Electrical Power and Energy Systems 142 (2022) 108303

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Data analysis and management for optimal application of an advanced ML-based fault location algorithm for low voltage grids

P. Stefanidou-Voziki [a,*], D. Cardoner-Valbuena [b], R. Villafafila-Robles [c], J.L. Dominguez-Garcia [a]

[a] Institut de Recerca en Energia de Catalunya (IREC), Jardins de les Dones de Negre 1, 08930 Sant Adrià de Besòs (Barcelona), Spain
[b] Ciklum, Severo Ochoa 5, 29590, Malaga, Spain
[c] Centre d'Innovació Tecnològica en Convertidors Estátics i Accionaments (CITCEA-UPC), Departament d'Enginyeria Eléctrica, Universitat Politécnica de Catalunya, ETS d'Enginyeria Industrial de Barcelona, Av. Diagonal 647, Pl. 2, 08028, Barcelona, Spain

ARTICLE INFO

Keywords:
Data analysis
Dimensionality reduction
Fault location
Low voltage grid
Machine learning

ABSTRACT

As the need for automatization of the electricity grid's fault diagnosis schemes is rising, the application of technologies such as the artificial intelligence (AI) can provide practical solutions to the problem. AI can overcome the challenges that complex topologies like those of the low voltage (LV) smart grids pose and prove to be a powerful tool in the development of advanced fault diagnosis methods. An important parameter for the success of any AI-based method is the quality of data. Therefore, in this paper a data analysis is performed in order to evaluate the type of data produced by a small LV grid and an representative AI algorithm's response to those. In the context of this analysis, the most important features and meters were identified. Furthermore, as a response to the large volume of available data, a data management strategy is proposed. The strategy combines original and reshaped features. For this purpose, five dimensionality reduction methods are tested and compared. Truncated-SVD is deemed the most appropriate and is subsequently utilized for the reshaping of the dataset that is introduced to the XGBoost fault location model. The integration of the dimensionality reduction technique in the algorithm results in the decrease of the computational time and the dataset's size and in a higher generalizability of the algorithm. Thus, the application of the proposed method is not limited by the grid's topology. The method's robustness was verified against various influencing parameters such as the fault resistance, the size of the dataset, the loss of data and the photovoltaics' penetration level. The overall algorithm achieved a mean squared error of 13.26 and a training and test accuracy of more than 99% when tested on the CIGRE LV benchmark grid.

1. Introduction

The transformation of the traditional electricity grids into smart grids is well underway, mandating the redefinition of the grid operation principles. One of the vital operating parts of the grid requiring redesign is the protection system and more specifically the fault diagnosis schemes, since the bidirectionality of power flows and the intermittency of generation sources pose additional challenges. Fault diagnosis refers to the detection, classification and location of a fault. Rapid and automatized fault diagnosis leads to increased reliability of the electricity grid, aligned with the needs of the modern society. With the vast changes in the grid's topology, the traditional fault diagnosis methods have become outdated and inefficient. Therefore, the necessity for novel accurate and fast fault diagnosis methods has soared.

This study focuses on the fault location part of the process, as it is considered the most challenging and can assist significantly the minimization of power outage times. An accurate location of a fault can result in automated fault isolation or early repair from the Operations and Maintenance (O&M) technicians on the field and fast restoration of power supply is an important factor in the evaluation of transmission system operators' (TSOs) and distribution system operators' (DSOs) quality of services. One of the main points, though, that differentiates this research from other similar ones is the analyzed part of the grid. Each part of the electricity grid presents distinct characteristics, mainly related to the grid topology, the loads, the line parameters and the type of energy generation. Hence, different fault location approaches are required, adjusted to the particularities of each voltage level. The majority of the related studies have developed methods suited for high voltage (HV) [1–3] and medium voltage (MV) grids [4–16]. Nevertheless, due to the smart grid-transition, the complexity as well as the importance of the low voltage (LV) grid have risen, thus the focus of this research is centered in on that part of the grid.

[*] Corresponding author.
E-mail address: pstefanidou@irec.cat (P. Stefanidou-Voziki).

https://doi.org/10.1016/j.ijepes.2022.108303
Received 24 January 2022; Received in revised form 19 April 2022; Accepted 27 April 2022
Available online 12 May 2022

Traditionally, the LV grid is characterized by its tree-shaped form, the mix of types of conductors, the great variance in the grids' lengths, the residential loads and the multi-phase operation. In addition to these, the latest advancements in the energy sector have led to the gradual transformation of consumers to prosumers, in line with the growing integration of electric vehicles, batteries and renewable energy sources (RES), and to the improvement of the grid's observation and control devices and strategies. Therefore, the extensive available research on the HV and MV grids either cannot be applied to the LV grid or needs to be tailored to its particular properties.

1.1. State-of-the-art

The part of the grid presenting the most common characteristics with the LV grid is the MV one. The main methodologies found in the fault location methods developed for MV grids can be categorized as: impedance-based [4–6], traveling-wave-based [7,8], sparse measurements [9–11], artificial intelligence (AI) [12–14] and hybrid [15–17]. The first two were also the first to be developed and are still very frequently applied as they clearly reflect the physical laws of electric circuits. The impedance-based methods consist of analytical equations based on the Kirchhoff and Ohm laws. Thus, they are considered to be simpler to implement, nevertheless they can result in multiple location estimation or inaccuracies in the presence of RES. The traveling-wave-based methods locate the faulted point by interpreting the waves' reflections. They usually lead to accurate and fast fault location, however, their implementation requires specialized equipment and personnel. Furthermore, in grids with many lateral branches such as the LV grid, the signal retrieval could prove challenging.

On the other hand, the broad installation of smart meters on the grid, combined with their increasing capabilities, led to the development of fault location methods relying on collected measurements. A representative example is the sparse measurements methods that are based on the comparison between the voltages recorded in multiple points throughout the grid and the voltages collected from the simulation of all possible faults. The case presenting the highest convergence between the two indicates the fault location. Sparse measurements are usually combined with an impedance-based method for the elimination of the multiple location estimation problem. Such method combinations lead to the formation of hybrid methods, that aim at maximizing the accuracy and efficiency of the faulted point's location. Because of their nature, however, hybrid methods tend to be complex and demanding schemes. The most popular hybrid methods so far are those combining sparse measurements with impedance-based methods as well as those utilizing wavelets as inputs to AI algorithms.

AI is another example of a method gaining popularity due to the growing data availability. The applications of AI algorithms in the fault diagnosis field have multiplied over the last years, providing innovative and highly accurate solutions. The main drawbacks of AI are considered to be the data and computational requirements. Nevertheless, the large volume of necessary training data can be minimized with the use of dimensionality reduction or feature selection techniques, while the technological progress gradually provides solutions to the computational requirements. Overall, in the majority of cases the AI's performance outweighs its drawbacks.

From the aforementioned methods, the ones finding applications in the LV grid so far are the signal-based, the measurement-based, the AI and some hybrid methods. In the signal-based methods, one of the first developed was the time domain reflectometry (TDR). Examples of its application can be found in [18] and the more elaborated method of [19]. TDR's results are considered reliable, however, it requires specialized equipment and personnel in order to be implemented. In the recent years, there have been further advancements on methods utilizing signals for the location of faults. Some of them are based on the introduction of a test signal to the grid [20], while others on the analysis of vectors such as the Park's vector [21].

As commented earlier, the installation of smart meters and sensors as well as the continuous investments in the grid's O&M, have created new opportunities for the development of novel methods taking advantage of the collected data. Therefore, also in the case of LV grids, an increase was observed in fault location methods relying on smart devices. E.g. in [22], fault indicating devices are used to locate the faulted section. In [23], the negative-sequence voltage variations are recorded, however their utilization entails the establishment of thresholds for each particular grid. Then, the proposal of [24] to compare the current magnitude and angle between the sides of each section in order to locate the faulted section is simple but implies big infrastructure investments. The common vulnerability of the aforementioned methods lies in their performance's dependence on various grid parameters that were, additionally, not taken into consideration during the research. A technique with higher robustness is presented in [25,26], were emphasis is being given to the local sensors and the grid's communication channels for an accurate and fast location of the faulted branch. However, as in the rest of the aforementioned papers, with the exception of [23], only the faulted section is located and not the exact fault point.

Following the pattern observed in the MV grid fault location techniques, there was also a development of hybrid methods utilizing the versatile data that can be obtained by smart devices as inputs in impedance-based methods in order to increase their accuracy. In [27] novel devices operating in very high frequencies are employed for the collection of data to be used in the location of arc faults. Then, in [28] the obtained voltage measurements are compared with thresholds that once again require adjustment based on the specific application. In an effort to counterbalance this negative characteristic and increase the flexibility of the method, the grid's zone division was proposed.

Finally, AI has been applied lately as a solution to the fault location problem in LV grids as well. In [29] a statistical approach of machine learning (ML) is presented, that can locate the faulted point of the line and not just the node closest to the fault. Even though this approach tries to avoid the black-box side of AI, it lacks in accuracy and only refers to three phase faults with a resistance up to 1 Ω. On the other hand, in [30,31] two more widely known AI tools are utilized, that both lead to high accuracy. More specifically, a gradient boosting tree (GBT) model and a deep learning (DL) algorithm are implemented respectively. The GBT requires less training data, has less computational requirements, it is faster and its performance is not as depended on the hyperparameter tuning as that of the DL. However the GBT was not tested in the location of the exact point, only in the identification of the faulted section, so it was used for classification and not regression purposes. Apart from that, the two proposed methods follow a similar logic, starting with the fault detection and continuing with the identification of the faulted branch. In that last part, the DL algorithm led to better results than the GBT. Furthermore, the proposed DL algorithm is also independent of the number of meters on the grid. Both methods, showcase the robustness of AI methods, with the thorough sensitivity analysis they include, however they exclusively refer to single phase and three phase faults and have high data demands.

Based on the literature review, it is obvious that despite the abundance of fault location methods, only a small portion of them is addressing the particularities of the LV grid. Additionally, AI emerges as a powerful tool also in the field of fault location, combining high accuracy with practicality. The available AI data management and analysis techniques counterbalance the most important limitations associated with its application while enabling the deeper understanding of complex topologies such as those of LV grids. Moreover, many new technological tools facilitate the practical implementation of AI algorithms rendering its use even more appealing for both research and commercial purposes.

1.2. Contribution

The subject of the current study is the optimization of an AI algorithm's application for the location of shunt faults in LV grids. The method includes an all-important analysis of the data collected from the simulation of a small LV grid, in this case of a modified version of the CIGRE European LV benchmark, and a study on the best data form to be used as an input. Data analysis is crucial for the selection of the most suitable methodology – and, in the case of AI, the selection of the fitting prediction model – as well as for the correct interpretation of the results.

