# An ensemble learning framework for snail trail fault detection and diagnosis in photovoltaic modules

Edgar Hernando Sepúlveda-Oviedo, Louise Travé-Massuyès, Audine Subias, Marko Pavlov, Corinne Alonso

Engineering Applications of Artificial Intelligence 137 (2024) 109068

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

# An ensemble learning framework for snail trail fault detection and diagnosis in photovoltaic modules

Edgar Hernando Sepúlveda-Oviedo [a,b,*], Louise Travé-Massuyès [a], Audine Subias [a], Marko Pavlov [b], Corinne Alonso [a]

[a] LAAS-CNRS, Université de Toulouse, CNRS, INSA, UPS, Toulouse, France
[b] Feedgy, Paris, France

ARTICLE INFO

*Keywords:*
Photovoltaic fault diagnosis
Ensemble learning
Support vector machines
K-nearest neighbors
Decision trees
Time–frequency feature selection

ABSTRACT

This research proposes a method for detecting subtle faults named *snail trails* for their visual similarity with the trail of a snail in photovoltaic modules. *Snail trails* do not significantly reduce panel performance but they are the main cause of serious panel deterioration such as microcracks and delamination and can go so far as to set the panel on fire. To detect these faults, this research uses an ensemble learning framework, named ensemble learning for diagnosis, which combines several complementary learning algorithms, namely Support Vector Machines, K-Nearest Neighbors, and Decision Trees. A set of features is obtained by extracting the time–frequency characteristics and statistics from the photovoltaic current signal of the photovoltaic panel. This is followed by a feature selection and dimensionality reduction step that delivers the input to the learning algorithms. The approach presented in this study is experimentally validated, independently for the 4 seasons of the year, with data from a real photovoltaic string of 16 panels. The results demonstrate that the proposed approach can efficiently classify healthy panels and panels with *snail trails* efficiently. Interestingly, the method only requires the electrical current signal, measured on panels with data acquisition systems that are standard in the photovoltaic industry. The genericity of the approach makes it a good candidate for detecting other photovoltaic faults and for solving diagnosis problems in other domains.

## 1. Introduction

The development of the Solar photovoltaic (PV) industry has been reinforced since the 1970s with the fossil energy crisis and the need to operate a drastic energy transition. The National Renewable Energy Laboratory (NREL) predicts that 231 $GW$ of PV will be installed in 2023 and according to GlobalData the installed PV capacity worldwide will exceed 1,500 GW in 2030. To guarantee continuity in the production of energy from these PV systems, the development of new sensors with more PV conversion efficient materials and innovative technologies has been promoted for the construction of PV cells. However, despite these progresses, PV systems still present today strong problems of production continuity linked to frequent occurrence of faults in their components (protection systems, box junction, inverter, PV generator, etc.). Facing the complexity of a PV power plant, it is not easy to detect which component is faulty even though power losses can be as bad as 17.5% of the total power output (Dhere and Shiradkar, 2012).

Complete PV systems are composed of multiple components to enable the delivered power to loads or grids as switching converters, DC and AC fuses, all of which can suffer faults. Nevertheless, the studies have identified that most faults occur at the PV panel level such as corrosion, delamination, hot spots, among others (Santhakumari and Sagar, 2019). Focusing on the PV generator and its basic units, namely PV cells, responsible for the transformation of sunlight into electrical energy, the main faults are linked to cell cracks, discoloration, *snail trails* or delamination (Jordan et al., 2017). It is for this reason that this study focuses on fault detection in PV panels rather than in strings and PV complete systems.

In the field of fault detection applied to PV systems, multiple methods have been proposed, ranging from conventional methods based on visual inspection, image analysis, electrical detection to more recent methods based on machine learning (Sepúlveda-Oviedo et al., 2023a). However, most of these works have focused on detecting faults whose energy reduction in the PV system is critical. Damages caused by these faults induce the reduction of the generated power and can even cause a complete shutdown of the system. However, few works have been concerned with detecting faults whose electrical signature remains similar to that of a PV panel without fault (Sepúlveda-Oviedo et al., 2022). *Snail trail* faults are in this category (see Fig. 1).

* Corresponding author at: LAAS-CNRS, Université de Toulouse, CNRS, INSA, UPS, Toulouse, France.
*E-mail address:* ehsepulved@laas.fr (E.H. Sepúlveda-Oviedo).

https://doi.org/10.1016/j.engappai.2024.109068
Received 3 May 2023; Received in revised form 22 July 2024; Accepted 28 July 2024
Available online 21 August 2024

**Nomenclature**

| | |
|---|---|
| $DT$ | Decision Trees |
| $DWT$ | Discrete Wavelet Decomposition |
| $EL$ | Ensemble Learning |
| $FT$ | Fourier Transform |
| $Isomap$ | Isometric Mapping |
| $kNN$ | K-Nearest Neighbor |
| $MSD$ | Multiresolution Signal Decomposition |
| $MV$ | Majority voting |
| $PCA$ | Principal Component Analysis |
| $PSD$ | Power Spectral Density |
| $SVM$ | Support Vector Machine |

Our article aims to demonstrate the following hypothesis:

- Snail fault detection can be achieved with little data and without requiring additional imaging or instrumentation beyond existing components in contemporary solar power plants.

### 1.1. Snail trails fault

Snail trails (also named as snail tracks or worm marks) are lines of local discoloration that occur on PV panels after long-term usage. These brown or black lines appear near bus bars on solar edges or close to microcracks (Li, 2021). The name of this effect originates from the illusion like snails or worms have passed over the PV panel. The exact cause of this phenomenon remains unclear, but some reports suggest it may be associated with silver particles containing sulfur, phosphorus, or carbon (Li, 2021). Other studies indicate that factors such as cell cracks, additives in the EVA film, chemical agents applied to the cell surface, or humidity can expedite the formation of *snail trails* (Kim et al., 2016). The progression of this fault can be slow and may propagate through the PV cell or, in some cases, stabilize after its initial appearance (Köntges et al., 2014). Although they do not cause a significant energy loss, they can trigger other faults that may ultimately lead to fires in the PV plant, as reported in Kim et al. (2016). Furthermore, as indicated in Li (2021), a *snail trail* fault can result in localized temperature increases (hot spots), non-uniform degradation, or corrosion, all of which can have a severe impact on power production.[1]

### 1.2. Motivation

The detection of *snail trail* faults has recently been explored using machine learning methods. However, the proposed approaches are primarily based on image analysis. This is reasonable on a small scale due to its apparent visual signature, as observed in Fig. 1.

However, conducting such large-scale detections in high-power solar plants becomes impractical due to the need for drones for image capture, multiple specific conditions to ensure image capture under the same training conditions, and extensive training databases. Moreover, it is essential to capture images rapidly for comparative purposes, as irradiation conditions exhibit rapid diurnal variations. This dynamic nature of conditions makes it challenging to compare images taken at different time points, even if they originate from the same PV panel. All of this exponentially increases the cost of the detection solution, and a method that is hardly scalable to large-scale industrial environments. Therefore, detecting such faults without additional sensors or instrumentation has been the motivation behind this study.

**Fig. 1.** Example of *snail trail* fault. Note the visual similarity with the trail of a snail.

To address this issue, this study proposes the extraction of features to identify such faults from the electrical current signal of the PV system. This signal is already routinely collected for monitoring in most of these plants. Therefore, the detection system proposed in this study is capable of operating without the need for additional sensors or instrumentation. However, it is evident that the complexity of detecting this type of fault using an electrical signal is higher compared to approaches that use images. To illustrate the challenge of detecting these types of faults using only electrical signals compared to faults with critical energy reduction, Fig. 2 presents the current signal of a panel with a fault resulting in a significant power loss, specifically a Partial Shading fault, a *snail trail* fault, and a healthy panel.

As it can be seen in Fig. 2, achieving fault detection with large power loss using the current signal can be an easy task using conventional machine learning algorithms as the difference with healthy current is important. However, managing to detect almost unnoticeable flaws due to the *snail trail* fault is really a challenge. It is important to emphasize that the sooner a fault such as a *snail trail* is detected and classified, in order to carry out maintenance on the defective part, the greater the production guarantee. This means that *snail trail* detection has a heavy impact on the useful life of the PV plants and the cost of maintenance. Moreover, it is crucial to highlight that a machine learning approach can overcome the limitations of conventional methods such as visual inspection (which is highly dependent on the observer) and traditional laboratory methods that require the system to be completely disconnected and production to be stopped in order to perform diagnostics.

### 1.3. Contributions

The Ensemble Learning (EL) framework, named ELDIAG, proposed in this study uses only the electric current signal linked to Maximum Power Point (MPP) obtained with a manufactured optimizer detailed in Section 3. Importantly, the measure of the current does not require cutting the operative electrical production of the PV string. Fault signatures are extracted through techniques inspired by time–frequency signal processing and statistical analysis. A feature selection and dimensionality reduction step using two algorithms, Principal Component Analysis (PCA) and Isometric Mapping (Isomap), then delivers the input to the learning algorithms. ELDIAG is experimentally validated, independently for the 4 seasons of the year, with data from a real PV string of 16 panels.

The main contribution of this study is a generic ensemble learning framework for fault detection and diagnosis that leverages the complementarities of three machine learning algorithms, Support Vector Machines, K-Nearest Neighbors and Decision Trees as developed in Section 3.2. The learning step is efficiently organized by temporal slices with characteristic variations.

The impacts of the work are the following:

[1] For more detailed information on this type of fault, please refer to the following Refs. Li (2021).

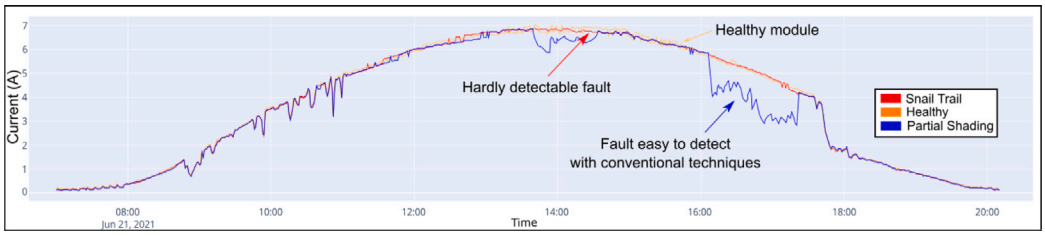


**Fig. 2.** Example of current of three PV panels with health states healthy (orange), Partial Shading (blue) and *snail trail* (red). This data was captured on June 21, 2021.

- *Snail trail* fault detection is achieved based on the MPP current sensor only and it does not require to cut the operative electrical production of the PV string,
- Detection is efficient even in low irradiation conditions,
- ELDIAG can be readily integrated into microprocessors, pre-existing devices within monitoring systems, or electrical interfaces such as inverters. It has been implemented and tested within a monitoring hardware device developed by the company Feedgy Solar, a company specializing in solar energy solutions.
- ELDIAG is generic and can extend to faults detectable from time series data.

The structure of the study is as follows: Section 2 presents the related work and methodology. Section 3 introduces the data acquisition and pre-processing. Section 4 and Section 5 are dedicated to the feature extraction method and the feature selection method respectively. Section 6 presents the classification algorithms, promising research topics and challenges. Section 7 gives the results and discussions. Finally, Section 8 provides conclusions and research prospects.

## 2. Related work

In the last years, multiple fault detection works have been proposed in PV systems, for instance, using neural networks (Hong and Pula, 2022) or fuzzy systems (Xu et al., 2022). Some works are based on the principle of Model-Based Difference Measurement, where the measured value of the variable (current, voltage, power, irradiation, etc.) is compared with the value predicted by a statistical model (Garoudja et al., 2017), others are based on image processing (Amaral et al., 2021) or visual inspection (Packard et al., 2012). However, they have problems to be updated or trained, or they have large components of subjectivity and/or a high cost of implementation.

