2019 3rd International Conference on Smart Grid and Smart Cities (ICSGSC)

A Hybrid Approach of Solar Power Forecasting Using Machine Learning

Arpit Bajpai
fortiss GmbH
Munich, Germany
e-mail: bajpai@fortiss.org

Markus Duchon
fortiss GmbH
Munich, Germany
e-mail: duchon@fortiss.org

Abstract—**Photovoltaic (PV) power generation is prone to fluctuations and it is affected by different weather conditions. Therefore, accurate forecasting becomes vital for grid operators to manage grid operations. In this study, multiple machine learning models are used, and different weather parameters are analyzed for PV forecasting. The performance of different models is compared. The proposed forecasting models are tested with 4 kWp PV device installed at our institute, fortiss GmbH and the comparison of the models is done through error analysis. The accuracy is evaluated using historical weather data. The study proposes a novel approach of PV forecasting where a model is built based on the principle of pipelining clustering, classification and regression algorithms. Clustering and classification algorithms group the data-points with similar weather conditions. Subsequently, segmented regression is applied to these groups. Weather forecast of the next day is utilized to determine the models used for forecasting. The results show that the proposed forecasting model for PV systems is effective and promising.**

Keywords—**Machine Learning, Photovoltaic, Forecasting**

I. Introduction

The increasing electrification of the world due to industrialization is resulting in an exponential increase in demands with every passing day. This leads us to the question about how to satisfy the requirement with an uninterrupted supply. Renewable/green energies are used to replace fossil fuels to reduce carbon emission. One of the most feasible options of energy generation from renewable sources is solar energy. Solar energy is abundant and that makes it the most practical type of energy. Due to a decrease in the cost of PV modules when compared to increasing costs of energy generation from fossil fuels, it is more practical to use renewable energy resources such as solar energy. This would make PV power generation one of the significant power resource. The prediction of the amount of energy generated by renewable energy sources is useful in terms of satisfying the energy demand and supply planning to avoid critical conditions in these systems. PV power generation forecasting can help in reducing the impact of system output uncertainty on the grid. Therefore assists in making the system more reliable and maintaining the power quality. Hence to make use of PV power more efficiently, forecasting becomes vital [1].

Energy generated by PV system is directly proportional to weather conditions such as cloud cover, solar intensity, site-specific conditions etc. However, due to fluctuating solar radiation, the output from PV generators is sporadic which makes it difficult to manage the power dispatch. The limitation is that the output of the PV cannot be planned as it is dependent on various external factors.

This study forecasts the PV power generation by the device deployed at our institute. The result of the study will be verified and incorporated with the Smart microgrid (SMG) project for providing the energy management infrastructure at the Smart Energy Living Lab [2], that relies on measured data and on forecasting algorithms to predict the future patterns of both local energy generation and power loads [3]. We propose a new algorithm to improve the prediction accuracy.

The goal of the study is to use different machine learning techniques to generate models that can forecast the energy generation for thin-film PV cells. This study is based on the PV power generation data obtained from Solarlog device installed at fortiss for the years from 2013 to 2016. We obtained the weather data from Meteorological Institute Munich [4] and also used the weather data purchased from Meteoblue [4].

The paper is structured as follows. In the next section we provide a survey on the related work focusing on prediction methods. The third section provides information on the data basis, as well as on the initial analysis. Section four describes the approach and methods applied including a performance evaluation. Based on these results we propose a hybrid model utilizing a clustering technique to identify similar conditions regardless of the season. Finally, last section summarizes the results and and shows current limitations.

II. Related Work

PV power generation forecasting can be classified into two approaches which are based on the use of solar insolation. The two methods are known as direct and indirect methods. In direct method of forecasting, the system output of the PV cell is forecasted directly based on the input parameters. In the indirect method, the input weather parameters obtained from numerical weather prediction are used to predict to solar intensity, which is further used to forecast the PV system output. Numerical weather prediction (NWP) method uses differential equations and simulates the atmosphere to make weather predictions. The indirect method of forecasting is used in [5], [6] while the authors in [7], [8] and [1] utilized the direct method of forecasting.

