Fault diagnosis in low voltage smart distribution grids using gradient boosting trees

Sapountzoglou, Nikolaos; Lago, Jesus; Raison, Bertrand

DOI
10.1016/j.epsr.2020.106254

Publication date
2020

Document Version
Final published version

Published in
Electric Power Systems Research

Electric Power Systems Research 182 (2020) 106254

Contents lists available at ScienceDirect

Electric Power Systems Research

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

Fault diagnosis in low voltage smart distribution grids using gradient boosting trees

Nikolaos Sapountzoglou[a,1,*], Jesus Lago[b,c,1], Bertrand Raison[a]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab, 38000 Grenoble, France
[b] Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, Delft, The Netherlands
[c] Algorithms, Modeling, and Optimization, VITO, Energyville, ThorPark, Genk, Belgium

ARTICLE INFO

Keywords:
Fault diagnosis
Fault detection
Fault identification
Low voltage distribution system
Machine learning
Gradient boosting trees

ABSTRACT

In this paper, a gradient boosting tree model is proposed to detect, identify and localize single-phase-to-ground and three-phase faults in low voltage (LV) smart distribution grids. The proposed method is based on gradient boosting trees and considers branch-independent input features to be generalizable and applicable to different grid topologies. Particularly, as it is shown, the method can be estimated in a specific grid topology and be employed in a different one. To test the algorithm, the method is evaluated in a simulated real LV distribution grid of Portugal. In this case study, different fault resistances, fault locations and hours of the day are considered. In detail, the algorithm is evaluated at eighteen fault resistance values between 0.1 and 1000 Ω; similarly, nine fault locations are considered within each one of the 32 sectors of the grid and the faults are simulated across different hours of a day. The developed algorithm showed promising results in both out-of-sample branch and fault resistance data especially for fault detection, demonstrating a maximum fault detection error of 0.72%.

1. Introduction

In order to tackle the climate change threat, more and more forms of renewable energy sources are being installed in the grid. The integration of those distributed sources comes with many challenges that increase the complexity of the grid and introduce a lot of uncertainty. On the other hand, advances in the rapidly evolving field of smart grids and the increased functionalities they bring, e.g. installation of smart meters, enhance the monitoring capabilities of the system operators. In this context, fault diagnosis processes are needed in order to pave the way towards a self-healing electrical network.

Faults at the distribution level account for eight out of ten cases of customer electricity interruptions [1]. While their societal and economical impact is huge, it is very difficult to calculate the cost of a power outage as it is a multivariant equation with many factors that are difficult to estimate, e.g. customer behavior or company reliability loss [2]. The famous blackout of 2003 in the USA and Canada resulted in an estimated cost of $6 billion. Another outage event the same year in Italy was reported to have caused a damage of €120 million to the local economy [2]. In an attempt to measure the impact of the faults on the customers, the value of lost load (VoLL) is used. The VoLL (€/kWh) is defined as the ratio of the economic value of leisure in households over the total household consumption. An annual average of 8.37 €/kWh was measured in Europe in 2013 [3]. Despite the grave effects of electricity interruptions described above, even today, many utilities are relying on customer phone calls to detect or localize a fault [4].

In the literature, several attempts have been made to automatize the fault detection and location process and minimize human interference [5]. The two most widely used fault location methods are the impedance-based and traveling wave methods; these methods are thorough analyzed by the IEEE standard C37.114-2014 [6]. A study reviewing the different available impedance-based methods is also provided in [7]. Besides impedance-based and traveling wave methods, other methods also exist: sparse measurements [8–11], artificial intelligence [12–15], as well as hybrid methods [16,17] that have attracted the researchers' attention over the last years.

While several methods have been proposed in the literature, they all have several underlying problems. Impedance-based methods, although being the most widely used method for fault location applications, they have a big problem: when using them, there is an underlying risk of identifying multiple fault locations belonging to different branches but of the same distance from the beginning of the feeder [4,18]. The latter, can be quite misleading in reality when a crew is sent to restore the power. While traveling wave methods have a higher accuracy, they also

* Corresponding author.
E-mail address: nikolaos.sapountzoglou@g2elab.grenoble-inp.fr (N. Sapountzoglou).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.epsr.2020.106254
Received 4 September 2019; Received in revised form 28 January 2020; Accepted 29 January 2020
Available online 12 February 2020

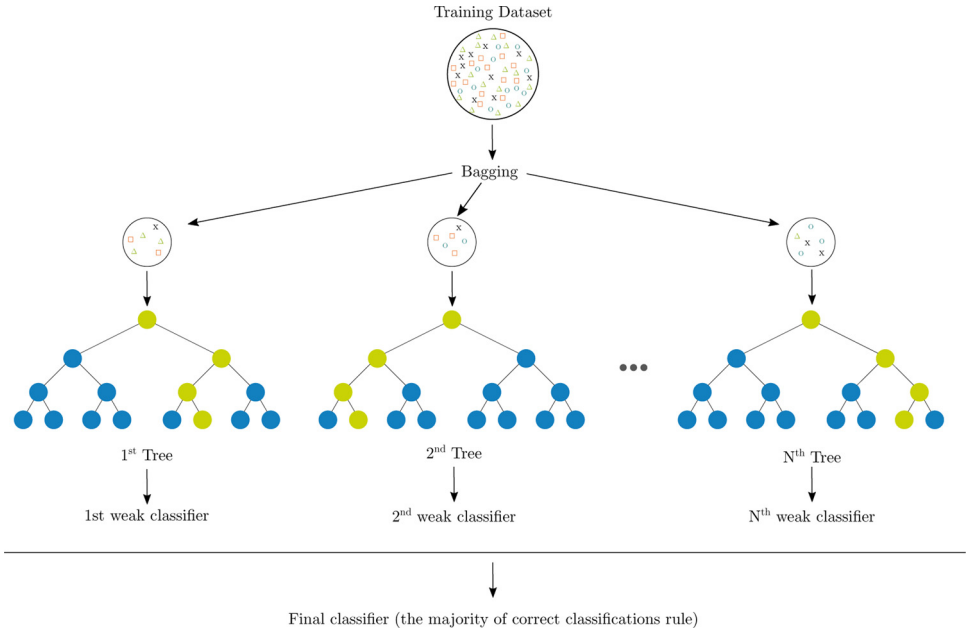

Fig. 1. Example of a random forest.

present several disadvantages: (a) they rely on the detection of the wavehead, which sometimes can be quite challenging, (b) they depend on the line parameters, which on distribution systems vary a lot since the speed of the wave is based on the inductance and capacitance of the lines, and (c) they are vulnerable to external signal interference [4,19,18]. Similarly, other methods like sparse measurements or hybrid methods can be very demanding in terms of equipment and quite costly [4]. Finally, knowledge-based methods face the danger of not identifying the fault in the case where it has not been part of the training scenarios [4,18].

A second problem with the existing work in the literature is that it has been limited to *medium voltage (MV)* distribution grids and low-impedance faults. In particular, a very limited amount of studies examined fault cases of a fault resistance higher than 100 Ω [15,20–22]. Similarly, to the best of the authors' knowledge, only a few methods were applied to *low voltage (LV)* grids [23–26] of which the maximum studied fault resistance was 6 Ω [23]. Considering that a large amount of faults appear in LV distribution grids and that the fault resistance range is in practice between 1 and 1000 Ω, it becomes clear that the existing literature is very limited. Moreover, as the larger the fault resistances are the harder it becomes to detect and identify the fault, there is a pressing need for LV fault diagnosis methods that can detect and identify high resistance faults.

The aim of this work is to address the issues raised above and explore aspects of fault detection, identification and location in a LV distribution grid under both low and high resistance faults with fault resistances ranging from 0.1 to 1000 Ω. In particular, to overcome the disadvantages of the traditional methods, a new artificial intelligence method is proposed in this paper based on *gradient boosting trees (GBT)*. The proposed method can detect, identify and locate both single-phase-to-ground faults, the most frequent ones, and three-phase faults, the most severe ones. The main advantage of GBT is its very fast estimation, which in turn makes it implementable in real-time applications. The contribution of this paper is fourfold and is summarized below:

(a) A method for fault detection and faulty feeder identification: the occurrence of the fault is detected with a simultaneous identification of the feeder under fault.
(b) A method for fault type identification: a distinction of the faulty and non-faulty phases is achieved thus identifying the fault type, single-phase-to-ground (AG, BG or CG) or three-phase fault.
(c) A method for faulty branch identification: following the feeder and phase identification, the faulty branch within a faulty feeder is also identified.
(d) A method that is topology-independent: unlike literature methods, the proposed approach is generalizable and applicable to different grid topologies. Particularly, the method can be estimated in a specific grid topology and be employed in a different one.

