applied sciences

MDPI

Article

Partial Discharge Pattern Recognition of Transformers Based on MobileNets Convolutional Neural Network

Yuanyuan Sun [1,*], Shuo Ma [1], Shengya Sun [1], Ping Liu [2], Lina Zhang [3], Jun Ouyang [4] and Xianfeng Ni [4]

[1] School of Electrical and Engineering, Shandong University, Jinan 250061, China; mashuosdu@foxmail.com (S.M.); shengyasun@foxmail.com (S.S.)
[2] CNOOC Energy Development Equipment Technology Co., Ltd., Tianjin 300452, China; liuping5@cnooc.com.cn
[3] CNOOC Research Institute Ltd., Beijing 100027, China; zhangln@cnooc.com.cn
[4] CNOOC(China) Tianjin Branch Co., Ltd., Tianjin 300452, China; ouyj3@cnooc.com.cn (J.O.); nixf2@cnooc.com.cn (X.N.)
* Correspondence: sunyy@sdu.edu.cn

Abstract: The power system on the offshore platform is of great importance since it is the power source for oil and gas exploitation, procession and transportation. Transformers constitute key equipment in the power system, and partial discharge (PD) is its most common fault that should be monitored and identified ın a timely and accurate manner. However, the existing PD classifiers cannot meet the demand for real-time online monitoring due to their disadvantages of high memory consumption and poor timeliness. Therefore, a new MobileNets convolutional neural network (MCNN) model is proposed to identify the PD pattern of transformers based on the phase resolved partial discharge (PRPD) spectrum. The model has the advantages of low computational complexity, fast reasoning speed and excellent classification performance. Firstly, we make four typical defect models of PD and conduct a test in a laboratory to collect the PRPD spectra as the data sample. In order to further improve the feature expression ability and recognition accuracy of the model, the lightweight attention mechanism Squeeze-and-Excitation (SE) module and the nonlinear function hard-swish (h-swish) are added after constructing the MCNN model to eliminate the potential accuracy loss in PD pattern recognition. The MCNN model is trained and tested with the pre-processed PRPD spectrum, and a variety of methods are used to visualize the model to verify the effectiveness of the model. Finally, the performance of MCNN is compared with many existing PD pattern recognition models based on convolutional neural network (CNN), the results show that the proposed MCNN can further reduce the number of parameters of the model and improve the calculation speed to achieve the best performance on the premise of good recognition accuracy.

Keywords: transformer; partial discharge (PD); pattern recognition; MobileNets convolution neural network

Citation: Sun, Y.; Ma, S.; Sun, S.; Liu, P.; Zhang, L.; Ouyang, J.; Ni, X. Partial Discharge Pattern Recognition of Transformers Based on MobileNets Convolutional Neural Network. *Appl. Sci.* **2021**, *11*, 6984. https://doi.org/10.3390/app11156984

Received: 19 May 2021
Accepted: 27 July 2021
Published: 29 July 2021

1. Introduction

As one of the core pieces of equipment, the transformer directly affects the safe, efficient and economic operation of the power system, so it is especially important to ensure the safe and stable operation of the transformer. Partial discharge (PD) is the main cause of transformer insulation damage, and the effect of different types of partial discharge on the transformer insulation system is different, so it is of great practical significance to effectively identify the discharge pattern and take corresponding measures [1,2].

The traditional PD pattern recognition method mainly includes three parts, which include obtaining the PD data, extracting the PD features and training the pattern recognition classifier [3]. Scholars have done a lot of research on the feature extraction and pattern recognition classifier. The common characteristic parameters include statistical parameters, waveform parameters, fractal characteristic parameters, wavelet parameters, moment characteristic parameters and so on [4–9]. For pattern recognition methods, back-propagation

neural network (BPNN) and support vector machine (SVM) are widely used [10–12]. However, these machine learning algorithms, which rely on the artificial construction of PD features, are highly subjective and have large recognition errors [13].

With the development of artificial intelligence and computer vision technology, the application of deep neural network to PD pattern recognition can be used to overcome the shortcomings of traditional methods. Although the convolutional neural network (CNN) has the advantage of requiring no artificial computation of signal features, it has not been widely used in the PD pattern recognition. At present, the basic research of CNN includes analyzing the performance differences of convolutional network under different layers, activation functions and pooling modes.

The CNN model has a better recognition effect, which improves the accuracy compared with SVM and BPNN [14]. Some literature uses the time domain waveform image of PD as data samples to conduct supervised training on the constructed one-dimensional CNN model [15]. In addition, PRPS spectra obtained from the field experiments and simulation are used as input data, and classical CNN model LeNet-5 is used as classifier [16]. However, these models have fewer layers and cannot fully extract the PD features. The depth model has good recognition accuracy, but it has a large number of parameters and high complexity in terms of model size, which leads to huge network capacity, slow operation speed and high computational resource consumption. Therefore, it is not suitable to embed the deep learning models into devices and it is difficult to deploy them on miniaturized hardware platforms.

In order to solve the above problems, a transformer PD pattern recognition method based on MobileNets convolution neural network model is proposed. The model introduces the lightweight attention mechanism Squeeze—and—Excitation (SE) module and h-swish activation function with stronger nonlinear performance based on the use of depth separable convolution module and inverse residual structure. Then, the MCNN network is trained and tested by different types of PD data obtained from laboratory experiments. Analysis results show that the MCNN can significantly reduce the number of parameters and the computational complexity of the model. In addition, MCNN can further improve the recognition accuracy compared with the existing convolutional neural network, which is often used for PD type classification. Finally, the feature maps of the convolution output of the last layer in the MCNN are visually calculated to explain the classification process and verify the function of the model.