Furthermore, a data management strategy is presented in order to optimize the data collected by the measurements devices. As the measuring devices and available measurements are expected to multiply in the future, their efficient management is imperative for the computational systems. The proposed strategy combines the utilization of the most important measurements, as indicated by the data analysis, with the dimentionality reduction of the less useful part of the data, thus decreasing the computational time (CT) and complexity of the algorithm. For this purpose, five dimensionality reduction algorithms are compared in order to select the appropriate one for the application. By reducing the size of the dataset, the algorithm becomes independent of the number and the location of the measuring devices, hence, generalizable. Thus, its overfitting to the specific grid data is decreased and it can be applied to every grid regardless of its topology.

Finally, an eXtreme Gradient Boosting (XGBoost) ML regression model is applied for the location of the faulted point. Tree-based algorithms have proven to be highly efficient for and easily applicable to problems with tabular data [32], such as the fault diagnosis. Hence, the performance of a state-of-the-art tree model, i.e. the XGBoost prediction model, is tested in the location of faults in active LV grids.

Counter to [30], in this paper the XGBoost model is used as a regressor, for the location of the exact faulted point and not only the faulted node. Moreover, this is the first ML-based method that is able to locate all types of shunt faults in LV grids and the first method that includes a data analysis and management strategy.

Overall, the contribution of the presented research can be summarized in four main points:

A. Evaluation of the quality of data collected by a small scale LV grid and analysis of the optimum dataset composition.
B. Development of an efficient data management strategy, independent of the grid's topology.
C. Minimization of the features' dimensionality and, as a result, of the CT, with a simultaneous increase of the method's generalizability.
D. Application of an advanced ML model for the accurate location of the faulted point in an active LV grid.

2. Fault location method

The presented fault location method refers to all types of shunt faults and can be applied to all active LV distribution grids, regardless of their topology. It is a ML-based method, thus, a prediction model is trained for the location of the exact faulted point. In order to increase the method's accuracy emphasis is given to the data and their management. Therefore, first the data are analyzed, based on that they are appropriately processed and finally they are combined in a new improved dataset. Following that, the model's parameters are carefully evaluated and tuned for optimum results.

Fig. 1 summarizes the outline of the proposed fault location method, which will be thoroughly described in this section. The shaded boxes represent the parts of the algorithm that are implemented only during the development phase and are omitted during its application. Moreover, the red block corresponds to an important part of the algorithm, the identification of the faulted branch, that is not being analyzed here since an accurate model that performs this task was presented in [33]. The faulted branch classification is an indispensable part of the method as it tackles the multiple location estimation problem, which occurs when the calculated distance between the fault and the feeder points to locations in more than one branches.

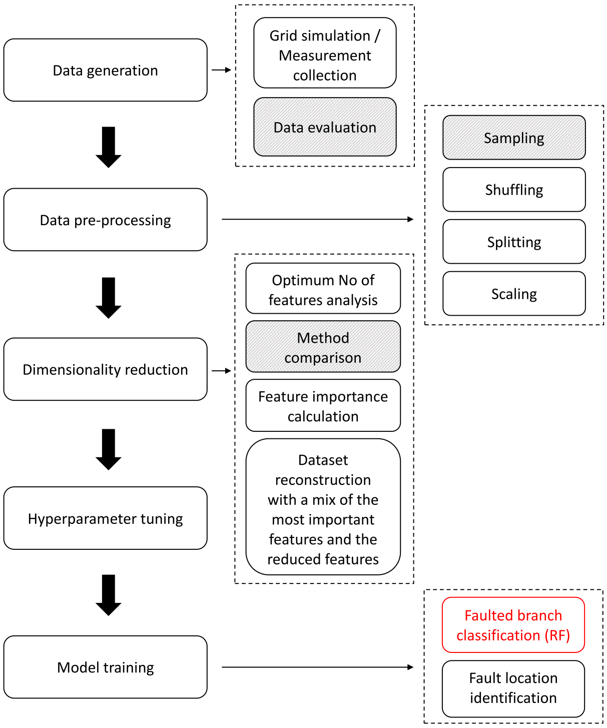

Data generation
Grid simulation / Measurement collection
Data evaluation
Data pre-processing
Sampling
Shuffling
Splitting
Scaling
Dimensionality reduction
Optimum No of features analysis
Method comparison
Feature importance calculation
Dataset reconstruction with a mix of the most important features and the reduced features
Hyperparameter tuning
Model training
Faulted branch classification (RF)
Fault location identification

Fig. 1. Structural diagram of the proposed fault location method.

2.1. Data analysis tools

The performance of a ML method depends heavily on the proper interpretation of the available data. Nevertheless, the importance of data evaluation is not limited to the ML applications. It is imperative for simulation data to be evaluated irregardless of the employed fault location methodology; the reliability of a method's test results depends on the reliability of the test data. The majority of the existing fault diagnosis methods rely on data generated by simulations, as there is a general lack of real data, especially in relation to fault events in the grid. Even though with the evolution of O&M solutions for electricity grids the DSOs' observability of the grid has significantly increased, there is still a long way to go until real data become vastly available to the research community. Hence, the evaluation of data generated from benchmarks like the one used in this study are of great value.

Therefore, the first step in any research should be the data analysis. In the case of ML methods, the data analysis should precede the model selection as the latter could be affected by the data distribution. Even though there are models such as the tree-based ones that are not affected by the data distribution this can still reveal patterns that could lead to the development of a custom model or to an appropriate data transformation. Furthermore, the data analysis can point to the features that are the most informative for the prediction of the target as well as, in the case of electricity grids, the position of the most important meters. Overall, the data analysis facilitates the researchers' methodology-related decisions and the interpretation of the algorithms' outcome, and it leads to the formation of an efficient dataset and the development of an accurate model.

The associated parameters studied in this research are the data normality, the correlation between the features and the target value, and the feature importance. The data normality was tested with the use of the Shapiro–Wilk normality test [34]. This test originally assumes that the provided data are normally distributed, then compares them with an actual normal distribution and finally calculates the probability of the provided data being similar to those belonging to a normal distribution. If the probability is higher than 0.05 then the data is categorized as normal.

Then, in order to test the data correlation three different approaches were used. First, Pearson's correlation coefficient r [35] was used for the study of the linearity between the features and the target value. The r coefficient is the fraction of the covariance of a random pair of samples (x_i, y_i) and their respective standard deviations:

$$r = \frac{cov(x_i, y_i)}{\sigma_{x_i}\sigma_{y_i}} \tag{1}$$

Pearson's correlation coefficient refers to Gaussian data. For non-Gaussian data two ranking coefficients are mainly used. The first one is Kendall's correlation coefficient τ [36], that is used for the comparison of concordant and discordant pairs in the data set. This is calculated as follows:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{2}$$

where (x_i, y_i) and (x_j, y_j) are two random pairs of observations and n is the total amount of observations. The other ranking coefficient is Spearman's ρ correlation coefficient [37] which is used for the calculation of the strength between two variables. More specifically, the ρ coefficient examines the existence of a monotonic relation between two variables, i.e. if one decreases/increases with the decrease/increase of the other, and is defined as follows:

$$\rho = \frac{cov(rank(x_i), rank(y_i))}{\sigma_{rank(x_i)}\sigma_{rank(y_i)}} \tag{3}$$

It can be observed that the definition of the ρ coefficient is similar to that of the r; their difference lies in the use of the variables' values rank in the calculation of ρ, instead of the actual sample values. Hence, the Spearman correlation is less affected by outliers compared to the Pearson correlation.

The data analysis is completed with the examination of the features' importance. The features' importance is an indication of each feature's contribution to the prediction of the target value. It is not another approach for the evaluation the features' correlation with the target value but rather a measure of each feature's influence on the model's decision-making process. Therefore, it can serve as a starting point for the employment of feature selection or dimensionality reduction techniques. Moreover, it provides information regarding the response of the prediction model to the specific type of data, since the calculation of the feature importance depends on the estimator fitted to the data. In decision trees the feature importance indicates which values are mostly used for the splitting of trees into branches. The method applied here for the calculation of the features' importance is the computation of the mean and the standard deviation of the impurity decrease accumulation in each tree. It was selected as an appropriate method as it is less computationally expensive than other alternatives and its use is only discouraged in the case of datasets containing high cardinality features. This does not apply to the generated dataset, as it does not contain a large number of unique values.

2.2. Prediction model

ML models are divided into classification and regression. Classification models predict categories, while regression models predict continuous values. In the presented method both a classification and a regression model are utilized; first a classification model is employed for the prediction of the faulted branch, and then a regression model is applied for the prediction of the fault's exact distance from the main feeder. Thus, the multiple location estimation problem that follows many fault location methods is resolved and an accurate solution is proposed.

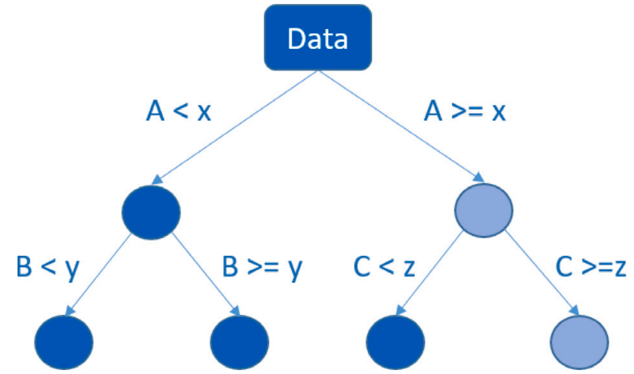


Fig. 2. The structure of a decision tree.

Both the classification and the regression models utilized are tree-based. The classification of the faulted branch with the use of a Random Forest (RF) has already been successfully validated in [33], therefore, only the regression model will be analyzed in this study. Tree-based prediction models offer high flexibility and efficiency, with reduced computational demands [38] and have been proven accurate in fault location problems for the MV [12]. Moreover, they have less pre-processing and data requirements than the employment of other popular AI techniques such as the NN, and perform well with tabular data, such as the ones used here. Apart from those basic advantages, nevertheless, each individual tree-based model has its own additional merits.

The tree-based models are named after their shape, which is similar to that of a tree; they start from a single node and by making decisions regarding an attribute of the dataset each node is split into two or more new sub-nodes. The objective of the decisions leading to a tree's splitting is the creation of homogeneous sub-nodes while the pruning that follows aims at the minimization of overfitting by removing terminal nodes that do not provide useful information related to the prediction result. The first node of the tree contains all or a bootstrapped sample of the available data, depending on the model. Based on indexes such as the entropy and the information gain scores the node split is performed. The process is repeated for each node and only for attributes that have not been previously selected by the algorithm. The selection of the attributes, the number of features considered in each split and the number of splits differ for the different tree-based models and their selection depends on the hyperparameter tuning which will be discussed later and on complex algorithms that fall beyond the scope of this research. The simplest tree model is the decision tree, illustrated in Fig. 2, based on which the rest of the tree-based models were developed.