The paper is organized as follows. In the following section, an explication of the developed method is provided. In the third section, the LV distribution grid case study is analyzed. Furthermore, the obtained results are presented in the fourth section. Finally, the conclusions are drawn in the last section.

2. Method

2.1. Model definition

The GBT algorithm [27] is a prediction model based on the principle of combining several regression trees. In particular, regression trees are models characterized by either having high bias and low variance errors if the tree is shallow, or low bias and high variance errors if the tree is deep. To solve this issue, there are two families of algorithms that combine several regression trees to reduce high errors.

The first family is random forests and it is based on the principle of bagging [28], i.e. combining models with low bias and high variance error in order to reduce the variance while keeping a low bias. In particular, the original training dataset is first sampled with replacement to create different bagged samples. Then, for each of the samples, a deep tree is trained, i.e. a model with high variance and low bias error; as the bagged samples are all different from each other, the prediction of each tree is different. Finally, the final prediction is built using the majority voting rule of all the decision trees. Fig. 1 depicts an example of a random forest algorithm.

The second family is gradient boosting trees and it is based on the principle of boosting [28], i.e. combining models with high bias and low variance error in order to reduce the bias while keeping a low variance. In detail, instead of using deep trees and different training datasets, boosting trees employ shallow trees that are trained in the

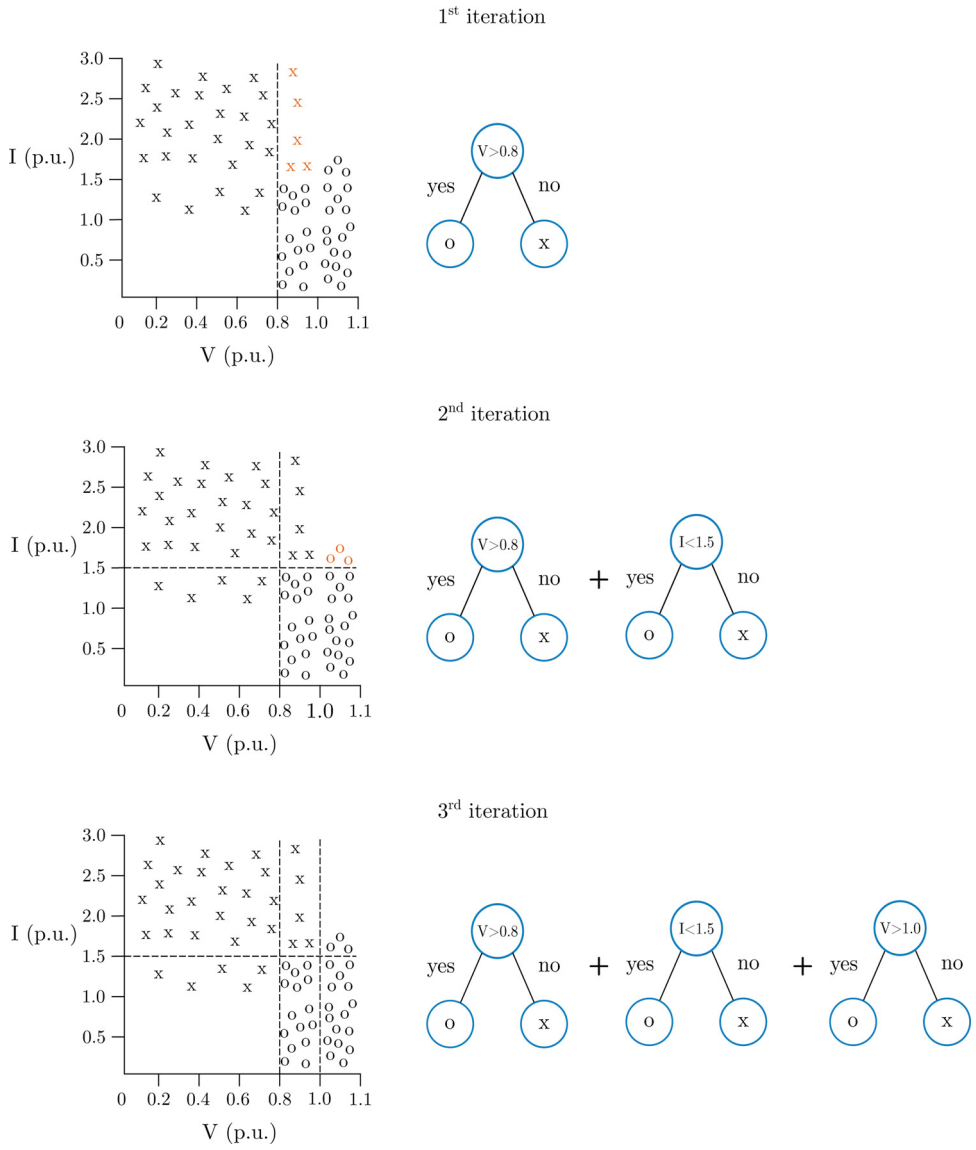

Fig. 2. Gradient boosting tree working principle for a simplified detection task. Healthy data are marked with circles while data under faulty operation (characterized mainly by a current increase) with an x. Missclassified datapoints are marked with red.

same dataset but where each tree is specialized in a specific characteristic of the input–output relation. In particular, successive shallow trees are trained in series, where the nth tree is trained with the goal of reducing the prediction errors of the previous $n-1$th trees. Fig. 2 illustrates a simplified example of fault detection using boosting trees. In the figure, there are datapoints representing healthy and faulty operation scenarios, together with the voltage–current measurements for each scenario. To distinguish between faulty and healthy, a first tree is estimated: the tree draws a decision boundary based on a voltage value. Next, a second tree is estimated to correct the missclassified samples of the first tree; this tree draws a second boundary using a current value. Finally, a third tree is estimated to correct the errors of the first two. At the end, the prediction of the model is based on the serial combination of the three trees.

The reason for selecting this algorithm and no other, e.g. a neural network, was threefold: (a) this has been shown to outperform other regression tree methods and has recently become the winner of several challenges in Kaggle, a site that hosts machine learning competitions; (b) it has been successfully used in other energy-based applications, e.g. forecasting electricity prices [29] or solar irradiance forecasting [30]; (c) it is a very fast model to train which allows real-time applications.

2.2. Algorithm functionality

The proposed algorithm has three distinct functionalities:

1. **Fault detection:** the first functionality of this algorithm is the detection of a fault occurrence with a simultaneous identification of the feeder under fault.
2. **Fault type identification:** an extra element which is often omitted by fault location algorithms is the fault type identification process. In this study, the proposed algorithm can also differentiate faulty from non-faulty phases.
3. **Faulty branch identification:** the last functionality of this proposed method is the faulty branch identification, which is the faulty branch within a faulty feeder.

2.3. Working principle

The main idea of the algorithm is to make use of its prediction capabilities to diagnose the grid faults. In particular, the algorithm uses a training dataset $\mathbb{S} = \{X_i, Y_i\}_{i=1}^{N}$, where X are the inputs of the GBT model and Y are the desired predicted output. For all the identification

tasks, the inputs X are the same: specific data corresponding to a specific branch, e.g. voltage on that branch. The outputs Y however depend on the specific task. Particularly, the output Y changes with the task as the algorithm has a slightly different working principle for each of the three tasks:

- **Fault detection:** to identify a faulty feeder, the algorithm considers data from healthy branches in healthy feeders and data from faulty branches. Then, it labels the healthy branches with a 0 and faulty branches with a 1 and the algorithm is trained to predict 0 or 1 to indicate the existence of a fault in a branch. In real time, to identify a faulty feeder, the algorithm is simply tested on all the branches of a feeder.
- **Fault type identification:** to identify the type of fault, the algorithm considers only data from faulty branches. Then, it labels each branch datapoint with 1, 2, or 3 to respectively denote single-phase fault in phase A, B and C, and uses a label 4 to denote three-phase faults. In real time, to identify the fault, the algorithm is simply tested on the faulty branch.
- **Faulty branch identification:** to identify the faulty branch within a faulty feeder, the algorithm considers data from healthy branches in a faulty feeder and data from faulty branches. Then, the algorithm is trained to distinguish between the two cases using two labels, i.e. 0 and 1. In real time, to identify the branch, the algorithm is tested on the branches of a faulty feeder.

