Applied Energy 263 (2020) 114642

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Detection of cleaning interventions on photovoltaic modules with machine learning

Matthias Heinrich[a,*], Simon Meunier[b,c], Allou Samé[a], Loïc Quéval[b,c], Arouna Darga[b,c], Latifa Oukhellou[a], Bernard Multon[d]

[a] IFSTTAR | COSYS-GRETTIA, 14-20 Boulevard Newton, 77420 Champs-sur-Marne, France
[b] University of Paris-Saclay, CentraleSupélec, CNRS, Group of Electrical Engineering Paris, 91192 Gif-sur-Yvette, France
[c] Sorbonne University, CNRS, Group of Electrical Engineering Paris, 75252 Paris, France
[d] SATIE, | Systèmes et Applications des Technologies de l'Information et de l'Energie, ENS Rennes, Univ. de Rennes 1, CNRS, 35170 Bruz, France

HIGHLIGHTS

- Optimally scheduling the cleaning of remote photovoltaic systems is challenging.
- Low cost detection of cleaning interventions helps to decide on further operations.
- 4 machine learning algorithms are applied on data from a remote photovoltaic system.
- Reliability of 95% is reached with 3.5 mHz voltage, current and temperature signals.
- 3 implementation strategies are proposed to meet low cost and accuracy goals.

ARTICLE INFO

Keywords:
Photovoltaics
Soiling
Monitoring
Maintenance
Machine learning
Detection

ABSTRACT

Soiling losses are a major concern for remote power systems that rely on photovoltaic energy. Power loss analysis is efficient for the monitoring of large power plants and for developing an optimal cleaning schedule, but it is not adapted for remote monitoring of standalone photovoltaic systems that are used in rural and poor regions. Indeed, this technique relies on a costly and dirt sensitive irradiance sensor. This paper investigates the possibility of a low-cost monitoring of cleaning interventions on photovoltaic modules during daytime. We believe that it can be helpful to know whether the soiling is regularly removed or not, and to decide if it is necessary to carry out additional cleaning operations. The problem is formulated as a classification task to automatically identify the occurrence of a cleaning intervention using a time window of temperature, voltage and current measurements of a photovoltaic array. We investigate machine learning tools based on Logistic Regression, Support Vector Machines, Artificial Neural Networks and Random Forest to achieve such classification task. In addition, we study the influence of the temporal resolution of the signals and the feature extraction on the classification performance. The experiments are conducted on a real dataset and show promising results with classification accuracy of up to 95%. Based on the results, three implementation strategies addressing different practical needs are proposed. The results may be particularly useful for non-governmental organizations, governments and energy service companies to improve the maintenance level of their photovoltaic facilities.

1. Introduction

In the world, more than 1 billion people still have no access to electricity in 2017 and most of them live in rural areas [1]. The cost of photovoltaic (PV) energy has been decreasing [2]. Being an accessible and relatively clean source of renewable energy [3], a growing number of facilities are adopting it, such as autonomous photovoltaic domestic or communal power plants, municipal lighting, photovoltaic water pumps, photovoltaic fridges [4,5]. The scientific community has a growing interest in finding techno-economic optimum for these [6,7].

However, the full potential of standalone photovoltaic systems (SAPVS) is far from being exploited in these regions, mainly because of the lack of maintenance, which can lead to degraded performance, breakage and abandon if no local capacity is present to repair them

* Corresponding author at: IFSTTAR | COSYS-GRETTIA, 14-20 Boulevard Newton, 77420 Champs-sur-Marne, France.
E-mail address: matthias.heinrich@ens-rennes.fr (M. Heinrich).

https://doi.org/10.1016/j.apenergy.2020.114642
Received 1 October 2019; Received in revised form 25 January 2020; Accepted 8 February 2020

Nomenclature

Symbols

I_{pv}	Current of the photovoltaic array (A)
V_{pv}	Voltage of the photovoltaic array (V)
T_{pv}	Temperature of one photovoltaic module (°C)
T_{amb}	Ambient temperature (°C)
$I_{SC,STC}$	Short-circuit current of the photovoltaic array under standard test conditions (A)
$V_{OC,STC}$	Open-circuit voltage of the photovoltaic array under standard test conditions (V)

Abbreviation

PV	Photovoltaic
MPPT	Maximum Power Point Tracker
LR	Logistic Regression
SVM	Support Machine Vector
RF	Random Forest
ANN	Artificial Neural Network
TS	Time Series
SF	Simple Features
PCA	Principal Component Analysis

[8,9]. In addition, several studies have shown that SAPVS owners in developing countries are likely to overestimate the services that the system is able to provide [10,11] and evidence show that they need social, economic and technical support to use properly and maintain service-oriented photovoltaic installations [12]. Innovative economic approaches of energy service concessions providing preventive and corrective maintenance have been tested in Zambia [13]. However, innovations that increase the reliability and the rate of return - like decreasing the maintenance costs of SAPVS - are still needed [14,15].

Soiling is known to deteriorate severely the performance of photovoltaic modules [16–18] through various mechanisms such as pollution deposition, accumulation of dust or organic particles, or bird droppings that cover the surface of the modules. Soiling effect is particularly impacting in desert areas close to the equator, where photovoltaic modules are only slightly inclined and very exposed to dust deposits, and where rains are rare during the dry season. In these areas, the efficiency of a solar photovoltaic array exposed to dust might decrease down to −6% over one day and −18% over one month without any cleaning of dust deposition [19]. The quantification of soiling is crucial since is allows one to compute economically optimal intervals for cleaning interventions [20,21]. There are currently three approaches to soiling estimation.

The first approach to quantify soiling is through power losses estimation, where the measured power output is compared with the expected power output calculated from the measured irradiance [22]. The second approach consists in measuring directly the soiling rate by comparing the short circuit current of a “soiled” PV cell with the one of a “clean” cell [23]. But for these two cases, the cost of the soiling measurement station and the cost of the sensors to measure the irradiance level, the pollutants concentrations or even atmospheric parameters, are too high to be affordable for small-scale SAPVS. And, the irradiance sensor is also sensitive to dirt, which can affect the effectiveness of the soiling estimation in remote locations. Another possibility could be to use satellite irradiance data as in [24].

The third approach is to model the soiling mechanisms [25] and soiling losses [18,26]. However, the amount of soiling losses is difficult to assess because it depends on several parameters such as the meteorological conditions [27], the nature and the distribution of particles [28], the surface condition of the modules [29] and their tilt angle [30,31]. This is probably why a large number of publications use statistical tools or machine learning algorithms to improve the accuracy of the soiling losses calculation [32]. For instance, Pulipaka et al. [33,34] proposed an Artificial Neural Network (ANN) model that is able to accurately compute the soiling losses and the photovoltaic power output using the irradiance signal and some pollutant characteristics. Massi Pavan et al. [18] used a regression model to calculate the soiling losses using the irradiance and the cell temperature as inputs, and compared it to Bayesian Neural Networks [35] which were found to outperform. ANN have also been used to predict PV soiling as a function of environmental variables such as particulate matter concentration, relative humidity and wind speed [26].

It is noticeable that, fault detection systems recurrently integrate detection of excessive soiling [36]. In [37] ANN is used for the identification of different types of faults in a photovoltaic array that cannot be easily distinguished with If-Then rules. Li et al. [38] implemented an ANN that takes the maximum power point array current and voltage, and the module temperature as inputs and which detects different types of defaults. Fuzzy logic and Neuro-Fuzzy systems have also been implemented for fault diagnostic [39]. It is noticeable that an Adaptive Neuro-Fuzzy Inference System (ANFIS) may perform a better fault classification compared to an ANN [40]. Numerous others fault diagnostic models based on semi supervised learning techniques [41], decision trees [42], Kalman Filter [43], K-Nearest Neighbors [44] already exist. In order to overcome the non-linearity of fault detection problems and the different operating environments of photovoltaic systems, some articles explore other methods to build models that are accurate, reliable and transposable. For instance, Chen et al. used a kernel extreme machine learning algorithm based on current–voltage (I-V) characteristics [45], a deep residual network using current and voltage curves and ambient irradiance and temperature [46], and a Random Forest model using array voltage and string current [47] to achieve a high diagnostic performance with better generalization performance. However, these tools do not separate the effect of soiling from other causes of power losses (e.g. mismatch effects, inverter's power limitation, Maximum Power Point Tracker (MPPT) failures, temperature effect [37]). This makes the scheduling of cleaning interventions difficult.

In the present article, we propose a low-cost monitoring tool for the detection of cleaning interventions on a photovoltaic array. It can be used as a complement to soiling estimation in order to improve the optimal scheduling of cleaning interventions. Indeed, the detection of cleaning interventions can be helpful to know whether the soiling is regularly removed or not, and to decide if it is necessary to carry out additional cleaning operations. This detection will also provide a true picture of the regular care of a remote installation, promoting then *Human in the Loop* strategies for the maintenance of SAPVS. We investigate the feasibility of building a binary classifier for cleaning detection using usually measured electrical signals or low-cost temperature sensors. No precise dating of the cleaning intervention is necessary. We discuss the minimal temporal resolution, the type of inputs and the nature of the algorithm. We decided to use machine learning models to perform this classification task because the modeling of cleaning interventions is complex, in particular due to the variety of cleaning techniques, the complexity of the irradiance dynamics and the diversity of isolated PV installations. In order to explore the complexity of this problem, we study the use of various linear (Logistic Regression and Support Vector Machine) and nonlinear (Artificial Neural Networks, Random Forest) machine learning algorithms which present various levels of complexity and explicability. The classifiers are tested using labelled datasets collected on a remote photovoltaic water pumping system located in a rural village in Burkina Faso.

