IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 October 2022, accepted 27 December 2022, date of publication 12 January 2023, date of current version 18 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3236315

RESEARCH ARTICLE

Verification of Interpretability of Phase-Resolved Partial Discharge Using a CNN With SHAP

RYOTA KITANI AND SHINYA IWATA, (Member, IEEE)
Osaka Research Institute of Industrial Science and Technology, Izumi, Osaka 594-1157, Japan
Corresponding author: Ryota Kitani (kitanir@tri-osaka.jp)

This work was supported by the JSPS (Japan Society for the Promotion of Science) KAKENHI under Grant JP20H02140.

ABSTRACT Deep neural networks can be used to distinguish partial discharge (PD) signals despite their complexity. This study analyzes the appropriateness of interpreting phase-resolved partial discharge (PRPD) signals using a convolutional neural network (CNN) through the Shapley additive explanation (SHAP) method. The generated PRPD signals were accumulated by applying AC voltage to four types of electrodes with a polyethylene sheet, followed by their conversion into scattered images to construct a classification model, CNN. The SHAP values for each pixel in the test images were then calculated. The result indicated that the pixels around the 0 V line retained high absolute SHAP values in every label, and the average of the summation of absolute SHAP values over all labels and all test images, which indicates the weight of each pixel, shows a similar tendency. Additionally, insight tests of the two CNN models were conducted, and the results showed that some structural defects could be detected by visualizing the SHAP values for each pixel. Finally, the verification of parameter-and-data vulnerability showed that SHAP has sufficient endurance against some types of instability in the data and model. Although the SHAP method lacks a perfect causal model because of its origin, the results imply that in appropriate use cases, weights on classifications of PD signals could be described by SHAP's interpretability.

INDEX TERMS Convolutional neural network, diagnosis, machine learning, phase-resolved partial discharge, Shapley additive explanations.

I. INTRODUCTION

In recent years, an increasing number of electric power devices have been associated with higher voltages, higher electric fields, and more complex environments. Therefore, electric power equipment and devices should be more reliable and durable. These circumstances have increased the need for research on the diagnosis of electrical insulators.

Partial discharge (PD) measurement is a diagnostic technique used for electrical insulators in devices such as mortar and other equipment [1], [2]. PD is recognized as a discharge phenomenon, which emits a slight electrical signal and light and causes incomplete electrical breakdown [3] inside insulators. In general, the deterioration of the insulator affects the PD signals, which are characterized by the number, phase angle, magnitude, and repeating discharge pattern [4], [5], [6], [7], [8], and are sometimes depicted in the schematic shown in Figure 1 and/or its variation. For example, the abc-model shown in Figure 1 represents the solid insulator and its deteriorated part as a capacitor. C_b, C_c, and C_a in Figure 1 indicate the capacitances of the void/impurity inside the insulator, the insulating part serially connected to the void, and the rest of the insulator (healthy part), respectively. PD occurs when the applied voltage exceeds the dielectric strength of C_b. Once the discharge occurs, the charge accumulation in C_b disappears, causing the potential between both sides of the void to become equal. However, when an oscillating voltage (for instance, AC voltage) is applied, electric charges accumulate, then the threshold voltage is exceeded, discharge occurs again. In other words, the deterioration of insulating materials and/or that of electric products can be diagnosed by measuring the PD signals under a high AC voltage (or pulsed inversion of DC voltage). In particular, research on discharging phase patterns has been

The associate editor coordinating the review of this manuscript and approving it for publication was Pavlos I. Lazaridis.

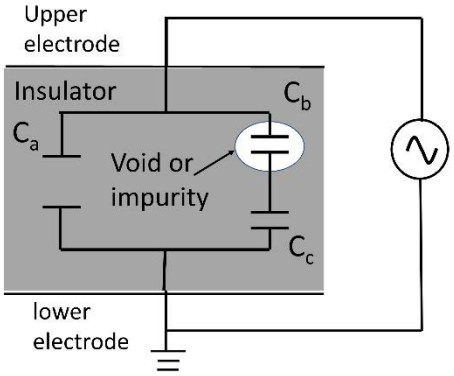


FIGURE 1. Analytical electrical circuit model of PD diagnosis.

conducted for many types of targets. Therefore, the phase-resolved PD pattern is considered an important diagnostic research technique.

On the other hand, in recent years, research, development, and evaluation involving the automation of certain processes of PD diagnoses have been increasingly conducted. This is mainly because highly experienced workers are required to perform diagnoses that exploit the PD patterns. For example, the process of judging whether the signals are benign or malignant for the system. Reference [9] reviewed and compared lots of such diagnostic methods.

Recently reported studies that used deep neural networks (DNN) and/or their variational techniques achieved high performance, with an accuracy score more than 90 %. These research absorbed and assimilated the latest machine learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GAN). For example, a UHF PD signal dataset created by a finite-difference time-domain simulation was transformed into a two-dimensional spectral frame representation using the short-time Fourier transform (STFT) method [10]. The STFT results were processed using a CNN model to accomplish feature extraction and classification with an accuracy of 0.967, which was higher than that of the processes using other machine learning methods. A GAN is used as an additional option for data augmentation, which decreases the cost of collecting and labeling data sources. Wang et al. used a GAN and achieved a higher score than that of the original datasets [11]. From another viewpoint, some researches focused on the relation between learning data and learning process. Mantach et al. discussed the difference of classified single-sourced and multiple sourced Phase Resolved PD (PRPD) patterns when training [12]. As suggested by the abovementioned results, it has been revealed that analyzing signals that involve PD and applying deep learning has sufficient potential to achieve high performance. Reference [13] surveyed and arranged the tendency of the combination of PD diagnoses and DNNs/CNNs and/or their variation.

On the other hand, most studies using deep learning have failed to consider the causes of PD in associated electrical models. In the actual use case, the prediction/classification insights are equally significant as the result itself. This is mainly because reasonable explanations are required to insist on the indispensability of the maintenance of electrical infrastructure owing to its value and impact.

Thus, in our previous work [14], we discussed the relationship between the physical phenomena associated with electrical circuit models and several machine-learning methods. The efficiency of conventional machine-learning methods for application to phase-resolved PD signal data was clarified and arranged. The present study is a variant of the previous study, which attempts to connect the input data and DNN/CNNs.

However, the present study did not use electrical circuit models. This is mainly because DNN/CNNs are black-box models and are not necessarily analyzed completely in a mathematical context. In other words, considering the current advancements, it is extremely difficult to describe the relationship between DNN/CNNs and physical phenomena. This implies that no methods enable us to directly connect a DNN/CNN model and an electrical circuit model; Thus, alternative methods should be used to evaluate the relationship and/or usage [15].

This study used model-agnostic methods to explain the DNN/CNN models, with a focus on the interpretability of the models of the phase-resolved PD signals using a CNN for different electrode systems. To exploit model-agnostic methods, the relationship between the prediction results and the input data is described. More important signals tend to be visualized for each model and/or sample. Several methods can be used to visualize the reasons for the pre-learned model and/or prediction/classification results, such as local interpretable model-agnostic explanations (LIME) [16], Shapley additive explanations (SHAP) [17], and gradient-weighted class activation mapping (grad-CAM) [18]. On the contrary, there are methods that edit model itself and make it interpretable. Mantach et al. reported the potential of attention-based model and discussed the better usefulness of its combination with grad-CAM [19].