A simplified representation of the proposed model for the three diagnosis tasks is depicted in Fig. 3. As can be seen, in all three tasks, sequential shallow trees are trained to identify the correct label in each task, where successive trees are estimated to improve upon the error of previous trees. Independently of the task, the trees take all type of variables into account to correctly identify the labels: voltage threshold, current values, voltage in one node larger than a voltage in another node, etc. The main difference between the tasks is the output of the model: while the fault detection and faulty branch identification tasks output 0 or 1 depending on whether a fault exist, the fault type identification task outputs 1, 2, 3 or 4 to indicate which of the four possible faults has occurred.

2.4. Training

Independently of the task, as they are all classification tasks, the algorithm is trained to minimize the cross-entropy loss of the training dataset. Moreover, to optimize the structure of the algorithm, all the boosting tree hyperparameters, e.g. number of branches or tree depth, are optimally selected using the Bayesian optimization [31]. In particular, the dataset is divided in three subsets: a training dataset, a validation dataset and a test dataset. The training dataset is used to estimate the algorithm parameters, the validation dataset is used to estimate the algorithm hyperparameters and finally the test dataset is used to evaluate the quality of the algorithm.

2.5. Input features

In terms of the inputs of the model several design choices were made. In particular, to make the model general enough, i.e. to make the model applicable to different grid topologies with various number of branches and available measurements, two design choices were made:

- First, the use of branch-specific features was avoided, e.g. the branch length or the branch resistances and reactances.
- Second, all branch-specific measurements were substituted with a fixed number of interpolated values so that each branch could have the exact same number of features. For instance, independently of the number of voltage measurements in a branch, five equally spaced points within the branch were selected and the voltage values from the voltage measurements were interpolated to these five locations.

With that motivation, in order to identify if a fault occurs at time t the following input features were considered:

1. *Time*: the hour of the day corresponding to t. This is important because the load and microgeneration penetration in the grid change along the day.
2. *Load*: the load in the grid at time t.
3. *Generation*: the microgeneration in the grid at time t.
4. *Current at time t*: the current at the beginning of each feeder at time t was considered as shown in Fig. 4. In particular, the current through the three phases and the neutral.
5. *Current 5 min before t*: the current at the beginning of each feeder five minutes before t was also considered. As before, current through the three phases and the neutral was considered. These features are important to have a comparison between two points close in time so that if a fault occurs at time t, the method can compare the current at time t with the values of the current during normal operation.
6. *Voltages at time t*: voltage values across each branch at time t were considered. More specifically, as mentioned before, five virtual/interpolated equally spaced measurements that were obtained from the real measurements in the branch were considered. Moreover, the voltages for each phase were considered, i.e. in total fifteen voltage points per branch.
7. *Voltage 5 min before t*: voltage values across the branch five minutes before t were also considered. The same fifteen voltage points as in time t were used. As with the current, the motivation behind these input features is to provide the method with voltage measurements during normal operation.

2.6. Computation time

A key advantage of the current algorithm is that it only needs to be trained periodically. In particular, for real-time fault diagnosis, the method simply evaluates a boosting tree model. As a result, the computational cost of the method is independent of the training dataset and nearly-independent of the grid size. Moreover, its computation cost in real time is in the order of milliseconds, which makes it very suitable for real-time applications.

In terms of training, the algorithm is also very fast: training a boosting tree model is done in less than 1 min. Therefore, as new data become available, the method is also very suitable for continuous adaption, e.g. hourly or daily, to environmental changes.

These two properties are key as they lead to a simple, yet accurate, fault diagnosis method that does not require complex techniques, e.g. data clustering or data reduction, to decrease the computational cost.

2.7. Representation

To provide a better understanding of the method, Fig. 5 represents the different components of the proposed methodology and how they relate to each other.

3. Case study

3.1. Grid structure

In order to apply the developed method to a real case scenario, the semi-rural LV distribution grid of Portugal that was provided by Efacec and is presented in Fig. 4, was used. This grid is a three-phase-four-wire one where the neutral is solidly grounded. Moreover, it incorporates eighteen single-phase photovoltaic installations and forty eight also single-phase loads in different nodes, attributing thus an unbalanced nature to the grid in terms of topology. Heterogeneity is yet another

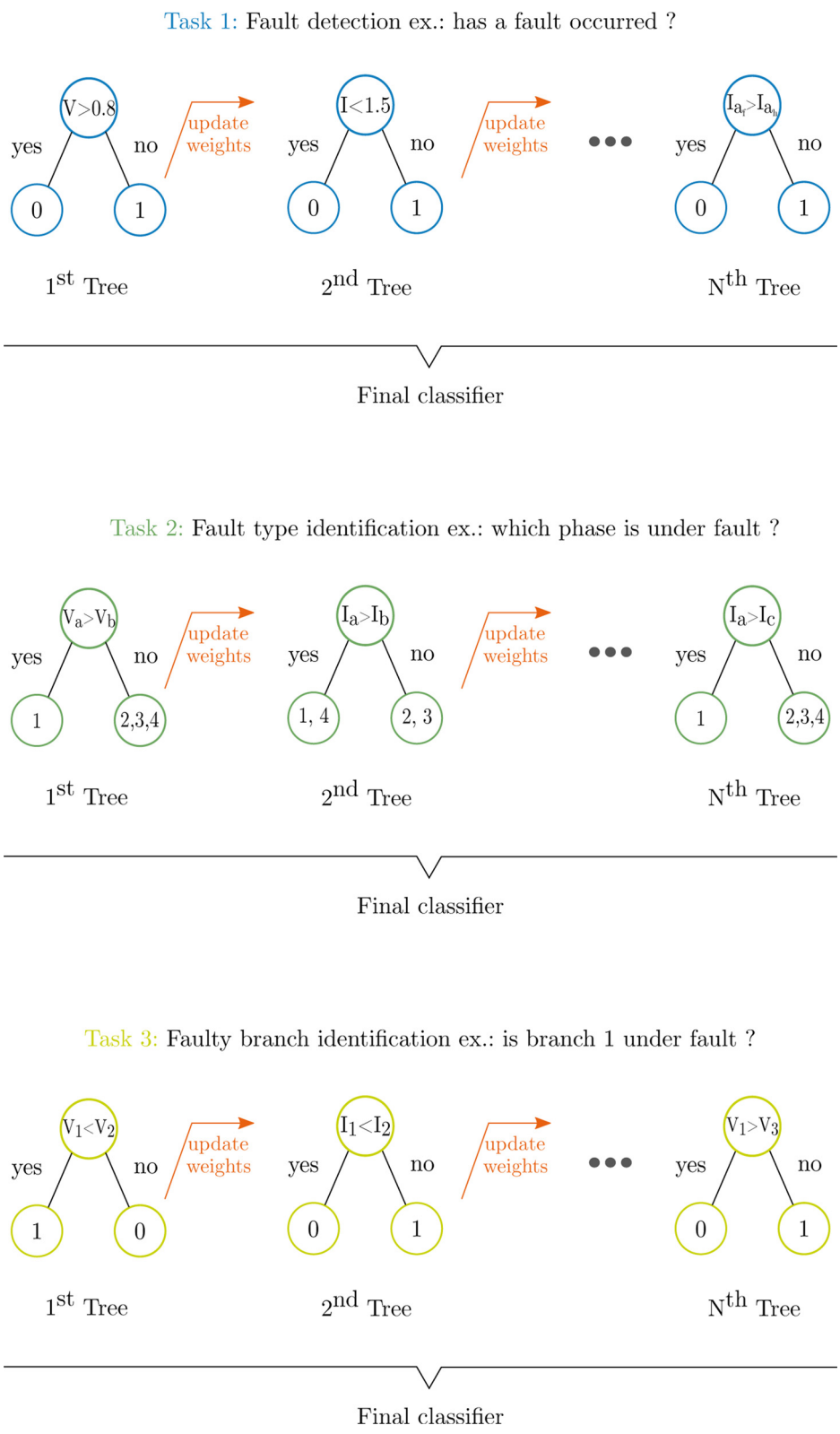

Fig. 3. Gradient boosting tree examples per task where 0 and 1 label a healthy and faulty state respectively for tasks 1 and 3. For the fault type identification task the number 1 to 4 correspond the faulty phase (1 for AG fault, 2 for BG, 3 for CG and 4 for ABC).

feature of this grid since conductors of various lengths, resistances and reactances are used to connect the nodes.

Fig. 4 is also helpful to define the grid sector and branch. A sector is the segment of the grid between two nodes, e.g. the part of the grid connecting nodes three and six would be a sector. In this grid of 33 nodes, 32 sectors can be defined. At the same time, a branch is a unique chain of sectors and in this grid topology nine different branches can be identified (Fig. 4).

