SAND2022-11707C

Shapley Additive Explanations for Traveling Wave-based Protection on Distribution Systems

Miguel Jiménez-Aparicio
Sandia National Laboratories
Albuquerque, NM, USA

Matthew J. Reno
Sandia National Laboratories
Albuquerque, NM, USA

Felipe Wilches-Bernal
Sandia National Laboratories
Albuquerque, NM, USA

***Abstract*—This paper proposes a framework to explain and quantify how a Traveling Wave (TW)-based fault location classifier, a Random Forest, is affected by different TW propagation factors. The classifier's goal is to determine the faulty Protection Zone. In order to work with a simplified, yet realistic, distribution system, this work considers a use case with different configurations that are obtained by optionally including several common distribution elements such as voltage regulators, capacitor banks, laterals, and extra loads. Simulated faults are decomposed in frequency bands using the Stationary Wavelet Transform, and the classifier is trained with such signals' energy. SHapley Additive exPlanations (SHAP) are used to identify the most important features, and the effect of different fault configurations is quantified using the Jensen-Shannon Divergence. Results show that distance, the presence of voltage regulators and the fault type are the main factors that affect the classifier's behavior.**

***Index Terms*—Power System Protection, Traveling Waves, Distribution Systems, Machine Learning, Stationary Wavelet Transform, SHapley Additive exPlanations**

I. Introduction

The development of Traveling Wave (TW)-based protection schemes for the distribution level is a common topic in the research community. Advanced signal processing techniques are used to get insightful information on those fast transients [1]. Most of the proposed approaches use Machine-Learning (ML) or Deep-Learning models (DL) when it comes to predicting the fault location or type. However, to our knowledge, this technology is still not used in commercial relays.

One of the main disadvantages of ML/DL approaches is that the models are black boxes: given some inputs, the model returns some distance to the fault or some fault type classification. This situation is considered too risky for the power system protection industry. Our daily lives rely on electricity supply, and the reliability and security of the power system cannot rest on black-box types of solutions. Besides these concerns on the ML/DL applicability, there are also debugging and interpretability issues. First, the model may be biased or wrong, but usually no insights can be directly obtained from its internal behavior. Second, these models may be able to unveil patterns in the data that are not obvious even to human experts. However, if their behaviors are not analyzed, the opportunity to expand our knowledge is lost.

In the last decades, there has been some effort in the ML community to develop techniques that help to explain the models' behavior. What is perhaps the most advanced of these, the SHapley Additive exPlanations (SHAP) [2], has been selected for this work. The main contributions are:

- Proposing a framework, based on SHAP, to identify the most important features in a TW fault location classifier.
- Using the SHAP values to describe the classifier's behavior on different TW propagation scenarios, and quantifying the effect of several factors in the classifier's response using the Jensen-Shannon Divergence (JSD).

II. Background

The application of ML for the detection and classification of faults has taken multiple and heterogeneous approaches in recent years. While either traditional distance protection [3], or more recent TW relays [4], provide reliable and consistent protection schemes for the transmission level, the intrinsic complexity of the distribution network requires more advanced techniques to protect distribution systems.

On the one hand, some approaches try to adapt legacy distance protection to the new challenges [5]. On the other hand, there is a trend to use faster time-domain methods, which usually depend on the detection and analysis of TWs, in order to find the fault location. The TWs are wide-band signals that propagate at almost the speed of light when a fault occurs [6]. In this regard, multiple approaches have been developed to get insightful information on those fast transients [1]. It has been shown that ML-based protection approaches can return an accurate fault location using a shorter window of time than more traditional approaches (as low as 100 μs) [7]–[9]. These papers use Random Forests (RFs), an algorithm is based on Decision Trees (DTs), to predict the fault location. This work revisits the method exposed in [7].

