A hybrid supervised learning model for a medium-term MV/LV transformer loading forecast with an increasing capacity of PV panels

A hybrid supervised learning model for a medium-term MV/LV transformer loading forecast with an increasing capacity of PV panels

George Rouwhorst*, Phuong Nguyen*, Han Slootweg*†
*Department of Electrical Engineering
Eindhoven University of Technology, Eindhoven, the Netherlands
Email: g.d.g.rouwhorst@tue.nl
†Asset Management
Enexis Netbeheer, 's-Hertogenbosch, the Netherlands

***Abstract*—The share of photovoltaic (PV) generation has increased quickly in the last decade. Many PV panels are connected behind-the-meter (BTM), so that they can not be identified with measurement equipment at MV/LV transformers. This poses a challenge for a medium-term MV/LV transformer loading forecast if the capacity of PV panels is increasing over time. Therefore, this paper proposes a hybrid approach for a medium-term load forecast (MTLF) of a MV/LV transformer with an increasing capacity of PV panels that are not separately measured. This approach combines a supervised learning model (data-driven approach) with a model to estimate the generation profile of the PV panels (model-based approach). The results indicate that the accuracy of the forecast improves significantly, while an accurate generation profile of the PV panels connected BTM or a disaggregation of the net load is unnecessary.**

***Index Terms*—behind-the-meter PV generation, distribution network, medium-term, net load forecasting, supervised machine learning**

I. Introduction

Photovoltaic (PV) generation has been increasing its share in electricity generation in the past decade rapidly due to the current energy transition. Many rooftop PV panels are connected behind-the-meter (BTM) to the low-voltage (LV) network unlike conventional centralized generators connected to the transmission network with an individual, dedicated grid connection and meter. This can have a large impact on the distribution network, because the electricity flows become more volatile and negative congestions could occur if a surplus amount of electricity is generated [1]-[2].

In order to strengthen monitor and control functionalities to cope with the explained issues, Distribution System Operators (DSOs) are installing measurement equipment at MV/LV transformers. However, the electricity generation by the PV panels BTM $P(t)$ and the load profile $L(t)$ can not be measured separately with measurement equipment installed at the MV/LV transformer. The DSO is only able to measure the net load over the transformer $L_{net}(t)$ at a given time t according to

$$L_{net}(t) = L(t) + P(t) \tag{1}$$

in [kW]. Consequently, the DSO is unable to monitor the impact of the PV panels on the transformer loading directly. This can cause problems for the DSO regarding the accuracy of load forecasts, which the DSO increasingly needs for a variety of applications, such as network reinforcement decisions for long-term planning, day-ahead scheduling and (real-time) integration of flexibility services [3]-[4].

Load forecasts can be classified based on their forecast horizon. Short-term load forecasting (STLF) is extensively studied for applications regarding network operation. The time horizon of STLF varies typically up to a week-ahead with a varying time resolution of the forecasted profile of an hour. Two approaches are commonly found in literature for this type of forecast to cope with the issue of being unable to separately measure consumption and production by PV panels connected BTM [5]-[6]. The first approach aims to disaggregate the generation profile $P(t)$ from the measured net load $L_{net}(t)$ using a model-based approach. This requires a rather complex method to calculate the unknown generation profile by the PV panels with accuracy, because the generation profile depends on local meteorological information and accurate knowledge of the physical properties of the PV panels [3],[7]-[8]. The second approach aims to disaggregate the net load by using other measurement data such as smart meters using a data-driven approach. However, this approach requires additional data sources, which are not always readily available. In both approaches, the generation profile and load profile are forecasted separately and subsequently added to forecast the net load after disaggregation. Other studies have also proposed alternative STLF methods by forecasting the generation profile of the BTM PV panels from the net load directly without disaggregating the net load first [9]-[10].

Long-term load forecasting (LTLF) is extensively studied for network planning applications. The time horizon is longer than a year and aims on forecasting the peak loads, which is the main driver to determine the required capacity of the network and the necessity of grid reinforcement and extension.

