Method of inter-turn fault detection for next-generation smart transformers based on deep learning algorithm

Lian Duan¹, Jun Hu¹, Gen Zhao¹, Kunjin Chen¹, Shan X. Wang¹,², Jinliang He¹

¹Department of Electrical Engineering, Tsinghua University, Beijing, People’s Republic of China
²Center for Magnetic Nanotechnology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

E-mail: hjun@tsinghua.edu.cn

Abstract: In this study, an inter-turn fault diagnosis method is proposed based on deep learning algorithm. 12-channel data is obtained in MATLAB/Simulink as the time-domain monitoring signals and labelled with 16 different fault tags, including both primary and secondary voltage and current waveforms. An auto-encoder is presented to classify the fault type of the abundant transformers. The proposed method under deep learning framework increases the accuracy and robustness in transformer fault diagnosis, indicating its potential and prospect in the next-generation smart transformers.

1 Introduction

Power transformer is one of the most essential and expensive equipment in the power transmission and distributed system. As the power rate per transformer grows higher and higher, loss of power transformer not only causes an interruption in power supply but also leads to disruption on the stability of the whole system. Therefore, it is very important to detect incipient inter-turn fault [1]. In general, detection of inter-turn short circuit fault can be classified into two categories, online monitoring and offline test.

Compared to online method, though offline test of a transformer has to be performed while it is out of service, its accuracy and reliability is much higher than the online monitoring. Winding resistance test can evaluate the current winding status, which is able to find out the existence of inter-turn fault [2, 3]. Partial discharges test is also capable of evaluating insulation status. Partial discharge is the early stage of the insulation break down which means short circuit fault can be discovered on its occurrence using this method [4, 5]. After the internal short circuit happens, electromagnetic force would cause winding deformation and frequency response analysis (FRA) is very sensitive to this [6]. Accuracy and application of FRA method have been extensively discussed in [7–9].

Online monitoring and diagnosis is the future trend in this field, since fault can be discovered as early as possible and the detection process will not cause power failure. Differential protection and digital relay is the classical and robust method to solve this problem. Several modifications are proposed in [10, 11] based on this method to make this method adapt to different application scenarios.

As the nature of the inter-fault detection of the transformer is a classification mission, machine learning technology seem to be very suitable for this application. Many researchers have tried to find signals containing fault information and applied them to existing machine learning classifier. The authors in [12, 13] make use of the differential current and put it into the particle swarm optimisation based probabilistic neural network. Wavelets are used in [14, 15] to do feature extraction on the neutral current. A novel DGA technique based on Parzen window estimation has been presented in [16].

Despite of the advantages of the above-mentioned methods, they all contain limitations. All off-line methods can only be conducted during the maintenance. Differential protection and digital relay diagnostic system has a higher reliability and can be realised by online monitoring. However, setting of the threshold value between normal and fault status is troublesome. All the machine learning methods are based on a shallow model which means the model can only do the simple job of classification. The extracted feature is designed by human. This feature designing process is the most time-consuming task and requires many experts’ opinions.

In this paper, a novel fault diagnosis system is proposed based on the deep learning (DL) method auto-encoder (AE) designed by Hinton and Salakhutdinov [17]. With the powerful DL tool, features can be extracted automatically without much expertise. Voltage and current waveforms of both primary and secondary side are put together to form a multi-channel signal matrix for the proposed system that is able to diagnosis 15 types of inter-turn short circuit fault. Section 2 introduces the MATLAB/Simulink simulation system and how to pre-process the data to transform the original fault signal to the data block to be learnt. Section 3 presents the proposed classification framework and its basic principle. Section 4 shows the parameter tuning procedure to further promote the performance of the framework. Section 5 studies the influence of input data frequency and channels.

2 Simulation system and data acquisition

2.1 Simulation system layout

Several simulation methods of internal winding fault have been proposed, some of which are based on circuit theory and others are based on electromagnetic theory. Since only current and voltage signals are necessary in this paper, circuit model is enough and easy to carry out. In [18], a transformer model based on primary and secondary coil has been proposed, which is capable of simulating faults between any turn and the earth or between two
turns of the transformer windings. The simulation is implemented in MATLAB/Simulink.

A simple three-phase system is illustrated in Fig. 1. Considering the convenience of laboratory experiment in the future, these two transformers are both modelled based on real transformer structure. The first transformer is oil-immersed, which is utilised to boost the three-phase 380 V to 10 kV. The second transformer is the dry-type one to be deployed faults. The whole system is 50 Hz. The open circuit and short circuit test results of the two transformers are list in Table 1.

2.2 Fault deploy

In this paper, a total of 15 kinds of faults will be studied and they could be summarised in three categories, which are modelled in Fig. 2. In Simulink, primary and secondary winding modules can be used to build a simple transformer, as shown in Fig. 2a. Pink line denotes the magnetic circuit while the blue line stands for the electrical circuit. All the windings are in one close magnetic loop, which means the mutual inductance is considered in this model which accords to the theory induced in [18].

When a single winding to ground fault occurs, short-circuit point 1 split the winding into two parts, shown as Fig. 2b. In the same way, when there are two short-circuit points in the fault, such as double winding fault and single winding fault, shown in Figs. 2c and d, respectively, eight windings are used in the simulation. By tuning the winding turns, different position of short-circuit can be obtained.

In Fig. 2b, only high-voltage winding to ground fault is simulated because insulation between low-voltage winding to ground is not strong enough to breakdown the system. Thus, three kinds of faults, A-g, B-g and C-g are simulated by this model.

