A Feature Selection and Generalization Analysis for High Impedance Fault Classification Based on Support Vector Machine

Maurício Pavani da Silva, Gabriela Nunes Lopes, José Carlos de Melo Vieira Junior

***Abstract*—High Impedance Faults (HIFs) are common events in Electrical Power Distribution Systems (DSs) and are responsible for imposing danger to public safety due to the potential to ignite fires and cause electric shocks. Many recent studies have applied methodologies based on intelligent algorithms (IA) and machine learning (ML) to classify these events. However, these studies did not address feature selection and collinearity analysis to investigate the quality of methods' input metrics, resulting in high-dimensional data inputs that may hinder the method training and generalization, as well as its interpretability. Therefore, the purpose of the present study is to perform a feature selection evaluation for HIFs. Then, the metrics are used as inputs to a newly proposed classification method based on Support Vector Machine (SVM). The main goal is to compose a model for practical implementations, therefore a generalization study was conducted in two stages: training the detection algorithm under normal system conditions and validating it with unseen system conditions and events. The model was validated with a wide variety of tests that indicate an accuracy greater than 97% in the generalization analysis. The dataset for training, testing, and validation was obtained using Alternative Transients Program (ATP) software. This study proposes a robust method for classifying HIF with potential for practical application, in addition to providing a model for developing new studies on HIF classification using intelligent algorithms.**

***Keywords*—Fault Classification, Feature Selection, High Impedance Fault, Machine Leaning, Power Distribution System, Support Vector Machine.**

I. Introduction

HIGH Impedance Faults (HIF) are common events in Electrical Power Distribution Systems (DS) and are responsible for imposing danger to public safety due to the potential to ignite fires and cause electric shocks. These events are caused by the contact between an energized conductor and surfaces such as bushing trees, gravel, sand, grass, or asphalt, which have low conductivity and impose a low fault current during HIFs [1]. Moreover, the current measured during the event has low magnitude, can be intermittent and has non-linear behavior due to electric arc formation [2]. Consequently, conventional protection devices are incapable of appropriately detecting an HIF, which can keep occurring for hours until being reported to the utility.

Most strategies to detect an HIF uses signal processing tools, such as Fourier Transform (FT), Wavelet Transform (WT), or Stockwell Transform (ST) [1], [2], [3]. These methodologies are capable of detecting the event, however, they can present difficulties in differentiating the HIF from other events in DS, such as capacitor bank and load switching, transformer inrush current, non-linear load, and distributed energy resources operation. This limitation reduces the reliability of the methods and can result in protection misoperation.

Many recent studies have applied methodologies based on intelligent algorithms (IA) to address the issue of differentiating HIFs from other events based on predefined metrics extracted from current and voltage signals. The authors in [4] proposed an HIF detection method based on Support Vector Machine (SVM) applied to metrics obtained through Variational Mode Decomposition (VMD) to distinguish between HIFs and non-HIF events. The authors in [5] also proposed an HIF detection method using WT-based metrics and SVM. The authors in [6] and in [7] performed a comparative study between different IA methods for HIF detecting. In [8], the authors presented a detection method based on the energy of WT and investigated the application of various IA methods for decision-making. In [9], the authors proposed a method based on KNN to classify events on the distribution system. The method was based on the energy and standard deviation of the detail and approximation components of the WT. The authors in [10] proposed a detection method using the standard deviation of the WT and neural network. Although the discussed papers propose algorithms with high accuracy, they did not explore the features selected for their methods or explored how these methods perform during events that occur in conditions that are not on the training dataset. Furthermore, these methods were not evaluated using signals from actual events or including noise. This type of generalization analysis is crucial to ensure the reliable operation of these methods in real-world applications.

Additionally, the mentioned HIF detection methods investigate the application of IA and ML over a limited set of predefined metrics to detect an HIF. However, they do not address the feature selection process or collinearity analysis to investigate the quality of these metrics, which results in

Financed by the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES)-Finance Code 001, Brazilian National Council for Scientific and Technological Development (CNPq). This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Grant Number 2024/16135-7

Maurício Pavani da Silva is with Department of Electrical and Computing Engineering, University of São Paulo (USP), Brazil (e-mail of corresponding author: mauricio_pavani@usp.br). Gabriela Nunes Lopes is with Department of Electrical and Computing Engineering, University of São Paulo (USP), Brazil (e-mail: nuneslopesgabriela@gmail.com). José Carlos de Melo Vieira Junior is with Department of Electrical and Computing Engineering, University of São Paulo (USP), Brazil (e-mail: jcarlos@sc.usp.br).