978-1-7281-3848-0/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSGSC.2019.00-10

There are a variety of methods being applied in the field of solar power forecasting. Most of these methods differ based on the time-horizon being forecasted and also on the above mentioned classifications.

A paper by Jenesius et al. [9] made use of Model Output Statistics (MOS) technique to forecast daily solar radiation. In [10], the authors proposed different approaches of forecasting solar radiation depending on different time scales. The image data from Meteosat satellites was used for prediction of cloud motion and very short-term forecasting was done in a temporal range of 30 minutes to 6 hours. The authors used the data from NWP for day-ahead-forecasting.

In [11] the researchers studied both physical models and statistical models and presented case studies conducted in China. The authors deduced that the forecasting performance is impacted by the seasons and the forecasting error directly depends on the accuracy of weather prediction information obtained from NWP

In [12] the author described physical, statistical, Artificial Intelligence (AI) based and Hybrid models. They discussed the advantages and disadvantages of these prediction methodologies. The authors also threw light on different prediction horizons that are helpful in decision-making activities for the management of smart grids.

Sharma Pranshu et. al [5] used Support Vector machines (SVM) to predict solar irradiance. They used different regression techniques to do the forecasting. SVM with multiple kernels outperformed linear least squared regression model. The results showed that SVM based prediction model with seven input weather parameters obtained from National Weather Service was 27% more accurate than other models researched by the authors. The authors also defined the correlation between different weather parameters

The use of neural networks to predict insolation by computer simulations was proposed by Yona et al. [6]. This paper employed the indirect way of one-day-ahead forecasting. Radial basis function neural network (RBFNN) and recurrent neural network (RNN) were used instead of feed forward neural network. RBFNN was preferred for its simplicity and RNN for its accuracy for time-series data forecasting. The contradistinction was shown in [7] where the researchers used feed-forward artificial neural networks with 14 input weather variables and claimed that it performed better than recursive neural networks.

It is difficult to predict power output of the PV cells due to weather irregularities. PV system output is dependent on solar insolation and fluctuates along with solar intensity which is randomly affected based on the weather condition and geographical location. In addition to solar intensity, PV output also depends on the temperature of the PV cell. Temperature has a negative correlation with the output, i.e. higher temperature will reduce the conversion efficiency [1]. However, in this study we ignore the effect of cell temperature on system output forecasting.

Much research has been done in the direction of prediction of solar radiation and the algorithms have improved the prediction precision. However, researchers are still exploring the field of PV power output forecasting. Since the short-term prediction of sunshine intensity is somewhat accurate and can be obtained by sophisticated weather stations. Therefore, we can assume that the weather forecast is reliable and focus on predicting the output of the PV system. Also, most of the studies discussed here focus on building one unified model rather than having multiple models for prediction.

III. Dataset

We collected the historical weather data from the weather station managed by Meteorology department of Ludwig Maximilian University (LMU) located at Theresienstr. 37, Munich (48.147992N 11.573496E) [13]. We also obtained the high quality 30 year historical weather data for Munich from Meteoblue [4]. It offers high precision weather data in hourly resolution from the weather station located at Alfons-Goppel-Str., Munich (48.14N 11.58E).

The weather data includes multiple weather metrics for the year from 2013 to 2016 observed in 1 minute interval. Different weather parameters were analyzed. The weather metrics include wind temperature, wind speed, wind direction, precipitation, solar radiation and air pressure. Along with the weather metrics, we include other input features that are described below.

a) Data Description:

- Day
- Month
- Year
- Hour
- Minute
- Wind Temperature at 2m height (Celcius)
- Wind Speed at 30m height $(m/second)$
- Wind Direction at 30m height
- Precipitation (mm/minute)
- Solar Radiation (W/m^2)
- Air Pressure at sea level (hPa)
- Relative humidity at 2m above ground (%)
- Snowfall amount (cm/m^2)
- Total cloud cover (%)

The PV output data was obtained from small PV system of 4 KW capacity installed at fortiss, Guerickestr. 25, Munich (48.174924N 11.596140E) [14].