Considering double winding fault, inter-phase low-voltage winding will not be taken into consideration because the low-voltage winding is covered by the high-voltage one. For the same reason, interphase fault between different sides of the transformer, such as A-b, will not be simulated either. A-a, B-b, C-c, A-B, B-C and C-A, a total of six kinds of fault data are obtained in Fig. 2c.

There are six more faults circumstance simulated in Fig. 2d, such as A-A, B-B, C-C, a-a, b-b and c-c. By tuning the different parameters of the system, a total of 61,225 records of raw dataset are generated. Specific parameters such as fault position, load angle, load magnitude and fault occurrence angle are listed in Table 2. System frequency is 50 Hz.

2.3 Pre-processing

Each record of raw data contains a 12-channel waveform with 0.5 s, which can be seen as a 12 × I_{seg} matrix. U_A, U_B, U_C, I_A, I_B, I_C, U_{av}, U_{pk}, I_{av}, I_{pk} correspond to the 1st to 12th row of the data record. A time span of 0.5 s contains both pre-fault and post-fault information. Zero degree of fault occurrence angle is aligned to 0.38 s. Only post-fault data fragment is of interest, so each record is cut into a I_{seg} matrix with only post-fault data. The beginning time is randomly selected from 0.4 to 0.42 s. Since the number of A-g, B-g, C-g and no fault data are far less than that of others (i.e. 1200 and 25), the final number of data segment is multiplied to 4800. The dataset is finally comprised of 76,800 records and each records is a 12 × I_{seg} matrix. Suppose that the

| Fault type | Fault position 1, % | Fault position 2, % | Load angle, deg | Load magnitude, kW | Fault occurrence angle, deg | Total number | Fault code |
|------------|---------------------|---------------------|-----------------|------------------|-----------------------------|-------------|-----------|
| A-A, B-B, C-C | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 4800 × 6 | 1, 5, 9 |
| a-a, b-b, c-c | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 4800 × 6 | 2, 6, 10 |
| A-a, B-b, C-c | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 4800 × 6 | 3, 7, 11 |
| A-B, B-C, C-A | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 1200 × 3 | 4, 8, 12 |
| A-g, B-g, C-g | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 1200 × 3 | 13, 14, 15 |
| no fault | 10, 20, 50, 70 | 10, 20, 50, 70 0 | 10, 20, 30, 40 | 1, 2, 3, 4, 5 | 0, 30, 60, 90, 120, 150, 180, | 210, 240, 270, 300, 330 | 1200 × 3 | 25, 16 |

The data analysis is conducted on the MATLAB platform. A new training dataset matrix is obtained by combining the pre-processing dataset and 76,800 training samples. The dataset is split into 90% for training and testing, respectively.
Given a set of input data \( \{x^{(1)}, x^{(2)}, ..., x^{(N)}\}, x^{(i)} \in \mathbb{R}^n \), \( i = 1, 2, ..., N \), this input data will be mapped to the hidden layer \( h_0 \).

Therefore, only half of the AE, input layer and hidden layer, will be used during application and the features are extracted by the weight matrix of the hidden layer. In [19], a sparse AE (SAE) is used to extract features of the various types of ferroresonance overvoltage.

\[
h(x) = f(Wx + b) \tag{1}
\]

where \( W, b \in \mathbb{R}^{m \times n} \) is the weight matrix connecting the visible and hidden layer and \( b \in \mathbb{R}^m \) is the bias vector. \( N \) is the total number of training vector \( x^{(i)} \); \( n \) is the dimension of each input vector, \( m \) is the size of hidden nodes. In (1), function \( f() \) is called the activation function. Typically, it is sigmoid function which is defined as

\[
f(a) = \frac{1}{1 + \exp(-a)} \tag{2}
\]

There are also other activation functions, such as tanh, ReLu, ELU and PreLU [20]. Evaluation of the effect of these activation function will be discussed in Section 4. After encoding, the second step, decoding, is completed by the reconstructing layer using

\[
\hat{x} = f(Wf(x) + b) \tag{3}
\]

where \( W, b \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^m \) are also called the weight matrix and bias vector associating output layer with hidden layer. The accuracy of a simple AE is assessed by the cost function defined as follows:

\[
J_{AE} = \frac{1}{2N} \sum_{i=1}^{N} \left\| \hat{x}^{(i)} - x^{(i)} \right\|^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \|W_j\|^2 \tag{4}
\]

In a simple way, training a SAE means the optimisation of \( J_{AE} \). When \( J_{AE} \) is optimised to an expected small value or a certain number of iterations is reached, the training process will be completed and \( W_1 \) is the feature learnt from the dataset \( x \) automatically.

Several improved AE have been proposed in recent years such as SAE and de-noise AE (DAE) [19]. The difference between the original AE and improved AE is adding an additional penalty term in (4) to adapt AE to different applications.

In the case in this paper, the process of AE application of a multi-channel-signal learning is shown in Fig. 4. In Section 2.3, all the dataset are transformed to a \( 12 \times l_{seg} \times 76,800 \) data cube, which means there are 76,800 records in the dataset and each of them is a \( 12 \times l_{seg} \) matrix. AE is trained by \( n_{patches} \) patches cut from this matrix. A \( 12 \times l_p \times n_{patches} \) patch cube is randomly cut out from data cube, where \( l_p \) is defined as the patches length in the training process. Since each layer of AE is a one-dimension vector, each patch matrix is stack to a vector, the length of which is \( 12l_p \). Thus, input training patches is a \( 12l_p \) by \( n_{patches} \) matrix. Supposing that there are \( m_{hidden} \) hidden nodes in the hidden layer, transforming matrix \( W_1 \), the extracted feature, is a \( m_{hidden} \) by \( 12l_p \) matrix. The feature \( W_1 \) matrix can be reshaped to \( m_{hidden} \times 12l_p \) small feature matrices. Moreover, these features can be visualised using greyscale images by mapping the matrix value to an interval of 0 and 255. By setting \( l_{seg}, l_p \) and \( m_{hidden} \) to 401, 30 and 40, a preliminary AE could be trained, the optimisation method of which is limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [21]. Under most conditions, L-BFGS has the advantages of higher convergence speed and lower cost of computing resources.

