Analysis of Interpretability and Generalizability for Power Converter Fault Diagnosis Based on Temporal Convolutional Networks

Tongyang Ren, Tao Han, Qun Guo, and Gang Li, *Member, IEEE*

***Abstract*— Despite the success of data-driven converter fault diagnosis methods, interpretability and generalizability limit the further promotion of data-driven methods in industrial applications. Therefore, to improve the accuracy in face of out-of-distribution problems and increase the confidence of power converter fault diagnosis, it is essential to understand the change and decision mechanism inside the deep model. First, we construct a general temporal convolutional network (TCN) to visualize the diagnostic process, which has been proven effective in power converter fault diagnosis. Then, the effect of hyperparameters on generalizability is analyzed under typical power converter disturbances. Finally, the concern area of the model for the current in the fault decision is interpreted intuitively by gradient-weighted class activation mapping (Grad-CAM), and the feature maps generated by the different channels are analyzed from multiple perspectives. The visualization results help to understand the complex structure of neural networks and can support the design of model to improve generalizability.**

***Index Terms*— Fault diagnosis, generalizability, interpretability, power converter, temporal convolutional network (TCN).**

I. Introduction

NOWADAYS, electrical energy has become an essential motor of technology and economy. Energy conversion is an indispensable part of the production, storage, and utilization of electrical energy, which relies on power converters to achieve [1]. At the same time, power converters play an important role in applications, such as electric vehicles, industrial manufacturing, and aerospace systems [2], [3], [4], [5]. However, power converters are required to operate in harsh environments for a long time and frequently switch between operating states, which challenges the reliability of the system. Efficient and accurate state detection, fault diagnosis, and fault-tolerant control methods for power converters are widely required [6]. In fact, many literatures have tried to explore this field, which has established a relatively well-developed knowledge system for fault monitoring and diagnosis of power converters [7], [8], [9], [10].

So far, the mainstream methods for power inverter fault diagnosis are model-based methods and data-driven methods. Model-based methods (observers or estimators) usually develop an accurate mathematical model based on the system parameters and compare the measured values with the output generated by the model to diagnose faults [11]. As electrical and electronic systems become more complex, traditional model-based diagnostic methods have difficulty achieving competitive results [12], [13]. In addition, with the rapid development of smart sensors, the Internet-of-Things, and high-speed communication, data from different lifecycles of power electronic converter systems can be collected and stored, which brings potential possibilities for data-driven fault diagnosis methods to face complex and challenging scenarios. Moradzadeh et al. [14] review the application of data mining techniques in fault diagnosis of power electronics system in recent years and point out that the data-driven approach is an ideal application tool in the power electronics industry in the future.

Shallow machine learning methods were first introduced for power converter fault diagnosis and are often combined with signal processing methods. Xia et al. [15] used fast Fourier transform and ReliefF algorithm to select most correlated features, and hybrid ensemble learning scheme combining extreme learning machine and random vector functional link network is proposed for extracting mapping relationship between fault modes and the selected features. Cai et al. [16] used principal component analysis to reduce the dimensions of data and Bayesian networks to diagnose fault. Recently, deep learning has swept other fields, such as machinery fault diagnosis, due to its powerful nonlinear mapping capabilities, and many studies have attempted to explore the role of deep learning in power converter fault diagnosis. Yating et al. [17] first introduced temporal convolutional network (TCN) to the three-phase converter diagnostics and extend it to the diagnosis of unknown faults. Guo et al. [18] used adaptive chirp mode decomposition to reconstruct the subsignal from the fault current and TCN is used to extract features from the subsignal to locate the fault of modular multilevel converters. Zhang et al. [19] improved the first-layer convolution kernel

Manuscript received 9 March 2023; revised 21 April 2023; accepted 8 May 2023. Date of publication 29 May 2023; date of current version 12 June 2023. This work was supported by the Natural Science Foundation of Jilin Province, China, under Grant 20220101171JC. The Associate Editor coordinating the review process was Dr. Gabriele Patrizi. *(Corresponding authors: Qun Guo; Gang Li.)*

Tongyang Ren is with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China.

Tao Han and Qun Guo are with the College of Instrumentation and Electrical Engineering, Jilin University, Changchun 130012, China (e-mail: guoqun6512@foxmail.com).

Gang Li is with the Key Laboratory of Geophysical Exploration Equipment, Ministry of Education, and the College of Instrumentation and Electrical Engineering, Jilin University, Changchun 130012, China (e-mail: ligang2013@jlu.edu.cn).

Digital Object Identifier 10.1109/TIM.2023.3280535

and pooling layer of the 1-D convolutional neural network (CNN) and apply it to three-phase voltage-source inverters. Zhao et al. [20] proposed a multibranch 1-D CNN combining the soft threshold and attention mechanism for the fault diagnosis problem of the dual active bridge (DAB) converter.

The above work demonstrates that deep learning is promising for fault diagnosis of converters, which provides a new perspective for robust fault diagnosis of complex electrical systems. However, the essence of deep learning is still a black-box approach, which means that the mechanisms inside the model cannot be understood although the optimization direction of the model can be constrained by different regularization methods [21]. In fact, the black-box approach has hindered the further application of deep learning methods. On the one hand, the uninterpretability of the model is crucial for the reliability of the diagnostic method [22]. Incomprehensible decision mechanisms may undermine user trust in application scenarios that require high credibility. On the other hand, understanding the decision mechanism helps to improve the generalization ability of the method, which is the most key prerequisite for realizing industrial applications. Although data-driven fault diagnosis methods have been successful in the field of power electronics for specific converters and specific operating conditions, the performance of the method inevitably degrades under unknown environments and conditions [23]. Therefore, research on the generalizability and interpretability of power converter fault diagnosis based on deep learning is essential. In fact, a number of interpretable and generalizable studies have been reported in the field of mechanical fault diagnosis [24]. Li et al. [25] proposed a novel wavelet-driven deep neural network WaveletKernelNet with the continuous wavelet convolutional (CWConv) layer to help the CNNs discover kernels with a certain physical meaning. An et al. [26] proposed an interpretable neural network to provide high-performance and credible mechanical fault diagnosis results, which is mainly generated by unrolling the nested iterative soft thresholding algorithm for a sparse coding model. Zhu et al. [27] embed feature information and temporal information into the model and explain the model's decision through the gradient-based approach. However, few similar studies have been reported in the power electronics fault diagnosis field so far.

In power converter fault diagnosis, the current data in normal conditions show a high degree of periodicity, while the current variation during faults is usually related to the parameters of the circuit system. Therefore, it is worth investigating whether the hyperparameter values associated with the design of the model structure change under different operating conditions for a particular type of power inverter. In this article, we work on analyzing the influence of hyperparameters related to the model structure on the generalizability and visualizing the black box model of the general diagnostic method to understand the internal working mechanism. We use TCN as the basic model to be visualized due to the successful application of TCN in power converter fault diagnosis. Analyzing the impact of general model structural design on power electronics fault diagnosis through visualization rather than model performance is a novel perspective and can show interesting findings. The contributions of this article are as follows.

1) The typical uncertainty factors for three-phase inverters in industrial applications are considered for the first time, while the impact of the hyperparameters related to the model structure on the model performance is evaluated under different uncertainty factors.
2) First attempts to analyze the decision mechanism of the model by gradient-weighted class activation mapping (Grad-CAM) in the field of power converter fault diagnosis and interpret the influence of batch normalization (BN) on the concern region from a visual perspective.
3) Adaptively select important channels via the efficient channel attention (ECA) module while visualizing the feature maps of different residual blocks within the model and the high-dimensional data distribution generated by different channels, which helps to understand the role of the different channels of the model in decision making.

The rest of this article is organized as follows. First, some basic definitions and concepts are introduced in Section II. Then, the converter fault diagnosis process for visualization is presented in Section III. Generalizability and interpretability analysis based on practical applications are introduced in Section IV. Finally, Section V concludes this article.