However, as ML algorithms become more complex, getting insights into their behavior and the data patterns that they learn

M. Jiménez-Aparicio, M. J. Reno, and F. Wilches-Bernal are with the Electric Power Systems Research Department at Sandia National Laboratories, Albuquerque, NM (email: mjimene@sandia.gov). This material is based upon work supported by the Laboratory Directed Research and Development program at Sandia National Laboratories and the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 36533.Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

978-1-6654-9921-7/22/$31.00 ©2022 IEEE

is much more difficult. There are several model-independent (agnostic) approaches to compute such feature importance, which vary in complexity and explanatory capabilities. This can be done either globally or by aiming just at individual samples. Some basic techniques are sensitivity analysis, Individual Conditional Expectation (ICE) or Permuted Feature Importance (PFI). Another option is to train a surrogate model [10]. Another technique, called Local Interpretable Model-agnostic Explanations (LIME), focuses on training interpretable models to approximate the model behavior for individual predictions [11]. However, there is one technique that stands above the rest for the amount of information that it retrieves: SHAP. This tool is further explained in Section III-C1 and is the one selected in this work. In power systems, there are just a handful of works that uses SHAP. Some works use SHAP for dimensionality reduction [12], but most of them use SHAP for feature importance calculation [13], [14].

III. The Method

A. Data processing and feature creation

Provided that the TWs' frequency spectrum varies as the wave propagates through the system, a time-frequency decomposition of the fault signals is justifiable for feature creation. The first step is the decomposition of the 3-phase voltages and currents into decoupled modes (one ground mode and two aerial modes) using the Karrenbauer Transform. Some interesting properties can be attributed to these modes as outlined in [15]: the ground mode is more sensitive to attenuation due to distance and seems especially suitable for Single-Line-to-Ground (SLG) faults. However, aerial modes are less prone to show the effect of attenuation and, in general, the fault signals are more energetic. In parallel, the TW arrival times are computed using the approach based on the dynamic mode decomposition presented in [16]. Each modal component is then fed into a time-frequency decomposition stage, based on the Stationary Wavelet Transform (SWT), to divide the signal in frequency bands, extracting the portion of the signal that is associated with each of them. In this regard, 6 decomposition levels to analyze the high-frequency part of the spectrum over 100kHz. The frequency ranges, for F_s equal to 10 MHz, are detailed in Table I. Finally, the Parseval's Energy (PE) of the decomposed voltage and current signals are calculated.

TABLE I: SWT Boundaries for Frequency Bands

Decomposition Level	Lower Frequency	Upper Frequency
1	2.5 MHz	5 MHz
2	1.25 MHz	2.5 MHz
3	625 kHz	1.25 MHz
4	312.5 kHz	625 kHz
5	156.25 kHz	312.5 kHz
6	78.125 kHz	156.25 kHz

B. Machine-Learning training

The selected algorithm is a RF, which is composed of several DTs. In each tree (or "estimator"), nodes are iteratively split into leafs in such a way that the purity of the resulting dataset division increases (i.e., the decision rules split the dataset better into classes). For this application, the RF selects energy values in the time-frequency decomposed signals to predict the faulty PZ. In order to study the fault type sensitivity to the employed decoupled modes, 3 different training approaches are compared in this work:

- A single RF model, trained only with the ground mode.
- A single RF model, trained with the 3 decoupled modes.
- 3 RF models, each one trained with the cases of a specific type of fault (which are SLG, LL or 3P). All of them are trained with 3 decoupled modes.

C. Interpretation of results

This work relies on SHAP for analyzing the classifier's behavior, and on the JSD to quantify such differences.

1) SHAP: It is an efficient implementation of the Shapley values theory, which was developed under the context of cooperative game theory as a method for finding out the players' contribution to a final payout [17], [18]. In this project, SHAP quantitatively determines how much the relay prediction would have changed depending on the values of the energy levels, which provides suggestions about how much is the relative importance of each level in a given decision. These results are especially insightful when these contributions are analyzed under different fault scenarios.

For a given prediction, SHAP studies what is the average contribution of each feature in the result under several coalitions of features. By coalition, it is meant for a certain subset of features. By assessing the predictions of each of the coalitions and averaging out the differences, the Shapley value for that level is calculated. The Shapley value ϕ for a certain feature j can be calculated using the following expression:

$$\phi_j = \frac{1}{M}\sum_{m=1}^{M}(f(x^m_{+j}) - f(x^m_{-j})), \tag{1}$$

where M is the total number of coalitions, x^m_{+j} is the subset of features including feature j, x^m_{-j} is the subset of features excluding feature j and $f(x)$ is the algorithm prediction function [19]. Note that using both sets of predictions, what is actually computed is the difference from the average prediction in the dataset when the feature j is considered.

