IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE POWER & ENERGY SOCIETY SECTION

Received 25 May 2025, accepted 17 June 2025, date of publication 23 June 2025, date of current version 30 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3582359

RESEARCH ARTICLE

High Impedance Fault Location in Distribution Systems: A Novel Approach With Enhanced Metrics and Intelligent Algorithms

GABRIELA NUNES LOPES[1], PEDRO I. N. BARBALHO[2], JOSÉ CARLOS MELO VIEIRA[2], (Member, IEEE), AND DENIS V. COURY[2]

[1]Department of Electrical Engineering, School of Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais 31270-901, Brazil
[2]Department of Electrical and Computer Engineering, São Carlos School of Engineering, University of São Paulo, São Carlos, São Paulo 13566-590, Brazil

Corresponding author: Gabriela Nunes Lopes (gabrielanuneslopes@ufmg.br)

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) under Grant 001; in part by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant 445374/2020-9, Grant 406453/2021-7, Grant 140235/2021-3, Grant 402334/2023-0, and Grant 308979/2022-2; and in part by São Paulo Research Foundation (FAPESP) under Grant 2020/06935-5 and Grant 2024/18735-1.

ABSTRACT Locating faults in a distribution system (DS) is a challenging task, especially with the increasing integration of distributed generation (DG). Among the various fault types in DSs, high impedance faults (HIFs) are particularly difficult to address due to their low fault currents and erratic behavior, often resulting from contact between a conductor and a low-conductivity surface. Despite of the numerous papers addressing HIFs, there is still a lack of consensus on the most effective metrics and techniques for their detection and location. This article proposes a comprehensive analysis aiming to identify the adequate metrics for artificial intelligent algorithms to locate HIFs in a DS in the presence of DG. This analysis provides valuable insights into developing more efficient HIF location methods. Therefore, these insights are applied to develop a new HIF location method, based on artificial intelligence. The analysis is carried out by evaluating the correlation between almost 11,000 metrics with the HIF distance. The metrics showing the strongest correlation with HIF location are then used as input data for intelligent algorithms designed to accurately estimate the HIF position. Subsequently, these algorithms are evaluated and compared based on their error rates in locating HIFs. Finally, a new intelligent-based method for HIF location in DSs is proposed and extensively evaluated, including a comparison with existing approaches and a generalization analysis, which showed high successful rates and strong potential for practical application.

INDEX TERMS High impedance faults, fault location, distributed generation, artificial intelligence.

I. INTRODUCTION

High impedance fault (HIF) diagnosis is a challenging issue in distribution systems (DSs). Firstly, HIF detection has attracted researchers' attention given the fault's characteristics, e.g., low current, randomness, and similarity to other events (such as transformer inrush current and non-linear load switching). Furthermore, the HIF location in DSs still needs to be improved as the research field is recent and only a few approaches consider the diverse conditions to which a DS may be exposed [1], [2]. The potential hazards posed by HIFs to living beings, the rising risks of bushfires, and the consequent impact on reliability indices for power utilities show the need to devise advanced and efficient HIF location methods (HIFLMs).

Existing HIFLMs predominantly rely on techniques designed for short-circuit events, which are not suitable for locating HIFs due to the distinct characteristics between HIFs and low-impedance faults [3]. In this context, [1] presents a review on HIFLMs. However, the article lacks an analysis of the metrics' implementation and a discussion of specific challenges related to distribution system applications. Moreover,

The associate editor coordinating the review of this manuscript and approving it for publication was Elizete Maria Lourenco.

it is noteworthy that the aspects related to the HIFLMs metrics and decision methods still need to be thoroughly explored or addressed in the existing literature. Some methodologies perform attribute selection, as in [2] and [4], but consider only a few limited options.

In order to present a summarized theoretical review on HIFLMs, Table 1 presents an overview of the main characteristics of HIFLMs documented in the literature. It highlights the different input electrical quantities, as existing algorithms utilize voltages, currents, or both as input signals. Furthermore, a wide array of metrics are utilized to identify the fault location. Some methodologies use signal pre-processing techniques such as Fourier, Wavelet, or Hilbert-Huang transforms to extract the metrics of interest [5], [6], [7], while others rely solely on the signal amplitude [8], [9]. The latter is predominantly used to build the system's impedance matrix to compute the system's power flow, and it depends on the correct modeling of the system and the HIF.

Table 1 also highlights three primary types of decision methods in HIFLMs. The first type relies on metrics calculated using different meters, enabling fault location determination through data comparison or database matching [5], [10], [11]. The second type is parameter-change-based, requiring prior knowledge of system parameters for fault location [12], [13], [14]. The third type involves intelligent algorithms [7], [9], [15], [16], offering high accuracy but requiring training and parameter definition, which remains unaddressed in the literature. They are based on different models, such as Artificial Neural Networks, Gaussian process regression, Support Vector Machines, and Extreme Learning Machines. Table 1 also specifies whether the HIFLMs determine the fault region or the fault-to-measurement distance, an essential consideration for utility maintenance crews. In summary, [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] are classified according to the decision method and metric utilized.

Recent works, as seen in [25], [26], and [27], have focused on analysing the zero-sequence component of voltage and current but have provided only the HIF region and not the location estimation. In [28], the authors proposed a location method using two Deep Neural Networks (DNNs), one to identify the fault branch and the other to estimate the location. However, the method requires a high computation burden as it uses the image of voltage-ampere curves as input. Overall, there is no consensus regarding the best metrics and decision methods for HIF location, highlighting the gap in this research field. Recent advances in HIF studies include [29], which delivers a combined detection and location scheme, but capable of identifying only the main section of the network. Reference [30] focuses on conductor level location by discriminating the faulted phase rather than performing the fault detection or location. Moreover, [31] employs a simple fault model based only on a high-resistance to test the proposed method. The algorithm uses a neural network to identify the faulty zone, but does not provide the fault distance.

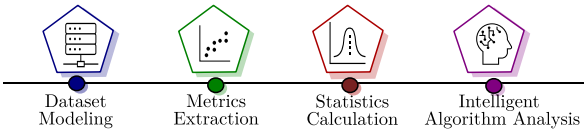


FIGURE 1. Flowchart of the study methodology.

This paper presents a comprehensive study that can help develop new tools, and also proposes a new approach for HIF location based on a step-by-step verification of the best tools to do so. It aims to conduct a comprehensive analysis of input metrics extracted from measurements in a system with distributed generation (DG) and background noise. The Pearson Correlation Coefficient (PCC) is used to select the most relevant metrics for HIF location. Subsequently, the best inputs are used to comprise the dataset of different intelligent algorithms, with a further comparison of the fault's location error. The primary objective is to provide a critical and detailed analysis aimed at identifying the most effective metrics and intelligent algorithms. Additionally, as a result of the performed analysis, this paper proposes a new High Impedance Fault Location (HIFL) method based on machine learning, with new generalization analysis. This work aims to assist researchers in developing more efficient HIFL methods applicable to systems with Distributed Generation (DG). The key contributions of this paper are:

- Proposing a comprehensive methodology to develop new HIF location methods by providing a critical and thorough assessment of new metrics for HIFLMs, and employing the Pearson correlation coefficient to determine the highest correlation amongst 10,800 metrics and fault distances. Moreover, the study includes comparing various intelligent algorithms in terms of their error rates when applied to HIF location;
- Designing a data-driven-based HIFLM, comprising outcomes (metrics and intelligent algorithms) of the comprehensive methodology mentioned in the previous item;
- Performing a generalization analysis of the proposed data-driven method, testing it in new scenarios not considered in the training set, an analysis not performed in existing methodologies.

A comparative analysis with existing methodologies is conducted as optimal solutions have been attained. This comparison demonstrates the potential contributions of the current study to the future development of HIFLMs.

II. METHODOLOGY

The main aim of this paper is to perform a step-by-step evaluation of metrics and decision methods to compose a new and effective HIF location method based on machine learning. To do this, a detailed analysis is carried out by following the methodology depicted in Fig. 1 and explained in the following sections. It starts with the dataset modeling,

TABLE 1. Electrical inputs, metrics and decision methods of existing HIF location methods.

Reference	Electrical Quantity	Metric	Decision Method	Output
[10]	Voltage and Current	Energy of First Detail Level of WT	Comparison among meters	region
[17]	Voltage and Current	Impulse among devices	Comparison among meters	region
[11]	Voltage	Voltage drop at the fault location	Comparison among meters	region
[18]	Current	Zero sequence current	Comparison among meters	region
[5]	Voltage	3 detail levels of WT	Comparison with database	region
[13]	Voltage and Current	Signals' amplitude	Comparison with database	distance
[19]	Voltage and Current	Lines Capacitive Current using power flow	Parameter Change	distance
[3]	Voltage and Current	Signal Amplitude to calculate power flow	Parameter Change	distance
[12]	Voltage	Harmonic source impact on voltage measurement	Parameter Change	distance
[20]	Voltage and Current	Least Square Estimation	Parameter Change	region
[14]	Voltage and Current	HIF influence on lines impedance using FT	Parameter Change	distance
[8]	Voltage	Signal amplitude	Parameter Change	distance
[21]	Current	Zero-sequence current decline	Parameter Change	distance
[22]	Voltage and Current	3^{rd} harmonic order using Fourier Transform	Parameter Change	distance
[23]	Magnetic Field	Magnetic Field	Parameter Change	distance
[15]	Voltage and Current	Line Capacitive Current	Artificial Neural Network (ANN)	distance
[6]	Current	5^{th} Detail Level of WT	ANN+Gaussian Process Regression	distance
[9]	Voltage and Current	Signals' amplitude	Adaptive Neurofuzzy Inference	distance
[16]	Voltage and Current	Signals' RMS and Angle	Artificial Neural Network	distance
[4]	Voltage and Current	Harmonics extracted by FT	HIF impedance probabilistic distribution	distance
[24]	Voltage	Wavelet coefficient entropy	Random Search Multilevel SVM	region
[7]	Current	Differential energy of EEMT-HHT components	Multi-kernel extreme learning machine	distance
[25]	Voltage and Current	Zero sequence current and undervoltage	Comparison among meters	region
[26]	Voltage and Current	Zero sequence current and voltage	Comparison among meters	region
[28]	Voltage and Current	Voltage-ampere curves	Deep Neural Networks	distance
[27]	Current	Zero-sequence current	Comparison among meters	region

WT: Wavelet Transform/ FT: Fourier Transform/ EEMT-HT: Ensemble empirical mode decomposition - Hilbert Huang Transform / SVM: Support Vector Machine

and then performs the metrics extraction, statistics calculation and intelligent algorithms analysis for the HIF location task. In this study, it is assumed that HIF detection is carried out by a separate module. This approach aligns with the majority of existing methodologies in the literature, which also operate under the premise that the fault has already been detected prior to the application of the location method.