Paper submitted to the International Conference on Power Systems Transients (IPST2025) in Guadalajara, Mexico, June 8-12, 2025.

methods that require input data with high dimensionality, hindering the training, generalization, and interpretability of the method. Moreover, a weak correlation between the metrics and the classes can negatively impact the performance of the algorithms. To try to address the problem of the correlation between the metrics and the classes, the authors in [11] proposed a framework for feature selection for HIF detection methods. The framework was based on an information gain ranking of a set of metrics extracted from current and voltage signals using Discrete Fourier Transform (DFT) and the Kalman filter. However, the authors did not present a collinearity analysis for the selected metrics, nor did they conduct a generalization analysis to ensure that the selected features were sufficient for accurately detecting and classifying an HIF in new event scenarios.

Therefore, the contributions and the purpose of this paper are to propose a complete methodology based on SVM for HIF classification involving feature selection and generalization analysis in order to distinguish them from other events. First, a feature selection study is performed on a set of low computational cost metrics obtained through the Short-Time Fourier Transform (STFT) of voltage and current to ensure the minimum dimensionality of the input data and avoid collinearity of the metrics. The feature selection analysis includes evaluating the length of the signal window, the Pearson correlation between each metric, and the mutual information between each metric and the target classes to rank them. Additionally, an optimal selection of hyperparameters of the SVM was applied based on the Grid Search Cross Validation (CV) algorithm. The method's generalization capacity is assessed regarding exposure to event conditions not present in the training database and includes analysis such as the different loading of the system and noisy signals. In the end, the methodology is presented to differentiate HIF and non-HIF events with a precision greater than 97%. The main contributions of this paper can be summarized as follows:

- Presenting a framework for feature selection based on mutual information applied to IA-based HIF classification. This framework analyzes the quality of the metrics and collinearity;
- Proposing the application of a Grid Search CV methodology for optimal selection of IA hyperparameters for classifying HIF;
- Investigating the generalization of the method by two stages: training the classification algorithm under system normal conditions and validating it with unseen system conditions and events;
- Developing an HIF classification method with low dimensionality data input and high generalization capacity, which are relevant characteristics for a practical implementation.

This paper is organized as follows: Section II presents the test system and explains how the HIFs were modeled. Section III presents the proposed feature selection methodology. Section IV presents the proposed SVM-based HIF classification method and the generalization analysis, and Section VII concludes the paper.

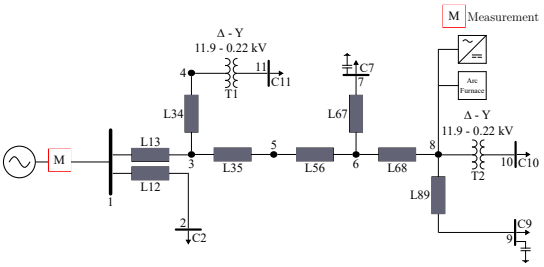

Fig. 1. Power distribution test system single-line diagram.

II. Test System Model

In this study, the performance of the proposed method is tested on a test system based on a real Brazilian DS. The topology of the test system is presented in Figure 1, which shows 9 buses with a nominal voltage of 11.9 kV and two buses with a nominal voltage of 220 V. This test system has power transformers, capacitor banks, non-linear loads, and linear loads that sum a nominal balanced loading of 5 MVA. These devices make it possible to evaluate different switching events and assess the performance of the detection method for false positives. The detailed modeling was given in [12]. The simulations were carried out with the system operating at both nominal loadings of 5 MVA and with only 10% of them. Moreover, it includes scenarios with balanced and unbalanced load conditions. In normal conditions, the system has a balanced loading of 5 MVA and noisy-free signals. In all these operating conditions, events such as capacitor bank switching, load switching, power transformer switching, rectifier operation, arc furnace operation, low impedance single-phase faults, and HIFs were simulated. The non-HIF events represent short-duration transients and non-linear loads with the potential for false positive detection. The simulations were carried out in the Alternative Transients Program (ATP) with 128 samples per cycle. All these simulations composed a database with 3,792 events, in which were 1,896 HIF and 1,896 non-HIF events.