Other authors have proposed approaches based on machine learning (Sepúlveda-Oviedo et al., 2023a). Using unsupervised methods, the authors have proposed approaches based on Fuzzy C-means (Xu et al., 2022) or hierarchical clustering (Sepúlveda-Oviedo et al., 2021). However, because they do not have labels, these systems are limited in differentiating or identifying incipient faults or electrical behavior very similar to that of a panel without a fault.

Alternatively, and recently, other authors have used supervised learning. In this learning scheme it is necessary to have a database with inputs (predictors) and outputs (labels or targets). These algorithms try to discover the relationship between inputs and outputs. They first produce a function that assigns data to labels. Then, this function is used to predict the label of new unlabeled data. In this type of machine learning, many approaches have been tested in the detection of faults in PV systems (Sepúlveda-Oviedo et al., 2023a) such as Support Vector Machines (SVM), decision trees, neural networks, Local outlier factor, Naive Bayes Classifier, also deep learning.

Specifically, to our knowledge, in the detection of *snail trail* faults, there are not numerous works. The few works addressing this issue can be divided into two major categories. The first category, encompassing the majority of *snail trail* fault detection studies, focuses on utilizing various artificial intelligence approaches, primarily deep learning, on thermal images, electroluminescence, RGB, etc. The second category, which has been growing in the literature, consists of articles whose approach works on electrical signals for *snail trail* fault detection. To position our study in relation to existing work in the literature, we constructed Table 1. For an objective comparison of the proposed approaches, Table 1 compares the few existing documents in the literature, dividing them into two categories based on the nature of the input data, *i.e.*, whether they use images or electrical signals for *snail trail* fault detection. The aspects compared in Table 1 include approach, type and quantity of data, how many scenarios under different irradiations are evaluated, advantages, disadvantages, and accuracy. The number of scenarios under different irradiance levels is of paramount importance since the performance of PV systems is closely tied to solar irradiance, which, in turn, can change rapidly. In cases of low irradiance, fault detection becomes extremely challenging.

As can be observed in Table 1, in a general context, image-based approaches, although highly efficient, are strongly limited to the detection of faults with a noticeable visual expression. However, snail trail-type faults are not strongly noticeable under low irradiation conditions, which, in turn, poses a significant limitation for image-based approaches. Moreover, these approaches require image capture under certain conditions, which represents an increase in cost due to the need for drones, image transmission, storage, among others, without assuring that the approach can be directly applied to another new solar plant. Generally, retraining would be necessary due to the change in irradiation reference, which results in a change in panel temperature and, therefore, modifications in image characteristics. Furthermore, solar plants would also need an additional fault detection approach that lacks a visual signature, increasing the overall solution cost. Finally, these approaches necessitate a large number of samples (images) for fault detection.

On the other hand, in the same Table 1, highlights a study that uses electrical signals as input for *snail trail* fault detection algorithms (Sepúlveda-Oviedo et al., 2022). This work focuses on overcoming the limitations of using images and proposes an approach that can be readily extrapolated to detect other faults. Since the signals currently collected by PV data acquisition systems also capture time series, by merely modifying the training phase, it is possible to detect other types of failures in PV systems or other time series-based systems. It is important to clarify that the faults of interest to the industry are those that start to reduce PV production, or that generate other errors of that type, and such faults invariably represent a typical modification in the electrical signals of a panel. Therefore, even if the impact level of the fault is low, there will inevitably be a change in the electrical output signal of the PV system. This implies that an approach based on electrical signals can cover a broader range of detectable faults than an image-based approach. Furthermore, it demonstrated that a large amount of data is not necessary for successful fault detection when using electrical signals. However, the limitation of this work is that it does not evaluate its approach under different irradiation conditions, such as different seasons of the year.

Therefore, based on the hypothesis that utilizing the electrical signal of the PV system allows the detection of a greater number of faults, a reduction in the required number of samples for detection, resulting in lower storage capacity, reduced computation time, and computing power, this study proposes an approach that was also tested under varying irradiation conditions (different seasons of the year).

**Table 1**
Comparison of existing works in the literature that propose approaches for detecting snail trails type faults.

| Ref | Year | Approach | Data type and samples | # Scenarios | Advantages | Disadvantages | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| This study | 2024 | ELDIAG: An Ensemble Learning Framework for Snail Trail Fault Detection and Diagnosis in Photovoltaic Modules | 16 electrical signals | 4 | High precision for detecting snail trails, requiring little data, avoiding changes in irradiation. Easily extrapolated to other faults. | Conditioned on the capture of electrical signals in the form of a time series | [88, 89] |
| Oulefki et al. (2023) | 2023 | Unsupervised sensing algorithms and 3D Augmented Reality (AR) visualization | 277 Thermo-graphic aerial images and 934 infrared images | 1 | High-precision detection of abnormalities, such as hot spots and snail trails. | It fails to differentiate between hotspots and snail trails. It requires retraining when the PV plant conditions change. It was only analyzed under a specific meteorological scenario. Limited to faults with visual signatures. | – |
| Venkatesh S et al. (2023) | 2023 | A deep learning-based technique involving convolutional neural networks (CNNs) | 600 RGB images | 1 | High precision for six different types of faults. | Strong internet dependency for data transmission. No testing under different irradiation scenarios. High cost. High data requirements. Requires additional instrumentation (drones). | 98.66 |
| Vasanth et al. (2023) | 2023 | DenseNet-201+kNN | 600 RGB images | 1 | High Precision in Snail Trail Detection | High computational cost, internet dependency for transmission, large data volumes, and the propagation of irrelevant information due to artificially augmented database from 600 to 3150. | 100 |
| Lestary et al. (2022) | 2022 | Deep learning with the YOLO (You Only Look Once) algorithm version 3 | 277 Thermo-graphic aerial images | 1 | High Precision in Snail Trail Detection | Challenging to implement on a large scale due to its high cost. Not tested using the same panels under varying irradiation conditions. Limited solely to faults with visual signatures. | 99.7 |
| Naveen Venkatesh et al. (2022) | 2022 | An ensemble-based deep neural network (DNN) + Random Forest using Majority Voting | 600 RGB images | 1 | High accuracy compared to approaches that solely use deep learning. | It employs only one type of machine learning algorithm in ensemble learning, which limits the final behavior of the ensemble framework. Tested under a single environmental scenario. High cost due to the need for drones to capture photos. | 99.86 |
| Sepúlveda-Oviedo et al. (2022) | 2022 | DTW Hierarchical clustering + PLS regression | 12 electrical signals | 1 | High accuracy and low computational cost. | It was not tested under different irradiation scenarios throughout the year. | 87.5 |
| Venkatesh and Sugumaran (2022) | 2021 | Multiple deep learning and machine learning algorithms are compared | 600 RGB images | 1 | High precision for six different types of faults. | High computational cost, internet dependency for transmission, and large data volumes are required for detection. | [88.25, 100] |
| Li et al. (2018) | 2018 | A convolutional neural network (CNN) | 3000 RGB images | 1 | High Precision for five Types of Faults | Strong dependence on wireless communication networks. High cost. Limitations related to image capture conditions. | [77.32, 84.40] |

## 3. Methodology

In an effort to take advantage of multiple learning algorithms, EL is a rather recent framework assembling properties of different techniques with good trade-offs (Ganaie et al., 2022). It has been applied for the diagnosis of renewable energy systems (Aizpurua et al., 2021). The main idea is to combine several base models in order to produce one optimal predictive model and to improve the classification results of any of the base fault detection techniques. Two main types of Ensemble Learning techniques can be considered: simple or advanced. In the first case, the predictions made by different models are taken as separate votes and the final prediction is obtained by combining the votes (Eskandari et al., 2023). For this there exists multiple options such as average, weighted average, majority voting and weighted majority voting. In the second case several models are used to make intermediate predictions and a meta classifier performs the final prediction based on the intermediate ones (Mellit et al., 2023). This study proposes a simple approach, as the goal is not to propose a new ensemble modeling but to demonstrate its great potential in this type of applications.

The first objective of this research is to enhance the fault detection and classification results in PV systems, even in cases where there is a limited number of PV panels (2 or more panels). The second objective of this study is to decrease the computation time required by conventional fault detection systems while ensuring the same or improved level of detection accuracy.

The proposed solution is an ensemble learning framework whose relevance is illustrated on a real PV platform.

The platform has 16 PV panels. The classifiers used are trained using 8 panels, 4 in a healthy state and 4 with the *snail trail* fault. The methodology is validated for the other 8 panels (4 healthy and 4 with traces of discoloration (snail tracks or snail trails). The considered signal is the optimal current linked to the MPP of each panel. This data is acquired using a commercial data acquisition module, known as the Tigo optimizer,[2] previously installed on the platform for production optimization and monitoring purposes. All signals used (for training, testing, validation) were captured for periods of one day, in each season of the year. To consider the influence of weather on irradiance, signals are divided into 4 temporal slices named: Morning, Midday, Afternoon and Evening (Sepúlveda-Oviedo et al., 2022).

[2] To obtain the description of this system, visit here.

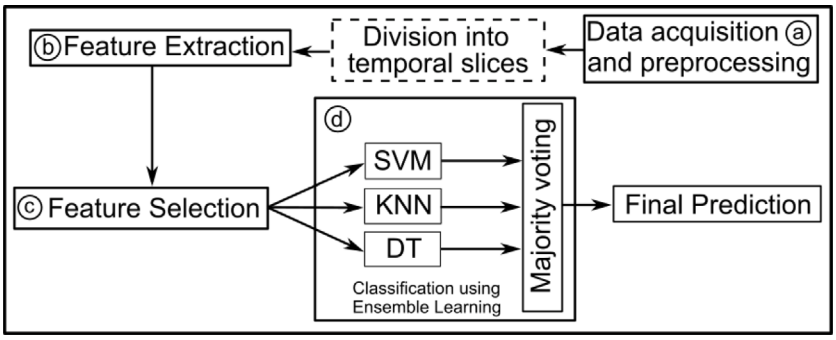


**Fig. 3.** Description of stages of the proposed Ensemble Learning approach.

A summary of the proposed methodology is presented in Fig. 3. Each of the stages of the methodology is explained in detail in the following sections.

### 3.1. Feature engineering

Although using EL improves classification results, the quality of results stays strongly linked to the quality of the databases used to train models. To increase the richness of the training data, some works have explored signal decomposition based on Multiresolution Signal Decomposition (MSD) (Yi and Etemadi, 2017b). However, despite its interest this methodology does not focus on the process of feature selection or dimensionality reduction and ignores the time dependence of PV measurements. Furthermore, these approaches have not been applied in real cases where electrical signatures of faulty panels are similar to those of healthy ones, as in the case of panels with *snail trail* phenomena. To reduce time dependency, an additional stage of statistical feature extraction (skewness, entropy, mean, among others) can be performed. With this type of non-time-dependent features, an important increase of the variance between the classes is achieved (Sepúlveda-Oviedo et al., 2022).

Extracting frequency time signatures and then statistical incremental signatures substantially increases the feature space dimension. This is where the use of feature selection algorithms such as PCA makes sense.

Working under the assumption that some faults may be visible in the time domain and others in the frequency domain, a multiresolution signal decomposition based on the discrete wavelet transform over each temporal slice is used to analyze them simultaneously. As a product of this decomposition, a set of detail and approximation coefficients are obtained on which the extraction of statistical characteristics is performed. For each temporal slice the features obtained are stored in a matrix of predictors that is then subjected to a feature selection process to reduce the feature space dimensionality keeping only the most relevant information. The reduced feature matrix is used as input to a set of machine learning algorithms, following an ensemble learning strategy, to detect and classify PV system faults.