The predictive capacity of the models is measured by the objective function, which consists of the training loss and the regularization term:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{4}$$

where θ is the vector of the weights added to each feature in order to predict the target value. In tree-based models the predicted target value can be expressed as follows:

$$\hat{y}_i = \sum_{k=i}^{K} f_k(x_i) \tag{5}$$

where $f_k \in F$ is a function containing the tree structure and leaf scores for all the possible trees, which form the functional space F. K is the number of trees employed.

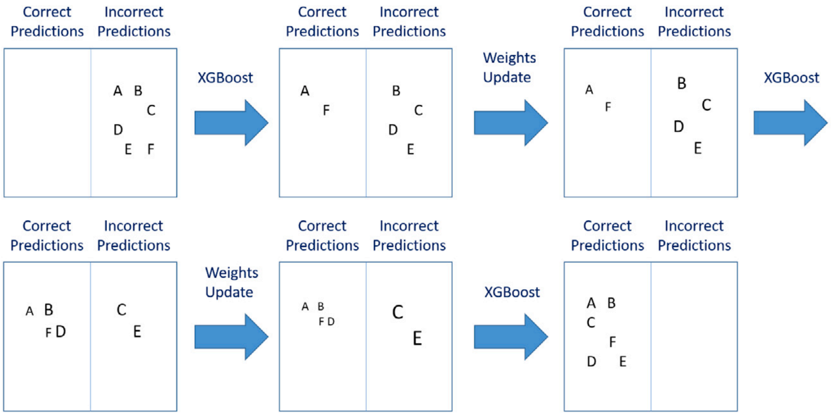


Fig. 3. Illustration of the XGboost model's sequential training and weight assignment.

Usually the training loss is calculated with the use of the mean squared error (MSE) and can be expressed as:

$$L(\theta) = \sum_{i} (y_i - \hat{y}_i)^2 \tag{6}$$

Thus, the form of the objective function for a tree-based model is the following:

$$obj(\theta) = \sum_{i}^{n} (y_i - f_k(x_i))^2 + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

The goal of the models is the optimization of the objective function, in this case of Eq. (7). Depending on the specific tree model the objective function can have a slightly different form.

2.2.1. Regression model

After reviewing the positive results presented in [30] regarding the use of gradient boosting in fault diagnosis in LV grids, as well as other applications of gradient boosting models, the tree-based model that was selected for the fault location's prediction is an advanced gradient boosting algorithm. In contrast with the aforementioned fault diagnosis method, however, in this case the XGBoost model [39] is not used for classification purposes but as a regressor. Thus, it is the first algorithm utilizing XGBoost for the location of faults. XGBoost combines the aforementioned advantages of the tree-based models with high computational speed. Like the other tree-based models, it also constitutes an ensemble model that is based on the combination of multiple weak learners[40]; here the weak learners are the decision trees. In this case though the weak learners are trained sequentially, with each next weak learner trying to improve the accuracy in a part of the data where the previous weak learner proved inaccurate. This is achieved by assigning a weight to each learner's results, as illustrated in Fig. 3, where each letter corresponds to a numerical value. The correctly predicted examples are given lower weights so that the weak learner that follows focuses on the falsely predicted examples which in turn are given higher weights.

So far XGBoost has not been applied to a fault location problem in an electricity grid. Nevertheless, it presents multiple benefits showcased in a variety of applications [41–43] and its multifarious characteristics could provide a robust solution to a complex problem such as the fault location. Among its noteworthy features that distinguish it from other tree-based and boosting models are:

1. The regularization; it is the most important characteristic of the XGBoost model as it decreases significantly overfitting.
2. The parallel processing; another important feature that reduces notably the execution time needed to built each tree and therefore to train the model.
3. The backwards pruning; commencing the pruning process after all splits have taken place optimizes the tree shaping.
4. The optimization possibilities; they exceed those offered by other algorithms, expanding the model's potential. Some examples of that include the built-in cross-validation and handling of missing values.

According to Eq. (5), the additive function of the boosting trees can be expressed by the prediction model equation as follows:

$$\begin{aligned} &\hat{y}_i^{(0)} = 0 \\ &\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ &\dots \\ &\hat{y}_i^{(n)} = \sum_{k=i}^{n} \hat{y}_i^{(n-1)} + f_n(x_i) \end{aligned} \tag{8}$$

where n is a step of the prediction process. The objective function of the model is formed in accordance with Eq. (7).

2.3. Data management strategy

The efficient management of collected data is crucial for any fault location method that relies heavily on data, especially for AI-based ones. As the data recording points in electricity grids are expected to increase in the future due to the broader installation of smart devices, their efficient management is of high importance; on one hand in order to relieve the pressure on the computational systems and on the other hand in order to extract all the available information.

The general norm in AI algorithms is that the larger the number of the available features the more accurate the prediction of the model, as long as these features are correlated with the target value. Nevertheless, a large number of features can also lead to overfitting and a failure of the algorithm to generalize to unseen datasets. This could deem any method unreliable and inaccurate. Moreover, the majority of datasets tend to contain a number of either uncorrelated features that add noise to the model's training or features that are too similar thus are redundant. This is common in datasets with many collected features.

The data management approach proposed here has two main goals: (a) the extraction of solely the useful features in order to optimize the model's performance and reduce the computational complexity and time and (b) the conservation of the physical meaning of the grid's most informative measurements. Hence, in the constructed dataset, the most informative measurements, as those were indicated by the features' importance analysis, are maintained in their original form while the rest of them are transformed according to the applied dimensionality reduction method. Thus, the quantity of information in the final dataset is enhanced while the generalizability of the algorithm remains high. The time added by the dimensionality reduction process is insignificant in comparison with the drop in the training time. This strategy is independent of the number and location of the meters, thus it is applicable to all grids regardless of their topology.

2.3.1. Dimensionality reduction

There are various dimensionality reduction techniques. Some are exclusively focused on the visualization of the dataset, thus they reduce the number of dimensions to a maximum of three, while others allow the selection of the dimensions' final number. In the present study the main goal of the dimensionality reduction was the optimum balancing of the minimum CT and the maximum accuracy. Hence, the methodologies applied here are the ones favoring a higher dimensionality space. Specifically, the techniques tested and compared are the Principal Component Analysis (PCA) [44], the Kernel PCA [45], the Fast Independent Component Analysis (FastICA) [46], the Truncated Singular Value Decomposition (T-SVD) [47] and the Isometric Feature Mapping (ISOMAP) [48].

1. PCA
   PCA is a well-known statistical method that aims at eliminating the correlated features of a dataset that carry little information and create a more compact one that is easier to analyze and manage during the training of ML model. Due to its efficiency it has been applied in numerous AI-based studies, including fault location methods [49]. The theory behind PCA lies in the projection of the data points onto a smaller feature space, whose every axis is perpendicular to the rest, thus uncorrelated with them. Each axis corresponds to one eigenvector, i.e. principal component. The eigenvectors are sorted based on their eigenvalues; in this case the higher the eigenvalue the more important the eigenvector. The first principal component is the vector representing the line with the minimum squared distance from all the data points, hence is characterized by the highest possible variance. Each new principal component aims at best fitting the data points and is orthogonal to the rest of the principal components. The total number of the principal components created by the PCA is the same as that of the original features, however, the biggest amount of information is compressed into the first components. Thus, the utilization of only the first few components can lead to the successful training of the ML model. Usually, the principal components selected are those with an explained variance adding up to 80% of the original dataset's.
2. Kernel PCA
   Kernel PCA is a variation of the traditional PCA that, contrary to the original method, is able to perform non-linear dimensionality reduction. Thus, before the application of the PCA's linear operations, the utilized kernels transform the dataset's dimensions into a space where it has a linear form. There are various available kernels; the selection of the most suitable one depends on the shape of the data. In this study the cosine kernel was used.
3. FastICA
   ICA methods aim at isolating the independent components of the dataset by finding the matrix that maximizes the non-gaussianity of the original features. The non-gaussianity metric is a means of measuring the statistical independence of the components. The difference between the traditional ICA methods and the FastICA lies in the calculation of the non-gaussianity. In the first case it is calculated with the use of the kurtosis whereas in the second case with the use of the negentropy. Hence, the FastICA is faster and more reliable.
4. Truncated SVD
   T-SVD is a dimensionality reduction method based on the factorization of the data matrix. Its operating principle is similar to that of the PCA with the exception that it does not center the data before performing the computations. While PCA transforms the covariance matrix, SVD transforms the data matrix. Thus, the computational complexity and time of the T-SVD are significantly lower. Furthermore, the truncating nature of the method is the one allowing the dimensionality reduction and differentiating it from the SVD method.
5. ISOMAP
   ISOMAP is another non-linear dimensionality reduction method. It can be considered an extension of Kernel PCA. The main goal of ISOMAP is the projection of the data into a lower-dimensional space where the geodesic distances between the data points are maintained unchanged. This is achieved with the application of the nearest neighbors methodology in order to distinguish the various manifolds of the dataset.

2.3.2. Measuring devices

As the only recorded data utilized by the algorithm are the three-phase voltage and current waveforms at the data collection points which are then transformed to the corresponding phasors, the method's data requirements are covered by the existing devices that provide synchronized measurements or a time stamp. These can be either measuring devices such as the Phasor Measurement Units or devices such as the data concentrators. These devices can be found in multiple points of the grid while the time stamp feature is expected to be included in the measuring devices that are planned to be installed by the DSOs. Therefore, the application of the method is considered to require minimum installation investments. Regarding the transmission of the data, the algorithm does not require the collection of data in real time, therefore, the existing technology is also sufficient for its application. Specifically, the algorithm was designed taking into consideration a data transmission every 15 min. The algorithm would locate the fault based on the last available measurements and there would be enough time to retrain the model, in case this is necessary.

2.4. Data pre-processing

As ML algorithms are data centered, the correct data processing before the model's training is important for the success of the prediction model. Data pre-processing consists of many steps and while some of them are necessary in all applications, e.g. the splitting of the data, others, e.g. the scaling of the data, depend on the employed model.

Splitting the data into smaller datasets is the first step that needs to be taken before training a model. The dataset is split into training, validation and test datasets. Some models such as the tree-based ones perform internal cross-validation, therefore there is no need for a separate validation dataset. Here the data were split with an 80/20 ratio between the training and test sets. Omitting the splitting leads to unreliable testing results of the prediction model, since the model is already fitted to the same data points that are used to test its accuracy. Thus, even though there is a high prediction accuracy in the testing, the model will be unable to perform well for new, unseen data. This phenomenon, known as overfitting, is one of the most important factors that require attention during the training of a prediction model. The dataset split is one of the ways to minimize it, nevertheless, further processing such as regularization may still be needed. Overall, the overfitting of a model to the training data can be tested by comparing the training and test accuracy of the model. A lack of similarity between the two values is an indication of overfitting.