The HIF events considered in this study were single-phase events that occur due to conductor rupture and contact with the ground at the substation side. To simulate this defect, 34 real HIF signals obtained through the cases presented in [13] were used. The authors in [14] explain that harmonics sources in electrical systems can be modeled as current sources. So, in this paper, the recorded signals were inserted into the simulation using a controlled current source and the Models environment available in ATP, as described in [15]. Figure 2 shows the topology of the circuit used in the simulation. The switches $sw1$ and $sw2$ were applied to simulate the rupture and contact of the cable with the ground. HIF events were simulated in all phases of the test system and in all 11.9 kV buses. Figure 3 illustrates a HIF current measured at the fault spot in the simulation, caused by the contact between phase A and the sand at bus 2, using the method presented in [15].

III. Proposed Methodology for Feature Selection

The proposed SVM HIF classification method is responsible for indicating whether the disturbance was an HIF or a non-HIF event. Therefore, the HIF classification conducted

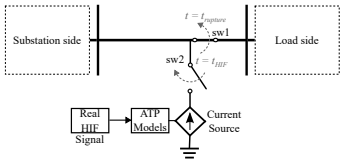

Fig. 2. HIF simulation scheme for real signals.

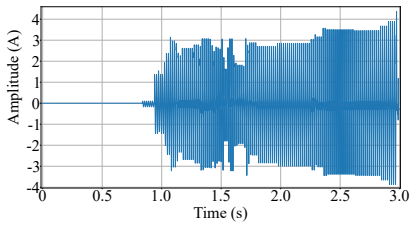

Fig. 3. HIF current measured at the fault spot caused by the contact between phase A and the sand at bus 2.

in this study was carried out by dividing the simulated dataset into HIF and non-HIF classes. This approach was chosen to minimize the computational burden of the final HIF classification method. For this classification process between HIF and non-HIF, three signal windows were evaluated to extract the metrics for the HIF classification method from the start of each event. The Window 1 had one cycle post-event, the Window 2 had three post-event cycles, and the Window 3 had five post-event cycles. Figure 4 illustrates these windows in the current signal measured at the fault spot during an HIF. For each window, the STFT was applied and the amplitudes of the harmonics from the second up to the 20th order were analyzed for the current and the voltage signals of the affected phase. This frequency range was chosen because it contains most of the energy of the signal measured during HIF, as can be seen in [15], and can still be obtained using devices with a low sampling rate. For each harmonic component and window, the maximum (max), mean, minimum (min), standard deviation (std), range, skewness, rugosity, kurtosis, and energy values were calculated. These statistical metrics resulted in 176 metrics for current and 176 for voltage totaling a pool of 352 candidate predictors. They were chosen to represent the randomness of the signal measured during an HIF and because of their low computational burden. Figure 5 presents a flowchart with the methodology to extract the metrics. Figure 5 presents a flowchart with the methodology to extract the metrics. In Figure 5, ATP simulated the events for the measurement voltage and current step. Python and its libraries handled COMTRADE, feature extraction, and implementing algorithms like FFT, SVM, and correlation methods.

For the feature selection analysis, mutual information and Pearson correlation were applied. Mutual information is a measure from information theory that quantifies how much information one random variable provides about another. A higher score indicates a stronger dependency between the metric and target variable, making the feature more useful for prediction [16]. In this paper, mutual information was calculated between each metric and the event classes by applying Equation (1), in which X and Y are metric classes and p is the probability function. The mutual information

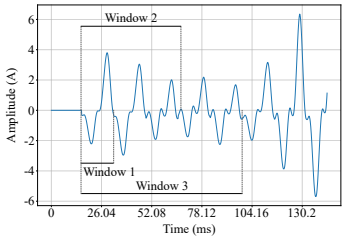

Fig. 4. Analyzed data window for the HIF classification method.

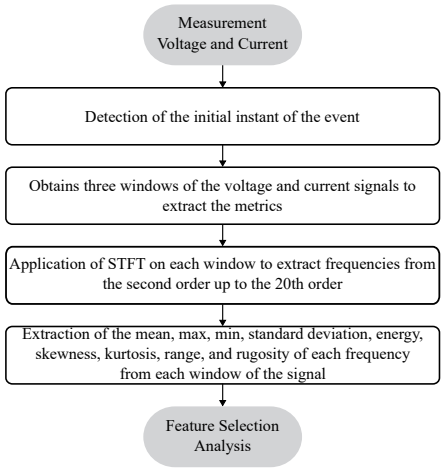

Fig. 5. Metric extraction process for feature selection, training, validating, and testing HIF classification method.

