Fault Classification in Active Distribution Systems using Explainable Convolutional Neural Networks

Mohammadreza Mirjafari[1], Peter B. Idowu[1], Md Habib Ullah[1],
Oluwadamilola Ajayi[2], and Akinwale Stephen Akintaro[1]
[1]Penn State Harrisburg, Middletown, PA, USA; [2]Schweitzer Engineering Laboratories, TX, USA
Emails: mzm7143@psu.edu, pbi1@psu.edu, mzu5071@psu.edu, damiajayit@gmail.com, asa6322@psu.edu

***Abstract*—One of the critical aspects of outage management in distribution networks is fault classification, which has become more complicated with the increasing penetration of distributed energy resources (DERs). Identifying the occurrence of a fault and then determining its type leads to more efficient system restoration, minimizing downtime and resource costs. This paper proposes a deep learning architecture based on convolutional neural networks (CNNs) for fault type classification in modern distribution systems. While CNNs offer strong capabilities for learning complex patterns and achieving high classification accuracy, their lack of interpretability remains a major concern in power system applications. To provide transparency into the predictions of the proposed CNN-based model, DeepSHAP is used as a post-hoc explainability method. This tool enables visualization of the contribution of features to the model's predictions, offering valuable insights to interpret the decisions made by the proposed model. In our experimental evaluation, the proposed deep learning model achieves 99.4% accuracy and demonstrates robust performance under 30% penetration of DER.**

***Index Terms*—Power distribution networks, Fault type classification, Convolutional Neural Networks, and Explainable Artificial Intelligence.**

I. INTRODUCTION

Distribution networks are responsible for delivering reliable and continuous electrical power to consumers. As a critical part of the power system, these networks with many branches and laterals are prone to different types of faults that may lead to power outages. Detecting and classifying fault types in distribution networks is an essential task that leads to more efficient system restoration, reducing both time and resource requirements. This task is also necessary for later fault analysis techniques, such as impedance-based fault localization, which require accurate fault type information to function effectively [1], [2].

Advances in monitoring devices and data-driven techniques have propelled a growing shift from conventional fault analysis methods toward artificial intelligence (AI)-based approaches in distribution systems [3]. AI-based methods can generally be classified into two main categories: machine learning and deep learning approaches. In recent years, traditional machine learning methods, such as tree-based techniques [2], support vector machines (SVM) [4], and artificial neural networks (ANN) [5], have been replaced by advanced deep learning techniques. This transition is mainly due to the strong capabilities of deep learning models, particularly CNNs, which are highly effective in extracting complex features directly from raw data.

In [6], a CNN-based approach is introduced for fault classification in distribution networks that include distributed generation (DG) units. The method uses raw three-phase voltage and current signals to classify fault types and normal conditions. In [7], a hybrid deep learning model is proposed for the fault classification in a grid integrated with DG units. Their proposed hybrid model combines a fully convolutional network (FCN) and a bidirectional long short-term memory (BiLSTM). In their study, raw voltage and current data from simulated fault events are used as input to the proposed model. In [8], an LSTM-DenseNet network is introduced for fault diagnosis and classification in power distribution networks. This method fuses two neural network architectures: the DenseNet component detects subtle variations in three-phase voltage and zero-sequence current, and the LSTM model captures crucial time-domain features from fault signals.

Although current studies have reached remarkable success and high accuracy in fault analysis of distribution networks, the interpretability of these models remains a significant challenge. As a result, there is a growing need for explainable AI (XAI) approaches to provide transparency for model decisions. Explainable AI techniques generally provide explanations at two different levels of scope: local and global. The local explanation focuses on a single data instance. The goal is to understand why the model made a specific decision for that particular input. On the other hand, global explanation aims to provide insights into the overall behavior of a model. This is typically achieved by analyzing a collection of inputs and generating an explanation map that reflects general model behavior.

In addition to the scope of explanations, XAI methods can also be broadly categorized based on how they are integrated into machine learning models. The first category is the intrinsic approach, in which the model is inherently interpretable by design, such as decision trees. The second category is the post-hoc approach, where explainability is applied after the model has been trained. This is particularly useful for complex models such as deep neural networks, where explanations are generated without modifying the model's architecture or training process [9]. The post-hoc approach can be further divided into two subtypes. Model-specific methods, such as Grad-CAM, need access to the model's internal structure to provide more accurate explanations. In contrast, model-agnostic methods like LIME can be applied to any trained model, regardless of its underlying structure [10].