Such methods can be useful, but some have complicated or elusive methods for determining hyperparameters, calculation burdens, or mathematical backgrounds [20], [21]. In addition, a considerable number of examples pertain to the image recognition of concrete objects in the actual world, for example, distinguishing an apple or a dog. However, in the present study, the targets are images of the dot pattern of electrical signals to prevent human beings from clearly and immediately distinguishing a picture from others. Then, by analyzing such methods, what the ‘‘interpretability’’ actually shows us of the models/results must be verified. This is a basic analysis to test the interpretability of the entire diagnostic system and its limitations. Moreover, it is important to consider the application of machine learning methods in actual appliances with respect to the mechanism through which the data should be treated for actual use.

Hence, the present study aimed to verify the conformity of what is explained in a CNN model. We attempted to arrange the relationship between the inputs and outputs using

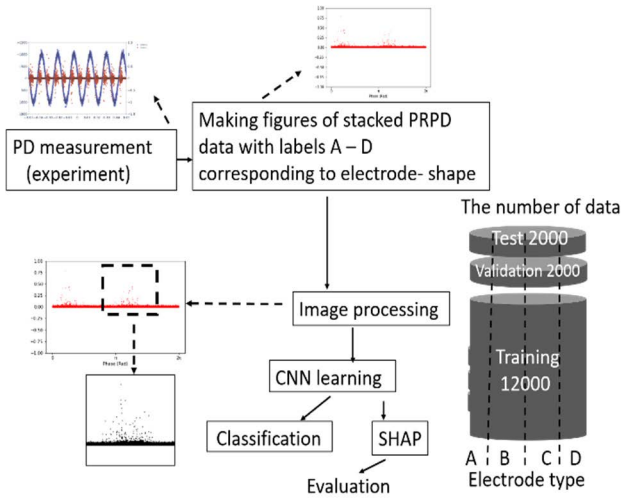


FIGURE 2. Analytical processing model.

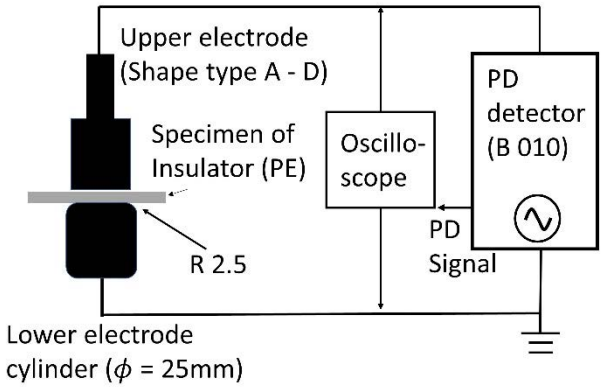


FIGURE 3. Schematic depiction of experimental PD measurement setup.

processed images of stacked PRPD data as input data for CNN training, followed by calculating the SHAP value. Moreover, the ''insight'' of the model, namely the results of visualization, is examined to determine whether the values are likely to be appropriate and/or useful. Finally, considering both the results and theoretical background, we conclude with the important points of the use of the SHAP method.

II. METHODS

An outline of the analysis process is shown in Figure 2. The process was divided into four parts: (1) PD measurement, (2) data preprocessing, (3) supervised learning with a CNN, and (4) evaluation of the resulting data.

A. PD MEASUREMENT

Figure 3 shows the experimental setup used for the PD measurements. A polyethylene sheet specimen with a thickness of 0.1 mm was used as the target for diagnosis. For the PD experiment, a cylindrical electrode was used as the lower electrode, as shown in the lower-left part of Figure 3. The upper electrode was an electrode of types A–D, as shown in Figure 4. The upper and lower electrodes were made of

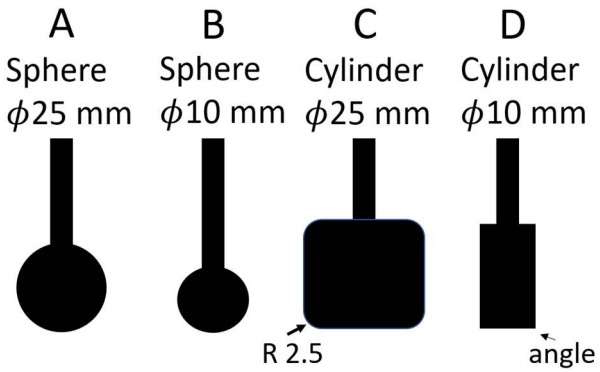


FIGURE 4. Schematic depiction of the upper electrodes.

brass. Note that the reason for using such different electrodes is to make verifying the factor affects on this measurement clearer. The sample's detail deterioration-status and/or the mode of PD (= different PD types, for examples, corona, surface, void, slot discharge, etc.) is not treated in this paper to simplify the verification process. B010 (band-pass range:15–150 kHz, 3 dB, manufactured by Fujikura Dia Cable Ltd.) was used as the PD detector for the measurement. The applied voltage was monitored along with the PD signal from B010, using a Tektronix DPO 4034 oscilloscope. The frequency of the applied power voltage was 60 Hz. The sample interval and recording length of the oscilloscope were 100 ms and 10,000 points, respectively. A total of 4,000 measurements were performed for each type of electrode, and the sample was exchanged after every 200 measurements to avoid unexpected deterioration. To maintain the maximum PD charge within the range of 800–3,000 pC for all measurements, the applied voltage was set in the range of 0.8 to 1.5 kV.

B. PD DESIGN OF CNN LAYERS

CNN is an advanced option for DNNs. Convolutional neural networks (CNN) are often utilized for image recognition. A highly complex network structure and several tuning techniques enable the learning model to have adequate weight coefficients to predict labels precisely. The layer designs of the CNN used in the present study are listed in Table 1. ''Normal network'' is the standard model in this paper, and ''narrowed network'' is the modified network which is notably similar to the normal network; however, some convolution parameters and the number of usage of max-pooling execution is slightly different from the normal.

C. DATA PREPROCESSING

An example of raw PD signal data and smoothed applied voltage data is shown in Figure 5. In the experiment, 4,000 measurements were recorded for each electrode. In the present study, the phase angle dependence of the PD signal is reflective of the electrode type. Considering the construction of the CNN models in this study, feature values are the binary

TABLE 1. Design of CNN layers. "Normal network" is used for majority of the analysis as the following chapters, and "Narrowed network" is used for the insight test in Chapter 3-C.

