Big data and predictive maintenance in PV – the state of the art

O&M | Big data-based predictive analytics techniques using artificial intelligence technologies offer exciting new possibilities in the field of solar operations and maintenance. Alessandro Betti, Fabrizio Ruffini, Lorenzo Gigoni and Antonio Piazzi examine how the power of data can be harnessed to safeguard the technical and economic performance of the PV fleet

Solar photovoltaic (PV) energy is nowadays one of the most effective alternatives to conventional dispatchable energy sources, mainly due to its increasing competitiveness, the growing energy demand of developing countries and the requirement of alternative technologies to alleviate pollution and reduce global warming. In 2017, the net capacity increased faster than any other power generating technology (Figure 1), reaching approximately 402GWdc globally [1]. Focusing on the European Union, solar additions increased by 36% to 8GW in 2018.

During the operative lifetime of the power plants, high-quality O&M activities are needed in order to maintain a high level of technical and economic performance over time. In the following, we will overview the O&M needs and the newest approaches to address them. Finally, we will discuss future trends and give conclusive remarks.

The importance of O&M activities in PV plants

O&M activities are undertaken by the O&M contractor, which generally shares its tasks with the technical asset manager, which is responsible for ensuring that the operations of the PV plant are compliant with the regulations and for reporting to the asset owner, and the O&M service provider, which instead has to monitor and supervise the PV site conditions and performances (operation team), as well as to carry out the necessary maintenance activities (maintenance service team). Supervision is usually done remotely in the operation centre (control room) by exploiting analytical monitoring software systems where all data collected by dataloggers at the PV site, down to inverter and string levels, are analysed to schedule short- to long-term operations to be followed by the maintenance team.

The procedure for fault management, when a failure is detected by the control room, is generally based on a three-levels support, ranging from restoring device functionality without the need for component replacement up to component substitution and software update, and relies on an escalation of corrective actions undertaken by professionals with increasing technical skills and access permissions, until the malfunction is solved, and the corresponding ticket is closed [2].

Key performance indicators (KPIs) are used for monitoring the operation of a PV plant and for comparing PV sites in a balanced fashion. They may be mainly divided between PV power plant KPIs, describing the PV site's production performances, and O&M contractor KPIs, which instead reflect the quality of the O&M service provided. Performance ratio (PR) [3] and availability [4] belong to the first group and are usually supervised by the asset manager by ensuring the optimal profitability of the plant over time. The availability, defined as the time percentage the plant operates over the whole time it should operate (usually required to be higher than 98% over a year), is also a striking indicator of the plant behaviour that needs continuous monitoring by the O&M service provider to undertake corrective actions when necessary. O&M contractor KPIs instead include the acknowledgement, intervention and resolution times, which monitor the time

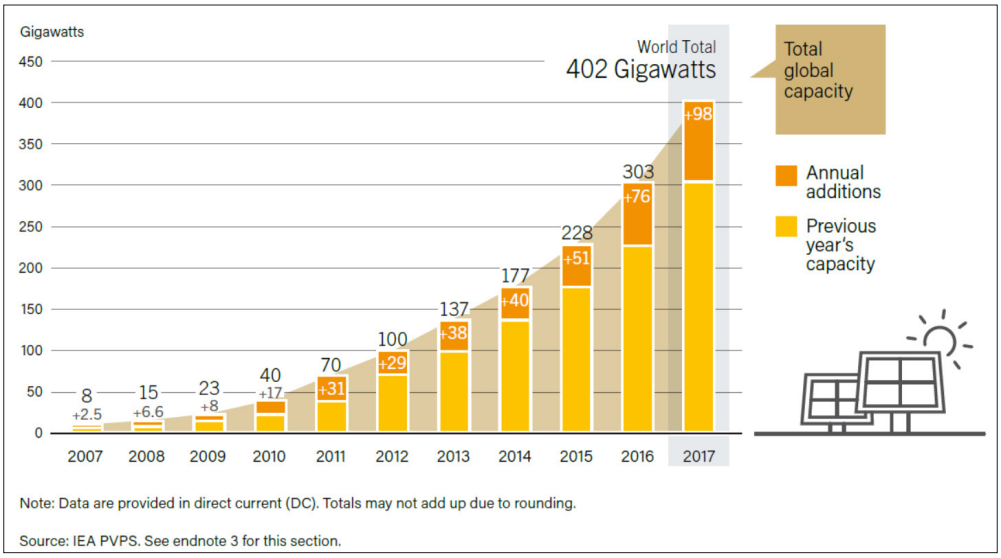


Figure 1. Solar PV global capacity and annual additions for the period 2007-2017 [1]

necessary to acknowledge an alarm, reach the plant and solve the issue, respectively.

Several kinds of maintenance strategies can be followed. **Preventive** maintenance is one of the typically followed: it includes regular visual and physical inspection of key components, such as string measurements or thermal scans of PV panels, in order to identify problems as soon as possible and start O&M activities. However, such a proactive strategy can involve expensive inspections with third parties and, since the frequency of such activities is not typically optimised, can lead to unsatisfactory results. On the other hand, the opposite **reactive** strategy, undertaking a corrective action only when a failure occurs, is usually more expensive as the losses due to downtime and repair or substitution of devices are higher.

In this scenario, the need for strategies able to predict incoming faults emerges, such as condition-based **predictive** maintenance approaches. Unlike preventive and reactive strategies, this kind of strategy optimises simultaneously the downtime periods, the lost production and the total cost of maintenance activities; therefore, it should be considered as the core strategy of the O&M contractor activities. A recent study for GE Oil & Gas of offshore gas facilities [5] showed a clear relationship between the followed maintenance approach and the unplanned downtime. In Figure 2 the annual financial impact of maintenance activities (blue bar, on the left) and the unplanned downtime rate (grey curve, on the right) are shown for the reactive, preventive and predictive approaches.