Finally, the available measurements considered were: (a) phase rms voltage measurements in every node of the grid and (b) phase rms

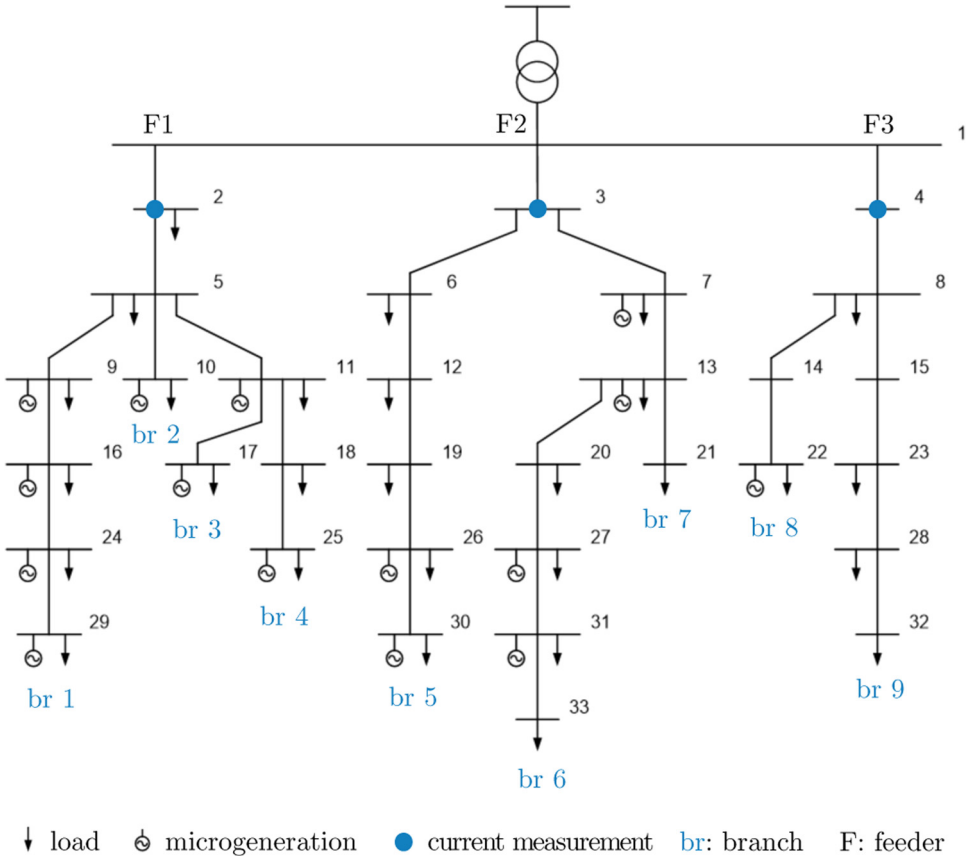

Fig. 4. LV distribution grid of Portugal.

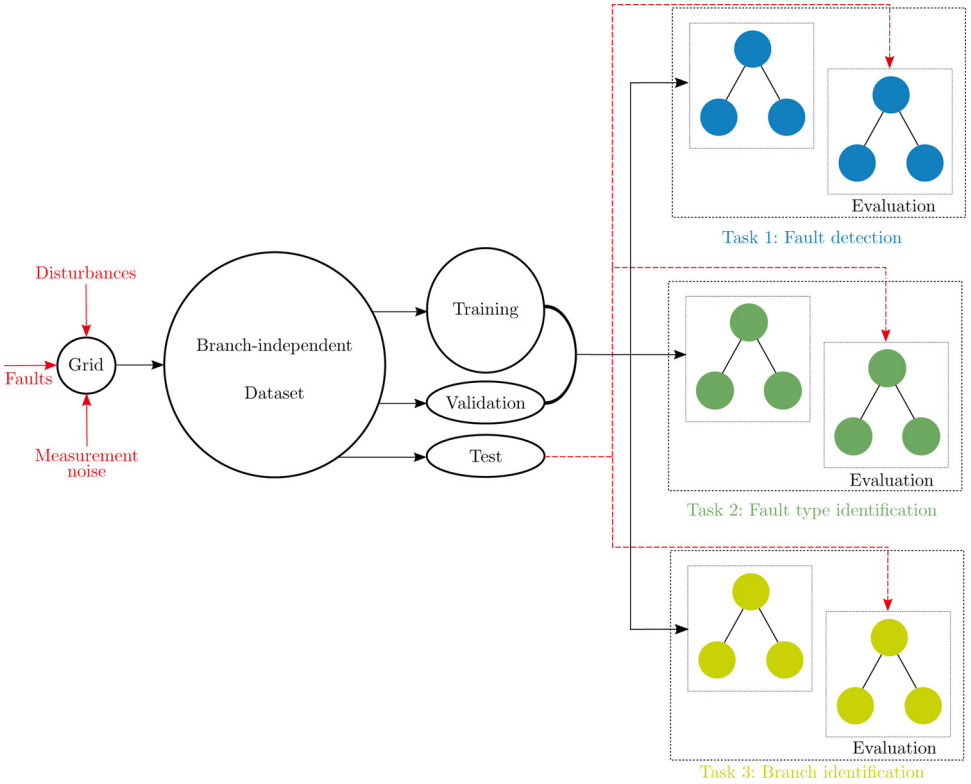

Fig. 5. Conceptual representation of the method implementation.

current measurements in the beginning of each feeder where in theory a sudden increase of the current is expected upon the occurrence of a fault.

3.2. Simulation environment

In order to obtain the necessary data, a realistic simulation model of the studied LV grid, designed in the MATLAB/Simulink environment, was employed. The simulation model was provided by the company Efacec [32]. The simulation environment provides as output: (a) nodal phase rms voltage measurements and (b) phase rms current measurements from the beginning of each feeder. Moreover, the use of phasor mode for the simulations reduces heavily the computational time without compromising the accuracy of the measurements. Additionally, the environment is suitable for both normal and faulty operation simulations. Finally, the simulation environment provides several configurable options such as: (a) the sampling frequency which in this case was set at 50 ms to further reduce the computation time, (b) erroneous measurements and (c) different daily generation and load profiles.

3.3. Grid effects

To simulate the most realistic conditions, five different effects were identified and considered in this study:

1. *Fault resistance*: As explained in the motivation, very few studies were reported in the literature that cover high resistance faults in LV grids. In this case, 15 different fault resistances were investigated: 0.1, 0.5, 1, 3, 5, 7.5, 10, 30, 50, 75, 100, 300, 500, 750 and 1000 Ω, covering the full spectrum of faults, both low and high resistance ones. In addition, to test the algorithm under unknown fault scenarios, three extra fault resistances were also considered: 4, 40 and 400 Ω.
2. *Fault location*: Faults in every sector of the grid were considered (32 sectors in total). In every sector, nine possible locations of fault occurrence were considered for distances of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% from the beginning of the sector.
3. *Fault types*: The following fault types were examined: single-phase-to-ground faults and three-phase faults. For every single-phase-to-ground fault case, all three phases: A, B and C were considered.
4. *Microgeneration penetration and load uncertainty*: The studied microgeneration and load profiles are provided in Table 1. During the simulations, these two variables were sampled assuming a uniform distribution with a 20% (generation) and 2% (load) interval length, and mean following the generation/load profiles defined in Table 1.
5. *Measurement noise*: As stated before, this study considered phase rms voltage measurements at every node and phase rms current measurements (including the neutral) at the beginning of each feeder. In order to approximate real measurement conditions as much as possible, a 2% underestimation measurement error was introduced.

3.4. Data recording and generation

To generate the data for the study, the grid was simulated using a *Monte-Carlo sampling (MCS)* technique. Particularly, while directly sampling faulty data is not possible (the distribution of that data is unknown), the distribution of the inputs affecting the grid are known and a simulated model of the grid is available. In this context, MCS can be performed for the inputs of the grid, and then the simulation of the grid with those inputs follows, leading to the generation of the desired data. In detail, as defined in Section 3.3, the distribution of the following variables is considered:

1. The noise in voltage measurements: modeled assuming a 2% underestimation uniform error.
2. The noise in current measurements: modeled assuming a 2% underestimation uniform error.
3. The location of the fault: modeled assuming nine uniformly distributed locations per sector.
4. The grid load: modeled using a uniform distribution with a 2% interval length and mean following the load profile defined in Table 1. During simulation, an hour of the day is first uniformly sampled and then the load is sampled using the uniform distribution.
5. The grid PV generation: modeled and sampled similarly to the load, but with a uniform distribution with a 20% interval length deviation.

Then, for each of the fault resistances, fault types and grid sectors (see Section 3.3), MCS were used to sample these five variables and simulate the grid. This sampling-simulation procedure is repeated multiple times to generate the required datasets.

