Interacting with Features: Visual Inspection of Black-box Fault Type Classification Systems in Electrical Grids

Carmelo Ardito, Yashar Deldjoo, Eugenio Di Sciascio, and Fatemeh Nazary*

Politecnico di Bari, Italy
firstname.lastname@poliba.it

Abstract. Automatic fault type classification is an important ingredient of smart electrical grids. Similar to other machine-learning models, methods developed for fault classification suffer from the issue of *lack of transparency*. This work sheds light on preliminary insights of an ongoing study, in which we show how feature importance measurement and feature interaction visualization using partial dependence plots (PDPs) can help *interpretability* of the classification outcomes. While the former, measures the role of each feature on the final predictions in isolation, the latter focuses on mutual interaction between pairs of features. We show the merits of these two complementary feature analysis mechanisms in facilitating interpretability of the fault type classification task.

Keywords: Fault type classification · Interpretability · Visualization.

1 Introduction and Context

Smart grids (SGs) are recognized as power distribution systems (PDSs) that need to possess traits including high reliability, efficiency, and penetration of renewable energy sources [1]. PDSs, however, are susceptible to a variety of electrical abnormalities and occasional failures, as the result of adverse weather conditions, equipment aging and degradation, security attacks among others. Over the last years, a set of machine-learned approaches have emerged that aim to detect and diagnose fault in a data-driven manner. This capability, known as *self healing*, is important to make electrical grids reliable and smart. In a nutshell, the goal in self healing is to restore and recover the interruption of electricity in the electrical grid automatically and reduce the interruption period for costumers [7] by performing *fault detection*, *fault type classification* and *fault location identification*. Fault type classification, the task we focus our attention in this work, classifies an occurred electrical fault in the three-phase electrical grid into one of the predefined classes according to (i) symmetrical faults, such as *LLL, LLLG*, which are related to three-phase faults, and (ii) asymmetrical

* Authors are listed in alphabetical order. Corresponding author: Fatemeh Nazary

faults, such as *LG, LL, LLG*, which show line-to-ground, line-to-line and line-to-line-to-ground faults respectively.

A common characteristic of the prior literature is that the nature of the empirical experiments carried out orients toward the *prediction* aspect of the fault event, aiming to find an answer to questions such as "*is it possible to detect a fault using ML techniques reliably*"? or "*which classification technique can more accurately predict a class type?*" and so forth. Regretfully, such trends for full automation of PDS's self-healing capability are not designed to inform human operators who have relied on manual/visual awareness for a long time. To keep humans involved in the control loop, it is crucial to design *interpretable* ML models that can replace these black-box prediction models and to produce rules that can be understood with little inspection.

Motivated by this observation, the work at hand puts its attention outside the subject of proposing another classification method for fault prediction, instead it tries to focus on the central question "*Given popular classification techniques already recognized by the community, is it possible to exploit the results of predictions in order to obtain more interpretable outcomes?*"

The contributions of this work are two-fold:

1. **Feature extraction and representation**: we rely on features extracted from the three-phase voltage signals, represented in both *time* and *frequency* (transform) domains. For feature representation, we compute the n-th moment of the probability distribution functions (PDFs) [11] ($n \in [1, 4]$) together with the *energy* and *max* of the signals on both time- and frequency-domain signals.
2. **Interpretability**: To better facilitate interpretability, we utilize *feature importance measurement* by employing the model-dependent technique based on decision tree [12], and further propose to utilize *visual analytic techniques* using partial dependence plots (PDPs) [8]. These two complementary visual analysis techniques measure/visualize the individual impact of features and their pairwise relationship on the final classification outcome, thereby helping user interpret the results of the classification model at hand.

The results of our empirical study show that in general, the computed features in this work are not only descriminative for our classification scenario, but are also easily interpretable, making the classification process transparent. While previous works have exploited features coming from signal or transform domains [4, 10, 5], our approach for *computing n-the PDF moments* of the both time and frequency signals, extracts rich information from signals that tend to be mutually complementary in some cases. In fact, by combining feature visualization (what is the relationship between features?) with attribution (how does it affect the output?), we can explore how the classifier decides between different fault types. The current work presented in this paper is the preliminary result of a larger ongoing study that makes advances to interpretability of ML models in the context of SGs, providing new insights on how to interpret results of fault prediction by proposing an inexpensive feature extraction, feature selection and visualization technique.