The data includes system output for the period from 8 February 2013 to 10 October 2016. The readings are divided in two parts, generated power and feed-in power in watts. Along with the measured power readings, the cell temperature is also recorded. The frequency of observed readings is random varying from 6 readings per minute to 1 reading in 5 minutes interval. We performed imputation for missing values in the data by defining a threshold on the percentage of values being present in the time horizon used for forecasting.

During the analysis phase, We found out that there were a lot of observations when the power generation was zero despite substantial solar insolation. This could be explained with snow deposit in the PV system. According to the report [15], it was found that the losses due to snowfall are dependent on the angle and technology of the PV system that occur with naturally accumulated snow. These samples should not be considered while building the model and should be treated as outliers. We removed outliers during our data analysis phase by defining the threshold on irradiance and checking the snowfall amount for those samples. We remove the observations from our model when a) the irradiance is greater than 180 $watt/m^2$ b) the system output is still zero c) there is a substantial amount of snowfall being observed in the last few days.

After analyzing the input features and studying the correlation, the insignificant weather parameters were removed from the input dataset. The final set of features were fed as an input to the model. The target variable is solar power. In addition to the weather data, we include the specific day of the year and time of the day as metrics.

After sensitivity analysis, the parameters that showed positive correlation were solar irradiation and temperature. The parameters that showed negative correlation were humidity, precipitation amount, cloud cover and snowfall amount. The parameters that showed low correlation were hour, day of year, air pressure, wind speed and wind direction. This signifies that power output can be high or low irrespective of aforementioned parameters. However, this does not guarantee that these parameters are not important. As we can see that time of the day (hour) and day of year are very important for forecasting which explains the high non-linear correlation of these parameters with PV system output. Therefore, we cannot rely solely on linear correlations to decide the input parameters for model building.

These complex relationships between different weather parameters and PV system output is the main motivation behind applying machine learning algorithms to automate intricate mathematical calculations required for predictions.

IV. Approach

In this section, we will discuss the methodology in detail. This section is divided into three parts. (a) The initial assumptions considered to solve the problem (b) the data preprocessing phase and (c) the forecasting model is discussed in detail.

a) Assumptions: The weather station is located at Theresienstrae 37 and the observed PV system is located at Guerickestr. 25. The distance between the two locations is around 4 km. Our results are based on the assumption that the weather conditions are similar, if not same, in both the locations without much deviation in the measured weather data.

b) Data Preprocessing: We first cleaned the data by removing the samples for the time when solar log was off and there were no readings from the PV system. We performed imputations to handle the missing values. We applied Simpson's rule to perform the numerical approximation to measure the amount of energy generated for the specified time interval i.e. 15 minutes and 1 hour in our case. We further cross checked our approximation with the readings measured by the solar log device. We then performed down sampling to aggregate the weather data into 15 minutes interval. We divided the dataset into two sets with the data from the year of 2013 until 2015 as training set and the data from 2016 is used as the test set.

Figure 1 shows the relationship between solar insolation with time where x-axis represents the time in hours and the y-axis represents the irradiance amount. We only consider the irradiance between 5 AM to 8 PM while training the model. From the figure, it is shown that solar irradiance increases with time and is maximum around midday and starts decreasing during the evening.