Another influencing parameter on the prediction model's performance is the scale of the features. Most models consider features measured in larger units to be more important, therefore, these features are weighed in more during the training and decision making process than features measured in smaller units, resulting in biased prediction results. An exception to this are the tree-based models that evaluate one feature at a time thus the different scale of the features does not affect the algorithm. Nevertheless, the effective training of most models as well as the application of other ML processes such as feature selection and dimensionality reduction rely on the scaling of the features to comparable sizes. Even though only tree-based models were employed in this study, the dimensionality reduction techniques applied to the dataset required the scaling of the data. There are various scaling methodologies, depending on the type of the data and the type of

Table 1
Hyperparameter values.

No. of gradient boosting trees	700
Maximum tree depth for base learners	7
Subsample ratio of the training instance	0.7
Min sum of instance weight needed in a child	3
Boosting learning rate	0.1
Subsample ratio of columns when constructing each tree	0.7
Objective	“reg:squarederror"

application. The scaler used in this case was the RobustScaler. It is a scaler focused on the elimination of outliers. This is accomplished by the scaling of the data based on the quantile range, i.e. the range between the 1st and 3rd quartile, and not the whole range of values. Furthermore, the RobustScaler places the mean of the data on point zero. This is a prerequisite for dimensionality reduction methods such as the PCA; point zero is the common cross point of all the linear subspaces formed by the PCA.

2.5. Hyperparameter tuning

Each AI model is characterized by various parameters called hyperparameters which define its basic properties and as a result the shape and performance of the model as well as the duration of the training process. The tuning of the model’s hyperparameters is one of the biggest challenges during the training process. The hyperparameter tuning does not rely on any clearly defined rules but rather on the programmer’s experience and intuition. The process has been partly optimized by the appearance of algorithms that evaluate the possible combinations of the hyperparameters’ potential values and return the set of hyperparameter values that leads to the highest accuracy. Still, the designer is required to select the hyperparameters to be tuned and provide the values to be tested for each one of them. In the RF and XGBoost models the most frequently tuned as well as crucial parameters for the model’s performance are the number of decision trees, the maximum depth of each tree, the number of features evaluated during each node split and the amount of data comprising the bootstrapped dataset.

The potential and selected values of these parameters for the XGBoost model are presented in Table 1. For the selection of the final values the RandomizedSearchCV meta-estimator was utilized. The meta-estimator creates a grid with all the possible combinations of the hyperparameter values given by the programmer. Then it randomly selects a number of cases that is previously set by the programmer to be tested and with the use of cross-validation it returns the hyperparameters leading to the model’s highest accuracy. Although this meta-estimator does not validate the performance of all the possible hyperparameter combinations, it is efficient and time saving. It should be noted that due to the randomness of the process, these parameters vary each time the code is run, nevertheless the changes are slight and do not affect the overall accuracy of the algorithm.

3. Case study

As commented earlier there is a lack of real data regarding the faulted behavior of LV grids. Therefore, for the validation of the presented fault location method, and thus, the training and testing of the AI models, a modified version of the CIGRE European LV benchmark [50] was simulated in Simulink in order to generate the required datasets. The selected grid with the additions of photovoltaics (PVs) and meters encloses the important characteristics of a real life small LV grid thus providing useful test data. The grid’s topology is presented in Fig. 4.

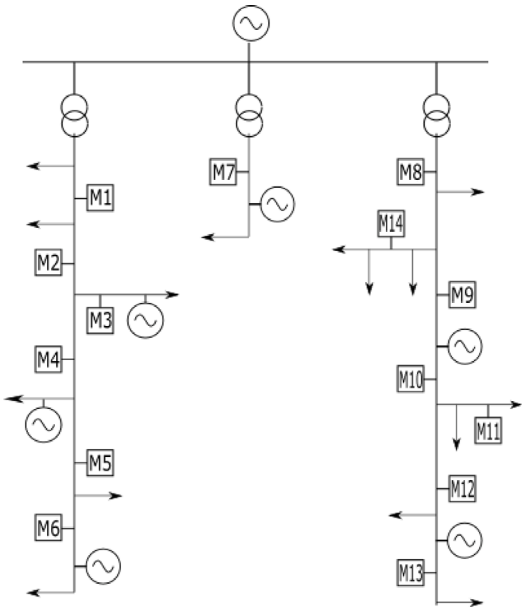


Fig. 4. Modified CIGRE European LV benchmark.

The layout of the original grid’s elements is maintained, with three PVs added in the first feeder, one in the second and two in the third. The nominal power of the PVs is different in each feeder and equals 5 kW for each PV on the first feeder, 13 kW for the PV on the second feeder and 10 kW for each PV on the third. Furthermore, various meters are placed throughout the grid, as indicated in Fig. 4 with the squared ‘M’ followed by a number. The meters were set in such a way as to cover the full length of the grid and enable the data and sensitivity analyses performed. Even though only 6 meters are the necessary ones, the data obtained from the rest were also used in order to showcase the potential and advantages of the proposed data management strategy. Finally, as mentioned in Section 2.3.2, when it comes to the collected variables, the algorithm has very low requirements that can be fulfilled by any measuring device, thus the method can be applied to any LV grid and it is not designed to accommodate specifically the particularities of this case study.

3.1. Data generation

For the generation of a realistic and diverse dataset several scenarios of the grid’s normal and faulty operation were simulated. The current research focuses on the location of shunt faults in LV grids. Specifically, the faults that are studied here are the single-phase, double-phase and three-phase faults that occur either between the phases or between the phases and the ground. The parameters modified during the simulations include the PV generation, the fault resistance, the fault location and the fault type. The ranges of the elements’ values are presented in Table 2. During the simulations, the PVs were considered to be operating in steady state while any transient phenomena related to them or the converters’ operation were ignored.

The generated dataset comprises the aforementioned variables, the node at the end of each faulted branch and the three phase current and voltage phasor measurements before and after the fault. The post-fault measurements were recorded within half cycle from the fault’s occurrence, before the activation of the protection devices. Furthermore, prior to the fault the grid was set to operate in steady state without transient phenomena. The number of simulation scenarios was 21250. The processor used during the simulations and the execution of the algorithms was an Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz.

Finally, in order to create a dataset as close to a real life one as possible, the generated dataset was sampled and some examples were randomly discarded. Thus, the final dataset was non-homogeneous. For

Table 2
Grid element values.

PV generation levels	Fault resistance	Fault location	Fault type
1st branch: 0, 800, 1700, 3000, 5000 W 2nd branch: 0, 4000, 7000, 10000, 13000 W 3rd branch: 0, 2500, 5000, 7500, 10000 W	17 values in the range of $[0, 40]\Omega$	25 different locations Distance from the feeder $[35, 315]m$	All 10 types of shunt faults

Table 3
Voltage and current values in each analyzed dataset.

1st dataset	2nd dataset	3rd dataset
$\overline{I}^{b}_{ph} = I^{b}_{ph} < \theta$ $\overline{V}^{b}_{ph} = V^{b}_{ph} < \theta$	$\overline{I}^{a}_{ph} = I^{a}_{ph} < \theta$	$\Delta I_{ph} = \frac{\overline{I}^{a}_{ph}}{\overline{I}^{b}_{ph}}$
$\overline{I}^{a}_{ph} = I^{a}_{ph} < \theta$ $\overline{V}^{a}_{ph} = V^{a}_{ph} < \theta$	$\overline{V}^{a}_{ph} = V^{a}_{ph} < \theta$	$\Delta V_{ph} = \frac{\overline{V}^{a}_{ph}}{\overline{V}^{b}_{ph}}$

b: value measured before the fault, a: value measured after the fault, ph: each of the three phases.

the evaluation of the data and the comparison of the dimensionality reduction techniques studied here, the size of the sampled dataset was randomly selected to be (10200, 176), where 10200 is the number of examples and 176 is the number of features. For the fault location part, the optimum dataset size is studied in Section 4.1 as part of the sensitivity analysis of the algorithm.

3.2. Data evaluation and correlation

As previously discussed, the analysis of a research's utilized data is invaluable for the validation and the interpretation of the results. Therefore, in order to verify the quality of the generated data, a statistical analysis of the utilized data from the simulated modified CIGRE grid was performed. This part of the study aims at providing useful information regarding the behavior of the basic variables measured in a small LV grid with RES and different kinds of loads. The presented data analysis refers to the fault location algorithm, so the target value is the distance between the fault and the main feeder. Regarding the recorded variables, i.e. the voltage and current, these were split into magnitude and angle values. The rest of the variables used as features were the generated PV power, the fault resistance and the type of fault.

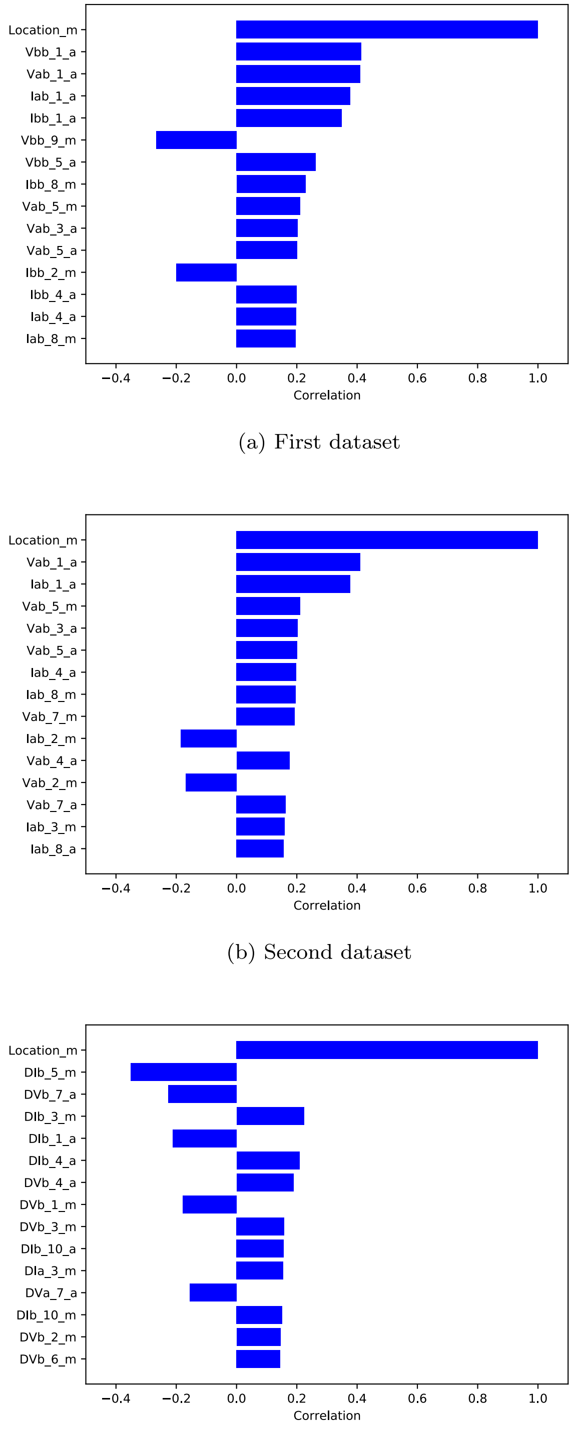
Three sets were formed from the collected data. The difference between them is the current and voltage values. As shown in Table 3, the first dataset contains the voltage and current before and after the fault, the second only the values measured after the fault and the third the ratio of each variable's values before and after the fault $\Delta I, \Delta V$. The variables' names in the first two datasets contain the letters a or b in the second position, with a denoting after and b before the fault. Then, the letter in the third position corresponds to the measured phase. The number contained in all the names points to the meter that performed the measurement, while the letters m and a at the end refer to the magnitude and angle of the measured value. In the case of the third dataset, the letter referring to whether the value was collected before or after the fault is missing, as the variables are the ratio of the two values. Finally, due to the large amount of features contained in the datasets, the plots illustrate only the features with the highest scores in each part of the analysis.