NWO ESI-bida project: 'Using small data and big data: Neighbourhood Energy & Data Management Integration System'

Therefore, forecasting the shape and pattern of the load profile is usually not considered in LTLF, but mainly different scenarios are studied driving the change in peak load. Examples of such scenarios involve changing policies and economic growth [11]. However, traditional LTLF is based on historic patterns and therefore unable to include the adoption of new technologies driven by the energy transition, such as PV generation [6], [12]. For this reason, [11] describes a LTLF approach to cope with this issue for distribution network planning.

Medium-term load forecasts (MTLF) are studied for operational planning applications. The time horizon varies generally from a week up to a year-ahead with usually an hourly time resolution [6]. To improve the utilization of the current network capacity over a period exceeding that of STLF, MTLF includes not only the forecasted peak load, but also the load profile to study applications such as the deployment of flexibility, PV curtailment and energy storage. LTLF are not applicable for this study, because the impact of these applications relies on the typical shape and frequency of the load profile. On the other hand, a forecasting period up to a week-ahead with STLF is not sufficient to capture the impact of long-term patterns, such as seasonality of the weather, holidays and the adoption and implementation of new technologies. Thus, there is a gap to forecast load profiles and corresponding peak loads for time horizons exceeding that of STLF, but shorter than those of LTLF. This study proposes an approach for MTLF taking into account the combined impact of load and PV generation connected BTM, which has not been well addressed in the literature so far by enriching existing STLF methods with long-term variables related to PV generation.

One of the extensively studied methods for STLF are supervised machine learning models. These data-driven models learn the correlation between features, such as time – and weather-related information, and load during a defined training period [13]-[14]. Subsequently, the model forecasts the load using the features of the period to be forecasted. If the forecast period is longer, a longer training period is generally required for an accurate forecast as well. As discussed, the capacity of BTM PV panels is rapidly increasing. As a consequence, this can lead to large errors of the MTLF, because the correlation between the load and the features changes during the training period, which is not considered in STLF as the variation on the short term tends to be rather limited.

The main contribution of this paper is to propose a general approach for a month-ahead forecasting (MTLF) for a MV/LV transformer loading directly at which the capacity of BTM PV panels is increasing over time. By combining a data-driven approach with a model-based approach, a hybrid approach is proposed to estimate the generation profile of the BTM PV panels without the requirement of disaggregating the net load or additional measurements as described.

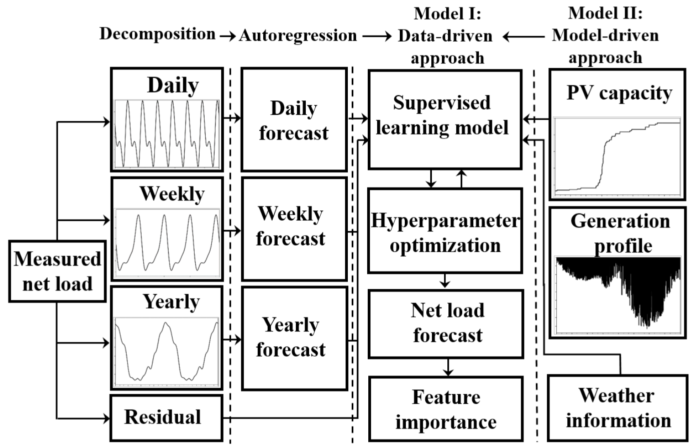

Figure 1: Flowchart of the proposed methodology to forecast the net load of a MV/LV transformer.

The remainder of this paper is organized as follows. Section II presents the proposed methodology of the model and section III describes the related implementation and evaluation of the model. Section IV presents and evaluates the results. Finally, in section V the main conclusions of the paper are drawn.