2) Jensen-Shannon Divergence: This metric quantifies the similarity between two probability distributions. In information theory, the Shannon entropy $H(X)$ is the amount of information required to describe a variable X (which is inherently a probability distribution). The JSD comes from the mixture of the Jensen divergence and the Shannon entropy theorems. First, the Jensen divergence states that, being Ψ a given concave function and X random variable, the expected value of P, $\mathbb{E}(X)$, evaluated on Ψ, is larger than the expected value of $\Psi(X)$. Therefore:

$$\Psi(\mathbb{E}(X)) - \mathbb{E}(\Psi(X)) \geq 0. \tag{2}$$

In the JSD, the concave function Ψ becomes the Shannon entropy H. For 2 variables, X and Y, the JSD becomes:

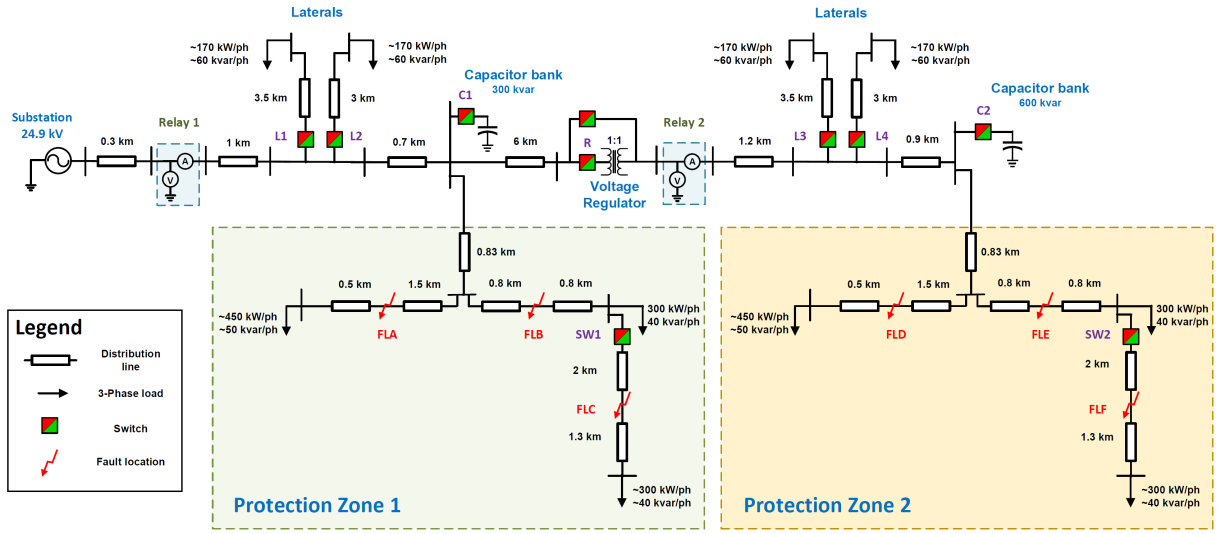


Fig. 1: Distribution system use case with relays and fault locations marked

$$JSD(X||Y) = H\left(\frac{X+Y}{2}\right) - \left(\frac{H(X)+H(Y)}{2}\right). \quad (3)$$

The JSD has 2 desirable properties: first, the result is bounded between 0 and 1 (the larger the mismatch, the larger the divergence) and, second, the result is symmetric (which means that $JSD(X||Y) = JSD(Y||X)$). In literature, the square root of the JSD is preferred over the base definition because it satisfies all the properties to be considered a metric of distance (or a "true" metric) [20]. Therefore, in this paper, the $\sqrt{JSD}$ will be the metric of comparison between the SHAP values distribution per energy level.

IV. The Use Case

The selected system aims to represent a simplified distribution system [15]. In order to provide additional variability to the data, several load scenarios and elements participation configurations are considered. Therefore, at the instant where the fault is produced, the number of laterals in the system, the presence of extra branches within the protection zones (PZs), and the presence of regulators and capacitor banks can be controlled by their respective switches. The combinations for the different elements can be observed in Tables II - V. The system is composed of 2 PZs (named "Protection Zone 1" or PZ1, and "Protection Zone 2" or PZ2). To make a fair comparison, the topology of both PZs is identical. The voltage and current data are recorded at the points labeled as "Relay 1" and "Relay 2" (R1 and R2, respectively), which are the locations of the considered protection devices. The sampling frequency is 10 MHz. The system is shown in Fig. 1.