II. Preliminaries

A. Fault Definition for Three-Phase Converters

Several surveys and researches have shown that capacitors and switching devices are the most fragile parts of power converters. Many factors affect the performance of switching devices, such as overvoltage, overcurrent, temperature, mechanical stress, and so on. The main fault forms in switching devices are open-circuit (OC) faults and short-circuit (SC) faults. SC faults occur quickly and have a short duration, which makes fault location, detection, and isolation by software very difficult. Due to the very obvious fault characteristics, SC faults are usually detected and isolated quickly by hardware protection circuits. Isolated SC faults can be regarded as OC faults. However, the current or voltage distortion caused by an OC fault is not obvious. Short-term operation of a system with an OC fault can damage the performance of the converter, and long-term operation can even damage other components in the system. Therefore, OC fault diagnosis methods for power converters have been an important research topic.

A schematic of a typical three-phase inverter is shown in Fig. 1, which contains six insulated gate bipolar transistors (IGBT). As shown in Fig. 1, each IGBT OC fault or normal operating condition can be considered as a category and the fault diagnosis of a three-phase converter can be formalized as a classification problem. The seven categories include IGBT1 OC faults, IGBT2 OC faults, IGBT3 OC faults, IGBT4 OC faults, IGBT5 OC faults, IGBT6 OC faults, and normal operating state. The relationship between operating status and fault label is shown in Table I. For a three-phase converter, an IGBT OC fault on any one bridge arm will also affect the phase currents of the other bridge arms. Thus, only phase A

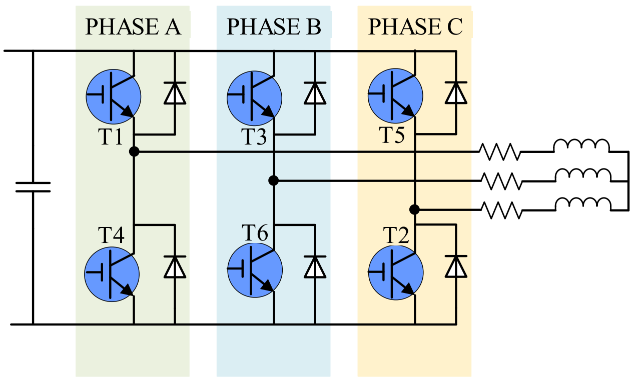


Fig. 1. Schematic of the three-phase converter.

TABLE I
RELATIONSHIP BETWEEN OPERATING STATUS AND FAULT LABEL

Operation Status	Label	Operation Status	Label
Normal	0	IGBT1 OC Fault	1
IGBT2 OC Fault	2	IGBT3 OC Fault	3
IGBT4 OC Fault	4	IGBT5 OC Fault	5
IGBT6 OC Fault	6		

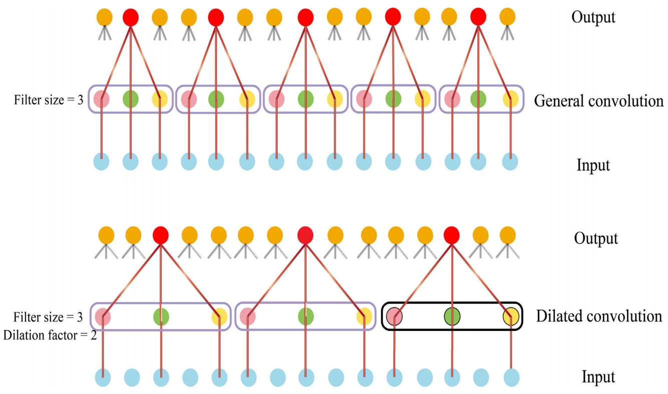


Fig. 2. Schematic of dilation convolution.

currents are collected as sample data in [17] and [19]. In this article, we try to analyze the link between model design and generalizability and interpretability.

1) Dilated Convolution: In contrast to the regular convolutional layer in CNN, the dilated convolution has a larger receptive field in TCN, which is more effective for converters fault diagnosis with long-time sequence inputs. The comparison between regular convolution and dilated convolution is shown in Fig. 2.

For the input sequence $x_0, x_1, \ldots, x_i$, the general formulation of the dilated convolution can be given as follows:

$$F(s) = \sum_{i=0}^{k-1} f(i) \cdot x(s - d \cdot i) \tag{1}$$

where f is the convolution kernel, k is the kernel size, and d is the dilation factor. Specifically, the larger d is, the more "holes" in the convolution kernel, which also means that the receptive field expands exponentially as the number of layers in the network deepens. Therefore, due to the periodic characteristics of current or voltage data, fault diagnosis of power converters can benefit from dilated convolution with fixed parameters.

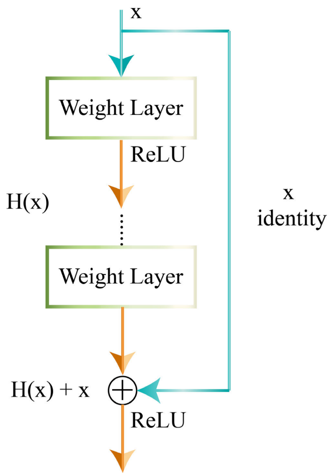


Fig. 3. Skip connection.

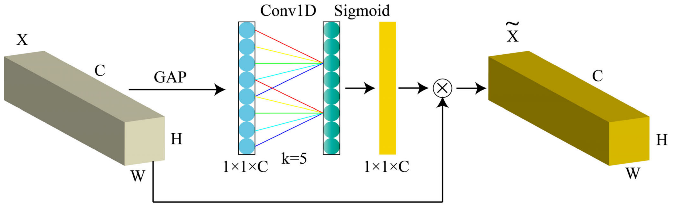


Fig. 4. Process of ECA.

2) Skip Connection: Skip connection is the core of the residual network, which is first proposed by He et al. [28] in 2015. Generally, as the layers of the network increase, the performance of the model decreases and is prone to gradient disappearance/gradient explosion. However, skip connections effectively improve the above problems.

The process of skipping connections is shown in Fig. 3. It can be formalized as follows:

$$y = \text{Activation}(x + H(x)) \tag{2}$$

where x is the input, y is the output, Activation is ReLU or sigmoid function, + is the element-by-element summation operation, $H(x)$ is a nonlinear mapping containing multiple convolutional layers, and the number of convolutional layers in the skip connection of TCN is 2.

B. Efficient Channel Attention

The channel attention mechanism has been shown to be effective in improving model performance, which also indicates that there is indeed variability between channels. In general, the attention module is inserted into the deep network to achieve better performance. However, in this article, we try to filter out the more important channels from the huge number of channels by the attention mechanism and observe the change of channels as the network deepens.