II. METHODOLOGY

A. Framework

This paper proposes a hybrid model consisting of a data-driven approach using a supervised learning model together with a model-based approach to estimate the generation profile at a MV/LV transformer where the capacity of BTM PV panels is increasing. The model-based approach is used as an extra feature for the supervised learning model. This enables the supervised learning model to learn the correlation between the increasing capacity of BTM PV panels and the net transformer loading without the need of disaggregating this transformer loading into generation and load first. Fig. 1 shows a flowchart of the proposed methodology. The first step is to decompose the measured transformer loading into stationary profiles (daily, weekly and yearly profiles) and non-stationary profiles (residual profile). These stationary profiles are forecasted using autoregression, while the non-stationary profile of the net load is forecasted using the supervised learning model. Subsequently, the electricity generation by the BTM PV panels during the period to be forecasted is estimated based on the model-based approach. This estimation is used with weather- and time-related features, to train the supervised learning model and predict the residual profile, which is aggregated with the stationary profiles to forecast the transformer loading.

B. Time series decomposition

Reference [4] describes the improvement in accuracy of the forecast if the measured net load signal is decomposed first. Therefore the transformer loading is decomposed into daily, weekly and yearly profiles, which are all stationary, while a non-stationary profile, the residual, is left [15].

C. Generation profile estimation

To estimate the generation profile of the BTM PV panels, the estimated power output of the sum of all BTM PV panels installed at the MV/LV transformer is calculated as follows

$$P \approx C \cdot \frac{I_{PV,t}}{10000} \cdot [1\text{-}\mu \cdot (T_{PV,t} - 25)]\,, \tag{2}$$

where C represents the capacity of the sum of all BTM PV panels connected to the MV/LV transformer [kWp], $I_{PV,t}$ the incoming radiation on the PV panels [W/m^2], μ the constant temperature parameter and $T_{PV,t}$ the temperature of the PV panel [℃] [3]. $T_{PV,t}$ is calculated as follows

$$T_{PV,\,t} \approx\ T_{A,t} + \frac{I_{PV,\,t}}{800} \cdot (NOCT\text{-}20)\,, \tag{3}$$

where $T_{A,t}$ represents the ambient air temperature [℃] and $NOCT$ the nominal operating cell temperature [℃] [3].

D. Net load forecasting

To forecast the transformer loading, the stationary profiles and non-stationary profiles are forecasted. The stationary profiles are forecasted using an autoregression model, which is a technique that uses a linear combination of past values to forecast the value ahead [16]. Based on the performance of the forecast accuracy as described in [4], a gradient boosting algorithm is used as supervised learning model to forecast the non-stationary residual profile [17]. To improve the accuracy of the forecast, a Bayesian optimization search is performed first to optimize the hyperparameters of this model. Finally, the MTLF is calculated by aggregating all separately forecasted profiles.

III. Implementation and Validation of the Supervised Learning Model

For this paper, a MV/LV transformer loading is measured during a period of 2.5 years with an hourly time resolution. The measurements from January 2018 until December 2019 are used as training set for the supervised learning model. The months from January 2020 until May 2020 are used to evaluate the forecast accuracy of the supervised learning model without and with the model-based approach. To estimate the generation profile of the BTM PV panels with this model-based approach, a database with all registered residential PV installations at the same MV/LV transformer is used, which includes the installation date and installed capacity.

A. Feature preprocessing

To forecast the month-ahead residual profiles, the following three types of features are implemented to train the model:

1. Time-related features, which include the hour of the day, the month of the year and whether it is a working day, weekend or a holiday including the difference of every variable with the previous hour.
2. Weather-related features measured at the nearest weather station, which are the temperature [℃], global irradiation [W/m^2], sunshine [*min./hour*] and rain duration [*min./hour*] every hour including the difference of every feature with the previous hour [18]. These used features are from 2008 until 2017, because it is assumed that the actual weather features of 2020 are unavailable for the period ahead.
3. PV-related features, which is the registered capacity of the BTM PV panels [kWp] and their estimated generation using the model-based approach [kW].
4. Stationary profiles as explained in section II.B, including the difference of every profile with the previous hour.

To improve the accuracy of the forecast, the features are preprocessed before they are applied to train the model. All time-related features are modelled as dummy features, because their values represent a qualitative value and not a quantitative value. The other features are scaled to a standardized scale according to (4)

$$Feature(t)' = \frac{Feature\ (t)\ \text{-}\ \mu}{\sigma}\,, \tag{4}$$

where $Feature(t)'$ represents the standardized value, $Feature(t)$ the unstandardized value, μ the median and σ the standard deviation [4].