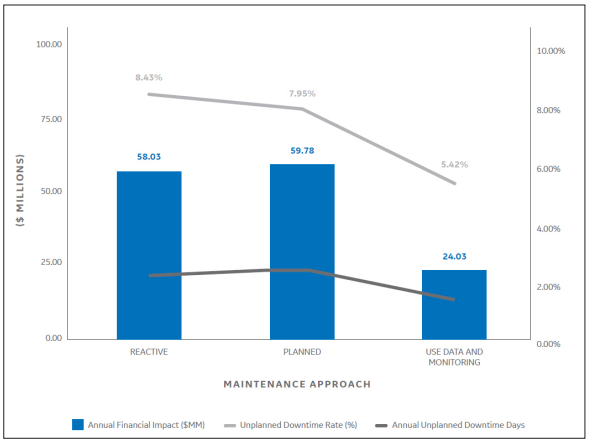


Figure 2. Annual financial impact of maintenance activities for offshore oil & gas facilities (in millions of dollars, on the left) and annual unplanned downtime rate and days (on the right, grey and black curves, respectively) for the reactive, preventive (planned) and predictive (use data and monitoring) approaches [5]

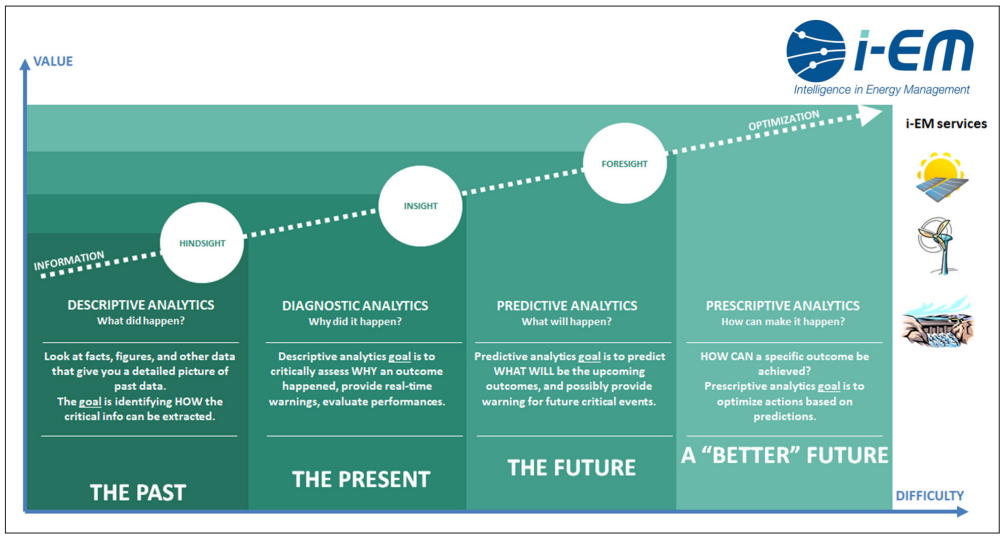


Figure 3. Chart describing the value added to the O&M activities undertaken by the four approaches – descriptive, diagnostic, predictive and prescriptive analytics – as a function of implementation difficulty (inspired by [6]). The higher is the service complexity, the larger is the value gained, up to predicting the future outcomes (predictive) and delivering the actions required to solve predicted failures (prescriptive)

According to this study, a 36% decrease in unplanned downtime activities, as well as a 60% drop in the annual financial impact, may be achieved by following a predictive strategy.

Towards a predictive maintenance strategy

Predictive maintenance usually requires the collection of a large amount of data at PV sites. Since data analytics is effective in extracting actionable insights from data, now companies are struggling to digitalise their processes. Data analytics may be split into four main groups depending on the growing level of added value they bring and also on the complexity they require for design and implementation: **descriptive, diagnostic, predictive** and **prescriptive** analytics (Figure 3). A descriptive solution provides valuable insights into the past by means of data mining algorithms and summarising raw data from multiple sources. No additional information is provided, such as the reasons why the event happens. Unlike the descriptive solution, the diagnostic approach discovers dependencies in data and identifies hidden patterns, getting into the causes of a past event or ongoing behaviour. However, this diagnostic level is still reactive, since it is applied to past/real-time events.

As information volume and quality increase, organisations may move towards the realm of a forward-looking (proactive) approach, adjusting their strategy according to what is predicted for the future. In particular, predictive analytics adds a level of complexity to the descriptive and diagnostic analytics, exploiting their findings to predict tendencies and future trends. It takes advantage of statistical models, either classical or machine learning, and it is based on a probabilistic foundation to forecast the likelihood of a future outcome and provide actionable insights to companies.

Prescriptive analytics, starting from predictive outcomes, finally suggests the action(s) to prevent a future issue of an asset or to take advantage of a predicted trend. It can be successfully used to optimise the production, the scheduling and inventory in the supply chain, or O&M actions (for example, operating only where and when necessary).

Predictive analytics on PV plants: benefits and challenges

While descriptive and diagnostic are well-established techniques in the PV energy sector, predictive and prescriptive maintenance are still at an embryonic stage and utilities and PV plant owners are only just beginning to turn their interest to such topics. These activities require a data-driven approach in O&M activities, where an accurate data collection including real-time data, historical data, data from similar assets and historical maintenance records is necessary, therefore delaying the shift from the feasibility-study level to real implementation.

The benefits of a proactive approach are manifold: a reduction in the component repair and replacement costs of factory equipment, a decrease in the

revenue loss due to plant downtime, a reduction of capital investment by extending the useful life of devices, better inventory management, and, in general, more effective O&M activities.

However, challenges for the real implementation of a predictive service are also manifold: first and foremost, the need to gather at a predefined acquisition rate a huge amount of data related to both electrical and environmental signals, as well as archiving either automatic or manual alarm logbooks describing past failure events. Furthermore, as a predictive model is generally trained over healthy component-related instances and tested against unknown periods in which the component status should be predicted, a preliminary separation of historical instances into normal or abnormal classes is mandatory but often difficult in practice. Clearly, a stable internet connection and a speedy and reliable IT infrastructure managing field data collection are required, starting from low-level Internet of Things (IoT) sensors and SCADA software installed at a PV site, up to a big data platform at the monitoring room level where data from different plants and countries are archived.