### 3.2. Ensemble learning strategy

The component methods of the ensemble learning strategy are K Nearest Neighbors (KNN), SVM and Decision tree learning (DT). These three algorithms have been selected because they operate based on three distinct principles, allowing the approach to harness the strengths of each principle. KNN relies on similarity measurement using a distance metric, the SVM method is based on finding an optimal separating hyperplane between two data classes, and, lastly, DT are founded on partitioning a dataset into smaller subsets to enable decision-making based on the information within each subset. The ensemble learning strategy utilized offers fresh insights into faults by leveraging the complementary strengths of its component methods when handling datasets. SVMs excel at identifying optimal non-linear decision boundaries, which makes them highly effective for complex datasets. Similarly, Decision Trees are adept at recognizing non-linear relationships among features, though they may be susceptible to overfitting. In contrast, KNN is vulnerable to noisy data due to its dependency on the proximity of data points in the feature space, making it less robust to noise than SVM and Decision Trees. The latter two are better at handling noise, particularly SVMs, which aim to maximize the margin between classes. Decision Trees, on the other hand, can partly overlook noise and outliers by concentrating on splits that maximize information gain across significant portions of the data.

To reach the optimal prediction, the 3 algorithms are combined based on the majority voting ensemble technique.

## 4. Data acquisition and preprocessing

The 16 PV panels of the experimental platform with reference *SLK60P6L*, can generate power between 205 and 240 $Wp$. Each panel is instrumented with a commercial monitoring system provided by an optimizer built by the company Tigo (Tigo, 2023) and able to be connected to other optimizers in series. The signals have been acquired with a sampling time of one minute in the year 2020 on August 6 between 7:00 a.m. and 8:00 p.m., November 6 between 7:45 a.m. and 5:15 p.m., February 6 between 8:00 a.m. and 6:00 p.m. and finally on May 6 between 7:00 a.m. and 8:00 p.m. These dates were carefully selected approximately in the middle of each of the seasons of the year, to measure the robustness of the proposed approach. Every day the data acquisition started as soon as the PV panel began to produce and ended once the panel stopped producing. The signals captured on each day of each season are shown in Fig. 4.

The *snail trail* fault considered in these experiments represents corrosion of the sheet of the encapsulation surface (Li, 2021). Although this fault does not cause severe or critical performance reduction of the PV panels at the beginning, the fault can evolve producing cracks or microcracks in the PV cells if the panels continue to be exposed to the same conditions of solar radiation. This can even lead to the production of the PV system (Kim et al., 2016). As it can be seen in Fig. 4, the behavior of PV panels with a *snail trail* (Red color) is very similar to that of healthy PV panels (orange) in all scenarios (Summer, Fall, Winter and Spring).

Once the data has been captured, feature extraction can be performed.

## 5. Feature extraction

This stage is based on the iterative wavelet decomposition, named Multiresolution Signal Decomposition, of the current signal $I_i$ for each PV panel $i = 1, \ldots, n_p$. It is combined with an extraction of statistical features on the decomposed signals as it has been proposed in other works previously (Dadhich et al., 2019) .

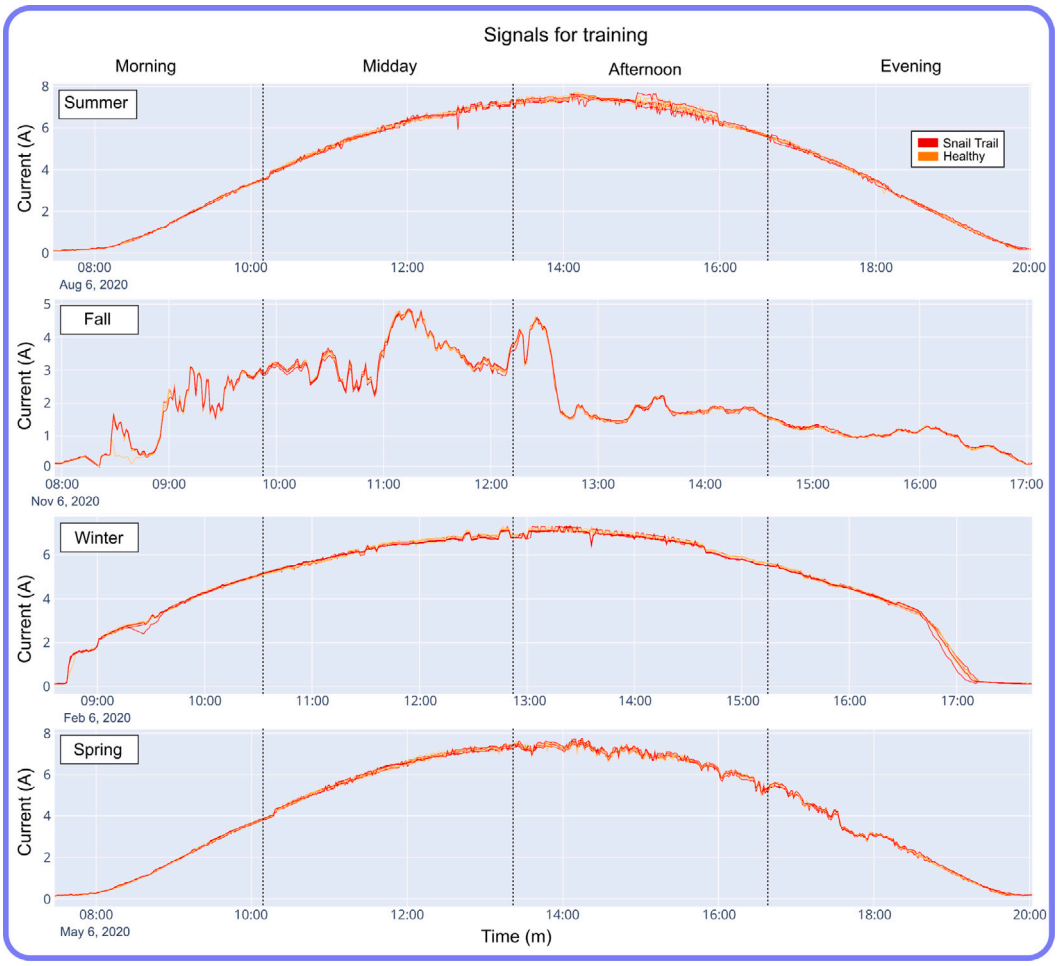


**Fig. 4.** Optimal current signals from 8 photovoltaic panels used for training (4 healthy (orange) and 4 with *snail trails* (red)).

### 5.1. Multi-resolution signal decomposition applied to PV system

In the PV domain, different signal processing and decomposition techniques have been studied and proposed for fault detection and classification (Lebreton et al., 2022). Among the proposed techniques, the continuous Fourier transform (FT) and the discrete FT that only provide information in the frequency domain are the best well-known.

Other research recommends using a FT with an associated window that provides time and frequency information. The problem with this technique applied on PV systems is that once the window is chosen, it cannot be changed if it becomes not suitable to new data. In some variations of weather conditions, it may induce a lack of efficiency to detect non-stationary disturbances having different signatures. These reasons have made the use of wavelet transform popular (Srikanta Murthy, 2018). The wavelet transform can deliver spatial frequency and time information, showing interesting improvements in the identification of different types of faults in PV systems (Ray et al., 2018). Along the same lines, other approaches have been proposed modifying the wavelet transform. Multi-resolution Signal Decomposition (MSD) applies the wavelet decomposition iteratively (Yi and Etemadi, 2017a). Slantlet transform (Mandal et al., 2012) is based on a modified version of the discrete wavelet with two zero moments and modified time localization. Finally, the wavelet packet transform uses an iterative decomposition on the approximation and detail coefficients of the signal (Ahmadipour et al., 2022).

Fault conditions in PV systems have associated some variations in the waveform of the current output signal that can be located in the frequency domain and/or in the time domain. Due to this, the MSD has become very relevant and is widely used (Lebreton et al., 2022). In general, the MSD based on discrete wavelet decomposition *DWT* divides the input signal into ranges with variable frequency. The decomposition of the signals depends directly on the pattern or *mother wavelet* selected for the decomposition. Each *mother wavelet* has associated or captures a particular frequency band. Additionally these *mother wavelets* have different computation speeds and decomposition quality. It is for this reason that the quality of the decomposition must be judged on the basis of the specific application. *Mother wavelets* can be defined as follows:

$$\Phi_{c,d}(t) = \frac{1}{\sqrt{c}}\Phi\left(\frac{t-d}{c}\right), \tag{1}$$

where $c$ and $d$ represent the scale and offset factor respectively. $t$ is the timestamp of the panel current signal and $\Phi$ is the *mother wavelet*. The values of $c$ and $d$ are obtained from the Equations (2) and (3) (Yi and Etemadi, 2017b)

$$c = c_0^{-(p_x/2)}, \tag{2}$$

$$d = q_x d_0 c_0^{p_x}, \tag{3}$$

where $p_x, q_x \in \mathbb{Z}$, $c_0 > 1$ and $d_0 > 0$. According to Yi and Etemadi (2017b) the discrete wavelet decomposition with the *mother wave* $\Phi_{c,d}(t)$, of a signal $S_{\{1:n_S\}}$, can be described as follows:

$$DWT(c,d) = \frac{1}{\sqrt{c}}\sum_{1:n_S} S(t)\Phi\left(\frac{t-d}{c}\right), \tag{4}$$

The challenge when performing the MSD is the selection of the relevant *mother wavelet* $\Phi_{c,d}(t)$. Therefore, some articles have dedicated

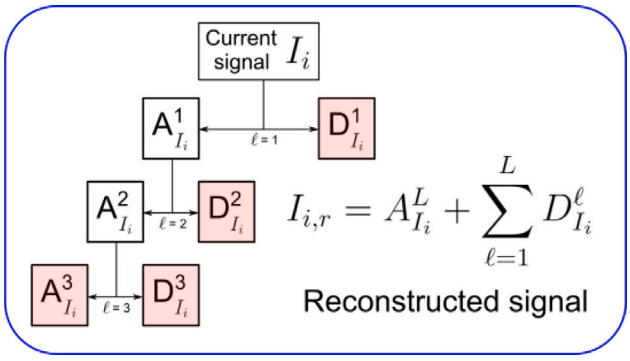


**Fig. 5.** Decomposition into 3 levels applied on the current signal $I_i$ of each faulty or healthy panel.

entire research to the development of complex algorithms for the optimal selection of *mother wavelet* (Jang et al., 2021). For the selection of the mother wavelet, in this research, all wavelet families were tested and the *mother wavelet* Daubechies38 (db38) has been selected due to its computational speed and good decomposition result. The decomposition into 3 levels of one of the current signals as a function of time is illustrated in Fig. 5. This is done for all the temporal slices. The three levels of decomposition $j = 1, 2, 3$, provide the coefficients of an approximated signal, noted $(A^j_{I_i})$, and the associated detail signal $(D^j_{I_i})$. Outlined in red, one can visualize the different required coefficients (approximation and detail coefficients) of the signals from which one can obtain the reconstructed signal $(I_{i,r})$, $i = 1, \ldots, n_P$, where $n_P$ is the number of panels. The extraction of statistical features is applied to these signals, responsible for rebuilding the signal according to the DWT theory.