The remainder of this paper is organized as follows. Data acquisition of PD in transformer is illustrated in Section 2.1; then, the CNN and NCNN (noisy convolution neural network) principle methodology are introduced in Section 2.2. Mobilenets Convolution Neural Network is illustrated in Section 2.3, and Lightweight MCNN Block is shown in Section 2.4. Training and Testing of MCNN are given in Section 2.5. Section 3 highlights the MCNN with a discussion about the experimental results and the comparison with a few previous pattern recognition methods. Finally, conclusions of the paper are given in Section 4.

2. PD Pattern Recognition Based on MCNN

2.1. PD Data Acquisition of Transformer

According to the common types of insulation defects in transformer operation, four typical PD models are made. These PD models are used to simulate the transformer tip discharge, surface discharge, air gap discharge and suspended discharge. In the experiment, a PD detection system based on ultra-high frequency (UHF) sensors is used. The sampling rate of the system is 4 GHz, and the bandwidth is 1 GHz.

During the experiment, the voltage is stepped up and (UHF) sensors are used to collect PD data from the four PD models. The collected PD data is used to build a sample database of PRPD maps for subsequent feature extraction and pattern recognition. The specific test steps are listed as follows:

(1) A PD insulation defect model is selected and connected to the partial discharge test circuit.
(2) The background noise of the laboratory is determined. Before the voltage is applied, the signal detected by the signal acquisition system is the background noise of the PD test environment.
(3) The inception voltage of the PD is determined. The voltage of the regulating transformer is increased uniformly and slowly in steps. When the PD signal is detected by the monitoring instrument, the corresponding voltage is recorded as the initial voltage U_1.
(4) The PD data is generated and collected at different voltage supplies. The voltage is continuously increased, and when it is greater than 1.5 times U_1, a stable and distinct PD signal series is recorded in the monitoring system. The voltage is then still increased to approximately twice U_1. The PD data generated during this process is collected by the UHF sensors. The sensors can record the partial discharge signals of different intensities generated by the partial discharge model at different voltage levels.
(5) Different types of insulation defect models are replaced. After completing the PD test of the current PD insulation defect model, the voltage is slowly lowered and the power is turned off. Different types of insulation defect models are replaced. After completing the PD test for the current PD insulation defect model, the voltage is slowly reduced and the power is turned off. The discharge bar is used to discharge the test circuit, and the insulation defect model is replaced with another one. Then, the experiment steps (2) to (5) are then repeated until all four types of PD data are collected.

The PRPD spectra obtained from the PD test are shown in Figure 1, in which each spectrum contains 50 power frequency cycles of discharge data. In the PRPD spectra, by observing the vertical and horizontal coordinates of a pulse point, we can directly obtain information about the phase and amplitude of the pulse.

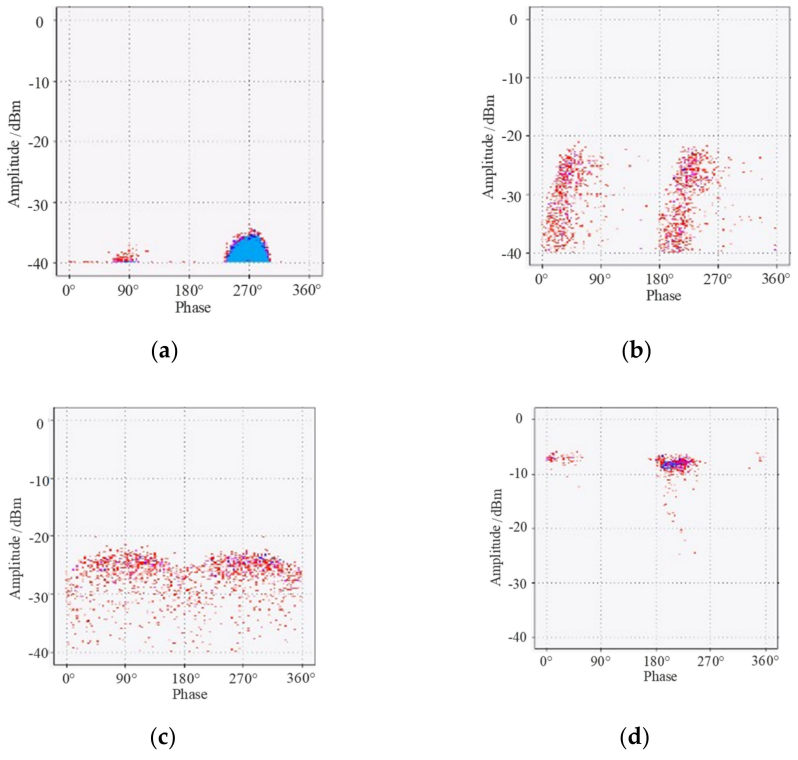


Figure 1. PRPD data of different PD types. (**a**) tip discharge, (**b**) surface discharge, (**c**) air gap discharge, and (**d**) suspended discharge.