values obtined were sorted in a ranking from the metric with the highest score to the lowest. The best metrics and window for HIF classification were selected based on this ranking. Furthermore, collinearity is an index based on the correlation between two random variables, and it indicates whether these metrics carry the same information about the target class. In cases in which collinearity is identified, one of the metrics can be dispensable to the IA-based method. In this paper, for identifying collinearities within the pool of candidates' metrics, the Pearson correlation between each metric was calculated by applying Equation (2). Also, the Pearson correlation between a voltage metric and a current metric was calculated to identify collinearity between different signals. Thus, the steps for feature selection presented in this study are the selection of the data window, collinearity analysis, and the selection of metrics.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \times log\left(\frac{p(x,y)}{p(x) \times p(y)}\right) \quad (1)$$

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \quad (2)$$

A. Signal Window Selection

The size of the window used to extract a metric is directly connected to the amount of information needed to identify an HIF correctly. Moreover, windows that have many samples can lead to a delay in HIF classification. Therefore, it is necessary to choose the size of a signal window in order to identify an HIF correctly and reduce this delay. For the purpose of identifying the best window size, the dataset

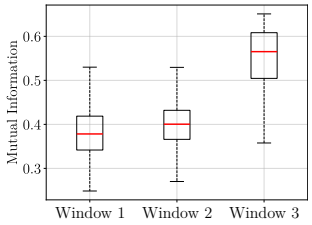

Fig. 6. Mutual information boxplot for each window size assessed.

with all 352 metrics extracted from HIF and non-HIF events during the normal operation of the test system was separated into three sets containing metrics extracted from each of the proposed windows. For each set, the mutual information between the metrics and the event classes was calculated. It was possible to identify the window size impacts by evaluating the mutual information of each group since the metrics and events were the same in all three sets and the only difference was the window size, consequently, the mutual information difference between the groups is due to signal window size. Measurements such as the mean, maximum, minimum, and standard deviation of the mutual information in each group were applied to illustrate the overall performance of the window and indicate the best choice for the method.

As mentioned previously, the application of mutual information results in a score capable of evaluating the metrics. Identifying the signal window that yields the highest mutual information values is a key parameter in determining its choice. Figure 6 illustrates the boxplot of the mutual information results of all metrics for each signal window. In this boxplot, the mean, max, min, and range of the mutual information for each window are available. The boxplot indicates that Window 3 presented better results than the other windows since its statistical values were higher. Due to the randomness of HIFs, a window that contains more information can often imply a better representation of the event, which may have caused the greater mutual information score of this window. This suggests that Window 3 is the most promising for application in the HIF classification method.

B. Collinearity Investigation

A collinearity analysis aims at identifying and removing metrics that provide the same information to the ML algorithm, reducing the dimensionality of the input data of the method. In this study, a window size, harmonic order, and type of signal measured during the events were fixed. Additionally, HIF and non-HIF events occurring during normal system operation were considered. These selected cases ensured that the impact on the metrics correlation depended only on the collinearity between them. In order to indicate the collinearity by applying the Pearson correlation, a threshold of 0.7 was adopted, which means if the correlation between two metrics was greater than 0.7 or less than -0.7, they have collinearity. Otherwise, the metrics were independent. The threshold of 0.7 indicates a high Pearson correlation.

The collinearity analysis provides an additional criterion for choosing metrics and aids in simplifying the complexity of ML models. In order to identify these metrics, Figure 7 illustrates this analysis on the 4th order harmonic current (i^{h4})

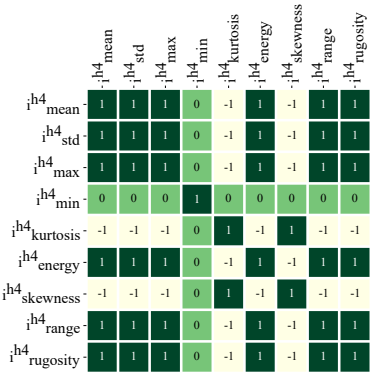

Fig. 7. Matrix of collinearity of current and 4th harmonic order metrics.

and contains a map that indicates 1 or -1 if the pair of metrics have collinearity and 0 if it does not based on this threshold. A negative signal in the Pearson correlation indicates that metrics are inversely proportional. Figure 7 illustrates that all metrics, except the minimum value (i^{h4}_{min}), are collinear. The minimum value not having collinearity with the other metrics may be because an HIF is an intermittent event. The minimum value is often the system load current value, which does not depend on the event. The collinearity between the metrics indicates that just one of them may be applied to HIF classification in order to avoid high dimensionality. The same analysis was used for metrics extracted from current and voltage and no collinearity was found. These results indicate that it is possible to choose only one metric from each signal, which contributes to the simplicity of the HIF classification implementation.