A. HIF MODEL AND THE DATASET MODELING

Simulations were performed using the IEEE 34-node test system [32], modeled in the Alternative Transients Program (ATP) with the addition of a 1.1 MVA synchronous generator and a 556 kW load, as shown in Fig. 2, with the control parameters set according to [33]. This system was chosen because it is a real-life system, located in Arizona, USA. It has long single-phase and three-phase side branches, increasing the rigor of the tests. Moreover, the network is naturally unbalanced, which poses a challenge for HIF diagnosis methods to overcome. In summary, the chosen system is a benchmark, widely used to test fault diagnosis algorithms with different purposes in the literature.

The present study evaluated current and voltage signals measured at the substation, as well as at buses 858 (medium distance from the substation) and 840 (farthest bus from the substation, except for the DG bus). HIFs occur after a broken conductor contacts a surface [34]. Thus, meters downstream

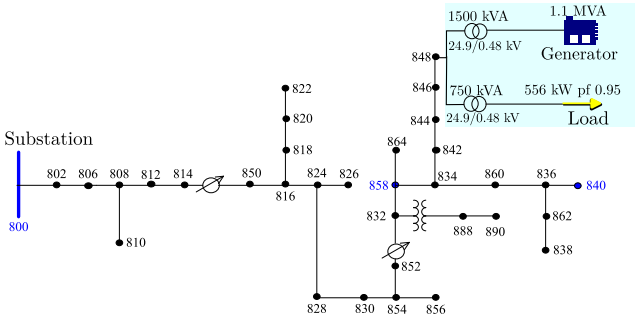


FIGURE 2. Modified IEEE 34-node test system.

of these faults often experience low voltages and currents. This characteristic enables the use of measurements from remote energy meters as key indicators for fault location. In our study, measurement points were selected at buses 800, 858 and 840 because these locations correspond to the 1st, 21st and 31st buses, respectively, when ranked by electrical distance from the system's substation. By choosing one bus near the substation, one in the mid-network region and one in the most remote area, we ensure that sensitivity and correlation analyses capture the full spatial extent of the distribution system. This approach allows us to evaluate fault location performance under near, intermediate and far conditions without adding measurements to every node.

The HIF simulations were performed using real HIF current signals obtained by the authors of [35]. These signals were included in the simulations according to the model shown in Appendix. Overall, 34 real HIF signals were separately inserted in each phase of these 33 buses of the system using a current source. As DSs usually have a signal-to-noise ratio between 50 and 60 dB [36], signals with no noise and 50 dB noise were evaluated. The aim is to evaluate the metrics considering realistic scenarios in terms of the presence of noise in DSs measurements.

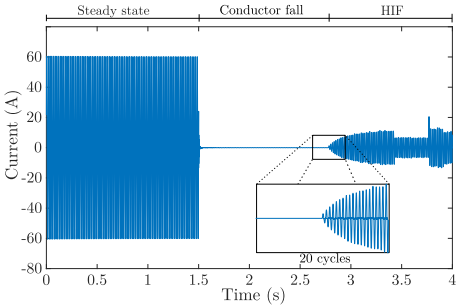


FIGURE 3. Example of the signal window used in the HIF location analysis.

B. INPUT METRICS EXTRACTION

In this study, conductor rupture followed by its contact with a high impedance surface is considered. The input metrics contain 20 cycles of the measured signals, 10-cycles pre-fault (during the conductor fall period), and 10 cycles after the HIF starts. This window size was chosen to highlight the changes in the measured signals before and after the HIF starts. One example is shown in Fig. 3 for a current signal measured at the fault spot (the current is zero when the conductor breaks until the HIF starts). We highlight that Fig. 3 was used for representation purposes to better explain the window extraction. For HIFs occurring downstream the meter location, the current may not be zero during the conductor fall period if there is load between the meter and the fault spot.

As HIFs can cause unbalance in test systems, in this study, we evaluate the signals measured at each phase, and also the sum of the three-phase signals. All parameters calculated from now on are obtained using the two types of input signals. Moreover, as HIFs imply non-linear characteristics on voltages and currents due to the electric arc, signal processing techniques can be used to extract features from them. In recent years, the Stockwell Transform (ST) has emerged as one of the most complete transforms used in fault diagnosis, as it is considered a combination of Fourier and Wavelet transforms. The ST extracts harmonics from the signals at each sample using a Gaussian window according to (1), and is minimally affected by background noise. Therefore, in this study, harmonics from the fundamental to the 20^{th} order were extracted by ST at each cycle of the signals acquired at each measurement spot. The low-order harmonics were evaluated because they can represent most HIF characteristics without requiring high sampling frequencies [37]. The ST can be calculated by:

$$ST\left[jT,\quad \frac{n}{NT}\right]=\sum_{m=0}^{N-1}H\left[\frac{m+n}{NT}\right]e^{-\frac{2\pi^2m^2}{n^2}}e^{-\frac{i2\pi mj}{N}}\quad n\neq 0 \tag{1}$$

in which j, m e $n = 0, 1, \ldots, N$-1 and the ST is calculated for each harmonic order n different from zero [38].

After extracting the harmonics, various parameters were calculated. The selected parameters included the module, angle, and energy of the current; the module, angle, and energy of the voltage; the active power; the impedance (voltage divided by current for each harmonic); and the power factor angle. These parameters were calculated for the 20 harmonic orders. They were chosen based on their potential to be influenced by the occurrence of HIFs in the system. Overall, the parameters were calculated at each cycle for each harmonic ($P^{cycle}_{harmonic}$), building the *Pmatrix* for each parameter:

$$Pmatrix=\begin{bmatrix} P^{first\ cycle}_{fund.} & \cdots & P^{last\ cycle}_{fund.} \\ \vdots & \ddots & \vdots \\ P^{first\ cycle}_{20^{th}\ order} & \cdots & p^{last\ cycle}_{20^{th}\ order} \end{bmatrix} \tag{2}$$

C. STATISTICS CALCULATION

After obtaining the parameters of each cycle, statistics were extracted from them to comprise the algorithms' dataset. Thus, the metrics were formed by statistics extracted from the parameters for each harmonic order $M^{statistic}_{harmonic}$. The *statistics* are the mean, maximum, standard deviation, variance, minimum, and amplitude (difference between the maximum and the minimum) of the 20-cycle parameters, and can be calculated by:

$$M^{statistic}_{harmonic}=statistic(\begin{bmatrix} P^{first\ cyle}_{harmonic} & \cdots & P^{last\ cyle}_{harmonic} \end{bmatrix}) \tag{3}$$

in which $harmonic = fund, 2^{nd}, \ldots, 20^{th} order$.

The analysis involved evaluating each metric individually and also considering the ratio between the metrics obtained from sparse measurements and those obtained at the substation. This ratio was used to normalize data collected at various locations with the data from the substation, providing a unique perspective on the input data. Nonetheless, this approach enabled the representation of changes occurring near sparse measurements. Both the specific metrics and the ratio metrics were utilized to train the AI-based algorithms.

Summarizing, the metrics were calculated according to the flowchart in Fig. 4. The analysis comprised two types of input signals (3-phase sum or faulted phase), 20 harmonic orders, 9 parameters of 3 specific measurements plus 2 regarding the ratio between the parameter calculated at a sparse measurement and at the substation, and 6 statistics, totaling 10,800 metrics. Each metric was evaluated for each fault signal (34 signals) at each bus (33 buses) of each system phase (3), totaling over 36,000,000 data.

In this work, over 10,800 candidate metrics were generated by combining signal types, harmonic orders, statistical descriptors and inter-measurement ratios. While a few metrics adapt traditional detection oriented measures for location, the vast majority are proposed here for the first time in the context of fault location. A systematic correlation-based selection process then distills these into a compact set of high-sensitivity indices, highlighting both the methodological novelty and the practical relevance of the proposed approach

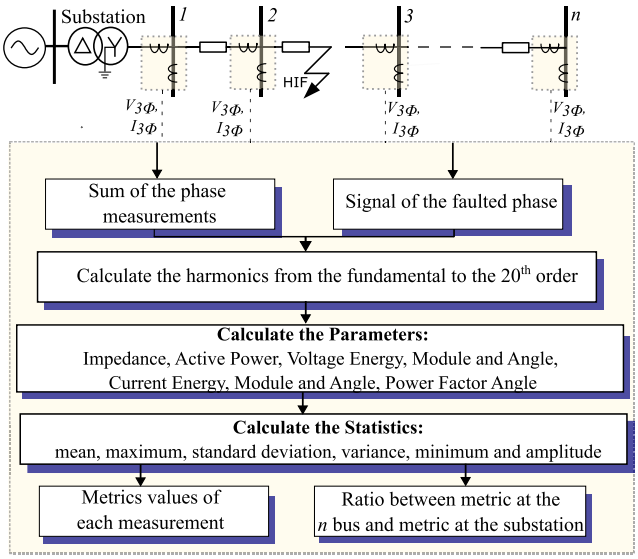


FIGURE 4. Flowchart of the process to obtain the metrics.

When developing fault location algorithms, it is important to verify if the selected metrics vary according to the fault location. In this study, this evaluation was performed using the PCC designed to evaluate the metrics correlation to the fault location. Pearson Correlation is a statistical measure that quantifies the strength and direction of the relationship between two variables [39]. It ranges from −1 (as one variable increases, the other decreases) to 1 (both variables increase together), with 0 indicating the absence of a linear correlation between the variables. The PCC is widely used for assessing the association between variables in statistics and data analysis and can be calculated according to (4) [39]. Therefore, the higher the PCC of a metric and the fault distance, the better it can represent the fault location when used as input on a fault location algorithm.

$$PCC = \frac{\sum (M_i - \bar{M})(d_i - \bar{d})}{\sqrt{\sum (M_i - \bar{M})^2 \sum (d_i - \bar{d})^2}} \tag{4}$$

where M_i and d_i are the individual values of the metrics and the fault distance for which that metric was obtained, respectively. $\bar{M}$ and $\bar{d}$ are their mean value.

