, the day-ahead aggregated load forecast of all MV/LV transformers in the area, which represents the related MV network loading, is calculated by summation of all day-ahead forecasts of the clusters. The accuracy of the day-ahead MV network loading forecast for the optimal number of clusters is calculated to determine the improvement of the day-ahead MV network loading forecast compared without clustering.

First of all, the proposed model enables the DSO to efficiently identify typical load profiles of many MV/LV transformer loadings for network planning in a neighborhood or city. The clustered profiles can be applied to efficiently and specifically estimate the different impact of increasing energy demands and adoption of PV generation driven by the energy transition on different types of MV/LV transformer loadings in distribution networks. Secondly, the day-ahead forecast of the proposed model also enables to be used for network operation applications of MV networks, such as flexibility, PV curtailment -and storage scheduling.

The paper studied the proposed methodology by clustering load profiles from 519 MV/LV transformers in a range from $K = 1, \ldots, 10$ clusters using six different clustering algorithms. The performance of the six algorithms was evaluated with four indicators, which all indicated $K_{opt} = 3$ clusters. The clusters in the range from $K = 1, \ldots, 6$ were forecasted after the forecast of each cluster was optimized with the described feature selection process. The accuracy of the day-ahead aggregated load forecast was evaluated using the NMRSE, MAPE and PCC. In accordance with the clustering results, these forecasts indicated $K_{opt} = 3$ clusters as well. Based on 30 day-ahead forecasts, the calculated average and 95 % confidence interval of the NRMSE, MAPE and PCC indicated that the day-ahead aggregated load forecast for $K_{opt} = 3$ clusters improved statistically significant compared with the day-ahead aggregated load forecast for $K = 1$ (without clustering).

REFERENCES

[1] R. Bernards, J. Morren, and H. Slootweg, "Development and implementation of statistical models for estimating diversified adoption of energy transition technologies," *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, pp. 1540–1554, Oct. 2018, doi: 10.1109/TSTE.2018.2794579.

[2] M. Salazar, I. Dukovska, P. H. Nguyen, R. Bernards, and H. J. G. Slootweg, "Data driven framework for load profile generation in medium voltage networks via transfer learning," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Eur.)*, Oct. 2020, pp. 909–913, doi: 10.1109/ISGT-Europe47291.2020.9248753.

[3] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015, doi: 10.1109/TSG.2014.2364233.

[4] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Zhang, "A data-driven game-theoretic approach for Behind-the-Meter PV generation disaggregation," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 3133–3144, Jul. 2020, doi: 10.1109/TPWRS.2020.2966732.

[5] R. Bernards, "Smart planning: Integration of statistical and stochastic methods in distribution network planning," Ph.D. dissertation, Dept. Elect. Eng., Eindhoven Univ. Tech., Eindhoven, The Netherlands, 2018.

[6] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, 2016, doi: 10.1016/j.ijforecast.2015.11.011.

[7] G. Rouwhorst, P. Nguyen, and H. Slootweg, "A hybrid supervised learning model for a medium-term MV/LV transformer loading forecast with an increasing capacity of PV panels," in *Proc. IEEE Madrid PowerTech*, Jun. 2021, pp. 1–6, doi: 10.1109/PowerTech46648.2021.9494854.

[8] N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers, and W. van Alst, "Deep learning versus traditional machine learning methods for aggregated energy demand prediction," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Eur. (ISGT-Eur.)*, Sep. 2017, pp. 1–6, doi: 10.1109/ISGTEurope.2017.8260289.

[9] E. Mocanu, "Machine learning applied to smart grids," Ph.D. dissertation, Dept. Elect. Eng., Eindhoven Univ. Tech., Eindhoven, The Netherlands, 2017.

[10] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy forecasting: A review and outlook," *IEEE Open Access J. Power Energy*, vol. 7, pp. 376–388, 2020, doi: 10.1109/oajpe.2020.3029979.

[11] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019, doi: 10.1109/TSG.2018.2818167.