The visualised 40 features are shown in Fig. 5. The darker the pixel is, the greater the value of the corresponding position in the matrix will be. Most of the features are in a moderate value, yet there are some exceptions. For example, features 9, 10, 17, 18 and 36 have more bright and dark pixels, showing a rapid fluctuation of the waveform, which will be activated when there is a sudden change in the input data.

3.2 Convolutional and pooling layer for classification

After training AE, the following convolutional and pooling layers help to transform the original data into a one-dimension vector used to train the softmax classifier. The whole process is illustrated in Fig. 6. This multi-channel convolution neural network is used by Zhang et al. [22] and Chen et al. [23] to solve multi-channel signal problem, which is called as convolutional AE (CAE) method or convolutional sparse AE (CSAE) method.
Each original segmented record is a $12 \times l_{\text{seg}}$ matrix and each extracted feature (shown in Fig. 6 as a greyscale picture in a frame with different colours) is a $12 \times l_p$ matrix. At the beginning, the feature is aligned to the left margin of the segmented record and a dot product is calculated with all the elements overlapped, which will be the first element of the horizontal vector. Then, as the feature moving to the right margin of the segmented record, dot products will be obtained one by one that will complete this horizontal vector. The size of this convoluted vector is $l_{\text{seg}} - l_p + 1$. After repeating this process using the rest of the features, a $n_{\text{hidden}}$ by $l_{\text{seg}} - l_p + 1$ matrix will be obtained.

Dimension of the original segmented data is transformed from $12 \times l_{\text{seg}} \times 76,800$ to $n_{\text{hidden}} \times (l_{\text{seg}} - l_p + 1) \times 76,800$ through convolutional layer. Generally, if the convoluted data is directly sent into the softmax classifier, the input data size would be so large that the model is easy to overfit. Thus, a pooling layer is used to prune the data followed by the convolutional layer. The commonly used method is max-pooling and mean-pooling. The basic theory is the same. In this paper, pooling is operated on each horizontal convoluted vector, so the pooling size is $1 \times s_p$. $s_p$ is defined as the pooling size. The vector will be divided into several disjoint segments and the length of each segment is $s_p$. It should be noted that it is not necessary that $(l_{\text{seg}} - l_p + 1)$ have to be divisible by $s_p$ because the remainder element can be discarded. For max-pooling, the max value of the $s_p$-length divided segment is calculated as the new element. For mean-pooling, the mean value of the $s_p$-length segmented vector is calculated as the new element. The final size of the pooled data is $n_{\text{hidden}} \times ((l_{\text{seg}} - l_p + 1)/s_p) \times 76,800$. In this paper, $\{ \}$ denotes the rounding operator towards minus infinity. Fig. 7 presents the data flow chart and their corresponding dimension.

The last layer of the framework is the softmax classifier. It is based on softmax regression model, which is a generalisation of logistic regression model and could deal with multi-class problems [24]. When softmax is applied to a $q$-class classification problem, the output vector of the classifier is made up with all $q$ probabilities. Thus, the input vector will be divided into the category with the high probabilities. Given a group of labelled input vector $\{x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}\}$, $x^{(\ell)} \in R^q$ and $y^{(\ell)}$ being the corresponding label, the probability that $x^{(\ell)}$ belongs to a certain class $j, j = 1, 2, \ldots, q$ can be calculated as follows [20]:

$$P(y^{(\ell)} = j|x^{(\ell)}) = \frac{e^{\theta^T_j x^{(\ell)}}}{\sum_{q=1}^{q} e^{\theta^T_q x^{(\ell)}}} \quad (5)$$

Since the input data of softmax has to be a one-dimensional vector, pooled $n_{\text{hidden}}$ by $\{(l_{\text{seg}} - l_p + 1)/s_p\}$ matrix has to be stack sequentially to a vector. 76,800 records are divided into two parts, training data and validating data, with the ratio of 0.7 and 0.3.

4 Hyper-parameter selection of the proposed method

Although the basic framework has been proposed in Section 3, some details should be determined step by step. AE type and its activation function, sparsity, depth, hidden nodes number and pooling strategy will be discussed one by one.

4.1 Choosing the type of AE and its activation function

SAE and DAE are two common varieties of the original AE. In SAE, sparsity parameter $\rho$ is defined to control the average activation level of all hidden nodes. For example, if $\rho$ is set to 0.1, only 10% of the hidden node is activated. DAE is generally used to make the model more robust to the noise. In this paper, all the data is based on Simulink simulation, noise is not taken into consideration. Thus, only SAE and original AE will be adopted.

The activation function enables a DL model to approximate to a non-linear model instead of a simple linear one, which makes the model more adaptive. In the former training process, sigmoid function is used as the default activation function, which is one of the most widely used activation functions at present. Several concerns are aroused by the application of sigmoid and the most important one is gradient vanishing. Relu is another kind of function that is widely used in DL realm [25]. It gives an output $x$ if $x$ is positive and 0 otherwise and this output characteristic will bring two advantages. For the input $x > 0$, the derivative of the function is a constant which will not cause vanishing gradient during the backpropagation and computation load is much less than sigmoid. For input $x < 0$, the output of the function is set to 0 and the sparsity of the network is therefore achieved directly.