To the best of our knowledge, this paper is the first one to study the detection of cleaning interventions over photovoltaic modules. Even

though the tool is applied here to a photovoltaic water pumping system, it may be applied to other photovoltaic systems. It could be particularly useful for helping non-governmental organizations, governments and Energy Service Companies (ESCOs) [13] to improve the maintenance level of SAPVS.

The test site is presented in Section 2. The experimental datasets are described in Section 3. The methodology for the classification is described in Section 4. The results of the different classification algorithms are presented and discussed in Section 5.

2. Experimental setup

2.1. Standalone photovoltaic system

The studied photovoltaic water pumping system is installed in Gogma, Burkina Faso (GPS coordinates: latitude 11.724444°; longitude −0.572222°) since January 2018. We described the operation of the system and modeled it in [48]. It consists (see Fig. 1) in:

- a photovoltaic array composed of 3 modules (total peak power of the array: 620 W_p STC);
- a motor-pump (Grundfos SQFlex 5A-7) which includes a maximum power point tracking (MPPT) controlled inverter;
- a controller which starts and stops the motor-pump, with hysteresis control of the water level in the tank;
- a water tank of 11.4 m^3;
- a fountain at which the inhabitants collect water.

An average local yearly horizontal irradiation of 2130 kWh/m^2 has been computed for this location using PVGIS [49]. Each day, around 7 m^3 of drinking water are pumped by this system for the domestic water uses of 280 people of the village. A local “Pump Management Office” has been created; it is in charge of ensuring access to clean drinking water, collecting the user payments, and managing the maintenance. The Pump Management Office hired one inhabitant to regularly clean the modules. He was trained by the team that installed the photovoltaic water pumping system. The cleanings must be more frequent during the dry season (November to May) than during the wet season (April to October) when precipitation contributes to the regular cleaning of the modules. Manual cleaning with water is used because it is a good compromise between costly water jets and low efficiency dry manual cleanings [50]. As water is a low-cost solvent, no demineralized water neither costly detergent is used. A wet smooth mop with water from the photovoltaic water pumping system is used (Fig. 2). The cleaner must climb on a ladder to brush and push the dust from the top to the bottom of the modules.

2.2. Monitoring system and cleaning interventions

A data logger has been monitoring several signals (Table 1). All the signals are measured with a time step of ∼2.2 s and are then rescaled to an evenly spaced temporal resolution of 3 s by linear interpolation.

Cleaning interventions influence the shape and the level of the signals. These variations depend strongly on the pump mode (pumping or standby), and on the instantaneous irradiance. The signals measured during two selected interventions are shown in Fig. 3 and in Fig. 4. The short-circuit current and the open-circuit voltage indicated by the manufacturer are plotted as reference levels of the installation.

In pumping mode (Fig. 3), the PV array current is initially proportional to the irradiance. The cleaning intervention causes a partial shading. As a result, the array voltage and current drop rapidly (−34% and −60% respectively). The significant voltage drop, and its variability are most likely due to the MPPT control in a dynamic partial shading context. The water projections cool down the modules (−28%). This has the effect of slightly increasing the array voltage which is visible just after the end of the cleaning. It should be noted that the irradiance increase during the cleaning intervention has a negligible effect on the array voltage.

In standby mode (Fig. 4), the current is zero. A cleaning intervention decreases the module temperature by −27% ($\Delta T = -14$ °C). King et al. [51] provided a normalized temperature coefficient of the open circuit voltage for mc-Si modules $\beta_{V_{oc}} \approx -0.0042$ °C^{-1}. The measured open circuit voltage increase of $\Delta V_{oc,m} = 6.9\ V$ is in good agreement with the theoretical value $\Delta V_{oc,th} = \beta_{V_{oc}}.\ V_{oc,STC}.\ \Delta T = 6,4\ V$.

The short dip in the measured irradiance is explained by the fact that the person performing the cleaning was asked to clean the irradiance sensor at each cleaning intervention. Naturally the shape of the signals is more complex during cloudy days, but the same trends are observed.

In order to train the detection model, we collected a database of labelled time series. As our study aims at detecting cleaning interventions on the photovoltaic array without using an irradiance sensor, we focus on the three measured signals I_{pv}, V_{pv} and T_{pv}. In agreement with the Pump Management Office, our team managed the cleaning of the photovoltaic modules from the 14th of March 2019 to the 7th of April 2019 (25 days). We planned 80 manual cleanings during that period, which were all performed by a local cleaner that we hired. Cleanings were done by manual wet cleaning with a wet smooth mop using water from the solar pump in the same way as the Pump Management Office cleaner. The date and time were recorded by the cleaner before and after each cleaning intervention. Before labelling the dataset, we double-checked manually, by looking at the data, that a cleaning intervention had indeed taken place during the time interval specified by the cleaner.

The cleanings were performed by series of 8 cleanings each spaced of 20 min, over 10 different days. We suppose that no other cleaning was done, so any other period is assumed to be “no cleaning” period. In order to guarantee the representativeness of the samples used for training of the model, the cleanings were carried out for different meteorological conditions and for different modes of operation of the SAPVS. Cleaning interventions were evenly distributed from 5:00 to 21:00 GMT during the test period. The measurements of 2 cleaning interventions were incomplete due to handling errors during data collection. The voltage and current signals are both zero at night. This prevents us from detecting any cleaning intervention at night using electrical signals. Thus, only cleaning interventions between sunrise 06:00 and sunset 18:00 GMT were finally kept. At the end, 57 cleaning interventions have been retained. Among these, 34 of them were done during pumping mode and the remaining 23 were done during standby mode.

The classifier is designed to determine if a cleaning intervention occurred in a temporal window of observation. Two classes are then considered for each window namely “Cleaning” (C) and “No cleaning” (NC). The maximum duration of a cleaning is about 12 min for the total

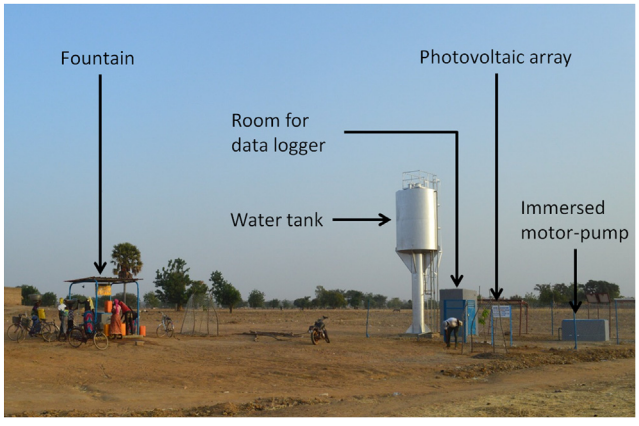


Fig. 1. Overview of the photovoltaic water pumping system in Gogma.

Fig. 2. Cleaning intervention on the photovoltaic array.

Table 1
Monitored signals and sensor characteristics.

Signal	Sensor (resolution)	Sensor location	Usually measured on SAPVS
Photovoltaic array voltage V_{pv}	LEM LV 25P (± 0.9%)	PV array output	Yes (integrated in MPPT)
Photovoltaic array current I_{pv}	LEM LA 55P (± 0.9%)	PV array output	Yes (integrated in MPPT)
Photovoltaic module temperature T_{pv}	PT1000 (± 0.05%)	On the back of one PV module	No
Ambient temperature T_{amb}	PT1000 (± 0.05%)	In the shadow, next to the PV array	No
Irradiance on the plane of the modules G_{pv}	Solems RG100 (± 10%)	On the plane of the PV array	No

array surface of 3.9 m^2, thus a temporal window of 30 min is considered in this article. For each cleaning intervention, 17 datasets were generated by sliding the window of 30 min over the signals by steps of 1 min (Fig. 5) so that every data point when cleaning occurs is included in the window. Our database has thus 969 datasets labelled C. To get a class-balanced learning database, 969 datasets are extracted randomly over the test period, except during the cleaning time range, and from 18:00 to 06:00 and they are labelled NC. Fig. 6 presents the shape of the database. X_{Cij} (respectively X_{NCij}) denotes the element j of dataset i of signal X of the C class (respectively NC class).

Basic descriptive statistics are computed by block over the dataset in order to get a first overview about the frontier between the two classes. The results are shown in Fig. 7. The near-zero level of the current signal medians for NC signals is a result of the intermittent operation of the solar pump. No significant and relevant distinction can be made between C and NC periods, using only voltage and current statistics. One trend that can however be observed is that temperature levels are likely to be lower during cleaning interventions.