Layer structure			Activation func.
Layer No.	Normal network	Narrowed network	
1st	Convolution (channel = 1 -> 32, kernel size = 5, stride = 2, padding = 3(reflection)) & batch normalization & maxpooling	Convolution (channel = 1 -> 32, kernel size = 5, stride = 3) & batch normalization & maxpooling	ReLU
2nd	Convolution (channel = 32 -> 128, kernel size = 5, stride = 2, padding = 3(reflection)) & batch normalization & maxpooling	Convolution (channel = 32 -> 128, kernel size = 4, stride = 2) & batch normalization	ReLU
3rd	Convolution (channel = 128 -> 256, kernel size = 3, stride = 1, padding = 3(reflection)) & batch normalization & maxpooling	Convolution (channel = 128 -> 256, kernel size = 3, stride = 2) & batch normalization	ReLU
4th	Convolution (channel = 256 -> 64, kernel size = 3, stride = 1, padding = 3(reflection)) & batch normalization	Convolution (channel = 256 -> 64, kernel size = 3, stride = 2) & batch normalization	ReLU
final	fully connection & softmax	fully connection & softmax	

values of each pixel in an image of the stacked PD signal data. Therefore, owing to the limitations of data acquisition and preprocessing, we exploited cut binary-images converted from the PRPD diagram data, whose black dots indicate the existence of an electrical signal at the phase point, as previously shown in the lower left of Figure 2. The total size of the PD image dataset (original image dataset) was 16,000, which did not contain any augmented data, and the numbers of training, validation, and test data were 12000, 2000, and 2000, respectively. Note that the limitation of the batch processing and batch size (32), some images are not actually used on each epoch (for example, the actually chosen number of test data is $32 \times 62 = 1984$ (< 2000) on each test).

D. SHAP AND "GRADIENT EXPLAINER"

The straightforward interpretability of some machine-learning methods, especially for DNNs, has not been proven. The straightforward interpretability of a machine learning method remains unproven, particularly for DNNs. One plausible way to connect the results of a prediction to the learning model is to overhaul the contribution of each element of a datum [22]. In this study, "Gradient explainer," a package from the SHAP methods was used [23]. Gradient explainer consists of "Integrated gradient" [24] and SHAP. SHAP is

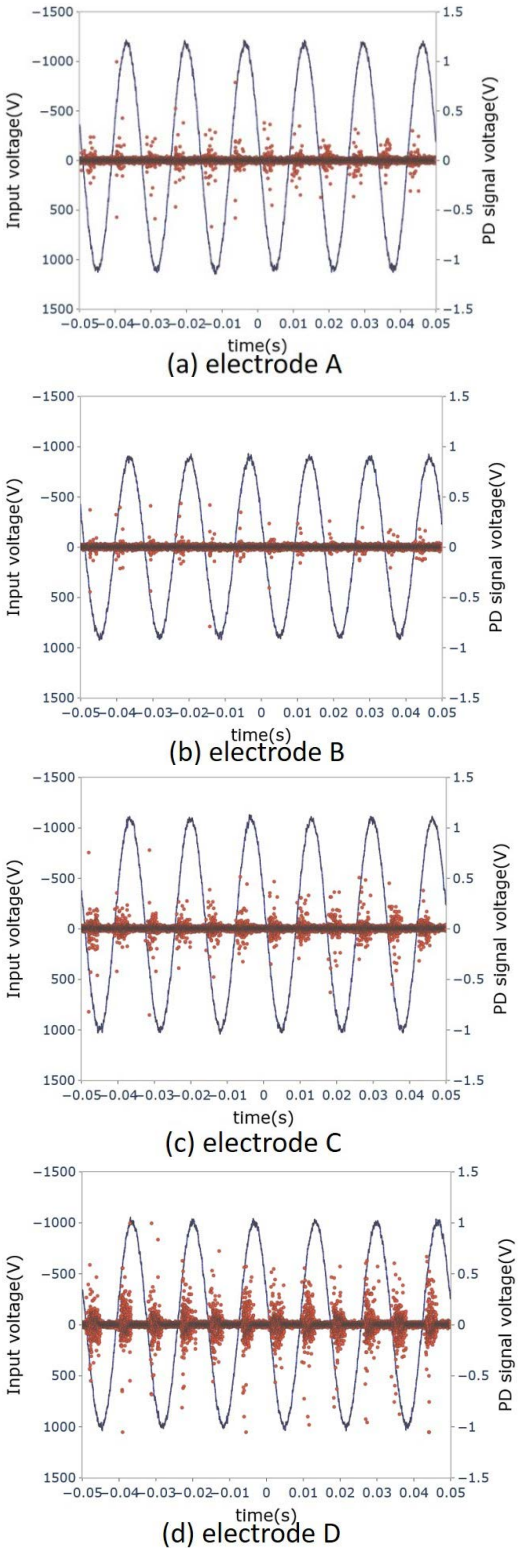


FIGURE 5. Examples of raw PD signal data from electrode types: (a) A, (b) B, (c) C, and (d) D. The applied voltage (set value) and PD signal voltage corresponds to the black oscillating line and red scatter plots, respectively.

an additive method so that it can be applied after the creation of networks is completed. A brief comparison of SHAP and similar methods' features is shown in Table 2. In short, SHAP

TABLE 2. Brief comparing of SHAP and relate methods.

Method name	(Brute force method)	SHAP	LIME
Computational time	Very long	Not so long	Short
Mathematical appropriateness	Less obvious than SHAP	Partially appropriate	Has a significant limitation
Difficulty to implement	Very easy	Not difficult	Easy
Main feature	Needs lots of computational time and the method to reasonably count each "points" of features.	Mathematically proved to some extent (local accuracy, missingness, consistency of each result)	Quite obvious, but contain some arbitrariness

has useful and unique mathematical properties that make the comparison of each result easier than other similar methods; Let g is a function of explanation model, $z'_j \in \{0, 1\}^M$ is called as the coalition vector, which means the existence or not-existence of each feature, M is the maximum size of the coalition vector, $\phi_j \in \mathbb{R}$ is Shapley value [22], which indicates the attribution of feature-value j, f is the original model, x is the target instance, and x' is x's corresponding simplified-expression as the coalition vector (simplified input). Then, as the definition of SHAP, g is described as followings;

$$g\left(z'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{1}$$

SHAP satisfies ''Local accuracy'', which requires the match of f(x) and g(x'), and indicates which the summation of feature contributions must be equal to,

$$f\left(x\right) = g\left(x'\right) = \phi_0 + \sum\nolimits_{j=1}^{M} \phi_j x'_j \tag{2}$$

''Missingness'' is

$$x'_j = 0 \Rightarrow \phi_j = 0 \tag{3}$$

and ''Consistency'' is, roughly speaking, SHAP value does not decrease when the original model f changes to its alternative f' so that some simplified input's contribution (= marginal contribution of a feature-value) increases or stays the same. Note that the detail of SHAP's mathematical property is described in [17] and [21].

The entire process of the gradient explainer is illustrated in Fig. 6. First, the data to be explained (data-to-explain) by SHAP, the machine-learning model to be explained (model to evaluate, as a pre-learned Model) and bulk data for training or testing the model (data-for-test) are prepared. Second, a datum of the compound of datum-to-explain and datum-for-test) with a rate of $1-\alpha$: α is created (samples_input). Third, after creating considerable data according to the abovementioned steps, the data are input to the model-to-evaluate. Then, the gradients corresponding to the pairs of the model and data are calculated, just as intermediate products of inferences.