Three types of AE with different kinds of activation function are trained in this section, such as AE with sigmoid, AE with Relu, SAE with sigmoid. Number of hidden nodes is set to 40, and sparsity $\rho$ is 0.1. Classification accuracies of different types of AE are listed in Table 3. Training process of each type of AE is shown in Fig. 8.

By checking the confusion matrix, it can be discovered that this kind of data is easy to be misclassified to fault 2, 6 and 10, which is the three kinds of low-voltage single winding fault, a-a, b-b and c-c. The learning process shown in Fig. 8 proves that AE with Relu is more efficient than sigmoid. However, under this size and dimension of the training data, the difference among them is not

---

**Fig. 6 Framework for fault detection and classification using CAE**

**Fig. 7 Data flow chart and their corresponding dimension**
that obvious. Thus, SAE is chosen as the basic feature extraction method in the following sections.

### 4.2 Sparsity parameters

Sparsity $\rho$ is the key value that will decide the classification performance of SAE. In this paper, $\rho$ is set to 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7. Typically, $\rho$ tends to be a small value \[19\] according to the fact of the bionics that only a little of neurons is activated when human brain receives an outside stimuli. So, the step of $\rho$ is below 0.1. The classification result is illustrated in Fig. 9.

When the sparsity is too small, for example 0.02 and 0.04, the accuracy of the classifier decreases to 96–98%. As $\rho$ increases to 0.2, the highest validating accuracy is 99.43%. However, further increasing $\rho$ from 0.4 to 0.7 will lead to a slight decline in the accuracy. Thus, sparsity is set to 0.2 in later training process.

### 4.3 Depth of AE

In \[26\], Hinon mentions that multiple-layer AE might make a feature extraction more effective, and it should be trained through top-down structures. Supposing a 360-200-40 SAE is going to be trained, the first 360-200 AE should be trained and the 200-40 SAE will be trained right after it. According to (2)

$$h_1(x) = W_1x + b_1$$  \hspace{1cm} (6)

$$h_2(x) = W_2h_1(x) + b_2 = W_2W_1x + (W_2b_1 + b_2)$$  \hspace{1cm} (7)

Let us define $W$ and $b$ as the overall weight matrix and bias vector and it can be easily derived that

$$W = W_2W_1, \hspace{0.5cm} b = W_2b_1 + b_2$$  \hspace{1cm} (8)

It is obvious that $W$ and $b$ are still a 40 by 360 matrix and 40 by 1 vector, the dimension of which is the same as that of the 360-40 SAE. In order to evaluate the influence of depth of AE, two multiple-layer SAEs, also called as the stacked SAE (SSAE), with different number of hidden nodes are implemented. The result of them is compared with that of single-layer SAE. All of the sparsity is set to 0.2. Under the circumstances in this paper, a final structure of 360-40 double layers will be enough.

Table 4 lists the comparison of different depths of SAE. The optimisation process is shown in Fig. 10. Solid and dashed lines in the same colour indicate the two steps of training a double-layer SAE. This double-layer structure of SAE improves the overall accuracy a little, from 99.4 to 99.5% at the cost of training complexity. It costs nearly as much as three times the time to train a double-layer SAE, since that the dimension of the training data matrix of first layer is much larger than that of the second layer. In Fig. 10, it can be concluded that the optimisation efficiency of the single-layer SAE is better than that of the secondary layer and worse than that of the first layer. Thus, considering the trade-off of training time and improved performance, single layer SAE is a better choice.

### 4.4 Number of hidden nodes

Number of hidden nodes affects the number of the extracted feature of SAE. More hidden nodes mean more information in the feature. It is foreseeable that more hidden nodes lead to a better classification accuracy because of sufficient information. Number of hidden nodes is set to 10 to 200 and the accuracies are list in Fig. 11.

When there are only ten hidden nodes in the SAE, accuracy of this diagnosis system drops to lower than 88%, because the

| Fault type | AE with sigmoid, % | AE with Relu, % | SAE with sigmoid, % |
|------------|-------------------|----------------|------------------|
| Train      | Validate          | Train          | Validate         |
| 1          | 100               | 100            | 100              |
| 2          | 100               | 100            | 100              |
| 3          | 100               | 100            | 100              |
| 4          | 98.24             | 98.19          | 99.02            | 99.24            | 98.27            | 97.57            |
| 5          | 99.97             | 100            | 100              | 99.93            | 100              | 100              |
| 6          | 100               | 100            | 100              | 100              | 100              | 100              |
| 7          | 100               | 100            | 100              | 100              | 100              | 100              |
| 8          | 94.52             | 93.89          | 100              | 100              | 100              | 100              |
| 9          | 100               | 100            | 100              | 100              | 100              | 100              |
| 10         | 100               | 100            | 100              | 100              | 100              | 100              |
| 11         | 100               | 100            | 100              | 100              | 100              | 100              |
| 12         | 100               | 100            | 100              | 100              | 100              | 100              |
| 13         | 100               | 100            | 100              | 100              | 100              | 100              |
| 14         | 100               | 100            | 100              | 100              | 100              | 100              |
| 15         | 100               | 100            | 100              | 100              | 100              | 100              |
| 16         | 81.10             | 81.11          | 84.4             | 86.81            | 87.68            | 88.47            |
| overall    | 98.36             | 98.33          | 98.97            | 99.12            | 99.12            | 99.13            |

Fig. 8 Cost value (normalised) optimisation process of the three types of AE

Fig. 9 Influence of sparsity parameters on the classifying accuracy
information extracted by only ten features is not enough for this circumstance. As the number goes up, the classifier becomes more and more accurate. There is a saturation effect of the accuracy as the increase of feature number for the reason that information is redundant. It should be noted that convolution calculation is a heavy burden for computer. The cost of time is linearly increasing with the number of features. So, it is necessary to reach a higher performance with fewer features. From Fig. 11, this number is 60 and the corresponding accuracy of validate data is 99.53%.