B. Error evaluation

The accuracy of the forecast a month-ahead is analyzed by performing an out-of-sample time series cross-validation starting from January 2020 until May 2020. The minimum ratio between the training set and forecast is therefore 0.96/0.04 for the month of January, which is increasing every subsequent month. Fig. 2 provides a schematic representation of the time series cross-validation carried out [19].

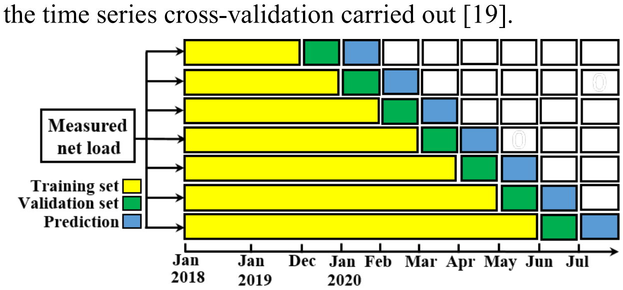

Figure 2: schematic representation of time series cross validation.

To analyze the accuracy of the forecast of every month, the related normalized root-mean-square-error (NRMSE) is calculated every hour according to (5)

$$NRMSE\ [\%] = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (L_{net,\,i} - L_{net,i}\,')^2}}{L_{net,\,max} - L_{net.min}} \cdot 100\,, \tag{5}$$

where n represents the number of forecasted timestamps, $L'_{net,i}$ represents the forecast of the net load at a timestamp [kW], $L'_{net,i}$ the measured net load at the same timestamp [kW] and $L_{net,max}$ and $L_{net,min}$ represent the maximum and minimum measured net load within the period of forecast [kW]. The average NRMSE is calculated for every month [20].

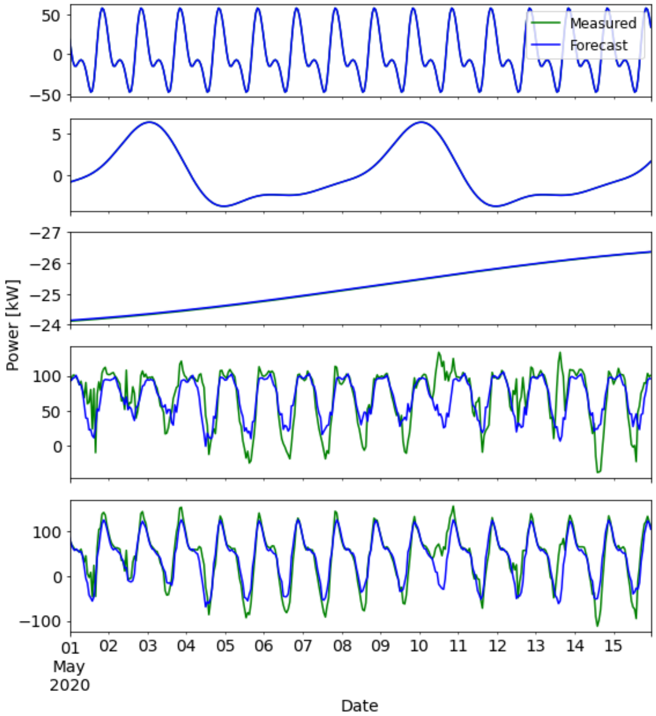

Figure 3: The decomposed profiles (green) and their related forecast (blue) during the first 15 days of May 2020.

C. Feature importance

The performance of the supervised learning model is not only evaluated on the calculated NRMSE, as explained in the previous section. After the NRMSE of the forecast is calculated for the supervised learning model with and without the model-based approach, the performance is also evaluated on the feature importance. This analysis indicates the relative importance of all the features applied with the supervised learning model. If the error of the forecast is reduced due to the additional features of the model-based approach, evaluation of the feature importance should indicate the relative importance of these features by the supervised learning model.