The total number of simulated fault cases is 6,440. This considers 7 types of faults (3 types of SLG faults, 3 types of Line-to-Line (LL) faults and 3-Phase (3P) fault), 5 resistance values between 0.01 and 10 ohms, 6 fault locations (3 in each PZ), 3 load combinations regarding laterals participation, 3 load combinations regarding the presence of extra branches in the PZs, 2 combinations for the regulator (either present or not in the system), and 4 combinations for the capacitor banks (same, either present or not in the system). In total, for this system, 32 different scenarios are considered.

TABLE II: Laterals Combinations Showing the State of Each Switch as Either Open (0) or Closed (1)

Combination	L1	L2	L3	L4
A	0	1	0	0
B	0	1	1	0
C	0	0	1	0

TABLE III: PZ Extra Branches Combinations Showing the State of Each Switch as Either Open (0) or Closed (1)

Combination	SW1	SW2
A	0	1
B	1	0
C	1	1

TABLE IV: Regulator Combinations Showing the State of Each Switch as Either Open (0) or Closed (1)

Combination	R1
A	0
B	1

TABLE V: Capacitor Banks Combinations Showing the State of Each Switch as Either Open (0) or Closed (1)

Combination	C1	C2
A	0	0
B	0	1
C	1	0
D	1	1

In order to study the behavior of a TW-based fault location classifier, 2 experiments are considered:

1) Area Protection: The model is trained with data from R1. The goal is to classify if the fault occurred in PZ1 (next PZ) or PZ2 (somewhere downstream). The data is split in 80/20% for training and testing purposes, respectively.
2) Extrapolation to PZ2 R2: The same classifier is deployed in R2. Only the faults in PZ2 (3,220 cases) are considered, and they are used for testing. It is expected that all the faults are predicted to be on the next PZ. However, due to a different location, the measured energy values are not similar and the models' accuracy is worse. This experiment shows the behavior of a classifier for fault cases for which it is not trained for, and where the TW propagation factors have a larger effect on it.

V. Results

A. Area Protection and Extrapolation to PZ2 R2 Results

For the first experiment, the 3 approaches have a 100% accuracy, as can be seen in Table VI, which means that the 7 km line in between the PZs leads to a clear attenuation of the faults coming from PZ2 [15]. For the second experiment, the 1 model/ 3 modes approach gives the best overall performance. For this reason, and for the sake of simplicity, the rest of the section is based only on this approach.

TABLE VI: Models' Accuracies

Model	Overall	SLG	LL	3P
Area Protection				
1 model/ 1 mode	100%	100%	100%	100%
1 model/ 3 modes	100%	100%	100%	100%
3 models/ 3 modes	100%	100%	100%	100%
Extrapolation to PZ2 R2				
1 model/ 1 mode	61.0%	86.2%	39.9%	48.9%
1 model/ 3 modes	63.4%	81.2%	50.0%	50.0%
3 models/ 3 modes	62.8%	80.8%	49.3%	50.0%

In order to gather more insights into the wrong predictions, a summary of the underlying combinations for the previously described factors is shown in Fig. 2. Overall, it seems that the voltage regulator has a large influence, as all of the failures correspond to combinations in which the regulator was present between the zones. Secondly, comparing the 3 fault types, SLG fault cases are less prone to cause a wrong prediction (as was previously introduced in Table VI). Regarding other factors, the errors are evenly distributed between all the combinations, so they are not as relevant.

B. Model behavior analysis with SHAP values

SHAP is applied on a per fault case basis, and the output is an array of SHAP values of length equal to the number of features. Gathering the values for all the cases, it is possible to rank the features. The magnitude of a SHAP value is proportional to the feature's importance in the decision. Fig. 3 shows the 10 most important features (from top to bottom) in the 1 model/ 3 modes approach. The color indicates whether that feature was high or low for that particular sample.

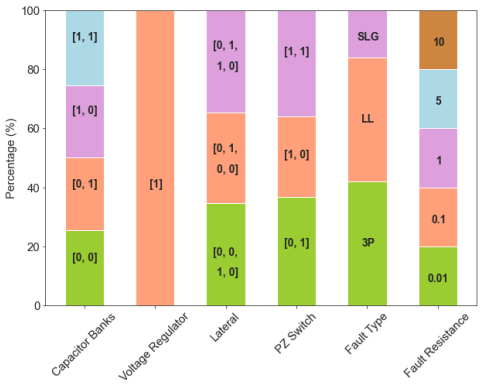


Fig. 2: Distribution of incorrect prediction per factor (for Extrapolation to PZ2 R2 of 1 model/ 3 modes approach)