The Shapiro normality test showed that the datasets do not follow a Gaussian distribution. The probability of similarity for all of them was smaller than 0.05, which is the general threshold set as a rule of thumb. As commented earlier, for datasets that do not follow a normal distribution the most frequent correlation calculation methods are either the Kendall's coefficient or the Spearman's coefficient. The Pearson's correlation coefficient was also calculated, but only for validation purposes, as it assumes a normal data distribution. Figs. 5 and 6 confirm the theoretical observation that the Kendall's coefficient is smaller than that of the Spearman's. Based on the same figures, it can be concluded that there is a stronger correlation between the rank of the features and the rank of the target value in the first dataset, i.e. the variables of the first dataset show stronger monotonic relationships with the target value. Nevertheless, the different types of faults lead to an overall weak monotonic relationship between the features and the target. Moreover, some of the features showing high correlation with the target value, such as the currents before the fault cannot be individually used for the prediction of the fault distance. In cases like the ones studied here where the fault is a sudden event the values before the fault do not have a direct physical relationship with the fault's location. Therefore, even though these variables seem to be correlated, the respective features should be ignored during the dataset selection.

Regarding Pearson's correlation coefficient, as illustrated in Fig. 7, in a small LV grid such as the one examined, the linear correlation between the features and the target value is rather weak. This is to be expected as the non-linearity of these relations is obvious already from the analytical equations presented in the impedance-based methods and is enhanced by the tree-shaped topology of LV grids and the RES added to the grid.

An important metric that can provide a valuable insight to the dataset and assist the successful development of an AI fault location method is the examination of the features' importance. In Fig. 8 it can be observed that the features assisting the most in the split of a tree-based predictive model are the ones measuring the difference between the voltages and currents before and after the fault, which are contained in the third dataset. The other two datasets also contain features with high importance, but not as high as those in the third dataset. Furthermore, both the correlation coefficients as well as the features' importance point to the same meters as the most informative ones. These are the meters found in the beginning of each feeder and a meter placed in the middle/end of the first feeder. This is a useful observation for the installation of new meters and the collection and analysis of only the necessary data. Finally, in the analyzed grid, the data coming from the phase b of the grid appear to be the most useful for the decision making process of the prediction model. This is stemming from the unbalanced line impedance. Hence, in grids with unbalances such as the one studied here the identification of the prevailing phase could be useful for the optimum utilization of the available measurements.

On one hand the results show the difficulty of conventional methods to generalize the complex relations between the voltage /current values and the fault location for different LV grids. On the other hand it appears that tree-based ML models are capable of providing accurate predictions for these datasets; this needs to be verified by the metrics of the final model. In relation to the selected dataset, due to the complexity of the problem, the decision was made based on the optimum balance between the correlated variables, the features' importance and the actual physical meaning of the variables. Emphasis was given to the last two factors since, as commented earlier, the correlation between the features and the target value was rather weak for all the datasets. Therefore, the third dataset was chosen as the most informative for the prediction model. The aforementioned dataset contains all the information regarding the state of the grid before and after the fault in a compact form and leads to better tree splits, hence more accurate results with less features.

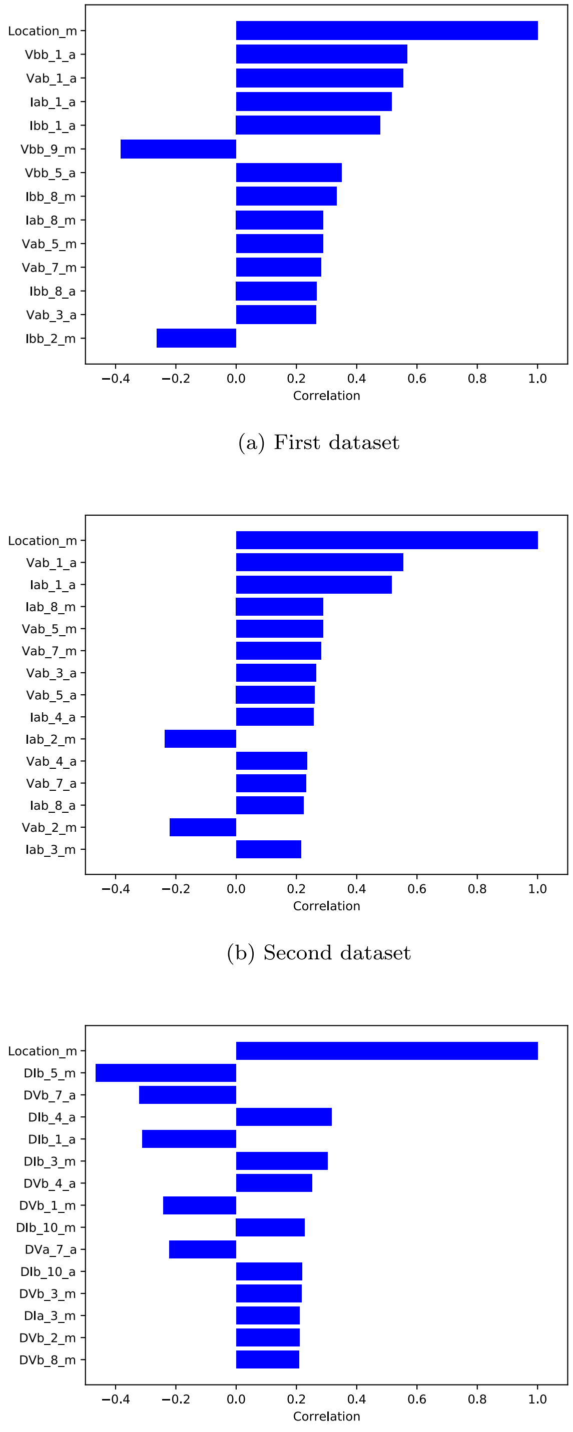


(a) First dataset

(b) Second dataset

(c) Third dataset

Fig. 5. Kendall's coefficient with the fault location.



(a) First dataset

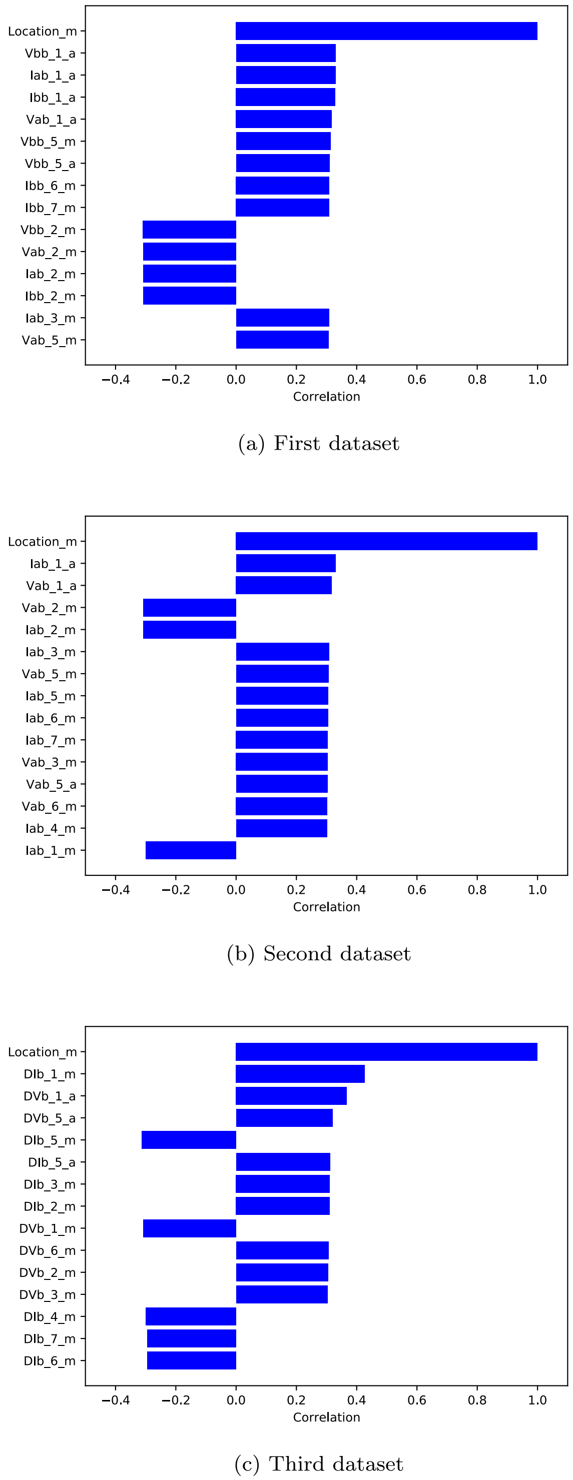
(b) Second dataset

(c) Third dataset

Fig. 6. Spearman's coefficient with the fault location.