In the experimental process of the tip discharge model, with the increase of applied voltage, a cluster of PD signals appear at the positions with the phase of 270°. This is because in 50 Hz AC, the AC voltage peaks at 270° in phase. The electrons are most active and more easily repelled at the moment when the voltage is at a negative peak, which coincides with the easy excitation of a negative polarity tip. With the further increase of voltage, the second cluster discharge signal appeared at the 90° phase position. This is because the voltage reaches a positive peak at 90° phase. When the voltage applied on the conductor is positive polarity, more energy is required for electron excitation, so the overall discharge amplitude at this time is lower than the first cluster signal at position 270°. Furthermore, the distribution phase width is narrower than that at position 270°.

Suspended discharge is produced by suspended potential. Taking loose bolts as an example: when bolts and conductors are not in contact, the voltage at both ends of the middle gap is the same. In this case, high energy will be released when PD occurs, resulting in a high discharge amplitude and a consistent discharge height. This is an obvious distribution characteristic of suspended discharge. In addition, in the PRPD spectra, both air gap discharge and surface discharge are evenly distributed signals with different amplitudes and wide phase distributions. It can be seen that the distribution of discharge points in the PRPD spectra of different types of partial discharges varies greatly, and this distribution difference can be translated into pixel space distribution features that can be processed and extracted by CNN for pattern recognition.

2.2. Convolution Neural Network

Convolution neural network (CNN) is a deep learning framework, which is proposed by the mechanism of the biological receptive field and developed rapidly in computer vision. The CNN reduces the parameters of model training and extracts the relevant semantic features of the input image effectively using sparse connectivity and weights sharing. A typical CNN structure consists of one input layer, one or more alternately connected convolution and pooling layers, a full connection layer, and an output layer. The input image is first processed continuously through the convolutional layer and pooling layer; then the input image is classified in the full connection layer; and, finally, the recognition result is output.

The convolution layer contains multiple convolution kernels, each of which is a filter to extract the features of the input data. Each neuron in the convolution layer is connected with the local receptive field of the upper layer, and the convolution operation is carried out to determine the relationship of each feature. The formula is shown in (1):

$$y_j^l = f\left(\sum_{i \in \mathrm{M}_j}^{n} x_i^{l-1} \times \omega_{ij}^l + \mathrm{b}_j^l\right) \tag{1}$$

where f is the activation function and $\times$ is the convolution operation. x_i^{l-1} the step is the output feature of the previous convolution layer. M_j is the set of input feature maps, ω is the convolution kernel, b is the offset value of each map, and l is the number of network layers.

The function of the pooling layer is used for down-sampling operation, which can select the output of the feature from the convolution layer, thus reducing the feature dimension. The nonlinear activation function is added between the convolutional layer and the pooling layer to improve the expression ability of the model. Common activation functions include Sigmoid, Tanh, and ReLU. In the fully connected layer, the final feature maps are reorganized into lower-dimensional vectors for the training and classification.

2.3. Mobilenets Convolution Neural Network

With the continuous increase of the number of layers of convolution neural network, the number of parameters and the amount of calculation also increase. This leads to a complex model structure and slow computing speed, which is not conducive to the application

and development of CNNs in the field of pattern recognition with real-time online requirements. To solve these problems, Google proposed the MobileNets convolution neural network (MCNN) in 2017 [17]. A Depthwise Separable Convolution structure is used in the unit block of the model, which greatly reduces the number of parameters and computation of the model and achieves more efficient performance through the reverse residual structure with a linear bottleneck. On this basis, we add the Squeeze-and-Excitation structure to the model used in the paper and adjust the activation function to further improve the learning ability and timeliness of the model [18].

In the MobileNets model, the depth separable convolution is used instead of traditional standard convolution. Depth separable convolution mainly includes two parts: depth convolution and 1×1 point convolution [19]. This decomposition is linear, which can reduce the computation and model parameters, as shown in Figure 2.

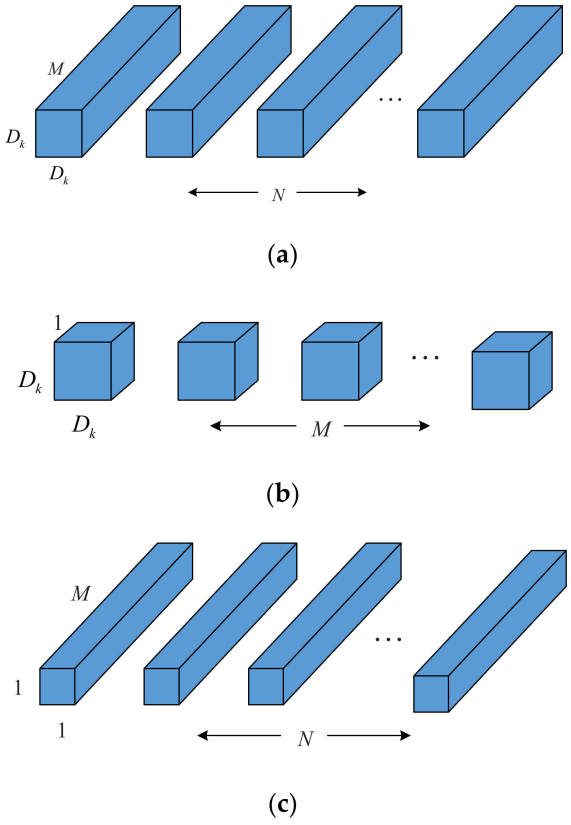


Figure 2. Standard convolution and depth detachable convolution. (**a**) Standard convolution filters. (**b**) Depth convolution. (**c**) 1×1 convolution.