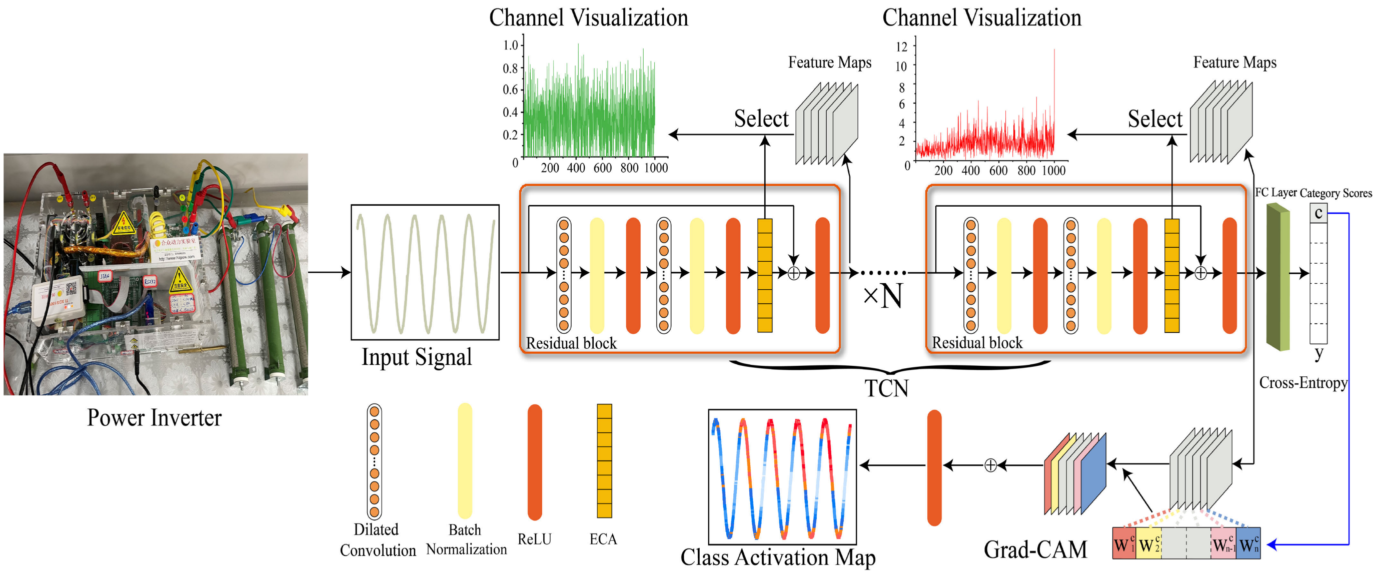


Fig. 5. General fault diagnosis process of power converters for the visualization.

Based on this motivation, a simple and effective channel attention mechanism, called ECA [29], is introduced into the TCN in this article. The working mechanism of the ECA module is shown in Fig. 4. ECA performs global average pooling (GAP) on each channel, which is a general operation to obtain the identifier of each channel in the attention mechanism. Let $X \in R^{H\times W\times C}$ denote the feature maps and H, W, and C denote height, width, and depth of the feature maps, respectively. The GAP can be formalized as follows:

$$y = G(X) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H} X_{i,j}. \tag{3}$$

ECA demonstrates that avoiding dimensionality reduction and learning local cross-channel interaction are more important than consideration of nonlinear channel dependencies. Thus, ECA discards the fully connected layer and uses fast 1-D convolution with k kernels to capture the relationship of each channel and its k neighbors. Specifically, the ECA can be formalized as follows:

$$\omega = \sigma(\mathrm{Conv1d}_k(y)) \tag{4}$$

where y is the channel identifier with the global information of each channel generated by GAP, σ is the sigmoid function, and Conv1d represents the 1-D convolution with a kernel size of k. Given channel dimension C, kernel size k can be adaptively determined by

$$k = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right| \tag{5}$$

where γ and b can be set empirically to 2 and 1.

The feature map after the ECA module can be recalibrated by ECA as

$$\tilde{X} = \omega X. \tag{6}$$

C. Grad-CAM

As a category-related localization technique, Grad-CAM can generate meaningful visual interpretations for networks based on CNN architecture [30]. In this article, we implement Grad-CAM with 1-D data as an input and apply it for the first time to the visual evaluation of power converter fault diagnosis. In general, the neurons of the convolutional layer in a fault diagnosis model try to find class-specific (distinguishing faults) information, and semantic information becomes clearer as the network deepens. Therefore, Grad-CAM uses the category scores to back-propagate to the last convolutional layer to obtain the gradient of the last feature map, which can represent the degree of importance in the decision process.

Specifically, for a certain fault c, Grad-CAM can generate explanations to classify the feature map as c. The formula for calculating the weights can be expressed as

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{7}$$

where A_{ij}^k is the value of the kth feature map at position (i, j), y^c is the category scores, and Z is the number of values in the feature maps. In addition, we only focus on positions that have a positive impact on the classification, while negative values are usually considered negative. Therefore, the ReLU function is introduced and Grad-CAM can be expressed as

$$L_{\mathrm{Grad-CAM}}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{8}$$

Unlike heat map overlay in images, color mapping is used to indicate the importance for classification in this article.

III. General Diagnostic Methods for Visualization

Since this research focuses on visualization and interpretability in power converter fault diagnosis, a basic fault

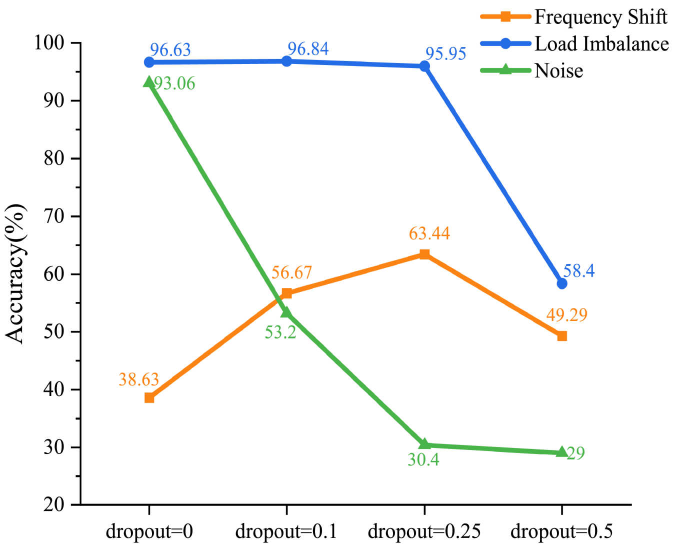


Fig. 6. Effect of dropout on generalization ability.

diagnosis framework based on TCN is developed, as shown in Fig. 5.

A. Basic Architecture

As shown in Fig. 5, the TCN is made up of multiple residual blocks stacked together. The optimal number of residual blocks is different for different tasks. The residual blocks in TCN contain two dilation convolution layers, and the feature maps after the dilation convolution layers both perform BN and nonlinear ReLU activation functions. In general, the BN layer can accelerate model convergence by regularizing the data distribution and ReLU function enables the model to have nonlinear mapping capabilities, which is also a standard structure in CNN. The output of the last residual block is sent to the fully connected layer to obtain a category score, which is used as a gradient back-propagation in Grad-CAM. Finally, the softmax module is used to generate the probability of the corresponding faults and perform cross-entropy operations with the one-hot label to update the gradient.

B. Diagnosis and Visualization Process

The fault diagnosis and visualization process in this article consists of data preprocessing, model training, and result analysis, and the specific steps are as follows.

Step 1 (Preprocessing): To eliminate the effects of amplitude inconsistencies in the original current data, the current data from the three-phase inverter are normalized. Fixed size windows are used to intercept data segments to generate the dataset and the start of the data segment is randomly changed to better match the actual sampling. Finally, the dataset is randomly divided into training set, validation set, and test set according to the ratio.

Step 2 (Model Training and Testing): We train the model on the training set, select the model by the validation set, and give the final diagnosis on the test set. In addition, during the testing process, the results that need to be visualized are saved, such as the attention values of ECA, the data of the different channels, the output of the different residual blocks, and so on, which will be displayed directly or intuitively interpreted through visualization methods.

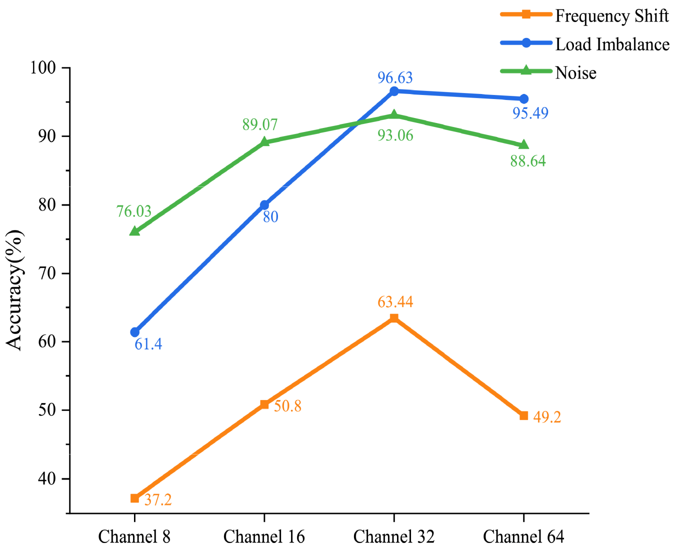


Fig. 7. Effect of the number of channels on generalization ability.