979-8-3315-8138-1

To address the growing need for explainable deep learning models capable of accurate and robust fault classification in distribution networks with DERs, this paper introduces a one-dimensional CNN model integrated with DeepSHAP. Compared to 2D CNNs, 1D CNNs offer a more efficient solution for directly analyzing 1D signals, such as voltage and current waveforms, with fewer parameters and reduced computational complexity [11]. The main contributions of this paper are as follows:

- A new deep learning model based on 1D CNNs is proposed for fault classification in modern distribution systems, achieving 99.4% accuracy and demonstrating high reliability and precision.
- To enable the interpretability of the proposed model, we employ DeepSHAP, which visualizes the contribution of each feature to the model's predictions.
- A Typhoon HIL–OpenDSS co-simulation framework is developed to generate a realistic dataset that captures the impact of DERs, including photovoltaic (PV) and battery systems. Using this dataset from a real-time simulation, the robustness of the proposed model is examined at 30% DER penetration.

The remainder of this paper is organized as follows. Section II discusses the methodology, including data generation and preprocessing, description of the proposed model, and explainable AI. In Section III, the performance of the proposed model is analyzed with several tests, with special attention to DER penetration and explainability. Finally, the conclusions are discussed in Section IV.

II. Methodology

In this study, a one-dimensional CNN approach is proposed for effective fault classification in active distribution networks integrated with DERs. To overcome the black-box nature of the proposed deep learning model and provide interpretability, DeepSHAP is applied to visualize the impact of individual features on the model's predictions. The proposed methodology is structured into four main phases: test system simulation and data generation, data preprocessing, implementation of the proposed CNN-based model, and explainable AI using DeepSHAP. The following sections provide a detailed explanation of each step in this methodology.

A. Data Generation

Distribution systems are susceptible to various fault types that can be broadly classified as symmetrical (balanced) or asymmetrical (unbalanced). Typical asymmetrical faults include single line-to-ground (LG), line-to-line (LL), and double line-to-ground (LLG) faults, while symmetrical faults involve three-phase (LLL) and three-phase-to-ground (LLLG) faults. Among these, LG faults are the most common, accounting for approximately 70% of occurrences, but are generally less severe, whereas three-phase faults are relatively rare, around 5%, but cause the highest damage [12].

In this work, the main fault dataset utilized for training and validation of our proposed 1D-CNN model is generated via co-simulation of MATLAB and OpenDSS, covering all fault

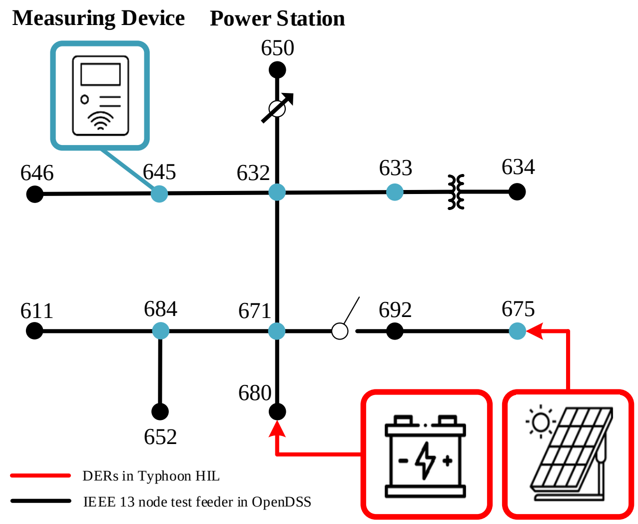

Fig. 1: Test system under study

TABLE I: Distribution of Samples Across Normal and Fault Classes

Fault type	Phases involved	Class	Samples
No fault	-	T0	8760
Line to ground	AG	T1	624
	BG	T2	624
	CG	T3	696
Line to line	AB	T4	552
	AC	T5	552
	BC	T6	624
Double line to ground	ABG	T7	552
	ACG	T8	588
	BCG	T9	624
Three phase	ABC	T10	552
Three phase to ground	ABCG	T11	552

types. In this framework, the three-phase voltage and current are measured using phasor measuring units (PMUs) before and during the fault, considering the inherent uncertainties in load and fault information. The test system in this study is the IEEE 13-node test feeder, operated at 4.16 kV, which can be used to test the features of the distribution networks. Due to cost limitations in placing PMUs at every bus, it is considered that only six PMUs would be installed at nodes 632, 633, 645, 671, 675, and 684, as shown in Fig. 1 [13]. Table I lists the total number of samples, including normal system operation as well as all fault classes.