Figure 2a is the traditional standard convolution, the number of convolution kernels is N, and the number of input channels is M. The traditional standard convolution can be decomposed into the sum of the depth convolution shown in (b) and the point-by-point convolution shown in (c). The MobileNets can reduce the number of parameters and improve the computational efficiency by using the depth separable convolution structure.

Theoretically, the recognition effect of the model with more layers of the network will be better. But practical experience shows that the difficulty of network training and optimization increases with the increase of layers, which leads to network degradation, and the effect will be worse than that of the relatively shallow network. In the traditional residual structure shown in Figure 3, the above problem is solved by introducing a shortcut connection structure. The dimension of the feature channel is first reduced to the lower dimension and extracted features, then extended to the high dimension. To improve the classification performance of the network and alleviate the problem of gradient vanishing

in multi-layer backpropagation, the reverse residual structure with linear bottleneck is introduced into the MobileNets, as shown in Figure 4. The feature channel dimension in the reverse residual structure is first extended to the higher dimension to extract features and then reduced to the lower dimension. This structure can increase the nonlinear expression capacity of each feature channel by deep convolution in the high-dimensional feature space and maintain the compact feature expression in the input and output information, which makes the structure of the model more efficient.

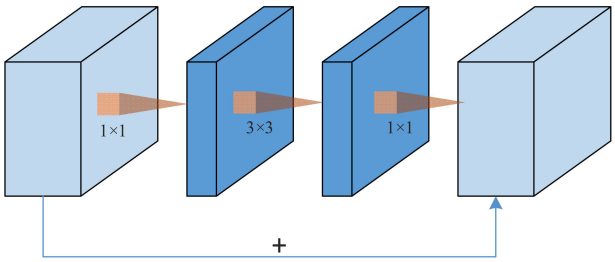


Figure 3. ResNet residual unit.

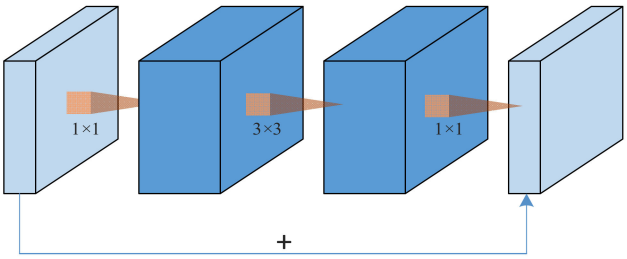


Figure 4. Reverse residual unit.

The feature transfer in the traditional convolution network is to transfer the weight of the feature graph to the next layer, and the Squeeze-and-Excitation (SE) module can adaptively correct the importance of features between channels according to the global loss function of the network, to increase the weight of effective features. This structure can strengthen the learning ability of the network and improve the accuracy of the model by strengthening the important features.

2.4. Lightweight MCNN Block

The lightweight and efficient convolution module adopted in the MCNN is shown in Figure 5. Firstly, the number of input feature channels is extended to a large number of intermediate layers by 1×1 point convolution, then features are extracted and optimized by 3×3 depth convolution. Finally, 1×1 point convolution is used to compress the features to the size given by the output channel. The specific implementation of each module of constructing this convolution network corresponding to Figure 6 is shown in Table 1. The number of input channels of the convolution module is N, and the number of output channels is M. For the input feature of a certain size $W \times H$, the expansion factor is t, the convolution kernel size is 3, s denotes stride, and NL denotes the nonlinear function.

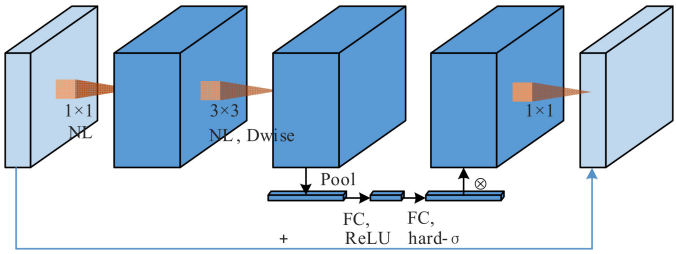


Figure 5. Structural Diagram of MobileNets Block.

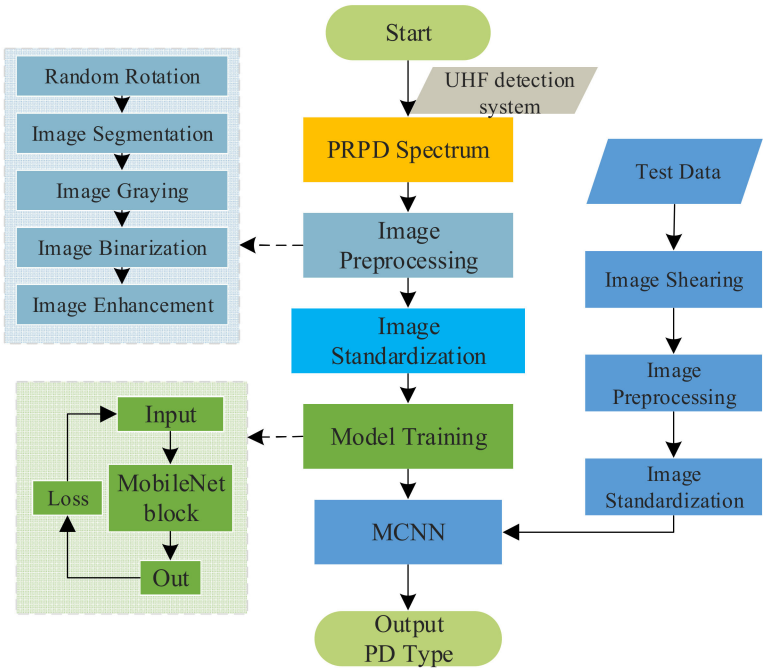


Figure 6. The procedure of PD pattern recognition by MCNN.