### 5.2. Extraction of statistical features

The statistical features extracted in this section such as mean, Power spectral density, Skewness, Entropy and Kurtosis are suggested for fault detection in PV systems (Malik et al., 2022). For a signal $S_{\{1:n_S\}}$, without loss of generality, *Mean* $\mu$ indicates a representative number of the time series calculated as follows:

$$\mu = \frac{1}{n_S}\sum_{t=0}^{n_S} S_t, \tag{5}$$

Power spectral density *PSD* represents the power content of the signal as a function of frequency. The amplitude is normalized per unit frequency as:

$$PSD = \lim_{T\to\infty} \frac{1}{T}|S_t|^2, \tag{6}$$

*Skewness* is a representation of the asymmetry of the data respect to $\mu$ and is described as follows (Esmael et al., 2012):

$$Skewness = \frac{\frac{1}{n_S}\sum_{t=0}^{n_S}(S_t-\mu)^3}{\left(\sqrt{\frac{1}{n_S}\sum_{t=0}^{n_S}(S_t-\mu)^2}\right)^3}, \tag{7}$$

*Entropy* is widely used in information theory to evaluate the uncertainty of a signal and even as a tool to identify the quality of the information or inherent surprise of the signal. Entropy can be defined as:

$$Entropy = -\sum_{t=1}^{n_S} p(S_t)log(S_t), \tag{8}$$

*Kurtosis* measures the maximum value and skewness of the probability distribution of the data and is defined as follows (Esmael et al., 2012):

$$Kurtosis = \frac{\frac{1}{n_S}\sum_{t=0}^{n_S}(S_t-\mu)^4}{\left(\sqrt{\frac{1}{n_S}\sum_{t=0}^{n_S}(S_t-\mu)^2}\right)^4}, \tag{9}$$

With this set of statistical features, the feature matrix $\mathbb{M}_{\mathbb{F},*}$ of dimensions $n_P \times ((L+1)\times n_F)$ is constructed adopting the notation of Sepúlveda-Oviedo et al. (2022).

$$\mathbb{M}_{\mathbb{F},*} = \begin{pmatrix} A^L_{I_{1,*}} & D^1_{I_{1,*}} & \ldots & D^\ell_{I_{1,*}} & \ldots & D^L_{I_{1,*}} \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ . & . & \ldots & . & \ldots & . \\ A^L_{I_{n_P,*}} & D^1_{I_{n_P,*}} & \ldots & D^\ell_{I_{n_P,*}} & \ldots & D^L_{I_{n_P,*}} \end{pmatrix}, \tag{10}$$

where the temporal slice $*\in$ {Morning, Midday, Afternoon and Evening}, $L$ is the number of decomposition levels and $n_F$ is the number of statistical features. $A^L_{I_{i,*}}$ and $D^\ell_{I_{i,*}}$ are the vectors that correspond to the coefficients of approximation and coefficients of detail. Each of these vectors has dimension $n_F$. It is possible that the high dimensionality of the matrix $\mathbb{M}_{\mathbb{F},*}$ increases the computational complexity and/or contains information not relevant for the classification of the panels. As a solution to this problem, a feature selection process that compresses the original feature matrix $\mathbb{M}_{\mathbb{F},*}$ for each of the temporal slices $*\in$ {Morning, Midday, Afternoon and Evening} is applied. Then, a generic and theoretical explanation of the combined methods in the approach proposed in this research is carried out. The explanation is made in a generic way to highlight that the method could be applied to any other system. Then, the method is exemplified using the case study.

## 6. Feature selection

Feature selection is a process that allows keeping only the relevant information for classification and therefore fault detection. One of the main advantages of this process consists in a drastic decrease of the computational time of the learning algorithms, hence increasing their efficiency to process complex big data. Multiple algorithms for feature selection, often accompanied by dimensionality reduction, have been proposed. Some of the best-known approaches are PCA, Isomap; Local Linear Embedding, and Singular Value Decomposition. This research uses PCA and Isomap algorithms that apply a transformation to the original features. As a result, a transformation of the matrix $\mathbb{M}_{\mathbb{F},*}$ into a new matrix $M_{F,*}$ of reduced dimensions $n_P \times U$ is obtained as:

$$M_{F,*} = \begin{pmatrix} C^1_{I_{1,*}} & \ldots & C^{u-1}_{I_{1,*}} & \ldots & C^U_{I_{1,*}} \\ . & \ldots & . & \ldots & . \\ . & \ldots & . & \ldots & . \\ . & \ldots & . & \ldots & . \\ C^1_{I_{n_P,*}} & \ldots & C^{u-1}_{I_{n_P,*}} & \ldots & C^U_{I_{n_P,*}} \end{pmatrix}, \tag{11}$$

where $U$ is the number of features obtained as a result of PCA or Isomap, also called latent components. Each row $M_{F,*}(i,.)$, $i = 1, \ldots, n_P$ of the matrix $M_{F,*}$ provides the signature of the health state of the PV panel $PV_i$. Each column $M_{F,*}(.,j)$, $j = 1, \ldots, U$ of the matrix $M_{F,*}$ provides each of the latent components.

In this case study, the signals from the 16 panels are decomposed into 4 levels both for training and testing. That is, the matrix $\mathbb{M}_{\mathbb{F},*}$ has dimensions of $8 \times (5 \times 5)$ (5 statistical features and 4 detail signals + 1 approximate signal). The 25 features are reduced to only the first three latent components ($U = 3$) which contain 95% of the explained variance, as shown in Fig. 6. In other words, they encapsulate the most relevant information from the original matrix $\mathbb{M}_{\mathbb{F},*}$.

This process reduces the amount of data keeping the relevant information for the classification of the health status of the panels. It allowed to reduce in a proportion 1/5 the training computational time and in

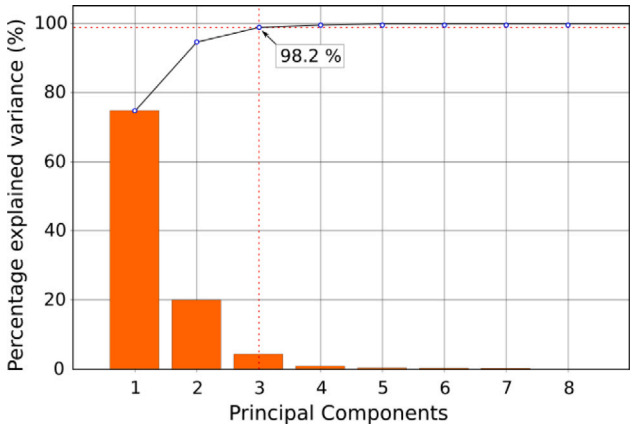


**Fig. 6.** Cumulative explained variance of the PCA method.

a proportion 1/3 the required disk space. The description of the two chosen algorithms used to create two independent matrices $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ is briefly presented.

### 6.1. Principal component analysis

Principal Component Analysis (PCA) is a powerful multivariate statistical technique that identifies and extracts uncorrelated features (named latent components) from the multidimensional original feature space. That is, the PCA algorithm uses a linear combination of the original features to construct new features while maintaining the maximum variance information (Fadhel et al., 2019). PCA represents the new features in two subspaces (Fadhel et al., 2019). The first is named the principal subspace or the "representation" subspace and the second is named the complementary or residual subspace in which noise and outliers are rejected.

Assuming $M$ original features and their covariance matrix, the first principal components, spanning the principal subspace, are given by the first $q$ dominant eigenvectors of the data covariance matrix. These later are associated with the highest $q$ eigenvalues. The last not retained eigenvectors $(M - q)$ define the residual subspace. The projection of the data onto the dominant eigenvectors provides principal component scores (Fadhel et al., 2019). The percentage of variance contained in each principal component is expressed by its corresponding eigenvalue. Each principal component is aligned in a direction corresponding to the largest variance in the data, starting with the first PC. Therefore, the principal components are ordered from the highest variance, or most energized, associated with the highest eigenvalue, to the least variance associated with the lowest eigenvalue (Fadhel et al., 2019).

Let us come to a case study. Without loss of generality and to simplify the notations, the dimensions of the matrix $\mathbb{M}_{\mathbb{F},*}$ are henceforth expressed as $N \times M$, where $N = n_P$ and $M = ((L+1) \times n_F)$. Each row of $\mathbb{M}_{\mathbb{F},*}$ provides the original features for one of the PV panels and one temporal slice. Each row is extended with the different observations for these features, which results in a data matrix of dimensions $(O \times N) \times M$, where $O$ is the number of observations. Let us denote the data matrix by $\mathbf{X}$. Each of the rows of the data matrix $\mathbf{X}$ represents a different repetition of the observations, and each of the columns is a particular feature.

The elements of $\mathbf{X}$ are first standardized by subtracting the mean of the observations for the corresponding feature and by dividing by the standard deviation of that same feature. This provides us with a standardized data matrix $\mathbf{X_c}$ from which the covariance matrix is calculated as follows:

$$\mathbf{C} = \frac{1}{N-1} \mathbf{X_c}^T \mathbf{X_c}, \tag{12}$$

where $\mathbf{X_c}^T$ denotes $\mathbf{X_c}$ transposed. The quality of the obtained representation depends on the latent components retained in the main representation space. Let us denote by $\mathbf{P}$ the column matrix of the retained $q$ dominant eigenvectors, *i.e.* the linear transformation matrix. Dominant eigenvectors are arranged in descending order of their corresponding eigenvalues (Fadhel et al., 2019). The principal component scores are obtained by projecting the original centered and reduced data onto the new space generated with $\mathbf{P}$, obtaining the matrix $\mathbf{T}$ of the principal component scores of dimensions $(O \times q)$. That is, the linear transformation matrix $P$ transforms $\mathbf{X_c}$ into a new matrix of latent components $\mathbf{T}$ as follows:

$$\mathbf{T} = \mathbf{P}\mathbf{X_c}, \tag{13}$$

where each column $\mathbf{T}(., i)$ of the matrix $\mathbf{T}$ provides a principal component for the set of PV panels $n_p$.

### 6.2. Isometric mapping

Isomap stands for isometric mapping. This method addresses dimensionality reduction as the problem of creating a transformation from high dimension to low dimension in a graph-theoretic framework (Samko et al., 2006). Isomap extends the metric multidimensional scaling (MDS) (Hout et al., 2013) by incorporating the concept of geodesic distances imposed by a weighted graph (Bouttier et al., 2003).

In graph theory, the distance between two vertices on a graph corresponds to the number of edges in the shortest path connecting them. This distance is also known as the geodesic distance (Bouttier et al., 2003). Isomap is intended to preserve pairwise geodesic distances between conformations in a graph, that is, in the lower dimension. The distances $d_{\mathbf{X}}(i, j)$ between all pairs $i$, $j$ of $O$ data points in the high-dimensional input space $\mathbf{X}$ are required as input to the Isomap algorithm, generally measured using the standard Euclidean distance. The algorithm outputs coordinate vectors $\mathbf{Y}_i$ in a (lower) d-dimensional Euclidean space $\mathbf{Y}$ that best represents the intrinsic geometry of the data. Dimensionality reduction or feature selection using Isomap is based on three steps:

**First step** – Isomap estimates the neighborhood graph by determining the neighbors of each input point in the manifold $\mathbf{M}$ based on the distances $d_{\mathbf{X}}(i, j)$ between pairs of input points $i$, $j$ in the input space $\mathbf{X}$. The neighbors can be determined with the $K$ nearest neighbors (K-Isomap) or with a neighborhood radius $\epsilon$ ($\epsilon$-Isomap) (Samko et al., 2006). Neighborhood relationships are plotted as a weighted graph $\mathbf{G}$ over the data points, with weighted edges $d_{\mathbf{X}}(i, j)$ between neighboring points.

**Second step** – Isomap computes the shortest path graph given the neighborhood graph. Isomap then estimates the geodesic distances $d_{\mathbf{M}}(i, j)$ between all pairs of points in the manifold by computing the shortest path lengths $d_{\mathbf{G}}(i, j)$ in $\mathbf{G}$. $d_{\mathbf{G}}(i, j) = d_{\mathbf{X}}(i, j)$ if $i$, $j$ are joined by an edge, and $d_{\mathbf{G}}(i, j) = \infty$ in otherwise.

**Third step** – Isomap constructs the lower dimensional embedding applying classical MDS to the graph distance matrix $\mathbf{D_G} = \{d_{\mathbf{G}}(i, j)\}$, constructing an embedding of the data in a d-dimensional Euclidean space that best preserves the estimated intrinsic geometry of the manifold.

The only free parameter of Isomap is the neighborhood factor $K$ or $\epsilon$ depending on the method used. Its efficiency lies in choosing an appropriate value for these parameters whose choice is generally left to the user.

## 7. Classification using ensemble learning

In this section, the principles of the three classification methods that constitute the Ensemble Learning method are presented. The classification methods are applied in parallel to the reduced matrices resulting from the PCA and Isomap methods $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ and to the matrix without feature selection $\mathbb{M}_{\mathbb{F},*}$. Performing the classification on the matrices $M_{F,*}^{PCA}$ and $M_{F,*}^{ISOMAP}$ significantly reduces the calculation time, since the high dimensionality of the features is reduced with a minimum loss of information.