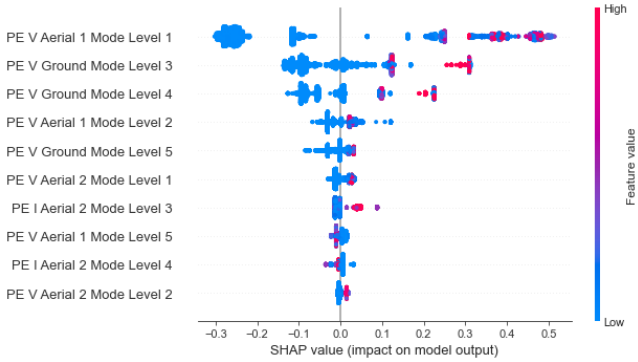
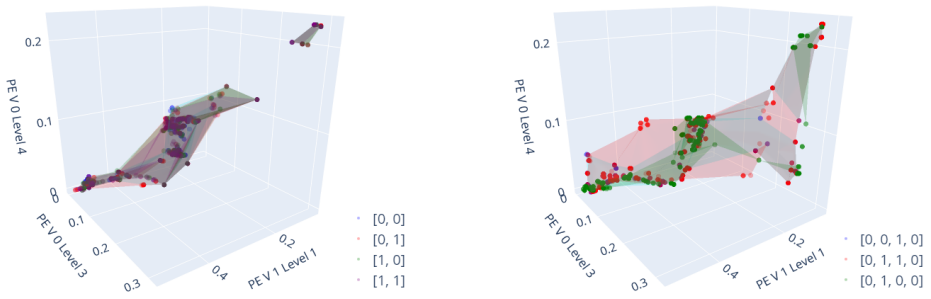


Fig. 3: Feature importance for the 1 model/ 3 modes according to SHAP using the training set's fault cases

If the 3 most important PE levels according to Fig. 3 are selected, the corresponding SHAP values form a set of spatial coordinates that can be visually analyzed. The rest of the subsection qualitatively analyzes the difference in behavior between a classifier that works correctly (as in the "Area Protection" experiment), and one that does not (as in the "Extrapolation to PZ2 R2" experiment). In order to analyze how relevant is the scenario in which the fault was produced, the different configuration cases are compared between themselves. It is considered that if the SHAP values point clouds are significantly different between 2 combinations of the same factor, therefore this factor is relevant for TW propagation because it requires a larger change in the classifier's behavior.

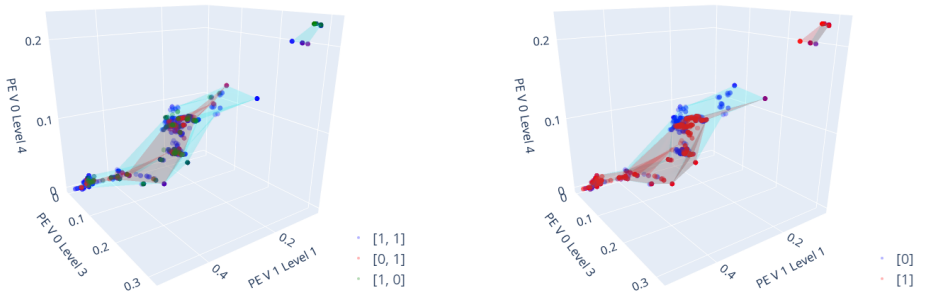
1) Area protection experiment: Fig. 4 show the behavior of a classifier that is able to predict the faulty PZ. Point clouds for different configurations are mostly overlapping. This classifier deals with faults under many different TW propagation conditions, which requires just a slight accommodation of the most important features to take into account in each case.

2) Extrapolation to PZ2 R2 experiment: The SHAP values between different combinations of the same factor tend to form different point clouds in Fig. 5, which shows how the classifier modifies its behavior to cope with slightly different input PE level values. The changes in behavior are especially noticeable