In order to have the same range of values for each of the inputs and thus improve convergence of weight and biases of the machine learning models, all values are centered by mean and scaled by the mean of the standard deviations as following;

$$V_{pv}(k,l) \leftarrow \frac{V_{pv}(k,l) - \bar{V}_{pv}}{\sigma_{V_{pv}}}$$

$$I_{pv}(k,l) \leftarrow \frac{I_{pv}(k,l) - \bar{I}_{pv}}{\sigma_{I_{pv}}}$$

$$T_{pv}(k,l) \leftarrow \frac{T_{pv}(k,l) - \bar{T}_{pv}}{\sigma_{T_{pv}}}$$

where σ_X and $\bar{X}$ are respectively the mean of the standard deviations and the mean of all the datapoints of signal X, and $X(k,l)$ is the element at line k, column l from the sub- matrix of the signal X.

3. Methodology

We address the problem of cleaning intervention detection by considering the supervised training of a binary classifier within the framework of machine learning. Fig. 8 summarizes the methodology used for the classification.

3.1. Data selection

Several combinations of input signals are used. All configurations used in this study are listed in Table 2. The temporal resolution of the input signals is directly linked to the frequency of the monitoring. A lower acquisition frequency will reduce the memory and the energy consumption requirements of the monitoring system but might also lead to a loss of information. Hence, we study the performance of the classifiers that use input signals with different temporal resolutions. For this purpose, input signals with low time resolution are obtained by nearest neighbor interpolation of the initial data (Fig. 9).

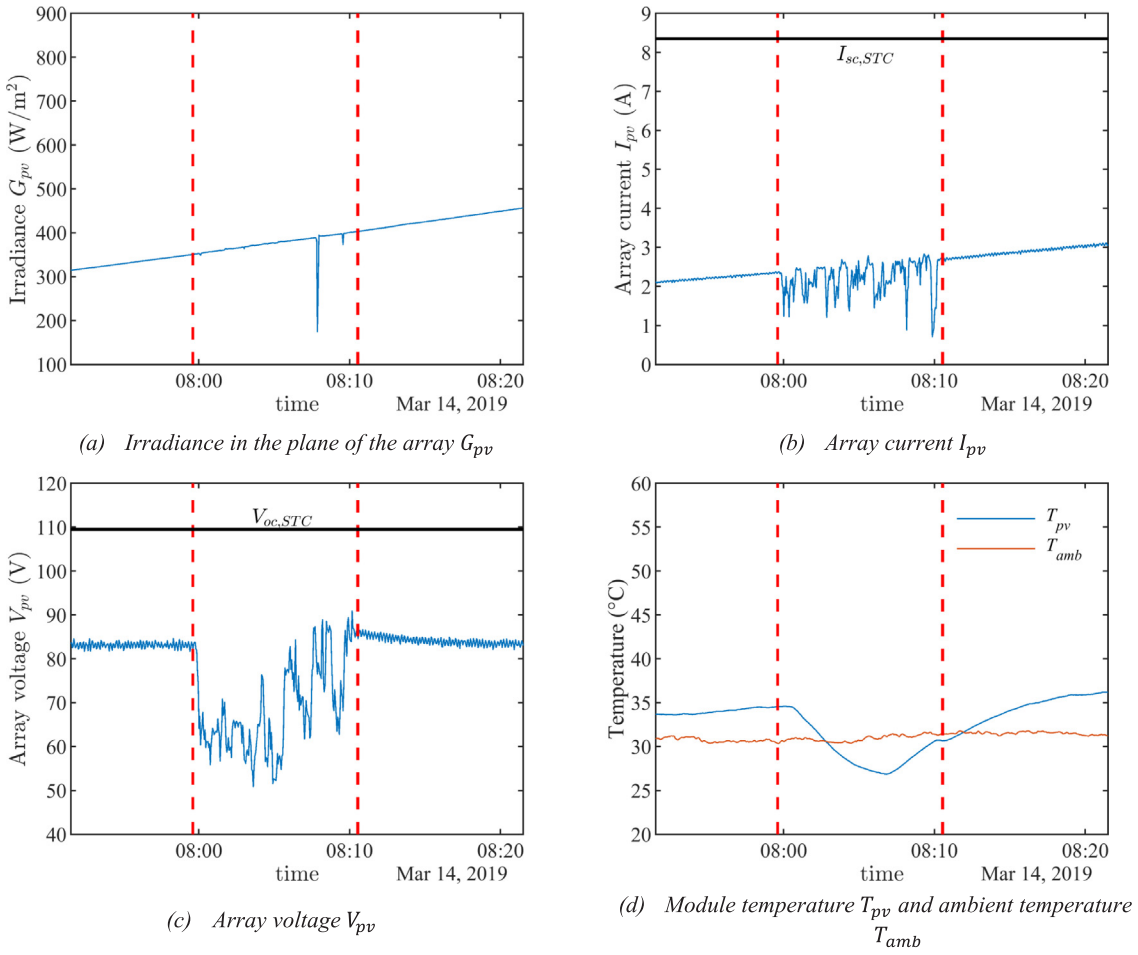
Within the framework of this study, a theoretical limit of the time resolution is given by the Nyquist–Shannon sampling theorem. The maximum duration of a cleaning intervention is 720 s. However, the module temperature keeps on evolving after the cleaning in itself is finished (see Fig. 3), and the temperature signal signature lasts up to 1200 s. Consequently, the Shannon boundary for the electrical signals is 360 s and for the temperature signal is 600 s. We therefore estimate the performances of the classifiers for the following temporal resolutions of the input signals, where we consider some temporal resolutions above the Shannon boundary: [3, 10, 50, 100, 150, 200, 300, 400, 500, 600, 900]s.

3.2. Feature extraction

The nature and the structure of the inputs of a classification model influence the amount of available information, the duration of trainings and the accuracy of the predictions. Three feature extraction methods are tested here:

- Time series (TS) of voltage, current and module temperature can be used as inputs for the classification model. However, if one considers three datasets having duration of 1800 s and a temporal resolution of 3 s, up to $3 \times 600 = 1800$ inputs are provided, which generates excessive computation time for classification.
- Simple Featuring (SF) consists in extracting the five following features from a given dataset: minimum, maximum, average, median, and standard deviation.
- Principal Component Analysis (PCA) can be used to build features which explain at least 95% of the total variance of all the datasets of a given signal. The numbers of features depend on the number, the nature and the temporal resolution of the selected input signals. PCA analysis is computed using the *pca* method from MATLAB [52].

In total, 16 different configurations of inputs signals and feature extraction methods are tested. They are listed in Table 2.



(a) *Irradiance in the plane of the array* G_{pv}

(b) *Array current* I_{pv}

(c) *Array voltage* V_{pv}

(d) *Module temperature* T_{pv} *and ambient temperature* T_{amb}

Fig. 3. Measured signals during a cleaning intervention in pumping mode. The cleaning intervention is delimited by the red dashed lines.

3.3. Model selection and training

We compare the performances of several common binary classifiers based on supervised learning [53,54]. Each has a different level of complexity of implementation and is capable to draw linear or non-linear classification border.

- **Logistic regression (LR)**

Logistic regression is a simple and robust model which is used here as the reference model. This model can handle both continuous and categorical variables. It computes the regression of the logit of conditionnal expectation of a detected cleaning $\mathrm{P}(C|X)$ as a linear combination of explanatory variables $X = (x_i)$ [55]. The decision boundaries are linear.

$$\mathrm{Logit}(\mathrm{P}(\mathrm{C}|\mathrm{X})) = \ln\left(\frac{\mathrm{P}(C|X)}{1-\mathrm{P}(C|X)}\right) = \mathrm{b}_0 + \sum_i \mathrm{b}_\mathrm{i}\, x_i \qquad \mathrm{b}_\mathrm{i} \in \mathbb{R} \tag{1}$$

For this model, the features are continuous variables and we assume that they are explanatory variables. The LR coefficients are computed by using the multinomial logistic regression function *mnrfit* of MATLAB [56]. They are initialized to 0. The model is computed only for SF and PCA features as the TS feature vectors have a high dimension that prevents the algorithm from converging.