Table 1. Core module construction of MOBILENETS.

Input	Operator	Output
W × H × N	1 × 1 conv2d, NL	W × H × t N
W × H × t N	3 × 3 dwise s = s, NL	W/s × H/s × t N
W/s × H/s × t N	Linear 1 × 1 conv2d	W/s × H/s × M

Table 2 shows the overall structural parameters of the compact and computationally efficient lightweight MCNN. The rows in Table 2 describe the specific structural configuration of each convolution layer in the model [19]. The size of the input data of the first layer is 224 × 224 × 3, in which 3 represents the 3-channel image in RGB format. The standard convolution operation is carried out on this input data, in which the number of channels of the kernel is set to 16, which can minimize the operation time on the premise of ensuring accuracy. The 2nd to 12th convolution layers in Table 2 all adopt the reverse residual module with a linear bottleneck. The 13th layer is point-by-point convolution. The 14th layer is the pooling layer, which converts the output feature from the convolution layer pooling to 1 × 1 × 576. Then, the fully connected layer is used to reconstruct the final feature image into a lower-dimensional vector. Finally, *k* types of identification information are output in the 16th layer.

Table 2. Core module construction of MOBILENETS.

Layer Num	Input	Operator	Exp Size	Out	SE	NL	s
1	2242 × 3	conv2d, 3 × 3	-	16	-	HS	2
2	1122 × 16	Bneck, 3 × 3	16	16	√	RE	2
3	562 × 16	Bneck, 3 × 3	72	24	-	RE	2
4	282 × 24	Bneck, 3 × 3	88	24	-	RE	1
5	282 × 24	Bneck, 5 × 5	96	40	√	HS	2
6	142 × 40	Bneck, 5 × 5	240	40	√	HS	1
7	142 × 40	Bneck, 5 × 5	240	40	√	HS	1
8	142 × 40	Bneck, 5 × 5	120	48	√	HS	1
9	142 × 48	Bneck, 5 × 5	144	48	√	HS	1
10	142 × 48	Bneck, 5 × 5	288	96	√	HS	2
11	72 × 96	Bneck, 5 × 5	576	96	√	HS	1
12	72 × 96	Bneck, 5 × 5	576	96	√	HS	1
13	72 × 96	conv2d, 1 × 1	-	576	√	HS	1
14	72 × 576	Pool, 7 × 7	-	-	-	-	1
15	12 × 576	FC, NBN	-	1280	-	HS	1
16	12 × 1280	FC, NBN	-	k	-	-	1

The "conv2d" in the Table 2 denotes the two-dimensional convolution operation; the "Bneck" is the bottleneck structure, the "NBN" indicates that the batch is not normalized, and the module with "SE" lightweight attention structure is marked in the "SE" column. The type of non-linear used in each convolution layer is introduced in "NL", in which "HS" means to use h-swish, "RE" means to use standard ReLU6.

In some convolution layers of MCNN, h-swish is used to replace swish. This is because although swish can effectively improve the accuracy of the network, it will lead to a large amount of calculation and is takes too much time to contain the sigmoid function due to its characteristics of no upper bound and lower bound, and non-monotonicity. The h-swish can not only better approximate the original swish to reduce the amount of computation but can also improve the computational efficiency in the quantization mode, which makes the model more suitable for the embedded, low-power environment. The mathematical definitions of the swish and the h-swish are as follows. x is the input signal.

$$\text{swish } x = X \cdot \frac{1}{1 + e^{-x}} \tag{2}$$

$$\text{h} - \text{swish}(x) = X \cdot \frac{\text{ReLU6}(x+3)}{6} \tag{3}$$

$$\text{ReLU6}(x) = \begin{cases} 0 & (x < 0) \\ x & (0 \le x \le 6) \\ 6 & (x \ge 6) \end{cases} \tag{4}$$

2.5. Training and Testing of MCNN

In this paper, MCNN is used to identify the type of PD in the transformer. To further improve the performance of the model and avoid over-fitting, we first expand the PD data obtained, to effectively increase the number and diversity of training samples. In the output layer of the model, the SoftMax is used as the classifier, and the one-hot encoding is used to encode the categories of four types of PD. The specific steps of PD pattern recognition of transformer based on MCNN are as follows:

(1) Data acquisition. According to the data collected by the discharge experiment, the PRPD spectrum is generated, and the collected PD images are randomly divided into the training set and test set, accounting for 80% and 20% of the total samples respectively.

(2) Data preprocessing. Firstly, the PRPD spectrum of PD is randomly allocated to obtain a 224 × 224 × 3 three-channel RGB image, and then the image is flipped randomly and converted into a floating-point tensor with a value between 0~1.
(3) Data enhancement. The sample of the original data is expanded, including random rotation, segmentation, scaling, and other operations. To improve the generalization ability of the model, 20% of the training data is randomly selected by using data enhancement for image generation.
(4) Data standardization. To establish the comparability of the data, the standardized method is used to normalize the input data.
(5) Model training. The loss function used in the training process is the Cross-Entropy Loss, and the optimization algorithm adopted is the Stochastic Gradient Descent. Dropout and Batch Normalization are used to improve the training performance.
(6) Model testing. The trained model is tested on the test set. The purpose of the model testing phase is to verify the generalization ability of the model, and the recognition accuracy of the model to PD data can be obtained.