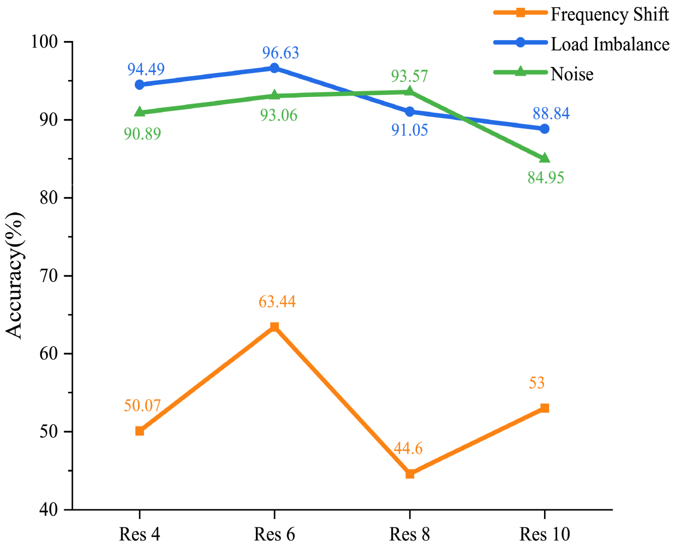


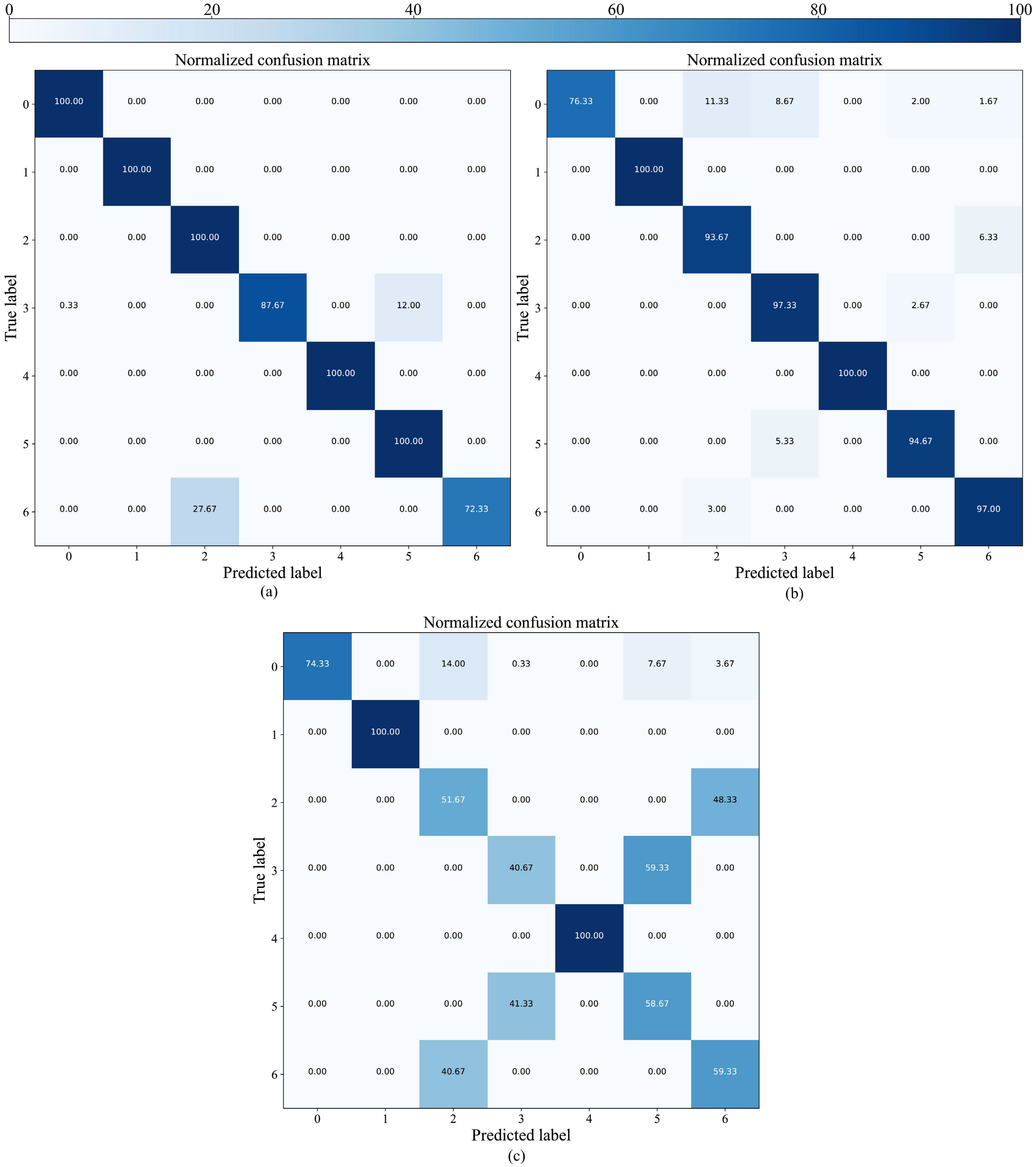
Fig. 8. Effect of the number of residual blocks on generalization ability.

Step 3 (Result Analysis): This part is the core of this article, where we try to explain visually the results and the model structure and the links between the input data through visualization methods. Furthermore, we hope that this interpretability will contribute to the implementation of a highly generalizable method for practical power converter fault diagnosis.

IV. Experiment

A. Data Acquisition

Sample data are collected from the three-phase inverter experimental platform in Fig. 5. The experimental platform parameters are: the filter inductance is 1 mH, the filter capacitance is 4.7 μF, the input voltage is 60 V, the output voltage is 30 V, the switching frequency is 20 kHz, the data sampling



(a)

(b)

(c)

Fig. 9. Confusion matrix. (a) Load imbalance. (b) Noise. (c) Frequency shift.

frequency of oscilloscope is 10 kHz, the current frequency is 50 Hz, and the IGBT type is FGH40N65. The digital signal processor TI TMS320F28069 is used to generate the SPWM signal. It should be noted that only the A-phase current is collected as a data sample.

Each signal sample consists of 1000 samples, i.e., five cycles. To be more in line with industrial applications, random overlapping sampling is used to split sample segments. We collect 300 samples for each state.