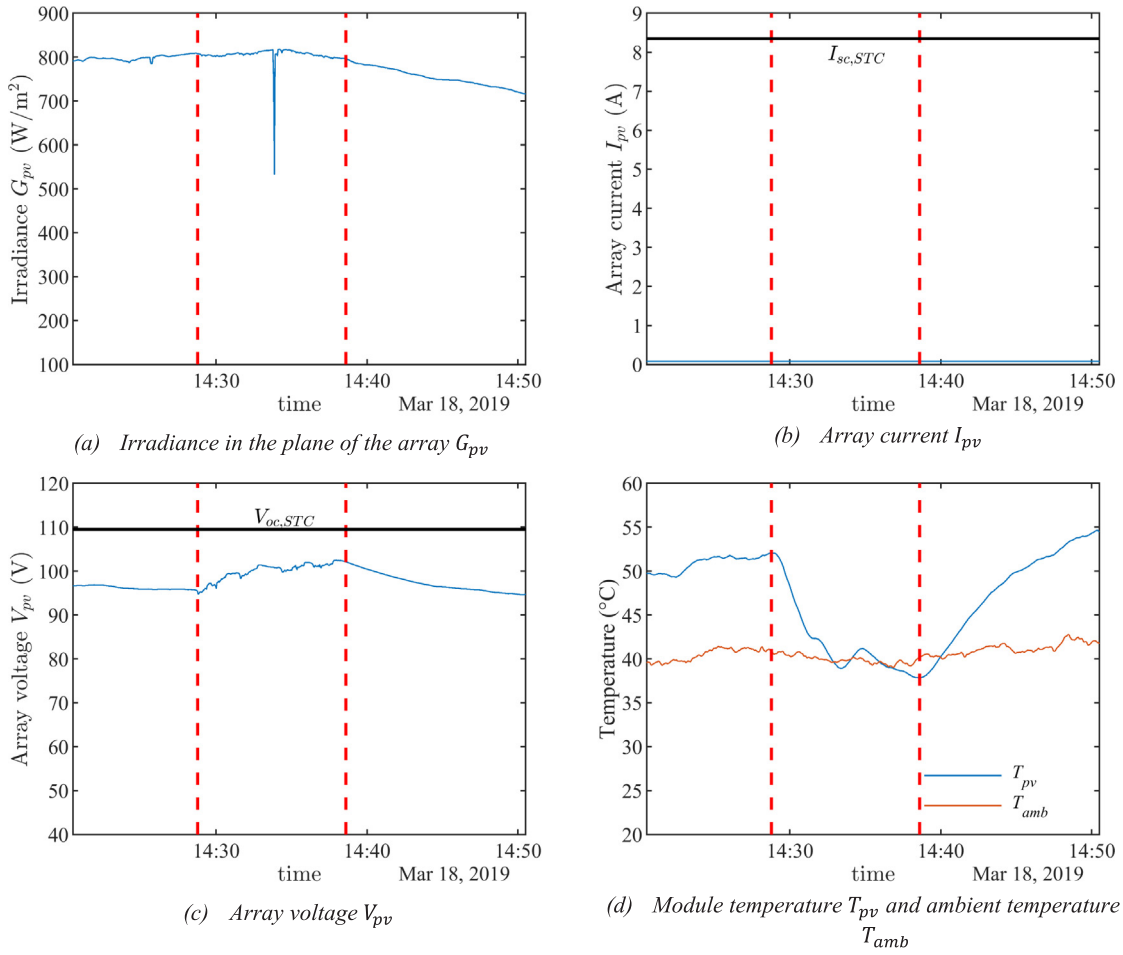
- **Support-Vector Machine (SVM)**

SVM are linear classifiers that are based on the calculation of a maximal margin hyperplane that separates the data into two classes. The data are mapped to a higher dimensional space using a kernel function. In this paper, the SVM is trained using the *fitcsvm* MATLAB function [57] with a Gaussian radial basis function as kernel function [58] with scale factor of 1. A non-linear kernel is used in order to study if the considered features can be projected in a larger space where the problem could become linearly separable and outperform Logistic Regression. The SVM is not trained for TS but only for SF and PCA features for the same reason as for the logistic regression.

- **Artificial Neural Networks (ANN)**

ANN is non linear classification model [53]. ANN can be viewed as a weighted graph where each neuron (node) computes an output thanks to an activation function [59]. In our case, and as a first approach, a simple feedforward structure with a single hidden layer of 100 neurons is used. The number of neurons was chosen by computing the results for neural networks with 5 to 2000 neurons hidden layer, and by keeping a good compromise between classification accuracy and model complexity. A sigmoid function is used as the activation function. Network building, training and testing are carried out using the Deep Learning Toolbox from MATLAB [60]. The data used for training are divided into three subsets: 70% for training, 15% for validating and 15% for testing. The standard scaled conjugate gradient backpropagation method is used for training [61]. Cross entropy is used as the objective function to measure training loss. Initialization of the weight and biases is performed with the Nguyen-Widrow initialization method [62]. The ANN is trained with TS, SF and PCA features.

- **Random Forest (RF)**



(a) *Irradiance in the plane of the array* G_{pv} (b) *Array current* I_{pv}

(c) *Array voltage* V_{pv} (d) *Module temperature* T_{pv} *and ambient temperature* T_{amb}

Fig. 4. Measured signals during a cleaning intervention in standby mode. The cleaning intervention is delimited by the red dashed lines.

Random Forest is a supervised learning technique based on two major ideas that are bagging and random features for training developed by Breiman [63]. The algorithm uses an ensemble of uncorrelated decision trees that are trained over different subsets. Output is computed as the principal mode of the classes given by the individual trees. Averaging the result of the forest results makes the RF more robust and accurate than a simple decision tree. Three main parameters must be defined for the construction of the RF model, including the number of random variables used at each split, the fraction of the original data that is used to build the bootstrap samples and the number of decision trees in the forest. Based on the literature for classification RF models [64], the number of predictors used for training is set to the square root of the total number of variables. Then, each tree is formed by randomly selecting rows with replacement of the total base. Therefore, each bootstrap sample contains on average approximately 2/3 of the observations. Sensitivity results show that accuracy results remain almost the same when sampling randomly over 66% and 80% of the total base. The remaining rows of the training data are called the Out-of-Bag samples. For all the RF built in our work, the number of trees has been set to 100. According to Breiman, a too high number of trees is not an issue since Random Forest does not overfit. This choice is based on the computation of the Out-of-Bag error rates with PCA VIT and SF VIT inputs, which showed that Out-of-Bag error rate does not decrease when adding more trees in these cases. The RF is trained only with SF and PCA features for the same reason as for the logistic regression. The RF algorithm is computed using the *TreeBagger* MATLAB function [65].

Each test configuration is thus entirely defined by a quadruplet {Input signals, Temporal resolution, Feature extraction method, Classification model}. For each quadruplet, 70% of the database is selected randomly for training and the remaining 30% is used for performance evaluation.

3.4. Performance evaluation

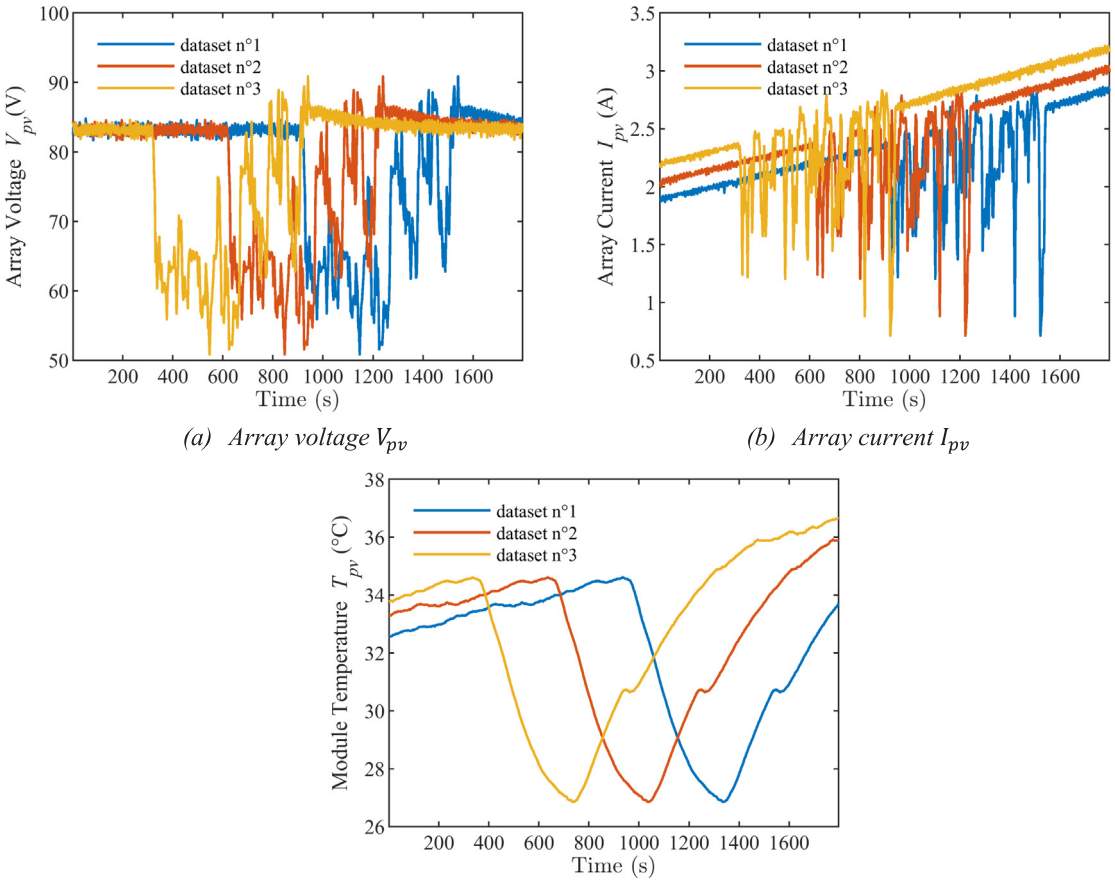
The best quadruplet is supposed to detect both “cleaning” (C) and “no cleaning” (NC) classes without misclassification. A common metric for classification algorithm is the percentage of correctly classified data also known as the classification accuracy (ACC) which is defined as:

$$ACC = \frac{Tr_P + Tr_N}{Tr_P + Tr_N + Fa_P + Fa_N} \quad (2)$$

where true positives (Tr_P) and true negatives (Tr_N) denote the number of correct classifications of positive and negative examples respectively. Positive example denotes here the occurrence of a cleaning intervention and negative the opposite. False positives (Fa_P) represent the number of incorrect classification of negative examples into the positive class, and false negatives (Fa_N) are positive examples incorrectly classified into the negative class.