IV. Results and Discussion

A. Net load forecasts

Fig. 3 shows the decomposed profiles during the first 15 days of May (green) together with their forecast (blue). The forecast of the stationary profiles are constant, while the shape of the residual forecast follows the measured profiles rather accurately. The load forecast of the residual profile is the average of the forecast using the weather-related featured from 2008 until 2017. The lowest graph of Fig. 3 shows the aggregated forecast during the same period.

Fig. 4 shows the measured transformer loading (green) with the average forecast of the transformer loading for all months using weather-related features from 2008 – 2017 in case when the PV-related features are excluded (red) and included (blue). Fig. 4 shows on the left graph that if the electricity generation is low (January and February), the difference between the forecasts is small. The error compared with the measurement is generally small as well, except during some days in February when the measured generation is increasing. Fig. 4 shows on the right graph that the accuracy improves significantly if the PV-related features are included when the electricity generation by the PV panels is increasing. The remaining error of the generation peaks can predominantly be explained due to the use of weather-related features from previous years instead of the actual weather-related features of 2020. The NRMSE of the peak loads is also increasing during the lockdown in the Netherlands caused by the COVID-19 pandemic from the second half of March until the end of May. Due to this lockdown, a higher peak load was measured during this period compared with previous years. As a consequence, the forecast by the model trained with the measurement of previous years will be lower as indicated in Fig. 4. After the lockdown has ended, the measured peak load reduced from June and the forecast of the peak load error reduced as well.

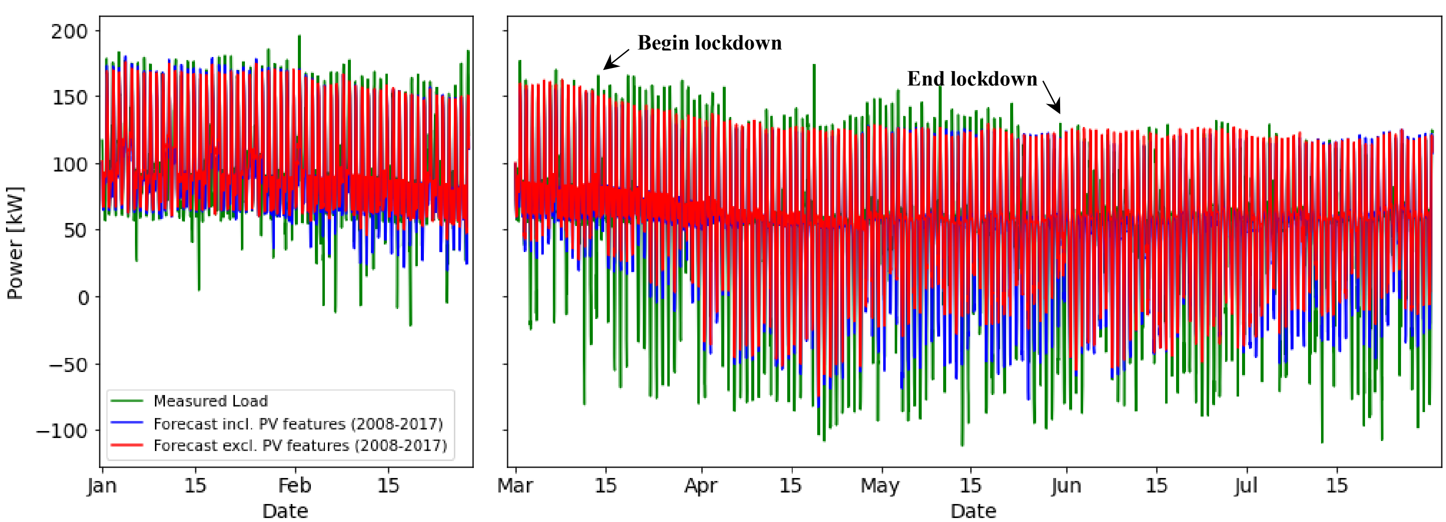

Figure 4: The measured net load (green) with the forecast of the net load incl. (blue) and excl. (red) the PV related features and weather-related features from 2008 – 2017.