After calculating the PCC amongst all metrics and fault distances, the metrics that achieved the highest correlation with the fault distance from the substation were selected as the best ones to evaluate the intelligent algorithms in the next phase of the present study.

D. INTELLIGENT ALGORITHM ANALYSIS

After choosing the best input metrics for the HIF location, the next step involved selecting intelligent algorithms suitable for decision-making, which was part of the proposed investigative methodology.

Intelligent algorithms are often used for fault location methods. These algorithms are mainly divided into supervised and unsupervised learning. The supervised algorithms are trained with a labeled dataset which associates the inputs to the desired outputs, whereas the unsupervised algorithms deal with unlabeled data to unveil patterns, structures, or relationships. Usually, supervised algorithms are more robust for hard-to-solve problems, and they were chosen for the present analysis.

Supervised learning is mainly comprised of regression algorithms, intended to estimate continuous values; and classification algorithms, which are focused on predicting a class instance. It also comprises deep learning methods, which use deep neural networks to extract and learn complex data representations, using both types of learning previously discussed. Among the three types, the Deep Learning algorithm requires a high processing rate. Therefore, regression-based algorithms were chosen for evaluation in this paper due to their vast applicability for fault location. Supervised learning algorithms, and the ones that combine supervised and unsupervised learning, are probably the best match for the HIF location problem.

There are many regression algorithms and some of them were chosen for a further performance comparison. The chosen algorithms are widely used for different purposes in the literature and are shown in the following section with a brief description and a list of their hyperparameters:

- **Artificial Neural Network (ANN):** It is a model inspired on the brain activity, comprised of interconnected neurons. ANNs can be used to learn complex and/or multidimensional relationships and are known as universal function approximators. Therefore, the ANNs are effective in regression tasks due to their ability to model non-linear relationships [40]. The ANNs used in this work are fully connected and feedforward neural networks. The hyperparameters are the size of fully connected layers and activation functions of the layers.
- **Decision Tree (DT):** They are structures that divide data according to their characteristics. Decision trees are interpretable and can handle non-linear relationships. A decision is made at each node, starting from the root node, using input feature vectors to form a binary tree. This binary tree grows by continually branching with the decisions of each node and ends by providing the predicted output [41], [42]. The hyperparameters are the splitting criterion and maximum tree depth.
- **Ensemble Tree (ET):** It combines multiple decision trees to improve predictive performance. Methods as Random Forest and Gradient Boosting are used to mitigate overfitting and capture more robust patterns

for complex and non-linear regression problems [42], [43]. The hyperparameters are the ensemble method (adaboost or bagging), number of trees, maximum tree depth, learning rate.

- **Support Vector Machine (SVM):** It finds an optimal separation hyperplane between classes to maximize the distance between any categories. In regression, SVM is adapted to estimate the regression function by mapping data to a feature space where linear separation is performed [42]. According to the results in [24] and [44], the SVM has several advantages that make it a suitable choice for the fault location problem. The hyperparameters are the Kernel type, Kernel parameters, and box constraints.
- **Gaussian Kernel Regression (GKR):** It utilizes kernel functions to estimate the relationship between variables, including non-linear relationships. A kernel function maps data from a low-dimensional space to a high-dimensional space, and then fits a linear model in the high-dimensional space by minimizing the regularized objective function [45]. The hyperparameters are the kernel type and kernel parameters.
- **Efficient Linear Regression (ELR):** It is a computationally efficient approach suitable for large datasets, offering a quick and scalable solution for linear regression problems [42]. These models are trained with high-dimensional predictor data, whether complete or sparse [45]. The hyperparameters are the learning model, regularization, and gradient tolerance.

As mentioned, all selected algorithms require the hyperparameters setting before using them. However, the empirical selection of hyperparameters can imply errors. As a solution, a Bayesian optimizer was employed to automate the hyperparameter tuning process, aiming to avoid a purely empirical approach and improve the model's accuracy. Additionally, it allows efficient exploration, speeding up the algorithm's convergence toward the optimal configuration [43], [45]. The Bayesian optimizer aims to minimize a scalar objective function in a bounded domain. In the context of this study, the hyperparameter optimization was based on the minimization of the models' distance estimation error [43], [45].

After determining the hyperparameters, the dataset was split between training and testing. Subsequently, the algorithms were trained, and the models were used to make predictions of the HIFs location using the test data. Their performance were compared using error-related indices calculated as follows:

1) **Mean Average Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

It represents the average deviations between predicted ($\hat{y}_i$) and actual (y_i) values. A MAE close to zero indicates a good fit, with predictions close to actual values;

2) **Mean Percentage Error (MPE):**

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_t}\right) \times 100$$

This expresses the average error as a percentage of the total system length y_t (59 km for the used test system);

3) **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

This provides a measure of the average deviation weighted by the magnitude of the deviations. A low RMSE indicates predictions close to the actual values, suggesting higher accuracy. It accounts large errors more than small errors due to the squaring operation.

The error indices were compared and shown in the next section, enabling the determination of the most appropriate HIF location algorithm.

III. METRICS AND INTELLIGENT ALGORITHMS EVALUATION

This section presents the results of the practical implementation methodology to investigate optimal metrics and AI-driven decision-making techniques for locating HIFs, amongst the ones selected in this paper. The analysis in this section consider the system loading and DG power at nominal rates. After obtaining the best metrics and intelligent algorithms, the tools for the proposed HIFLM are finally determined.

A. METRICS ASSESSMENT

The analysis in this paper encompassed two types of input signals (three-phase sum and faulted phase), 20 harmonic orders, nine direct parameters measured at three strategically spaced buses, two inter-measurement ratios, and six statistical descriptors—totaling 10,800 candidate metrics per phase. We then computed the Pearson correlation coefficient (PCC) of each metric against the electrical distance from the substation across 34 fault events, 33 buses, and three phases (over 36 million data points). Owing to space constraints, Table 2 reports only those metrics whose absolute PCC exceeded 0.9, indicating a very strong relationship to fault location. Notably, only the ratio-based metrics achieved PCC > 0.9; no individual direct-measurement metric surpassed this threshold.

Table 2 shows that the best parameters were the energy and module of the voltage harmonics. The best results achieved by metrics derived from voltage can be seen as an insight from the analysis conducted, as current signals are typically used to verify the effects of HIFs. Good results were achieved with both three-phase sum and considering only the faulty phase measurement. The best harmonics were between the 8^{th} and 20^{th} orders with different statistics. The highest PCC achieved by an

input combination was 0.93. All top performing metrics correspond to the ratio of the measurement at bus 858 to the substation measurement. By normalizing the remote bus value, which varies significantly with the proximity of the fault, compared to the relatively stable substation value, these ratio metrics suppress the effects of common modes and amplify the distance-dependent deviations, producing the highest Pearson correlation coefficients (> 0.9).

TABLE 2. Input combination that achieved a pearson correlation coefficient above 0.9.

Metric	Measurement	Harmonic	Statistic	Coefficient
Power	3-phase sum	8	max	0.91
V Energy	3-phase sum	8	var	0.90
V Energy	3-phase sum	9	var	0.90
V Energy	3-phase sum	12	mean, max, std, amp	0.90
V Energy	3-phase sum	13	mean, max, std, amp	0.91
V Energy	3-phase sum	14	mean, max, std, amp	0.92
V Energy	3-phase sum	15	mean, max, std, amp	0.91
V Energy	3-phase sum	16	max, std, amp	0.91
V Energy	Faulty Phase	16	mean, max, std, amp	0.91
V Energy	Faulty Phase	17	mean, max, std, amp	0.91
V Energy	Faulty Phase	18	mean, max, std, amp	0.92
V Energy	Faulty Phase	19	mean, max, std, amp	0.91
V Energy	Faulty Phase	20	mean, max, std, amp	0.91
V Module	3-phase sum	12	var	0.91
V Module	3-phase sum	13	max var amp	0.91
V Module	3-phase sum	14	mean, max, std, var, amp	0.92
V Module	3-phase sum	15	mean, max, std, var, amp	0.92
V Module	3-phase sum	16	mean, max, std, amp	0.93
V Module	3-phase sum	17	mean, max, std, amp	0.92
V Module	3-phase sum	18	mean, max, std, amp	0.92
V Module	3-phase sum	19	max, std, amp	0.92
V Module	3-phase sum	20	mean, max, std, amp	0.92
V Module	Faulty Phase	16	var	0.91
V Module	Faulty Phase	17	var	0.92
V Module	Faulty Phase	18	mean, max, std, var, amp	0.91
V Module	Faulty Phase	19	mean, max, std, var, amp	0.91
V Module	Faulty Phase	20	mean, max, std, var, amp	0.92

- : All the best metrics are the ones obtained by the ratio between the measurement at the substation and the measurement at bus 858 /V: Voltage/ max: maximum/ min:minimum/ std: standard deviation/ amp: amplitude/ var: variance.

Table 2 also shows that different statistics and harmonics were repeatedly mentioned among the best PCCs. To summarize this, Fig. 5a highlights the statistics for PCCs above 0.9, considering noisy and non-noisy input signals. It shows that the mean, maximum, standard deviation, and amplitude of the parameters were the best statistics. However, when considering the input signals with a 50 dB noise, the mean value did not imply a correlation above 0.9, which can hinder the fault location using the mean value of the metrics in real conditions. The metrics' variance and minimum values did not have a high correlation with the fault distance. Moreover, Fig. 5b presents an analysis of the harmonics in the best PCCs' list. It reveals that harmonics from the fundamental to the 7^{th} order had a low correlation to the fault distance. Harmonics from the 12^{th} to the 20^{th} order had the best PCCs, but the number of inputs above the 17^{th} harmonic order decreased, when noisy signals were considered.