The development of a predictive model must not forget to combine experience in the PV technical domain with data science expertise. This requires the inclusion of the O&M team in the whole lifecycle of the predictive service process and, in particular, in the phases of model features selection, model development and online model validation, where the accuracy of the model is compared to real-time events. But domain knowledge should not be limited to the interaction with the O&M team, as it should include also asset manufacturers in order to define, prior to the model design, a fault taxonomy table reporting, for each fault type that may be triggered in the field, at least the corresponding device involved, the alarm type, the manufacturer code, the event name and description, the event severity, as well as the potential cause of event and the actions to solve the issue. This is necessary to get the correspondence among the past events recorded in the fault logbook and the taxonomy file, to prioritise events and to customise and tune the predictive models against the failures considered more critical according to customer needs.

Figure 4. Artificial intelligence has infinite applications. Recently a research group adopted such techniques for predicting new patients affected by Alzheimer's disease [7]

Techniques for predictive maintenance on PV plants: the key role of artificial intelligence

The search for increasingly practical and accurate predictive maintenance tools has coincided with the simultaneous growth in artificial intelligence technology and the introduction of statistical methods based on machine learning. Such a discipline has infinite applications: some months ago, a research group from California claimed for example its application to diagnose patients with Alzheimer's disease based on brain scans made some years before [7].

A typical workflow of an artificial intelligence model is outlined in Figure 5. A dataset is typically extracted from an archive and preliminarily preprocessed to obtain clean data useful for further processing. Domain knowledge is usually required to select the best predictors for the problem of interest, as well as to combine different signals to achieve enhanced predictors (feature engineering). In order to handle signals concerning heterogeneous quantities, feature normalisation is typically applied. Then, traditional statistical or machine learning-based algorithms are used to create models based on these features and validated against a test set (historical and/or real time) to verify performances. Usually an iterative approach is applied in order to find the best coupling between preprocessing and processing phases which maximises performances. Finally, the developed model is deployed to deliver a periodical service to the customer.

Predictive maintenance models may be designed for different target PV components: PV module, string of PV panels, inverter, or the whole plant. They may be also grouped in three different categories, characterised simultaneously by an increasing level of details provided and by a shorter prediction horizon: prediction of generic faults and machine status, prediction of severity category of the incoming event and prediction of specific faults. In the first case, the model predicts a generic failure through a measure of deviations from normal operation, in the second scenario it returns the criticality of the fault event according to asset manufacturer taxonomy, whereas in the third one it provides the specific fault class among those available in the taxonomy archive. The algorithmic core of the model and its complexity, as well as the input features fed into the model and the training methodology adopted, change according to the level of detail required for the prediction. In particular, the prediction horizon may reach days or weeks in the first approach down to few days, or even hours or minutes, in the latter, depending on the statistics available and on the degree of correlations among the input predictor and the fault predicted [8].

An example of prediction of specific fault classes is shown in Figure 6 (next page) where the classification metrics accuracy, sensitivity and specificity, as

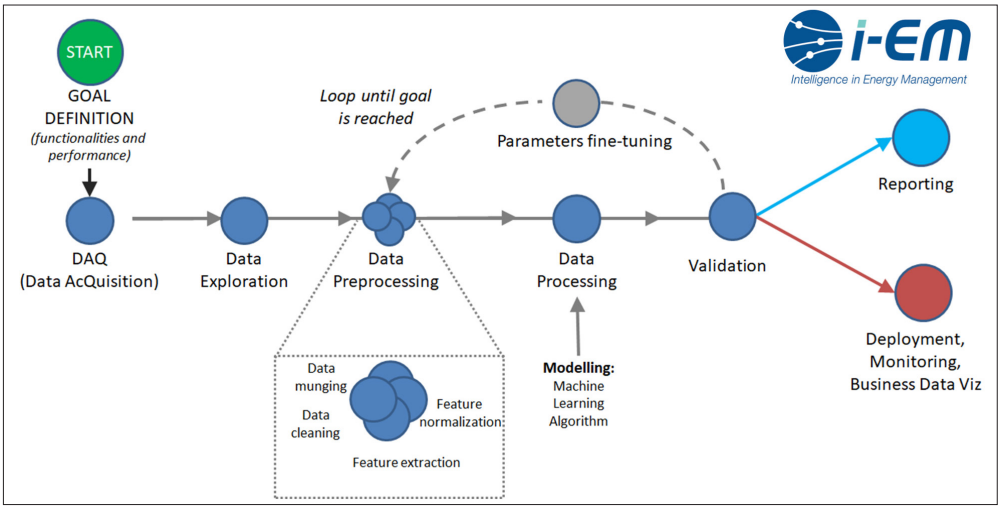


Figure 5. A typical machine learning workflow

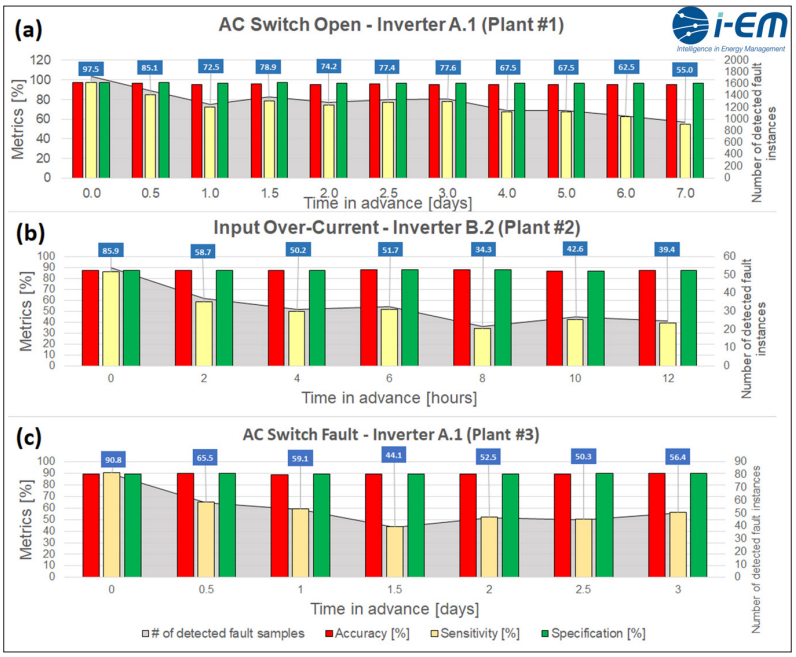


Figure 6. Classification metrics (bar plot on the left) and number of detected faults (grey area on the right) as a function of time in advance for predictive model developed by [8]. (a) fault class AC switch open (plant #1 in Romania); (b) fault class input over-current (plant #2 in Greece); (c): fault class AC switch fault (plant #3 in Greece)