The impact of DERs, which are a pivotal part of the modern grids, is not considered in our main dataset and the proposed CNN-based model is initially trained and tested with a dataset from the base network. This is because considering all possible DER operating scenarios during the training of AI-based methods is practically infeasible. However, to examine the robustness of our proposed model under DER penetration, a limited dataset is generated using the co-simulation of Typhoon HIL and OpenDSS. This framework accurately captures the influence of DERs on fault transients and associated measurement signals through high-fidelity real-time simulation in Typhoon HIL.

979-8-3315-8138-1

B. Data Preprocessing

Data preprocessing techniques are applied to the generated fault dataset to prepare it for input to the proposed model. Firstly, we use data splitting to separate the dataset into training, validation, and test sets. Secondly, a data augmentation approach called synthetic minority oversampling technique (SMOTE) is utilized to balance the dataset, preventing bias and ensuring reliable model performance [14]. This technique is applied to only the training dataset to prevent data leakage and optimistic results. Afterward, data normalization is performed to align the voltage and current scales to a common range, ensuring that these measurements contribute equally to the analysis. Finally, the categorical target variables are converted into a one-hot encoded format, helping the model to learn categorical data effectively.

C. Proposed CNN Model

As shown in Fig. 2, after preprocessing the generated fault dataset, it is fed into the proposed 1D CNN, which consists of three convolutional layers with rectified linear unit (ReLU) activation, followed by batch normalization and a pooling layer. Afterward, the outputs of these modules are flattened and fed into fully connected layers to extract information, followed by a dropout layer to prevent overfitting. Finally, a softmax activation is used in the output layer to make the final prediction.

D. Explainable AI

Although deep learning models can achieve high accuracy for fault analysis in power systems, their decision-making process is often not transparent to engineers and operators. In this paper, we use DeepSHAP to provide post-hoc interpretability for our proposed deep learning model. SHapley Additive exPlanations (SHAP) utilizes principles from game theory to interpret model predictions by attributing contributions to each feature. The SHAP explanation model can be formulated as follows [15]:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where M is the number of input features, ϕ_0 represents the base value of the model output, and $z' \in \{0,1\}^M$ demonstrates the presence or absence of features. ϕ_i is the Shapley value defined as follows:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{(|z'|!)\,(M - |z'| - 1)!}{M!} \times \left[f_x(z') - f_x(z' \setminus i)\right], \tag{2}$$

where the term in brackets represents the marginal contribution of feature i, quantifying the change in the model output when feature i is added to a given subset of features relative to the same subset without feature i. SHAP values are the unique additive feature attributions that satisfy the properties of local accuracy, missingness, and consistency [15].

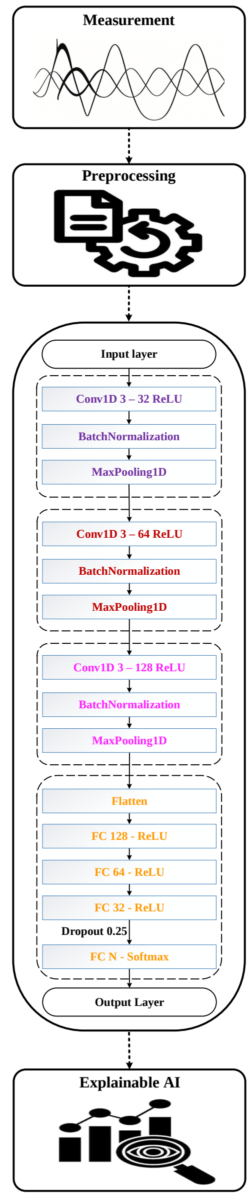

Fig. 2: Proposed explainable 1D CNN framework for fault type classification

While SHAP provides a theoretically rigorous measure of feature contributions, computing exact values for deep neural networks is computationally expensive. DeepSHAP overcomes this by leveraging the compositional structure of neural networks: it calculates SHAP values for simple network components and propagates their contributions through the entire network using multipliers and a chain rule inspired by DeepLIFT. The multipliers for the intermediate variable x_j relative to the output f, and the input feature y_i relative to the intermediate function g_j, are defined as follows, respectively:

$$m_{x_j f} = \frac{\phi_j(f,x)}{x_j - E[x_j]}, \tag{3}$$

$$m_{y_i g_j} = \frac{\phi_i(g_j,y)}{y_i - E[y_i]}, \tag{4}$$

979-8-3315-8138-1

TABLE II: Comparison of the Proposed Model with Baseline Methods

Metrics	CNN	ANN	SVM	DT
Accuracy	0.994	0.982	0.927	0.908
Precision	0.991	0.972	0.941	0.952
Recall	0.987	0.977	0.927	0.908
F1-score	0.989	0.974	0.924	0.916

where $E[x_j]$ and $E[y_i]$ represent the expected baseline values. The multipliers are propagated through the network via a chain rule that aggregates contributions along all computational paths. This rule is defined as follows:

$$m_{y_i f} = \sum_{j=1}^{n} m_{y_i g_j} \cdot m_{x_j f}. \tag{5}$$

Finally, the SHAP value is derived using a linear approximation as follows [15], [16]:

$$\phi_i(f, y) \approx m_{y_i f} \cdot (y_i - E[y_i]). \tag{6}$$

III. Experiments and Results

In this section, the performance of the proposed 1D CNN model is evaluated through a series of experiments. Next, the explainability of the proposed model is analyzed in detail. The proposed 1D-CNN model contains 108,332 parameters with a total model size of 423.17 KB. The model was trained for 20 epochs in 37.43 seconds on Google Colab.

A. Performance Evaluation of the Proposed CNN Model

To evaluate the performance of our proposed CNN-based model, we first use a confusion matrix to compare predicted and actual values. There are a total of twelve classes, including one representing the normal system and eleven corresponding to different fault types. As shown in Fig. 3, the proposed method demonstrates strong performance on the imbalanced test dataset, producing only a few misclassifications, as indicated by the off-diagonal elements of the normalized confusion matrix. Following the confusion matrix, an accuracy and loss graph is used to monitor trends in both metrics during training. As shown in Fig. 4, the proposed 1D CNN model demonstrates stable performance on the training and validation datasets, with no signs of overfitting or underfitting. The accuracy and loss on the test dataset are 0.994 and 0.019, respectively, confirming the strong performance of the proposed model. Furthermore, the proposed model's performance is evaluated against the ANN, SVM, and decision tree methods. As shown in Table II, the proposed model outperforms the others on the test dataset, achieving higher accuracy and superior precision, recall, and F1-score.

To evaluate the proposed CNN model's performance under DER penetration, including PV and battery, we generated a fault dataset with 100 samples using a co-simulation of Typhoon HIL and OpenDSS. In this framework, the IEEE 13-node test feeder is modeled in OpenDSS with a quasi-static simulation, and PV and battery are modeled in the EMT-based Typhoon HIL environment. Compared to modeling the entire system in the EMT domain, this framework enables faster simulations with high-fidelity observation of the characteristics

