Abstract—Integration of renewable energy resources (RES) in the power system is increasing. However, the variability of PV output poses several challenges in maintaining reliable grid operation. This variability is a function of several meteorological variables. An accurate PV forecasting can tackle this issue in maintaining and scheduling stable grid operation. In this paper, we conduct a study to analyze the impact of using optimum combination (feature selection and extraction) of meteorological features and low dimensional subspace (dimensionality reduction) on the forecasting accuracy. We also assess and compare the output of the forecasting model when it is fed with all the input features present in the dataset with the case when we use low subspace of the dataset as an input to the model.

I. INTRODUCTION

The transition from high-carbon energy production to green and sustainable renewable energy production is gaining momentum. This transition involves integration of RES in the power grid as a power generation source. Due to this transition, the installed PV generation capacity is expected to increase by more than 4 TW by 2025 and 21.9 TW by 2050 [1]. However, the integration of RES in the power grid brings several challenges. One of the key challenges is its high variability, which is typically a function of many factors, for example location, weather, time and other physical characteristics. Secondly, this leads to steep ramps in the difference between power generation and demand.

Estimating and forecasting PV output power has the potential to tackle these issues by enabling energy balancing as well as maintaining and scheduling reliable grid operation [2]. Moreover, this technique also helps in reducing the integration and operation cost.

In this work, we compare the output of the forecasting model when preprocessed and condensed data is fed to the model with the case when the model is fed with raw or original data. We assess the impact of using data mining techniques which focuses on discovery of unknown properties in data (knowledge discovery in databases) on the forecasting accuracy.

The paper is organised as follows: Section II addresses related work on estimating PV output forecasting. Data description is provided in Section III. Section IV explains the proposed methodology and presents the result discussion followed by conclusion in Section V.

II. RELATED WORK

Photovoltaic (PV) output power can be estimated either using direct or indirect methods. The direct method forecasts the output of PV directly whereas in the later case it is calculated by first forecasting the solar irradiance using irradiance data along with other meteorological variables.

There are several methods that can be used for estimating and forecasting PV output which includes statistical and time series based methods, physical methods and hybrid methods. Artificial neural network, support vector machine and regression methods lie in the category of statistical and time-series based methods. Time-series methods develop mathematical models that can forecast future observations based on available data. Physical methods uses mathematical equations to describe the physical state and the dynamic movement of the atmosphere for example numerical weather prediction (NWP). Hybrid methods combines different techniques with unique features to address the limitations of each techniques thus enhancing the forecasting accuracy. However, these models are designed for a particular location and PV plant.

Several methods can be found in the literature for predicting PV systems output using meteorological features as an exogenous input which aim for different time horizons. In [3] a procedure is proposed using a physical hybrid method (PHANN) that first identifies the optimum settings in terms of number of layers, neurons and trials and then perform day-ahead PV power forecasting using different sets for training and validation. Authors in [4] present Sun4Cast, a solar power forecasting system designed by National Center for Atmospheric Research (NCAR) which forecasts the expected sun’s irradiance and resulting power output from 15 min (nowcast) to 168h (up to a week ahead and beyond).

The two most commonly used machine learning based methods ANN (artificial neural network) and SVR (support vector regression) are applied in [5] for predicting energy predictions for 15 min, 1h and 24h ahead of time. Authors in [6] propose an ANN-based ensemble method for performing a day
ahead PV power forecasting and analyses its sensitivity with respect to input data sets. The results show that the ensemble error is smaller than error obtained by single trials. The work in [7] investigates the performance of data-driven methods for PV forecasting when different preprocessing techniques are applied to input datasets. In their work, a combination of PCA and wavelet decomposition showed the most promising results. Despite of some studies present in the literature dedicated to data cleaning and control, to the best of our knowledge, there is no study that directly observes the impact of data quality on the forecasts which causes interference with data analysis [8]. Moreover, there is lack of investigations evaluating the impact of using optimum combination (feature selection and extraction) of meteorological features and using low dimensional subspace (dimensionality reduction) present in the datasets on the forecasting accuracy. Therefore, this paper aims to contribute towards filling the research gaps present in the related work.