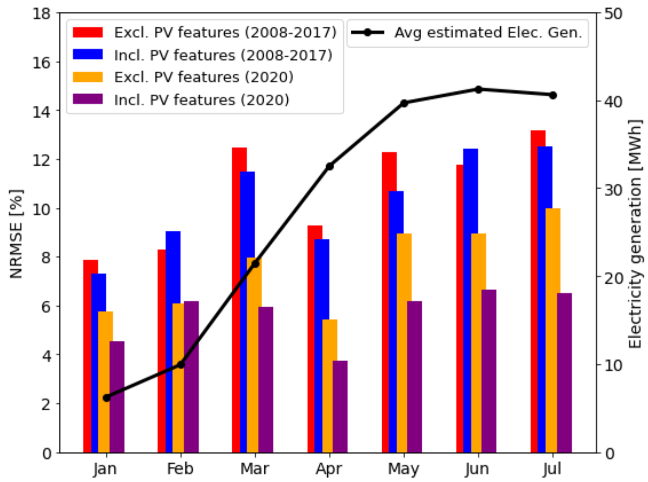

Figure 5: Average monthly NRMSE excluding and including PV-related features with weather-related features from 2008 – 2017 (red and blue) and with weather-related features from 2020 (orange and purple) together with the average estimated generation per month of the BTM PV panels (orange).

B. Model evaluation

Fig. 5 shows the calculated NRMSE according to (5) to analyze the forecasts using weather-related features from 2020 in case when the PV-related features are excluded (orange) and included (purple). Fig. 5 also shows the calculated NRMSE of the forecasts in combination with the weather-related features from 2008-2017 when the PV-related features are excluded (red) and included (blue). The NRMSE of the forecast is significantly reduced when the actual weather-related features from 2020 are used. This supports the discussion that the accuracy of the forecast is limited due to the use of weather-related features from previous years. However, this is from a practical perspective an unavoidable cause of error. The error of the forecast is also generally increasing if the estimated average electricity generation per month is increasing (black).

As also indicated by Fig. 4, the difference in calculated NRMSE shown in Fig. 5 is relatively small for the months of January and February independent from the used weather-related features. If the average estimated electricity generation is increasing, the difference between the forecast which includes and excludes the PV-related features is increasing, especially when the weather-related features from 2020 are considered. This difference is smaller when the weather-related features from 2008-2017 are used. For the month of June, the error of the forecast without the PV-related features is even slightly lower. This is mainly due to the variation in weather over years on individual days.

The lower graphs Fig. 6 shows that including the PV-related features with the weather-related features from 2020 enables to improve the forecast during peak generation. In general, the upper graph of Fig. 6 indicates that this also holds when comparing the forecasts with the weather-related features from 2008-2017. However, if the weather-related features from 2008-2017 are used, too much PV generation is forecasted on some specific days. For example, the measurements indicate that the electricity generation was significantly lower on the 4th, 5th and 14th of June, most likely due to cloudy days with a low amount of sunshine. As a consequence, a larger NRMSE is calculated according to (5) for the forecast with a higher peak generation. Nevertheless, the general shape and frequency of the profile in combination with the duration and size of the forecasted peak load have improved in general due to the PV-related features over the whole month of June, which is of more value than the exact timing on each specific day. In other words, interchanging the forecast of the 4th of June with the forecast of the 12th of June would reduce the calculated NRMSE of June, but would not be a relevant improvement for the MTLF considering the general goal and application of MTLF explained in section I.