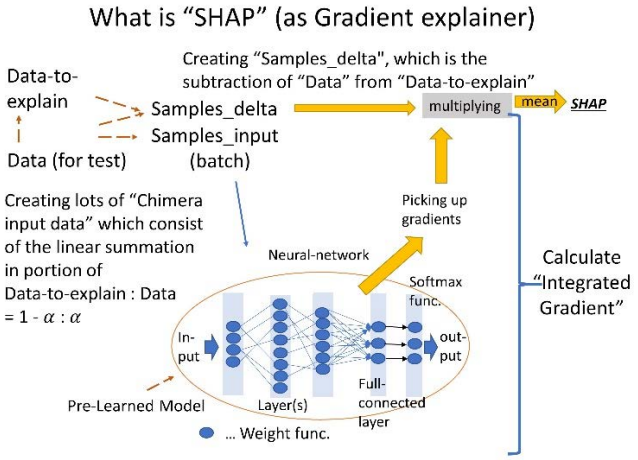


FIGURE 6. Data-flow of SHAP (Integrated Gradient).

TABLE 3. Precision, recall, F-value, and accuracy of each label of the PRPD image data (128 × 128 pixels).

Actual/Pred. Label	A	B	C	D	Recall	F value
A	460	1	0	6	98.5%	0.954
B	6	503	2	8	96.9%	0.975
C	0	1	490	5	98.8%	0.984
D	31	8	8	455	90.6%	0.932
Precision	92.6%	98.1%	98.0%	96.0%	96.2%	(=Accuracy)

On the other hand, the subtraction of the bulk data (here the same as the second step) from the data-to-explain is calculated (samples_delta.) In this case, the subtraction refers to that of the value of each pixel. Finally, the gradients were multiplied with the corresponding sample_delta(s), the mean was calculated, and the result of SHAP (SHAP value) is obtained.

Note that the gradient explainer shows a Shapley value for the result of each integrated gradient method so that it contains an approximation (Monte Carlo method-like parts) in the calculation to reduce the calculation burden.

E. LEARNING AND PREDICTION PROPERTY OF CNNS

Table 3 shows the scores of the prediction test performed by the CNN defined in chapter 2-C. Note that the accuracy is defined as the value obtained by dividing the number of correct predictions by the total number of predictions. TP refers to a true positive, the result of a positive prediction that is also equal to the actual label. Similarly, TN, FP, and FN refer to true negatives, false positives, and false negatives, respectively. Accuracy was denoted as (TP of all labels) / (all test data). Similarly, ''recall'' is defined as (TP) / (TP + FN), precision is defined as (TP) / (TP + FP), and the F-value is defined as 2 × recall × precision / (recall + precision).

According to the table, the models created by sufficient quality of data and contain a reasonable network structure

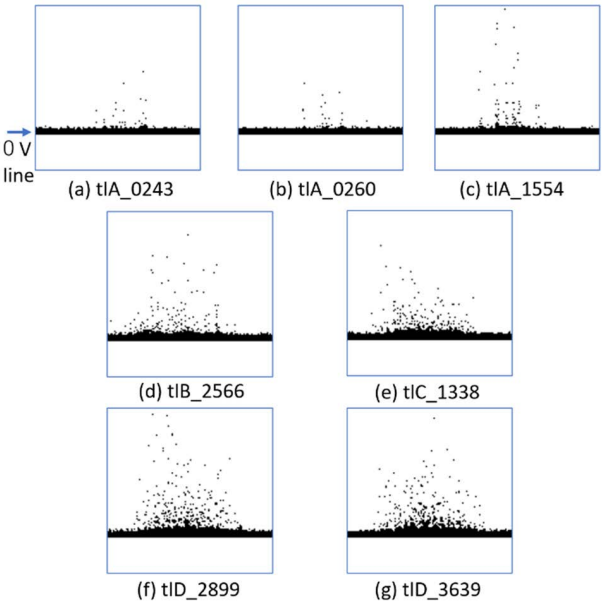


FIGURE 7. Binary images of PRPD, whose sign from (a) to (g) correspond to those of Figure 8. The label in each image is an ID name.

performs with an accuracy of over 96 %. It should be noted that the score itself has low validity here because of SHAP. The results of SHAP are not directly affected by whether the predicted and true labels are the same; In other words, there are no problems if the accuracy in the scope of this study is not high.

III. RESULTS AND DISCUSSIONS

A. BASIC RESULT OF SHAP

Figure 7(a)–(g) show the original test data (= cut binary-images of PRPD data) corresponding to the examples of the results of SHAP shown in Figure 8(a)–(g). In Figure 8, the upper left, upper right, lower left, and lower right figures correspond to electrodes A, B, C, and D, respectively. Each pixel position on the X-Y plane corresponds to the original test data in Figure 7, like the left-bottom point (x, y) = (0, 0). The red and blue pixels indicate high positive and negative values, respectively. White pixels are of little/no absolute value; therefore, they are not concerned about the prediction. In this case, the number of prediction classes is four (A, B, C, D), and each prediction-class-candidate can have SHAP values, although three of the four candidates are incorrect labels. In other words, although an image has a false-predicted result, the SHAP values can be calculated in accordance with the label chosen by the user. Therefore, in this study, we preferentially choose the predicted labels when there are no additional explanations. Notably, the summation of all the SHAP values at one pixel of all the candidates of an image (for example, the summation of the SHAP values at (64, 64) pixels in images A, B, C, and D) is almost equal to 0. This is because the SHAP values shown in this study were zero-adjusted in accordance with the properties of SHAP, as previously described in chapter 2.

Some typical results were obtained. Some parts of the dotted area, especially the area around the bottom baseline (= around (x, 36), 0 V line) and some spots in the upper half of the original picture data, tend to have large absolute values (and their distribution) of SHAP. This result is intuitively considered correct; However, the SHAP values of the correct label are not necessarily larger than those of the incorrect label, as shown in Figure 8(d)–(g). On the other hand, the lower-left and upper-half parts of images in Figures 8(a) and (b) show the dotted red and/or blue pixels in the area corresponding to the void space (= no signal detected) of the original picture data. Another typical feature is that the SHAP values corresponding to electrodes C and D are smaller than those of electrodes A and B. Figure 9 shows the average SHAP values of each label. This indicates that the aforementioned tendency is statistically correct.

The results of the SHAP values show a significant number of patterns; Thus, a comprehensive treatment is required to make practical use of the results.

B. EVALUATION OF THE WEIGHT OF EACH PIXEL

To introduce the weight of each pixel, it seems good that the absolute value of each pixel is added; If the SHAP value (not the absolute value) for each pixel is simply added, the addition becomes 0 for all pixels, as shown in Figure 10 and already explained in chapter 3-A. To prevent this, SHAP values must be converted into absolute values before addition, and the resulting values indicate the weight of each pixel. Hence, Figure 11 shows the average SHAP values of "the summation over all test data" of "the summation of absolute SHAP values over all labels" on each pixel (average values of the summation of absolute SHAP values over all labels and all test images on each pixel.)

$$\frac{\sum_{\text{test data}} \sum_{\text{label } A}^{D} |(\text{SHAP value})|}{(\text{number of test data})} \tag{4}$$

From the result, the weights of the pixels in the lower side of the picture are relatively high, particularly around the area corresponding to the 0 V line of the PD signals. Therefore, the contribution of the lower-amplitude and wide-phase parts of the images was generally proven to be high in this case. Note that this is a comprehensive tendency and every picture in the data has the same/similar tendency, and that individual evaluation should be considered at the same time in an actual use case.