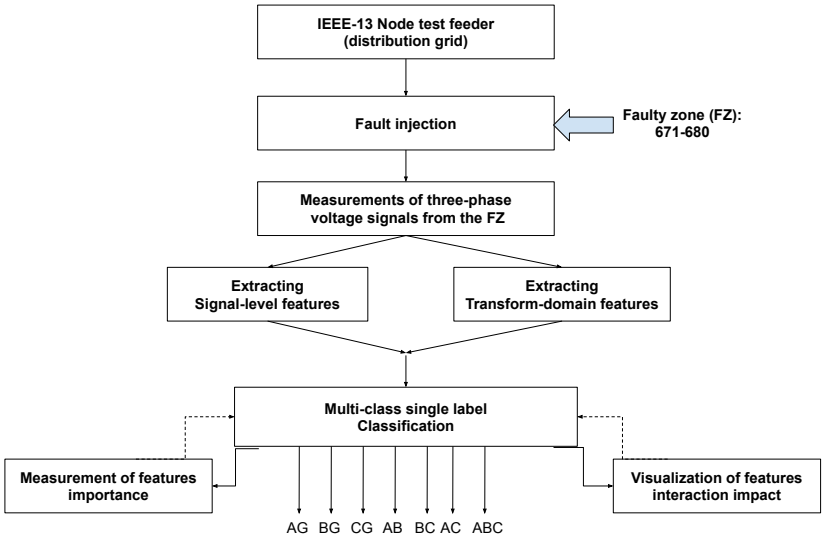

Fig. 1. The main processing stages in our proposed system

2 Proposed method

The goal of the proposed method is two-fold: (i) fault-type classification, and (ii) interpretability achieved via *feature importance measurement* and *data visualization*. The main processing stages involved in the proposed system are presented in Figure 1. The input to the system is the IEEE-13 node test feeder, while the output is one of the seven fault types, namely: line-to-ground (AG, BG,CG), Line-to-Line (AB, AC, BC), and three-phase fault (ABC).

2.1 Fault simulation and Feature Extraction

We chose IEEE-13 node test feeder, which includes a voltage generator of 4.16 kvlt and 13 buses for the simulation of fault and measurement of three-phase signals. One can divide this distribution system into four critical zones, zone 1: 632-671, zone 2: 632-633, zone 3: 692-675, and zone 4: 671-680. To collect data, faults were injected to one arbitrarily chosen zone, in this case zone 4, and then features were collected from three-phase voltage signals of this zone. We injected all the 7 different faults (i.e., AG, BG, CG, AB, BC, AC, ABC). These faults have been applied at a certain start time $t = 0.01$ and revoked at time $t = 0.02$ for all of the fault simulations. Thus, $t_f = [0.01 - 0.02]$ represents the *faulty period* while $t_h = [0 - 0.01]$ characterizes the *non-faulty (healthy) period*. All the features that were extracted were taken from the faulty period t_f were normalized by the same feature extracted from the healthy period t_h to obtain a relative score. The following two classes of features were extracted:

– **Signal-level features:** Six features were extracted from raw voltage data of three phases. They include the 1st to 4-th moments: *mean*, *standard deviation*, *skewness*, *kurtosis* together with the *energy* and the *maximum* level of the signal.
– **Transform-domain features:** In addition, we extracted features based on discrete Fourier transform (DFT), to obtain richer information about frequency of the signals. After applying DFT, from the computed spectrum we extracted similar features as signal-level features.

In total, 12 (6+6) features were collected to represent the features in our labelled training dataset. These two set of features constitute the backbone of many ML systems [3, 2]. To augment the training dataset with further data, the fault resistance value R_f in the fault detection module was varied by choosing 20 different values in the range of 0.001 to 2 as done in previous works [9, 6]. This resulted in 20 simulations for each of the fault types and a training dataset of 140 samples taking into account all the 7 fault types.

2.2 Fault type classification and interpretability analysis

Fault type classification was done by using two main classifiers: *decision tree* and *k-nearest neighbors*. We model the classification task as a multi-class signal label classification — instead of multi-label — since there are more classifiers' choices available for the single-label classification task. For interpretability experiment (see next section), we only use decision tree to keep the discussion simple.