To generate data representing faulty conditions, a total of 72 datapoints are sampled for each fault resistance, fault type and grid sector. That leads to a regular dataset (in-sample fault resistances) of 165,888 datapoints, and an extra dataset of 27,648 datapoints representing out-of-sample fault resistances.

To generate data representing healthy conditions (needed for the algorithm to distinguish between faulty and normal operation), a total of 300 datapoints are sampled for each hour of the day and for each grid branch. This leads to a dataset containing 64,800 datapoints representing healthy conditions.

It is important to note that faulty operation measurements are taken 150 ms after the fault occurrence. This choice was made for the fault to be as close to the steady-state as possible and to avoid corruption of the data by the activation of any protective element.

3.5. Implementation

The algorithm was implemented in `python` using the `XGBoost` [27] library for the GBT model, and the `hyperopt` [33] library to perform the hyperparameter optimization based on Bayesian optimization.

3.6. Model training and evaluation

The algorithm is trained and evaluated using the regular dataset (in-sample fault resistances) and the dataset of healthy data. Both dataset together comprise a total of 230,688 datapoints, which are randomly divided into the training, validation and test datasets as defined in Table 2.

The model was repeatedly trained with the training dataset and the algorithm was evaluated in the validation dataset for guiding the Bayesian optimization algorithm to find the optimal parameters. Then, after the optimal hyperparameters were found, the algorithm was evaluated in the test dataset.

In addition, to have a model that generalizes to different grid faults,

Table 1
Default microgeneration and load profiles from a typical day in Portugal expressed in percentages (%).

Hour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
μgen	0	0	0	2	9	30	54	60	86	88	73	60	100	83	49	44	14	16	3	0	0	0	0	0
Load	30	28	25	23	20	20	23	30	40	43	46	50	50	55	60	60	55	50	65	85	90	90	75	55

Table 2
Dataset sizes.

Dataset type	Size
Train	115,344 (50%)
Validation	46,138 (20%)
Test	69,206 (30%)

the algorithm was evaluated in out-of-sample fault resistances. In particular, to test the algorithm performance against unknown fault scenarios, an extra test dataset was created comprising out-of-sample 4, 40, and 400 Ω fault resistances. As defined in Section 3.4, this extra test dataset comprises 27,648 datapoints.

Similarly, to have a model that generalizes to different grid topologies, the algorithm was also evaluated against out-of-sample branches. Particularly, in order to test the generalizability of the method to other grid topologies, a training and validation datasets were built comprising only data from the first and third feeder, i.e. branches 1–4 and 8–9 respectively. Then, the method was evaluated in a test dataset comprising data from the second feeder, i.e. branches 5–7. This choice is justified as feeder one and three have the maximum and minimum number of branches respectively. In that way, as it will be shown in the next section, the algorithm was able to provide promising results not only on fault resistances and branches that appear in the training dataset, but also in out-of-sample fault resistances and branches.

3.7. Comparison with similar works

Due to a lack of available research papers in the LV grid, two methods designed for MV distribution grids were employed to compare the algorithm performance in addition to a conventional method for LV grids. These references were used to compare the faulty branch identification results.

In the first case [34], the authors developed a general fault location method based on voltage and current measurements at the point of common coupling of distributed generators. They considered all the different types of faults, i.e. single-phase-to-ground, double-phase-to-ground, phase-to-phase and three-phase. Moreover, they studied faults in three possible locations within a sector at distances of 5%, 50% and 95% from the beginning of each sector. However, the maximum fault resistance value for phase-to-phase and three-phase faults was 5 and 50 Ω for the rest.

In the second study that was used as reference [21], the authors developed a method based on real time state estimation that detects faults and identifies faulted lines. The authors considered single-phase-to-ground, double-phase-to-ground and three-phase faults. Furthermore, they considered only two possible fault locations within a faulty sector, at the middle of the line and at a distance equal to 25% of the sector's length. Although they investigated high–impedance faults of up to 1000 Ω, the data they presented for such high fault resistances were applicable only to a single fault case of an unearthed neutral. For the rest of the cases, the maximum fault resistance they tested was that of 100 Ω.

In the third study (the only one from the LV case) [35], the authors used a conventional criterion for determining the faulty branch within a faulty feeder. The faulty branch was considered to be the one that presented the highest voltage drop. This method was tested against the simulation scenarios that were described above (under the same case study) and the results are presented in the following section.

All these information are gathered in Table 3 where the considerably bigger number of fault scenarios that were considered in this study is demonstrated.

4. Results

As a first step, the algorithm was trained with data from all three feeders of the grid. Then, its performance was tested against out-of-sample fault resistance data. As it was mentioned before, the out-of-sample fault resistances that were chosen were the 4, 40 and 400 Ω. The next step was to expose the algorithm to out-of-sample branches. As explained in the motivation, the purpose of this test was to verify the robustness of the algorithm against changes in the topology of the grid.

For all the results, the following definition of accuracy was used:

$$\text{Acc } (\%) = \frac{(\text{true positives } + \text{ true negatives})}{\text{total number of samples}} \times 100 \quad (1)$$

where true positives are the correctly identified faulty branches and true negatives the correctly identified as non-faulty.

4.1. Fault detection

The first functionality of this algorithm is the detection of a fault occurrence with a simultaneous identification of the feeder under fault. The results are presented in Fig. 6 for the test dataset. An accuracy of 100% is achieved across all fault resistances. Such a level of accuracy renders the proposed method a completely reliable tool for fault detection and feeder identification problems.

4.2. Fault type identification

An extra element which is often omitted by fault location algorithms is the fault type identification process. In this study, the proposed algorithm was also implemented to differentiate faulty from non-faulty phases. The obtained results for the test dataset are depicted in Fig. 7.

In Fig. 7, the first effect of the increase of fault resistance is noticed. For low–resistance faults (below 10 Ω), the accuracy of faulty phase identification is maintained at a level higher than 98%. After that, a more and more significant decrease of accuracy is noticed with the increase of the fault resistance down to minimum of 86.7% for 1000 Ω. This was an expected result as the increase of fault resistance will decrease the voltage drop during a fault and thus bring the voltages across a faulty branch closer to the values of normal operation. The limit of 10% voltage drop proposed by the EN50160 standard for LV grids, is very likely to be violated making it more difficult for the algorithm to distinguish faulty from normal operating phase for fault resistance values higher than 10 Ω.

Moreover, Table 4 shows the accuracy of the method for each of the fault types in the same test dataset. A similar performance was noticed in all four types of faults: single-phase-to-ground (AG, BG, CG) and

Table 3
Comparison of different case studies of similar works.

Parameters	Brahma [34]	Pignati et al. [21]	This paper
Grid	12.4 kV (MV), U.S.A.	10 kV (MV), The Netherlands	400 V (LV), Portugal
Fault types	1ph-G, 2ph-G, ph-ph, 3ph	1ph-G, 2ph-G, 3ph	1ph-G, 3ph
Fault resistance	1–5 (ph-ph, 3ph), 1–50 (1ph-G, 2ph-G)	1, 100, 1000[a]	0.1, 0.5, 1, 3, 5, 7.5, 10, 30, 50, 75, 100, 300, 500, 750, 1000
Fault location within the sector	0.05, 0.5, 0.95	0.25, 0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Noise in measurements	–	0.016% for V, 1.2% for I	2%

[a] 1000 Ω only for one case.

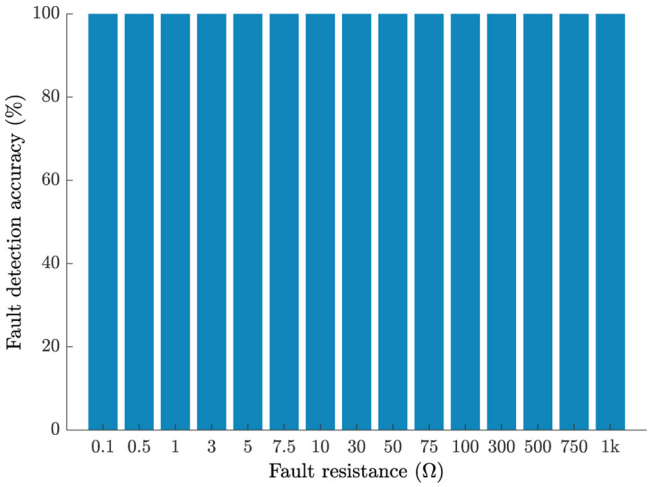

Fig. 6. Fault detection accuracy for different fault resistance values in the test dataset.