well as the number of detected historical faults, are shown as a function of the time prior to the fault occurrences for different inverter failure class. Three examples for three different plants, one located in Romania and two in Greece, are reported (Figure 6 (a), (b) and (c), respectively). As can be seen, a strong correlation between statistics available and model performance is evident, since machine learning algorithms are black-box computing units which learn the underlying non-linear relationship between input and output according to the training dataset available. When thousands of occurrences are available (Figure 6a), the prediction horizon is as large as seven days, with sensitivity decreasing down to almost 50-60% about one week ahead. On the other hand, when the statistics amount to almost one hundred instances, prediction capabilities degrade much faster on the time horizon from one hour to 12 hours in advance (Figure 6b). It is worth noticing that, however, a strong correlation between input predictors and fault class may enlarge the horizon in some exceptional cases (Figure 6c).

Control charts and machine learning algorithms

The category of models predicting generic failures includes essentially statistical process control approaches trained over nominal behaviour periods of the modelled component and then able to early detect or predict a not-nominal trend. Such approaches may be divided between univariate and multivariate, depending on how many signals are

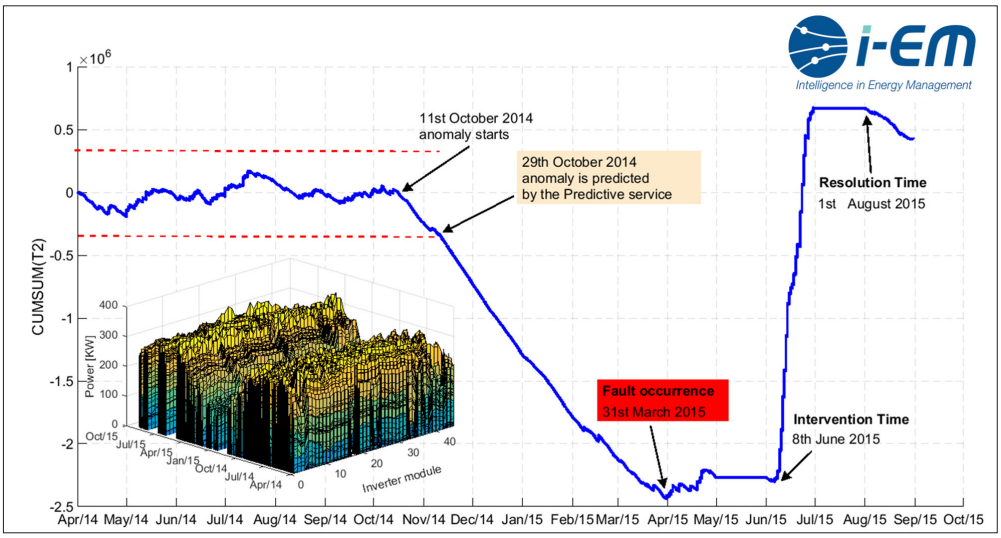


Figure 7. Cumulative sum of T2 as a function of time for a PV plant in Romania. Inset: AC power for the inverter module installed on site. Reproduced with permission of i-EM srl, Flyby Group (www.i-em.eu)

considered in the analysis. In particular, unlike the univariate method where a single signal at a time is analysed, the multivariate analysis accounts for a bunch of signals simultaneously, therefore including the effect of correlations.

The procedure in both cases generally consists of two phases: retrospective and prospective. In the retrospective phase a nominal period for the variable of interest is considered, where the variable may be a single signal or a KPI obtained by aggregating together different signals. In both cases, the model is built or trained against the nominal period identified by using domain knowledge (best case) or, at worst, by statistical considerations as explained earlier. Then, in the prospective phase, the algorithm is tested over an unknown period to predict out-of-control component behaviour, triggering a generic warning. Auxiliary information may be also provided along with the warning, such as the specific sub-class problem, if the tags more correlated to the anomaly were identified, and the fault severity, in order to simplify also the maintenance activities. Additionally, in the case that the sub-class problem was suggested, the action to solve the issue may be proposed by getting it from a look-up table storing the correspondence between issues and actions, thus turning the approach from predictive to prescriptive. However, since the information regarding specific failures are neither present in this approach, nor used for training, and, additionally, different tags may concur to the same anomaly or the same tags can lead to different anomalies, the prescriptive suggestions provided may have a limited effectiveness.

The most common statistical methods predicting generic failures include traditional approaches, such as Hotelling's T2 control chart [9, 10], and machine learning-based algorithms, both supervised and unsupervised. T2 is based on correlation analysis and describes the global system behaviour. It can be interpreted as a deviation of the process from a nominal condition. When deviation is below a threshold, the system is under control. On the contrary, when the threshold is exceeded, the process is declared out-of-control. Another control chart widely researched is the cumulative sum [10], which is efficient in detecting small shifts in the mean of a process. Cumulative sum is simply the partial sum of the variable of interest up to the current element and removing the mean value of the variable. By analysing its trend and unexpected and sudden slope changes, an out-of-control process can be easily detected or predicted.

An application example of cumulative sum is shown in Figure 7 for a historical failure of a PV plant in Romania, which suffered a severe thermal issue in 2015 that led to replacements of different inverters and to a prolonged downtime period. A deep valley with abnormal behaviour can be observed starting approximately at the end of October 2014, some months before the failure occurred at the end of March 2015. In particular such failure led to a plant downtime of some months, with the intervention time scheduled in June 2015 and the resolution time happening finally in August 2015, when the plant recovered its normal operation.