(a) Capacitor banks

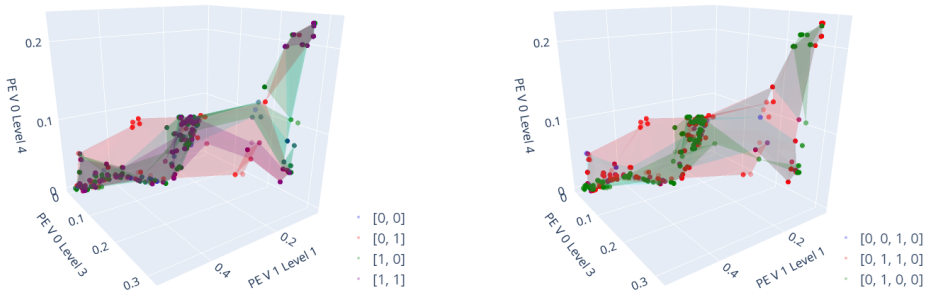
(b) Laterals



(c) PZ extra branch

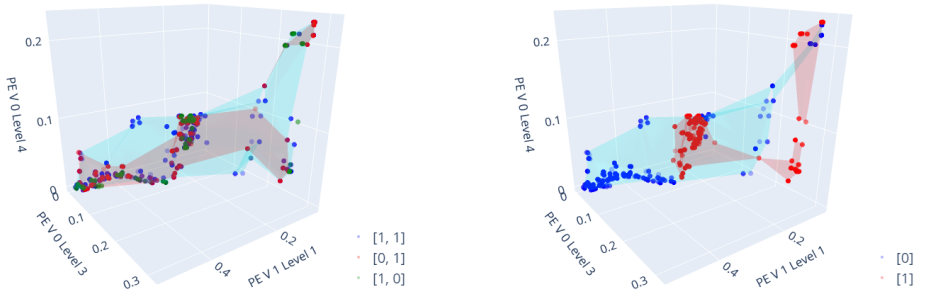
(d) Voltage regulator

Fig. 4: 3D analysis of SHAP distributions per combinations of each factor for PZ1 and PZ2 R1



(a) Capacitor banks

(b) Laterals



(c) PZ extra branch

(d) Voltage regulator

Fig. 5: 3D analysis of SHAP distributions per combinations of each factor for PZ2 R2

in the case of the voltage regulator, which leads to two clearly distinguishable point clouds.

C. Quantification using JSD

In order to quantify how similar is the behavior for different combinations, the pair-wise similarity between the distributions is calculated using the JSD.

TABLE VII: Top Divergences per Level PZ1 R1

PE Level	Factor	Comb. 1	Comb. 2	$\sqrt{\text{JSD}}$
V Aer. 1 Lev. 1	Lateral	[0, 0, 1, 0]	[0, 1, 0, 0]	0.772
V Gnd. Lev. 3	Lateral	[0, 0, 1, 0]	[0, 1, 1, 0]	0.606
V Gnd. Lev. 4	Lateral	[0, 0, 1, 0]	[0, 1, 0, 0]	0.526

TABLE VIII: Top Divergences per Level PZ2 R1

PE Level	Factor	Comb. 1	Comb. 2	$\sqrt{\text{JSD}}$
V Aer. 1 Lev. 1	Regulator	[1]	[0]	0.616
V Gnd. Lev. 3	Regulator	[1]	[0]	0.644
V Gnd. Lev. 4	Lateral	[0, 0, 1, 0]	[0, 1, 0, 0]	0.593

TABLE IX: Top Divergences per Level PZ2 R2

PE Level	Factor	Comb. 1	Comb. 2	$\sqrt{\text{JSD}}$
V Aer. 1 Lev. 1	Regulator	[1]	[0]	0.819
V Gnd. Lev. 3	Regulator	[1]	[0]	0.805
V Gnd. Lev. 4	Regulator	[1]	[0]	0.757

Tables VII-IX gather the top larger dissimilarities for each decomposition level and PZ and relay combination. The differences in the distributions in PZ1 R1 are mainly related to laterals or extra branches in the PZ. However, for PZ2 R1 and PZ2 R2, the voltage regulator is the most relevant factor that causes the large differences observed in the classifier's behavior. This matches the observations previously shown in Fig. 2. In addition, for PZ2 R1, the divergences are relatively smaller as faults are more attenuated due to the regulator.