III. INPUT DATASET

The dataset considered in this study is a collection of data recorded at SolarTech Laboratory, located in the Department of Energy at Politecnico di Milano, Italy, for the whole year of 2014 with an hourly resolution. In total, 29 PV modules with capacity of 285 Wp are installed at the geographic location with latitude 45.502941N and longitude 9.156577E. They are all oriented with an azimuth angle $\gamma$ of $-6^\circ30'$ (assuming that $0^\circ$ is the south direction) and a tilt angle of $30^\circ$. The environmental conditions are monitored by a meteorological station equipped with a solar irradiation sensor (solarimeter), temperature, humidity sensors, wind speed/direction sensors and rain collector. Solar irradiation is measured with three different sensors, a net radiometer for the measurement of direct normal irradiance (DNI), two pyranometers for the measurement of the total and diffuse irradiance on horizontal plane. The dataset utilized for conducting this study utilizes the following parameters as presented in Table I.

The dataset contains a wide range of meteorological variables but only ambient temperature, global horizontal irradiation (GHI) and output PV power is considered. Along with this, the theoretical irradiation computed according to the deterministic Clear Sky Radiation Model (CSRM) [9], [10] is also considered. In addition, the day of the year and the hour of the day are given. The historical forecasts for the next day are delivered by a weather service station at 11:00 pm.

IV. METHODOLOGY

A. Preliminary analysis of dataset

The analysis of the dataset begins with an exploratory data analysis, which is conducted basically to check five important measures of the dataset which includes count (number of measurements), mean (average of the measurements), standard deviation (measure of amount of variance), quartiles of each feature presented in Table II. Exploratory data analysis can be considered as a first step towards quality assessment of a dataset. The very first measure, which is count for each feature, shows the value 5184, which indicates no missing values in the dataset; it is followed by mean and standard deviation and the quartile which gives the information regarding the presence of any outlier in the dataset. For example the feature $S_0$ has a mean value of 199.159 and std. deviation value of 275.461, which indicates the presence of an outlier. Fig. 1 shows the box plot visualising the outliers present in the dataset (not included in the box of observations).

![Fig. 1. Box plot indicating data distribution through quartiles](image)

B. Correlation of input features

As stated in section III, the dataset under study includes 15 input parameters. Out of which, 3 parameters namely $DOY'$, $H$, $C_t$ are not used, since the objective is to analyze the association between the meteorological parameters. For the same we use a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship using Pearson’s coefficient of correlation. In Fig. 2, we can observe the strength of relationship i.e. the value of correlation coefficient varies between $+1$ and $-1$. We visualize the correlation in a heatmap matrix using the python library seaborn. The off-diagonal elements show the degree of reliance of for each pair of variables.

Starting from variable $T$, it is noticeable that it holds a positive correlation with $T_{amb}$, i.e. 0.94, which shows the forecasted value of ambient temperature obtained from the weather station is nearly same as the value recorded at the PV site. Similarly $S_0$ has a positive correlation with $GHI_{meas}$, i.e. 0.91.

Fig. 3 shows the pairwise relationship of variables $S_0$, $S_{30}$ and $GHI_{meas}$, validating the accuracy of measured value recorded at the site with respect to forecasted value obtained from the weather station. It uses two basic figures histogram showing the distribution of single variable and scatter plot on upper and lower triangles showing the relationship between two variables.

The association of $T$ with other meteorological variables indicates that it also holds a positive correlation with $S_0$, $S_{30}$, $P_{out}$ and $GHI_{cs}$ (0.46, 0.35, 0.39 and 0.48, respectively). Correlation with $P_{out}$, with value 0.39, is not very significant,
TABLE I
LIST OF AVAILABLE DATA