C. Selected Metrics for HIF classification

For the purpose of choosing the best metric, two analyses were conducted: the mutual information ranking and the graphical analysis. In the mutual information analysis, only events that occurred during normal system operating conditions were considered. Moreover, the study was performed considering the signal Window 3. After the mutual information was calculated, the metrics were sorted from high to low values and the best ones extracted from the current and voltage measured during the events were selected for graphical analysis. The graphical analysis consisted of analyzing the separability of the space formed by the metrics. The easier it is to separate classes in this space, the better the metrics.

Table I presents the 10 best metrics among the 176 extracted from the current measured during the HIF events and the 10 best extracted from the voltage. These 20 metrics compose a pool of the best candidates to be chosen. Additionally, the result of the collinearity analysis, which indicated that metrics extracted from the same signal provide the same information, was also used. Therefore, only one metric was selected from each type of signal to compose the HIF classification method. Figure 8 contains a graph with the distribution of events when characterized by the metrics i^{h4}_{range} and v^{h2}_{std}. It can be observed that these two metrics consistently distinguished most events, with only occasional instances of misclassification as an HIF. These results indicate that these two metrics have great potential in composing the HIF classification method.

TABLE I
TOP TEN METRICS OF CURRENT AND VOLTAGE SIGNALS SORTED BY MUTUAL INFORMATION RESULTS

Current signal		Voltage signal	
Metric	Value*	Metric	Value*
i^{h8}_{range}	0.6510	$v^{h4}_{kurtosis}$	0.6312
i^{h8}_{max}	0.6510	$v^{h4}_{skewness}$	0.6277
i^{h4}_{std}	0.6287	v^{h2}_{std}	0.6186
i^{h4}_{max}	0.6182	$v^{h6}_{skewness}$	0.6144
i^{h4}_{range}	0.6280	$v^{h14}_{skewness}$	0.6106
i^{h10}_{max}	0.6273	v^{h2}_{range}	0.6092
i^{h15}_{std}	0.6270	v^{h2}_{max}	0.6076
i^{h6}_{std}	0.6266	$v^{h4}_{rugosity}$	0.6008
i^{h10}_{range}	0.6266	v^{h10}_{range}	0.5924
i^{h20}_{std}	0.6256	$v^{h2}_{rugosity}$	0.5923

* - Mutual information result

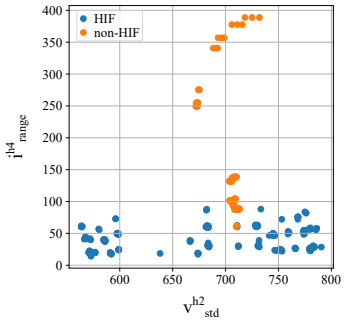

Fig. 8. Distribution of the HIF and non-HIF events at the v^{h2}_{std} x i^{h4}_{range}.

IV. HIF CLASSIFICATION METHOD

In this study, the challenge of the generalization of the method based on IA was addressed. Therefore, the HIF classification method developed was designed in two steps: optimal selection of hyperparameters and training and analysis of the generalization of the algorithm. The Grid Search CV methodology was used to select the optimal hyperparameters. This methodology presents a range of variations for the hyperparameters of the intelligent algorithm and, iteratively, trains and evaluates the performance of the method through cross-validation with different combinations of the hyperparameters in this range [17]. In the Grid Search CV, the combination of hyperparameters with the best performance in the cross-validation test is the chosen combination. The training and evaluation of the generalization capacity of the detection method was carried out by training the SVM with data from events that occurred under normal operating conditions and the validation was carried out with events that occurred outside the normal conditions without retraining. Additionally, events in the validation set were removed from the training set.