Finding important variables (features) helps to discover the main drivers in a supervised learning classification task. However, this approach does not produce information about the relationship between input variables and how this relationship impacts the ML model outcome (predictions). The approach envisioned in this work contemplates using: (i) a classical feature importance technique to show the contribution of each feature on predictions individually, and (ii) a partial dependence plot (PDP) to understand the relationship between pairs of input variables and predictions. PDP is calculated after the model is fitted on the training data; thus, it is a *model-specific* feature importance analysis technique (rather than model-agnostic). For example, in our context a PDP can show whether the probability of certain fault increases with signal energy and kurtosis of the frequency signal, a question whose answer does not seem to be trivial. Furthermore, PDP can establish the type relationship between two features: monotonic, linear, or not related. These are important cues that can help the human operator to better inspect/interpret the black-box fault classification predictions with little supervision.

3 Results and discussions

The discussion of results is organized into two sections. First, we describe the results of classification and next, we describe the impact of two feature analysis

techniques on the interpretability of classification predictions.

Classification: Table 1 summarizes the classification results using two classifiers, namely decision tree and k-nearest neighbors, on the basis of a hold-out setting (80%-20%) for training and test set. We can notice that in all the considered experimental cases the average classification accuracy is more than 92%, indicating the discriminative power of the features chosen. The best classification outcome is achieved for the decision tree with the accuracy of 96.42%. Thus, we use decision tree for the next step.

Table 1. Classification accuracy (%) using 12 features and two classifiers. For the k-nearest neighbors, $k = 5$ was used.

Classifier	**decision tree**	**k-nearest neighbors**
Accuracy	96.42	92.85

Feature analysis and interpretability: Results of feature importance analysis are shown in Fig. 2. In particular, Fig 2-a shows the impact of individual features on fault type classification predictions. According to the results, the most informative features are (i) from *signal-level features*: energy, mean and kurtosis, while (ii) from *frequency-level features*: energy and mean. Thus, the information that this analysis provides is that both signal-level and frequency-level features can play a role in the classification predictions.

Fig 2-b and Fig 2-c however provide a more meticulous interpretation of the results. These plots are results of utilizing the PDP approach (see Section 2.2) and visualize the impact of mutual feature interactions on the classification outcome. We can note that the two selected features (as an example) in Fig 2-b, i.e., **mean_dft** and **energy_sig** are NOT mutually informative; in other words, a change in the values of both of these features does not lead to the increase or decrease in the classification outcome. This is equal to say that **mean_dft** has all the necessary information encoded in the set {**mean_dft**, **energy_sig**}. Thus, we can safely use **mean_dft** for the classification task and expect to obtain good classification results. However, as shown in Fig 2-c, for what concerns the interaction between features {**mean_dft**, **kurtosis_sig**} a different relation is obtained. We can note that, in this case, both of the features monotonically impact the classification predictions. The highest classification is achieved when feature values are in the bottom-left portion of the figure.

We round off this discussion by highlighting that the results of our study show that the information provided by the PDP analysis for the SG fault type classification task offer new insights that could not be obtained from the classical feature importance analysis technique, as shown in Fig 2-a. For example, while Fig 2-a reports on the impact of the 12 employed features as a group, it does not provide specific insights if the same results could be obtained when a smaller set of features are used. We can see that while some pairs of features are mutually complementary such as **mean_dft** and **energy_sig**, there exist other

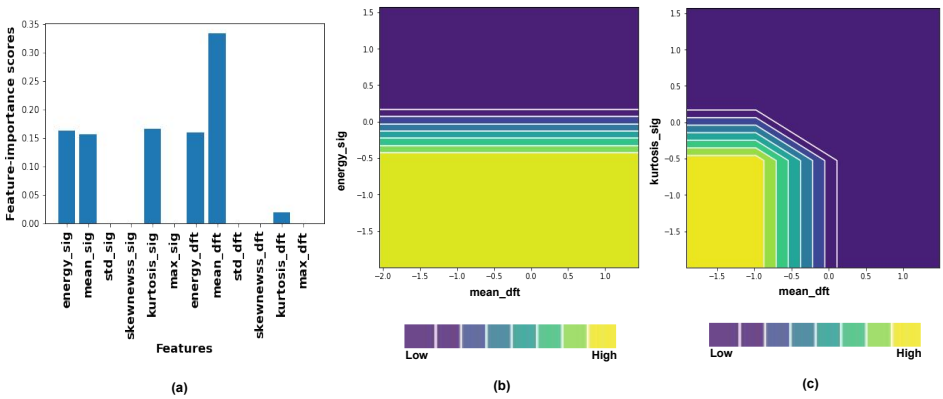

Fig. 2. Results of feature analysis (a) feature importance scores for 12 features by the decision tree (b-c) PDP interaction plots using two dominant features in part (a).

feature pairs that are correlated. This information could eventually be used by the system designer to know (i) which feature(s) to focus on for the extraction phase from the SG signals, (ii) how to represent the feature to obtain more informative features (e.g., n-th PDF moment we used), and (iii) by the system human operator to understand the root of specific faults in the system.