C. INSIGHT TEST WITH GRADIENT EXPLAINER

Figures 12 (a) and (b) show the results when the trained neural network was changed from a normal network to a narrowed network. A comparison of the properties of each layer of both networks is presented in Table 3. The narrowed network is intentionally "compressed" at its creating CNN layer stage so that the network can only affect a limited area in the original area.

With the SHAP values, Figure 12 successfully visualizes the accessible area of each network, such that the lower

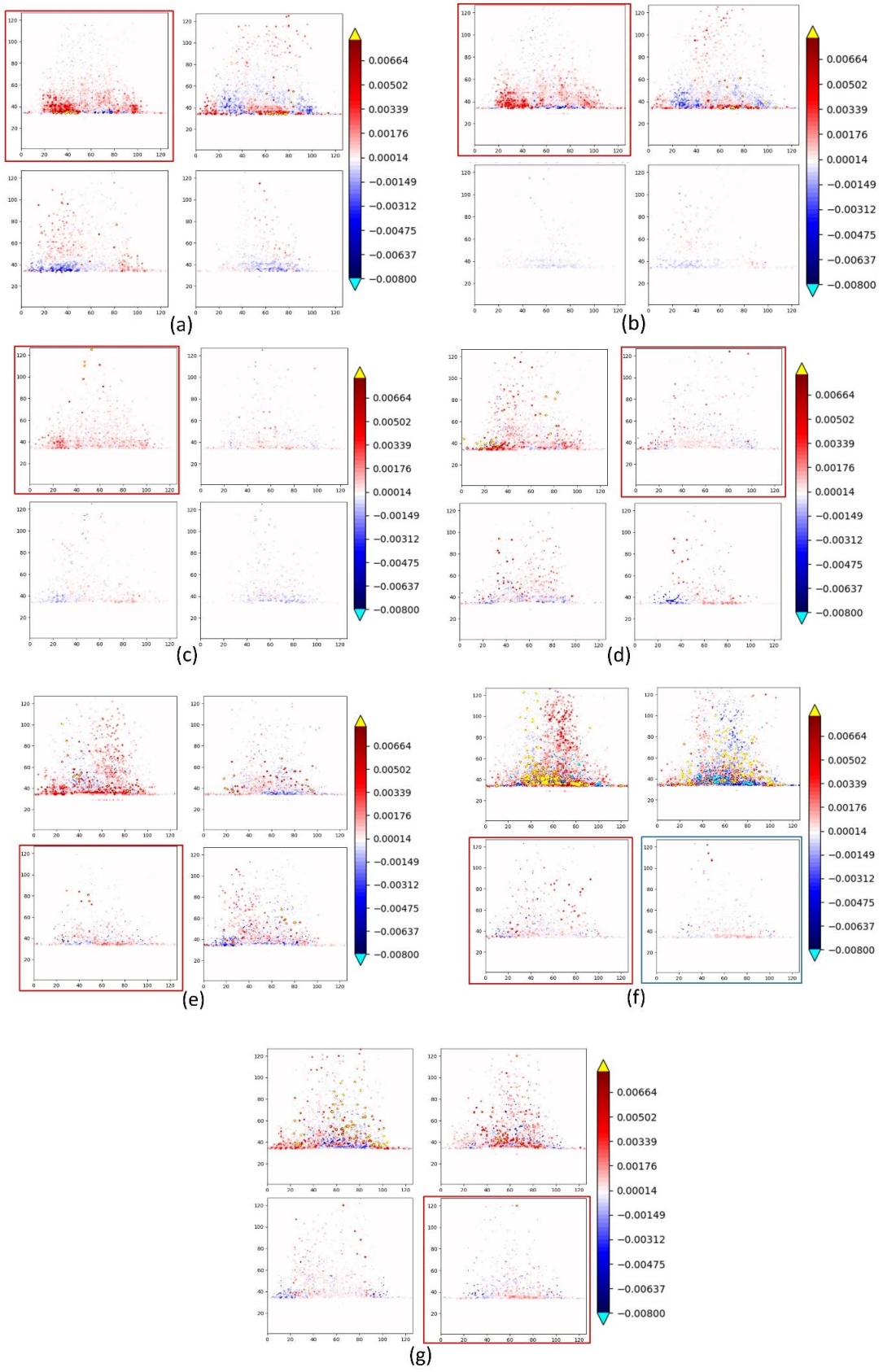


FIGURE 8. SHAP images of PRPD, whose sign from (a) to (g) is correspond to those of Figure 7. Red square surrounding an image indicates that its original PRPD image is predicted by the CNN model. Blue square shown in (f) indicates the image of true label's, which implies failure of prediction of the original image (f).

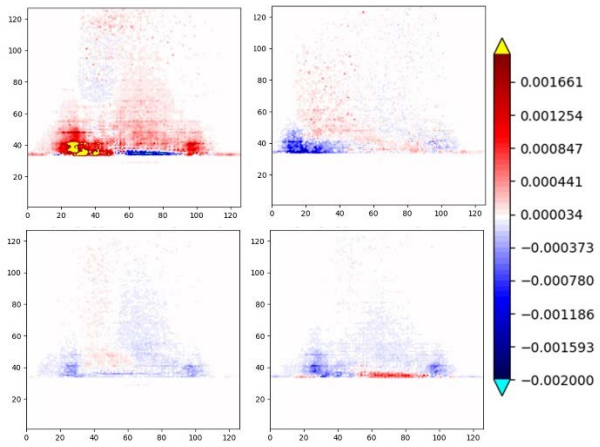


FIGURE 9. Average SHAP values of each label.

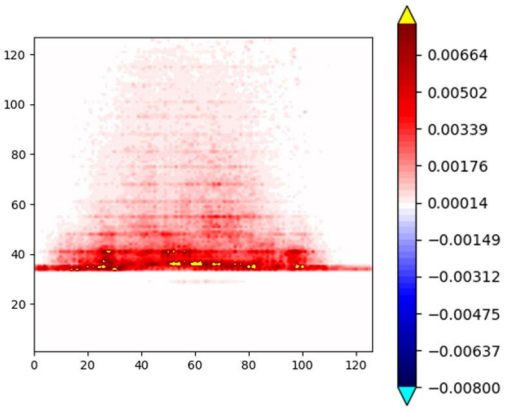


FIGURE 11. Average SHAP values of the summation of absolute SHAP values over all labels and all test images on each pixel.