4.5 Pooling strategy

Max pooling and mean pooling strategies are adopted in this paper. Considering the length of the convoluted data is 172, pooling length is set to 5, 10, 15, 20 and 25, respectively. Both training and validating accuracy are shown in Fig. 12.

When different pooling functions are used, such as max and mean function in this case, pooling size has a different impact on the accuracy. In general, classification accuracy decreases when the pooling size becomes larger, which can be attributed to that a larger pooling size will omit more information of the original data. Under this condition, mean pooling loses more information than the max pooling. However, mean pooling size has a better performance at low pooling size. The difference between training and validating accuracy is the lowest when pooling size is around 10 and 15. In this paper, mean pooling with a size of 5 is adopted because the model trained has the highest accuracy and the difference between training and validating is acceptable, which is 0.07%.

After the parameter selection in Section 4, the basic structure of the model is a 360-40 SAE with sigmoid function, the sparsity of which is 0.2. The pooling strategy is mean pooling with the size of 5.

5 Influence of signal characteristics

5.1 Influence of time window

In the discussion above, time window of the data is 20 ms, which corresponds to a whole cycle of the waveform and contains all the fault waveform information. It would be better if the model can be perfectly trained with smaller amount of data, making the computation load relieved. Thus, redundancy needs to be studied by narrowing the data window. The sampling rate is still fixed to 20 kS/s. The data window ratio is set between 10 and 90% with a step of 10%. When the ratio is set to 10%, the data length is 41, with a time window of 2 ms.

From the information shown in Fig. 13, it can be concluded that as data time span decreases from one cycle (20 ms) to half a cycle (10 ms), the training accuracy drops a little. When further decreasing, the accuracy drops drastically. Although overall accuracy seems to be acceptable, which is over 92%, some of the faults fail to be detected. For example, when the time span is 2 ms, 384 out of 1440 record of fault 4 is diagnosed as no fault and 357 out of it is diagnosed as fault 3. Thus, the accuracy of fault 4 is only 48.54%. Same situation happens to fault 8 and no fault data and the accuracy of them are only 65.54 and 53.61%, respectively.

### Table 4 Classification accuracies of different depths of SAE

| Fault type | SAE 360-40, % | SAE 360-200-40, % | SAE 360-100-40, % |
|------------|--------------|-----------------|------------------|
|            | Train Validate | Train Validate | Train Validate   |
| 1          | 100          | 100             | 100              |
| 2          | 100          | 100             | 100              |
| 3          | 100          | 100             | 100              |
| 4          | 100          | 100             | 100              |
| 5          | 100          | 99.97           | 100              |
| 6          | 100          | 100             | 100              |
| 7          | 100          | 100             | 100              |
| 8          | 100          | 100             | 100              |
| 9          | 100          | 100             | 100              |
| 10         | 100          | 100             | 100              |
| 11         | 100          | 100             | 100              |
| 12         | 100          | 100             | 100              |
| 13         | 100          | 100             | 100              |
| 14         | 100          | 100             | 100              |
| 15         | 100          | 100             | 100              |
| 16         | 90.88        | 91.21           | 92.29            |
| overall    | 99.45        | 99.43           | 99.52            |

Fig. 10 Cost value (normalised) optimisation process of the SAE and two SSAEs

Fig. 11 Influence of hidden node on the classifying accuracy

Fig. 12 Influence of pooling strategy on the classifying accuracy
Regarding the time of training time span 20 ms as base value, time ratio of others to it is also shown in Fig. 13. Considering the trade-off between training time and classification accuracy, 12 ms seems to be a suitable length of the data with half of the training time and 0.37% drops of the accuracy.

5.2 Verification dataset with different frequencies

In the cases studied above, the sampling rate of data matrix is set to 20 kS/s. If this sampling rate changes, the feature trained by AE will be altered. For example, if the time interval of patches is 1.5 ms used above, the length of it is 30 at 20 kS/s. When the sampling rate decreases, that of the patches will be dropped simultaneously. Once the feature is trained by 20 kS/s data, it is unknown that whether it is necessary to re-train the model if the input data is sampled by another frequency. In Fig. 14, two solutions are compared with each other, interpolating the original AE features or training a new AE.

The original AE is trained by 20 kS/s data. When the frequency does not drop too much, such as 15 and 17.5 kS/s in this case, the classification accuracy of interpolated AE and newly trained AE is about the same. When the frequency further decreases, such as 5, 8 and 10 kS/s, the difference between the accuracies varies a lot. Fig. 14 also shows that when the frequency decreases to 8 kS/s, the overall accuracy of the newly trained data is 98.48%, which means that in the proposed fault diagnosis system, sampling frequency of the input signal must be higher than 8 kS/s.

5.3 Influence of signal channel selection

During the training and validating process above, three-phase voltage and current waveform of both sides are used. However, it is necessary to study the scenario that only part of the channels is selected. Less channel number means less usage of the storage and a faster computation speed.