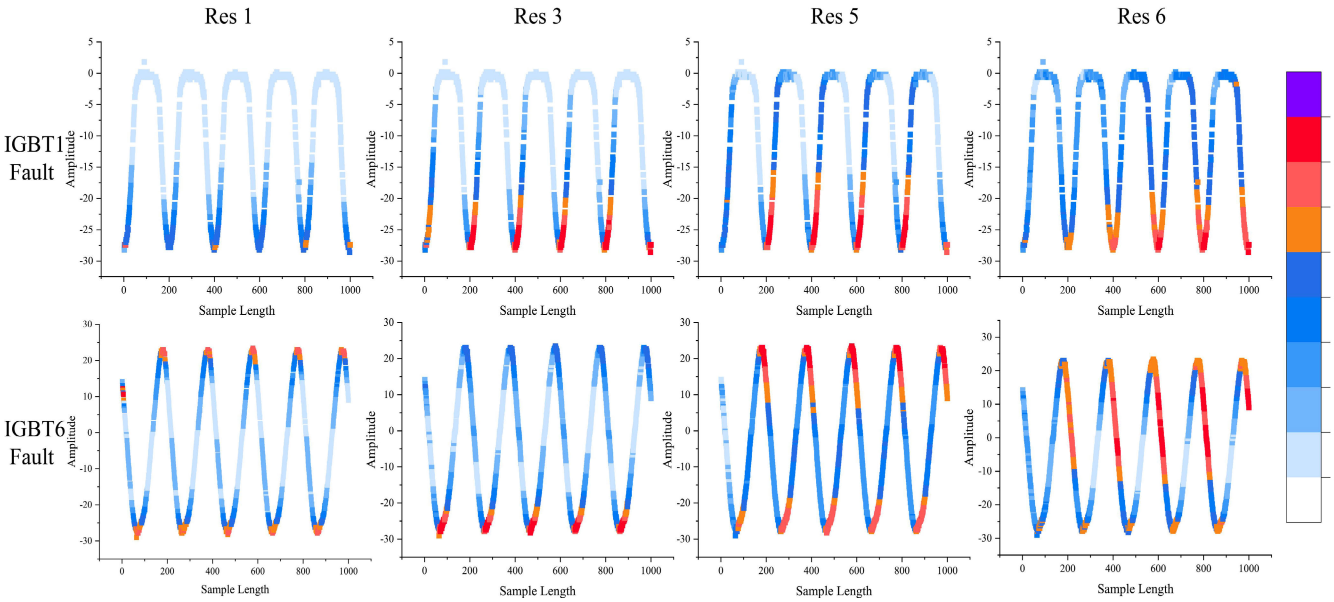


Fig. 10. Visual interpretability of different residual blocks.

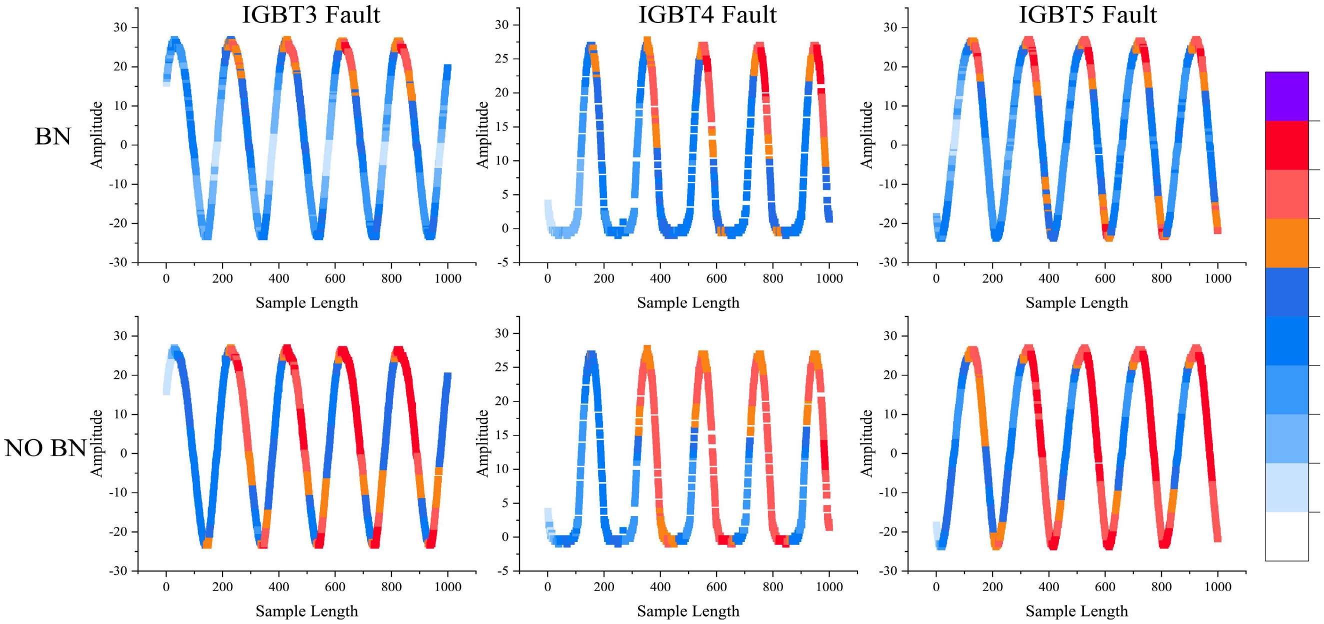


Fig. 11. Visual interpretability of models with BN and without BN.

B. Analysis on Hyperparameters

The effectiveness of data-driven fault diagnosis methods in the field of power converters has been proven. However, the generalization capability limits the implementation of data-driven methods in industrial manufacturing. It is foreseeable that generalizability will be a key research aspect of data-driven fault diagnosis methods. Therefore, we attempted to analyze the effect of key hyperparameters on generalization ability by detailed generalizability experiments, which could also provide a reference criterion for future research on generalizability.

The hardware environment in the experiment is Intel i7-10700KF + NVDIA RTX 3070, and the software environment is Windows10 + Python + Pytorch. The hyperparameters of the experiment refer to [17] and [18] and are set as follows: the epoch is 200, the batch size is 64, the kernel size is 3, the optimizer is Adam, the loss function is cross entropy, the dilation factor is $2^i (i = 0, 1, \ldots\ldots, N)$, and the learning rate is 0.001. The hyperparameters relevant to the model structure, the number of residual blocks, and the number of channels and dropout were analyzed to explore the effects on the generalization ability.

Typical uncertainty factors in power inverters including load imbalance, noise, and frequency shift are considered in the generalization experiments. The load of phases A and B is 30 Ω and the load of phase C is 15 Ω in the load unbalance experiment. The −5-dB noise is added to the original current in the noise experiment. The current frequency changes from 50 to 40 Hz in the frequency shift experiment, which is very common in inverters. It should be noted that the performance of the model inevitably degrades due to the perturbations only being added in the test set.

The three structure-related hyperparameters are set by the following [17]. Only specific hyperparameters are changed and other hyperparameters are kept fixed in the experiment. Each experiment was repeated five times and the final results were averaged over the five experiments to reduce randomness. Accuracy is used to evaluate model performance and can be defined as (the number of correctly classified samples)/(the number of all samples).

The effect of dropout on generalization is shown in Fig. 6. The number of channels and the number of residual blocks are fixed at 32 and 6 as in [17]. Low dropout does not improve the model and high dropout can even significantly impair the performance of the model in load imbalance experiments. The noise experiment also shows a significant negative correlation between dropout and accuracy, so we can infer that dropout cannot improve the generalization under unknown noise, which is different from the conclusion in [17]. The noise is added not only to the test set but also to the training set in [17], which can be considered as a diagnosis problem under the same distribution. However, to move toward more practical applications, the out-of-distribution generalization problem similar to the experiments in this article must be considered. In addition, a reasonable dropout can improve the generalization for frequency shifts, but there is a limit to this improvement. Therefore, considering the above factors, we believe that using a lower dropout or no dropout is more conducive to the research of the generalization problem under unknown interference.

The effects of the number of channels and the number of residual blocks on the generalization are shown in Figs. 7 and 8. It can be seen from Figs. 7 and 8 that the excessive number of channels or residual blocks increases the complexity of the model and degrades its performance. Therefore, the optimal number of channels is 32 and the optimal number of residual blocks is 6 for the data with a sample length of 1000.

To comprehensively evaluate the performance of the model, the confusion matrix of the model with optimal hyperparameters under generalization experiments is shown in Fig. 9. In the load imbalance experiment, for very similar IGBT2 Fault, IGBT6 Fault, IGBT3 Fault, and IGBT5 Fault, a bias is developed in the model. The model tends to diagnose IGBT3 Fault as IGBT5 Fault and IGBT6 Fault as IGBT2 Fault. This bias can be regularized by adding specific classes of fault data. However, misclassification exists simultaneously in IGBT2 Fault, IGBT3 Fault, IGBT5 Fault, and IGBT6 Fault for noise experiment. In addition, the accuracy of the normal state decreases due to the noise perturbation. This performance degradation reaches a maximum in the frequency shift experiment. The confusion matrix shows that the model cannot correctly distinguish between IGBT2 Fault and IGBT6 Fault and IGBT3 Fault and IGBT5 Fault, which indicates that even small frequency shift greatly affects the generalizability. Therefore, it is necessary to introduce domain generalization methods to achieve a fault diagnosis method with high generalizability.