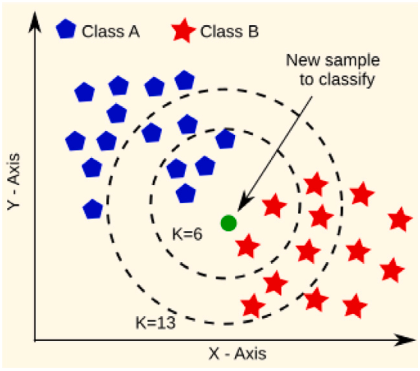


**Fig. 7.** Example of classifying a new sample using kNN.

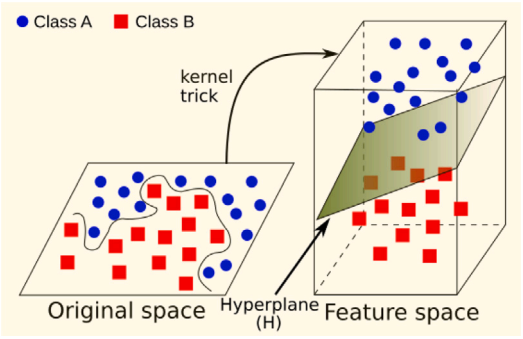


**Fig. 8.** Example of classification hyperplane representation of SVM. $H$ corresponds to the optimal hyperplane.

### 7.1. K-nearest-neighbors

The non-parametric algorithm K-Nearest-Neighbors (kNN) is one of the most used models for classification thanks to its simplicity (Zhang and Zhou, 2007). For a given sample $x' = [x'_1, \ldots, x'_{n_F}]$ with $n_F$ features to be classified, kNN finds its nearest neighbors among already classified samples based on some distance metric. This work uses the Euclidean distance as illustrated by Eq. (14) for a classified sample $x = [x_1 x_2, \ldots, x_{n_F}]$ (Wang et al., 2020).

$$d(x, y) = \sqrt{\sum_{i=1}^{n_F} (x_i - x'_i)^2}, \tag{14}$$

kNN assigns the sample to the class most common among its k nearest neighbors, as illustrated by Fig. 7.

The proper choice of the (only) free parameter $k$ has a significant impact on diagnosis performance. As the value of $k$ is increased, the model can tolerate more noise because the impact of variance caused by random error is reduced but there is a risk of missing a small but important pattern in the data. The key to choose an appropriate value of $k$ is to strike a balance between overfitting and underfitting.

### 7.2. Support vector machines

Support vector machines (SVM) are one of the most powerful, complex and widely applied classification algorithms for fault diagnosis (Cervantes et al., 2020). The SVM classifier searches for one or a set of optimal separation hyperplanes, maximizing the margin between classes. When classes are not linearly separable, the main idea of SVM is to map the input space into a higher dimensional feature space through a *kernel function* and then apply linear SVM in this space as illustrated in Fig. 8. For this purpose, it uses a set of data points, known as support vectors, that are close to the hyperplane and influence its position and orientation.

To formally understand SVM, let us consider the two-class support vector machines and a set of $n$ training samples $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. $x_i \in R^d$, where $y_i$ is the class label of the $x_i$ sample and $y_i \in [-1, +1]$. The optimal separating hyperplane (H) maximizing the "margin" of the classifier is given by the equation: $w^T x + b = 0$, where $w \in F$ and $b \in R$ are two parameters that determine the position of the decision hyperplane in the feature space $F$ (its orientation is tuned by $w$ and its displacement by $b$). This leads to the following decision function and the problem is to find $(w, b)$:

$$f(x; w, b) = sign(w^T x + b) \in \{-1, +1\}$$
$$\text{where: } sign(w^T x + b) = \begin{cases} +1, & \text{if } (w^T x + b) \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{15}$$

In soft-margin SVMs, certain samples can breach the margin, and a non-linear decision boundary can be achieved by projecting the data into a higher-dimensional space $F$ through a nonlinear function $\Phi(x)$. Although data points may not be linearly separable in their original space, they are mapped to a feature space $F$ where a hyperplane can separate them. To avoid over-fitting noisy data, slack variables $\xi$ are introduced, allowing some data points to lie within the margin. The parameter $C > 0$ in Eq. (16) regulates the balance between classification error on the training data and margin maximization. The objective function of SVM classifiers can be minimized as follows:

$$\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \xi_i$$
$$\text{Subject to: } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \tag{16}$$
$$\xi_i \geq 0, i = 1, \ldots, n$$

The minimization problem is solved using Lagrange Multipliers $\alpha_i, i = 1, \ldots, n$. The new decision function rule for a data point $x$ is:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b) \tag{17}$$

Every $\alpha_i > 0$ is weighted in the decision function and thus supports the machine. Since SVMs are considered to be sparse, there are relatively few Lagrange multipliers with a non-zero value.

The function $K(x, x_i) = \Phi(x)^T \Phi(x_i)$ is known as the *kernel function*. Since the outcome of the decision function only relies on the dot-product of the vectors in the feature space $F$ (*i.e.* all the pairwise distances for the vectors), it is not necessary to perform an explicit projection. As long as a function $K$ provides the same results, it can be used instead. This is known as the *kernel trick*.

### 7.3. Decision trees

Tree-based ML techniques are among the most widely used nonlinear models in many applications, where Random Forest (Sepúlveda-Oviedo et al., 2023b) and DT are the most popular having, in some cases, an accuracy greater than that of neural networks (Lundberg et al., 2020). The DT model uses two types of nodes, which are the decision node and the leaf node. Decision nodes have multiple branches and are used to make a decision, while leaf nodes are the result of these decisions. DTs are built from a recursive split of the set of samples based on a set of splitting rules that relate to the features (Mahesh, 2019). An illustration is presented in Fig. 9.

Many variants of DTs exist among which the Iterative Dichotomies 3 (ID3), Successor of ID3 (C4.5), Automatic Chi-Square Interaction Detector (CHAID), Classification and Regression Tree (CART), etc. In this work, the algorithm C4.5 has been selected due to its good results in detecting faults in PV systems (Benkercha and Moulahoum, 2018).

C4.5 builds decision trees from a set of training data using the concept of Information Gain that refers to entropy. Information Gain computes the change in entropy of a dataset before and after splitting

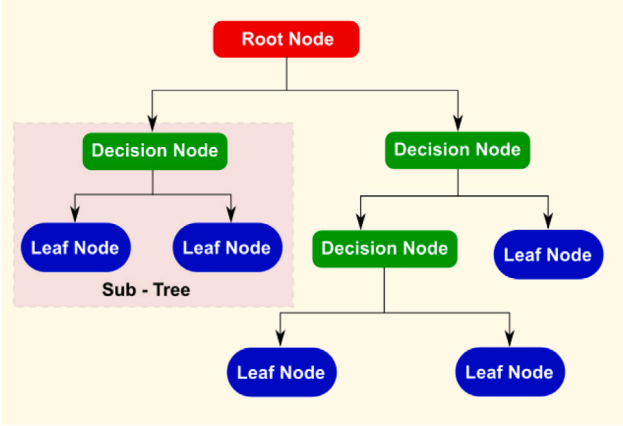


**Fig. 9.** Structure of the Decision Tree classifier. The DT classifier is composed of two types of nodes: decision nodes and leaf nodes.

the data on the values of a given feature. More precisely, the Information Gain of a split $IG_{split}$ equals the original entropy $H$ minus the weighted sum of the sub-entropies $H_j$, with the weights equal to the proportion of data samples being moved to the sub-datasets as given by Eq. (18).

$$IG_{split} = H - \left( \sum_j \frac{|D_j|}{|D|} * H_j \right) \tag{18}$$

where $|D|$ is the number of samples in the original data set $D$ and $|D_j|$ is the number of samples in the $j$th sub-dataset after being split.

C4.5 then attempts to lessen the bias of Information Gain $IG_{split}$ on highly branched predictors by introducing a normalizing term called the Intrinsic Information $II_{split}$ given in Eq. (19).

$$II_{split} = -\left( \sum_j \frac{|D_j|}{|D|} * log_2 \frac{|D_j|}{|D|} \right) \tag{19}$$

C4.5 finally obtains the Gain Ratio of a split $GR_{split}$ as follows:

$$GR_{split} = \frac{IG_{split}}{II_{split}} \tag{20}$$

The feature with the highest Gain Ratio is chosen to make the split.

### 7.4. Majority voting

The three classifiers (kNN, DT, and SVM) are integrated using the principle of majority voting (MV), as suggested in Zhang et al. (2020). To clarify, we refer to kNN, DT as unstable/stochastic classifiers and SVM as "stable learner" with the aim of achieving a better final detection result. In this context, a "stable learner" denotes a classifier with the capacity to make accurate predictions consistently, while an "unstable/stochastic learner" typically exhibits less consistent predictive capabilities.

This study takes advantage of the principle of majority voting, treating each "stable learner" and "unstable/stochastic learner" as a voter. In this approach, the prediction results or outputs of the three classifiers are compared to determine the final class assignment. Weighted majority voting, relative majority voting, and absolute majority voting are potential methods for conducting this vote.

Assume a set of $n_v$ weak classifiers or voters, each returning a singleton output set $O_i, i = 1, \ldots, v$ that contains one label value $l_j, j = 1, \ldots, n_l$ for each sample $x$. Hence $O_i \subset \{l_1, l_2, \ldots, l_{n_l}\}$ and $|O_i|$=1 for $i = 1, \ldots, v$. The voter output sets are gathered in the multiset $\mathcal{O} = O_1 \cup O_2 \cup \cdots \cup O_v$. Then the relative majority vote returns the label $l_v$ that has maximal frequency in $\mathcal{O}$ as follows:

$$l_v = \underset{l_j \in \mathcal{O}}{\operatorname{argmax}} \{\# l_j\} \tag{21}$$

where $\# l_j$ is the frequency of the label $l_j$.

Note that if the trust placed in each weak classifier is different, one can use a weighted variant of relative voting. In this case, the final label is obtained as follows:

$$l_v = \underset{l_j \in \mathcal{O}}{\operatorname{argmax}} \{ \sum_{i=0}^{v} w_i \underset{O_i}{\#} l_j \} \tag{22}$$

where $\underset{O_i}{\#}$ is the frequency of the label $l_j$ in the set $O_i$.

Finally, the absolute majority voting method only generates the final label if the highest voting rate of some label exceeds 50%, otherwise, it does not issue a prediction. A combination of relative and absolute majority voting is used in this work.

An example of classifying a current signal for a *snail trail* panel is presented in Fig. 10. At the first level, presented in Fig. 10, the relative majority voting is used to determine the health status in each temporal slice. Relative majority voting has been selected for this research due to its interesting results in fault detection (Zhang et al., 2020) and the non-use of arbitrarily assigned weights. At the second level, the absolute majority vote is used to determine the overall health status of the panel.

## 8. Results

This section presents the classification results for the $n_v = 3$ algorithms (kNN, SVM and DT) and the EL method through the optimal current signals of 8 PV panels different from those used for training. The tuning parameters for the algorithms used in our study were meticulously selected through multiple rounds of testing to ensure optimal performance. This iterative process was crucial for achieving the best individual results with each algorithm, underlining the importance of precise parameter tuning in our findings. This process can be automatized by selecting a multidimensional grid and testing every combination. The current signals are captured under the same conditions as the signals presented in Fig. 4.

The selected characteristics with significant variance (obtained with the dimensionality reducers) are a priori those that can be useful to solve problems of detection and classification of the health status of PV panels. These features are processed by the feature selection algorithms and then processed by the classification methods. All algorithms (kNN, SVM, DT and EL) are trained and tested with the same PV panels. Then, the algorithms are tested with the signals presented in Fig. 11.

In this case study, the first majority vote uses the relative majority voting. Then, the results obtained from the relative majority voting for the 4 temporal slices are submitted to an absolute majority voting in a second stage.