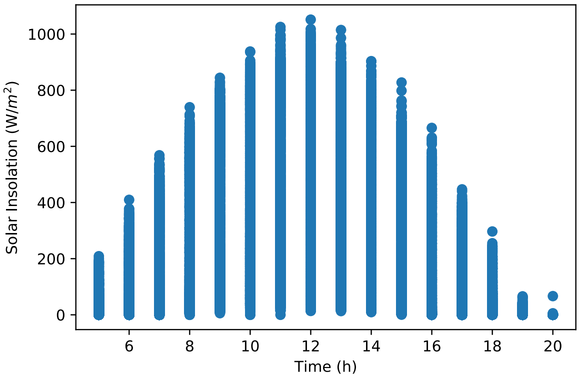

Fig. 1: Solar insolation with time

c) Forecasting: We applied different machine learning techniques to predict the PV power generation output of the system. The model focuses on doing one-day ahead forecasting with the frequency of both 15 minutes and 1 hour interval. Support vector regression (SVR) with radial basis function as kernel and random forest (RF) algorithms are used for building the models. Due to the non-linearity of the dataset, we used the models described above, as linear models will not be suitable. We used the features discussed in last section and then applied feature scaling to normalize the input data. The accuracy of SVR varies based on the selected kernel function and other parameters. We used Grid search method to find the optimal parameters.

We evaluated the performance of the models by computing Root-Mean-Square Error (RMSE) and the R squared values on test set. We fine tuned the hyperparameters for models and finally choose the models giving minimum root mean squared error and maximum R squared values. We then came up with the best model that predicts the PV system output using various weather factors explained in last section.

We obtained the best results with support vector regression and random forests on the dataset and these models were later used to forecast the PV system output for the year of 2016.

The predicted output varies from 0-1000 Watt hours in this scenario as we are considering the predictions for 15 minute interval instead of 1 hour interval. For one hour interval, PV system output range is between 0 to 4000 Watt hours.

We then used the test data to evaluate these models. The SVR model has the RMSE of 178.23, while the random forest model outperformed the SVR model with RMSE of 120.74.

Figure 2 and Figure 3 show the predicted vs actual scatter plot for both the models. It can be observed that for the random forest model, the points are close to the regressed diagonal line while for SVR model, they are more dispersed.

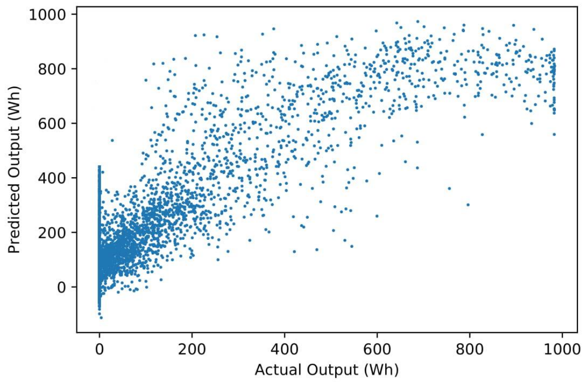

Fig. 2: Predicted vs actual photovoltaic system power output scatter plot for SVR

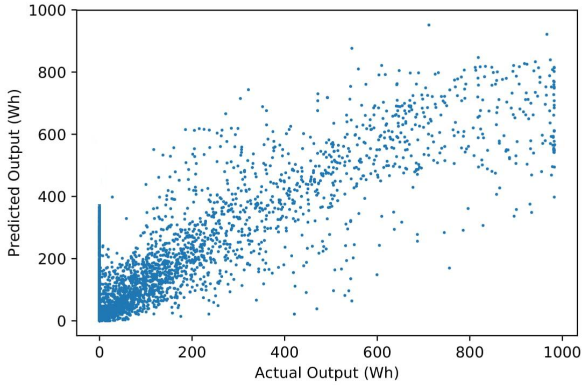

Fig. 3: Predicted vs actual photovoltaic system power output scatter plot for random forest

Figure 4 shows the performance of the model for 10 continuous days using the random forest model. It has to be noted that the days has been considered of 15 hours here as the energy generation is not significant during the early morning and late evening hours.