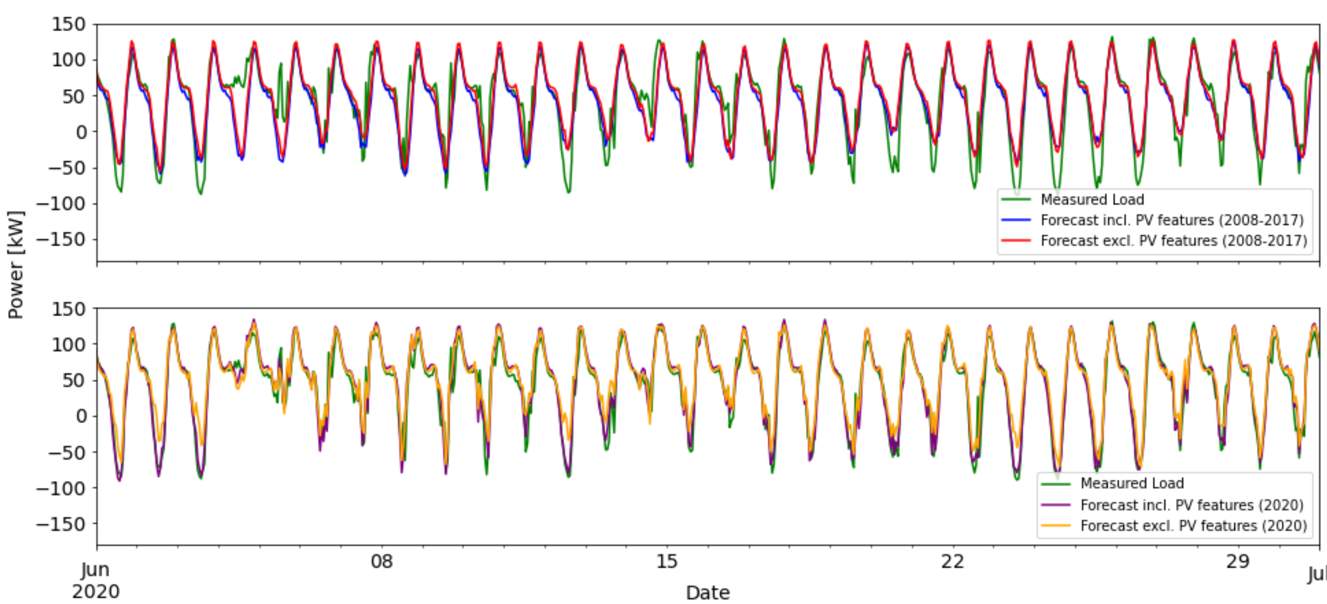

Figure 6: The measured net load (green) with the forecast of the net load incl. (blue) and excl. (red) the PV related features during the month of June.

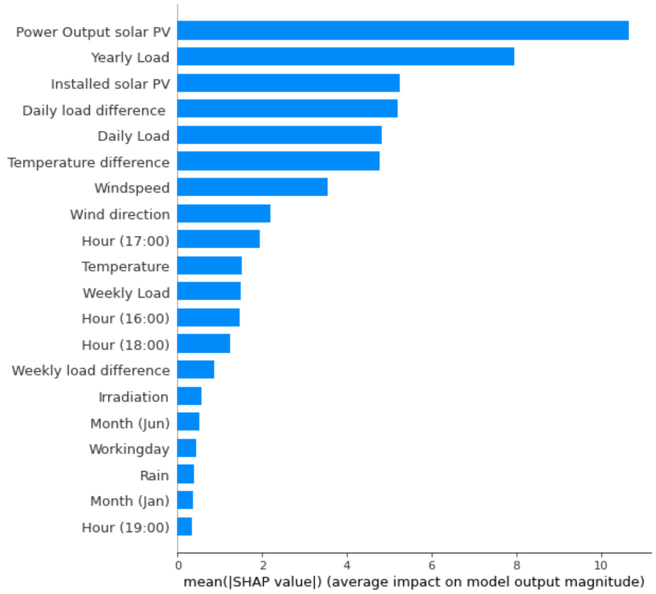

Figure 7: Overview of the 20 features with the largest impact on the accuracy of the forecast.

C. Feature importance

To gain more insight into the outcome of both forecast accuracies, Fig. 7 shows the 20 features with the largest impact on the forecast by the supervised learning model [21]. Fig. 7 supports the impact of the PV-related features on the accuracy of the forecast as noticed in Fig, 4 and Fig. 5. Many of the other important features are weather-related features which was also noticed by the improvement in accuracy shown in Fig. 5, when the weather-related features of 2020 are applied. Additionally, Fig. 7 also indicates that time-related features related to the usual periods of peak demand and the decomposed stationary profiles have a relative large impact on the accuracy forecast.