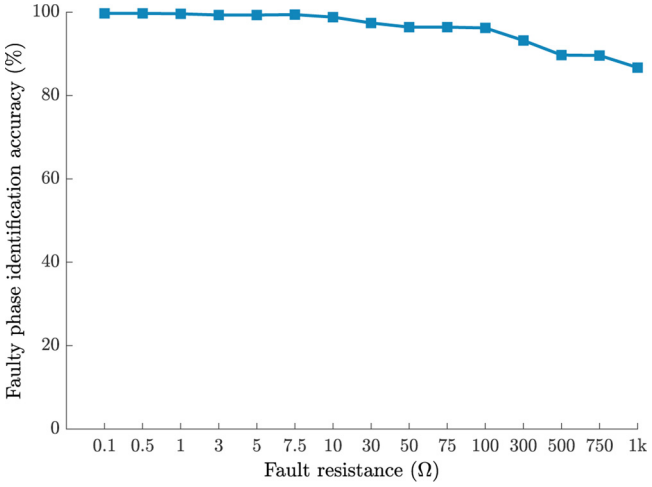

Fig. 7. Faulty phase identification accuracy for different fault resistance values in the test dataset.

Table 4
Faulty phase identification accuracy for each fault type in the test dataset.

Fault type	Phase identification accuracy (%)
AG	95.6
BG	96.0
CG	96.1
ABC	96.5

three-phase faults (ABC) with a maximum deviation of 0.9% between single-phase-to-ground and three-phase faults. This is an indication that besides the different combinations of generation and load penetration in the grid, the unbalanced nature of the grid, i.e. the topological and per phase asymmetry in the distribution of PVs and loads in the grid, did not affect the proposed method.

4.3. Faulty branch identification

The last functionality of this proposed method is the identification of the faulty branch. The results obtained with the proposed algorithm in the test dataset are presented in Fig. 8 and Table 5 for different fault resistance values and fault types respectively. It is shown in Fig. 8, that with an increase of the fault resistance the accuracy of the method

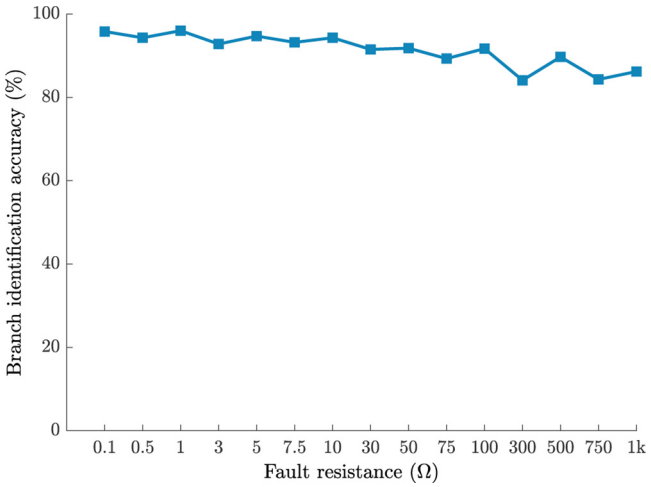

Fig. 8. Faulty branch identification accuracy in the test dataset for different fault resistance values.

Table 5
Faulty branch identification accuracy in the same test dataset for each fault type.

Fault type	Phase identification accuracy (%)
AG	93.1
BG	89.2
CG	91.2
ABC	91.7

decreases. More specifically, as can be seen from Fig. 8, a maximum accuracy of 95.8% is obtained for 0.1 Ω and a minimum of 84.1% for 300 Ω.

As it was mentioned in the previous subsection, the decrease of the accuracy of the method with an increase of the fault resistance is expected as the circulating fault current in the faulty branch will be less significant and thus the voltage drop smaller.

Furthermore, in Table 5, the accuracy of the faulty branch identification process is presented for each fault type. In all four of the presented cases, the proposed method is not affected by the fault type as the differences are really small. This renders the method immune to the unbalanced nature of the grid, i.e. the per phase unbalanced distribution of loads and PVs.

The above result is a key feature of this method. It was expected that since the loads and PV units are connected to the grid via single-phase connections, that a big difference in the accuracy of the method with regards to the fault type would be observable. However, this is not the case as it is demonstrated in Table 5.

4.4. Comparison with literature methods

As mentioned in the previous section, three similar studies were used to compare the algorithm results. Two methods designed for the MV grid and one conventional one for the LV grid. It should be noted that there are two factors that render the comparison with the MV cases not exactly fair: (a) the fact that the LV grid presents a more complex structure, highly heterogeneous and unbalanced, and (b) the fact that there are differences in the case studies between the available in the literature methods, e.g. studied grid, available measurements, noise in the measurements, fault location scenarios, studied fault resistances, fault types etc. These differences between the case study of this paper and the ones from the literature are presented in Table 3.

The results of this comparative analysis are gathered in Table 6. The results presented in Table 6 concern the fault types considered in this study: the single-phase-to-ground faults (most frequent) and three-

Table 6
Comparison of faulty branch identification accuracy of different MV methods to the performance of the proposed method for single-phase-to-ground and three-phase faults and specific fault resistance values/ranges.

Paper	Fault Resistance (Ω)	
	1–50	100
Brahma [34]	92.1%	–
Pignati et al. [21]	–	83.78–100%
Proposed method	93.8%	91.7%

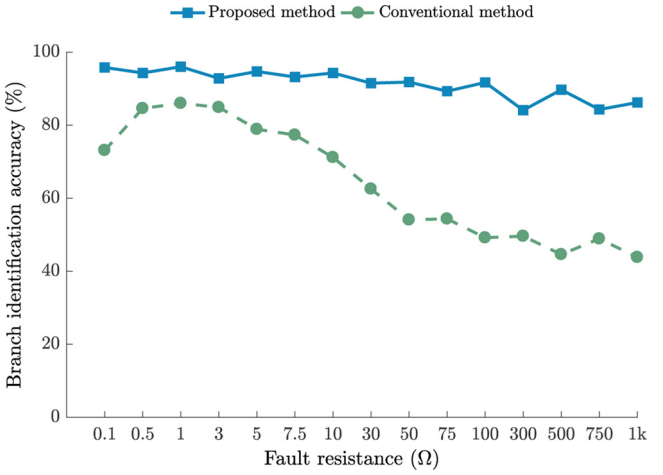

Fig. 9. Comparison of the branch identification accuracy of the proposed method with a conventional one for LV distribution grids [35].

phase faults (most severe). In order to compare the results for similar values of fault resistances, two cases were identified: (a) a range from 1 to 50 Ω which was used in [34] and b) the case of 100 Ω from [21]. For the first case, the proposed method of this paper outperforms the one from the literature by an average of 1.7%. For the second case, the proposed method outperforms the one of the literature in some cases. In general, taking into consideration the enlarged number of scenarios considered in this case study, i.e. the increased number of considered fault locations and the noise in the measurements, and the fact that the LV grid is a more complex case, the results are considered excellent.

To further test the performance of the proposed method, a conventional method for the LV grid case [35] was tested on the same dataset/case study of this paper. The results are presented in Fig. 9. The superiority of the proposed method is evident. An important remark is that although the accuracy of the conventional method decreases severely with the increase of the fault resistance, down to a minimum of 43.8% for 1000 Ω, the proposed method maintains high levels of accuracy as it was mentioned before.

4.5. Generalization to different fault cases

As a first step to test the robustness and generalization capabilities of the method, the algorithm was exposed to different fault cases, i.e. out-of-sample fault resistances. The values of these fault resistances were 4, 40 and 400 Ω. Then, to analyze the robustness of the method, its performance when exposed to out-of-sample fault resistances was compared with the one of the regular test dataset. In particular, the average accuracy in the test dataset between 3 Ω and 500 Ω was compared with the average accuracy of the three out-of-sample fault resistances. The results of this comparison are presented in Fig. 10.

It is shown that for the fault detection, the accuracy is identical in both cases (100%). For the fault type identification task, the difference in the accuracy is negligible (0.25% less accurate in the out-of-sample

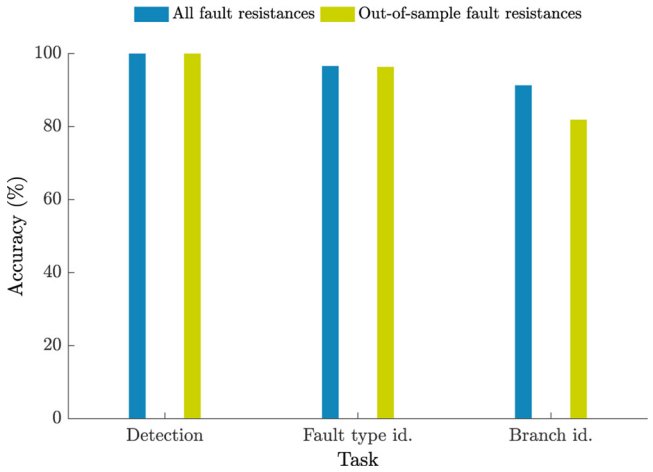

Fig. 10. Comparison of the average accuracy of the three tasks between all studied fault resistances between 3 and 500 Ω and the out-of-sample fault resistances.

case). For the faulty branch identification accuracy, while the accuracy drops from 91.3% to 81.9%, this level of accuracy is still considered excellent. Therefore, based on these results, it can be concluded that the method can generalize to unknown fault cases as its performance in out-of-sample fault resistances is either identical, negligible or very similar to the one of in-sample fault resistances.