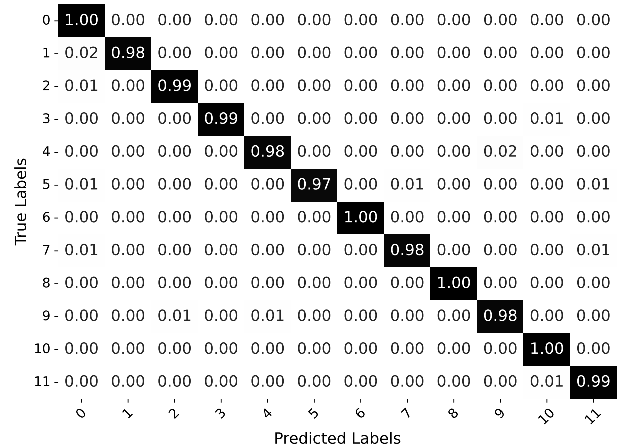

Fig. 3: Normalized confusion matrix

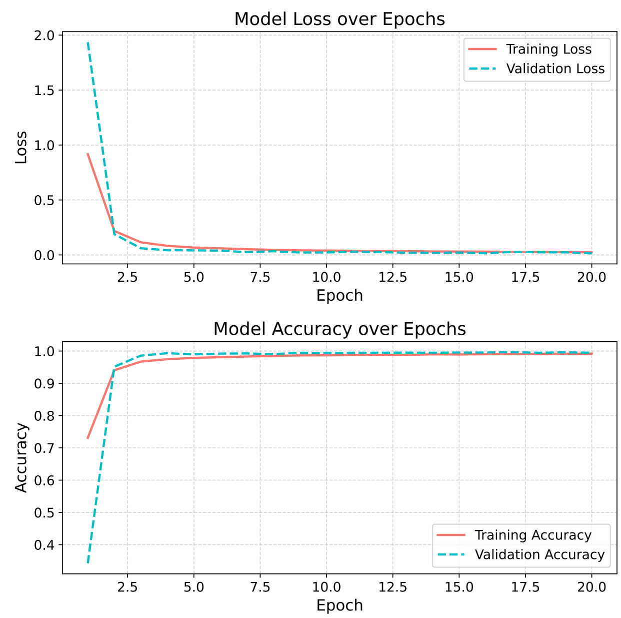

Fig. 4: Accuracy and loss curves of the proposed fault classification model

of DERs in Typhoon HIL. At 30% DER penetration, the proposed 1D CNN model demonstrated robust performance, misclassifying only 3 samples across classes 8, 10, and 11, and achieving an overall accuracy of 97%.

B. Model Explainability

In this study, we employ DeepSHAP waterfall plot to provide local explainability for our proposed CNN-based model by quantifying the contribution of each input feature to individual predictions. This analysis highlights the contributions of voltage and current measurements from different PMUs with particular attention to the role of each individual phase. In the waterfall plot, red bars depict features with positive SHAP values that raise the model's prediction, while blue bars illustrate features with negative SHAP values that reduce it. The length of each bar indicates the magnitude of the corresponding feature's impact.