All the considered binary classifiers provide class membership probabilities. The optimal classification threshold is computed for each quadruplet, at each computation, as the one which minimizes the distance between the receiving operating characteristic (ROC) and the optimal classifier point. In an effort to limit uncertainties ACC is computed 100 times by using random test sample for each quadruplet and the average value $\overline{ACC}$ is kept. The standard deviation of the accuracies is computed in percent of the average accuracy.

Other indicators may be used for classification characterization like recall, F1 score, or ROC curve analysis [53]. However, for the sake of



(a) Array voltage V_{pv} *(b) Array current I_{pv}*

(c) Module temperature T_{pv}

Fig. 5. Example of 3 cleaning intervention datasets. Datasets are generated by sliding a 30 min window over the three signals.

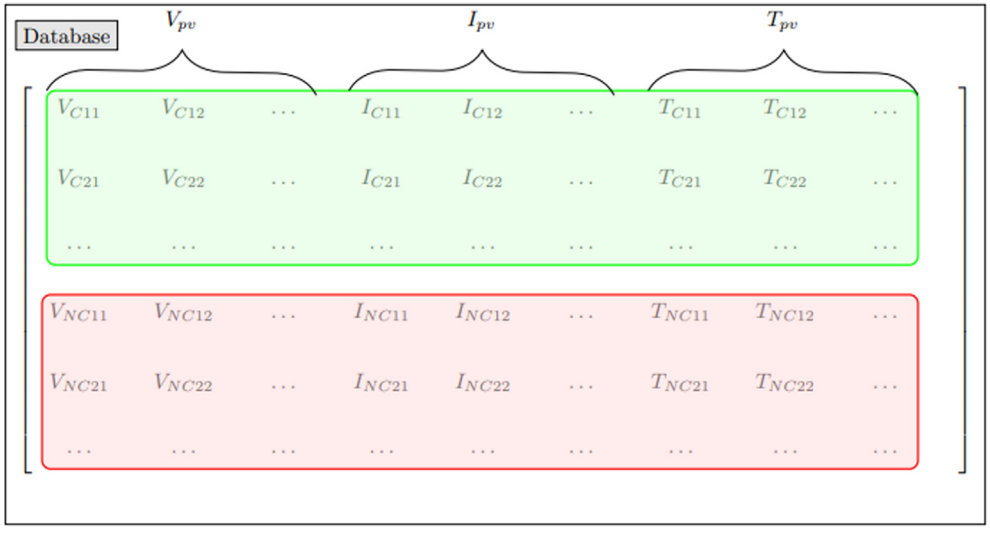


Fig. 6. Structure of the database. Each line contains a dataset generated by a sliding 30 min temporal window. Green (resp. red) block contains datasets belonging to "C" (resp. "NC") class. Each block contains 969 lines with voltage, current and temperature time series.

clarity and consistency only the classification accuracy index is used here.

4. Results and discussion

4.1. Accuracy of cleaning detection

The average classification accuracy $\overline{ACC}$ for each quadruplet {Input signals, Temporal resolution, Feature extraction method, Classification model} is plotted in Fig. 10. The corresponding standard deviations are

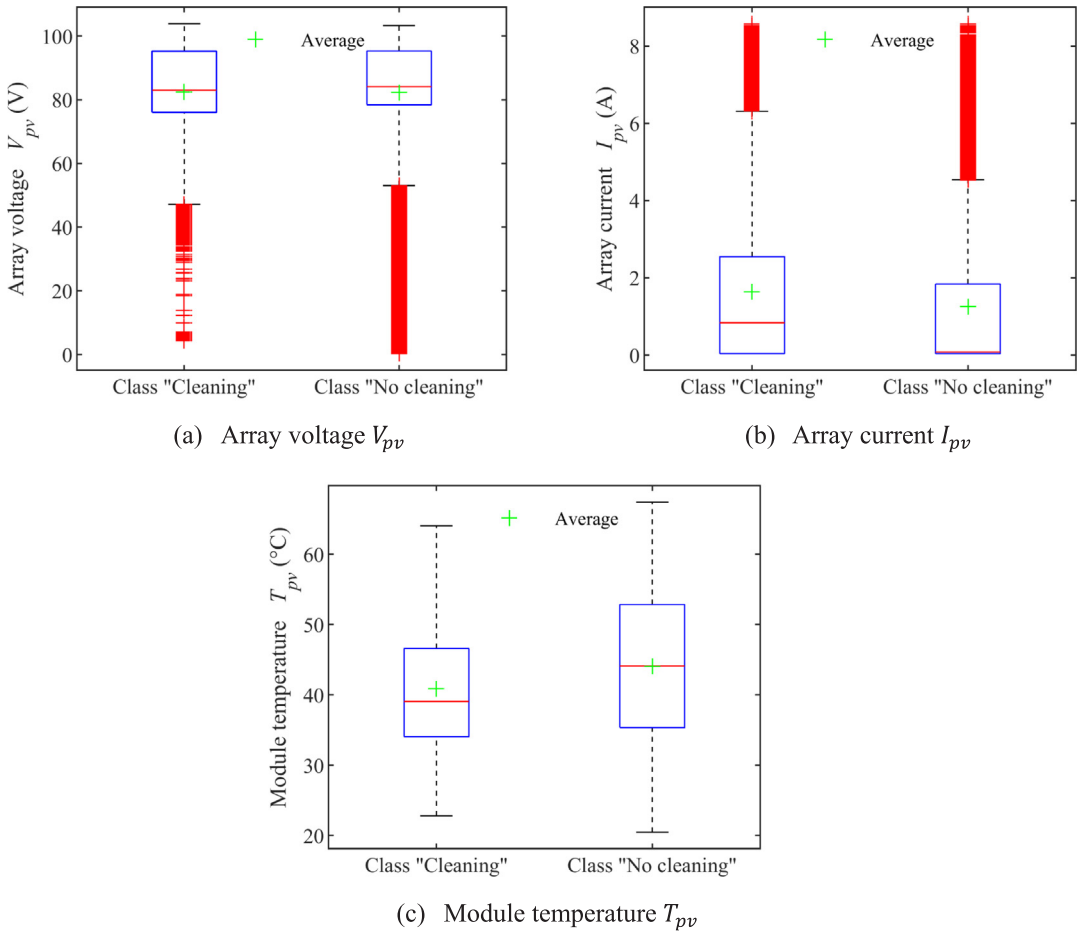


Fig. 7. Basic descriptive statistics for the database. For each diagram, we plot the median and the average. The bottom and the top edges of the box indicate the 25th and 75th percentiles. Outliers are plotted individually.

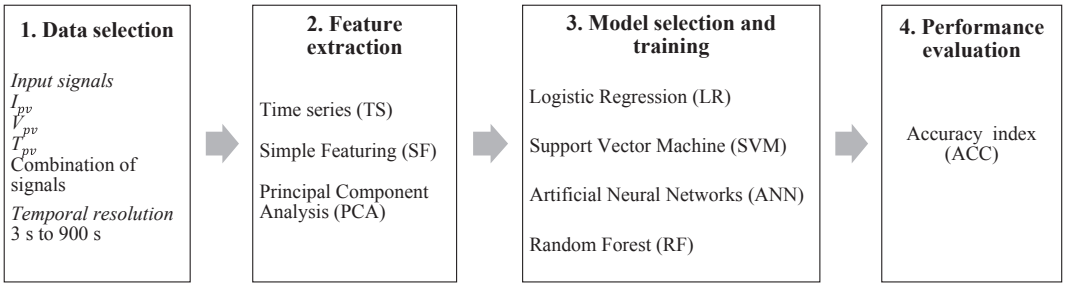


Fig. 8. Methodology used for the classification of the windows of observation.

generally below 2.5%, for temporal resolutions below the Shannon limit. Noteworthy exceptions are quadruplets with current as single input signal, or above the Shannon limit, where standard deviations are observed up to 10%.

Regarding data selection, under the Shannon limit (360 s for the electrical signals and 600 s for the temperature signal), an accuracy of 95% is achievable when using voltage or temperature signal. Current signal seems however to be ineffective to detect cleanings accurately, as in its best configuration an accuracy of only 70% is achieved. This may be due to confusions between cleaning interventions and passing clouds – where irradiance and array current may vary rapidly -, and also to cleanings which occur in standby mode – where the array current is zero. As expected, the accuracy drops with the temporal resolution for all configurations.

Regarding feature extraction methods, TS can be effective with ANN even when the temporal resolution is degraded close to the Shannon boundary. Indeed 80% accuracy for TS V is observed even at a 300 s resolution. SF provides accuracies over 85% when using voltage and temperature signals with SVM, ANN or RF and temporal resolution below 300 s. PCA featuring improves slightly the accuracy and provides robustness for all models to the price of an increased computing time.

Regarding classification models, LR (Fig. 10a) provides accuracy results over 85% for temporal resolution signals below 10 s with SF V and over 90% for temporal resolutions below 300 s for SF T. SVM, ANN and RF always outperform LR. LR and SVM may provide worse results when using combinations of features of different signals. On the contrary, ANN and RF tend to improve their accuracy when combining features from different signals. For instance, at a 150 s resolution SF V, SF VI and SF VIT provide respective accuracies of 85%, 88% and 90% with the RF classification model.