C. Visual Interpretability

In this section, we aim to provide an intuitive visual interpretability to demonstrate the decision mechanism of the power converter fault diagnosis method. We constructed a general six-layers diagnostic model and extracted the visual interpretability of the fault from different residual blocks by Grad-CAM. To demonstrate the reliability, the visual interpretability of the two faults is compared in Fig. 10. As shown in Fig. 10, the horizontal axis represents the length of the sample and the vertical axis represents the amplitude of the fault current. Different color maps represent different levels of importance in fault diagnosis. Red indicates that the model pays more attention to the sample point, and blue indicates that the model pays less attention to the sample point. It can be seen in Fig. 8 that the model can focus on the regions of the same trend from the randomly sampled periodic current data. For example, the valley of the current is paid more attention to the IGBT1 fault, while both the peak and the valley are paid more attention to the IGBT6 fault. In addition, the shallow visual interpretability of the model fails to show a distinct region of concern. However, as the layers deepen, the visual interpretability indicates that a decision mechanism is developed by focusing on distinctive sample regions.

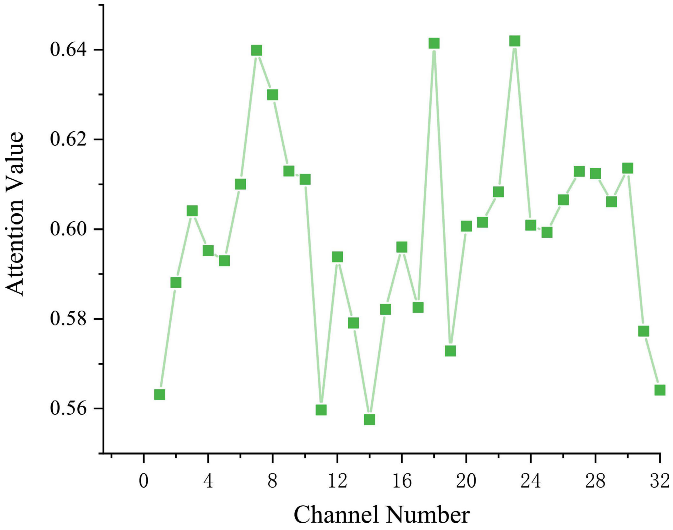


Fig. 12. Effect of the number of residual blocks on generalization ability.

Visual interpretability of faults not only enhances the credibility of fault diagnosis but also supports the design of model structures. BN is a popular trick in deep learning, which is used to help models converge faster and enhance performance. However, it has recently been shown that BN limits the generalization ability of the model for out-of-distribution problems [31]. The visual interpretability of the model with BN and the model without BN is shown in Fig. 11. We try to understand the role of BN in model training through an intuitive visual explanation. In Fig. 11, the visual interpretability of multiple faults shows that the model without BN usually has a larger concerned area and the concerned area of the model with BN is only a subset of the above area. From the perspective of interpretability, BN accelerates convergence by forcing the model to narrow the area of concern and mining deeply into the area for discriminative information related to the fault. However, this decision mechanism, which relies only on subareas, may fail under large gaps in data distribution. Furthermore, we can

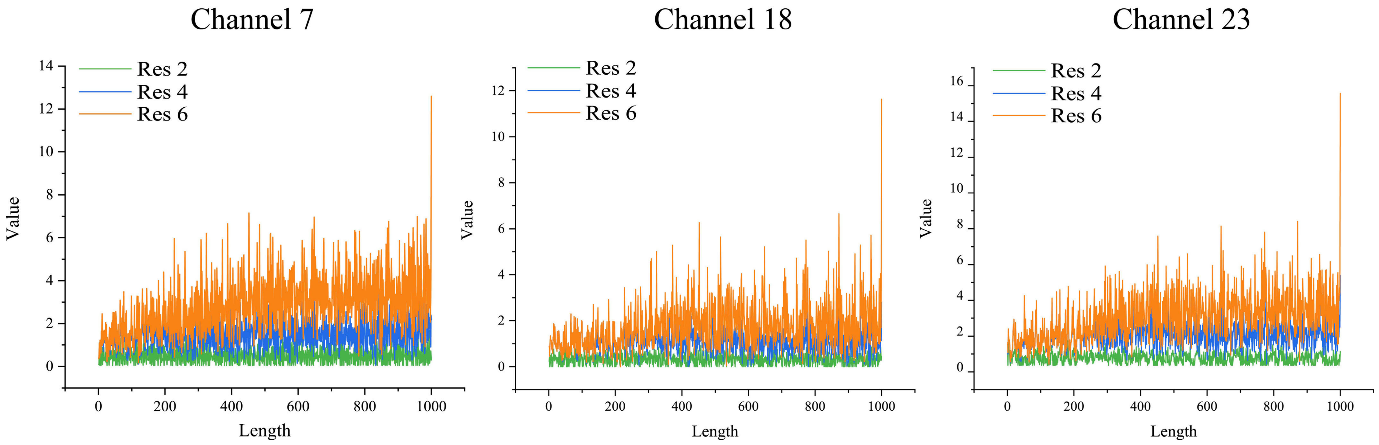


Fig. 13. Feature maps visualization.

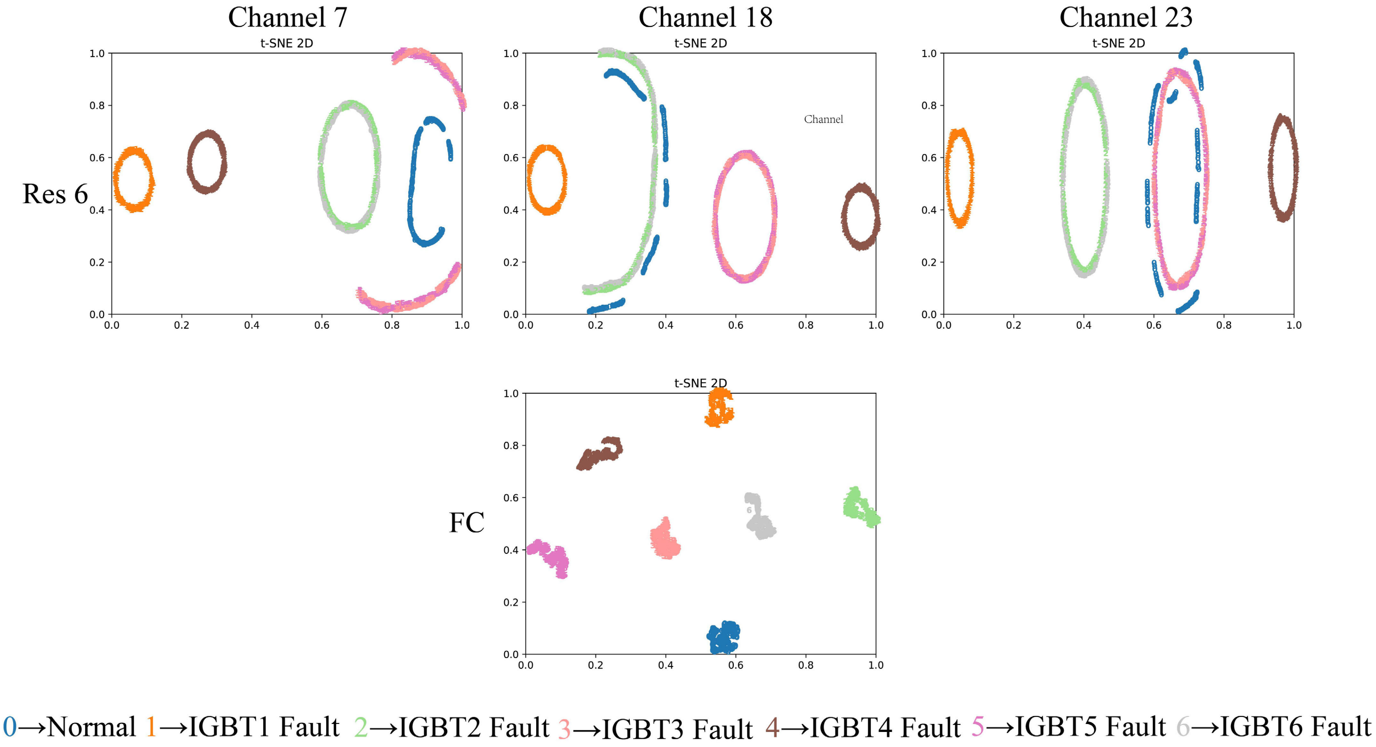


Fig. 14. High-dimensional data visualization.