A. Selection of Hyperparameters by Grid Search CV

The SVM-based HIF classification method developed in this study used a Gaussian kernel to ensure the separability of event classes. The search space in the selection of hyperparameters was defined based on this kernel and on the general SVM hyperparameters. The SVM algorithm has a regularization hyperparameter (C) that increases the robustness of training against noise in the dataset. For optimization, four possible values of hyperparameter C were evaluated: 1, 10, 100, and 1000. Additionally, the Gaussian kernel requires an adjustment in the gamma hyperparameter inherent to the Gaussian function. For the optimization, six possible gamma values were evaluated: 0.001, 0.01, 0.1, 1, 10, and 100. After adjusting the search region of the Grid Search algorithm, the training and validation dataset was defined. In order to evaluate the generalization of the method later, the hyperparameters were optimized only for the normal operating conditions of the system. Also, the z-score normalization was applied to the data to eliminate the influence of the difference between the magnitudes of the metrics. It is presented in Equation (3), in which x is the value to be normalized, $\overline{x}$ is the mean, and σ is the standard deviation of the dataset. The z-score normalization makes the mean of the metrics equal to zero and the standard deviation equal to one. In the next step, the dataset was divided into 5 folds, and for each hyperparameter combination, the cross-validation method was applied for training, testing, and validating the method.

$$z = \frac{x - \overline{x}}{\sigma} \tag{3}$$

As discussed previously, the Grid Search CV algorithm was applied to ensure the maximum accuracy of the method with the data set and metrics used. The best SVM accuracy obtained by applying this method was 98.7%. This high accuracy value obtained with the application of SVM could suggest overfitting during training, however, the cross-validation methodology was applied to minimize the possibility that this has occurred. Therefore, the high performance of the algorithm may be related to the application of a hyperparameter optimization algorithm and the selection of the best metrics from the data set. The SVM with this accuracy was obtained with the combination of hyperparameters with C equal to 100 and gamma equal to 10. This combination of hyperparameters will be used to parameterize the SVM for detecting HIF in the method generalization evaluation step.

B. HIF Classification Method Generalization

For the practical application of the classification method, it is necessary to evaluate its generalization against system operating conditions not present in the model training. To address this problem, a generalization study of the detection method was conducted, consisting of two stages: training the classification algorithm under normal system operating conditions and validation for conditions and events not present in the training dataset. Therefore, the training dataset consisted of HIF and non-HIF events during system operation with loading equal to 100% and 10%. This training dataset did not contain noisy signals or unbalanced loads. The SVM algorithm with the Gaussian kernel and the hyperparameters selected by the Grid Search CV algorithm was trained with this dataset corresponding to 70% of the set. After training, the method was tested with the remaining 30% of the set. The method was also evaluated, without retraining, with a new dataset

composed of events with unbalanced loads and 50 dB Gaussian White noise. Moreover, to increase the rigor of the tests, the events were not repeated in the databases.

To evaluate the performance of the method and its generalization, four indexes were used. The first index evaluated was accuracy, which indicates the ratio between the number of correctly classified events and the total number of events tested (Equation (4)). The next metric was the recall, shown in Equation (5), which evaluates the performance of the intelligent method in correctly classifying only the target class. In the case of the detection described in this study, the target class is HIF. It varies between 0 and 1, and the closer to 1, the better the method's performance. The next index is balanced accuracy, which is calculated as the average of the recall values for a possible class in the database. This metric addresses the issue of unequal event distribution across classes. Finally, the last metric evaluated is the F1-score, which is interpreted as an average between accuracy and recall. Equation (6) shows how this metric is calculated. It also varies between 0 and 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1_{score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

In these equations, TP and TN are the true positive and negative, FP and FN are false positive and negative.

Table III presents the validation results of the SVM algorithm training in the different system operation scenarios. Eight scenarios were evaluated considering the combination of noise parameters (no noise and 50 dB Gaussian white noise), loading (100% and 10% of nominal), and state (balanced and unbalanced loads). The conditions named "Balanced and 100% Loading" and "Balanced and 10% Loading" correspond to the validation of the method for the same training conditions. In both scenarios, accuracies higher than 99% were obtained, which indicates a high performance of the method. The high balanced accuracy, F1 score, and recall confirm the effectiveness in distinguishing HIF from non-HIF, highlighting the reliability of the chosen metrics and hyperparameters. The other system conditions of the table provide results for the scenarios unknown to the detection method. The lowest accuracy obtained was 97.6% for the system operating conditions that were not in the training. Even for the worst case evaluated, the method maintained a performance higher than 97%. These generalization results of the method suggest a good choice of metrics and selected hyperparameters, and they illustrate the effectiveness of the practical application.

V. Comparison

In this section, the proposed method has been compared with the methods reported in [4] and [9], recent methods that use IA to detect HIFs. The method proposed in [4] implements an algorithm based on a SVM and features extracted from current signals through Variational Model Decomposition. The method proposed in [9] implements an ensemble classifier based on K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Tree (RT), and features extracted from current signals with Wavelet Transform. Both papers validated the proposed methods in several tests. However, the generalization of the classification models was not analyzed under different conditions of the system. Table II shows the classification accuracy for the nominal conditions of the test system and training the method with 70% and testing with 30%. All methods in Table II were implemented and tested in the same conditions. The results reveal that the proposed method performs better.