All things considered, RF provides the best classification accuracy for cleaning intervention detection when using a combination of PCA

Table 2
List of the input signals and their associated feature extraction methods. The size of the feature vectors used for classification vary with the temporal resolution of the input signals.

Input signals	Feature extraction	Symbol	Dimension of feature vector (Temporal resolution = 3 s)	Dimension of feature vector (Temporal resolution = 900 s)
V_{pv}	Time Series (TS)	TS V	600	2
I_{pv}		TS I	600	2
T_{pv}		TS T	600	2
I_{pv} & V_{pv}		TS VI	1200	4
T_{pv} & V_{pv}		TS VT	1200	4
I_{pv} & T_{pv}		TS IT	1200	4
V_{pv} & I_{pv} & T_{pv}		TS VIT	1800	6
V_{pv}	Simple Featuring (SF)	SF V	5	5
I_{pv}		SF I	5	5
T_{pv}		SF T	5	5
I_{pv} & V_{pv}		SF VI	10	10
V_{pv} & I_{pv} & T_{pv}		SF VIT	15	15
V_{pv}	Principal component analysis (PCA)	PCA V	11	2
I_{pv}		PCA I	6	2
T_{pv}		PCA T	4	2
V_{pv} & I_{pv} & T_{pv}		PCA VIT	21	2

features of current, voltage and temperature signals (PCA VIT). In this configuration, an average accuracy of 97% is achieved at a temporal resolution of 3 s. This observation underlines the high interest of ensemble learning techniques that use the mode of the classes of the individual models as the output class. It can be guessed that the formulated problem of cleaning detection requires detecting very different types of curves for the same class. This confirms the intuition that the result given by a combination of high number of simple models is likely to be better than the one given by a single model.

4.2. Synthesis & implementation

In the previous section, we compared the accuracy of different quadruplets. But other constraints must be considered for a real-world implementation. Table 3 provides a summary of the different advantages and drawbacks of selecting a given input. The choice of the feature extraction method, the temporal resolution and the classification model might be relative to the case study. Indeed, it could depend on the configuration of the photovoltaic array (bypass diodes, MPPT controller), the duration of the cleaning and the movements done by the cleaner. Table 4 presents a comparison of the classification models for this case study.

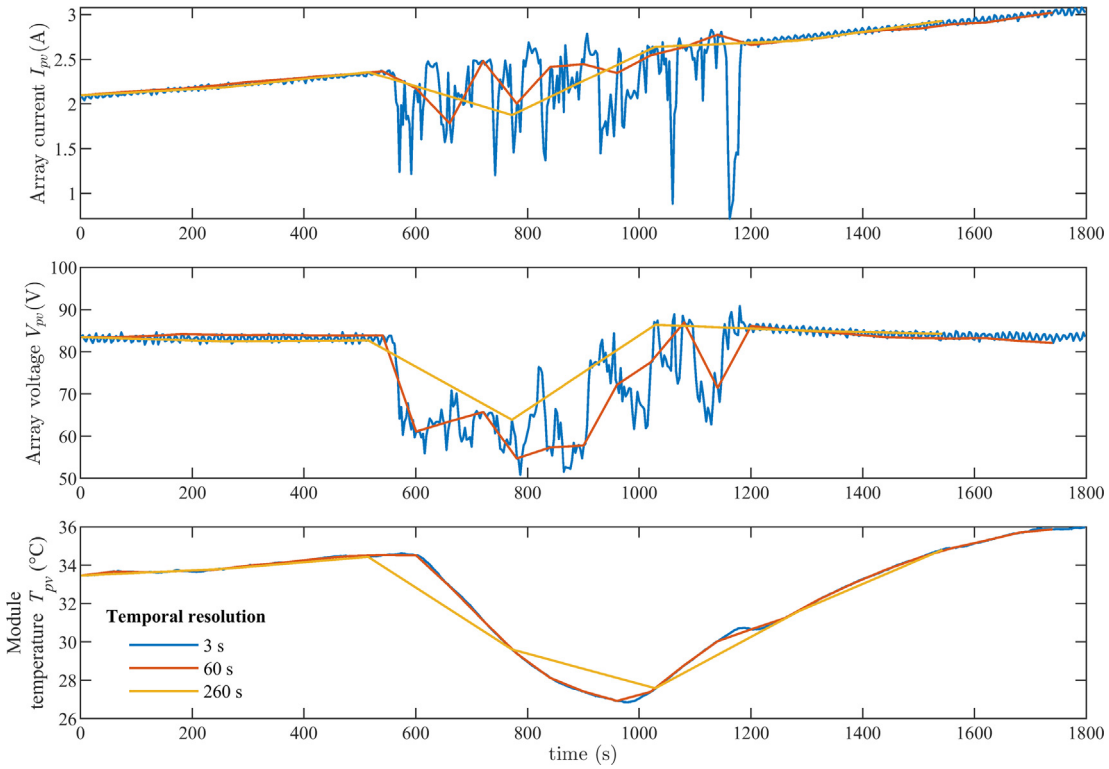


Fig. 9. Example of signals with different time resolutions, during a cleaning intervention. Inputs signals with lower time resolution are generated by nearest neighbor interpolation of the original 3 s resolution signals.

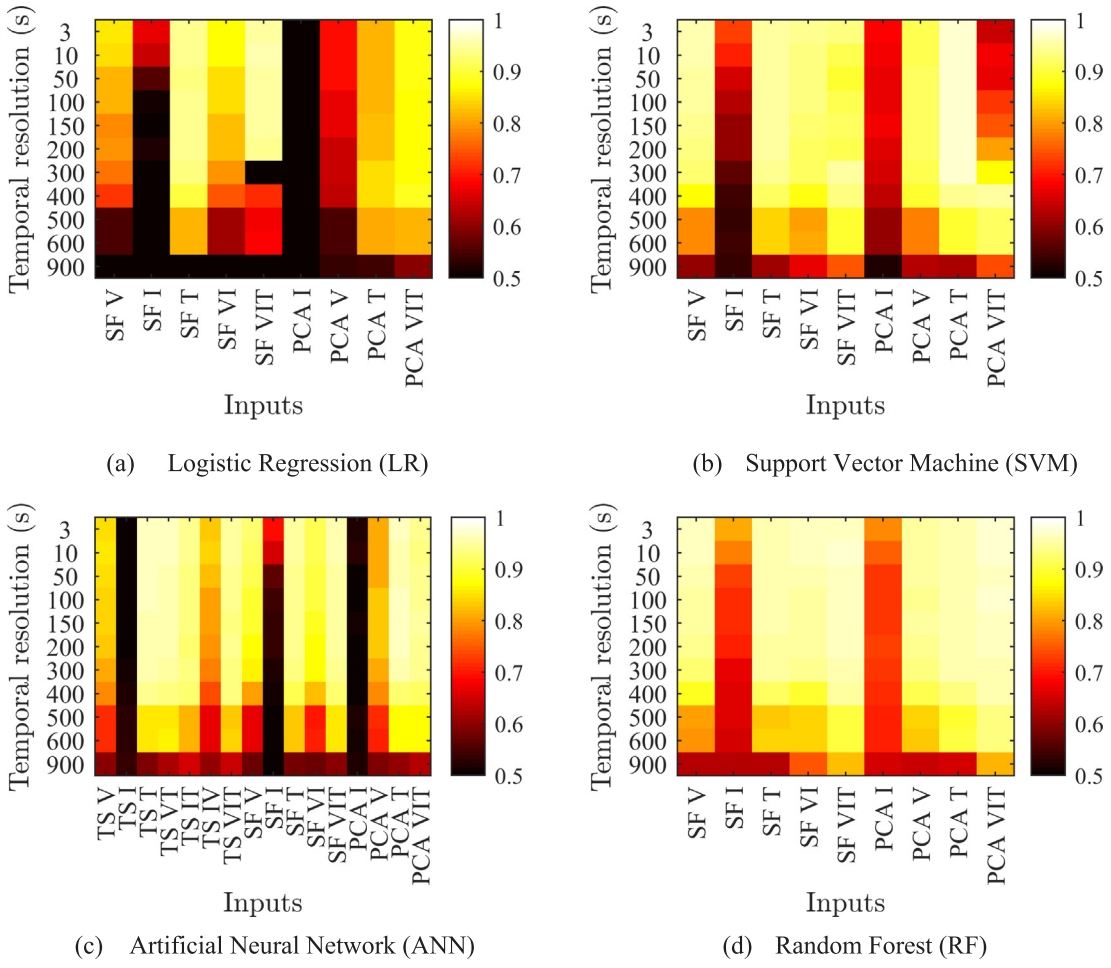


Fig. 10. Average classification accuracy $\overline{ACC}$ for each quadruplet {Input signals, Temporal resolution, Feature extraction method, Classification model} over 100 trainings.

Table 3
Advantages and drawbacks of the different signals for practical use.

Input signals	Advantages	Drawbacks
I_{pv}	Common measure	Low accuracy because of confusions during cloudy days. Useless in standby mode because current is zero. Useless by night because array current is zero.
V_{pv}	Common measure	Useless by night because array voltage is zero.
T_{pv}	Best accuracy for wet cleaning	A temperature sensor must be installed.Assumes that the temperature of one module is representative of the temperature of each module of the array. The cleaner must use enough water.

Table 4
Synthesis of the best-case classification accuracy of the different classification models.