speculate that ensemble learning is effective because it can improve generalization performance by assembling multiple models with BNs that focus on different subareas. In general, the visual interpretability of faults can be used as a reference to evaluate models.

D. Channel Visualization

Visualization of feature maps can reveal the changing mechanism inside the model. However, there are numerous channels in the model. To extract meaningful feature maps, ECA is employed to evaluate the importance of each channel. For the normal state samples, the attention values of the 32 channels generated by ECA are presented in Fig. 12. Channels 7, 18, and 23 are selected for visualization according to the size of the attention value.

The feature maps of the three channels are shown in Fig. 13. In addition to residual block 6, the feature maps of residual blocks 2 and 4 are presented together for comparison. Generally, the style information of the data is kept in the shallow layer of the model, while the semantic information exists in the deep layer of the model. The shallow feature map is chaotic and many values are close to zero in Fig. 13. As the layers increase (green–blue–orange), the value of each channel increases significantly and shows a different waveform pattern. It can be inferred that the unique waveform patterns of each channel finally together form the feature representation of the sample, which is the core of fault diagnosis.

The visualization of feature maps can show the details of signal transformation in the model, but it is hardly reflecting the data distribution discrepancy in global. To visualize the distribution of high-dimensional data, t-distributed stochastic neighbor embedding (t-SNE) is applied to reduce the high-dimensional sample features to 2-D data [32]. The 1000-D data generated by the last residual block for the different channels and the 32-D data generated by the fully

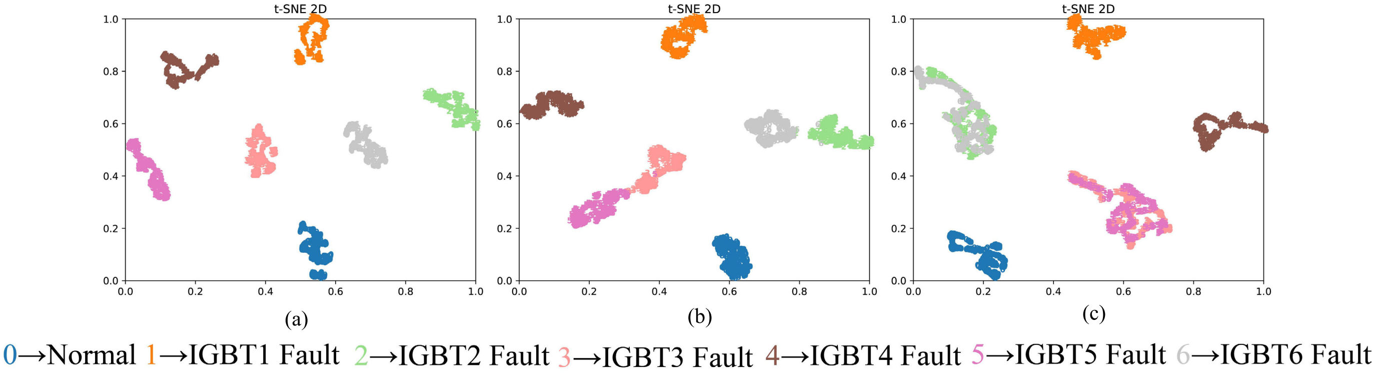


Fig. 15. Visualization of fully connected layer in generalization experiments.

connected layer are visualized in Fig. 14. An interesting finding is that the fault clusters of the different channels appear to be different and complementary. Since the current waveforms of IGBT1 Fault and IGBT4 Fault are sufficiently distinct, the IGBT1 Fault and IGBT4 Fault samples of all channels form respective clusters with clear boundaries, which illustrates the commonality between channels. While clusters formed by IGBT2 Fault and IGBT6 Fault in channel 7, clusters formed by IGBT3 Fault and IGBT5 Fault in channel 18, and four clusters in channel 23 demonstrate the discrepancy and complementarity of the different channels, which explains the reason why the model needs enough number of channels to acquire generalizability. In addition, the data distribution of IGBT2 Fault, IGBT6 Fault, IGBT3 Fault, and IGBT5 Fault is mixed up due to minor discrepancies in the current waveforms, which indicates that it is difficult to obtain enough fault information from a single channel to distinguish all faults. However, information from all channels is aggregated at the fully connected layer and forms a unique feature representation for each fault. In contrast to the sparse clusters in the channel visualization, seven tight clusters with clear boundaries are formed in the fully connected layer visualization, which shows that the different faults have been well separated.

Following the same settings, the visualization results of the fully connected layer in the generalization experiment are shown in Fig. 15. The seven fault clusters of the load imbalance experiment are close to Fig. 14, which indicates that the impact of load imbalance on data distribution is relatively small. Although the performance of the model is well in the noise experiment from the perspective of accuracy and confusion matrix, IGBT3 Fault clusters and IGBT5 Fault clusters have partially overlapped in the visualization results of noise experiment, which indicates a potential risk of the model. IGBT2 Fault clusters, IGBT6 Fault clusters, IGBT3 Fault clusters, and IGBT5 Fault clusters have been completely mixed up in the visualization results of the frequency shift experiment. Therefore, it can be inferred that the diagnostic results are more influenced by the randomness of the model rather than the extracted discriminative features in the frequency shift experiment, which also corresponds to the accuracy of about 50% in the confusion matrix of Fig. 9.

V. Conclusion

As the first attempt to research the interpretability of data-driven methods in the field of power converter fault diagnosis, we only focus on the interpretability and generalizability of the model in this article. A general TCN is constructed to perform interpretability and generalizability experiments. The effect of hyperparameters associated with the model structure on generalizability is analyzed. We believe that there is an optimal number of channels and number of residual blocks for fixed length current data. However, the value of dropout must be considered comprehensively, which greatly affects the performance of the model. To intuitively explain the decision mechanism of the model, Grad-CAM is used to visualize the concern areas of the model in the fault diagnosis. Visual interpretability shows that the model can systematically focus on specific areas in the electric current for different faults. In addition, the role of BN in model training is interpreted from the visualization perspective, which shows promise for supporting the design of model structures from a visualization perspective. Finally, the variation of the feature map with deeper layers is analyzed and the distribution of high-dimensional data for different channels is visualized, which intuitively explains the connection between channels and fully connected layers. This article presents an exhaustive visual analysis of the basic diagnostic models in the field of power converter fault diagnosis with the aim of promoting more interpretable and generalizable research, but this analysis can only be used as an aid to the research of interpretable and generalizable methods. For the fault diagnosis of power converters, generalizability and interpretability studies will be important topics for future research, e.g., domain generalization methods for unknown operating conditions and interpretable methods with physical meaning.

References

[1] G. J. Kish, M. Ranjram, and P. W. Lehn, “A modular multilevel DC/DC converter with fault blocking capability for HVDC interconnects,” *IEEE Trans. Power Electron.*, vol. 30, no. 1, pp. 148–162, Jan. 2015, doi: 10.1109/TPEL.2013.2295967.