The voltage energy and module exhibited a strong correlation with the fault distance. These metrics were computed using either the faulty phase or the sum of the voltage signals from all three phases. Furthermore, the calculations considered the ratio between measurements from a sparse location and the substation. The most relevant inputs are centered around harmonics ranging between the 12^{th} and 16^{th} orders, encompassing statistics such as maximum values, standard deviation, and amplitude, totalling 15 metrics. In Fig. 6, these metrics are numbered from 1 to 15, corresponding to the amplitude, maximum value, and standard deviation for each of the 12^{th} to 16^{th} harmonics of the three-phase voltage sum. From Fig. 6, the metric value increase with the fault distance. Additionally, each metric is affected by the distance with varying intensity. Therefore, the machine learning algorithm is expected to combine these metrics to estimate the fault location.

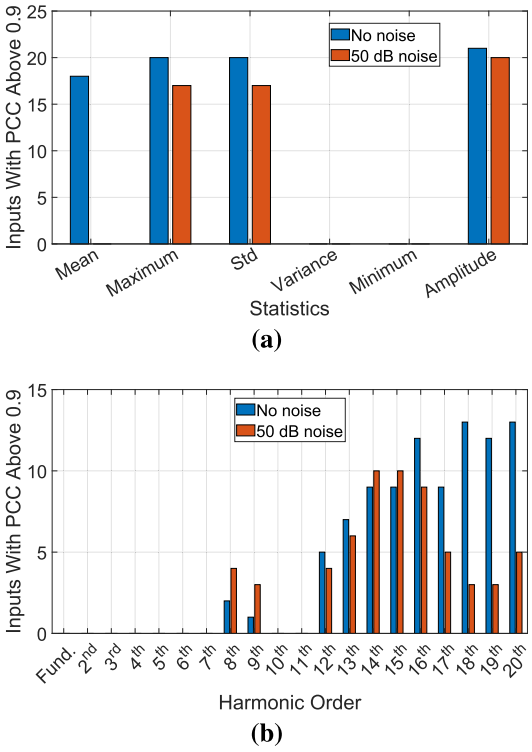


FIGURE 5. Number of metrics with PCC above 0.9 with and without noise considering the (a) Statistics and (b) Harmonics.

Overall, as the mentioned metrics had a high correlation with the fault location, they can be used as input to decision methods. It is important to highlight that the PCC is a simple and widely used technique. Thus, if the metrics had a high linear correlation with the fault location, there is no need to use more sophisticated correlation techniques.

B. EVALUATION OF THE INTELLIGENT ALGORITHMS

After determining the best metrics, they were used as inputs for the selected intelligent algorithms. The study in

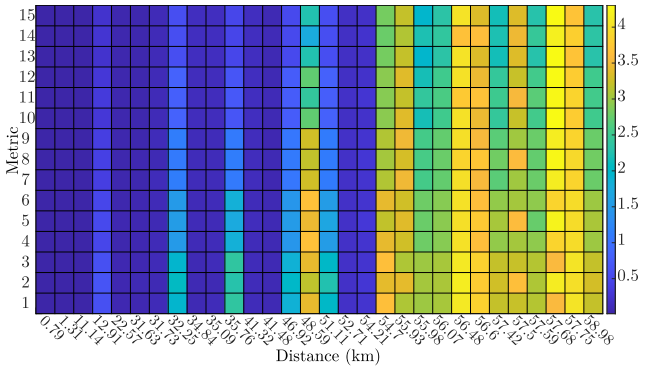


FIGURE 6. Metrics values for different HIF locations.

the previous section revealed that using the faulty phase voltage and the sum of the three-phase voltages as input signals for HIF location resulted in satisfactory outcomes. However, when implementing an HIFLM, using the faulty phase voltages requires a fault phase classification algorithm, which could introduce errors to the HIF location method. For this reason, the sum of the three-phase voltages was used as the input signal. Moreover, the metrics' evaluation revealed that both the energy and module of the input signal's harmonics are suitable for HIF location. In this case, the voltage harmonic's module was selected due to its lower processing requirements.

The final *DataSet* contains harmonics from the 12^{th} to the 16^{th} order, and the evaluated statistics were amplitude, standard deviation, and maximum value, as they showed promising results regardless of the noise level. These metrics were determined using the ratio between the parameters calculated on bus 858 (located between the substation and the DG) and those calculated at the substation. Summarizing, Table 3 presents the best metrics used to develop the *DataSet*. They can be organized in a matrix, where the column of the *DataSet* contains a metric $M^{harmonic}_{statistic}$, and the last one is the fault distance in the condition the metrics were obtained. Thus, each row has the metrics for an observation (an occurrence of a fault in a specific distance), as shown in (5). In general, the number of columns is fixed, and the number of rows depends on the number of n cases simulated.

$$DataSet = \begin{bmatrix} M^{12h}_{amp} & M^{12h}_{max} & M^{12h}_{std} & \cdots & M^{16h}_{std} & Dist.\,1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ M^{12h}_{amp} & M^{12h}_{max} & M^{12h}_{std} & \cdots & M^{16h}_{std} & Dist.\,n \end{bmatrix} \tag{5}$$

Thus, the final *DataSet* was formed by 34 HIF signals at 33 different locations in the test system across the 3 phases. Thus, there were 3,366 observations for each noise level with 15 metrics each. All algorithms were implemented using the Machine Learning Library of the Matlab software. Moreover, 80% of the data was randomly selected for training and 20% for testing. During training, the methods were evaluated using the k-fold method, dividing the dataset into five parts to avoid overfitting, and the models' hyperparameters were determined through Bayesian optimization.

TABLE 3. Best metrics for HIF location.

Electrical Quantity	Voltage
Signal	Three-phase sum
Harmonics	12^{th} to 16^{th} order
Parameter	Harmonic's Module
Statistics	Amplitude, maximum and standard deviation
Consideration	Ratio between a sparse measurement and substation

Table 4 shows the results of error indices calculated for the test data of each of the trained and tested models with and without noise. Overall, the results show that training the models with the expected noise level in the system can contribute to their operation. It implies that when considering the test data with 50 dB noise, training with 50 dB noise implies better results than training with non-noisy data. Overall, the lowest errors achieved by the models are highlighted in bold in Table 4.

Table 4 also shows that the algorithm with the highest error rate was the ELR. This occurred because the fault distance in the evaluated system did not have a completely linear relationship with the metrics. On the other hand, the Ensemble Trees algorithm resulted in low errors. The MAE index, which represents the estimated average error, showed that the ET algorithm resulted in a maximum error of 2.8 km when only the noise-free signal was used in training, but the test contained noisy signals. However, when the noisy signal was used in training, the error dropped to 1.9 km. This was more evident when evaluating the MPE metric, which deals with the error percentage relative to the total size of the test system (59 km). The MPE showed a maximum error of 4.7% of the system size considering a critical condition (model training with no noise and tested with noise), contrasting with linear regression, with an error of 13.24% of the system in the same condition. Finally, the RMSE index, which highlighted the worst errors, also demonstrated the superiority of the ET model, proving that these methods also resulted in fewer outliers, indicating high accuracy in comparison to the other methods.

In order to better analyze the algorithms, Fig. 7 provides a comparison among the models that resulted in the best and worst cases. Fig. 7 illustrates the real fault distances and the average distance estimated by the models considering the different HIF signals applied at the same location. Thus, despite some occasional errors, the commendable performance of ET was evident across all regions of the system.

In order to better analyze the results obtained by the ET method, Fig. 8 presents the standard deviation of the

TABLE 4. Comparison of the ML models results.

Training	No noise						50 dB noise					
Testing	No noise			50 dB			No noise			50 dB		
Error	MAE (km)	MPE (%)	RMSE (km)	MAE (km)	MPE (%)	RMSE (km)	MAE (km)	MPE (%)	RMSE (km)	MAE (km)	MPE (%)	RMSE (km)
DT	1.02	1.73	2.70	3.07	5.20	7.12	1.45	2.45	3.60	2.04	3.46	4.65
ANN	1.29	2.19	2.87	2.63	4.47	5.57	1.48	2.51	2.64	2.06	3.50	3.96
ELR	8.93	15.13	13.71	7.88	13.36	12.30	8.86	15.01	13.55	7.81	13.24	12.26
ET	**0.96**	**1.63**	**2.45**	2.81	4.76	6.06	1.33	2.25	2.91	1.98	3.36	4.04
GKR	3.48	5.89	5.60	5.34	9.06	8.39	4.56	7.73	6.47	5.02	8.50	7.49
SVM	1.78	3.02	3.79	3.36	5.69	6.75	1.81	3.07	3.23	2.81	4.76	5.24

DT: Decision Tree, ANN: Artificial Neural Networks, ELR: Efficient Linear Regression, ET: Ensemble Tree, GKR: Gaussian Kernel Regression, SVM: Support Vector Machine.

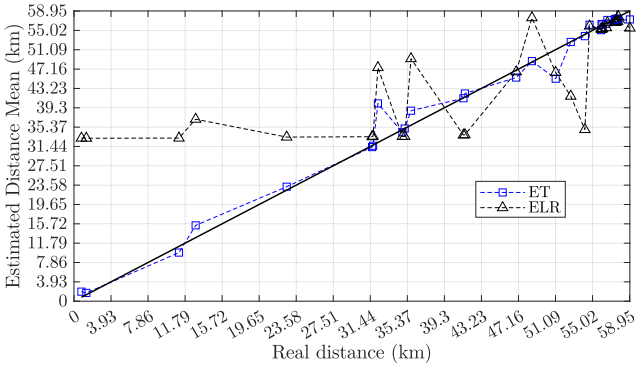


FIGURE 7. Mean of estimated values.

estimated distances considering different real HIF signals at the same location.

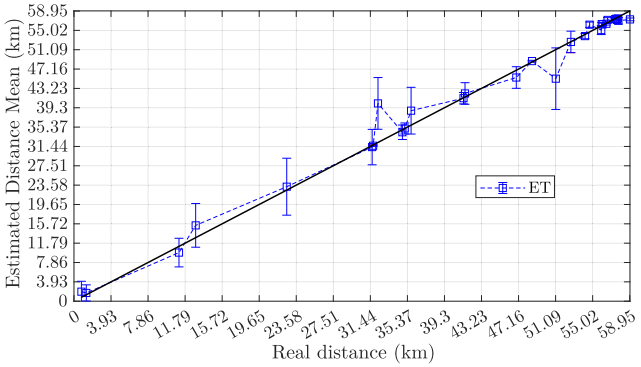


FIGURE 8. Standard deviation of the estimated distances.