The overall flow chart of PD recognition based on the MobileNets convolution neural network designed above is shown in Figure 6.

3. Results

3.1. Recognition Results and Performance Analysis

3.1.1. The Effect of Initial Parameters on Network Performance

In this paper, the PyCharm framework based on Python is used to build the network model; the platform is configured in the computer with Inteli5-6500CPU, 8 GB of memory, and win10 64-bit operating system. Based on the PRPD images measured by the transformer simulation experiment, the model is trained according to the steps in Section 2.5; finally, the model construction is completed.

The initial network parameters such as learning rate (the method for adjusting input weights of neural networks) and epoch (the process of training all training samples once) will have a certain impact on the classification performance of the model. To prevent the step size from jumping out of the optimal area in the later stage of training, the learning rate is dynamically adjusted during the training process. The decay factor of the weight is set to 1×10^5, the momentum of the optimizer is set to 0.9, the learning rate is reduced in the specified iterative period. When the number of the training cycle is 100 and the learning rate is set to be reduced by 10 times in the 30th, 60th, and 90th cycles, the comprehensive performance of the model is optimal. When the training period is greater than 100, the accuracy is slightly improved, but the generalization ability of the model will be reduced due to overfitting. Therefore, the training cycle of the model is set to 100, and different initial learning rates are adjusted in turn to obtain the variation trend of classification performance under the corresponding conditions, as shown in Figure 7.

From the analysis of the above results, it can be concluded that when the learning rate is greater than 0.2, each training process will cross the optimal area due to the large step, resulting in the poor overall performance of the model and even non-convergence. Because the learning rate will be dynamically adjusted, the too-small initial value will also reduce the accuracy of recognition to some extent, and with the decrease of the learning rate, the training time will also increase greatly. For the MCNN using PD data for training designed in this paper, when the initial learning rate is set to 0.15 and the number of iterations is 100, the model achieves the best classification effect and has the strongest generalization ability.

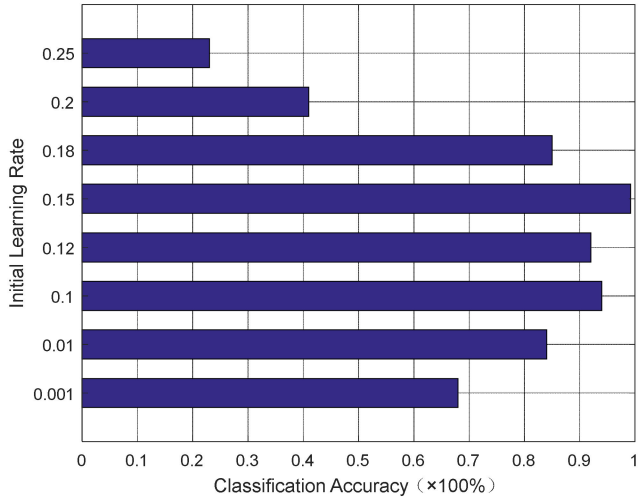


Figure 7. The influence of initial learning rates on the accuracy of MCNN.

3.1.2. Classification Performance Comparison of Models

To fully verify the accuracy of this method, a total of 1200 groups of data obtained from four kinds of defect models are selected as test samples. The MCNN is compared with the other six models in terms of performance, including three deep neural network models, AlexNet; VGG16; Resnet18; and three lightweight convolutional networks, including SqueezeNet1.0, DenseNet121, and MobileNetsV2. The recognition accuracy (Accuracy) is used as a unified index to quantitatively analyze the performance of each model, and the pattern recognition results are shown in Table 3.

Table 3. Core module construction of MOBILENETS.

Methods	Recognition Accuracy/%				Average Accuracy
	Tip Discharge	Surface Discharge	Air-Gap Discharge	Suspended Discharge	
AlexNet	92.52	82.42	81.81	96.71	88.37%
ResNet-18	94.32	88.42	90.22	100	93.24%
VGG16	93.28	88.34	86.79	97.36	91.44%
SqueezeNet1.0	91.01	82.17	80.64	95.33	87.29%
DenseNet121	100	94.37	95.54	100	97.48%
MobileNetsV2	99.41	90.51	93.24	100	95.79%
OurModel	100	96.31	98.53	100	98.71%

As can be seen from the data in Table 3, the overall recognition rate of the MCNN is 98.55%, which is significantly higher than that of the three kinds of depth convolution models: AlexNet, ResNet18, and VGG16 model. Compared with the MobileNetsV2 network without SE module, the MCNN designed in this paper has improved the recognition ability. Among the four compressed convolution models, the recognition accuracy of the model designed in this paper is the highest. In addition, the classification results obtained from the seven models are ideal, because the PRPD spectra are used as the input sample, and the spectra contain PD information over a period of time, so they can better reflect the characteristics of different types of PD and promote the recognition rate of each model generally higher. Among all the types of PD, the recognition rate of suspended discharge and tip discharge is relatively high. This is because the data analysis of PD detection in Section 2.1 shows that the signal spectra of these two types have obvious characteristics of amplitude and phase, while the statistical spectra of air gap discharge and surface discharge are similar.

3.1.3. Complexity Analysis of Models

As PD detection instruments may be carried by operators to various environments for signal measurement, under the premise of ensuring the accuracy of network identification, the resource consumption and running speed of the model are also key indicators to measure the comprehensive performance of the model. To further compare and analyze the performance advantages of the MCNN model proposed in this paper compared with other models, the total parameters of seven models and the storage space occupied by the models are counted, and the results are shown in Table 4.