<table>
<thead>
<tr>
<th>Group</th>
<th>Parameters</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>DOY-Day of the year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H-Hour of the day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GHI&lt;sub&gt;cs&lt;/sub&gt;-Global horizontal irradiation in clear sky condition</td>
<td>(W m&lt;sup&gt;-2&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Weather Forecast</td>
<td>T-Ambient temperature</td>
<td>(°C)</td>
</tr>
<tr>
<td></td>
<td>S&lt;sub&gt;0&lt;/sub&gt;-Global horizontal irradiance</td>
<td>(W m&lt;sup&gt;-2&lt;/sup&gt;)</td>
</tr>
<tr>
<td></td>
<td>S&lt;sub&gt;30&lt;/sub&gt;-Global irradiance at tilted plane</td>
<td>(W m&lt;sup&gt;-2&lt;/sup&gt;)</td>
</tr>
<tr>
<td></td>
<td>W&lt;sub&gt;s&lt;/sub&gt;-Wind speed</td>
<td>(m s&lt;sup&gt;-1&lt;/sup&gt;)</td>
</tr>
<tr>
<td></td>
<td>W&lt;sub&gt;d&lt;/sub&gt;-Wind direction</td>
<td>(°)</td>
</tr>
<tr>
<td></td>
<td>P&lt;sub&gt;amb&lt;/sub&gt;-Ambient pressure</td>
<td>(hPa)</td>
</tr>
<tr>
<td></td>
<td>P&lt;sub&gt;pt&lt;/sub&gt;-Precipitation</td>
<td>(mm)</td>
</tr>
<tr>
<td></td>
<td>C&lt;sub&gt;c&lt;/sub&gt;-Percentage of cloud cover</td>
<td>(%)</td>
</tr>
<tr>
<td></td>
<td>C&lt;sub&gt;t&lt;/sub&gt;-Cloud type</td>
<td>(Height-low/medium/high)</td>
</tr>
<tr>
<td>Measured</td>
<td>GHI&lt;sub&gt;meas&lt;/sub&gt;-Measured global horizontal irradiation</td>
<td>(W m&lt;sup&gt;-2&lt;/sup&gt;)</td>
</tr>
<tr>
<td></td>
<td>T&lt;sub&gt;amb&lt;/sub&gt;-Ambient temperature</td>
<td>(°C)</td>
</tr>
<tr>
<td></td>
<td>P&lt;sub&gt;out&lt;/sub&gt;-Output PV power</td>
<td>(kW)</td>
</tr>
</tbody>
</table>

TABLE II
EXPLORATORY DATA ANALYSIS

| T | S<sub>0</sub> | S<sub>30</sub> | W<sub>s</sub> | W<sub>d</sub> | P<sub>amb</sub> | P<sub>pt</sub> | C<sub>c</sub> | C<sub>t</sub> | P<sub>out</sub> | GHI<sub>cs</sub> | GHI<sub>meas</sub> | T<sub>amb</sub> |
|---|--------------|---------------|--------------|-------------|---------------|--------------|--------------|-------------|---------------|----------------|-----------------|-----------------|---------------|
| Count | 5184         | 5184          | 5184         | 5184        | 5184          | 5184         | 5184         | 5184        | 5184          | 5184           | 5184            | 5184          |
| Mean  | 15.9212      | 199.1591      | 249.990      | 1.452       | 165.821       | 1014.16      | 0.075        | 48.415      | 40.082        | 233.168        | 164.482         | 16.218         |
| Std. Dev. | 7.448        | 275.461       | 341.399      | 1.140       | 112.918       | 6.438        | 0.307        | 44.096      | 61.209        | 305.455        | 246.304         | 6.522          |
| 25 % | 9.9          | 0             | 0            | 0.56        | 71            | 1010         | 0            | 1           | 0             | 0              | 0               | 11             |
| 75 % | 21.4         | 355           | 448          | 1.94        | 71            | 1010         | 0            | 100         | 64            | 467.25         | 273.4           | 20.98          |

Fig. 2. Pearson correlation map
which seems reasonable due to the fact that increase in temperature till a certain degree increases the PV output however, afterwards it starts to decrease. T<sup>°</sup>C is negatively correlated with P<sub>amb</sub> due to fact that pressure decreases with an increasing temperature. S<sub>0</sub> shows very small value of correlation with W<sub>s</sub> and W<sub>d</sub> i.e. 0.20 and 0.12 respectively which indicates wind speed has a little influence on global horizontal irradiance. S<sub>0</sub> shows a strong correlation with P<sub>out</sub> which shows as the solar irradiance increases, the power generated by PV plant also increases. S<sub>30</sub> shows nearly the same relationship with the other meteorological variables as S<sub>0</sub>. W<sub>s</sub> is negatively correlated with P<sub>amb</sub>, i.e., when ambient pressure is high wind speed is usually low. C<sub>c</sub> shows positive correlation (0.28) with P<sub>pt</sub> which indicates the chances of rain and snow increase on an overcast day. Looking at the P<sub>out</sub>, it can be observed that the meteorological variables which possess the positive correlation include T, S<sub>0</sub>, S<sub>30</sub> and GHI<sub>cs</sub>.

Conducting the correlation test on the meteorological variables is an approach to assess their impact on the PV output estimation. On the basis of the obtained values and its correlation with the target variable, the list of features having positive degree of association with variable P<sub>out</sub> are listed in Table III.