3.3. Dimensionality reduction method selection

As part of the data management strategy described in Section 2.3, a dimensionality reduction process was applied before the training of the prediction model. The goal of this process was to increase the algorithm's accuracy with out-of-sample data, minimize the overfitting and lower the CT and complexity. For the selection of the appropriate dimensionality reduction technique five methods were compared, as analyzed in Section 2.3.1. Their comparison was conducted on the basis of (a) the required CT for the dimensionality reduction, (b) the required CT for the prediction of the final result with the use of the reduced dataset, without taking into consideration the tuning of the hyperparameters, and (c) the prediction accuracy. The results for each method, for a 30-dimensional space, are presented in Table 4. Both the number of dimensions and the dimensionality reduction method were selected based on the optimum trade-off between the CT and the accuracy. Fig. 9 illustrates the mean square error (MSE) in relation to the CT for the T-SVD method for a range of [20, 90] dimensions. It can be observed

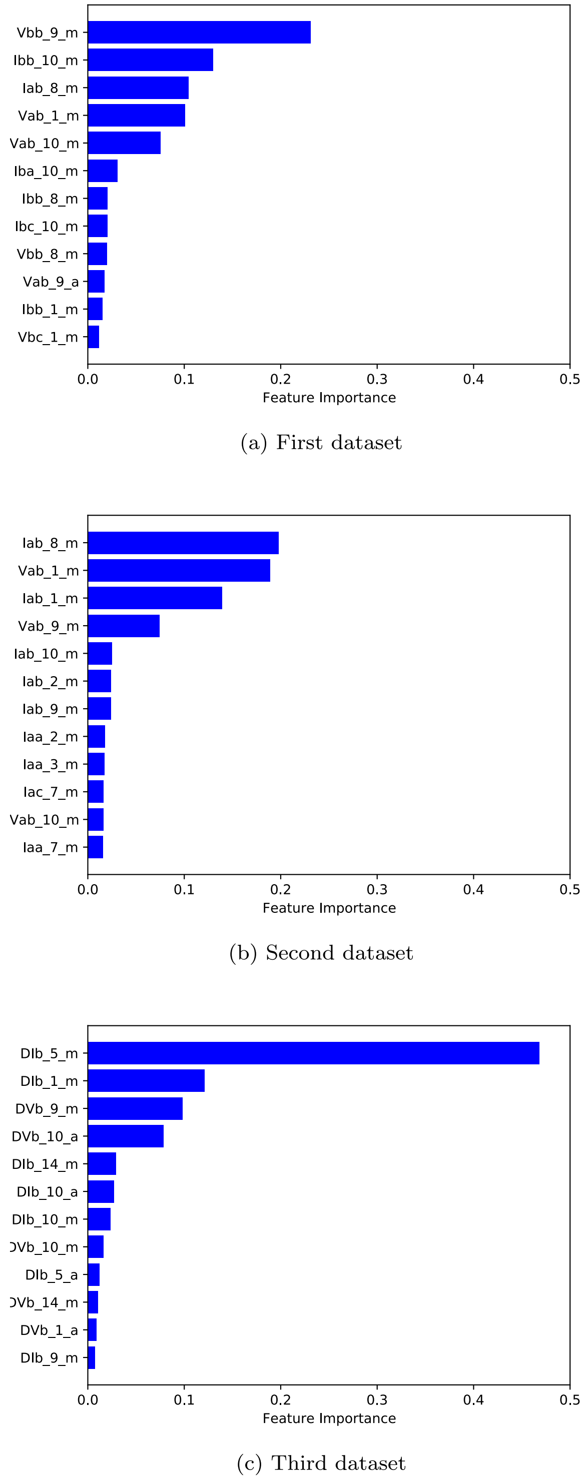


(a) First dataset

(b) Second dataset

(c) Third dataset

Fig. 7. Pearson’s coefficient with the fault location.



(a) First dataset

(b) Second dataset

(c) Third dataset

Fig. 8. Features’ importance.

that there is an almost inversely exponential relation between the two variables. According to the multi-objective optimization process that was followed, the number of dimensions selected was 30. Moreover, based on the results presented in Table 4 the selected dimensionality reduction methodology was the T-SVD. Even though Kernel PCA leads to the lowest MSE, T-SVD combines a similar error with a much lower execution time.

In accordance with the proposed data management strategy, in this study 10 of the most important features were kept intact and the rest

Table 4
Comparative table of the dimensionality reduction methods.

Method	CT of dimensionality reduction (s)	MSE (m^2)	Total CT (s)
PCA	0.38	119.25	10.08
KPCA	76.6	87.8	85.8
FastICA	1.3	155.44	11.1
T-SVD	0.18	95.61	11.65
ISOMAP	102.78	696.66	111.7

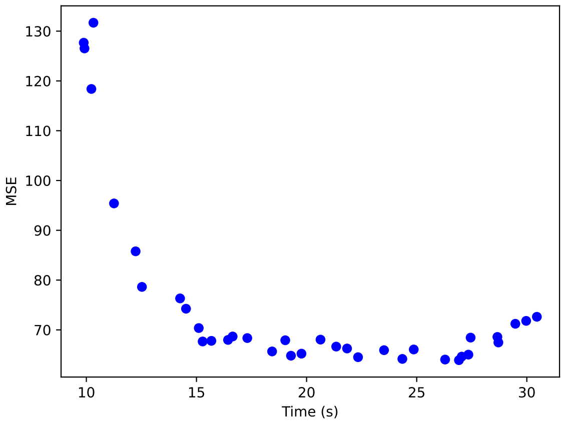


Fig. 9. The MSE in relation to the CT of the algorithm for a range of [20, 90] dimensions with the use of T-SVD.

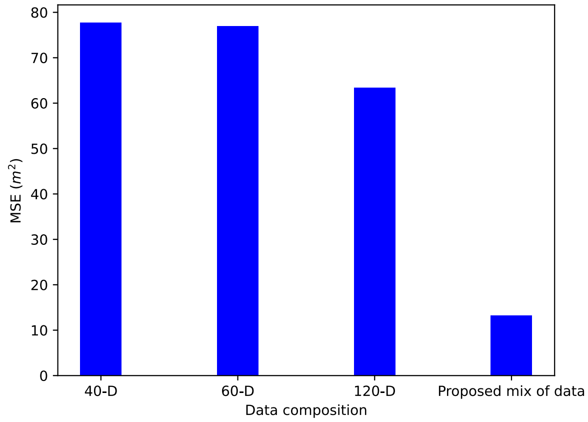


Fig. 10. The MSE of the algorithm in relation to the dataset constitution.

were reduced to 30 dimensions, forming a new dataset with a total of 40 features. The most important features were selected according to the features' importance analysis presented in the previous section. As illustrated in Fig. 10, the proposed technique is much more efficient than the use of a higher number of dimensions in the dimensionality reduction process and it manages to concentrate the most informative parts of the original dataset in less than a third of its size.

3.4. Fault location

For the location of the faulted point an XGBoost regressor was trained with the dataset created after the data pre-processing. The performance of the prediction model was evaluated based on the MSE, the mean absolute error (MAE), the mean percentage error (MPE) and the CT. Additionally, the overfitting of the model was examined by comparing the train and test accuracy, as calculated by the RandomizedSearchCV meta-estimator with the use of the R^2 metric. The aforementioned metrics are defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^*)^2 \tag{9}$$

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - Y_i^*|}{n} \tag{10}$$

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{(Y_i - Y_i^*)}{Y_i} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2} \tag{12}$$

Table 5
Fault location prediction model results.

MSE (m^2)	13.26
MAE (m)	1.69
CT without hyperparameter tuning (s)	11.89
CT with hyperparameter tuning (s)	735.99
Train accuracy (%)	99.9
Test accuracy (%)	99.8

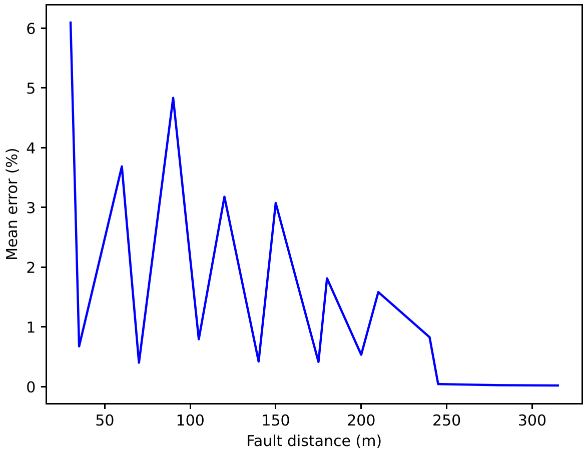


Fig. 11. MPE of the algorithm in relation to the fault's distance from the feeder.

where Y is the real value of the fault distance, Y^* is the predicted value, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and n is the total number of examples.

Table 5 presents the results for the trained model. The use of the processed dataset has lead to a radical decrease of the MSE and even a further decrease of the algorithm's CT. It should be also pointed out here that the CT for the algorithm trained with the original dataset, without the use of the dimensionality reduction but with the tuning of the hyperparameters, leads to a CT of 1970,1s or 32 min. This is more than twice the CT of the method after the use of dimentionality reduction. Furthermore, the train and test accuracy are both high and almost identical, thus indicating the lack of overfitting.

Finally, the exceptional performance of the developed method is illustrated also in Fig. 11, as the MPE of the method is less than 6% in all cases. The spikes in the plot correspond to the grid locations that are within the same distance from the feeder but are placed on different branches. The more branches with equally distant points from the feeder, the higher the algorithm's error. This is particularly noticeable for shorter distances where the voltage values are similar in all the branches.

4. Sensitivity analysis

For the verification of the algorithm's robustness an analysis of its sensitivity to a set of highly influencing parameters was performed. These parameters are the number of examples included in the dataset, the loss of measurements, the fault resistance and the injected power from the PVs to the grid.

4.1. Dataset size

The volume of data required for the effective training of an AI model constitutes one of the biggest concerns related to its application. It is often believed that there is a need for vast data storage as a prerequisite for the utilization of an AI model. Therefore, the dependence of the proposed algorithm on the number of training and testing examples was

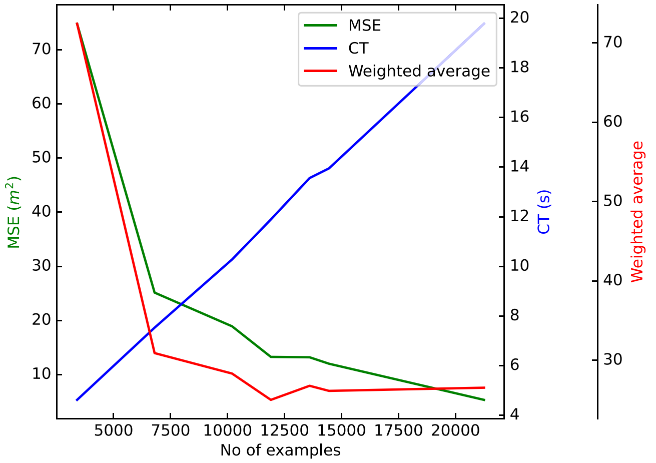


Fig. 12. The MSE and CT of the algorithm in relation to the number of utilized examples.