To evaluate the prediction results of the classification algorithms, i.e the number of panels correctly classified, the confusion matrix and the $F_{value}$ are used. For example, in a classification example with two classes (class Positive ($Pos$) and class Negative ($Neg$)), the $F_{value}$ metric does not consider true negatives. A true negative is generated when a sample (in this case, a PV panel) that belongs to class $Neg$ is effectively classified in class $Neg$. The $F_{value}$ takes its value between 0 and 1, with 1 being the best performance and 0 being the worst. The $F_{value}$ is defined as follows:

$$F_{value} = 2 * \frac{pr * re}{pr + re}, \tag{23}$$

where the term $pr$, represents the precision that can be seen as the cost of false positives and is defined as follows:

$$pr = \frac{True_{Pos}}{(True_{Pos} + False_{Pos})}. \tag{24}$$

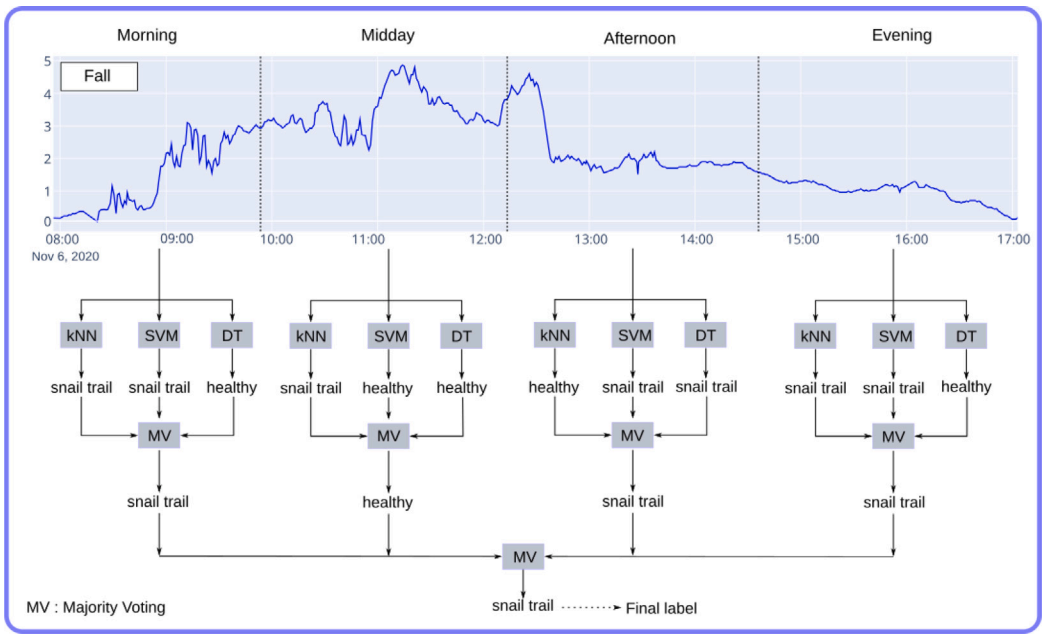


**Fig. 10.** Example of fault classification of a PV panel with *snail trail* using EL based on Majority voting.

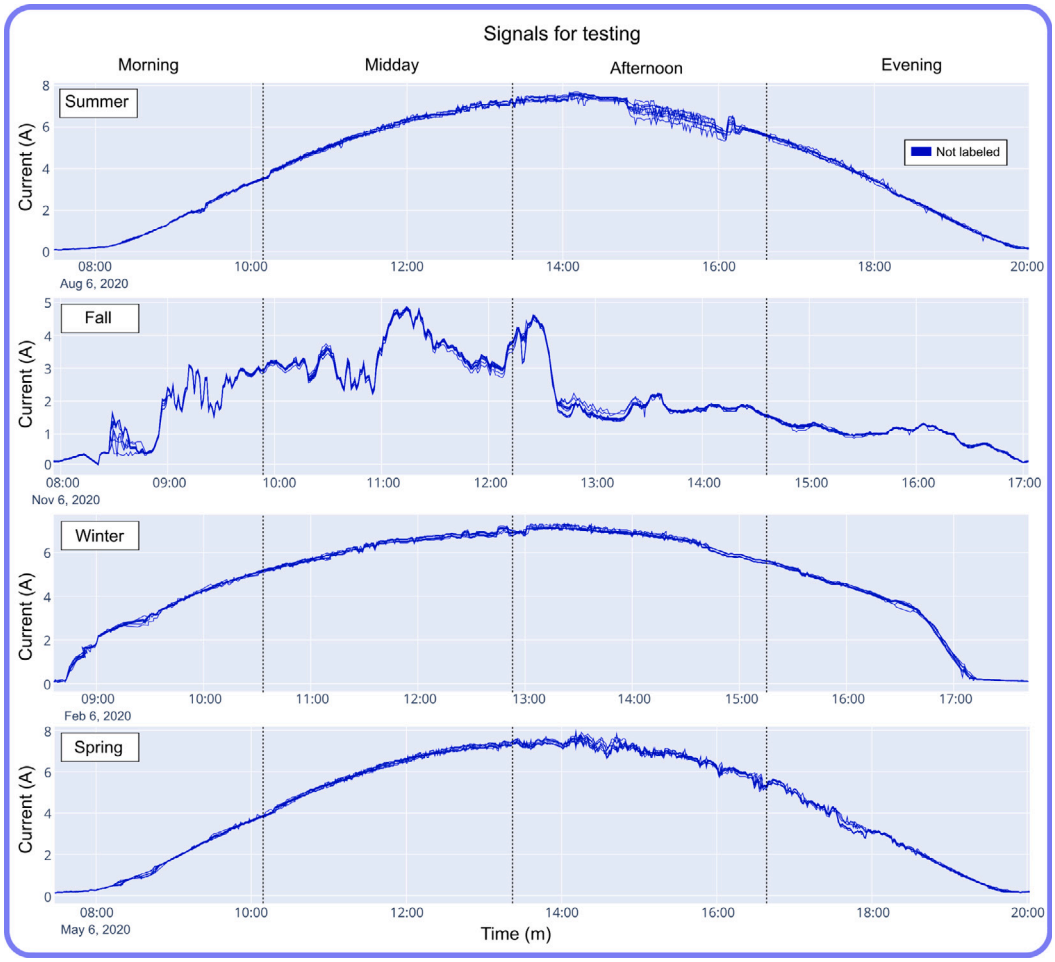


**Fig. 11.** Optimal current signals from 8 photovoltaic panels used in the testing of the proposed methodology. The signals were captured during 4 full days, one day for each season of the year. selected for each season of a year. The data is captured with a frequency of one minute. The 4 temporal slices proposed (Sepúlveda-Oviedo et al., 2022) and adopted in this research are represented using dotted lines.

**Table 2**
Fault detection and classification results ($F_{value}$) for signals captured in Summer. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

| Season | Temporal slice | Methodology | | kNN | SVM | DT | EL |
|---|---|---|---|---|---|---|---|
| Summer | Morning | Without approach | NSD_NDR | 0.62 | 0.62 | 0.5 | **0.71** |
| | | New approach | PCA | 0.73 | 0.7 | 0.53 | **0.88** |
| | | | Isomap | 0.7 | 0.71 | 0.54 | **0.8** |
| | Midday | Without approach | NSD_NDR | 0.63 | 0.65 | 0.53 | **0.72** |
| | | New approach | PCA | 0.63 | 0.71 | 0.54 | **0.80** |
| | | | Isomap | 0.66 | 0.65 | 0.54 | **0.75** |
| | Afternoon | Without approach | NSD_NDR | 0.62 | 0.64 | 0.54 | **0.72** |
| | | New approach | PCA | 0.63 | 0.71 | 0.59 | **0.83** |
| | | | Isomap | 0.68 | 0.64 | 0.55 | **0.74** |
| | Evening | Without approach | NSD_NDR | 0.63 | 0.63 | 0.51 | **0.7** |
| | | New approach | PCA | 0.67 | 0.66 | 0.53 | **0.85** |
| | | | Isomap | 0.67 | 0.67 | 0.53 | **0.7** |

**Table 3**
Fault detection and classification results ($F_{value}$) for signals captured in Fall. In Without approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

| Season | Temporal slice | Methodology | | kNN | SVM | DT | EL |
|---|---|---|---|---|---|---|---|
| Fall | Morning | Without approach | NSD_NDR | 0.63 | 0.64 | 0.5 | **0.72** |
| | | New approach | PCA | 0.67 | 0.73 | 0.53 | **0.88** |
| | | | Isomap | 0.63 | 0.63 | 0.5 | **0.76** |
| | Midday | Without approach | NSD_NDR | 0.62 | 0.64 | 0.54 | **0.72** |
| | | New approach | PCA | 0.66 | 0.67 | 0.6 | **0.88** |
| | | | Isomap | 0.67 | 0.71 | 0.51 | **0.8** |
| | Afternoon | Without approach | NSD_NDR | 0.62 | 0.65 | 0.53 | **0.7** |
| | | New approach | PCA | 0.68 | 0.64 | 0.59 | **0.87** |
| | | | Isomap | 0.62 | 0.65 | 0.54 | **0.76** |
| | Evening | Without approach | NSD_NDR | 0.62 | 0.62 | 0.5 | **0.7** |
| | | New approach | PCA | 0.7 | 0.69 | 0.57 | **0.88** |
| | | | Isomap | 0.67 | 0.7 | 0.5 | **0.79** |

**Table 4**
Fault detection and classification results ($F_{value}$) for signals captured in Winter. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

| Season | Temporal slice | Methodology | | kNN | SVM | DT | EL |
|---|---|---|---|---|---|---|---|
| Winter | Morning | Without approach | NSD_NDR | 0.63 | 0.64 | 0.54 | **0.71** |
| | | New approach | PCA | 0.7 | 0.75 | 0.54 | **0.8** |
| | | | Isomap | 0.67 | 0.65 | 0.57 | **0.75** |
| | Midday | Without approach | NSD_NDR | 0.63 | 0.63 | 0.54 | **0.72** |
| | | New approach | PCA | 0.73 | 0.74 | 0.54 | **0.8** |
| | | | Isomap | 0.65 | 0.67 | 0.5 | **0.72** |
| | Afternoon | Without approach | NSD_NDR | 0.64 | 0.65 | 0.51 | **0.71** |
| | | New approach | PCA | 0.64 | 0.74 | 0.57 | **0.79** |
| | | | Isomap | 0.71 | 0.68 | 0.6 | **0.73** |
| | Evening | Without approach | NSD_NDR | 0.61 | 0.64 | 0.5 | **0.7** |
| | | New approach | PCA | 0.69 | 0.72 | 0.59 | **0.85** |
| | | | Isomap | 0.72 | 0.67 | 0.54 | **0.74** |

**Table 5**
Fault detection and classification results ($F_{value}$) for signals captured in Spring. In Without Approach scenario, the statistical characteristics are extracted directly from the raw current signal. In New approach scenario, the full approach (signal decomposition, statistical feature extraction, and dimensionality reduction using PCA and Isomap) is performed.

| Season | Temporal slice | Methodology | | kNN | SVM | DT | EL |
|---|---|---|---|---|---|---|---|
| Spring | Morning | Without approach | NSD_NDR | 0.61 | 0.62 | 0.5 | **0.7** |
| | | New approach | PCA | 0.66 | 0.67 | 0.61 | **0.87** |
| | | | Isomap | 0.63 | 0.65 | 0.61 | **0.76** |
| | Midday | Without approach | NSD_NDR | 0.63 | 0.63 | 0.55 | **0.71** |
| | | New approach | PCA | 0.65 | 0.72 | 0.58 | **0.88** |
| | | | Isomap | 0.7 | 0.65 | 0.6 | **0.71** |
| | Afternoon | Without approach | NSD_NDR | 0.62 | 0.63 | 0.53 | **0.72** |
| | | New approach | PCA | 0.69 | 0.64 | 0.59 | **0.85** |
| | | | Isomap | 0.63 | 0.67 | 0.6 | **0.76** |
| | Evening | Without approach | NSD_NDR | 0.63 | 0.64 | 0.55 | **0.71** |
| | | New approach | PCA | 0.63 | 0.71 | 0.62 | **0.89** |
| | | | Isomap | 0.67 | 0.69 | 0.57 | **0.79** |

The term $re$ represents the recall, this recall is the estimate of the number of panels correctly classified based on the total number of panels belonging to the class. The recall is defined as follows:

$$re = \frac{True_{Pos}}{(True_{Pos} + False_{Neg})}. \tag{25}$$

Tables 2–5 present the results of the classification methods, for each season of the year, as a function of $F_{value}$. The values reported in Tables 2–5 are divided by each temporal slice. In the first scenario (without treatment approach), the extraction of statistical features is performed directly on the database of current signals. That is, no signal decomposition and no dimensionality reduction (NSD_NDR). In the second scenario, signal decomposition is performed using MSD, statistical feature extraction, and dimensionality reduction using PCA and Isomap.