[12] E. Mocanu, P. H. Nguyen, and M. Gibescu, "Deep learning for power system data analysis," in *Big Data Application in Power Systems*. Amsterdam, The Netherlands: Elsevier, 2018, ch. 7, pp. 125–158, doi: 10.1016/B978-0-12-811968-6.00007-3.

[13] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016, doi: 10.1109/TSG.2016.2548565.

[14] M. Sun, I. Konstantelos, and G. Strbac, "C-vine copula mixture model for clustering of residential electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2382–2393, May 2017, doi: 10.1109/TPWRS.2016.2614366.

[15] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007, doi: 10.1109/TPWRS.2007.901287.

[16] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006, doi: 10.1109/TPWRS.2006.873122.

[17] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012, doi: 10.1016/j.energy.2011.12.031.

[18] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005, doi: 10.1109/TPWRS.2005.846234.

[19] G. A. Susto, A. Cenedese, and M. Terzi, "Time-series classification methods: Review and applications to power systems data," in *Big Data Application in Power Systems*. Amsterdam, The Netherlands: Elsevier, 2018, ch. 9, pp. 179–220, doi: 10.1016/B978-0-12-811968-6.00009-7.

[20] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.

[21] X. Wang, W. J. Lee, H. Huang, R. L. Szabados, D. Y. Wang, and P. V. Olinda, "Factors that impact the accuracy of clustering-based load forecasting," *IEEE Trans. Ind. Appl.*, vol. 52, no. 5, pp. 3625–3630, Sep. 2016, doi: 10.1109/TIA.2016.2558563.

[22] P. Laurinec and M. Lucká, "Clustering-based forecasting method for individual consumers electricity load using time series representations," *Open Comput. Sci.*, vol. 8, no. 1, pp. 38–50, Jul. 2018, doi: 10.1515/comp-2018-0006.

[23] A. Cini, S. Lukovic, and C. Alippi, "Cluster-based aggregate load forecasting with deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207503.

[24] I. P. Panapakidis, "Clustering based day-ahead and hour-ahead bus load forecasting models," *Int. J. Elect. Power Energy Syst.*, vol. 80, pp. 171–178, Sep. 2016, doi: 10.1016/j.ijepes.2016.01.035.

[25] C. Zhang and R. Li, "A novel closed-loop clustering algorithm for hierarchical load forecasting," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 432–441, Jan. 2021, doi: 10.1109/TSG.2020.3015000.

[26] Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia, "An ensemble forecasting method for the aggregated load with subprofiles," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3906–3908, Jul. 2018, doi: 10.1109/TSG.2018.2807985.

[27] H. H. H. Aly, "A proposed intelligent short-term load forecasting hybrid models of ANN, WNN and KF based on clustering techniques for smart grid," *Electr. Power Syst. Res.*, vol. 182, May 2020, Art. no. 106191, doi: 10.1016/j.epsr.2019.106191.

[28] Y. Yang, W. Hong, and S. Li, "Deep ensemble learning based probabilistic load forecasting in smart grids," *Energy*, vol. 189, Dec. 2019, Art. no. 116324, doi: 10.1016/j.energy.2019.116324.

[29] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Optimized clusters for disaggregated elecricity load forecasting," *RESTAT Stat. J.*, vol. 8, no. 2, pp. 105–124, Nov. 2010.

[30] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000, doi: 10.1109/34.868688.

[31] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, pp. 103–114, Jun. 1996, doi: 10.1145/235968.233324.

[32] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963, doi: 10.1080/01621459.1963.10500845.

[33] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annu. Rev. Statist. Appl.*, vol. 6, pp. 355–378, Jan. 2019, doi: 10.1146/annurev-statistics-031017-100325.

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[35] M. J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[36] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Sci. Technol.*, vol. 20, no. 2, pp. 117–129, Apr. 2015, doi: 10.1109/TST.2015.7085625.

[37] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974, doi: 10.1080/01969727408546059.