Fig. 8 illustrates that the highest standard deviation occurred for the ET method when the fault was approximately 31.44 km far from the substation. This is because the test system contains lateral branches that increase the difficulty to the HIF location at this distance. In contrast, the ET showed a low standard deviation for most of the fault locations analyzed.

Moreover, Fig. 9 shows the distances estimated by the Ensemble Tree for the worst-case scenario when the models were trained with non-noisy data and tested with noisy data. It can be inferred that the method's highest error occurred when the fault was close to the substation. Additionally, it is easier for power utilities maintenance crews to locate the fault when it occurs near the substation, as a large portion of the system loads are disconnected from the feeder. However, faults that occur many kilometers away pose a greater difficulty, and the ET resulted in high accuracy in these circumstances.

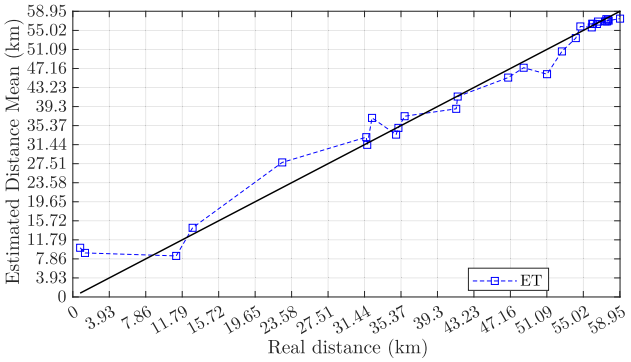


FIGURE 9. Mean of estimated values when the models are trained with non-noisy data and tested with noisy data.

Fig. 9 shows that there was no considerable performance degradation of the ET algorithm and it still estimated reasonable values in this worst-case scenario. Finally, the noise impact in the estimations closer to the substation is noticeable. Considering this, the ET is the decision method chosen for the final algorithm of this paper.

IV. PROPOSED HIF LOCATION METHOD

Based on the methodology developed in the previous sections, this article provided a step-by-step guide for developing a new High Impedance Fault Location Method (HIFLM). Overall, this methodology can be replicated by other researchers for the development of new tools. Specifically, in this work, the analysis was conducted to ensure that the proposed HIFLM is grounded in a comprehensive study, supported by a rigorous analysis.

The results presented in Section III, led to the proposition of a new HIFLM, composed by the metrics in Table 3 and the ET as the intelligent algorithm.The pseudocode presented in Algorithm 1 outlines the step-by-step process of the fault location method proposed in this study. It begins with voltage

signal measurements across the three phases of the system at the substation and at a midpoint bus within the system. Next, the magnitudes of harmonics from the 12th to the 16th order are calculated for each cycle within a 20-cycle window. The amplitude, maximum value, and standard deviation of each harmonic order are then computed to build the dataset. This dataset is subsequently split, with a percentage of the data randomly assigned for training using the ensemble tree algorithm, with hyperparameters determined via Bayesian optimization. Finally, the algorithm's performance is evaluated using the test dataset with the same decision method.

Algorithm 1 Proposed HIF Location Method

Input Variables:
Sum of three-phase voltage { $V_{in} = V_a + V_b + V_c$}

Metrics Extraction:

Data Processing: Obtain the module of harmonics from 12^{th} to 16^{th} order;

Metrics: Calculate the amplitude, maximum and standard deviation of each harmonic in a 20-cycle window

Dataset: Calculate the ratio between the metrics calculated in the middle of the system and the metrics obtained at the substation

Decision Method:

Model a test system and obtain the dataset. Separate it between training and testing.

Train the Ensemble Tree algorithm with hyperparameters determined by the Bayesian Optimization algorithm

Test the Ensemble tree algorithm with the testing data and evaluate its performance.

Generalization test for comprehensive performance analysis.

To assess computational efficiency, Table 5 reports the execution time of each stage of the proposed algorithm, measured in MATLAB on a Windows-based workstation with AMD Ryzen 9 5900X 12 and Core Processor of 3.70 GHz. The results demonstrate that the overall runtime is modest and well within practical limits. Moreover, performance can be further improved by parallelizing the meter-wise data processing or by porting the implementation to more performance-oriented environments—such as optimized Python libraries or a C++ implementation—thereby reducing computational overhead even further.

A. COMPARISON WITH EXISTING METHODS

To demonstrate the contributions of the proposed HIFLM, Table 6 compares the results achieved using the selected metrics and recent HIF location methods in the literature. It is important to highlight that the aim of this paper is not only to propose a new algorithm, but mainly to provide a complete methodology to use effective metrics and models to develop robust new methods for HIF location. Consequently, the main goal of this section is to show that the critical evaluation of metrics and intelligent algorithms used to compose the proposed method can imply contributions to the state-of-the-art and future research, considering the rigor of the tests

TABLE 5. Execution time of each step of the algorithm.

Operation	Time (s)
Data loading	0.3650
Window extraction	0.0009
Harmonic extraction	4.5966
Statistics calculation	0.0036
Feature matrix construction	0.0004
Trained Model loading	0.2096
Classification	0.1907

performed. Moreover, only results obtained by the authors were considered, as relying on external implementations could introduce inaccuracies due to missing details, potentially compromising the validity of the comparison.

Therefore, Table 6 shows a comparison among different papers on HIF location. It highlights the DS model utilized to test the methodology, its total length in km, if the study considered background noise and DG in the system, the HIF model used in the analysis and the MPE (average error in relation to the total system length) for each method presented. Notice that two of the methods, [3] and [13], were tested in the same test system of this paper's method, allowing a fair comparison.

Table 6 reveals that although the existing solutions published in [3], [4], [13], and [22] present low errors, they were not tested in systems with DG. For example, the method in [13] presented a lower MPE than the evaluated methods, but it was tested with modeled HIF signals and in a system without DG. The method in [22] implied a 0.87 MPE, but it did not consider real HIF signals or the presence of DG, and it used a shorter system to evaluate the method when compared to the present study. The method in [4] was tested in a system with a DG and it displayed a 9.54% error, compared with 1.63% error the proposed HIFLM, based on ET. Moreover, the existing methods modeled the HIFs using simple models that do not represent all the randomness in this type of fault. In contrast, the proposed approach was evaluated using real HIF currents. Table 6 also highlights that some algorithms were evaluated in small systems, which facilitates a low error rate.

TABLE 6. Comparison of the MPE of existing HIF location methods and the solutions assessed in this paper.

Ref.	DS Model	Length (km)	Noise	DG	HIF Model	MPE (%)
[3]	IEEE34	59.0	✓	✗	Diode based	2.38
[13]	13bus/IEEE34	15/59	✓	✗	Diode based	1.08/1.25
[22]	IEEE13	1.524	✓	✗	Diode based	0.87
[4]	23 Bus Syst.	4.167	✓	✓	Diode based	9.54
proposed (ET)	IEEE34	59	✓	✓	Real signals	1.63

B. GENERALIZATION ANALYSIS

In the literature, HIFLMs are typically tested using a fixed dataset, divided into training and test groups. However, this approach provides a limited perspective on the generalization capacity of the method. Since training and test groups are

randomly sorted, they may share scenarios with similar characteristics, potentially inflating performance metrics. Therefore, conducting a more comprehensive generalization test using a dataset containing distinct operating scenarios from those in the initial dataset is recommended. This section presents a generalization analysis considering variations in scenarios commonly encountered in DSs with DG. This analysis is particularly critical because the final method proposed in this paper relies on data-driven algorithms, which inherently depend on the quality and diversity of the data used during training.

For the generalization analysis, the same test system was used. However, we controlled the DG power to vary among 100%, 66% and 33% of its rated power. Additionally, the loads - modelled with constant impedance - were changed, implying on a 100% and 30% of the nominal system loading. While the measurement was kept at the system substation, the sampling frequency used to acquire the signals was changed, in order to verify its impact on the algorithm. Table 7 shows the scenarios division for the generalization analysis, totalling 36 combinations.

TABLE 7. Division of train and test scenarios for the generalization analysis.

	Scenarios	Types
Train	DG power	100%, 66% and 33%
	System Loading	100% and 30%
Test	DG power	100%, 66% and 33%
	System Loading	100% e 30%
	Total	**36 combinations**

It is important to highlight that, in addition to the analysis of training and testing across the combinations, additional evaluations were conducted considering three different sampling frequencies, and considering or not noise in the measurement. In the analysis, the data was randomly divided between training (80%) and testing (20%). Therefore, even when the algorithm is trained and tested on the same scenario, they have different HIF signals.

This proposed method was evaluated using the different scenarios combinations in the modeled test system. The results regarding the error in kilometers from the substation, the mean percentage error considering the whole system and the R^2 are summarized in Tables 8 and 9 for input signals without and with noise, respectively. It is important to highlight that these results were obtained using only one measurement device at the system substation, with HIFs occurring along with the whole system.

The generalization results show that when the training and testing occurred within the same scenario, the lowest error rates were observed. For instance, in scenarios with nominal DG penetration and loading conditions, the average error was 0.88 km for noise-free signals and 1.99 km for noisy signals when sparse meters were used—only 1.49% and 3.3% of the system's total length, respectively. Table 6 has already demonstrated that these normalized errors are lower than those reported by other state-of-the-art HIF-location methods on comparable network sizes. However, when training and testing were conducted under different scenarios, the performance degraded, highlighting the intrinsic difficulty of high-impedance fault location, a task fundamentally more challenging than conventional short-circuit location and often simplified in the literature by placing meters adjacent to the fault. The exhaustive evaluation on the real IEEE 34-bus feeder—with lateral branches, actual HIF waveforms, and distributed generation penetration confirms the method's robustness and practical applicability. Incorporating noise into the training phase proved crucial for improving success under varied conditions. For applications requiring tighter accuracy, we recommend extending the training dataset to include the specific operating scenarios expected in practice. Regarding sampling frequency variations, higher sampling

TABLE 8. Generalization tests of the proposed method without noise for different training and test DG power (DGP), system loading and sampling frequencies.