VI. Final Remarks for the Proposed Method

The analyses in this paper proposed a methodology to distinguish HIF from other events in distribution systems using SVM algorithm. It included a feature selection study on a set of low-cost voltage and current metrics to minimize input data dimensionality and avoid metrics with collinearity. The feature selection process considered the signal window length, Pearson correlation between metrics, mutual information between metrics and target classes for ranking, and graphical analysis. Additionally, optimal SVM hyperparameters were selected using the Grid Search CV algorithm to maximize the accuracy of the SVM and avoid overfiting. The method's generalization ability was also evaluated under event conditions absent from the training set and including different system loads and noisy signals. The final version of method trains and validates in 7s, and classifies the dataset in 1ms after SVM training, using a computer with an AMD Ryzen 7 2700 processor (3.2 GHz) and 32GB RAM for efficiency. Figure 9 presents the comprehensive methodology proposed in this paper, encompassing each step.

VII. Conclusions

To distinguish HIF from other events is a hard-to-solve problem in DSs. Previous studies have presented HIF detection methods based on intelligent algorithms. However, these studies did not apply feature selection techniques to assess the quality of the input metrics or even the existence of collinearity among them. This can result in models with high dimensionality. DSs present variable conditions, which can hinder the real-life applicability of AI-based classification methods. In this sense, existing solutions did not assess the generalization capacity of the models when subjected to fault conditions and events not present in the training database. To address these gaps, this paper presented an SVM-based HIF classification method with feature selection techniques, optimal hyperparameter selection, and training generalization analysis aligned with practical implementation bias.

Initially, 352 candidate metrics were proposed for the SVM algorithm. Using feature selection techniques, only

TABLE II
Accuracy comparison with other methods in the literature

Method	IA Method	Accuracy
Proposed	SVM	99.3%
[4]	SVM	92.5%
[9]	KNN	97.2%
	LR	60.4%
	RT	97.2%

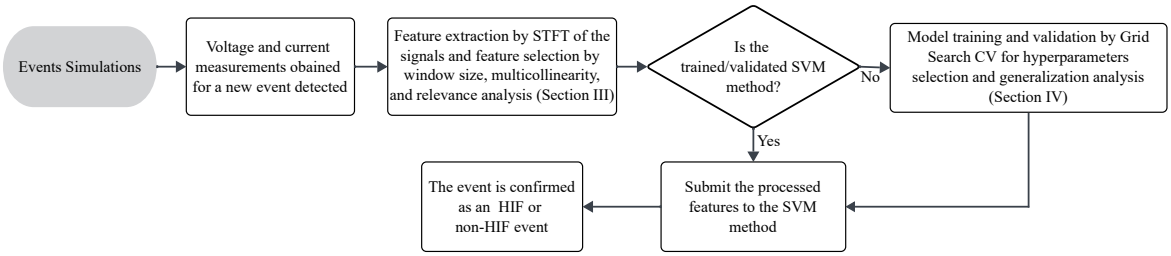

Fig. 9. Practical implementation flow chart of the proposed HIF classification method based on SVM and feature selection.

TABLE III
GENERALIZATION ANALYSIS OF THE PROPOSED METHOD

Event conditions	Accuracy	Balanced Accuracy	F1-score	Recall
Noise free				
Balanced and 100% Loading	99.3%	0.993	0.993	0.987
Balanced and 10% Loading	99.3%	0.994	0.994	0.988
Unbalanced and 100% Loading	98.6%	0.985	0.985	0.971
Unbalanced and 10% Loading	97.6%	0.979	0.980	0.973
Noise level of 50 dB				
Balanced and 100% Loading	97.9%	0.980	0.980	0.961
Balanced and 10% Loading	98.6%	0.986	0.988	0.988
Unbalanced and 100% Loading	98.6%	0.985	0.985	0.971
Unbalanced and 10% Loading	97.6%	0.979	0.980	0.973

two were selected. The selected metrics were capable of characterizing the evaluated events without collinearity and with high correlation with the HIF classification. This result demonstrates the importance of the study of feature selection in the construction of ML models.