VI. Conclusions

The behavior of a Traveling Wave (TW)-based fault location classifier has been analyzed, using SHapley Additive exPlanations (SHAP) and the Jensen-Shannon Divergence (JSD), under several TW propagation scenarios. SHAP provides the feature contributions that led to a prediction in each fault case. The difference in such contributions is quantified using the square root of the JSD for the 3 most important features. A small realistic system is employed to simplify the analysis. Faults are simulated according to several system configurations to enrich the classifier's behavior (varying the number of laterals, if the capacitor bank or the voltage regulator are switched on or off, and if the protection zones had the extra branch or not). A Random Forest model is trained to provide area protection, which shows that the attenuation due to distance is enough even in short distribution lines. The classifier's behavior is consistent for different fault and system configurations. When this model is extrapolated to a fairly similar task, there is a certain loss of accuracy in Line-to-Line and 3-Phase faults due to the voltage regulator effect. This is translated to erratic behavior, which can be appreciated both visually in the SHAP values distributions and on the largest JSD values in the most important energy levels. In conclusion, the proposed framework analyzes and quantifies how certain TW propagation factors disturbs the classifier's behavior.

References

[1] F. Wilches-Bernal et al., "A Survey of Traveling Wave Protection Schemes in Electric Power Systems," *IEEE Access*, vol. 9, pp. 72949–72969, 2021.

[2] C. Molnar, *Interpretable Machine Learning*.

[3] B. Kasztenny and D. Finney, "Fundamentals of Distance Protection," *2008 61st Annual Conference for Protective Relay Engineers*, 2008.

[4] E. O. Schweitzer, B. Kasztenny, A. Guzman, V. Skendzic, and M. V. Mynam, "Speed of line protection - can we break free of phasor limitations?," in *2015 68th Annual Conference for Protective Relay Engineers*, pp. 448–461, IEEE, mar 2015.

[5] F. Hariri and M. Crow, "New Infeed Correction Methods for Distance Protection in Distribution Systems," *Energies*, vol. 14, no. 15, 2021.

[6] A. Greenwood, *Electrical Transients in Power Systems*. New York, NY, USA: Wiley-Interscience, 2nd ed., 1991.

[7] M. Jiménez Aparicio, M. J. Reno, P. Barba, and A. Bidram, "Multi-Resolution Analysis Algorithm for Fast Fault Classification and Location in Distribution Systems," *9th International Conference on Smart Energy Grid Engineering (SEGE)*, 2021.

[8] F. Wilches-Bernal, M. Jiménez Aparicio, and M. J. Reno, "An Algorithm for Fast Fault Location and Classification Based on Mathematical Morphology and Machine Learning," in *2022 IEEE Innovative Smart Grid Technologies North America (ISGT NA)*, 2022.

[9] F. Wilches-Bernal, M. Jimenez-Aparicio, and M. J. Reno, "A Machine Learning-based Method using the Dynamic Mode Decomposition for Fault Location and Classification," in *Thirteenth Conference on Innovative Smart Grid Technologies (ISGT)*, 2022.

[10] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, p. 832, jul 2019.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 1135–1144, Association for Computing Machinery, 2016.

[12] M. Zhang, W. Chen, Y. Zhang, F. Liu, D. Yu, C. Zhang, and L. Gao, "Fault Diagnosis of Oil-Immersed Power Transformer Based on Difference-Mutation Brain Storm Optimized Catboost Model," *IEEE Access*, vol. 9, pp. 168767–168782, 2021.

[13] K. Zhang, P. Xu, and J. Zhang, "Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control," in *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, pp. 711–716, 2020.

[14] V. Hoffmann, J. R. A. Klemets, B. N. Torsæter, G. H. Rosenlund, and C. A. Andresen, "The value of multiple data sources in machine learning models for power system event prediction," in *2021 International Conference on Smart Energy Systems and Technologies (SEST)*, pp. 1–6, 2021.

[15] M. Jiménez-Aparicio, M. J. Reno, and F. Wilches-Bernal, "Traveling Wave Energy Analysis of Faults on Power Distribution Systems," *Energies*, vol. 15, no. 8, 2022.

[16] F. Wilches-Bernal, M. J. Reno, and J. Hernandez-Alvidrez, "A Dynamic Mode Decomposition Scheme to Analyze Power Quality Events," *IEEE Access*, vol. 9, pp. 70775–70788, 2021.

[17] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.0, 2017.

[18] L. S. Shapley, "17. A Value for n-Person Games," in *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318, Princeton University Press, dec 1953.

[19] E. and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, pp. 647–665, 2013.

[20] T. M. Osán, D. G. Bussandri, and P. W. Lamberti, "Monoparametric family of metrics derived from classical Jensen–Shannon divergence," *Physica A: Statistical Mechanics and its Applications*, vol. 495, pp. 336–344, 2018.