5.3.1 Selection of types and sides: In Section 4, after the final parameter tuning, the classification accuracy of the validating dataset is 99.5%. Compared to the result in Table 5, it can be found that removal of any channel from the original 12-channel signal will lead to a decrease of the accuracy.

More specifically, deleting primary voltages has the least influence, 99.34%. If the faulted transformer is connected to an infinite power system, the voltage barely changes when the fault occurs. Deleting the secondary voltage has the same impact as deleting the secondary current, which can be attributed to that in this case current is highly related to voltage because the secondary side of the transformer is connected to a linear load. Further removing the input signal to six channels, the fault identification accuracy is still 100%, but the non-fault data is more easily misclassified to other kinds of fault data. When only three channels of signals are selected, the accuracy is too low to be accepted.

5.3.2 Selection of phases: Based on the nine-channel signal, primary current, secondary voltage and current, shown in the second column in Table 5, signal of some phases is removed in this section. The accuracies are listed in Table 6.

When two phases are selected, the overall accuracy drop from 99.5% to around 99% and performance of single phase to ground identification drops from 100 to 99%. However, when single phase signal is chosen, the results are unacceptable.

6 Experiment on a real 10 kV transformer

6.1 Platform introduction

The experiment platform shown in Fig. 15 is a realisation of simulated system in Fig. 1, except that there is no load considered. The one on the right is a 10 kV 10 kVA oil-immersed transformer stepping voltage from 380 V to 10 kV. The other transformer on

Table 5 Classification accuracies of different selections of channels (types and sides) (%)

| Fault type | 9-C (1I & 2UI) | 9-C (1UI & 2I) | 9-C (1UI & 2UI) | 6-C (1I & 2UI) | 3-C (1I) | 3-C (2I) |
|------------|---------------|---------------|----------------|---------------|----------|----------|
| 1          | 100           | 100           | 100            | 100           | 100      | 100      |
| 2          | 100           | 100           | 100            | 100           | 100      | 100      |
| 3          | 100           | 100           | 100            | 100           | 93.6     | 100      |
| 4          | 100           | 100           | 100            | 100           | 100      | 93.0     |
| 5          | 100           | 100           | 99.9           | 100           | 99.9     | 100      |
| 6          | 100           | 100           | 100            | 100           | 98.9     | 100      |
| 7          | 100           | 100           | 100            | 100           | 100      | 100      |
| 8          | 100           | 100           | 100            | 100           | 100      | 52.7     |
| 9          | 100           | 100           | 100            | 100           | 100      | 100      |
| 10         | 100           | 100           | 100            | 100           | 100      | 95.3     |
| 11         | 100           | 100           | 100            | 100           | 100      | 99.5     |
| 12         | 100           | 100           | 100            | 100           | 100      | 57.0     |
| 13         | 100           | 100           | 100            | 100           | 100      | 95.9     |
| 14         | 100           | 100           | 100            | 100           | 100      | 96.6     |
| 15         | 100           | 100           | 100            | 100           | 100      | 88.5     |
| 16         | 89.4          | 88.5          | 88.0           | 82.1          | 77.0     | 75.5     |
| overall    | 99.34         | 99.28         | 99.25          | 98.88         | 97.2     | 89.98    |

9-C (1I & 2UI) means 9 channels are selected, which is 3-phase primary current and secondary voltage and current.
the left is a resin-insulated one that is easy to be set to different malfunctioning conditions by shorting high-voltage winding taps. Fig. 16 demonstrates the inner structure of a high-voltage winding of single phase.

From Table 7 and Fig. 16, it can be seen that taps 1–6 divide the whole winding into six parts with different number of turns. Connecting taps 1 and 2 will make use of the whole winding and the voltage of the primary side is 10.5 kV. Similarly, if taps 5 and 6 are connected, the voltage will be 9500 V. Define A1 as tap 1 of phase A for convenience.

3 Pintech HVP-15HF voltage probes are used to measure the primary voltages (channels 1–3). The ratio of the probes is 1000:1 and the maximum voltage is 20 kVpp. Waveforms are recorded by Lecroy 8254. 3 Pintech 6010A voltage probes are applied on the secondary side, the ratio of which is 100:1 and the signal is input into Lecroy 3054. Two oscilloscopes are both triggered by the same signal, voltage of phase A. Three current sensors made by Tsinghua University are chosen to monitor the current. Highest frequency of the custom-made sensor is 1 MHz which is capable of sampling the signal at the frequency of 20 kHz. The data is transmitted to a pad via Bluetooth. Time axis of the pad and the oscilloscopes are aligned together to ensure the accuracy of phase information.

6.2 Fault detection of different kinds of short-circuit cases

When normally operating, taps 1 and 2 are connected and the voltage is 10.5 kV. By manually short other taps, nine kinds of fault related to high-voltage windings can be easily deployed. For example, when setting A-A fault, taps 4 and 6 will be shorted. Turns between taps 4 and 6 are 82 and considering safety, it is the minimum turns that can be chosen. Under this circumstance, 2.4% of the inter-turn fault is deployed, noted as A4-A6 in Table 8.