Table 4. Core module construction of MOBILENETS.

Methods	Parameter/Million	Weight Storage/MB
AlexNet	57.02	217.51
ResNet-18	11.18	42.65
VGG16	134.29	512.28
SqueezeNet1.0	0.74	2.82
DenseNet121	6.96	26.55
MobileNetsV2	2.23	8.51
OurModel	0.80	3.05

As can be seen from the data in Table 4, the number of parameters of the MCNN model is only 8000, which is significantly smaller than that of the three types of depth convolution neural networks, while the other three compression models have a good performance in terms of spatial complexity. In addition, although the MCNN model takes up only 3.05 MB of memory, slightly more than the SqueezeNet1.0 model, it is still relatively small. Therefore, it is suitable for mobile devices or intelligent terminal instruments with limited volume and processor performance.

If the reaction time of the model is too long in practice, it will be unable to deal with the PD timely and quickly. Therefore, in addition to the recognition accuracy and memory footprint of the model, the corresponding test time of different models will directly determine their application performance in the actual situation.

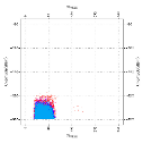
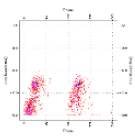
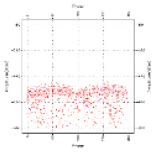
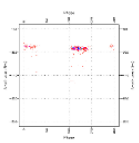
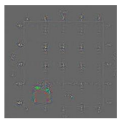
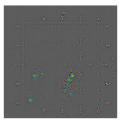
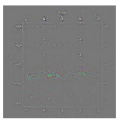
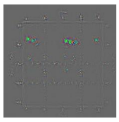
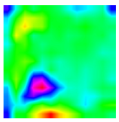
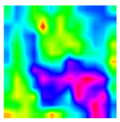
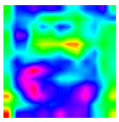
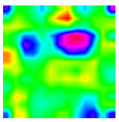
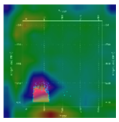
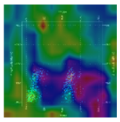
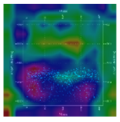
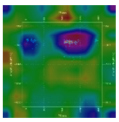
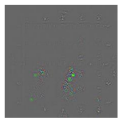
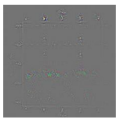
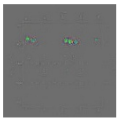
3.1.4. Visual Analysis of MCNN

At present, the deep convolutional neural network has been widely used in classification and recognition based on massive data in various fields. However, the processing process of input data and results in CNN is similar to that in a "black box". When we apply each model, we tend to ignore the explanation of the feature training process between the layers within the network, which leads to the questioning of the reliability of the network model after training. Therefore, it is necessary to further clarify whether the model can truly identify the target object. We use four visualization algorithms to show the distribution of the eigenvalues extracted by MCNN.

After many times of convolution and pooling (feature extraction layer), the deep convolution neural network often contains the most abundant spatial and semantic information in the last convolution layer. The last convolution layer can extract the features contained in the input image to the greatest extent, and it has the most direct impact on the final classification results. Therefore, to verify the function of the model and more vividly understand the classification process of the MCNN model for PD pattern recognition, the Guided-Backpropagation (Guided-BP) algorithm is used to visually calculate the feature map of the convolution output of the last layer of the MCNN model. The CAM method can show the decision basis of the model in the form of a class activation map to further analyze the interpretability of the model. To fully explain the classification results, we further explored which region of the model is used to judge the category of the input, and the CAM method is used for convolution feature up-sampling processing and a gradient backpropagation. Guided-BP, original CAM, Grad-CAM based on gradient, and Guided-

Grad-CAM based on gradient guidance are used to visualize the features of MCNN. The results are shown in Table 5.

Table 5. Core module construction of MOBILENETS.

Methods	Processing Results			
Original image				
Guided-BP				
CAM				
Grad-CAM				
Guided-Grad-CAM				

The visualization graphs in Table 5 show all the features that can be extracted from the MCNN. The feature maps obtained by Guided-BP are noise-free, and it can be seen that the features extracted by the model are concentrated on the discharge data of all kinds of PRPD spectra. Through the reverse visualization calculation, the pixels used to judge the network are marked. The figures in Table 5 show that the eigenvalue regions selected by MCNN in the training process are consistent in the four visualization algorithms, and the eigenvalues are all in the discharge pulse aggregation region. Therefore, the training process and recognition results of MCNN are reliable.

The weighted sum of the feature maps contained in the last convolutional layer in the MCNN model and their corresponding weights are obtained as the result of CAM. The Grad-Cam thermal diagram is obtained by superimposing the original image with the up-sampling processing of these features. On this basis, the back-propagation gradient is used as the weight of the feature map; it can be concluded that the visualization results obtained by the Guided-Grad-CAM method are highly consistent with those obtained by the Guided-BP. Therefore, through the above four methods to visualize the recognition process of all types of PD data in the MCNN, it is very intuitive to explain the classification results of the model. The visualization graphs in Table 5 further show that the MCNN designed in this paper can extract the image features of different discharge types and has strong feasibility for the realization of PD pattern recognition.