4 Conclusion and future work

This work presented preliminary results of a large study, in which we focused on the central question of interpretability of ML models in the context of fault prediction for smart grids. First, we classified fault types using two different classifiers, k-nearest neighbors and decision tree, and identified decision tree as the best choice; afterwards, for the interpretability task, we studied the role of two complementary feature analysis techniques, namely feature importance measurement and feature interaction visualization using partial dependence plots (PDPs). We provided insights that can be obtained from the PDP technique on the relationship between features, that could not be found in the classical approach. Our study acknowledges merits of the two complementary feature analysis mechanisms in facilitating offering explanations. For the future work, we plan to extend our dataset by injecting fault to other critical zones, and using a wider set of features. We plan to experiment with larger electrical grids, e.g., IEEE-34, 37 and 123 that are commonly used in the literature [3]. Finally, we consider to study more interpretable models for the core prediction task.

Acknowledgments

This work has been partially funded by *e-distribuzione S.p.A* company, Italy, through a PhD scholarship granted to Fatemeh Nazary.

References

1. Cremer, J.L., Konstantelos, I., Strbac, G.: From optimization-based machine learning to interpretable security rules for operation. IEEE Transactions on Power Systems **34**(5), 3826–3836 (2019)
2. Deldjoo, Y., Schedl, M., Cremonesi, P., Pasi, G.: Content-based multimedia recommendation systems: Definition and application domains. In: Proceedings of the 9th Italian Information Retrieval Workshop, Rome, Italy, May, 28-30, 2018. CEUR Workshop Proceedings, vol. 2140. CEUR-WS.org (2018)
3. Gilanifar, M., Cordova, J., Wang, H., Stifter, M., Ozguven, E.E., Strasser, T.I., Arghandeh, R.: Multi-task logistic low-ranked dirty model for fault detection in power distribution system. IEEE Transactions on Smart Grid **11**(1), 786–796 (2019)
4. Jamehbozorg, A., Shahrtash, S.: A decision tree-based method for fault classification in double-circuit transmission lines. IEEE transactions on power delivery **25**(4), 2184–2189 (2010)
5. Kashyap, K.H., Shenoy, U.J.: Classification of power system faults using wavelet transforms and probabilistic neural networks. In: Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03. vol. 3, pp. III–III. IEEE (2003)
6. Lwin, M., Min, K.W., Padullaparti, H.V., Santoso, S.: Symmetrical fault detection during power swings: An interpretable supervised learning approach. In: 2017 IEEE Power & Energy Society General Meeting. pp. 1–5. IEEE (2017)
7. Mohammadi-Hosseininejad, S.M., Fereidunian, A., Shahsavari, A., Lesani, H.: A healer reinforcement approach to self-healing in smart grid by phevs parking lot allocation. IEEE Transactions on Industrial Informatics **12**(6), 2020–2030 (2016)
8. Molnar, C.: Interpretable Machine Learning. Lulu. com (2020)
9. Onaolapo, A.K., Akindeji, K.T., Adetiba, E.: Simulation experiments for faults location in smart distribution networks using ieee 13 node test feeder and artificial neural network. In: Journal of Physics: Conference Series. vol. 1378, p. 032021. IOP Publishing (2019)
10. Saleh, K.A., Hooshyar, A., El-Saadany, E.F.: Hybrid passive-overcurrent relay for detection of faults in low-voltage dc grids. IEEE Transactions on smart grid **8**(3), 1129–1138 (2015)
11. Spanos, A.: Probability Theory and Statistical Inference: Empirical Modeling with Observational Data. Cambridge University Press (2019)
12. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., et al.: Top 10 algorithms in data mining. Knowledge and information systems **14**(1), 1–37 (2008)