[2] T. Soong and P. W. Lehn, “Assessment of fault tolerance in modular multilevel converters with integrated energy storage,” *IEEE Trans. Power Electron.*, vol. 31, no. 6, pp. 4085–4095, Jun. 2016, doi: 10.1109/TPEL.2015.2477834.

[3] M. Hagiwara, K. Nishimura, and H. Akagi, “A medium-voltage motor drive with a modular multilevel PWM inverter,” *IEEE Trans. Power Electron.*, vol. 25, no. 7, pp. 1786–1799, Jul. 2010, doi: 10.1109/TPEL.2010.2042303.

[4] A. Antonopoulos, L. Angquist, S. Norrga, K. Ilves, L. Harnefors, and H.-P. Nee, “Modular multilevel converter AC motor drives with constant torque from zero to nominal speed,” *IEEE Trans. Ind. Appl.*, vol. 50, no. 3, pp. 1982–1993, May 2014, doi: 10.1109/TIA.2013.2286217.

[5] G. Rojas-Dueñas, J. Roger Riba, and M. Moreno-Eguilaz, "Modeling of a DC–DC bidirectional converter used in mild hybrid electric vehicles from measurements," *Measurement*, vol. 183, Oct. 2021, Art. no. 109838.

[6] Q. Guo, J. Li, F. Zhou, G. Li, and J. Lin, "An open-set fault diagnosis framework for MMCs based on optimized temporal convolutional network," *Appl. Soft Comput.*, vol. 133, Jan. 2023, Art. no. 109959, doi: 10.1016/j.asoc.2022.109959.

[7] D. Binu and B. S. Kariyappa, "RideNN: A new rider optimization algorithm-based neural network for fault diagnosis in analog circuits," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 2–26, Jan. 2019, doi: 10.1109/TIM.2018.2836058.

[8] S. Chen, H. Ge, H. Li, Y. Sun, and X. Qian, "Hierarchical deep convolution neural networks based on transfer learning for transformer rectifier unit fault diagnosis," *Measurement*, vol. 167, Jan. 2021, Art. no. 108257.

[9] J. A. Reyes-Malanche, F. J. Villalobos-Pina, E. Cabal-Yepez, R. Alvarez-Salas, and C. Rodriguez-Donate, "Open-circuit fault diagnosis in power inverters through currents analysis in time domain," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2021.3082325.

[10] G. Patrizi et al., "Electrical characterization under harsh environment of DC–DC converters used in diagnostic systems," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2021.3129513.

[11] H. Nademi, A. Das, and L. E. Norum, "Modular multilevel converter with an adaptive observer of capacitor voltages," *IEEE Trans. Power Electron.*, vol. 30, no. 1, pp. 235–248, Jan. 2015, doi: 10.1109/TPEL.2014.2301879.

[12] Y. Zhang, "Current behavior-based open-switch fault on-line diagnosis of inverters in PMSM drive systems," *Measurement*, vol. 202, Oct. 2022, Art. no. 111810.

[13] T. Ameid, A. Menacer, H. Talhaoui, and I. Harzelli, "Rotor resistance estimation using extended Kalman filter and spectral analysis for rotor bar fault diagnosis of sensorless vector control induction motor," *Measurement*, vol. 111, pp. 243–259, Dec. 2017.

[14] A. Moradzadeh, B. Mohammadi-Ivatloo, K. Pourhossein, and A. Anvari-Moghaddam, "Data mining applications to fault diagnosis in power electronic systems: A systematic review," *IEEE Trans. Power Electron.*, vol. 37, no. 5, pp. 6026–6050, May 2022, doi: 10.1109/TPEL.2021.3131293.

[15] Y. Xia, Y. Xu, and B. Gou, "A data-driven method for IGBT open-circuit fault diagnosis based on hybrid ensemble learning and sliding-window classification," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5223–5233, Aug. 2020.

[16] B. Cai, Y. Zhao, H. Liu, and M. Xie, "A data-driven fault diagnosis methodology in three-phase inverters for PMSM drive systems," *IEEE Trans. Power Electron.*, vol. 32, no. 7, pp. 5590–5600, Jul. 2017.

[17] G. Yating, W. Wu, L. Qiongbin, C. Fenghuang, and C. Qinqin, "Fault diagnosis for power converters based on optimized temporal convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2020.3021110.

[18] Q. Guo, X. Zhang, J. Li, and G. Li, "Fault diagnosis of modular multilevel converter based on adaptive chirp mode decomposition and temporal convolutional network," *Eng. Appl. Artif. Intell.*, vol. 107, Jan. 2022, Art. no. 104544.

[19] S. Zhang, R. Wang, Y. Si, and L. Wang, "An improved convolutional neural network for three-phase inverter fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022, doi: 10.1109/TIM.2021.3129198.

[20] Y. Zhao et al., "Multibranch 1-D CNN based on attention mechanism for the DAB converter fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022, doi: 10.1109/TIM.2022.3203445.

[21] H. Yang, X. Li, and W. Zhang, "Interpretability of deep convolutional neural networks on rolling bearing fault diagnosis," *Meas. Sci. Technol.*, vol. 33, no. 5, May 2022, Art. no. 055005.

[22] S. Yu, M. Wang, S. Pang, L. Song, and S. Qiao, "Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network," *Measurement*, vol. 196, Jun. 2022, Art. no. 111228.

[23] Y. Xia and Y. Xu, "A transferrable data-driven method for IGBT open-circuit fault diagnosis in three-phase inverters," *IEEE Trans. Power Electron.*, vol. 36, no. 12, pp. 13478–13488, Dec. 2021, doi: 10.1109/TPEL.2021.3088889.

[24] R. Yan, F. Shen, C. Sun, and X. Chen, "Knowledge transfer for rotary machine fault diagnosis," *IEEE Sensors J.*, vol. 20, no. 15, pp. 8374–8393, Aug. 2020, doi: 10.1109/JSEN.2019.2949057.

[25] T. Li et al., "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 4, pp. 2302–2312, Apr. 2022, doi: 10.1109/TSMC.2020.3048950.

[26] B. An, S. Wang, Z. Zhao, F. Qin, R. Yan, and X. Chen, "Interpretable neural network via algorithm unrolling for mechanical fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3188058.

[27] C. Zhu, Z. Chen, R. Zhao, J. Wang, and R. Yan, "Decoupled feature-temporal CNN: Explaining deep learning-based machine health monitoring," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, doi: 10.1109/TIM.2021.3084310.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[31] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. ECCV*, vol. 11208, 2018, pp. 484–500, doi: 10.1007/978-3-030-01225-0_29.

[32] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

Tongyang Ren received the B.S. and M.S. degrees from Jilin University, Changchun, China, in 2016 and 2019, respectively.

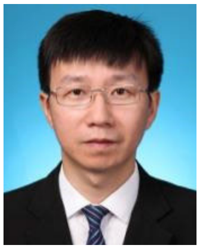
He is now an Assistant Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include circuit design, artificial intelligence, fault diagnosis, prognostics health management, and its applications in photoelectric and electromagnetic detection equipment.

Tao Han was born in Jinzhou, Liaoning, China, in 1998. He received the B.E. degree from Jilin University, Changchun, China, in 2020, where he is currently pursuing the master's degree.

His research interests include the development of high-power transmitter and simulation and control of renewable power microgrid.

Qun Guo received the B.S. and M.S. degrees from Jilin University, Changchun, China, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree.

His research interests include artificial intelligence, fault diagnosis, dc/dc power converter, and its applications in electromagnetic detection instruments.



Gang Li (Member, IEEE) received the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2009.

He is currently an Associate Professor with the College of Instrumentation and Electrical Engineering, Jilin University, Changchun, China. His research interests include modeling of semiconductors and li-batteries, matrix converters, dc/dc power converter and its applications in electromagnetic detection instruments, and design and implementation of high-power source.