4. Conclusions

In this paper, a method to identify the transformer PD patterns based on MobileNets convolution neural network (MCNN) is designed. Based on adopting a deep separable convolution module and inverse residual structure, the basic module of MCNN introduces lightweight attention mechanism SE module and h-swish function with stronger nonlinear ability, which can effectively avoid the problems such as the need for human interven-

tion and low recognition rate existing in the traditional shallow learning algorithm. In addition, the model solves the defects of the current depth model applied in PD pattern recognition, such as high complexity, a large number of parameters, large memory, and slow running speed.

The MCNN is trained and tested through different types of PRPD spectra obtained by experiments, and the classification performance of the model is compared with many existing deep convolution neural network models and compression network models. A variety of visualization methods are used to further understand the internal solution process of the MCNN model and explain the feasibility of the model in PD recognition. The results show that the MCNN can significantly reduce the number of parameters and the computational complexity of the model, and at the same time, further improve the accuracy of PD type recognition. Therefore, PD pattern recognition based on the model can be used in small mobile devices or integrated systems, and alleviate the pressure of space resources shortage on offshore platforms.

Author Contributions: Conceptualization, S.M. and S.S.; methodology, S.M. and S.S.; software, S.M. and S.S.; validation, Y.S., S.M. and S.S.; formal analysis, S.M.; investigation, Y.S., P.L., L.Z., J.O. and X.N.; resources, Y.S., L.Z., J.O. and X.N.; data curation, S.M. and S.S.; writing—original draft preparation, Y.S., S.M. and S.S.; writing—review and editing, Y.S., S.M. and S.S.; visualization, Y.S.; supervision, Y.S.; project administration, Y.S.; funding acquisition, L.Z., P.L., J.O. and X.N. All authors have read and agreed to the published version of the manuscript.

Funding: The National Key Research and Development Program of China under grant # 2018YFB0904800.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported by The National Key Research and Development Program of China under grant # 2018YFB0904800.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shang, H.K.; Yuan, J.S.; Wang, Y. Partial discharge pattern recognition in power transformer based on multi-kernel multi-class relevance vector machine. *Trans. China Electrotech. Soc.* **2014**, *29*, 221–228.
2. Zhong, L.P.; Ji, S.C.; Cui, Y.J. Partial discharge characteristics of typical defects in 110 kV transformer and live detection technology. *High Volt. Appar.* **2015**, *51*, 15–21.
3. Li, J.H.; Han, X.T.; Liu, Z.H. Review on partial discharge measurement technology of electrical equipment. *High Volt. Eng.* **2015**, *41*, 2583–2601.
4. Li, J.; Sun, C.X.; Liao, R.J. Study on statistical features used for PD image recognition. In Proceedings of the CSEE, Kunming, China, 13–17 October 2002; Volume 22, pp. 104–107.
5. Zheng, Z.; Tan, K.X. Partial discharge recognition based on pulse waveform using time domain data compression method. In Proceedings of the 6th International Conference on Properties and Applications of Dielectric Materials, Xi'an, China, 21–26 June 2000.
6. Gong, Y.P.; Liu, Y.W.; Wu, L.Y. Identification of partial discharge in gas insulated switchgears with fractal theory and support vector machine. *Power Syst. Technol.* **2011**, *35*, 135–139.
7. Ren, X.W.; Xue, L.; Song, Y. The pattern recognition of partial discharge based on fractal characteristics using LS-SVM. *Power Syst. Prot. Control.* **2011**, *39*, 143–147.
8. Ju, F.L.; Zhang, X.X. Multi-scale feature parameters extraction of GIS partial discharge signal with harmonic wavelet packet transform. *Trans. China Electrotech. Soc.* **2015**, *30*, 250–257.
9. Zhou, S.; Jing, L. Pattern recognition of partial discharge based on moment features and probabilistic neural network. *Power Syst. Prot. Control.* **2016**, *44*, 98–102.
10. Schaik, N.V.; Czaszejko, T. Conditions of discharge-free operation of XLPE insulated power cable systems. *IEEE Trans. Electr. Insul.* **2008**, *4*, 1120–1130. [CrossRef]
11. Yin, J.L.; Zhu, Y.L.; Yu, G.Q. Relevance vector ma-chine and its application in transformer fault diagnosis. *Electr. Power Autom. Equip.* **2012**, *32*, 130–134.
12. Zhao, L.; Zhu, Y.L.; Jia, Y.F. Feature extraction for partial discharge grayscale image based on Gray Level Co-occurrence Matrix and Local Binary Pattern. *Electr. Meas. Instrum.* **2017**, *54*, 77–82.

13. Liu, B.; Zheng, J. Partial discharge pattern recognition in power transformers based on convolutional neural networks. *High Volt. Appar.* **2017**, *53*, 70–74.
14. Yuan, F.; Wu, G.J. Partial discharge pattern recognition of high-voltage cables based on convolutional neural network. *Electr. Power Autom. Equip.* **2018**, *289*, 130–135.
15. Wang, X.Q.; Sun, H.; Li, L.G. Application of convolutional neural networks in pattern recognition of partial discharge image. *Power Syst. Technol.* **2019**, *47*, 130–135.
16. Hui, S.; Dai, J.; Sheng, G.; Jiang, X. GIS partial discharge pattern recognition via deep convolutional neural network under complex data source. *IEEE Trans. Dielectr. Electr. Insul.* **2018**, *25*, 678–685.
17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M. Hartwig AdamMobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
18. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2017**, arXiv:1709.01507.
19. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. *arXiv* **2019**, arXiv:1905.02244v5.