Besides traditional control charts, machine learning methods may be also applied: they include, for example, neural network (NN) and self-organising map (SOM) [8]. While NN belongs to the class of supervised algorithm, i.e. a target is present, SOM is unsupervised, i.e. it does not have this information. In a supervised learning model, the algorithm learns on a labelled dataset, for example represented by a set of input instances tagged with binary values identifying nominal or abnormal behaviour (the target), and provides an output that the algorithm can use to evaluate its accuracy on training (retrospective) data, before inferring over test data.

An unsupervised model, in contrast, provides unlabelled data that the algorithm tries to make sense of by extracting features and hidden structures on its own. In particular a SOM makes a non-linear mapping from an input N-dimensional space to a 2D space and preserves input topology by exploiting a competitive learning process. Changes in clusters emerging in the SOM map may be monitored by means of a KPI measuring a process variation from normal state towards abnormal operating conditions when a threshold is crossed [8].

Prediction of specific faults needs, first of all, an amount of information, such as alarm logbooks and categorised taxonomy files, which at the present stage is only sometimes available. Indeed, this entails not only a speedy, reliable and fault-tolerant acquisition chain but also demands cooperation of the O&M team and asset manufactures. From a design point of view, only supervised algorithms may be applied since they must be trained against specific fault classes: typical suitable architectures are pattern recognition feed-forward NNs [8] or deep learning structures such as a stack of auto-encoders.

A common issue is the so-called class unbalancing because the number of samples available for nominal class (the so-called negative or majority class) generally is much larger than that available for the faulty one (the so-called positive or minority class). Since training is done by minimising a cost function where the contribution of minority class is small, the model prediction is biased toward the majority class and, on average, misclassification of minority instances occurs with a higher rate. Different techniques may be applied to overcome such a problem,

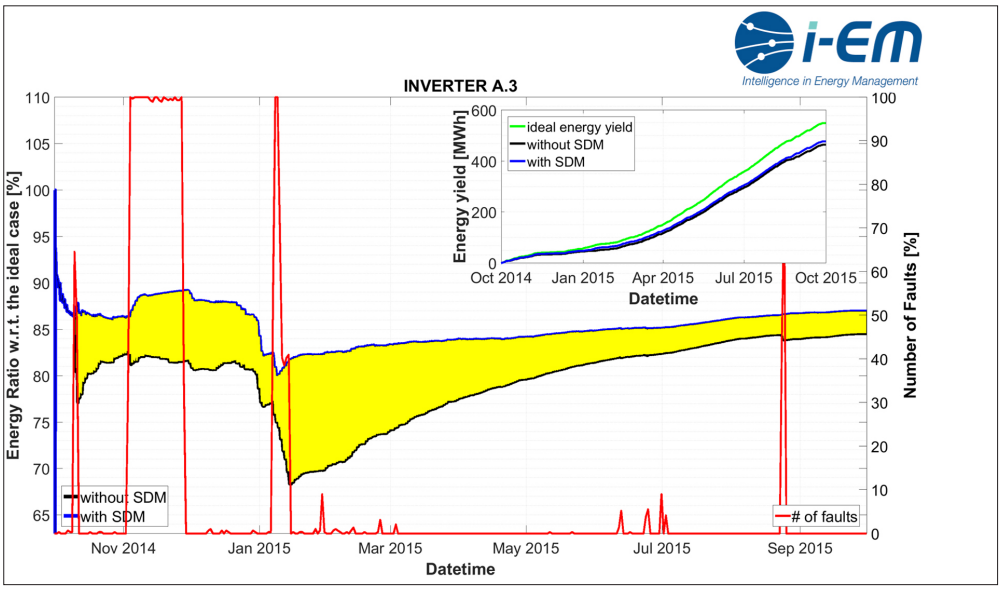


Figure 8. Energy yield time series with respect to the ideal case with (blue curve) or without (black) the predictive service: yellow area represents the energy gain enabling such service. Fault occurrences percentage is also shown on the right (red). Inset: energy yield in the ideal case, as well as with or without the predictive service. Data are referred to an inverter module of a PV plant located in Romania [8]

which may be grouped essentially in undersampling and oversampling. In the first case the number of instances of majority class is undersampled to make their number comparable to that of minority class; in the second case instead artificial instances are created to oversample the amount of minority faulty instances (e.g. SMOTE technique).

Prediction of specific faults formally closes the gap between predictive and prescriptive approaches because the fault taxonomy file may be used as a look-up table to suggest the action to solve the issue in correspondence with the predicted failure.

Further potentially effective techniques are those usually applied in the wind energy field, including bivariate analysis based on power curve modelling [11] and condition parameter-based models [12]. The latter approach, roughly speaking, consists in training over healthy instances and predicting the status of the component by monitoring the residuals between the forecasted and the measured target parameters: if such a residual exceeds a threshold, a warning is triggered. Such a method has also been applied in the PV sector to analyse fault classes affecting inverter power production (e.g. "lack of isolation" failure mode in [13]), using, as input predictors to a NN, the electrical and environmental signals correlated to production (e.g. internal inverter temperature, accumulated active energy of the inverter, irradiance, ambient temperature, etc.) and as a target the power production at the AC side of the inverter [13]. However, such methods fail to provide a comprehensive picture of the correlations among the component parameters. In addition, they cannot identify all the component failures, as their root causes may be classified according to their impact as affecting the system performance, the system availability and the cost of operations and reporting.