Fig. 3. Pairwise relationship of S<sub>0</sub>, S<sub>30</sub> and GHI<sub>meas</sub>
C. Principle component analysis

Principle component analysis (PCA) is a technique used for dimensionality reduction of a dataset. It reduces the computational complexity of the forecasting model along with the reducing computational effort. The PCA technique decomposes a multivariate dataset in sets of successive orthogonal components known as principle components (PC). Basically PCA works with variance-covariance matrix and involves the steps which are elaborately explained in [11].

In this work PCA is implemented using scikit-learn library, sklearn.decomposition.PCA.

In Fig. 4, percentage of variance captured by each principle components is presented. We considered 12 input features which are mainly meteorological features namely $T$, $S_0$, $S_{30}$, $W_s$, $W_d$, $P_{amb}$, $P_{pt}$, $C_c$, $C_{t}$, $GHI_{cs}$, $GHI_{meas}$, $T_{amb}$. We can observe from Table IV, 37% of variance is captured by the first PC and 14% of variance is captured by second PC. Cumulatively, around 86% of variance in dataset is captured by first 6 PCs which indicates PCA does reduce the amount of input variables.

D. Regression methods

The following machine learning models are used to forecast the PV output power: The first three algorithms mentioned in Table V are implemented using class linear_model.LinearRegression from scikit-learn library. For implementing SVR there are different methods present in scikit learn, out of which in this work a gaussian kernel called radial basis function (RBF) is used. Multiple linear perceptron (MLP) is implemented using the class sklearn.neuralnetwork.MLPRegressor from scikit learn library. For conducting the training and testing of the forecast models the data is split in training and testing dataset. 80% of the dataset is used for training and rest 20% is used for testing.

E. Performance indicators

To assess the effectiveness of forecast, most commonly used evaluation metrics are present in the literature [12], [13]. In this work we used two metrics which are defined here:

1) Normalized mean absolute error - defined as mean absolute error based on net capacity of the plant $C$.

$$NMAE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{p_{pred} - p_{meas}}{C} \right| \cdot 100$$

2) Normalized root mean square error - function of model residuals

$$nRMSE = \frac{1}{\max(p_{meas})} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (p_{pred} - p_{meas})^2} \cdot 100$$

F. Results and discussion

The implemented learning models are evaluated using above-mentioned performance metrics. The results report one more parameter testing time. The results of the evaluation for five models are shown in Table VI and Table VII. Table VI presents the results when used all the features as input to the forecasting models. The model based on SVR shows the better performance in terms of NMAE and nRMSE, while linear, lasso and ridge regression model are similar in the test period and perform a little worse than SVR and MLP model. In terms of time required for testing the model, linear and ridge perform the best, lasso regression takes longer. However, the time spent for testing the model based on SVR takes a bit
long time as compared to other techniques. Table VII presents the evaluation results when selected features (obtained after conducting correlation analysis and dimensionality reduction technique PCA), are used as an input to the forecasting models. The model based on SVR obtains the best result in terms of NMAE and nRMSE, however the processing time is still more than what consumed by other techniques, although it is half of what the model based on SVR consumed in the previous case. Overall the MLP model is a better choice considering both accuracy and testing data. After comparing both tables we can observe that depending on the location under study and the regression methods, using less variables as input to the forecasting models are enough to generate nearly similar results without affecting the performance. However, it is necessary to conduct the tests under different climatic conditions so as to ensure the reliability of the results.

V. CONCLUSION AND FUTURE WORK

In this paper we discussed and presented a detailed approach to perform analysis of data as part of exploratory data analysis for energy yield production. This includes first evaluating meteorological parameters and investigating optimum combination of meteorological parameters and features which impacts forecasting models accuracy. Subsequently we compared the output of the forecasting model when fed with all the input features with the case where it is fed with selected lower subspace of the features. We compared five models, out of which the MLP-based model shows the best result. Additionally we observed that using lower subspace resulted in nearly similar results when compared to using full set of features. The future work aims at validating the methodology on different locations having different climatic conditions.

VI. ACKNOWLEDGEMENTS

This work was partially supported by Lakeside Labs via the Smart Microgrid Lab. The work was based on research activity carried out at the laboratory Solar Tech Lab, Department of Energy, Politecnico di Milano, Campus Bovisa, Milano.

REFERENCES