V. Conclusion

This paper proposes a combination of a data-driven approach and a model-based approach for a MTLF of a MV/LV transformer, while the capacity of BTM PV panels connected behind the transformer is increasing. The performance of the model is evaluated based on the accuracy of the forecast, which is compared with the forecast if the model-based approach is excluded. Additionally, the importance of the features from the model is evaluated.

The applied data-driven approach involved a supervised learning model, which is trained with time-related features, weather-related features and the PV-related features. The latter is estimated using the model-based approach. If the PV-related features are included, the relative improvement in accuracy is increasing if the average estimated electricity generation by the BTM PV panels is increasing. Analysis of the feature importance also indicated the importance of the PV-related features for the forecast. Therefore, this approach enables a general method to improve the MTLF of a MV/LV transformer at which the amount of BTM PV panels is increasing on an hourly time resolution.

References

[1] F. Wang, K. Li, X. Wang, L. Jiang, J. Ren, and Z. Mi, "A distributed PV system capacity estimation approach based on support vector machine with customer net load curve features," *Energies*, vol. 11, no. 7, pp. 1–19, 2018.

[2] D. W. van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renew. Sustain. Energy Rev.*, vol. 81, no. December 2016, pp. 1484–1512, 2018.

[3] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-Driven Probabilistic Net Load Forecasting With High Penetration of Behind-the-Meter PV," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3255–3264, 2018.

[4] R. Fonteijn, T. Castelijns, and J. Morren, "Short-term load forecasting on MV / LV transformer level," in *CIRED 2019*, 2019, pp. 1–5.

[5] 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, 2020.

[6] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecast.*, vol. 32, no. 3, pp. 914–938, 2016.

[7] A. Kaur, L. Nonnenmacher, and C. F. M. Coimbra, "Net load forecasting for high renewable energy penetration grids," *Energy*, vol. 114, pp. 1073–1084, 2016.

[8] V. P. Menon, S. Lokhande, and Y. K. Bichpuriya, "Correcting forecast of utility's demand with increasing solar PV penetration," *2017 IEEE Innov. Smart Grid Technol. - Asia Smart Grid Smart Community (ISGT-Asia 2017)*, pp. 1–6, 2017.

[9] H. Shaker, H. Zareipour, and D. Wood, "Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2456–2465, 2016.

[10] H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2466–2476, 2016.

[11] R. Bernards, "Smart planning: integration of statistical and stochastic methods in distribution network planning," Ph.D. disstertation, Dept. of El. Eng., Technische Universiteit Eindhoven, 2018.

[12] K. B. Lindberg, P. Seljom, H. Madsen, D. Fischer, and M. Korpås, "Long-term electricity load forecasting: Current and future trends," *Util. Policy*, vol. 58, no. May, pp. 102–119, 2019.

[13] N. G. Paterakis, E. Mocanu, and M. Gibescu, "Deep Learning Versus Traditional Machine Learning Methods for Aggregated Energy Demand Prediction," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2017, pp. 1-6.

[14] E. Mocanu, P. H. Nguyen, and M. Gibescu, "Deep Learning for Power System Data Analysis," in *Big Data Application in Power Systems*, R. Arghandeh and Y. Zhou, Eds. Elsevier, 2018, pp. 125–158.

[15] S. J. Taylor and B. Letham, "Forecasting at Scale," *PeerJ Prepr. 5e3190v2*, vol. 35, no. 8, pp. 48–90, 2017.

[16] "statsmodels.tsa.ar_model.AutoReg — statsmodels." [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.ar_model.AutoReg.html. [Accessed: 08-Oct-2020].

[17] T. Chen, "XGBoost : A Scalable Tree Boosting System," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference*, 2016, pp. 785–794.

[18] "KNMI - Uurgegevens van het weer in Nederland - Download." [Online]. Available: http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi. [Accessed: 08-Oct-2020].

[19] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and review," *Int. J. Forecast.*, vol. 16, no. 4, pp. 437–450, 2000.

[20] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," pp. 1–49, 2018.

[21] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, and B. Nair, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.