[38] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[39] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.*, vol. 3, no. 1, pp. 1–27, Sep. 1974, doi: 10.1080/03610927408827101.

[40] R. Fonteijn, T. Castelijns, M. Grond, P. Nguyen, J. Morren, and H. Slootweg, "Short-term load forecasting on MV/LV transformer level," in *Proc. 25th Int. Conf. Exhib. Electr. Distrib.*, Madrid, Spain, 2019, pp. 1–5, doi: 10.34890/135.

[41] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[42] H. I. Erdal, "Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1689–1697, Aug. 2013, doi: 10.1016/j.engappai.2013.03.014.

[43] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.*, vol. 77, no. 4, pp. 802–813, Jul. 2008, doi: 10.1111/j.1365-2656.2008.01390.x.

[44] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[45] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and review," *Int. J. forecasting*, vol. 16, no. 4, pp. 437–450, Oct. 2000, doi: 10.1016/S0169-2070(00)00065-0.
[46] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*.
[47] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971, doi: 10.1080/01621459.1971.10482356.
[48] D. S. de O. Santos, J. F. L. de Oliveira, and P. S. G. de Mattos Neto, "An intelligent hybridization of ARIMA with machine learning models for time series forecasting," *Knowl.-Based Syst.*, vol. 175, pp. 72–86, Jul. 2019, doi: 10.1016/j.knosys.2019.03.011.
[49] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1394–1401, doi: 10.1109/ICMLA.2018.00227.

GEORGE ROUWHORST (Member, IEEE) was born in Nieuwegein, The Netherlands, in 1995. He received the B.S. degree in science and chemistry from Radboud University, Nijmegen, The Netherlands, in 2017, and the M.S. degree in sustainable energy technology from the Eindhoven University of Technology (TU/e), Eindhoven, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Electrical Energy Systems (EES) Group.

His current research interest includes forecasting the impact of the energy transition on the load profiles in the distribution network for planning applications.

EDGAR MAURICIO SALAZAR DUQUE (Member, IEEE) received the B.S. degree in electrical and electronic engineering from the Universidad de Los Andes, Bogotá, Colombia, in 2008, and the M.S. degree *(cum laude)* in smart electrical grids and systems from Kungliga Tekniska Högskolan (KTH), Stockholm, Sweden, and the Eindhoven University of Technology (TU/e), The Netherlands, in 2018. He is currently pursuing the Ph.D. degree with the Electrical Energy Systems (EES) Group, TU/e.

His research interest includes data analysis and applications of machine learning techniques on power distribution grids for planning and operation.

PHUONG H. NGUYEN (Member, IEEE) received the Ph.D. degree from the Eindhoven University of Technology (TU/e), The Netherlands, in 2010.

During his one-year sabbatical leave in 2019, he took up a group leader position at the Sustainable Energy Systems (SES) Group, Luxembourg Institute of Science and Technology (LIST). Since January 2020, he has been back to TU/e as an Associate Professor with the Electrical Energy System (EES) Group. He has committed his research effort to realize synergies of advanced monitoring and control functions for the distribution networks along with emerging digital technologies. This distinctive combination of competences allows him to develop a research pathway crossing over various domains of mathematical programming, stochastics, data mining, and communication networks. His research interests include data analytics with deep learning, real-time system awareness using (IoT) data integrity, and predictive and corrective grid control functions.

HAN SLOOTWEG (Senior Member, IEEE) received the M.S. degree *(cum laude)* in electrical power engineering and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 1998 and 2003, respectively.

He is currently the Director of the Asset Management Department, Enexis Netbeheer B.V., 's-Hertogenbosch, The Netherlands, one of the largest Distribution Network Operators of the Netherlands. Its spearheads are the strategic goals of Enexis: accelerating the transition towards a more sustainable energy supply and excellent, state of the art network operation. He also holds a professorship at the Smart Grids at the Electrical Energy Systems (EES) Group, Eindhoven University of Technology (TU/e). He has coauthored more than 200 articles, covering a broad range of various aspects of electrical power systems.