Only on the classification results using the proposed methodology with the PCA method, the confusion matrix is used because in all the scenarios presented in the Tables 2–5, the PCA method outperforms the Isomap method. The confusion matrix is a widely known tool that allows visualizing the performance of a supervised learning algorithm or classification algorithm (Mahesh, 2019). In this matrix, each column represents the number of predictions of each class, while each row represents the instances in the actual class. This allows to see what types of successes and errors the model of this research is having when going through the learning process with the current data as a function of the time of each PV panel of the string. Fig. 12 shows the results of the classification algorithms for each season of the year, after dimensionality reduction using PCA. The results are presented in the form of a confusion matrix where 0 is the healthy class and 1 is the class of the panels with snail trail. In this research, it is considered that if at least in a temporal slice the sample is classified as faulty, the final label is assigned as a faulty panel.

## 9. Discussion

As observed in Tables 2–5, the performance of classifiers significantly improves with the method proposed in this study, namely PCA and Isomap. Additionally, in the same Tables 2–5 and Fig. 12, the EL algorithm demonstrates the ability to accurately discriminate between the two panel types (healthy and *snail trail*). This reduction of information to the essential elements effectively eliminates redundant or irrelevant data, enhancing the overall efficiency of the classification process.

It can be seen how the Ensemble learning approach proposed in this work is superior to that of the kNN, SVM and DT algorithms. Thus, the algorithm for the use of temporal slices includes the analysis of the evolution of the faults detected over time. In addition, results have been obtained with drastic reduction of time calculation (up to 35% less computing time).

Although the accuracy results presented in Tables 2 to 5 may not match those achieved by image-based approaches as shown in Table 1, it is important to highlight the following aspects. Firstly, our approach

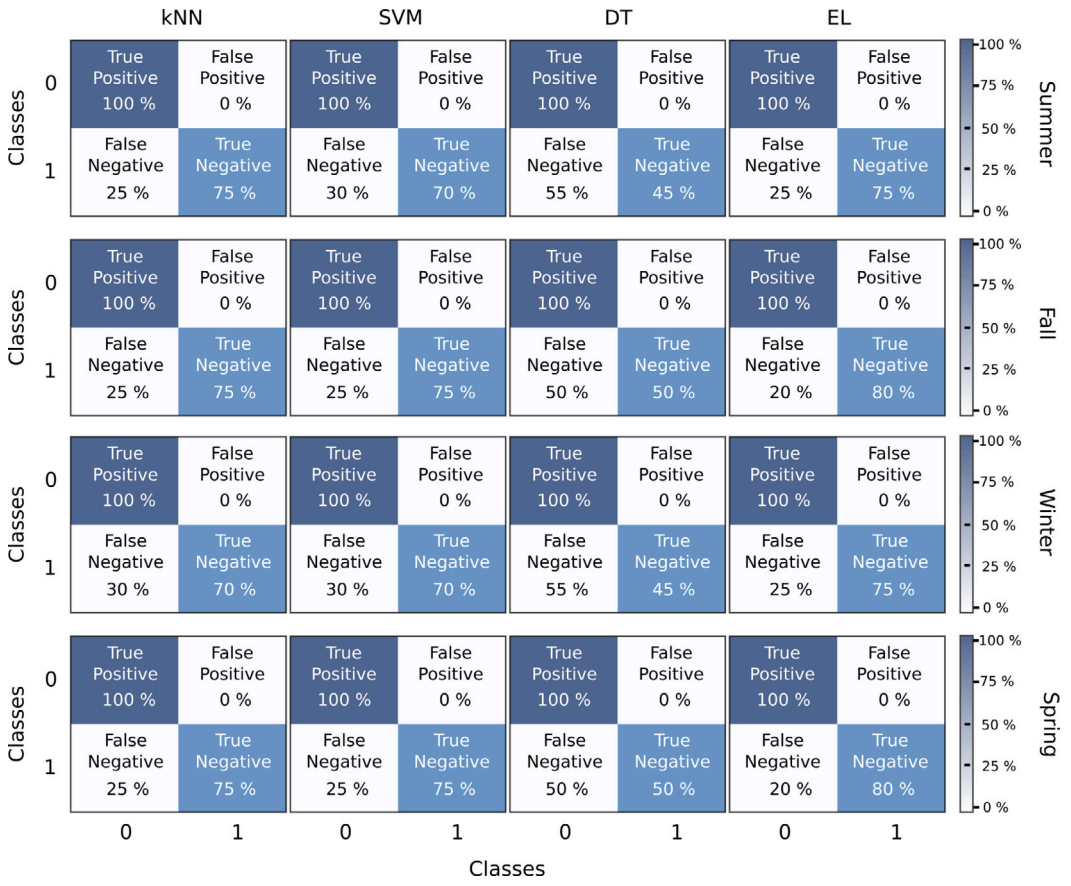


**Fig. 12.** Confusion matrix of the results of the classification algorithms for each season of the year, after dimensionality reduction using PCA. The class 0 corresponds to healthy panels and the class 1 corresponds to panels with a snail trail.

robustly identifies *snail trail* faults under varying irradiation conditions using only one current signal per PV panel. This significantly reduces the required number of samples while maintaining the robustness of fault detection. Additionally, our approach does not require any additional installation to function. In economic terms, it is superior, as it offers an industrially scalable solution. Furthermore, by operating using electrical signals, it can detect a large number of faults, surpassing the limitations of image-based approaches.

Moreover, due to data dimensionality reduction and effective feature extraction using the Multi-resolution signal decomposition, our approach has the potential not only to detect permanent faults like *snail trails* but also to identify rapid changes in electrical signals (such as those generated by arc faults). Finally, in comparison to image-based approaches, our method significantly reduces the computational and memory requirements for fault detection. When comparing the detection results to those achieved by the study in Sepúlveda-Oviedo et al. (2022) presented in Table 1, it can be observed that the results are comparable. However, our approach ensures this level of performance consistently across different seasons with low computational cost. Notably, the algorithm even outperforms the results in Sepúlveda-Oviedo et al. (2022) in the Spring season during low irradiation conditions at the end of the day as seen in Table 5.

The various contributions highlighted make both the proposed approach effective to detect the faults of PV systems and is likely to reduce maintenance costs significantly. As the EL approach is a generic approach, it can be easily extrapolated to a multiple detection of faults with an adaptation of some decision rules or other diagnosis problems in other domains. Finally, the approach proposed in this research can be easily embedded in microprocessors or numerical devices existing in monitoring systems or electrical interfaces like inverters that are able to capture several electrical measurements of strings, panels, or arrays as a function of time.

## 10. Limitations

The main limitation of this approach lies in the selection of the mother wavelet, the decomposition levels for Multi-resolution signal decomposition, and the number of latent components obtained from PCA. In this case, all these aspects were determined through multiple trials. However, the values may vary when detecting a different type of fault, by increasing or reducing the number of features generated by Multi-resolution signal decomposition or the amount of explained variance contained in the first three latent components of PCA.

In the context of our current study, we did not employ cross-validation, to measure the performance of our approach, due to several specific challenges associated with our dataset and the nature of our case study. These challenges include the high specificity and low variability of our snail trails dataset, which is difficult to replicate synthetically and cannot be effectively split for cross-validation without compromising its coherence or representativeness. This is particularly problematic for data with inherent sequentially or time series data. Furthermore, our dataset, while small, exhibits distinct characteristics that the model consistently recognizes, suggesting it can generalize well even in the absence of cross-validation. We also faced the potential risks of overfitting and introducing bias with cross-validation in such a small dataset, alongside the "curse of dimensionality" in high-dimensional data spaces, which can diminish the effectiveness of cross-validation (Debie and Shafi, 2019). Despite these challenges, multiple field verifications and the use of evaluation metrics like the F-score and confusion matrix have demonstrated the model's stability and consistency, supporting its efficacy.

Like any approach based on machine learning, the accuracy of our method is significantly influenced by the correct selection of hyperparameters for each algorithm. This underscores the critical role of parameter tuning in defining the efficacy and reliability of our results.

## 11. Future work

Due to the interesting results obtained, potential directions for future work include the following aspects:

- Integrating the proposed approach into multiple real-time monitoring devices.
- Collecting various electrical signals with parasitic noise signals to assess the algorithm's robustness to noise.
- Testing the proposed approach with other types of permanent and/or temporary faults.
- Modifying the size of the temporal slices to identify the minimum amount of data required for fault detection.
- Adapting the algorithm for testing using a single sliding temporal slice for fault detection.
- Proposing a data normalization algorithm for PV data that allows the algorithm to be applied to new PV plants without the need for retraining. This algorithm will consider operational aspects of the plant (number of panels, orientations, technologies, etc.) and environmental factors (humidity, wind speed, irradiance, and ambient temperature).
- Exploring the use of heuristic or metaheuristic optimization algorithms to fine-tune the parameters of our approach.
- Building an extended database containing the electrical signature of multiple faults.
- Exploring approaches similar to the one proposed in this study for other key elements in energy generation systems, such as storage, conversion, and transmission components.
- Performing a sensitivity analysis to ascertain the influence of hyperparameters on the diagnostic outcome.
- Exploring and comparing other ensemble approaches such as fusion via incremental learning or ensembles using Dempster Shafer.
- Exploring classifiers such as edRVFL - randomized networks as potential replacements for the DT. edRVFL - randomized networks has proven to be faster than some RF algorithms such as Oblique Random Forest or Double Rotation Forest that are themselves much more efficient than DT.

## 12. Conclusions

This research proposes a method for detecting subtle faults known as *snail trails* using an ensemble learning framework called ELDIAG. ELDIAG combines several complementary learning algorithms, including Support Vector Machines, K-Nearest Neighbors, and Decision Trees. It extracts time–frequency characteristics and statistics from the PV current signal of the panel, followed by feature selection and dimensionality reduction. ELDIAG is experimentally validated for all four seasons using data from a real PV string of 16 panels. The results show efficient classification of healthy panels and those with snail trails. Notably, the method relies solely on the electrical current signal, making it suitable for standard PV data acquisition systems.

Based on the results obtained, we consider that our proposal provides unique key elements.:

- The algorithm works without disrupting PV system production.
- It allows temporal analysis of faults' evolution.
- The combination of relative and absolute majority voting aids in fault detection and characterization.
- It detects rapid signal changes through multi-resolution signal decomposition.
- The approach is computationally efficient and can be integrated into various data acquisition systems.
- It reduces the data storage space required for the predictor matrix.
- It successfully detects faults even under low irradiation conditions.