V. Hybrid Model

After the analysis and based on the evaluation of above models, it was observed that the prediction model performs best for sunny weather and not so good for cloudy/rainy weather. Therefore, we tried a new approach to overcome this problem. We propose an hybrid or a mixed approach where clustering, classification and regression techniques are combined to form a forecasting model. Based on the weather forecast of the next day, the model (with closest weather condition) is selected to forecast the power output using cluster-wise regression. The idea is to first use classification for the samples that have to be tested to find out the cluster (similar weather conditions) they belong to. Then use the specific regression model for the samples.

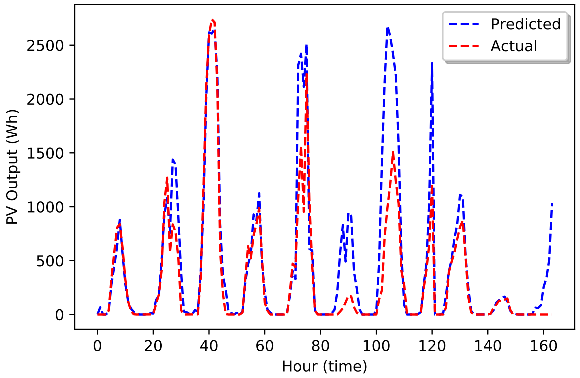

Fig. 4: Predicted vs actual photovoltaic system power output

The flowchart of the hybrid model is shown in the figure 5. There are placeholders for different models and they don't necessarily have to be the algorithms discussed in this section. The various steps of the algorithm are:

a) Clustering: Clustering technique is used to group similar days in a cluster as there were no labels present for the dataset. We used unsupervised learning because the points have no external classification.The idea is to group samples with similar weather conditions, which does not necessarily have a seasonal correlation.

K-means algorithm is used for clustering. The first step is to choose the number of clusters. Within-cluster sums of squares (WCSS) metric is used to find the number of clusters. The aim is to use the optimal value of WCSS. We used the elbow method to find the significant drop in WCSS and subsequently find the optimal number of clusters. We started with 3 clusters. However, due to an insufficient number of data samples in one of the clusters, we decided to use only two clusters. These are labelled as Cluster 1 and Cluster 2. To deal with random initialization trap, we used K-Means ++ algorithm.

b) Classification: A classifier is trained using the cluster labels as a target variable. The aim is to match new days with already existing clusters. For classification task, k-nearest neighbors (k-NN) algorithm is used. In k-NN classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors [16]. The idea is to classify the test data points and find the cluster they belong to.

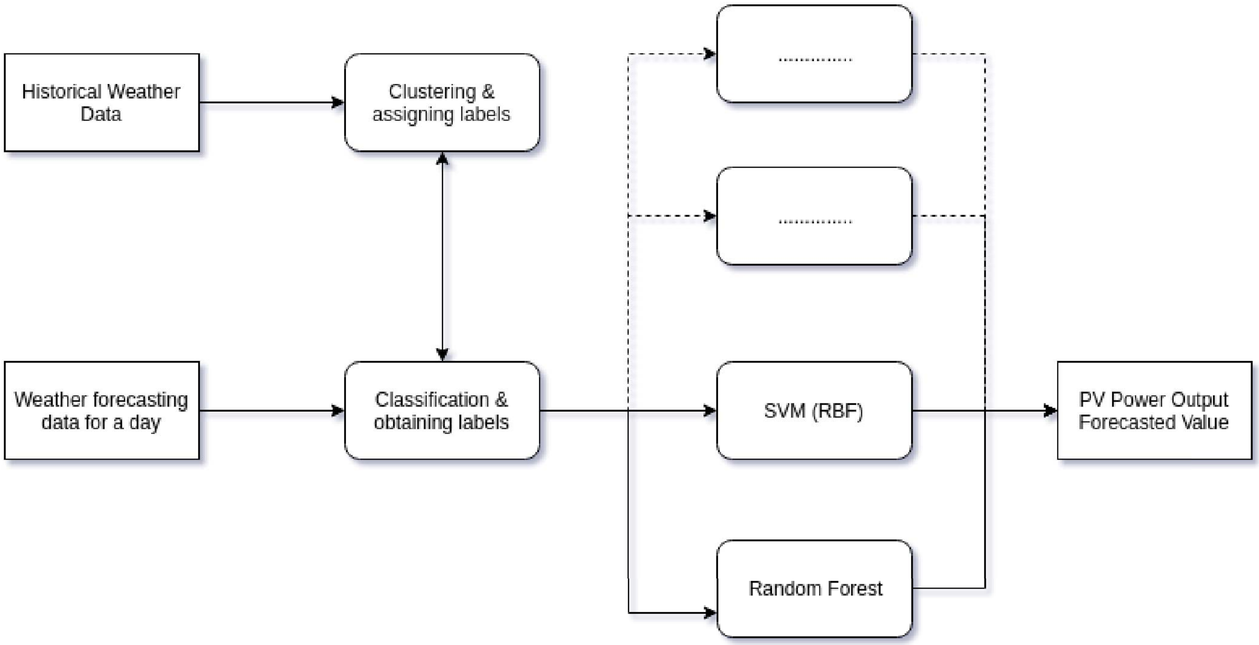
Historical Weather Data
Clustering & assigning labels
....
....
Weather forecasting data for a day
Classification & obtaining labels
SVM (RBF)
PV Power Output Forecasted Value
Random Forest