Model	Classification accuracy
LR	Good (> 75%)
SVM	High (> 85%)
ANN	High (> 85%)
RF	Very high (> 90%)

Based on the above analysis, three quadruplets {Input signals, Temporal resolution, Feature extraction method, Classification model} are proposed in Table 5 corresponding to different user objectives. Monitoring of autonomous systems requires a compromise between accuracy of classification, energy consumption and system cost. It is worth underlining that the temporal resolution is tightly linked to the hardware acquisition frequency, and therefore to the power consumption, data storage capacity, data transfer capacity and computing power for online/offline post processing.

Careful considerations are to be made concerning the use of the temperature signal. Firstly, the location and the number of temperature sensors are of primary importance to get a true picture of the photovoltaic array temperature. Secondly, the temperature signal can be used only for wet cleanings. Indeed, several dry cleanings have been carried out and no temperature variation was observed. In any case, dry cleanings are inefficient and are not recommended for SAPVS.

5. Conclusion

Several machine learning models have been implemented and succeeded in detecting wet cleaning interventions during daytime. A classification accuracy of 97% was obtained when using a real database

Table 5
Proposition of quadruplets adapted to different user objectives and associated for this case study.

User objective	Input signals	Temporal resolution (s)	Feature extraction method	Classification model	Average accuracy ($\overline{ACC}$)
High accuracy	V_{pv} & I_{pv} & T_{pv}	10	PCA	RF	96%
Low cost hardware	V_{pv} & I_{pv}	100	SF	RF	84%
Low energy consumption	V_{pv} & I_{pv} & T_{pv}	300	SF	RF	90%

of 1938 labelled datasets. Extracting simple features of the temperature signal alone and using a logistic regression model gives very good detection performances (> 90%) for a temporal resolution lower than 300 s. However, more complex tools such as random forest or artificial neural networks are needed to maintain good performances for a temporal resolution > 300 s. Array current and voltage are often already measured on SAPVS and are likely to provide also good accuracy results (about 85%) for cleaning detection with acquisition time resolution around 100 s. Finally, the combination of simple characteristics of module temperature, array voltage, and array current signals with a random forest (RF) model appears to provide the most accurate classifier (95%). Three different strategies for the on-site implementation were proposed to answer several user objectives.

This demonstrates that only a few signals (array current and voltage, module temperature) must be monitored with a relatively low temporal resolution (300 s i.e. a frequency of 3.5 mHz) to get a good picture of the regular care of a remote photovoltaic installation and of its soiling state. This low-cost tool could help manage the schedule of the cleaning. An interesting novelty of this approach is that it allows including people in a retroactive loop to take care of a system.

The variation of the performance with the climatic conditions, the characteristics of the SAPVS (different PV modules configuration and load), and the cleaning sequence (duration, quantity of water used, efficiency of movement) will notably be studied. Transfer learning algorithms or deep learning techniques could be investigated as interesting pathways to avoid the need for large learning databases. The achieved good accuracy of cleaning intervention detection opens the possibility to monitor the soiling state of a SAPVS (by comparing the efficiency before and after the cleaning). From an application perspective, socio-technical studies would be interesting to better understand local social organizations around SAPVS, especially in terms of routine maintenance. These studies would make it possible to build a sustainable and suitable management scheme, with low environmental impact, which would help communities better maintain PV systems thus increasing their socio-economic impact.

CRediT authorship contribution statement

Matthias Heinrich: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing - original draf, tWriting - review & editing. **Simon Meunier:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Writing - original draft, Writing - review & editing. **Allou Samé:** Methodology, Resources, Validation, Writing - review & editing. **Loïc Quéval:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Visualization, Writing - review & editing. **Arouna Darga:** Funding acquisition, Writing - review & editing. **Latifa Oukhellou:** Methodology, Resources, Validation, Writing - review & editing. **Bernard Multon:** Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the « Investissement d'Avenir » program, through the "IDI 2016" project funded by the IDEX Paris-Saclay, ANR-11-IDEX0003-02. We wish to thank Eric Pilat (CEA-INES) for the enlightening discussions about soiling mechanisms. We would also like to thank Séverin Darga for his patient and ongoing commitment in Gogma, Burkina Faso. We would like to thank the reviewers and the editor for their comments and suggestions which allowed to significantly improve the article.

References

[1] World Bank, Access to electricity, rural (% of rural population) | Data. [Online]. Available: https://data.worldbank.org/indicator/EG.ELC.ACCS.RU.ZS. [Accessed: 19-Sep-2019].
[2] Fraunhofer ISE, Current and future cost of photovoltaics. Long-term scenarios for market development, system prices and LCOE of utility-scale PV systems. Study on behalf of Agora Energiewende, 2015.
[3] Edenhofer O. Intergovernmental Panel on Climate Change, and Working Group 3, Renewable energy sources and climate change mitigation: summary for policy-makers and technical summary: special report of the intergovernmental panel on climate change. New York: Cambridge University Press; 2011.
[4] Semaoui S, Hadj Arab A, Bacha S, Azoui B. The new strategy of energy management for a photovoltaic system without extra intended for remote-housing. Sol Energy 2013;94:71–85. https://doi.org/10.1016/j.solener.2013.04.029.
[5] Modi A, Chaudhuri A, Vijay B, Mathur J. Performance analysis of a solar photovoltaic operated domestic refrigerator. Appl Energy 2009;86(12):2583–91. https://doi.org/10.1016/j.apenergy.2009.04.037.
[6] Zubi G, Dufo-López R, Pasaoglu G, Pardo N. Techno-economic assessment of an off-grid PV system for developing regions to provide electricity for basic domestic needs: A 2020–2040 scenario. Appl Energy 2016;176:309–19. https://doi.org/10.1016/j.apenergy.2016.05.022.
[7] Mazzola S, Astolfi M, Macchi E. The potential role of solid biomass for rural electrification: A techno economic analysis for a hybrid microgrid in India. Appl Energy 2016;169:370–83. https://doi.org/10.1016/j.apenergy.2016.02.051.
[8] Serpa P, Zilles R. The diffusion of photovoltaic technology in traditional communities: the contribution of applied anthropology. Energy Sustain Dev 2007;11(1):78–87. https://doi.org/10.1016/S0973-0826(08)60566-9.
[9] Foley G. Photovoltaic Applications in Rural Areas of the Developing World. World Bank 1995.
[10] Wamukonya N, Davis M. Socio-economic impacts of rural electrification in Namibia: comparisons between grid, solar and unelectrified households. Energy Sustain Dev 2001;5(3):5–13. https://doi.org/10.1016/S0973-0826(08)60272-0.
[11] Blodgett C, Dauenhauer P, Louie H, Kickham L. Accuracy of energy-use surveys in predicting rural mini-grid user consumption. Energy Sustain Dev 2017;41:88–105. https://doi.org/10.1016/j.esd.2017.08.002.
[12] Martinot E, Cabraal A, Mathur S. World Bank/GEF solar home system projects: experiences and lessons learned 1993–2001. Renew Sustain Energy Rev 2001;5(1):39–57. https://doi.org/10.1016/S1364-0321(00)00007-1.
[13] Lemaire X. Fee-for-service companies for rural electrification with photovoltaic systems: The case of Zambia. Energy Sustain Dev 2009;13(1):18–23. https://doi.org/10.1016/j.esd.2009.01.001.
[14] Akinyele DO, Rayudu RK, Nair NKC. Development of photovoltaic power plant for remote residential applications: The socio-technical and economic perspectives. Appl Energy 2015;155:131–49. https://doi.org/10.1016/j.apenergy.2015.05.091.
[15] Ahlborg H, Hammar L. Drivers and barriers to rural electrification in tanzania and mozambique - grid extension, off-grid and renewable energy sources. World Renewable Energy Congress, Sweden, Linköping 2011. p. 2493–500. https://doi.org/10.3384/ecp110572493.
[16] Sayyah A, Horenstein MN, Mazumder MK. Energy yield loss caused by dust deposition on photovoltaic panels. Sol Energy Sep. 2014;107:576–604. https://doi.org/10.1016/j.solener.2014.05.030.
[17] Said SAM. Effects of dust accumulation on performances of thermal and photovoltaic flat-plate collectors. Appl Energy Jan. 1990;37(1):73–84. https://doi.org/10.1016/0306-2619(90)90019-A.
[18] Massi Pavan A, Mellit A, De Pieri D. The effect of soiling on energy production for large-scale photovoltaic plants. Sol Energy May 2011;85(5):1128–36. https://doi.org/10.1016/j.solener.2011.03.006.

[19] Saidan M, Albaali AG, Alasis E, Kaldellis JK. Experimental study on the effect of dust deposition on solar photovoltaic panels in desert environment. Renew Energy 2016;92:499–505. https://doi.org/10.1016/j.renene.2016.02.031.

[20] Hammad B, Al-Abed M, Al-Ghandoor A, Al-Sardeah A, Al-Bashir A. Modeling and analysis of dust and temperature effects on photovoltaic systems' performance and optimal cleaning frequency: Jordan case study. Renew Sustain Energy Rev 2018;82:2218–34. https://doi.org/10.1016/j.rser.2017.08.070.