Samples per cycle				256			128			64		
DGP Train	Load. Train	DGP Test	Load. Test	Error (km)	MPE (%)	R^2	Error (km)	MPE (%)	R^2	Error (km)	MPE (%)	R^2
100%	100%	100%	100%	0.88	1.49	0.98	1.23	2.08	0.97	2.09	3.54	0.94
100%	100%	100%	30%	3.62	6.13	0.85	3.95	6.69	0.83	4.48	7.60	0.82
100%	100%	66%	100%	0.70	1.18	0.99	0.84	1.42	0.98	1.24	2.10	0.98
100%	100%	66%	30%	4.37	7.40	0.77	4.33	7.34	0.80	4.57	7.74	0.81
100%	100%	33%	100%	0.69	1.17	0.99	1.07	1.82	0.98	1.43	2.43	0.97
100%	100%	33%	30%	4.31	7.30	0.78	3.77	6.39	0.87	4.47	7.57	0.81
100%	30%	100%	100%	3.61	6.11	0.84	4.25	7.20	0.82	4.52	7.66	0.82
100%	30%	100%	30%	1.32	2.24	0.97	1.70	2.87	0.94	3.05	5.17	0.90
100%	30%	66%	100%	3.25	5.50	0.89	4.11	6.97	0.83	4.42	7.49	0.82
100%	30%	66%	30%	0.91	1.54	0.98	0.97	1.64	0.98	1.87	3.16	0.96
100%	30%	33%	100%	3.40	5.77	0.87	4.43	7.50	0.81	4.74	8,03	0.82
100%	30%	33%	30%	0.81	1.38	0.99	1.09	1.84	0.97	1.85	3.13	0.96
66%	100%	100%	100%	0.61	1.04	0.99	0.89	1.50	0.98	1.34	2.28	0.97
66%	100%	100%	30%	3.64	6.18	0.84	3.97	6.73	0.82	4.56	7.73	0.80
66%	100%	66%	100%	1.15	1.95	0.97	1.26	2.14	0.96	2.01	3.40	0.93
66%	100%	66%	30%	4.35	7.37	0.77	4.29	7.27	0.80	4.69	7.95	0.79
66%	100%	33%	100%	0.73	1.24	0.99	1.00	1.69	0.98	1.26	2.14	0.98
66%	100%	33%	30%	4.25	7.20	0.78	3.72	6.31	0.87	4.58	7.76	0.79
66%	30%	100%	100%	3.61	6.12	0.84	3.99	6.75	0.84	4.47	7.57	0.83
66%	30%	100%	30%	0.99	1.68	0.98	0.96	1.62	0.98	1.83	3.10	0.96
66%	30%	66%	100%	3.24	5.50	0.89	3.81	6.46	0.86	4.42	7.49	0.83
66%	30%	66%	30%	1.37	2.31	0.96	1.65	2.79	0.94	3.04	5.15	0.90
66%	30%	33%	100%	3.42	5.79	0.87	4.18	7.09	0.82	4.69	7.94	0.83
66%	30%	33%	30%	0.84	1.43	0.99	1.03	1.74	0.98	1.81	3.06	0.96
33%	100%	100%	100%	0.70	1.18	0.99	1.07	1.81	0.98	1.77	3.00	0.96
33%	100%	100%	30%	3.69	6.26	0.85	4.24	7.19	0.81	4.84	8,20	0.80
33%	100%	66%	100%	0.89	1.51	0.98	1.03	1.75	0.98	1.48	2.52	0.97
33%	100%	66%	30%	4.50	7.63	0.77	4.58	7.76	0.78	4.93	8,36	0.79
33%	100%	33%	100%	0.89	1.52	0.98	1.08	1.82	0.96	1.76	2.98	0.95
33%	100%	33%	30%	4.35	7.37	0.79	4.08	6.91	0.85	4.76	8,06	0.79
33%	30%	100%	100%	3.64	6.16	0.84	3.89	6.60	0.85	4.51	7.65	0.82
33%	30%	100%	30%	0.81	1.37	0.99	1.06	1.79	0.97	1.82	3.09	0.96
33%	30%	66%	100%	3.32	5.63	0.88	3.79	6.43	0.86	4.52	7.66	0.81
33%	30%	66%	30%	0.80	1.35	0.99	1.07	1.81	0.97	1.85	3.14	0.96
33%	30%	33%	100%	3.43	5.82	0.87	4.11	6.96	0.84	4.74	8,04	0.82
33%	30%	33%	30%	1.25	2.11	0.97	1.63	2.76	0.94	3.07	5.20	0.89

frequencies consistently led to greater accuracy. Nonetheless, if data processing constraints are a critical consideration, the results showed that reducing the sampling rate from 256 samples per cycle to 64 samples per cycle increased the error by a maximum of 1.5 km in the worst-case scenarios. Thus, lower sampling rates may still be a viable solution depending on the application requirements.

In general, the results show the importance to train the algorithm for all the scenarios expected to occur in DSs (train and test in similar scenarios) to improve the results. However, not considering it does not hinder using the proposed

TABLE 9. Generalization tests of the proposed method with noise for different training and test DG power (DGP), system loading and sampling frequencies.

Samples per cycle				256			128			64		
DGP Train	Load. Train	DGP Test	Load. Test	Error (km)	MPE (%)	R^2	Error (km)	MPE (%)	R^2	Error (km)	MPE (%)	R^2
100%	100%	100%	100%	1.99	3.38	0.93	2.08	3.53	0.90	3.45	5.85	0.85
100%	100%	100%	30%	4.45	7.55	0.77	4.55	7.71	0.73	5.13	8,70	0.76
100%	100%	66%	100%	1.83	3.11	0.93	2.28	3.86	0.92	2.70	4.58	0.91
100%	100%	66%	30%	5.00	8,48	0.70	4.37	7.40	0.77	5.57	9.45	0.69
100%	100%	33%	100%	2.30	3.91	0.92	2.61	4.43	0.91	3.66	6.21	0.83
100%	100%	33%	30%	4.47	7.58	0.74	4.28	7.26	0.79	5.20	8,81	0.77
100%	30%	100%	100%	4.01	6.79	0.82	4.34	7.36	0.77	5.15	8,72	0.75
100%	30%	100%	30%	1.86	3.15	0.94	2.32	3.94	0.90	3.29	5.57	0.89
100%	30%	66%	100%	4.52	7.66	0.75	4.54	7.69	0.79	5.02	8,51	0.77
100%	30%	66%	30%	1.96	3.32	0.91	2.03	3.44	0.90	3.03	5.14	0.88
100%	30%	33%	100%	4.90	8,31	0.73	4.87	8,25	0.76	5.60	9.50	0.69
100%	30%	33%	30%	1.67	2.83	0.94	2.31	3.91	0.89	3.11	5.27	0.90
66%	100%	100%	100%	1.87	3.16	0.94	2.08	3.52	0.91	3.06	5.18	0.88
66%	100%	100%	30%	4.45	7.54	0.77	4.60	7.79	0.72	5.18	8,78	0.77
66%	100%	66%	100%	1.97	3.34	0.93	2.47	4.19	0.91	3.08	5.22	0.88
66%	100%	66%	30%	4.97	8,43	0.71	4.35	7.37	0.77	5.46	9.25	0.72
66%	100%	33%	100%	2.31	3.91	0.92	2.51	4.26	0.92	3.48	5.90	0.85
66%	100%	33%	30%	4.54	7.69	0.74	4.37	7.40	0.78	5.36	9.09	0.77
66%	30%	100%	100%	4.06	6.88	0.82	4.27	7.24	0.79	4.67	7.91	0.78
66%	30%	100%	30%	1.68	2.85	0.94	2.03	3.45	0.91	2.64	4.48	0.92
66%	30%	66%	100%	4.62	7.83	0.75	4.28	7.25	0.82	4.52	7.66	0.79
66%	30%	66%	30%	2.34	3.96	0.90	2.25	3.81	0.91	3.77	6.40	0.85
66%	30%	33%	100%	4.53	7.68	0.78	4.75	8,04	0.80	4.88	8,27	0.75
66%	30%	33%	30%	1.71	2.90	0.94	2.31	3.92	0.90	3.12	5.29	0.90
33%	100%	100%	100%	2.18	3.69	0.92	2.28	3.86	0.92	3.62	6.14	0.84
33%	100%	100%	30%	4.37	7.41	0.78	4.71	7.98	0.72	5.54	9.40	0.74
33%	100%	66%	100%	1.83	3.10	0.94	2.64	4.47	0.90	3.24	5.50	0.88
33%	100%	66%	30%	4.89	8,28	0.72	4.38	7.42	0.78	5.90	10,0	0.69
33%	100%	33%	100%	2.04	3.46	0.93	2.31	3.91	0.93	3.01	5.10	0.88
33%	100%	33%	30%	4.38	7.42	0.75	4.34	7.35	0.79	5.64	9.55	0.75
33%	30%	100%	100%	4.24	7.18	0.82	4.12	6.99	0.81	5.04	8,54	0.77
33%	30%	100%	30%	1.58	2.67	0.96	2.00	3.40	0.93	3.03	5.14	0.91
33%	30%	66%	100%	4.75	8,05	0.74	4.64	7.87	0.81	4.77	8,09	0.80
33%	30%	66%	30%	1.95	3.31	0.93	2.04	3.46	0.92	3.21	5.44	0.88
33%	30%	33%	100%	4.74	8,03	0.77	4.94	8,37	0.78	5.50	9.33	0.71
33%	30%	33%	30%	1.83	3.11	0.95	2.53	4.29	0.89	3.95	6.70	0.86

methodology. Nonetheless, a previous analysis regarding the common DG power, loading and measurement noise, can improve AI-based algorithms' performance.