979-8-3315-8138-1

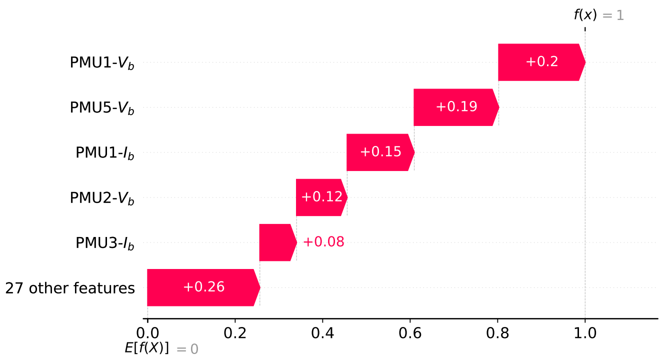

Fig. 5: Waterfall plot of feature contributions for the BG fault

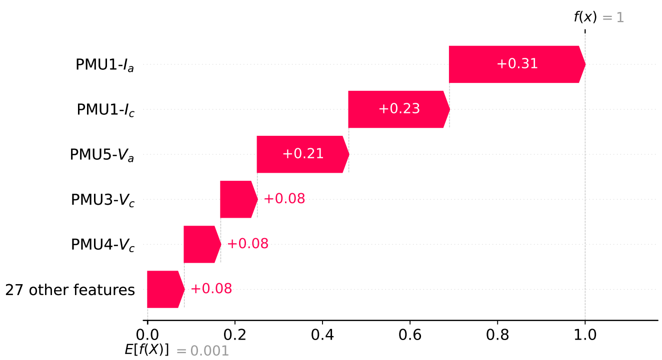

Fig. 6: Waterfall plot of feature contributions for the AC fault

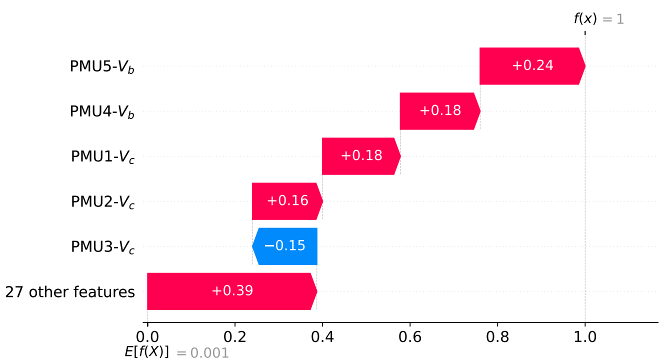

Fig. 7: Waterfall plot of feature contributions for the BCG fault

Fig. 5 illustrates the model's explainability for an LG fault. Since a phase B-to-ground fault causes voltage drop and overcurrent in this phase, the results align with expectations, with voltage and current measurements from phase B identified as the most influential features. Fig. 6 shows the waterfall plot explaining the prediction of an AC fault. As can be seen, this figure highlights a greater influence of voltage and current measurements from phases A and C that aligns with our expectations. Fig. 7 shows the model's explainability for a BCG fault. The results emphasize the contribution of voltage measurements from phases B and C, consistent with expected fault behavior.

IV. Conclusion

This paper has introduced a deep learning model based on 1D CNNs for fault classification in active distribution networks. The DeepSHAP waterfall plot was used to interpret the predictions of the proposed CNN-based model, helping to mitigate the black-box nature of deep learning methods. This tool visualizes the contribution of features to the model's predictions, providing a clearer understanding of its decision-making process. The results showed that the proposed CNN-based model achieves 99.4% accuracy in fault classification and demonstrates robust performance under DER penetration.

Acknowledgement

This work is supported in part by the Appalachian Regional Commission (ARC) [grant MU-21579-302-23, 2023].

References

[1] P. Stefanidou-Voziki, N. Sapountzoglou, B. Raison, and J. L. Dominguez-Garcia, "A review of fault location and classification methods in distribution grids," *Electric Power Systems Research*, vol. 209, p. 108031, 2022.

[2] O. Ajayi, M. Mirjafari, P. B. Idowu, and M. H. Ullah, "Explainable ai for fault detection and classification in microgrids," in *2024 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2024, pp. 1835–1840.

[3] M. Gilanifar, H. Wang, J. Cordova, E. E. Ozguven, T. I. Strasser, and R. Arghandeh, "Fault classification in power distribution systems based on limited labeled data using multi-task latent structure learning," *Sustainable Cities and Society*, vol. 73, p. 103094, 2021.

[4] L. J. Awalin, K. Naidu, and H. Suyono, "Fault types classification using support vector machine (svm)," in *AIP Conference Proceedings*, vol. 2129, no. 1. AIP Publishing LLC, 2019, p. 020132.

[5] A. Onaolapo and K. Akindeji, "Application of artificial neural network for fault recognition and classification in distribution network," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. IEEE, 2019, pp. 299–304.

[6] P. Rai, N. D. Londhe, and R. Raj, "Fault classification in power system distribution network integrated with distributed generators using cnn," *Electric Power Systems Research*, vol. 192, p. 106914, 2021.

[7] P. K. Bais, N. D. Londhe, K. Chintapalli, V. Nallana, and S. Mittal, "Fault classification using hybrid fcn-bilstm model in grid integrated with distributed generators," in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*. IEEE, 2024, pp. 1–6.

[8] L. Ji, X. Tian, Z. Wei, and D. Zhu, "Intelligent fault diagnosis in power distribution networks using lstm-densenet network," *Electric Power Systems Research*, vol. 239, p. 111202, 2025.

[9] R. Machlev, L. Heistrene, M. Perl, K. Y. Levy, J. Belikov, S. Mannor, and Y. Levron, "Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities," *Energy and AI*, vol. 9, p. 100169, 2022.

[10] K. Devireddy, "A comparative study of explainable ai methods: Model-agnostic vs. model-specific approaches," *arXiv preprint arXiv:2504.04276*, 2025.

[11] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[12] K. V. S. M. Babu, S. Nayak, D. Dwivedi, P. Chakraborty, C. N. Bhende, P. K. Yemula, and M. Pal, "Efficient fault detection and categorization in electrical distribution systems using hessian locally linear embedding on measurement data," *Measurement: Energy*, vol. 5, p. 100035, 2025.

[13] M. Spitzer, C. Eerdmans, A. S. Nair, and P. Ranganathan, "Evaluation of pmu placements with sori and orc indices for ieee test feeders," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*. IEEE, 2018, pp. 0687–0690.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] F. Eilers and X. Jiang, "Deepcshap: Utilizing shapley values to explain deep complex-valued neural networks," *arXiv preprint arXiv:2403.08428*, 2024.

979-8-3315-8138-1