4.6. Generalization to different grid topologies

As explained in the introduction, a key property of the proposed method is that it is generalizable and independent of the grid topology. Particularly, the method can be trained in a specific grid topology and then be used in a different one.

In this section, to study this specific property, the accuracy of the method is analyzed when it is trained in a specific grid topology and then employed in a different one. For that, the method is trained using feeders one and three of the considered grid, and then evaluated using data from the second feeder. Feeder one and three are selected as a training basis since they contain the maximum and minimum amount of branches (see Fig. 4). The results of all three method tasks when exposed to out-of-sample branches are depicted in Fig. 11.

For the fault detection task, a reduction of the accuracy from 100% to 99.15% can be noticed. Since the maximum error is 0.85 % for both single-phase-to-ground and three-phase faults, it is safe to assume that

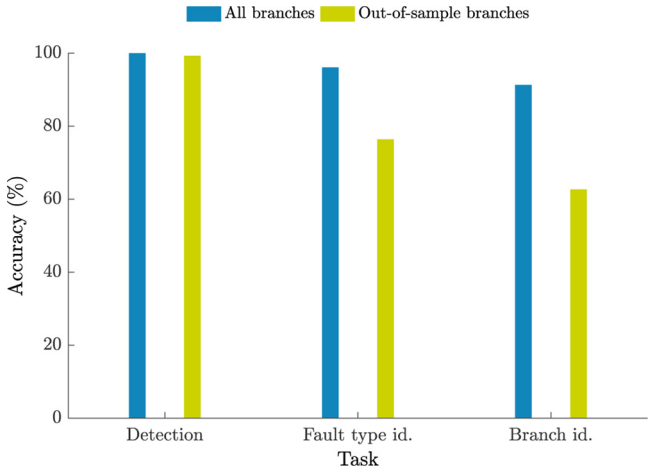

Fig. 11. Comparison of the average accuracy of the three tasks between all branches being part of the training and out-of-sample branches.

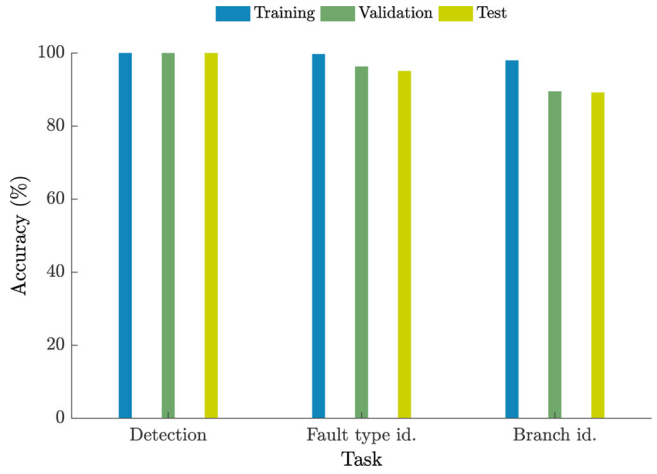

Fig. 12. Average accuracy for the three fault identification tasks for the training, validation and test datasets.

the algorithm will also detect out-of-sample double-phase-to-ground and phase-to-phase faults. Therefore, it can be stated that the fault detection results even under unpredictable circumstances are considered excellent. For the following two tasks however, the method is not as successful as before.

For the fault identification task, the accuracy decreases from 96.1% to 76.4%. At the same time, for the faulty branch identification task, the average accuracy drops from 90.15% to 62.76%; however, although not depicted in Fig. 11, the decrease of the accuracy is bigger for high fault resistance faults. Particularly, for faults up to 10 Ω an accuracy higher than 70% is achieved for the faulty branch identification task. In this case, to further improve the attained results, a retraining of the method is advised. The same is also advised in the case of a microgrid that needs to operate in isolated mode. The rapid training time of the algorithm facilitates that process and makes it ideal for real-time applications.

4.7. Overfitting

A standard issue with computational intelligence methods is overfitting [36]. Particularly, unless regularization techniques are used, computational intelligence methods can easily overfit the training dataset and perform poorly in out-of-sample data.

In the proposed method, to prevent that, the GBT model is evaluated during training using a validation dataset so that the hyperparameters and model structure do not become too complex. In this section, to show that the proposed method does not overfit, the performance of the method is compared with the training, validation and test datasets.

Fig. 12 displays this comparison for the three tasks. As can be clearly seen, the method does not overfit for fault detection, i.e. the accuracy is exactly the same across the three datasets. Similarly, the method does not overfit either for fault type identification: while a minor decrease in accuracy can be observed between the training and validation test, and validation and test dataset, this behavior is expected. Particularly, while the accuracy of three methods should ideally be the same, this is not possible since the number of datapoints in each dataset is finite. In practice, since the data distribution in each dataset is slightly different, the method always performs slightly better in the datasets used during training, and minor differences between the datasets are expected. A similar reasoning can be applied to the branch identification task: while the accuracy in the training dataset is slightly better, the method does not overfit.

5. Conclusions

In this paper, a gradient boosting tree model was proposed to detect, identify and locate faults in *low voltage (LV)* smart grids. To estimate the model, a set of non-branch-specific input features was employed to ensure the robustness of the algorithm against different grid topologies and available number of voltage measurements per branch. The proposed method was evaluated in a case study of a real case semi-rural LV distribution grid of Portugal. In detail, the case study comprised: (a) fault resistances between 0.1 and 1000 Ω, (b) different fault locations inside each sector (c) single-phase and three-phase faults, and (d) a 2% of underestimation error in the phase rms current and voltage measurements.

To test the accuracy of the proposed algorithm, the method was tested in an out-of-sample dataset. In addition, to analyze the robustness and generalization capabilities of the algorithm, the method was also tested against out-of-sample fault resistances and branches (resistance values and grid branches not included in the training dataset).

An excellent accuracy for fault detection and fault type identification was achieved. Faulty branch identification showed very promising results. In comparison with other studies in the literature, the algorithm accuracy for identifying a faulty branch, was found to be superior to a conventional method for the LV grid but also better than two methods from the medium voltage case. A great feature of the algorithm was that, as can be seen in the symmetrical performance in all the phases, the asymmetrical distribution of loads and photovoltaics across the phases and branches does not really affect the algorithm performance. In addition, as it could be expected, the increase of the fault resistance decreased the accuracy of fault type and branch identification tasks.

In detail, the algorithm obtained an accuracy of 100% when identifying the faulty feeder, an accuracy between 99.7% and 86.7% (the higher the fault resistance the lower the accuracy) when identifying the fault type, and an accuracy between 95.8–86.2% when identifying the faulty branch. These results show a clear superiority of the proposed method with regards to the other methods of the literature.

In future work, the omitted fault types will be included: double-phase-to-ground faults and phase-to-phase faults as well as the extension of the method to an exact fault location estimation. In addition, the algorithm will be applied in different grids and/or experimental setups.

Conflict of interest

None declared.

CRediT authorship contribution statement

Nikolaos Sapountzoglou: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Jesus Lago:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Bertrand Raison:** Supervision.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 675318 (INCITE). Thanks are also due to Mr. Konstantinos Kotsalos, ESR of INCITE-ITN and the company Efacec in Maia, Portugal, beneficiary partner of the INCITE-ITN, for providing us with the available data of a real semi-rural LV distribution grid of Portugal.

References

[1] T. Gönen, Electric Power Distribution Engineering, 3rd ed., Taylor & Francis, Boca Raton, 2014.
[2] H. Bjørnebye, Investing in EU Energy Security: Exploring the Regulatory Approach to Tomorrow's Electricity Production, Kluwer Law International B.V., 2010.