The research focuses on early fault diagnosis, particularly for faults like *snail trails* that may lead to severe issues affecting PV system production. The results underscore the capability of our method to identify panels with subtle faults, even under low irradiation conditions. This ability is crucial for enhancing the production reliability of PV systems. Furthermore, adopting approaches such as the one presented in this study enables early intervention in the PV system (at a stage where power loss is not substantial). Such preemptive actions help prevent the occurrence of more severe failures that could be triggered by snail trails, including hot spots or cracks. These issues, unlike the initial snail trail faults, can lead to significant power losses and pose potential safety risks to operational staff. Moreover, this approach simplifies data collection, storage, and computation, making it cost-effective and easily integrable into various data acquisition systems, including PV inverters and monitoring systems.

The PV industry can significantly benefit from this work by integrating this approach into their monitoring and maintenance systems. It enables early detection of faults, improves the operational efficiency of solar panels, reduces maintenance costs, and extends the lifespan of the equipment. Moreover, it provides a foundation for the development of smarter and more automated monitoring systems. The primary users would be operators of PV plants, maintenance engineers, and solar technology companies that develop monitoring and diagnostic solutions. It may also be of interest to researchers in the field of solar energy and artificial intelligence applied to renewable energy systems.

## CRediT authorship contribution statement

**Edgar Hernando Sepúlveda-Oviedo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Louise Travé-Massuyès:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Audine Subias:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Marko Pavlov:** Writing – original draft, Validation, Resources, Project administration, Funding acquisition. **Corinne Alonso:** Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

## 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.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This project is partially supported by the 3IA Artificial and Natural Intelligence Toulouse Institute (ANITI), French program ‘‘Investing for the Future – PIA3’’ under the Grant agreement n° ANR-19-PI3A-0004, Agence Nationale de la Recherche. It is also related to the SticAmSud project HAMADI 4.0 “Hybrid Algorithms based on Models and Data in Industry 4.0 ”, n° 22-STIC-06.

## References


Ahmadipour, M., Murtadha Othman, M., Alrifaey, M., Bo, R., Kit Ang, C., 2022. Classification of faults in grid-connected photovoltaic system based on wavelet packet transform and an equilibrium optimization algorithm-extreme learning machine. Measurement 197, 111338.

Aizpurua, J.I., Catterson, V.M., Stewart, B.G., McArthur, S.D.J., Lambert, B., Cross, J.G., 2021. Uncertainty-aware fusion of probabilistic classifiers for improved transformer diagnostics. IEEE Trans. Syst. Man Cybern. 51 (1), 621–633.

Amaral, T.G., Pires, V.F., Pires, A.J., 2021. Fault detection in PV tracking systems using an image processing algorithm based on PCA. Energies 14 (21).

Benkercha, R., Moulahoum, S., 2018. Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system. Sol. Energy 173, 610–634.

Bouttier, J., Di Francesco, P., Guitter, E., 2003. Geodesic distance in planar graphs. Nuclear Phys. B 663 (3), 535–567.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408, 189–215.

Dadhich, K., Kurukuru, V.S.B., Khan, M.A., Haque, A., 2019. Fault identification algorithm for grid connected photovoltaic systems using machine learning techniques. In: International Conference on Power Electronics, Control and Automation. ICPECA, pp. 1–6.

Debie, E., Shafi, K., 2019. Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. Pattern Anal. Appl. 22.

Dhere, N.G., Shiradkar, N.S., 2012. Fire hazard and other safety concerns of photovoltaic systems. Journal of Photonics for Energy 2 (1), 1–14.

Eskandari, A., Aghaei, M., Milimonfared, J., Nedaei, A., 2023. A weighted ensemble learning-based autonomous fault diagnosis method for photovoltaic systems using genetic algorithm. Int. J. Electr. Power Energy Syst. 144, 108591.

Esmael, B., Arnaout, A., Fruhwirth, R., Thonhauser, G., 2012. A statistical feature-based approach for operations recognition in drilling time series. Int. J. Comput. Inform. Syst. Indus. Manag. Appl. 4 (6), 100–108.

Fadhel, S., Delpha, C., Diallo, D., Bahri, I., Migan, A., Trabelsi, M., Mimouni, M., 2019. PV shading fault detection and classification based on I-V curve using principal component analysis: Application to isolated PV system. Sol. Energy 179, 1–10.

Ganaie, M., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Eng. Appl. Artif. 115, 105151.

Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., Silvestre, S., 2017. A statistical-based approach for fault detection and diagnosis in a photovoltaic system. In: 2017 6th International Conference on Systems and Control. ICSC, pp. 75–80.

Hong, Y.-Y., Pula, R.A., 2022. Detection and classification of faults in photovoltaic arrays using a 3D convolutional neural network. Energy 246, 123391.

Hout, M.C., Papesh, M.H., Goldinger, S.D., 2013. Multidimensional scaling. WIREs Cognitive Science 4 (1), 93–103.

Jang, Y.I., Sim, J.Y., Yang, J.-R., Kwon, N.K., 2021. The optimal selection of mother wavelet function and decomposition level for denoising of DCG signal. Sensors 21 (5).

Jordan, D.C., Silverman, T.J., Wohlgemuth, J.H., Kurtz, S.R., VanSant, K.T., 2017. Photovoltaic failure and degradation modes. Prog. Photovolt., Res. Appl. 25 (4), 318–326.

Kim, N., Hwang, K.-J., Kim, D., Lee, J.H., Jeong, S., Jeong, D.H., 2016. Analysis and reproduction of snail trails on silver grid lines in crystalline silicon photovoltaic modules. Sol. Energy 124, 153–162.

Köntges, M., Kurtz, S., Packard, C., Jahn, U., Berger, K., Kato, K., Friesen, T., Liu, H., Van Iseghem, M., Wohlgemuth, j., Miller, D., Kempe, M., Hacke, P., Reil, F., Bogdanski, N., Herrmann, W., Buerhop, C., Razongles, G., Friesen, G., 2014. Review of failures of photovoltaic modules. IEA-PVPS.

Lebreton, C., Kbidi, F., Alicalapa, F., Benne, M., Damour, C., 2022. PV fault diagnosis method based on time series electrical signal analysis. Eng. Proc. 18 (1).

Lestary, F.D., Syafaruddin, Areni, I.S., 2022. Deep learning implementation for snail trails detection in photovoltaic module. In: 2022 FORTEI-International Conference on Electrical Engineering. FORTEI-ICEE. pp. 41–46.

Li, B., 2021. Health monitoring of photovoltaic modules using electrical measurements (Ph.D. thesis). Université Paris-Saclay.

Li, X., Yang, Q., jun Yan, W., 2018. Intelligent fault pattern recognition of aerial photovoltaic module images based on deep learning technique. Syst. Cybernet. Inform. 16 (2), 2–5.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67.

Mahesh, B., 2019. Machine learning algorithms—A review. Int. J. Sci. Res. 381–386.

Malik, A., Haque, A., Kurukuru, V.B., Khan, M.A., Blaabjerg, F., 2022. Overview of fault detection approaches for grid connected photovoltaic inverters. e-Prime - Adv. Electr. Eng. Electro. Energy 2, 100035.

Mandal, P., Madhira, S.T.S., haque, A.U., Meng, J., Pineda, R.L., 2012. Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques. Procedia Comput. Sci. 12, 332–337.

Mellit, A., Zayane, C., Boubaker, S., Kamel, S., 2023. A sustainable fault diagnosis approach for photovoltaic systems based on stacking-based ensemble learning methods. Mathematics 11 (4).

Naveen Venkatesh, S., Rebecca Jeyavadhanam, B., Moradi Sizkouhi, A., Esmaili-far, S., Aghaei, M., Sugumaran, V., 2022. Automatic detection of visual faults on photovoltaic modules using deep ensemble learning network. Energy Rep. 8, 14382–14395.

Oulefki, A., Himeur, Y., Trongtiraku, T., Amara, K., Agaian, S., Benbelkacem, S., Guerroudji, M.A., Zemmouri, M., Ferhat, S., Zenati, N., Atalla, S., Mansoor, W., 2023. Unveiling the invisible: Enhanced detection and analysis of deteriorated areas in solar PV modules using unsupervised sensing algorithms and 3D augmented reality. Comput. Vis. Pattern Recognit. 1–13.

Packard, C.E., Wohlgemuth, J.H., Kurtz, S.R., 2012. Development of a Visual Inspection Data Collection Tool for Evaluation of Fielded PV Module Condition. National Renewable Energy Lab.(NREL), Golden, CO, USA.

Ray, P.K., Mohanty, A., Panigrahi, B.K., Rout, P.K., 2018. Modified wavelet transform based fault analysis in a solar photovoltaic system. Optik 168, 754–763.

Samko, O., Marshall, A., Rosin, P., 2006. Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognit. Lett. 27 (9), 968–979.

Santhakumari, M., Sagar, N., 2019. A review of the environmental factors degrading the performance of silicon wafer-based photovoltaic modules: Failure detection methods and essential mitigation techniques. Renew. Sustain. Energy Rev. 110, 83–100.

Sepúlveda-Oviedo, E.H., Travé-Massuyès, L., Subias, A., Alonso, C., Pavlov, M., 2021. Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems. In: X Congreso Internacional IngenierÍa MecáNica, MecatrÓNica Y AutomatizaciÓN (XCIMM). Bogotá, Colombia, pp. 1–2.

Sepúlveda-Oviedo, E.H., Travé-Massuyès, L., Subias, A., Alonso, C., Pavlov, M., 2022. Feature extraction and health status prediction in PV systems. Adv. Eng. Inform. 53, 101696.

Sepúlveda-Oviedo, E.H., Travé-Massuyès, L., Subias, A., Pavlov, M., Alonso, C., 2023a. Fault diagnosis of photovoltaic systems using artificial intelligence: A bibliometric approach. Heliyon 9 (11), e21491.

Sepúlveda-Oviedo, E.H., Travé-Massuyès, L., Subias, A., Pavlov, M., Alonso, C., 2023b. Detection and classification of faults aimed at preventive maintenance of PV systems. In: XI Congreso Internacional de Ingeniería Mecánica, Mecatrónica y Automatización 2023. Universidad Nacional de Colombia, Cartagena, Colombia, pp. 1–2.

Srikanta Murthy, A., 2018. Islanding detection technique using slantlet transform and ridgelet probabilistic neural network in grid-connected photovoltaic system. Appl. Energy 231, 645–659.

Tigo, 2023. Tigo TS4-A-O. Tigo Energy, Available from: URL https://fr.tigoenergy.com/product/ts4-a-o (Accessed on 18 October 2023).

Vasanth, J., Venkatesh S, N., Sugumaran, V., Mahamuni, V., 2023. Enhancing photovoltaic module fault diagnosis with unmanned aerial vehicles and deep learning-based image analysis. Int. J. Photoenergy 2023, 1–17.

Venkatesh, S.N., Sugumaran, V., 2022. A combined approach of convolutional neural networks and machine learning for visual fault classification in photovoltaic modules. Proc. Inst. Mech. Eng. O 236 (1), 148–159.

Venkatesh S, N., Balaji, A., Chakrapani, G., Annamalai, K., Aravinth, S., Anoop, P., Sugumaran, V., Mahamuni, V., 2023. Photovoltaic module fault detection based on deep learning using cloud computing. Sci. Program. 2023, 1–10.

Wang, Y., Pan, Z., Pan, Y., 2020. A training data set cleaning method by classification ability ranking for the $k$-nearest neighbor classifier. IEEE Trans. Neural Netw. Learn. Syst. 31 (5), 1544–1556.

Xu, L., Pan, Z., Liang, C., Lu, M., 2022. A fault diagnosis method for PV arrays based on new feature extraction and improved the fuzzy C-mean clustering. IEEE J. Photovolt. 12 (3), 833–843.

Yi, Z., Etemadi, A.H., 2017a. Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. IEEE Trans. Smart Grid 8 (3), 1274–1283.

Yi, Z., Etemadi, A.H., 2017b. Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine. IEEE Trans. Ind. Electron. 64 (11), 8546–8556.

Zhang, Z., Han, H., Cui, X., Fan, Y., 2020. Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems. Appl. Therm. Eng. 164, 114516.

Zhang, M.-L., Zhou, Z.-H., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognit. 40 (7), 2038–2048.