When fault is deployed, there will be a serious loop current in the shorted windings that would cause a permanent damage in the winding. After several times of experiments, the time duration of short circuit must be <1 s. Randomly selected 200 0.02 s-length data records for one kind of fault, which are re-sampled as 20 kHz. In this way, the input data is transformed to the same structure as the nine-channel data introduced in Table 5, 1VI & 2V (voltages

| Fault type | 3-C (a) | 3-C (b) | 3-C (c) | 6-C (ab) | 6-C (bc) | 6-C (ac) |
|------------|---------|---------|---------|----------|----------|----------|
| 1          | 100     | 100     | 97.7    | 100      | 100      | 100      |
| 2          | 97      | 100     | 98.4    | 100      | 99.8     | 100      |
| 3          | 77.3    | 99.6    | 89.6    | 100      | 99.2     | 99.8     |
| 4          | 94.7    | 93.1    | 86.9    | 100      | 99.2     | 99.8     |
| 5          | 96.1    | 100     | 100     | 100      | 100      | 100      |
| 6          | 100     | 100     | 100     | 100      | 100      | 100      |
| 7          | 99.9    | 100     | 100     | 100      | 100      | 100      |
| 8          | 73.9    | 86      | 87.9    | 100      | 99.8     | 100      |
| 9          | 100     | 94.7    | 100     | 100      | 100      | 100      |
| 10         | 100     | 100     | 98.5    | 100      | 98.9     | 100      |
| 11         | 93.4    | 93.6    | 96.5    | 100      | 100      | 100      |
| 12         | 84.1    | 64      | 72.3    | 98.0     | 99.2     | 100      |
| 13         | 99.2    | 99.7    | 99.2    | 100      | 100      | 100      |
| 14         | 96.7    | 97.5    | 79.7    | 100      | 100      | 100      |
| 15         | 88.4    | 93.9    | 92.2    | 100      | 100      | 100      |
| 16         | 46.3    | 43.8    | 42.3    | 85.2     | 85       | 84.7     |
| overall    | 90.47   | 91.66   | 90.93   | 99.01    | 98.95    | 98.96    |

Selection of channel is based on the 9-channel signal of primary current and secondary voltage and current (1I & 2VI). 6-C (ab) means 6 channels are selected, which are phase a and phase b.
and currents of primary side and secondary voltage). In this way, testing dataset is comprised by 2000 records, which is summarised in Table 8. It should be noticed that the training and validating data is the same as those used in the fourth column in Table 5 and these 2000 records are merely used as practical dataset to examine the robustness of the model trained by simulated dataset.

Since the confusion matrix is too big and most of the elements are zero, it is simplified in the last column in Table 8 as misclassified information. For example, misclassified information of A4-A6 is 1(4), 1(8), and 3(16), which means that 1 data record is misdiagnosed to fault 4, 1 data record is misdiagnosed to fault 8 and 3 data records are misdiagnosed to fault 16. From Table 8 it can be concluded that all accuracies of the fault testing data are higher than 96% and some of them are even close to 100%. The overall testing accuracy is 97.25%, which is 2% lower than the validating accuracy. For the reason that the signal-to-noise ratio of the testing data is about 25 dB and three-phase voltage is a little unbalanced, this 2% reduction in accuracy is acceptable.

7 Conclusion and future work

This paper proposed an inter-turn fault detection method for transformers based on DL method that could be used for next-generation smart transformer. A 12-channel signal is combined sequentially from voltage and current waveforms of the primary and secondary side. Based on the simulation method proved by previous studies, over 76,000 data records are obtained which are used as the training and validating dataset. Sixty features combined in a matrix are trained by a SAE with a sparsity of 0.2. A CSAE method is applied by convoluting the 0.2 ms data segment with the trained features. With the help of pooling and softmax layer, different kinds of faults can be classified. The final training recognition accuracy is 99.5%.

The influence of signal characteristic in this model is studied. The minimum time window of the data segment is 12 ms and the lowest frequency should be 8 kS/s. Removal of certain channels will lead to a decrease of classification accuracy. Signal made up by 9 out of 12 channels including primary current and secondary voltage and current has an accuracy of 99.34%. Nine kinds of faults are deployed on an operating resin-insulated transformer one by one and 2000 data records are collected as the testing dataset. Due to the influence of slight operating noise and unbalanced voltage, testing accuracy is 2% lower than the overall accuracy of training and testing data.

Simulated dataset and preliminary experimental data used in this paper prove the feasibility of this classification algorithm, future implementation of the method may use the practical data collected by a faulted transformer operation in the power grid, under a more complex operating situation.

8 Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant 51720105004, and by the Research Project of State Grid Corporation of China under grant 5202011600U.

9 References

[1] Sun, P., Sima, W., Jiang, X., et al.: ‘Review of accumulative failure of winding insulation subjected to repetitive impulsive voltages’, High Volt., 2019, 4(1), pp. 1–11.

[2] Okabe, S., Ueta, G., Nojima, K.: ‘Resistance characteristics and electrification characteristics of GIS epoxy insulators under DC voltage’, IEEE Trans. Dielectr. Electr. Insul., 2014, 21(3), pp. 1260–1267.

[3] Zhirui, L., Xufeng, Z., Shengguo, N.: ‘DC resistance measurement for low voltage side winding of large-scale transformer’, Elect. Power Autom. Equip., 2007, 27, pp. 27–30.

[4] Janani, H., Kordi, B.: ‘Towards automated statistical partial discharge source classification using pattern recognition techniques’, High Volt., 2018, 8(3), pp. 162–169.

[5] Qingguan, L., Simeng, L., Wen, S., et al.: ‘Analysis of the key problem about insulation condition assessment of oil-paper in power transformers based on partial discharge’, High Volt. Eng., 2017, 43(8), pp. 2558–2565, (in Chinese).

[6] Bagheri, M., Salay Naderi, M., Blackburn, T., et al.: ‘Frequency response analysis and short-circuit impedance measurement in detection of winding deformation within power transformers’, IEEE Trans. Electr. Insul. Mag., 2013, 29(3), pp. 33–40.