[21] Jones RK, et al. Optimized cleaning cost and schedule based on observed soiling conditions for photovoltaic plants in central Saudi Arabia. IEEE J Photovolt 2016;6(3):730–8. https://doi.org/10.1109/JPHOTOV.2016.2535308.

[22] Chouder A, Silvestre S. Automatic supervision and fault detection of PV systems based on power losses analysis. Energy Convers Manag 2010;51(10):1929–37. https://doi.org/10.1016/j.enconman.2010.02.025.

[23] Gostein M, Düster T, Thuman C. Accurately measuring PV soiling losses with soiling station employing module power measurements, in: presented at the IEEE 42nd Photovoltaic Specialist Conference (PVSC), New Orleans, USA, 2015, pp. 1–4, http://doi.org/10.1109/PVSC.2015.7355993.

[24] Drews A, et al. Monitoring and remote failure detection of grid-connected PV systems based on satellite observations. Sol Energy 2007;81(4):548–64. https://doi.org/10.1016/j.solener.2006.06.019.

[25] Darwish ZA, Kazem HA, Sopian K, Al-Goul MA, Alawadhi H. Effect of dust pollutant type on photovoltaic performance. Renew Sustain Energy Rev 2015;41:735–44. https://doi.org/10.1016/j.rser.2014.08.068.

[26] Javed W, Guo B, Figgis B. Modeling of photovoltaic soiling loss asa function of environmental variables. Sol Energy 2017;157:397–407. https://doi.org/10.1016/j.solener.2017.08.046.

[27] Goossens D, Van Kerschaever E. Aeolian dust deposition on photovoltaic solar cells: the effects of wind velocity and airborne dust concentration on cell performance. Sol Energy 1999;66(4):277–89. https://doi.org/10.1016/S0038-092X(99)00028-6.

[28] Maghami MR, Hizam H, Gomes C, Radzi MA, Rezadad MI, Hajighorbani S. Power loss due to soiling on solar panel: A review. Renew Sustain Energy Rev 2016;59:1307–16. https://doi.org/10.1016/j.rser.2016.01.044.

[29] Piliougine M, et al. Comparative analysis of energy produced by photovoltaic modules with anti-soiling coated surface in arid climates. Appl Energy 2013;112:626–34. https://doi.org/10.1016/j.apenergy.2013.01.048.

[30] Cano J, John JJ, Tatapudi S, TamizhMani G. Effect of tilt angle on soiling of photovoltaic modules, in: presented at the IEEE 40th Photovoltaic Specialist Conference (PVSC), Denver, USA, 2014, pp. 3174–3176, http://doi.org/10.1109/PVSC.2014.6925610.

[31] García M, Marroyo L, Lorenzo E, Pérez M. Soiling and other optical losses in solar-tracking PV plants in navarra. Prog Photovolt Res Appl 2011;19(2):211–7. https://doi.org/10.1002/pip.1004.

[32] Mellit A, Kalogirou SA. Artificial intelligence techniques for photovoltaic applications: A review. Prog Energy Combust Sci 2008;34(5):574–632. https://doi.org/10.1016/j.pecs.2008.01.001.

[33] Pulipaka S, Mani F, Kumar R. Modeling of soiled PV module with neural networks and regression using particle size composition. Sol Energy 2016;123:116–26. https://doi.org/10.1016/j.solener.2015.11.012.

[34] Pulipaka S, Kumar R. Power prediction of soiled PV module with neural networks using hybrid data clustering and division techniques. Sol Energy 2016;133:485–500. https://doi.org/10.1016/j.solener.2016.04.004.

[35] Massi Pavan A, Mellit A, De Pieri D, Kalogirou SA. A comparison between BNN and regression polynomial methods for the evaluation of the effect of soiling in large scale photovoltaic plants. Appl Energy 2013;108:392–401. https://doi.org/10.1016/j.apenergy.2013.03.023.

[36] Mellit A, Tina GM, Kalogirou SA. Fault detection and diagnosis methods for photovoltaic systems: A review. Renew Sustain Energy Rev 2018;91:1–17. https://doi.org/10.1016/j.rser.2018.03.062.

[37] Chine W, Mellit A, Lughi V, Malek A, Sulligoi G, Massi Pavan A. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renew Energy 2016;90:501–12. https://doi.org/10.1016/j.renene.2016.01.036.

[38] Li Z, Wang Y, Zhou D, Wu C. An intelligent method for fault diagnosis in photovoltaic array. System Simulation and Scientific Computing. 2012. p. 10–6.

[39] Bonsignore L, Davarifar M, Rabhi A, Tina GM, Elhajjaji A. Neuro-fuzzy fault detection method for photovoltaic systems. Energy Procedia 2014;62:431–41. https://doi.org/10.1016/j.egypro.2014.12.405.

[40] Belaout A, Krim F, Mellit A, Talbi B, Arabi A. Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification. Renew Energy 2018;127:548–58. https://doi.org/10.1016/j.renene.2018.05.008.

[41] Zhao Y, Ball R, Mosesian J, de Palma J, Lehman B. Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. IEEE Trans Power Electron 2015;30(5):2848–58. https://doi.org/10.1109/TPEL.2014.2364203.

[42] Zhao Y, Yang L, Lehman B, de Palma J, Mosesian J, Lyons R. Decision tree-based fault detection and classification in solar photovoltaic arrays, in: presented at the 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), Orlando, USA, pp. 93–99, http://doi.org/10.1109/APEC.2012.6165803.

[43] Kang B-K, Kim S-T, Bae S-H, Park J-W. Diagnosis of output power lowering in a PV array by using the Kalman-filter algorithm. IEEE Trans Energy Convers 2012;27(4):885–94. https://doi.org/10.1109/TEC.2012.2217144.

[44] Harrou F, Taghezouit B, Sun Y. Improved kNN-based monitoring schemes for detecting faults in PV systems. IEEE J Photovolt 2019;9(3):811–21. https://doi.org/10.1109/JPHOTOV.2019.2896652.

[45] Chen Z, Wu L, Cheng S, Lin P, Wu Y, Lin W. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics. Appl Energy 2017;204:912–31. https://doi.org/10.1016/j.apenergy.2017.05.034.

[46] Chen Z, Chen Y, Wu L, Cheng S, Lin P. Deep residual network based fault detection and diagnosis of photovoltaic arrays using current-voltage curves and ambient conditions. Energy Convers Manag 2019;198:111793https://doi.org/10.1016/j.enconman.2019.111793.

[47] Chen Z, et al. Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. Energy Convers Manag 2018;178:250–64. https://doi.org/10.1016/j.enconman.2018.10.040.

[48] Meunier S, et al. A validated model of a photovoltaic water pumping system for off-grid rural communities. Appl Energy 2019;241:580–91. https://doi.org/10.1016/j.apenergy.2019.03.035.

[49] Joint Research Centre, Photovoltaic geographical information system. [Online]. Available: https://re.jrc.ec.europa.eu. [Accessed: 13-Jul-2019].

[50] Sarver T, Al-Qaraghuli A, Kazmerski LL. A comprehensive review of the impact of dust on the use of solar energy: History, investigations, results, literature, and mitigation approaches. Renew Sustain Energy Rev 2013;22:698–733. https://doi.org/10.1016/j.rser.2012.12.065.

[51] King DL, Kratochvil JA, Boyson WE. Temperature coefficients for PV modules and arrays: measurement methods, difficulties, and results, in: presented at the Conference Record of the Twenty Sixth IEEE Photovoltaic Specialists Conference, Anaheim, USA, 1997, pp. 1183–1186, http://doi.org/10.1109/PVSC.1997.654300.

[52] MathWorks - Principal component analysis of raw data - MATLAB pca. [Online]. Available: https://fr.mathworks.com/help/stats/pca.html. [Accessed: 07-Sep-2019].

[53] Bishop CM. Pattern recognition and machine learning. Springer; 2006.

[54] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning vol. 103. New York, NY: Springer, New York; 2013.

[55] Jr DWH, Lemeshow S, Sturdivant RX. Applied Logistic Regression. John Wiley & Sons; 2013.

[56] MathWorks - Multinomial logistic regression- MATLAB mnrfit. [Online]. Available: https://www.mathworks.com/help/stats/mnrfit.html. [Accessed: 07-Sep-2019].

[57] MathWorks - Train support vector machine (SVM) classifier for one-class and binary classification - MATLAB fitcsvm. [Online]. Available: https://www.mathworks.com/help/stats/fitcsvm.html. [Accessed: 07-Sep-2019].

[58] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media; 2013.

[59] Jain AK, Mohiuddin KM. Artificial neural networks: a tutorial. Computer 1996;29(3):31–44. https://doi.org/10.1109/2.485891.

[60] MathWorks - Deep Learning Toolbox - MATLAB. [Online]. Available: https://www.mathworks.com/products/deep-learning.html. [Accessed: 07-Sep-2019].

[61] Møller MF. A scaled conjugate gradient algorithm for fast supervised learning, p. 21.

[62] Nguyen-Widrow layer initialization function - MATLAB initnw - MathWorks France. [Online]. Available: https://fr.mathworks.com/help/deeplearning/ref/initnw.html. [Accessed: 19-Jan-2020].

[63] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[64] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer series in statistics; 2009.

[65] MathWorks- Create bag of decision trees - MATLAB. [Online]. Available: https://www.mathworks.com/help/stats/treebagger.html. [Accessed: 07-Sep-2019].