[3] A. Shivakumar, M. Welsch, C. Taliotis, D. Jakšić, T. Baričević, M. Howells, S. Gupta, H. Rogner, Valuing blackouts and lost leisure: Estimating electricity interruption costs for households across the European Union, Energy Res. Soc. Sci. 34 (2017) 39–48, https://doi.org/10.1016/j.erss.2017.05.010.
[4] A. Bahmanyar, S. Jamali, A. Estebsari, E. Bompard, A comparison framework for distribution system outage and fault location methods, Electr. Power Syst. Res. 145 (2017) 19–34, https://doi.org/10.1016/j.epsr.2016.12.018.
[5] A. Zidan, M. Khairalla, A.M. Abdrabou, T. Khalifa, K. Shaban, A. Abdrabou, R.E. Shatshat, A.M. Gaouda, Fault detection, isolation, and service restoration in distribution systems: state-of-the-art and future trends, IEEE Trans. Smart Grid 8 (5) (2017) 2170–2185, https://doi.org/10.1109/TSG.2016.2517620.
[6] IEEE, Guide for Determining Fault Location on AC Transmission and Distribution Lines, IEEE Std C37.114-2014 (Revision of IEEE Std C37.114-2004), (2015), pp. 1–76, https://doi.org/10.1109/IEEESTD.2015.7024095.
[7] J. Mora-Flòrez, J. Meléndez, G. Carrillo-Caicedo, Comparison of impedance based fault location methods for power distribution systems, Electr. Power Syst. Res. 78 (4) (2008) 657–666, https://doi.org/10.1016/j.epsr.2007.05.010.
[8] M. Majidi, M. Etezadi-Amoli, M.S. Fadali, A novel method for single and simultaneous fault location in distribution networks, IEEE Trans. Power Syst. 30 (6) (2015) 3368–3376, https://doi.org/10.1109/TPWRS.2014.2375816.
[9] A. Teninge, C. Pajot, B. Raison, D. Picault, Voltage profile analysis for fault distance estimation in distribution network, 2015 IEEE Eindhoven PowerTech (2015) 1–5, https://doi.org/10.1109/PT.C.2015.7232746.
[10] S. Jamali, A. Bahmanyar, A new fault location method for distribution networks using sparse measurements, Int. J. Electr. Power Energy Syst. 81 (2016) 459–468, https://doi.org/10.1016/j.ijepes.2016.02.046.
[11] C. Grajales-Espinal, J. Mora-Flórez, S. Pérez-Londoño, Advanced fault location strategy for modern power distribution systems based on phase and sequence components and the minimum fault reactance concept, Electr. Power Syst. Res. 140 (2016) 933–941, https://doi.org/10.1016/j.epsr.2016.04.008.
[12] D. Thukaram, H.P. Khincha, H.P. Vijaynarasimha, Artificial neural network and support vector machine approach for locating faults in radial distribution systems, IEEE Trans. Power Deliv. 20 (2) (2005) 710–721, https://doi.org/10.1109/TPWRD.2005.844307.
[13] H.A. Darwish, A precise fault locator algorithm with a novel realization for MV distribution feeders, 2006 IEEE Power Engineering Society General Meeting (2006) 8, https://doi.org/10.1109/PES.2006.1709357.
[14] H. Zayandehroodi, A. Mohamed, H. Shareef, M. Farhoodnea, A novel neural network and backtracking based protection coordination scheme for distribution system with distributed generation, Int. J. Electr. Power Energy Syst. 43 (1) (2012) 868–879, https://doi.org/10.1016/j.ijepes.2012.06.061.
[15] P.E. Farias, A.P. de Morais, J.P. Rossini, G. Cardoso, Non-linear high impedance fault distance estimation in power distribution systems: a continually online-trained neural network approach, Electr. Power Syst. Res. 157 (2018) 20–28, https://doi.org/10.1016/j.epsr.2017.11.018.
[16] J. Mora-Florez, V. Barrera-Nunez, G. Carrillo-Caicedo, Fault location in power distribution systems using a learning algorithm for multivariable data analysis, IEEE Trans. Power Deliv. 22 (3) (2007) 1715–1721, https://doi.org/10.1109/TPWRD.2006.883021.
[17] R.H. Salim, K.R.C. de Oliveira, A.D. Filomena, M. Resener, A.S. Bretas, Hybrid fault diagnosis scheme implementation for power distribution systems automation, IEEE Trans. Power Deliv. 23 (4) (2008) 1846–1856, https://doi.org/10.1109/TPWRD.2008.917919.
[18] S.S. Gururajapathy, H. Mokhlis, H.A. Illias, Fault location and detection techniques in power distribution systems with distributed generation: a review, Renew. Sustain. Energy Rev. 74 (2017) 949–958, https://doi.org/10.1016/j.rser.2017.03.021.
[19] L. Andrade, T. Ponce de Leao, travelling wave based fault location analysis for transmission lines, EPJ Web of Conferences, vol. 33 (2012) 04005, https://doi.org/10.1051/epjconf/20123304005.
[20] A.N. Milioudis, G.T. Andreou, D.P. Labridis, Enhanced protection scheme for smart grids using power line communications techniques–Part II: location of high impedance fault position, IEEE Trans. Smart Grid 3 (4) (2012) 1631–1640, https://doi.org/10.1109/TSG.2012.2208988.
[21] M. Pignati, L. Zanni, P. Romano, R. Cherkaoui, M. Paolone, Fault detection and faulted line identification in active distribution networks using synchrophasors-based real-time state estimation, IEEE Trans. Power Deliv. 32 (1) (2017) 381–392, https://doi.org/10.1109/TPWRD.2016.2545923.
[22] S.H. Mortazavi, Z. Moravej, S.M. Shahrtash, A searching based method for locating high impedance arcing fault in distribution networks, IEEE Trans. Power Deliv. 34 (2) (2019) 438–447, https://doi.org/10.1109/TPWRD.2018.2874879.
[23] M.M. Alamuti, H. Nouri, N. Makhol, M. Montakhab, Developed single end low voltage fault location using distributed parameter approach, 2009 44th International Universities Power Engineering Conference (UPEC) (2009) 1–5.
[24] M.M. Alamuti, H. Nouri, R.M. Ciric, V. Terzija, Intermittent fault location in distribution feeders, IEEE Trans. Power Deliv. 27 (1) (2012) 96–103, https://doi.org/10.1109/TPWRD.2011.2172695.
[25] G.A. Orcajo, J.M. Cano, M.G. Melero, M.F. Cabanas, C.H. Rojas, J.F. Pedrayes, J.G. Norniella, Diagnosis of electrical distribution network short circuits based on voltage Park's vector, IEEE Trans. Power Deliv. 27 (4) (2012) 1964–1972, https://doi.org/10.1109/TPWRD.2012.2210448.
[26] N. Silva, F. Basadre, P. Rodrigues, M.S. Nunes, A. Grilo, A. Casaca, F. Melo, L. Gaspar, Fault detection and location in Low Voltage grids based on distributed monitoring, 2016 IEEE International Energy Conference (ENERGYCON) (2016) 1–6, https://doi.org/10.1109/ENERGYCON.2016.7514000.
[27] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'16, ACM Press, San Francisco, California, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[28] T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York, New York, NY, 2001, https://doi.org/10.1007/978-0-387-21606-5.
[29] J. Lago, F. De Ridder, B. De Schutter, Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms, Appl. Energy 221 (2018) 386–405, https://doi.org/10.1016/j.apenergy.2018.02.069.
[30] J. Lago, K. De Brabandere, F. De Ridder, B. De Schutter, Short-term forecasting of solar irradiance without local telemetry: a generalized model using satellite data, Solar Energy 173 (2018) 566–577.
[31] J.S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: J. Shawe-Taylor, R.S. Zemel, L. Bartlett, F. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24, Curran Associates, Inc., 2011, pp. 2546–2554.
[32] L. Marques, N. Silva, I. Miranda, E. Rodriges, H. Leite, Detection and localisation of non-technical losses in low voltage distribution networks, Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower 2016) (2016) 1–8, https://doi.org/10.1049/cp.2016.1079.
[33] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D.D. Cox, Hyperopt: a python library for model selection and hyperparameter optimization, Comput. Sci. Discov. 8 (1) (2015) 014008.
[34] S.M. Brahma, Fault location in power distribution system with penetration of distributed generation, IEEE Trans. Power Deliv. 26 (3) (2011) 1545–1553, https://doi.org/10.1109/TPWRD.2011.2106146.
[35] N. Sapountzoglou, B. Raison, N. Silva, Fault Detection and Localization in LV Smart Grids, 2019 IEEE Milan PowerTech (2019), https://doi.org/10.1109/PTC.2019.8810799.
[36] J. Lago, F. De Ridder, P. Vrancx, B. De Schutter, Forecasting day-ahead electricity prices in Europe: the importance of considering market integration, Appl. Energy 211 (2018) 890–903, https://doi.org/10.1016/j.apenergy.2017.11.098.