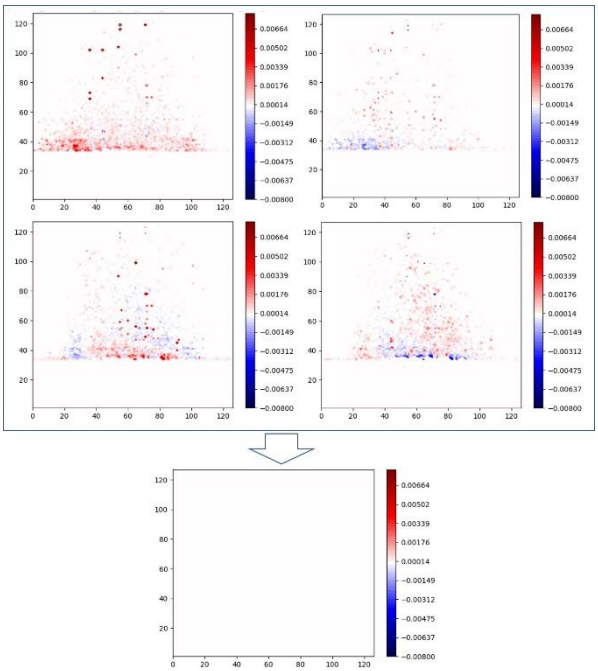


FIGURE 10. Examples of the simple summation of SHAP values from electrode A to D. It turns to be almost 0 in every image.

and right parts of the images on the narrowed network cannot present the SHAP value; Therefore, both parts are not included in the insight of the narrowed network.

D. PARAMETER AND DATA INVULUNERABILITY

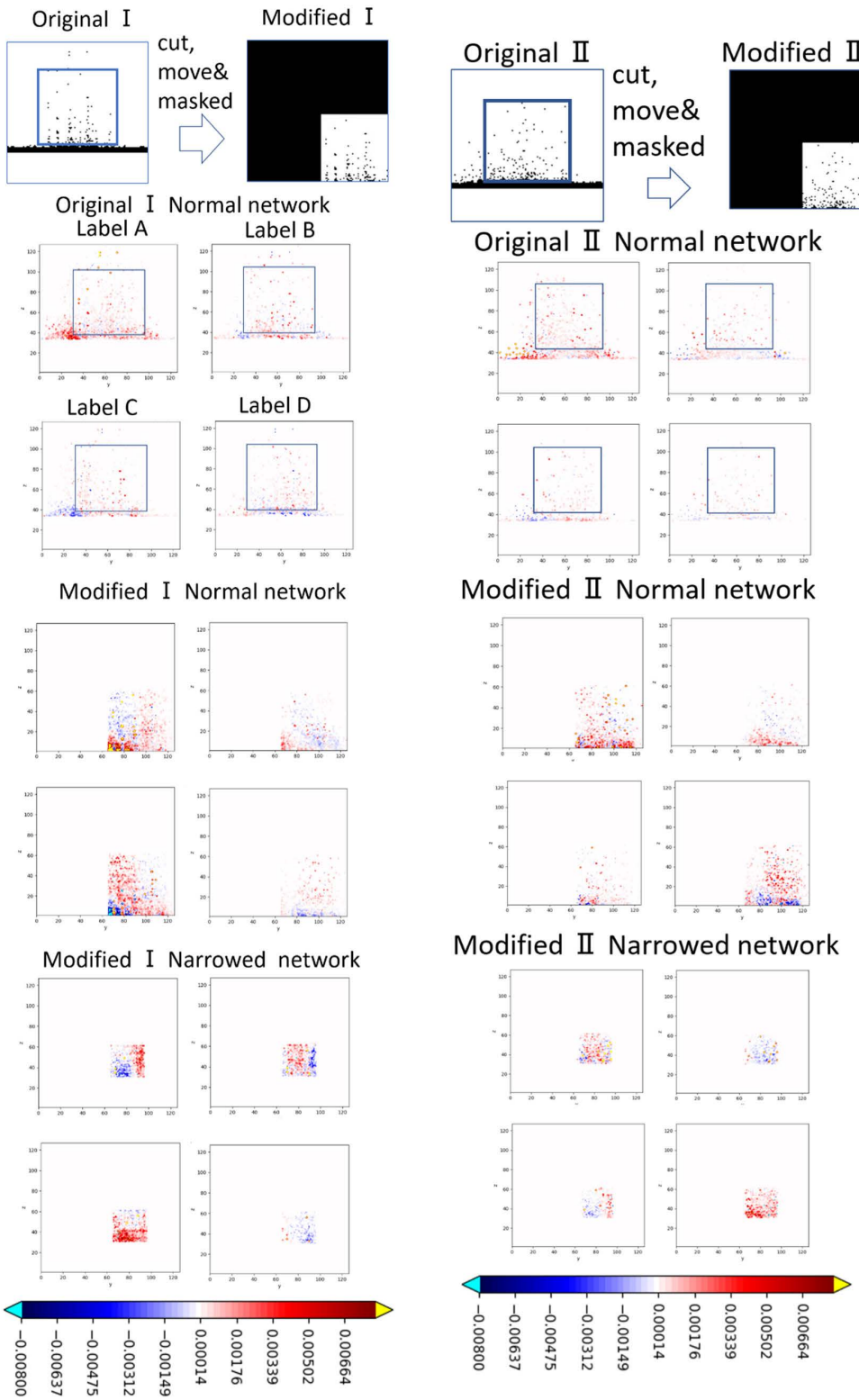
In this section, we verify the results in terms of the hyperparameter invulnerability. For example, LIME [21], which is a technique similar to SHAP, consists of arbitrary parameters; Therefore, it is difficult to use LIME as an invulnerable visualization method for machine learning. On the other hand, two typical hyperparameters are used for calculating SHAP: (1) random seed number, which is used to shuffle and randomize the choice of samples, and (2) number of ''Chimera input data,'' as shown in Figure 6 (in general, approaches similar to Monte Carlo methods require a significant number of samples to stabilize the results.) Hence, parameter tests were conducted according to these parameters.

The results of the verification of the effects of changing a random number used in the calculation algorithm and the number of chimera samples used in the algorithm are shown in Figure 13 and 14, respectively. The extracted results indicate average values of the summation of absolute SHAP values over all labels and all test images on each pixel. (in the similar way as Figure 11.) In addition, Figure 15 shows the result of the effect of changing the learning order of the CNN model in a similar way. These results indicate that the algorithm used in the gradient explainer was sufficiently vulnerable. In other words, the process of the gradient explainer only requires 50 randomized chimera samples for each test and is little affected by the choice of a random number. Specifically, the required number in this case is significantly small compared to the brute-force method, which requires at least $O(2^N)$ per datum-division size N, and choosing a gradient explainer as a method of evaluating the insight of CNN is reasonable in terms of saving time and calculation costs.

The properties shown above indicate that SHAP may allow the vulnerability depending on the creation stage of a CNN itself just like the case of chapter 3-C, however, once the model is created properly, it has a good parameter invulnerability, and there seem to be few arbitrary factors compared to LIME [21].