Profitability of a predictive service

Current literature regarding profitability of predictive service is still lacking and some revenue gain estimations have started to appear only recently. According to [14], application of a smart predictive service may increase the annual revenue of a typical standard-performing fleet of 100MW PV plants from €128,000/year up to €240,000/year and from €368,000/year up to €948,000/year for a low-performance PV portfolio. Considering a typical lifetime of 20 years, the cumulative impact will range from €2.5 million to €20 million. Similar estimations are presented in [15], where a 2% performance increase of a 100MW PV plant leads to €500,000 of additional annual revenue and a €420,000 saving in the annual O&M activities. However, it is worth noting that these benefit analyses address a suite of tools which includes not only the application of machine learning-based models, but also additional performance trend studies

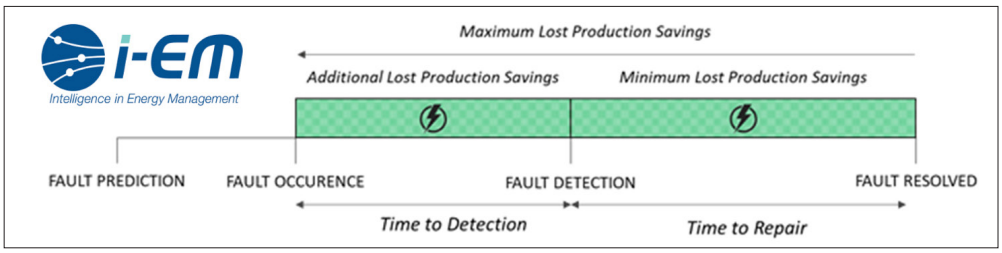


Figure 9. Fault events timeline. The positions of fault prediction by the predictive service, as well as the fault occurrence, fault detection by standard monitoring platform and nominal behaviour restored are shown. The predictive service allows a lost production saving from one month (assumed as the mean duration of time to repair) up to almost two months (assuming three weeks as the mean duration of time to detection according to current monitoring systems)

that support proactive maintenance. As a consequence, since predictive maintenance is not the unique benefit provided, the estimated values cannot be assumed as a precise indication of the profitability of a "standalone" predictive maintenance service based on machine learning techniques.

An interesting assessment of the maximum energy gain achievable, if a predictive service had been installed on-site and the ideal energy production instantly restored in correspondence of the fault prediction, is presented in Figure 8 of work [8]. On the left axis the energy yield with respect to the ideal case (i.e. where AC power is the theoretical one achievable according to actual irradiance) is shown as a function of time with and without the predictive service (here called SDM) enabled (blue and black curves, respectively). On the right axis the historical number of normalised daily fault occurrences is shown as a function of time (red curve). Such a number is computed according to the time duration of fault events recorded in the alarm logbook available for the PV site. The component under consideration is an inverter module of the same plant in Romania observed in Figure 7. As can be seen, if the predictive service had been applied in such a plant, energy yield would have been increased ideally by up to almost 10-15%, saving also the costs of inverter replacements and maintenance activities.

By considering a portfolio of PV assets and focusing on a predictive service at inverter level, the impact of its costs and benefits on the net yearly revenue gain may be assessed. Such benefits can be mainly grouped in decrease of revenue loss due to device downtime and reduction of O&M costs.

Figure 9 helps to address the impact of the first factor, depicting a typical situation encountered, with the predictive service anticipating the failure occurrence, unlike a typical monitoring platform where the failure is detected only after it has happened. In particular, a mean value of about three weeks has been assessed by considering standard operations activities. In addition, according to [16], the time-to-repair (i.e. intervention plus resolution times) is about one month. According to these assumptions, the predictive service may therefore enable saving periods ranging from one month (time-to-repair) up to almost two months (time-to-repair + time-to-detection). According to domain expertise, a further reduction of O&M activities cost of about 20-30% may be achieved.

Table 1 summarises the main hypothesis and shows the benefits provided by the predictive service, both in terms of revenue gain, ranging from €90,000/y up to €145,000/y, and O&M cost saving, ranging from €280,000/y up to €420,000/y (considering O&M activity costs of €14/kW). It has been assumed a portfolio of 100MW, a specific yield of 1,500KWh/KWp (utilisation factor of 17%), an energy price of €100/MWh, an inverter failure occurrence probability of 8% [12, 14] and a model sensitivity of 85%. In short, the total gain enabled by a predictive service ranges approximately from 2.7% up to 4.2% of the yearly net revenue of a solar asset.

Benefit Assumption	
Time to Repair	744h (1 month)
Time to Detection	504h (3 weeks)
Predictive Model Sensitivity	85%
Failure Probability	8%
O&M Savings	20-30 %
Benefit Evaluation	
Revenue Gain	*from* €90k/y up to €145k/y *
O&M Cost Savings	*from* €280k/y up to €420k/y **
Benefit impact on net revenue	*from* 2.7% up to 4.2% ***

Table 1. Benefit assumptions used for the evaluation of predictive maintenance service profitability and benefit values. The latter are reported both as absolute values and as percentages of the net revenue of a PV portfolio.

Internet of Things and the big data challenge

But why data are the new wealth just now? And how to extract value from such data?

Nowadays we live in the so-called era of the Internet of Things (IoT): a broad range of devices and objects are "smart", i.e. linked to the internet and to each other, and able to acquire and manage data. It is estimated that by 2020 the accumulated volume of big data will increase up to roughly 44 zettabytes (ZB), i.e. 44 trillion GB [17], due to the huge increase in things creating data and the refined granularity of data being produced. Big data are not characterised only by the Volume, but also by Variety, Velocity, Veracity, and Variability (the so called five "Vs" of big data"). Such data are generated by a great variety of heterogeneous sources, from social media to sensors and mobile devices, both in structured and unstructured forms. They are also collected at a high rate (velocity): every 60 seconds, it is estimated that there are 72 hours of footage uploaded to YouTube, more than 2 million Instagram posts and 204 million emails sent. In addition, data need veracity, i.e. they should be of good quality that is continuously updated in real-time. Finally, the meaning of data depends on the context in which they are collected, making important the use of technical domain knowledge (variability).

The IoT revolution occurs also in the PV energy sector: all components are now instrumented and data loggers allow the monitoring of many heterogeneous parameters thanks to specialised sensors. They include, for example, inverter internal parameters, generation data and meteorological data. Such data are typically pushed to a cloud server, which gives the flexibility of preserving a huge volume of historical plant data. The data are also stored at a local control room and can be retrieved in case of communication failure with the cloud server, thus ensuring reliable and accurate data availability.