Fig. 5: Flowchart of hybrid approach

c) Regression: Finally, we created and trained different regression models for each cluster. We checked both random forest (RF) and support vector regression (SVR) models for each cluster. Subsequently, we tested the hybrid model on our test set, and the results are shown in table I.

Cluster 1 showed good accuracy with RF model while for cluster 2, SVR showed better results. The RMSE was reduced from 103.85 to 92.90 for cluster 2 when we used SVR instead of RF model. For the sake of visualization, we only plot the results for cluster 2. Figure 6 shows how the prediction error goes down when we consider SVR model instead of RF model in case of cluster 2 for the test data. There is a significant improvement in the prediction accuracy when we use SVR model instead of RF model.

The overall error is reduced from 120.74 to 109.48 with the hybrid approach instead of just using a single model for forecasting. The consolidated results are shown in table I and it can be seen that the hybrid approach further reduced the forecasting error by around 9%.

Model	RMSE
SVR (RBF)	178.23
Random Forest	120.74
Hybrid Approach	109.48

TABLE I: Summary of forecasting errors

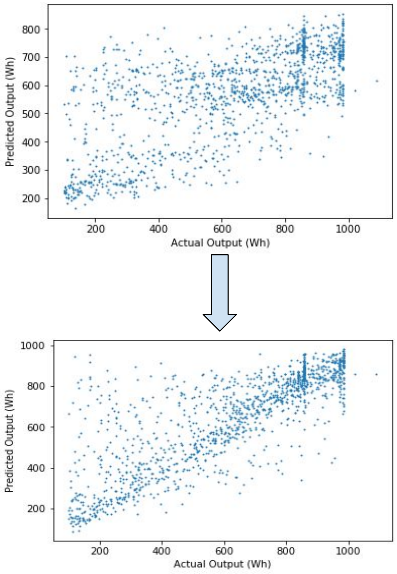

Fig. 6: Predicted vs Actual graph for Cluster 2. The upper one is random forest and the lower one is SVR

VI. Conclusion

The paper is aimed at forecasting the power generation by PV cells using machine learning techniques. We found out that random forest model outperformed other models used before for PV system output forecasting.

The study emphasized on analyzing different weather metrics. It was shown that it is complex to determine the relationship between the weather parameters and PV output.

The study found that the forecasting accuracy depends on the weather and the model performs better for sunny days as compared to the cloudy/rainy days in general, therefore a hybrid model is proposed to solve this problem.