E. PRACTICAL AND THEORETICAL APPROPRIATENESS

Unfortunately, as mentioned in Chapter 1, there is no way to reveal the real reasons for the judgement made by DNN/CNN, although SHAP can easily provide obvious visualization. In other words, no causal relation was proven, and a correlation between the result of the SHAP value and the actual input data exists [25]; A relationship between the creation process of each network and its output is also required. This is critical for the diagnosis of PD signals because, as mentioned



(a) Insight test or Image I. (b) Insight test of Image II

FIGURE 12. **Insight test of the Two Images. Both images indicate that SHAP can evaluate the defects of the network structure because the outputs of the narrowed network are not fully depicted.**

in Chapter 1, such signal data cannot be distinguished by humans at first glance.

However, from chapters 3-A to 3-C, SHAP is useful in some cases, particularly in making the first test of

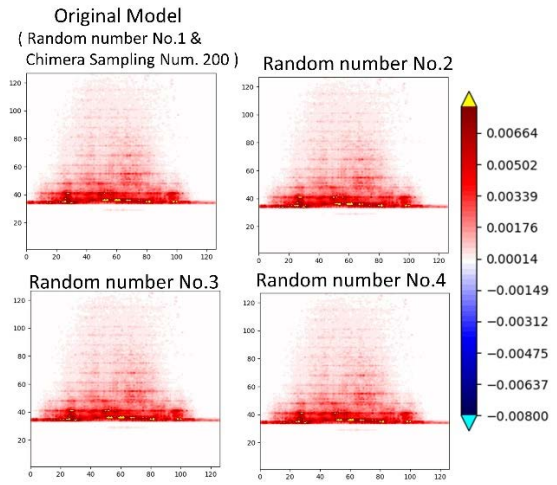


FIGURE 13. Verification of the effect of changing a random number. Average values of the summation of absolute SHAP values over all labels and all test images on each pixel. when the random number inside the calculation algorithm is changed.

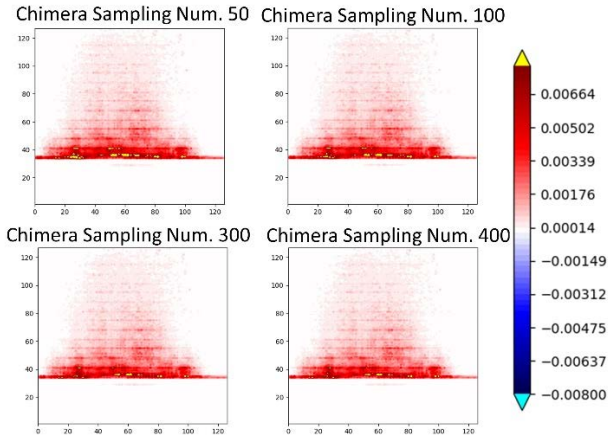


FIGURE 14. Verification of the effect of changing the number of chimera samples. Average values of the summation of absolute SHAP values over all labels and all test images on each pixel. when the number of chimera samples created and used inside the calculation algorithm is changed.

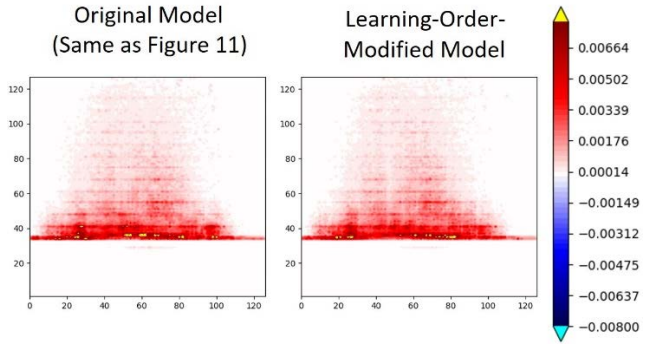


FIGURE 15. Verifying/Comparing the effect of changing the learning order of the CNN model. Average values of the summation of absolute SHAP values over all labels and all test images on each pixel.

a DNN/CNN as easy as possible. Moreover, other similar methods are less theoretically appropriate than SHAP, owing to the mathematical properties, as explained in chapter 2-D.

Ultimately, at this moment, a visualization method for the insight of CNN/DNN should be used to simply "support" the explanation of the obtained prediction results; For example, as shown in chapter 3-C, the SHAP values can evaluate the CNN/DNN whether the structure of the network is correct.

To improve the clarity of the relation between the input and output, more research on structural causal models [26] and/or transfer learning [27] may have the potential to alleviate this disparity. Moreover, checking the prediction results and SHAP values in another way, for example, a random sampling test, seems important until the theoretical analysis of DNN/CNN is sufficiently improved.

IV. CONCLUSION

In this study, phase-resolved PD signals using a CNN for different electrode systems were analyzed, and the relationship between their input and output was evaluated and visualized through SHAP. The motivation of the research was to validate the efficacy of interpretability, even to the targets whose appropriateness human beings cannot easily understand. After explaining the basic properties of the SHAP, the results of a CNN model were validated. Moreover, we determined what should be done in the PRPD analyses with SHAP. Our conclusions are as follows:

1. SHAP can be used to visualize the importance of each pixel in PRPD images. Specifically, some discrete dotted areas in the upper half of the image and around the 0 V line, where the signal values are almost equal to 0, have relatively high moduli of SHAP values.

2. To observe this tendency statistically, data manipulation (summation and taking an average), was conducted. Subsequently, the SHAP values retained a similar tendency to 1. Additionally, the result of the average SHAP values of "the summation over all test data" of "the summation of absolute SHAP values over all labels" on each pixel shows the weight of each pixel, and the area around the 0V line maintains a high modulus of the SHAP values.

3. Insight tests of normal and narrowed networks were conducted. The results indicate that SHAP reflects the network structure well, so that a strange network can be detected via visualization.

4. The effect of changing a random number, the number of chimera samples used in the algorithm and the learning order of the data of the CNN model was tested to verify the parameter/learning-order vulnerabilities. Ultimately. The vulnerabilities are low so that the SHAP has universality in the range of after-preparation of the neural network.

5. Due to the limitation of theoretical appropriateness, SHAP cannot explain the cause of the relationship between input and output. However, SHAP is significantly useful in some use cases, for example, a simplified test in a brute force test's stead or an early test of the network structure.

The above results indicate that it may be possible to order and evaluate the impact of each model based on accuracy

scores, despite not being perfectly quantitative. However, some problems regarding visualization and comprehension remain in CNN/DNN models. Future issues to be studied include the following:

1. Revealing the relationship among the physical phenomena, typical PD circuit model, and data science more directly, including the deterioration of samples and/or types of PD.

2. Connecting what SHAP indicates to the structural causal models and transfer learning.

In addition, actual use also requires the development of ''white box'' machine learning eagerly. The accuracy of a specific model should be improved, and the best explainable method to implement the model for actual use cases, such as the maintenance of electrical appliances and equipment, should be considered. Moreover, to clarify the difference in each insight, the distinguished results obtained by human experts should be compared with the results of the present study. This would help to create a complementary inspection system for humans and machines.