But such a volume of data requires software and hardware infrastructure suitable for analysing in real-time this continuous stream of information in order to extract meaningful insights, and

then convert such insights into actions useful to improve the overall business value. Application domains are manifold: transportation industry, media and entertainment industry, health industry, government industry, energy sector and many others. Here the big data analytics tools come into play, due to their capacity to handle large volumes of data generated from IoT devices. And this is also the right time to analyse such data by means of machine learning and deep learning techniques; indeed, the availability of massive amounts of data proceeds along with the advances in machine learning algorithms and the dramatic progress in computer processing capabilities.

Hadoop and modern big data platforms

It has been calculated that a PV plant, with an installed capacity of 500MW and single panels generating around 200W of DC power, produces almost 8GB of data every second [18]. Such a volume can be neither stored in conventional databases, nor processed by a single local computing resource.

In 2008 Yahoo released Apache Hadoop as an open-source project and in the last few years large companies have adopted it as a next-generation platform, collecting massive data assets in Hadoop Data Lakes. In particular, Hadoop is an open-source platform and framework for storage and large-scale processing of datasets.

Hadoop offers many advantages: first, it can store and process quickly huge amount of heterogeneous data and it can archive both traditional structured data and challenging unstructured data.

Secondly, Hadoop has an enormous computing power and may also handle virtually limitless concurrent tasks or jobs. In addition, it is fault-tolerant and flexible, since both unstructured and structured data may be stored without the need of pre-processing. It is finally low-cost and scalable.

However, while using Hadoop for broad predictive analytics, companies discovered limits in its use concerning both performance and complexity. For this reason, a new platform called Apache Spark has been developed on top of Hadoop, leveraging Hadoop's big data management capabilities while achieving higher performances by running predictive analytics in Apache Spark.

Many different Hadoop distributions exist. Top tier includes solutions such as Cloudera, Hortonworks, MapR, IBM and Pivotal. They may be deployed either on customers' premise, in a private cloud or in a public cloud. Additional cloud-based Hadoop distributions exist, such as for example Amazon Web Services and Microsoft Azure HD Insights: unlike the previous distributions, such solutions run on public clouds and cannot run on the customer's hardware.

Paradigm shift: "data to computation" to "computation to data"

The complexity of using Spark and Hadoop to develop predictive analytics applications on large data assets, however, makes it challenging for companies to find or train human resources with the right skills. For these reasons, recently Microsoft has launched a new flexible enterprise platform called Microsoft Machine Learning Server (previously known as Microsoft R Server), which allows R developers to conduct the different steps of data science, from data exploration to predictive modelling, on large data assets stored in Hadoop, but without the need to become Hadoop experts themselves. The solution is the result of the acquisition in 2015 of the company Revolution Analytics by Microsoft and of further improvements of the product already available in that year. A recent version supports also the Python language. Thanks to the availability of the RevoScaleR Package, such language may manage large datasets and develop machine learning algorithms without the need of loading them all at once in the memory. Additionally, it makes possible to run code in an efficient, parallel and scalable fashion, finally deploying the model on a remote server such as SQL Server or a Spark cluster with minimal effort, thus reducing the time-to-market of the product or service developed. R Server therefore shifts the computation methodology from the traditional paradigm "Data to Computation", i.e. data moved from the environment where they reside to that of computing unit, to the new one "Computation to Data", i.e. computation performed just where the data live. In this manner, the time to move data is avoided and, additionally, it may be taken advantage of the computing power, as well as of the scalability of the environment where data are located.

Future trends

As discussed, the digital revolution of the PV sector is just happening: due to the need of reducing the cost of maintenance activities, solar operations and maintenance vendors are now turning to innovative technologies to remain competitive and profitable. In addition to big data analytics, deep learning and augmented reality will be the next key innovations to enhance maintenance capabilities, by improving the efficiency of operational processes, and by strengthening the digitalisation process. PV systems require frequent diagnosis to analyse the effect of external agents on PV panels. Thermographic inspections are the most effective methods for PV module failure detection. However, manual analysis of massive amounts of images acquired by cameras mounted on drones or car roofs and resulting from inspection of large-scale PV power plants is time consuming and prone to human errors. Deep learning, which is a machine learning technique generally applied to classification and/or detection (i.e. classification and localisation) of objects inside images, may help in automating such analysis and locating earlier potential defects at cell, module and string levels such as hot spots, cracks or abnormal soiling, as well as classifying failures in real-time.

Automatic object detection may be combined also with augmented reality (AR) tools, in order to overlay on the detected asset its corresponding virtual object and support maintenance activities (see Figure 10 [19]). In particular AR and virtual reality (VR) tools have multiple benefits in maintenance activities: they reduce downtime costs, due to quicker intervention times, and enhance employee capabilities, by augmenting and speeding up their cognition by showing only the necessary information of the environment all around. In addition, they decrease travel costs worldwide of maintenance teams, allowing the operator to request online real-time involvement of remote specialists or to request online big data analytics processes to run on the interested site. Finally, they reduce costs and improve the effectiveness of training courses, allowing trainees to learn in an immersive VR environment synthesised from reality: for example, initially projecting augmented reality contents on to a virtual environment while the trainee is in its office and, as a second level of training, showing AR

Figure 10. Augmented reality for assistance in maintenance operations [19]

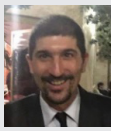
contents over the real environment as detected from a camera on the helmet of the user. An example of ARVR generic Software Development Kit (SDK) is that provided by Mapbox [20] that, based on most widespread 3D engine and libraries like Unity or OpenGL, allows developers to use their APIs with pay-per-use-fees (free for emerging applications). Instead Reflect Remote [21] is an interesting product of AR solution for remote assistance employed together with holographic or 3D glasses.

Additionally, Data Analytics as a Service (DAaaS) is now starting to attract attention also in the renewable energy field. It is an extensible analytical cloud-based platform approach where various tools for data analytics are available to users and can be configured by the users themselves to process and analyse massive amounts of heterogeneous data. It includes mainly two elements: a run-time environment, i.e. a platform for processing data, and a workbench environment where the users, from a skilled data scientist to a business user or a maintenance operator, may configure the system by using a set of analytics tools to handle different use cases. In this manner, it is possible for a maintenance team, which does not want to share their technical knowledge with external data analytics vendors, to analyse, interpret and predict underlying patterns in data even if they have no specific data science expertise. The direction is clear: make data science more democratic allowing everyone, even those not having a data science background, to analyse data and make predictions by means of automatic models built on raw data. Scientists at MIT are recently researching on this topic [22].