With the hybrid model, we proposed a novel approach by combining different models according to specific weather conditions improving the overall quality of the prediction. The hybrid model performs better than the universal random forest regression model and this can be improved further with the refinement of specialized models.

The limitation of hybrid approach is that it takes much more time and effort to train and tune that many specialized models as compared to training the universal one.

The PV system forecasting accuracy depends on the weather forecasts. The more accurate weather forecasts we use, the better will be the predictions.

The distance between the weather station and the PV device can affect the prediction accuracy, especially, during frequent weather variations.

In future, the clusters can be inspected and analyzed to show some light on the observed behavior. As a next step, the proposed forecasting model will be tested in other cites with different weather conditions. We plan to conduct the experiment in Oulu, Finland to test how the model performs.

Acknowledgment

The authors would like to thank Mr. Markus Garhammer and his team from the meterological department at Ludwig-Maximilians-University Munich for providing the detailed weather data. Futhermore, this work is carried out as part of the iE-Services project (FKZ: 01IS16033D), an ongoing joint research work supported by the Federal Ministry of Education and Research.

References

[1] J. Shi, W.-j. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines," vol. 48, no. 3, pp. 1064–1069, 2012.

[2] "Smart living lab." https://www.fortiss.org/forschung/projekte/smart_energy_living_lab/. Accessed: 2019-03-15.

[3] C. Rottondi, M. Duchon, D. Koss, A. Palamarciuc, A. Pití, G. Verticale, and B. Schätz, "An energy management service for the smart office," *Energies*, vol. 8, no. 10, pp. 11667–11684, 2015.

[4] "Historical weather data." https://www.meteoblue.com/en/weather/forecast/week/munich_germany_2867714. Accessed: 2017-12-25.

[5] N. S. Sharma Pranshu; Irwin, David; Shenoy, Prashant, "Predicting Solar Generation from Weather Forecasts Using Machine Learning," *SmartGridComm*, pp. 528–533, 2011.

[6] A. Yona, T. Senjyu, T. Funabshi, and H. Sekine, "Application of Neural Network to 24-hours-Ahead Generating Power Forecasting for PV System," *IEEJ Transactions on Power and Energy*, vol. 128, no. 1, pp. 33–39, 2008.

[7] M. Abuella and B. Chowdhury, "Solar Power Forecasting Using Artificial Neural Networks," *2015 North American Power Symposium (NAPS)*, no. October, pp. 1–5, 2015.

[8] C. Paper, N. Carolina, and N. Carolina, "Solar Power Forecasting Using Support," no. October 2016, 2017.

[9] J. C. Jenesius, M. CRONE, and C. KOCH, "The Still Elusive T Cell Receptor," *Scandinavian Journal of Immunology*, vol. 14, no. 6, pp. 693–704, 1981.

[10] D. Heinemann, E. Lorenz, and M. Girodo, "Forecasting of solar radiation," *Solar Energy Resource Management for Elitricity Generation from Local Level to Global Scale*, no. chapter 2, pp. 83–94, 2006.

[11] Y. Huang, J.Lu, C.Liu, X.Xu, W.Wang, and X.Zhou, "Comparative study of power forecasting methods for PV stations," pp. 1–6, 2010.

[12] C. Wan, J. Zhao, S. Member, and Y. Song, "Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management," *Journal of Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2015.

[13] "Historical weather data." https://www.en.meteo.physik.uni-muenchen.de/. Accessed: 2017-12-25.

[14] "fortiss gmbh." https://www.fortiss.org/home/. Accessed: 2019-03-15.

[15] L. Powers, J. Newmiller, and T. Townsend, "Measuring and modeling the effect of snow on photovoltaic system performance," in *2010 35th IEEE Photovoltaic Specialists Conference*, pp. 000973–000978, June 2010.

[16] T. N. Phyu, "Survey of Classification Techniques in Data Mining," *International MultiConference of Engineers and Computer Scientists*, vol. I, pp. 18–20, 2009.