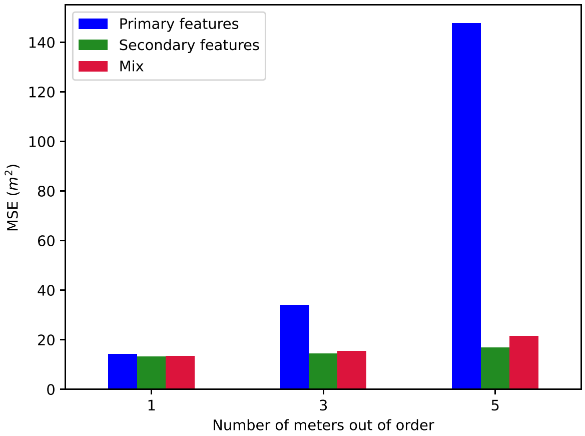


Fig. 13. The MSE of the algorithm in the case of data loss from one, three or five meters.

analyzed. The original dataset contained 21250 examples. In Fig. 12 the MSE and the training time in relation to the amount of examples is presented. As expected, the more the utilized examples the lower the MSE and the higher the CT. Even though the MSE is high for the lower end of the data volume, it drops significantly for more than 6800 examples. Such an amount of data is easily collected and stored. Hence, the presented algorithm combines high accuracy with low data and storage requirements.

The optimum dataset size is approached another case of multi-objective optimization, with the target being the balance point between the lowest CT and the lowest MSE. Both parameters are considered almost equally important with CT being given a slightly higher importance factor. Specifically, the weights selected were 1.1 for the CT and 0.9 for the MSE. The red line in Fig. 12 illustrates the weighted average curve of the CT and the MSE in relation to the number of examples. The minimum of the curve constitutes the optimum point and corresponds to 11900 examples. This is the number of examples used for the training and testing of the presented fault location algorithm.

4.2. *Data loss*

Another factor that could have an important impact on the performance of a fault location method is the loss of data due to a communication error or the malfunction of one or more measuring devices. Specifically, the possibility that the measurements from one, three or five of the grid's meters could not be collected was studied. The missing values were replaced with 0. For each of the three studied cases five random combinations of failing meters were used and the mean value of the method's MSE was calculated. As this algorithm relies more on certain measuring devices than others, three different scenarios were tested regarding the meters that failed. The first scenario explored the possibility that the failing devices were the ones providing the more important data that were utilized by the algorithm in their original form. These are characterized as primary devices. In the second scenario it was considered that the missing values came from the devices providing the less important data, thus characterized as secondary devices. Finally, in the third the data loss originated from both the primary and secondary devices. Meters 1, 4, 5, 9, 10 were selected as the failing primary devices and meters 2, 6, 8, 13, 14 as the failing secondary devices. In the case that three of both primary and secondary devices failed, one of them was considered to be a primary meter and two secondary meters. In the case of five failing meters the ratio was two primary and three secondary meters.

As it can be seen in Fig. 13, the algorithm depends heavily on the primary measuring devices and loss of data from them can result in high errors in case three or more of them fail simultaneously. Nevertheless, this is a very improbable scenario that does not characterize the performance of the algorithm. In the more probable cases of failures, those only in secondary meters or both kinds of meters, the MSE is almost the same as that under normal conditions. Therefore, the algorithm is considered to be robust against possible data loss.

4.3. *Fault resistance*

Fault resistance is a defining parameter for all fault diagnosis methods. Its effect on the fault current, a fundamental variable to related calculations, can have a great impact on the method's accuracy. Thus, in this study a broad range of fault resistances was simulated in order to identify the algorithm's sensitivity to it. More specifically, the tested fault resistances were in the range of [0, 40] Ω. Higher impedance faults are out of this research's scope, therefore, this range was selected as the most representative for detectable faults in LV grids [51]. The magnitude of the fault resistance depends mainly on the fault and ground type. Fig. 14 depicts the high accuracy of the algorithm for the whole range of fault resistances. It can be observed that the accuracy is lower for the lower values of the fault resistance. The reason behind that is the greater variation of the current's values in this area of resistances, as it can be seen in Fig. 15 for measurements obtained from meter 8. Hence, it is challenging for the algorithm to distinguish between the different cases and predict the correct target value. Nevertheless, the accuracy is constantly above 99% deeming the effect of the fault resistance on the proposed algorithm negligible.

4.4. *PV penetration level*

Finally, a major parameter in the smart grids era is the integration of RES. In fault location methods the power injected by the RES can affect significantly the result's accuracy. Therefore, in this study the effect of five different PV power generation levels was analyzed. The generation levels are presented in Table 2. As illustrated in Fig. 16 there is no clear pattern between the generated power and the algorithm's accuracy. Nonetheless, the consistently high accuracy indicates that the algorithm is practically unaffected by the various PV penetration levels.

5. Conclusions

In this study a novel artificial intelligence (AI) – based fault location method for low voltage grids is presented. The transformation of traditional electricity grids to smart grids has rendered most of the conventional fault location methods obsolete, however, it has also offered opportunities for innovation due to the increased observability over the

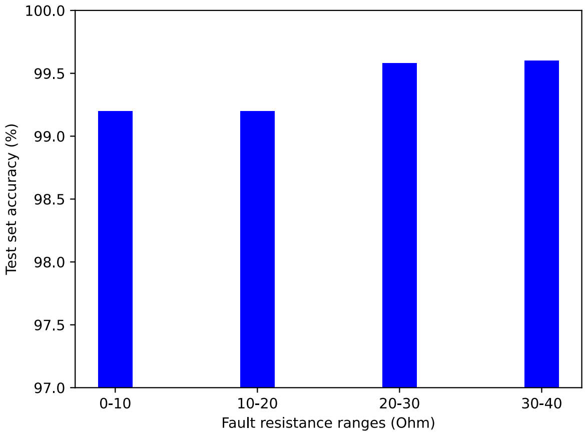


Fig. 14. The accuracy of the algorithm in relation to the fault resistance value ranges.

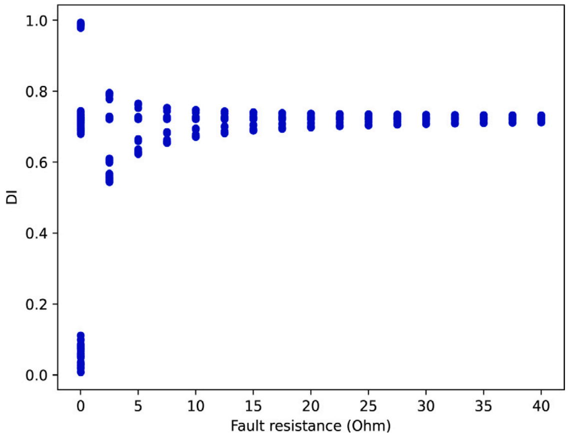


Fig. 15. The ratio between the current before and after the fault in relation to the fault resistance as measured by meter 8.

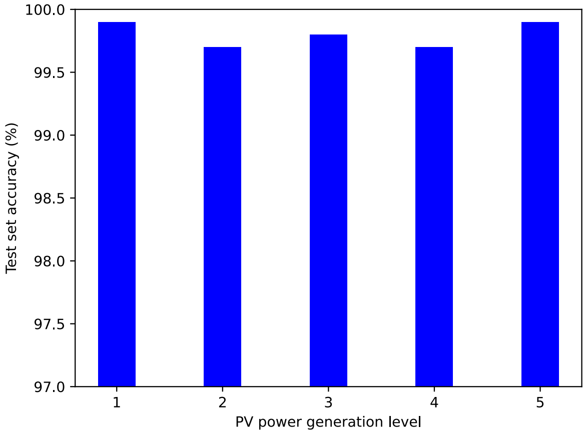


Fig. 16. The accuracy of the algorithm in relation to the PV power generation levels.

grid. The measuring devices installed throughout the grid are expected to multiply and allow the collection of large amounts of data. This favors the implementation of AI methods but at the same time it raises the need for appropriate data management. AI has proven to be highly accurate in a variety of applications and, as a continuously developing field, constitutes a flexible and sophisticated tool. Nevertheless, when applied without proper data analysis and processing it can lead to high computational times (CT), reduced accuracy or overfitting of the model.

The proposed algorithm aims at solving these problems by optimizing the application of AI in a fault location method. This is achieved by evaluating the collected data and developing a data management strategy. More specifically, first, the data analysis results point to the form with which the recorded variables should be included in the dataset. The dataset is then processed following the data management strategy proposed. This part of the algorithm reduces the data volume, and thus the CT of the method, by transforming the least informative features with the use of the Truncated-SVD technique while keeping the 10 most informative features in their original form. This step results in the efficient exploitation of all the available data and, at the same time, the generalization of the algorithm. The reduced dataset is used as an input for the training of an XGBoost model, which combines low overfitting with high computational speed and accuracy. The final algorithm is characterized by superior performance, with a mean squared error of 13.26 and a training and testing accuracy above 99% when evaluated with data generated from the simulation of the CIGRE European LV benchmark.

Overall, the proposed method has low input data requirements, that can be fulfilled by the existing measuring devices, hence it is easily applicable. Additionally, the location of the measuring devices is not taken into consideration in the data management scheme, therefore, the method is independent of the grid's topology and can be applied to all LV grids. Furthermore, it is robust against parameters that could affect its performance such as the fault resistance, the PV penetration levels and the loss of data. Regarding the first two parameters, the accuracy of the algorithm for the wide range of tested values remains above 99%. Finally, in the case of lost data due to a malfunction in one or more meters or a communication failure, it is highly improbable that such a disruption will significantly affect the response of the algorithm.

CRediT authorship contribution statement

P. Stefanidou-Voziki: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **D. Cardoner-Valbuena:** Methodology, Software, Formal analysis. **R. Villafafila-Robles:** Supervision, Validation, Writing – review & editing. **J.L. Dominguez-Garcia:** Supervision, Validation, Writing – review & editing.

Declaration of competing interest

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

References

[1] Gaur VK, Bhalja BR, Saber A. New ground fault location method for three-terminal transmission line using unsynchronized current measurements. Int J Electr Power Energy Syst 2022;135:107513. http://dx.doi.org/10.1016/j.ijepes.2021.107513.

[2] Lin T-C, Xu Z-R, Ouedraogo FB, Lee Y-J. A new fault location technique for three-terminal transmission grids using unsynchronized sampling. Int J Electr Power Energy Syst 2020;123:106229. http://dx.doi.org/10.1016/j.ijepes.2020.106229.

[3] Dobakhshari AS. Fast accurate fault location on transmission system utilizing wide-area unsynchronized measurements. Int J Electr Power Energy Syst 2018;101:234–42. http://dx.doi.org/10.1016/j.ijepes.2018.03.009.

[4] Bretas A, Orozco-Henao C, Marín-Quintero J, Montoya O, Gil-González W, Bretas N. Microgrids physics model-based fault location formulation: Analytic-based distributed energy resources effect compensation. Electr Power Syst Res 2021;195:107178. http://dx.doi.org/10.1016/j.epsr.2021.107178.

[5] Yang H, Liu X, Guo Y, Zhang P. Fault location of active distribution networks based on the golden section method. Math Probl Eng 2020;2020:1–9. http://dx.doi.org/10.1155/2020/6937319.