The future is just around the corner. ■

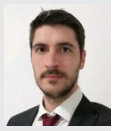
Authors

Alessandro Betti received an MSc degree in physics and a PhD degree in electronics engineering from the University of Pisa, Italy, in 2007 and 2011, respectively. His main field of research was modelling of noise and transport in quasi-one dimensional devices. His work has been published in 10 papers in peer-reviewed journals in the field of solid state electronics and condensed matter physics and in 16 conference papers. In September 2015 he joined the company i-EM in Livorno, where he currently works as a senior data scientist developing power generation forecasting, predictive maintenance and deep learning models, as well as solutions in the electrical mobility fields and managing a data science team.



Fabrizio Ruffini received a PhD degree in experimental physics from University of Siena, Italy, in 2013. His research activity is centred on data analysis, with a particular interest in multidimensional statistical analysis. During his research activities, he was at the Fermi National Accelerator Laboratory (Fermilab), Chicago, USA, and at the European Organisation for Nuclear Research (CERN), Geneva, Switzerland. Since 2013, he has been working at i-EM as a data scientist focusing on applications in the renewable energy sector, atmospheric physics and smart grids. Currently, he is working as senior data scientist with a focus on international funding opportunities and dissemination activities.

Lorenzo Gigoni received an MSc degree in energy engineering from the University of Pisa, Italy in 2016. He is employed at i-EM S.r.l., where he worked for three years as a data scientist in the R&D department developing predictive models for renewable energy systems. His main research activities were on predictive maintenance, forecasting and nowcasting models applied to wind and solar plants. Currently, his activities are focused on technical support to the sales and business development departments.



Antonio Piazzi received an MSc degree in electrical engineering from the University of Pisa, Pisa, Italy, 2013. An electrical engineer at i-EM since 2014, he gained his professional experience in the field of renewable energies. His research interests include machine learning and statistical data analysis, with main applications on modelling and monitoring the behaviour of renewable power plants. Currently he is working on big data analysis applied on hydro and photovoltaic power plants.

Turn to next article, p.32, for insights into how the solar industry is taking advantage of the opportunities offered by big data

References

[1] Renewable Energy Policy Network for the 21st Century (REN21), "Renewables 2018 Global Status Report", 2018
[2] Solar Power Europe, "Operation & Maintenance. Best Practice Guidelines", Version 3.0, 2018
[3] PR is the ratio between the system's final yield *Yr* and the reference yield *Yr*. Definition based on "Woyte, Achim, Mauricio Richter, David Moser, Nils Reich, Mike Green, Stefan Mau, and Hans Georg Beyer. 2014. "analytical Monitoring of Grid-Connected Photovoltaic Systems - Good Practice for Monitoring and Performance analysis." Report IEA-PVPS T13-03: 2014. IEA PVPS, in line with IEC 61724-1:2017, and are common practice.
[4] The technical availability is the parameter that represents the time during which the plant is operating over the total possible time it is able to operate, without taking any exclusion factors into account. Definition from SolarPowerEurope report "OM-Best-Practices-Guidelines-V3.0", December 2018.
[5] The Impact of Digital on Unplanned Downtime, an offshore Oil and Gas perspective. Baker Hughes, a GE Company, October 2016
[6] Gartner, "Data and Analytics Leadership Vision for 2017", 2016
[7] https://tech-news.websawa.com/researchers-say-training-an-artificial-intelligence-to-figure-out-alzheimers-in-patients/
[8] A. Betti, M. L. Lo Trovato, F. S. Leonardi, G. Leotta et al, Predictive Maintenance in Photovoltaic Plants with a Big Data Approach, 33rd European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC), pp. 1895-1900, 2017
[9] Process or Product Monitoring and Control, Engineering Statistics Handbook (www.itl.nist.gov/div898/handbook/pmc/pmc.htm)
[10] The application of Hotelling's T2 control chart in an automotive stamped parts manufacturing plant, Muzalwana Abdul Talib et al (umexpert-um.edu)
[11] A. Kusiak, Monitoring Wind Farms with Performances Curves, IEEE Transaction of Sustainable Energy 4, pp. 192-199 (2013)
[12] M. Schlechtingen, I. Santos, Comparative Analysis of Neural Network and Regression Based Condition Monitoring Approaches for Wind Turbine Fault Detection, Mechanical System and Signal Processing 25, pp. 1849-75, 2010
[13] F. A. O. Polo, J. F. Bermejo, J. F. Gomez Fernandez, A. C. Marquez, Failure mode prediction and energy forecasting of PV plants to assist dynamic maintenance tasks by ANN based models, Renewable Energy 81, pp. 227-238, 2015
[14] White Paper, Beyond Standard Monitoring Practice, 3E (http://www.3e.eu/white-paper-beyond-standard-monitoring-practice/)
[15] Solar Plant Asset Performance Management (APM). Improve Solar Asset Performance and Reliability, Reduce Operating Cost and Risk (file:///G:/best_practices/Solar%20APM_Overview_Final20180906_2.pdf)
[16] Solar Bankability, "Technical Risks in PV Projects - Report on Technical Risks in PV Project Development and PV Plant Operation", 2017.
[17] https://www.newgenapps.com/blog/big-data-statistics-predictions-on-the-future-of-big-data
[18] S. Begum, A. Ahamed, R. Banu and P. B. D., A Comparative Study on Improving The Performance Of Solar Power Plants Through IOT and Predictive Data Analytics, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), pp. 89-91, 2016
[19] https://www.forbes.com/sites/bernardmarr/2018/07/30/9-powerful-real-world-applications-of-augmented-reality-ar-today/
[20] https://www.mapbox.com/
[21] https://www.re-flekt.com/reflekt-remote
[22] http://news.mit.edu/2019/nonprogrammers-data-science-0115