The sensitivity of fault-location metrics to varying load levels is a critical concern for practical deployment. Because fault location is normally performed offline, the proposed method can be trained under nominal generation and loading conditions and then evaluated in alternate scenarios. While the highest accuracy occurs when training and testing conditions match, the minimal performance loss under altered loading demonstrates the method's robustness and applicability across realistic operating scenarios. Overall, the results obtained through the presented approach demonstrate that the proposed algorithm exhibits high applicability, making it a robust solution for fault location challenges. It is important to highlight that existing methodologies in the literature do not present this type of generalization analysis, hindering a comparison.

V. CONCLUSION

This paper presented a comprehensive study by analyzing relevant metrics and intelligent algorithms for HIF location in DSs with DG. The step-by-step evaluation enables researchers to the other attributes, in different systems, using various algorithms to achieve a novel and effective selection.

By performing this evaluation, the study results revealed a strong correlation between the maximum, standard deviation, and amplitude of the energy and module of harmonics between the 12^{th} and 16^{th} order extracted from the voltage signals. They were obtained by the ratio between the metric at a sparse measurement and the metric at the substation. These are some of the novel insights obtaining by performing this study. After selecting them, they were used as input for different intelligent regression algorithms.

Most existing HIF location methodologies select one machine-learning algorithm without presenting a thorough analysis of that choice. Considering the HIF's unique characteristics, the present study underscores the importance of carefully selecting a machine learning model for HIF location algorithms. Thus, a baseline database and an optimization tool were used to set each algorithm's hyperparameters, making it possible to determine which ones were more suitable for the problem. The analysis showed the superiority of the ET model when evaluating the error indices. Compared to existing solutions, it was shown that the present evaluation analysis can contribute to future HIF location methods, as it was tested on a larger system and with DG integration.

Furthermore, the generalization analysis of the proposed algorithm has shown the importance of testing data-driven algorithms in different scenarios to find their limitations. This test has demonstrated the extent of the proposed method's ability to locate HIFs, considering distinct scenarios from the ones used in the training/test dataset. The proposed method accurately located HIFs even when tested with new scenarios.

In general, the study presented an unprecedented investigation of the metrics and decision methods that can be used in HIF location algorithms. The proposed approach uses only one measurement device and can work with HIFs occurring along with the whole system in different scenarios. This investigation can help researchers develop increasingly efficient tools for this task, which has many challenges to overcome and is amplified when considering systems with DGs. Future work may include the evaluation of the proposed method under momentary loading variations and different network configuration.

APPENDIX
HIF SIMULATION PROCEDURE

This appendix presents the steps to simulate HIFs using real measurements in the ATPDraw software.

The first step is to acquire the real signals dataset. In this study, the real measurements were performed by the authors of [35], and the experiments consisted of the energized conductor rupture followed by the contact with different soil surfaces, such as sand, grass, clay, cement, and asphalt, both wet and dry.

According to [46], most harmonic-producing loads can be modeled by a current source. Thus, the same principle was used to simulate HIFs in this paper, as it enables reproducing the behavior found in real HIF signals within the simulations, as proven in the analysis in [37]. Thus, the second step to simulate the HIFs was to build a model to insert them in the system using a controlled current source via the MODELs

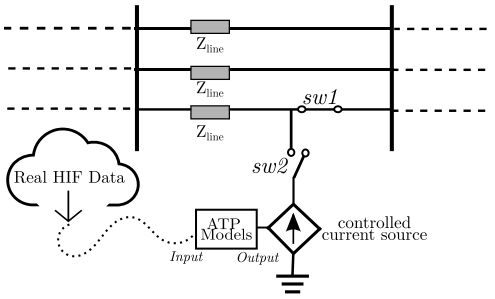


FIGURE 10. HIF simulation in ATP.

environment in the ATPDraw software. The scheme used to perform this simulation is shown in Fig. 10. The real HIF simulation scheme includes one switch ($sw1$) to simulate the conductor rupture and another switch ($sw2$) to connect the HIF model to the system, emulating the contact with the high-impedance surface. The conductor rupture is simulated and, after a delay (conductor fall period), the HIF starts. This period was set as 1.28 s, which is the minimum amount of time the conductor would take to fall considering an 8 m average DS pole using the free-fall equation.

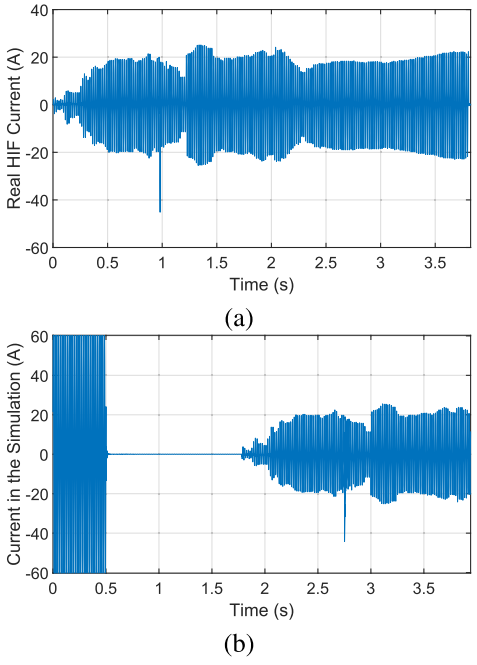


FIGURE 11. (a) Real HIF signal measured during the conductor contact with sand and the (b) signal measured in ATP when simulating this signal.

The models' input data consists of each sample of the real signals for each time step in a text format. The current source replicates each sample of the HIF current using the MODELS code shown in the pseudocode of Algorithm 2.

To better illustrate the key fault current dynamics, Fig.11a presents an example of a real HIF waveform measured during contact with sandy soil at the fault location. Fig.11b shows a simulation of an HIF at bus 802 (the closest to the substation) using the same input signal, with measurements taken at the substation. These plots are idealized to emphasize the essential phenomena: the pre-fault interval (up to 0.5 s) is shown as a constant current steady state, since this interval is under stable operating conditions. In the sequence, because the fault at bus 802 disconnects the downstream load, the measured current falls to zero upon conductor rupture. At 1.78 s, the HIF behavior begins, reliably reproducing the characteristic onset and distortion observed in the real signal.

Algorithm 2 Real HIF Signals Insertion Model

Input Variables:
t_{ref} { *Time variable incremented at each integration step*}
Ireal {*Real fault current*}

Function *Ireal*(POINTLIST):
INCLUDE <path to the real signal samples in your PC>

Initialization:
$t_{\text{ref}} := 0$

Execution:
while execution end time is not reached, do:
$t_{\text{ref}} := t_{\text{ref}} + \text{timestep}$
$I := Ireal(t_{\text{ref}})$ {*including the real current at each timestep*}

ACKNOWLEDGMENT

The authors would like to thank [35] from the Federal University of Uberlândia, Brazil, for providing them with the real HIF data. They would also like to thank São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil, for the facilities provided.

REFERENCES

[1] E. Baharozu, S. Ilhan, and G. Soykan, "High impedance fault localization: A comprehensive review," *Electr. Power Syst. Res.*, vol. 214, Jan. 2023, Art. no. 108892.
[2] G. N. Lopes, T. S. Menezes, D. P. S. Gomes, and J. C. M. Vieira, "High impedance fault location methods: Review and harmonic selection-based analysis," *IEEE Open Access J. Power Energy*, vol. 10, pp. 438–449, 2023.
[3] S. H. Mortazavi, Z. Moravej, and S. M. Shahrtash, "A searching based method for locating high impedance arcing fault in distribution networks," *IEEE Trans. Power Del.*, vol. 34, no. 2, pp. 438–447, Apr. 2019.
[4] Q. Cui and Y. Weng, "Enhance high impedance fault detection and location accuracy via μ-PMUs," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 797–809, Jan. 2020.
[5] M. S. Ali, A. H. A. Bakar, H. Mokhlis, H. Aroff, H. A. Illias, and M. M. Aman, "High impedance fault localization in a distribution network using the discrete wavelet transform," in *Proc. IEEE Int. Power Eng. Optim. Conf.*, Melaka, Malaysia, Jun. 2012, pp. 349–354.
[6] K. Moloi, J. A. Jordaan, and Y. Hamam, "High impedance fault classification and localization method for power distribution network," in *Proc. IEEE PES/IAS PowerAfrica*, Cape Town, South Africa, Jun. 2018, pp. 84–89.
[7] S. Sarangi, B. K. Sahu, and P. K. Rout, "High-impedance fault identification and location by using mode decomposition integrated adaptive multi-kernel extreme learning machine technique for distributed generator-based microgrid," *Electr. Eng.*, vol. 105, no. 1, pp. 383–406, Feb. 2023.
[8] S. Hossain, H. Zhu, and T. Overbye, "Distribution high impedance fault location using localized voltage magnitude measurements," in *Proc. North Amer. Power Symp. (NAPS)*, Pullman, WA, USA, Sep. 2014, pp. 1–6.
[9] A. Bouricha, T. Bouthiba, R. Boukhari, and S. Seghir, "High impedance faults location in the distribution networks using adaptive neuro-fuzzy inference system," in *Proc. Int. Conf. Electr. Sci. Technol. Maghreb (CISTEM)*, Algiers, Algeria, Oct. 2018, pp. 1–5.
[10] W. C. Santos, F. V. Lopes, N. S. D. Brito, and B. A. Souza, "High-impedance fault identification on distribution networks," *IEEE Trans. Power Del.*, vol. 32, no. 1, pp. 23–32, Feb. 2017.

[11] F. L. Vieira, J. M. C. Filho, P. M. Silveira, C. A. V. Guerrero, and M. P. Leite, "High impedance fault detection and location in distribution networks using smart meters," in *Proc. 18th Int. Conf. Harmon. Quality Power (ICHQP)*, Ljubljana, Slovenia, May 2018, pp. 1–6.

[12] J. Zhou, B. Ayhan, C. Kwan, S. Liang, and W.-J. Lee, "High-performance arcing-fault location in distribution networks," *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1107–1114, May 2012.

[13] M. J. S. Ramos, M. Resener, A. S. Bretas, D. P. Bernardon, and R. C. Leborgne, "Physics-based analytical model for high impedance fault location in distribution networks," *Electr. Power Syst. Res.*, vol. 188, Nov. 2020, Art. no. 106577.