[6] Aboshady F, Thomas D, Sumner M. A new single end wideband impedance based fault location scheme for distribution systems. Electr Power Syst Res 2019;173:263–70. http://dx.doi.org/10.1016/j.epsr.2019.04.034.

[7] Borghetti A, Bosetti M, Nucci CA, Paolone M, Abur A. Integrated use of time-frequency wavelet decompositions for fault location in distribution networks: theory and experimental validation. IEEE Trans Power Deliv 2010;25(4):3139–46. http://dx.doi.org/10.1109/TPWRD.2010.2046655.

[8] Goudarzi M, Vahidi B, Naghizadeh R, Hosseinian S. Improved fault location algorithm for radial distribution systems with discrete and continuous wavelet analysis. Int J Electr Power Energy Syst 2015;67:423–30. http://dx.doi.org/10.1016/j.ijepes.2014.12.014.

[9] Bountouris P, Guo H, Tzelepis D, Abdulhadi I, Coffele F, Booth C. MV faulted section location in distribution systems based on unsynchronized LV measurements. Int J Electr Power Energy Syst 2020;119:105882. http://dx.doi.org/10.1016/j.ijepes.2020.105882.

[10] Jamali S, Bahmanyar A. A new fault location method for distribution networks using sparse measurements. Int J Electr Power Energy Syst 2016;81:459–68. http://dx.doi.org/10.1016/j.ijepes.2016.02.046.

[11] Jiang Y. Data-driven fault location of electric power distribution systems with distributed generation. IEEE Trans Smart Grid 2020;11(1):129–37. http://dx.doi.org/10.1109/TSG.2019.2918195.

[12] Okumus H, Nuroglu FM. A random forest-based approach for fault location detection in distribution systems. Electr Eng 2020. http://dx.doi.org/10.1007/s00202-020-01074-8.

[13] Chen K, Hu J, Zhang Y, Yu Z, He J. Fault location in power distribution systems via deep graph convolutional networks. IEEE J Sel Areas Commun 2020;38(1):119–31. http://dx.doi.org/10.1109/JSAC.2019.2951964, arXiv:1812.09464.

[14] Sonoda D, de Souza AZ, da Silveira PM. Fault identification based on artificial immunological systems. Electr Power Syst Res 2018;156:24–34. http://dx.doi.org/10.1016/j.epsr.2017.11.012.

[15] Peng N, Liang R, Wang G, Sun P, Chen C, Hou T. Edge computing-based fault location in distribution networks by using asynchronous transient amplitudes at limited nodes. IEEE Trans Smart Grid 2021;12(1):574–88. http://dx.doi.org/10.1109/TSG.2020.3009005.

[16] Mirshekali H, Dashti R, Keshavarz A, Torabi AJ, Shaker HR. A novel fault location methodology for smart distribution networks. IEEE Trans Smart Grid 2021;12(2):1277–88. http://dx.doi.org/10.1109/TSG.2020.3031400.

[17] Silos-Sanchez A, Villafafila-Robles R, Lloret-Gallego P. Novel fault location algorithm for meshed distribution networks with DERs. Electr Power Syst Res 2020;181:106182. http://dx.doi.org/10.1016/j.epsr.2019.106182.

[18] Navaneethan S, Soraghan JJ, Siew WH, McPherson F, Gale PF. Automatic fault location for underground low voltage distribution networks. IEEE Trans Power Deliv 2001;16(2):346–51. http://dx.doi.org/10.1109/61.915506.

[19] Siew W, Soraghan J, Stewart M, Fisher D, Fraser D, Ltd P-E, et al. Intelligent fault location for low voltage distribution networks. No. 0327. 2007, p. 4.

[20] Pasdar AM, Sozer Y, Husain I. Detecting and locating faulty nodes in smart grids based on high frequency signal injection. IEEE Trans Smart Grid 2013;4(2):1067–75. http://dx.doi.org/10.1109/TSG.2012.2221148.

[21] Orcajo GA, Cano JM, Melero MG, Cabanas MF, Rojas CH, Pedrayes JF, et al. Diagnosis of electrical distribution network short circuits based on voltage park's vector. IEEE Trans Power Deliv 2012;27(4):1964–72. http://dx.doi.org/10.1109/TPWRD.2012.2210448.

[22] Sun K, Chen Q, Gao Z. An automatic faulted line section location method for electric power distribution systems based on multisource information. IEEE Trans Power Deliv 2016;31(4):1542–51. http://dx.doi.org/10.1109/TPWRD.2015.2473681.

[23] Jia K, Ren Z, Bi T, Yang Q. Ground fault location using the low-voltage-side recorded data in distribution systems. IEEE Trans Ind Appl 2015;51(6):4994–5001. http://dx.doi.org/10.1109/TIA.2015.2425358.

[24] Niu G, Zhou L, Pei W, Qi Z. A novel fault location and recognition method for low voltage active distribution network. In: 2015 5th International conference on electric utility deregulation and restructuring and power technologies. Changsha, China: IEEE; 2015, p. 876–81. http://dx.doi.org/10.1109/DRPT.2015.7432418.

[25] Silva N, Basadre F, Rodrigues P, Nunes MS, Grilo A, Casaca A, et al. Fault detection and location in low voltage grids based on distributed monitoring. In: 2016 IEEE international energy conference. Leuven, Belgium: IEEE; 2016, p. 1–6. http://dx.doi.org/10.1109/ENERGYCON.2016.7514000.

[26] Nunes M, Grilo A, Casaca A, Silva N, Basadre F, Rodrigues P, et al. Fault detection and location in low voltage grids based on RF-mesh sensor networks. In: CIRED workshop 2016. Helsinki, Finland: Institution of Engineering and Technology; 2016, p. 143 (4 .)–143 (4 .). http://dx.doi.org/10.1049/cp.2016.0743.

[27] Alamuti MM, Nouri H, Ciric RM, Terzija V. Intermittent fault location in distribution feeders. IEEE Trans Power Deliv 2012;27(1):96–103. http://dx.doi.org/10.1109/TPWRD.2011.2172695.

[28] Trindade FCL, Freitas W. Low voltage zones to support fault location in distribution systems with smart meters. IEEE Trans Smart Grid 2017;8(6):2765–74. http://dx.doi.org/10.1109/TSG.2016.2538268.

[29] Souto L, Meléndez J, Herraiz S. Fault location in low voltage smart grids based on similarity criteria in the principal component subspace. In: 2020 IEEE power energy society innovative smart grid technologies conference. 2020, p. 1–5. http://dx.doi.org/10.1109/ISGT45199.2020.9087707, ISSN: 2472-8152.

[30] Sapountzoglou N, Lago J, Raison B. Fault diagnosis in low voltage smart distribution grids using gradient boosting trees. Electr Power Syst Res 2020;182:106254. http://dx.doi.org/10.1016/j.epsr.2020.106254.

[31] Sapountzoglou N, Lago J, De Schutter B, Raison B. A generalizable and sensor-independent deep learning method for fault detection and location in low-voltage distribution grids. Appl Energy 2020;276:115299. http://dx.doi.org/10.1016/j.apenergy.2020.115299.

[32] Iman M, Giuntini A, Arabnia HR, Rasheed K. A comparative study of machine learning models for tabular data through challenge of monitoring Parkinson's disease progression using voice recordings. 2020, [Cs, Eess, Q-Bio] arXiv:2005.14257.

[33] Stefanidou-Voziki P, Cardoner-Valbuena D, Villafafila-Robles R, Dominguez-Garcia JL. Feature selection and optimization of a ML fault location algorithm for low voltage grids. In: 2021 IEEE international conference on environment and electrical engineering and 2021 IEEE industrial and commercial power systems Europe. 2021, p. 1–6. http://dx.doi.org/10.1109/EEEIC/ICPSEurope51590.2021.9584731.

[34] Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52(3/4):591–611. http://dx.doi.org/10.2307/2333709.

[35] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009, p. 1–4. http://dx.doi.org/10.1007/978-3-642-00296-0_5.

[36] Stepanov A. On the Kendall correlation coefficient. 2015, [Math, Stat] arXiv:1507.01427.

[37] Wissler C. The spearman correlation formula. Science 1905.

[38] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton: Routledge; 2017, http://dx.doi.org/10.1201/9781315139470.

[39] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016, p. 785–94. http://dx.doi.org/10.1145/2939672.2939785.

[40] Hastie T, Tibshirani R, Friedman J. Ensemble learning. In: The elements of statistical learning. New York, NY: Springer New York; 2009, p. 605–24. http://dx.doi.org/10.1007/978-0-387-84858-7_16.

[41] Wang W, Shi Y, Lyu G, Deng W. Electricity consumption prediction using XGBoost based on discrete wavelet transform. DEStech Trans Comput Sci Eng 2017;(aiea). http://dx.doi.org/10.12783/dtcse/aiea2017/15003.

[42] Abbasi RA, Javaid N, Ghuman MNJ, Khan ZA, Ur Rehman S, Amanullah. Short term load forecasting using XGBoost. In: Barolli L, Takizawa M, Xhafa F, Enokido T, editors. Web, artificial intelligence and network applications. Advances in intelligent systems and computing, Cham: Springer International Publishing; 2019, p. 1120–31. http://dx.doi.org/10.1007/978-3-030-15035-8_108.

[43] Yucong W, Bo W. Research on EA-Xgboost hybrid model for building energy prediction. J Phys Conf Ser 2020;1518:012082. http://dx.doi.org/10.1088/1742-6596/1518/1/012082.

[44] Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat 2010;2(4):433–59. http://dx.doi.org/10.1002/wics.101.

[45] Tenenbaum JB. Mapping a manifold of perceptual observations. In: Advances in neural information processing systems. Vol. 10. MIT Press; 1998, p. 682–8.

[46] Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 1999;10(3):626–34. http://dx.doi.org/10.1109/72.761722.

[47] Chan TF, Hansen PC. Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations. SIAM J Sci Stat Comput 1990;11(3):519–30. http://dx.doi.org/10.1137/0911029.

[48] Tenenbaum JB, Silva Vd, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23. http://dx.doi.org/10.1126/science.290.5500.2319.

[49] Thukaram D, Khincha HP, Vijaynarasimha HP. Artificial neural network and support vector machine approach for locating faults in radial distribution systems. IEEE Trans Power Deliv 2005;20(2):710–21. http://dx.doi.org/10.1109/TPWRD.2005.844307.

[50] Conseil international des grands réseaux électriques Comité d'études C6 and International Council on Large Electric Systems. Benchmark systems for network integration of renewable and distributed energy resources: Task force C6.04. In: [Brochures thématiques], CIGRÉ; 2014.

[51] Sapountzoglou N, Raison B, Silva N. Fault detection and localization in LV smart grids. In: 2019 IEEE milan powertech. Milan, Italy: IEEE; 2019, p. 1–6. http://dx.doi.org/10.1109/PTC.2019.8810799.