In the following, the Grid Search CV algorithm was applied to find the optimal combination of hyperparameters for the SVM algorithm, to maximize its performance. For this purpose, a search space was defined for each of the desired hyperparameters and a database with events under nominal system conditions. This study returned the best set of hyperparameters for the base case of events without overfitting during training, given that the cross-validation training method was applied. In the generalization evaluation, even in the worst-case scenario, the proposed classification method maintained an accuracy above 97%, illustrating the effectiveness of the presented methodology.

In general, this study can contribute to the advancement of the state of the art of HIF classification methodologies based on intelligent algorithms. It is also expected to contribute with a methodology for the development of new IA-based models to operate in electrical power systems.

REFERENCES

[1] A. Ghaderi, H. L. Ginn, and H. A. Mohammadpour, "High impedance fault detection: A review," *Electric Power Systems Research*, vol. 143, pp. 376–388, 2017.

[2] D. Gomes and C. Ozansoy, "High-impedance faults in power distribution systems: A narrative of the fields developments," *ISA Transactions*, vol. 118, pp. 15–34, 2021.

[3] M. Mishra and R. R. Panigrahi, "Taxonomy of high impedance fault detection algorithm," *Measurement*, vol. 148, p. 106955, 2019.

[4] B. K. Chaitanya, A. Yadav, and M. Pazoki, "An intelligent detection of high-impedance faults for distribution lines integrated with distributed generators," *IEEE Systems Journal*, vol. 14, no. 1, pp. 870–879, 2020.

[5] M. S. Attar and M. R. Miveh, "High-impedance fault detection in distribution networks based on support vector machine and wavelet transform approach (case study: Markazi province of iran)," *Energy Science and Engineering*, vol. 13, no. 3, p. 1171 1183, 2025.

[6] M. Sarwar, F. Mehmood, M. Abid, A. Q. Khan, S. T. Gul, and A. S. Khan, "High impedance fault detection and isolation in power distribution networks using support vector machines," *Journal of King Saud University - Engineering Sciences*, vol. 32, no. 8, pp. 524–535, 2020.

[7] A. T. Diefenthäler, A. T. Z. R. Sausen, M. De Campos, P. S. Sausen, and J. M. Lenz, "Artificial neural networks: Modeling and comparison to detect high impedance faults," *IEEE Access*, vol. 11, pp. 124 499–124 508, 2023.

[8] V. Veerasamy, N. I. A. Wahab, M. L. Othman, S. Padmanaban, K. Sekar, R. Ramachandran, H. Hizam, A. Vinayagam, and M. Z. Islam, "LSTM recurrent neural network classifier for high impedance fault detection in solar pv integrated power system," *IEEE Access*, vol. 9, pp. 32 672–32 687, 2021.

[9] K. Swarna, A. Vinayagam, M. B. J. Ananth, P. V. Kumar, V. Veerasamy, and P. Radhakrishnan, "A KNN based random subspace ensemble classifier for detection and discrimination of high impedance fault in pv integrated power network," *Measurement*, vol. 187, p. 110333, 2022.

[10] H. Bai, J.-H. Gao, T. Liu, Z.-Y. Guo, and M.-F. Guo, "Explainable incremental learning for high-impedance fault detection in distribution networks," *Computers and Electrical Engineering*, vol. 122, 2025.

[11] Q. Cui, K. El-Arroudi, and Y. Weng, "A feature selection method for high impedance fault detection," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1203–1215, 2019.

[12] M. P. da Silva, "Deteco de faltas de alta impedncia em sistemas de distribuio de energia eltrica baseada na entropia de shannon," Master's thesis, Escola de Engenharia de São Carlos - EESC/USP, São Carlos, Jan 2024.

[13] 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 Generation, Transmission and Distribution*, vol. 9, no. 16, pp. 2593–2601, 2015.

[14] R. C. Dugan, M. F. McGranaghan, S. Santoso, and H. W. Beaty, *Electrical Power Systems Quality*, 2nd ed. New York: McGraw-Hill, 2004, no. ISBN: 0-07-138622-X.

[15] G. Lopes, T. Menezes, G. Santos, L. Trondoli, and J. Vieira, "High impedance fault detection based on harmonic energy variation via S-transform," *International Journal of Electrical Power & Energy Systems*, vol. 136, p. 107681, 2022.

[16] S. Verdú, "alfa-mutual information," in *Information Theory and Applications Workshop (ITA)*, 2015, pp. 1–6.

[17] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-nearest neighbors and grid search cv based real time fault monitoring system for industries," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 2019, pp. 1–5.