[7] Mehdi, B., Maxim, L., Salay, N.M., et al.: ‘Transformer frequency response: a new technique to analyze and distinguish the low-frequency band in the frequency response analysis spectrum’, IEEE Electr. Insul. Mag., 2018, 34(5), pp. 39–49.

[8] Zhang, H., Wang, S., Yuan, D., et al.: ‘Double-ladder circuit model of transformer winding for frequency response analysis considering frequency-dependent losses’, IEEE Trans. Magn., 2015, 51(11), pp. 1–4, Article Number: 8402034.

[9] Pereira Portilla, W., Aponte Mayor, G., Pleite Guerra, J., et al.: ‘Detection of transformer faults using frequency-response traces in the low-frequency bandwidth’, IEEE Trans. Ind. Electron., 2016, 64(1), pp. 4971–4978.

[10] Silva, K., Bainy, R.: ‘Generalized alpha plane for numerical differential protection applications’, IEEE Trans. Power Deliv., 2016, 31(6), pp. 2565–2566.

[11] Daubni, H., Sanaye-Pasand, M.: ‘Power transformer protection using a multi-region adaptive differential relay’, IEEE Trans. Power Deliv., 2014, 29(2), pp. 777–785.

[12] Tripathy, M., Maheshwari, R.P., Verma, H.K.: ‘Power transformer differential protection based on optimal probabilistic neural network’, IEEE Trans. Power Deliv., 2010, 25(1), pp. 102–112.

[13] Tripathy, M., Maheshwari, R.P., Verma, H.K.: ‘Probabilistic neural network based protection of power transformer’, IET Electr. Power Appl., 2017, 1(5), pp. 793–798.

[14] Rao, M.R., Singh, B.P.: ‘Detection and localization of interturn fault in the HV winding of a power transformer using wavelets’, IEEE Trans. Dielectr. Electr. Insul., 2001, 8(4), pp. 652–657.

[15] Saleh, S.A., Aktabi, A., Ahshan, R., et al.: ‘The development of a d-q-axis WPT based digital protection for power transformers’, IEEE Trans. Power Deliv., 2012, 27(4), pp. 2255–2269.

Table 7 Classification accuracies of testing dataset (types and sides) (%)

| Voltage          | 1-2 | 2-3 | 3-4 | 4-5 | 5-6 |
|------------------|-----|-----|-----|-----|-----|
| 10,500 V         | 10,250 V | 10,000 V | 9750 V | 9500 V |

Table 8 Classification accuracies of testing dataset (types and sides) (%)

| Fault deploy | Fault type | Number of waveforms | Accuracy, % | Misclassified information |
|--------------|------------|---------------------|-------------|--------------------------|
| A4-A6        | 1          | 200                 | 97.5 (1(4), 1(8), 3(16)) |
| A4-g         | 4          | 200                 | 98.5 (3(16)) |
| B4-B6        | 5          | 200                 | 100 (—)     |
| B4-g         | 8          | 200                 | 99.5 (1(16)) |
| C4-C6        | 9          | 200                 | 99 (2(3))   |
| C4-g         | 12         | 200                 | 99 (2(10))  |
| A4-B4        | 13         | 200                 | 96.5 (2(4), 2(8), 3(12)) |
| B4-C4        | 14         | 200                 | 100 (—)     |
| A4-C4        | 15         | 200                 | 98 (1(3), 1(5), 1(7), 1(16)) |

1(4) in the last column means that 1 record has been misclassified to fault type 2.
Islam, M.M., Lee, G., Hettiwatte, S.N.: ‘Application of Parzen Window estimation for incipient fault diagnosis in power transformers’, High Volt., 2018, 3 (4), pp. 303–309

Hinton, G.E., Salakhutdinov, R.R.: ‘Reducing the dimensionality of data with neural networks’, Science, 2006, 313, (5786), pp. 504–507

Bastard, P., Bertrand, P., Meunier, M.: ‘A transformer model for winding fault studies’, IEEE Trans. Power Deliv., 1994, 9, (2), pp. 690–699

Ng, A.: ‘Sparse autoencoder’, CS294A Lecture Notes, 2011, vol. 72, pp. 1–19

Duan, L., Hu, J., Zhao, G., et al.: ‘Identification of partial discharge defects based on deep learning method’, IEEE Trans. Power Deliv., 2019, 34, (4), pp. 1557–1568, early access

Ngiam, J., Coates, A., Lahiri, A., et al.: ‘On optimization methods for deep learning’ Proc. 2011 Int. Conf. on Machine Learning (ICML), Algiers, Algeria, 2011, pp. 265–272

Zhang, R., Li, C., Jia, D.: ‘A new multi-channels sequence recognition framework using deep convolutional neural network’, Procedia Comput. Sci., 2015, 53, pp. 383–390

Chen, K., Hu, J., He, J.: ‘Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder’, IEEE Trans. Smart Grid, 2018, 9, (3), pp. 1748–1758

Do, C., Ng, A.Y.: ‘Transfer learning for text classification’, in Schölkopf, B., Platt, J., Hofmann, T. (Eds.): ‘Advances in neural information processing system’ (MIT Press, London, England, 2006), pp. 299–306

Wang, G.W., Giannakis, G., Chen, J.: ‘Learning ReLU networks on linearly separable data: algorithm, optimality, and generalization’, IEEE Trans. Signal Process., 2019, 67, (9), pp. 2357–2370

Hinton, G.E.: ‘Learning multiple layers of representation’, Trends Cogn. Sci., 2007, 11, (10), pp. 428–434