REFERENCES

[1] R. Bartnikas, ''Partial discharges. Their mechanism, detection and measurement,'' *IEEE Trans. Dielectr. Electr. Insul.*, vol. 9, no. 5, pp. 763–808, Oct. 2002, doi: 10.1109/TDEI.2002.1038663.

[2] G. Callender and P. L. Lewin, ''Modeling partial discharge phenomena,'' *IEEE Elect. Insul. Mag.*, vol. 36, no. 2, pp. 29–36, Mar. 2020, doi: 10.1109/MEI.2020.9070114.

[3] L. Niemeyer, ''A generalized approach to partial discharge modeling,'' *IEEE Trans. Dielectr. Electr. Insul.*, vol. 2, no. 4, pp. 510–528, Aug. 1995, doi: 10.1109/94.407017.

[4] M. Hikita, T. Kato, and H. Okubo, ''Partial discharge measurements in SF/sub 6/ and air using phase-resolved pulse-height analysis,'' *IEEE Trans. Dielectr. Electr. Insul.*, vol. 1, no. 2, pp. 276–283, Apr. 1994, doi: 10.1109/94.300260.

[5] Suwarno, Y. Suzuoki, F. Komori, and T. Mizutani, ''Partial discharges due to electrical treeing in polymers: Phase-resolved and time-sequence observation and analysis,'' *J. Phys. D, Appl. Phys.*, vol. 29, no. 11, pp. 2922–2931, Nov. 1996, doi: 10.1088/0022-3727/29/11/028.

[6] R. Altenburger, C. Heitz, and J. Timmer, ''Analysis of phase-resolved partial discharge patterns of voids based on a stochastic process approach,'' *J. Phys. D, Appl. Phys.*, vol. 35, no. 11, pp. 1149–1163, Jun. 2002, doi: 10.1088/0022-3727/35/11/309.

[7] E. Kasinathan, A. Mahajan, and N. Gupta, ''Phase resolved PD patterns in treeing in the presence of voids,'' *J. Electrostatics*, vol. 87, pp. 45–50, Jun. 2017, doi: 10.1016/j.elstat.2017.03.004.

[8] M. Kunicki and A. Cichon, ''Application of a phase resolved partial discharge pattern analysis for acoustic emission method in high voltage insulation systems diagnostics,'' *Arch. Acoust.*, vol. 43, pp. 235–243, 2018, doi: 10.24425/122371.

[9] J. Long, X. Wang, W. Zhou, J. Zhang, D. Dai, and G. Zhu, ''A comprehensive review of signal processing and machine learning technologies for UHF PD detection and diagnosis (I): Preprocessing and localization approaches,'' *IEEE Access*, vol. 9, pp. 69876–69904, 2021, doi: 10.1109/ACCESS.2021.3077483.

[10] G. Li, M. Rong, X. Wang, X. Li, and Y. Li, ''Partial discharge patterns recognition with deep convolutional neural networks,'' in *Proc. Int. Conf. Condition Monit. Diagnosis (CMD)*, 2016, pp. 324–327, doi: 10.1109/CMD.2016.7757816.

[11] X. Wang, H. Huang, Y. Hu, and Y. Yang, ''Partial discharge pattern recognition with data augmentation based on generative adversarial networks,'' in *Proc. Condition Monit. Diagnosis (CMD)*, 2018, pp. 1–4, doi: 10.1109/CMD.2018.8535718.

[12] S. Mantach, A. Ashraf, H. Janani, and B. Kordi, ''A convolutional neural network-based model for multi-source and single-source partial discharge pattern classification using only single-source training set,'' *Energies*, vol. 14, no. 5, p. 1355, Mar. 2021, doi: 10.3390/en14051355.

[13] Barrios, Buldain, Comech, Gilbert, and Orue, ''Partial discharge classification using deep learning methods—Survey of recent progress,'' *Energies*, vol. 12, no. 13, p. 2485, Jun. 2019, doi: 10.3390/en12132485.

[14] S. Iwata and R. Kitani, ''Phase-resolved partial discharge analysis of different types of electrode systems using machine learning classification,'' *Electr. Eng.*, vol. 103, no. 6, pp. 3189–3199, Dec. 2021, doi: 10.1007/s00202-021-01306-5.

[15] C. Rudin, ''Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,'' 2018, *arXiv:1811.10154*.

[16] M. Tulio Ribeiro, S. Singh, and C. Guestrin, '''Why should i trust you?': Explaining the predictions of any classifier,'' 2016, *arXiv:1602.04938*.

[17] S. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,'' 2017, *arXiv:1705.07874*.

[18] 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,'' *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2020, doi: 10.1007/s11263-019-01228-7.

[19] S. Mantach, P. Gill, D. R. Oliver, A. Ashraf, and B. Kordi, ''An interpretable CNN model for classification of partial discharge waveforms in 3D-printed dielectric samples with different void sizes,'' *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11739–11750, Jul. 2022, doi: 10.1007/s00521-022-07066-y.

[20] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, ''Sanity checks for saliency maps,'' 2018, *arXiv:1810.03292*.

[21] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. 2022. Accessed: Jan. 12, 2023. [Online]. Avaliable: https://christophm.github.io/interpretable-ml-book/

[22] L. Shapley, ''A value for N-person games,'' in *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, A. Roth, Ed. Cambridge, U.K.: Cambridge Univ. Press., 1988, pp. 31-40, doi: 10.1017/CBO9780511528446.003.

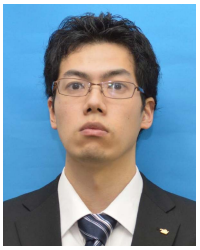
[23] S. Lundberg. Accessed: Oct. 20, 2022. [Online]. Avaliable: https://github.com/slundberg/shap

[24] M. Sundararajan, A. Taly, and Q. Yan, ''Axiomatic attribution for deep networks,'' 2017, *arXiv:1703.01365*.

[25] E. Dillon, J. LaRiviere, S. Lundberg, J. Roth, and V. Syrgkanis. (May 18, 2021). *Be Careful When Interpreting Predictive Models in Search of Causal Insights: A Joint Article About Causality and Interpretable Machine Learning With Eleanor Dillon, Jacob LaRiviere, Scott Lundberg, Jonathan Roth, and Vasilis Syrgkanis From Microsoft*. Accessed: Oct. 20, 2022. [Online]. Avaliable: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%C2%A0insights.html

[26] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[27] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2020.



RYOTA KITANI received the B.Eng. and M.S. (energy science) degrees from Kyoto University, in 2012 and 2015, respectively. Currently, he is a Research Scientist with the Osaka Research Institute of Industrial Science and Technology. His research interests include electrical degradation, continuum mechanics, computer science, and reliability of electrical product and systems.

SHINYA IWATA (Member, IEEE) received the B.Eng. degree from the Kyoto Institute of Technology, in 2005, and the M.Eng. and Ph.D. degrees from the University of Tokyo, in 2007 and 2011, respectively. Currently, he is a Senior Research Scientist with the Osaka Research Institute of Industrial Science and Technology. His research interests include electrical degradation, partial discharge, molecular dynamics, quantum chemistry, and reliability of electrical product and systems.