[14] M. J. S. Ramos, A. S. Bretas, D. P. Bernardon, and L. L. Pfitscher, "Distribution networks HIF location: A frequency domain system model and WLS parameter estimation approach," *Electr. Power Syst. Res.*, vol. 146, pp. 170–176, May 2017.

[15] P. E. Farias, A. P. de Morais, J. P. Rossini, and G. Cardoso, "Non-linear high impedance fault distance estimation in power distribution systems: A continually online-trained neural network approach," *Electr. Power Syst. Res.*, vol. 157, pp. 20–28, Apr. 2018.

[16] J. J. G. Ledesma, K. B. do Nascimento, L. R. de Araujo, and D. R. R. Penido, "A two-level ANN-based method using synchronized measurements to locate high-impedance fault in distribution systems," *Electr. Power Syst. Res.*, vol. 188, Nov. 2020, Art. no. 106576.

[17] A. N. Milioudis, G. T. Andreou, and D. P. Labridis, "Detection and location of high impedance faults in multiconductor overhead distribution lines using power line communication devices," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 894–902, Mar. 2015.

[18] J. T. A. Vianna, L. R. Araujo, and D. R. R. Penido, "High impedance fault area location in distribution systems based on current zero sequence component," *IEEE Latin Amer. Trans.*, vol. 14, no. 2, pp. 759–766, Feb. 2016.

[19] A. Dasco, R. Marguet, and B. Raison, "Fault distance estimation in distribution network for high impedance faults," in *Proc. IEEE Eindhoven PowerTech*, Jun. 2015, pp. 1–6.

[20] L. U. Iurinic, A. R. Herrera-Orozco, R. G. Ferraz, and A. S. Bretas, "Distribution systems high-impedance fault location: A parameter estimation approach," *IEEE Trans. Power Del.*, vol. 31, no. 4, pp. 1806–1814, Aug. 2016.

[21] J. Li, G. Wang, D. Zeng, and H. Li, "High-impedance ground faulted line-section location method for a resonant grounding system based on the zero-sequence current's declining periodic component," *Int. J. Electr. Power Energy Syst.*, vol. 119, Jul. 2020, Art. no. 105910.

[22] J. U. N. Nunes, A. S. Bretas, N. G. Bretas, A. R. Herrera-Orozco, and L. U. Iurinic, "Distribution systems high impedance fault location: A spectral domain model considering parametric error processing," *Int. J. Electr. Power Energy Syst.*, vol. 109, pp. 227–241, Jul. 2019.

[23] N. Bahador, F. Namdari, and H. R. Matinfar, "Tree-related high impedance fault location using phase shift measurement of high frequency magnetic field," *Int. J. Electr. Power Energy Syst.*, vol. 100, pp. 531–539, Sep. 2018.

[24] S. R. K. Joga, P. Sinha, and M. K. Maharana, "A novel graph search and machine learning method to detect and locate high impedance fault zone in distribution system," *Eng. Rep.*, vol. 5, no. 1, p. 12556, Jan. 2023.

[25] M. A. Ravaglio, L. F. R. Toledo, S. L. Santos, L. R. Gamboa, D. B. Dahlke, J. A. Teixeira, E. T. Yano, A. P. Silva, O. Kim, and M. G. Antunes, "Detection and location of high impedance faults in delta 13.8kV distribution networks," *Electr. Power Syst. Res.*, vol. 230, May 2024, Art. no. 110291.

[26] L. Li, H. Gao, T. Yuan, F. Peng, and Y. Xue, "Location method of high-impedance fault based on transient zero-sequence factor in non-effectively grounded distribution network," *Electr. Power Syst. Res.*, vol. 226, Jan. 2024, Art. no. 109912.

[27] H. Xu, X. Zhang, H. Sun, W. Wu, J. Zhu, and F. Wang, "Faulty section location method for high impedance grounding fault in resonant grounding system based on discrete Fréchet distance," *Electr. Power Syst. Res.*, vol. 238, Jan. 2025, Art. no. 111147.

[28] J. Wang, B. Zhang, D. Yin, and J. Ouyang, "Distribution network fault comprehensive identification method based on voltage–ampere curves and deep ensemble learning," *Int. J. Electr. Power Energy Syst.*, vol. 164, Mar. 2025, Art. no. 110403.

[29] M. Wei, F. Shi, H. Zhang, and W. Chen, "Wideband synchronous measurement-based detection and location of high impedance fault for resonant distribution systems with integration of DERs," *IEEE Trans. Smart Grid*, vol. 14, no. 2, pp. 1117–1134, Mar. 2023.

[30] X. Wang, J. Gao, X. Wei, L. Guo, G. Song, and P. Wang, "Faulty feeder detection under high impedance faults for resonant grounding distribution systems," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 1880–1895, May 2023.

[31] A. Mousavi, R. Mousavi, Y. Mousavi, M. Tavasoli, A. Arab, and A. Fekih, "Artificial neural networks-based fault localization in distributed generation integrated networks considering fault impedance," *IEEE Access*, vol. 12, pp. 82880–82896, 2024.

[32] IEEE Distribution System Analysis Subcommittee. (2010). *IEEE 34 Node Test Feeder*. EUA. [Online]. Available: http://sites.ieee.org/pes-testfeeders/resources/

[33] D. Motter and J. C. de Melo Vieira, "The setting map methodology for adjusting the DG anti-islanding protection considering multiple events," *IEEE Trans. Power Del.*, vol. 33, no. 6, pp. 2755–2764, Dec. 2018.

[34] K. Dubey and P. Jena, "A novel high-impedance fault detection technique in smart active distribution systems," *IEEE Trans. Ind. Electron.*, vol. 71, no. 5, pp. 4861–4872, May 2024.

[35] J. R. Macedo, J. W. Resende, C. A. Bissochi, D. Carvalho, and F. C. Castro, "Proposition of an interharmonic-based methodology for high-impedance fault detection in distribution systems," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 16, pp. 2593–2601, Dec. 2015.

[36] R. H. Tan and V. K. Ramachandaramurthy, "Numerical model framework of power quality events," *Eur. J. Sci. Res.*, vol. 43, no. 1, pp. 30–47, Jan. 2010.

[37] G. N. Lopes, T. S. Menezes, G. G. Santos, L. H. P. C. Trondoli, and J. C. M. Vieira, "High impedance fault detection based on harmonic energy variation via S-transform," *Int. J. Electr. Power Energy Syst.*, vol. 136, Mar. 2022, Art. no. 107681.

[38] R. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: The s transform," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 998–1001, Apr. 1996.

[39] R. Witte and J. Witte, *Statistics*. Hoboken, NJ, USA: Wiley, 2007. [Online]. Available: https://books.google.com.br/books?id=Mn5GAAAAYAAJ

[40] Y. Aslan, "An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks," *Electr. Eng.*, vol. 94, no. 3, pp. 125–134, Sep. 2012.

[41] M. Shaik, A. G. Shaik, and S. K. Yadav, "Hilbert–Huang transform and decision tree based islanding and fault recognition in renewable energy penetrated distribution system," *Sustain. Energy, Grids Netw.*, vol. 30, Jun. 2022, Art. no. 100606.

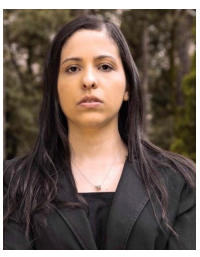
[42] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. Cham, Switzerland: Springer, 2013.

[43] A. Pragati, D. A. Gadanayak, S. Hasan, and M. Mishra, "Bayesian optimized ensemble decision tree models for MT-VSC-HVDC transmission line protection," in *Proc. Int. Conf. Adv. Power, Signal, Inf. Technol. (APSIT)*, Jun. 2023, pp. 385–390.

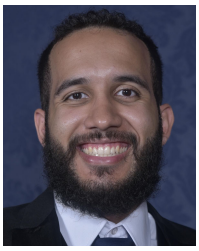
[44] H. Mirshekali, R. Dashti, A. Keshavarz, and H. R. Shaker, "Machine learning-based fault location for smart distribution networks equipped with micro-PMU," *Sensors*, vol. 22, no. 3, p. 945, Jan. 2022.

[45] The MathWorks Inc. (2023). *MATLAB*. Natick, MA, USA,. [Online]. Available: https://www.mathworks.com/products/MATLAB.html

[46] R. Dugan, *Electrical Power Systems Quality*, 2nd ed., New York, NY, USA: McGraw-Hill, 2003.



GABRIELA NUNES LOPES received the B.Sc. degree in electrical engineering from the Federal University of Mato Grosso, Cuiabá, Brazil, in 2018, and the M.Sc. and Ph.D. degrees from EESC-University of São Paulo (EESC-USP), Brazil, in 2020 and 2024, respectively. Currently, she is a Professor with the Federal University of Minas Gerais. Her research interests include high impedance faults, signal processing, artificial intelligence, and power quality.



PEDRO I. N. BARBALHO received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from EESC-University of São Paulo (EESC-USP), Brazil, in 2018, 2021, and 2025, respectively. In 2023, he was a Visiting Researcher with the University of Strathclyde, Glasgow, U.K. In 2025, he started as a Postdoctoral Researcher with the EESC-USP funded by The São Paulo Research Foundation, FAPESP. His research interests include machine learning algorithms for power system applications, microgrid control, and virtual power plant management.

JOSÉ CARLOS MELO VIEIRA (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Campinas, Campinas, Brazil, in 1999 and 2006, respectively. From 1999 to 2003, he was a Consulting Engineer with FIGENER, São Paulo, Brazil. Currently, he is an Associate Professor with the University of São Paulo, São Carlos, Brazil. His research interests include power distribution systems, integration of distributed energy resources, and power system protection.

DENIS V. COURY received the B.Sc. degree in electrical engineering from the Federal University of Uberlândia, Brazil, in 1983, the M.Sc. degree from EESC-University of São Paulo, Brazil, in 1986, and the Ph.D. degree from Bath University, England, in 1992. He joined the Department of Electrical and Computer Engineering, University of São Paulo, São Carlos, Brazil, in 1986, where he is currently a Full Professor with the Power Systems Group. He was with Cornell University, USA, from 1999 to 2000. His research interests include power system protection, expert systems, and smart grids.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614