J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 7627--7644

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Random forest algorithm and multi-source data fusion based on the converter valve electrical characteristics of time-varying law extraction method and condition monitoring technology research

Zhen Wang[1,✉], Linhao Xu[1], Ao Feng[2]

[1] *China Southern Power Grid Co., Ltd. Ultra High Voltage Transmission Company Electric Power Research Institute, Guangzhou, Guangdong, 510555, China*
[2] *Wuhan University, Wuhan, Hubei, 430072, China*

ABSTRACT
As the core equipment of high-voltage direct current transmission system, the operation status of the converter valve directly affects the safety of the power grid. In this paper, we first construct a multi-source data fusion system to realize the error-free fusion of fault information parameters. Then, combined with the random forest algorithm, the time-varying law of the electrical characteristics of the converter valve based on harmonic theory is extracted. Finally, the collected time-varying laws of electrical characteristics are input into the constructed Random Forest particle swarm optimization model, and the trained model is used to monitor the status of the converter valve. In the simulation experiment, the ±800kV UHV DC transmission system is built by PSCAD/EMTDC software, from which the current waveforms are collected when the converter valve fails, the time domain features of the current are extracted, and the obtained converter feature indicators are selected using the Random Forest algorithm, and 10 important features will be finally identified to construct the converter valve feature indicator set, and input into the Random Forest Particle Swarm Optimization model and the other comparative models for training and testing. The accuracy of this model is 97.5%, which is better than other comparative models. The study provides a high-precision solution for converter valve condition monitoring and effectively extends the application of multi-source data fusion in power equipment.

Keywords: converter valve, multi-source data fusion, random forest, harmonic theory, condition monitoring

✉ Corresponding author.
E-mail address: Wanguhvgz@163.com (Z. Wang).
Received 12 January 2024; Revised 05 March 2024; Accepted 25 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-422

1. Introduction

Currently, flexible DC transmission technology is widely used in the transmission of distributed energy sources, interconnection between regional power grids, and power supply to urban centers. Flexible DC transmission system generally adopts modular multilevel converter topology, and the number of half-bridge power modules is as few as several hundred or as many as several thousand [1-3]. During long-term operation, the voltage and current of the power modules change leading to aging of individual components. This aging to failure is parametric failure, also known as soft failure. At present, the monitoring function of the flexible DC transmission converter valve is less, and generally can only provide simple information such as the voltage, status word and fault word of the power module [4-7].

Flexible DC transmission converter valve consists of thousands of power sub-modules, each of which needs to be controlled independently, so the characterization data of the operating state of the flexible DC transmission converter valve is huge and complicated to analyze and process [8-9]. Currently, the state monitoring function of the flexible DC transmission converter valve is only limited to the detection and processing of the sub-module capacitance and IGBT (Insulate-Gate Bipolar Transistor) after the failure, as well as the leakage monitoring of the valve tower, and it does not have the monitoring function of the rest of the key components of the valve assembly, especially for the lack of on-line analyzing means of the sub-module in the operation [10-12]. The data acquisition, data analysis and processing, fault characterization, data caching, human-machine exchange, etc. of the operating parameters of the converter valve are very urgent in the electrical field, so it is necessary to carry out in-depth research [13-15].

How to timely and effectively find the various faults and hidden dangers of the equipment, such as the temperature, pressure, torque, and small leakage current of the key components of the converter valve, and the non-normal operation status of the cooling pipe leakage of the valve tower, to provide a reliable basis for analyzing and judging of the converter valve operators, and to avoid accidents from occurring or expanding further, is an important topic that needs to be seriously researched and solved at present [16-18]. Through the development of converter valve key components operating temperature and other physical quantities online monitoring with embedded miniaturized low-power intelligent sensors, so that the perfect integration into the converter valve, not only to ensure that the miniature sensors work reliably, but also does not affect the performance of the existing converter valve, is to achieve the converter valve operation reliability and safety enhancement is an important way to achieve the converter valve, and at the same time for the eventual realization of the converter valve intelligent condition maintenance to lay the foundation [19-21]. Converter valve is the core equipment to complete the AC/DC energy conversion in DC transmission system, and its operation status is related to the reliability of the whole DC project [22].

In this paper, the following steps are followed to carry out the condition monitoring of converter valves in high-voltage direct current transmission systems. (1) A novel multi-source data fusion system is designed with the support of random forest structurer. (2) When extracting the time-varying laws of the electrical characteristics of the converter valve, the harmonic characteristics (i.e., the electrical characteristics of the converter valve) are analyzed by integrating the time domain and wavelet analysis, the features are downscaled based on the principal component analysis, and the features are selected based on the Random Forest reordering of the importance. (3) Particle swarm algorithm is used to optimize the number of decision trees and the minimum number of leaves of the random forest to improve the classification accuracy of the random forest model for fault types. (4) Based on PSCAD/EMTDC, ±800kV UHV DC transmission system is constructed, and the effect of converter valve state monitoring is judged by simulation test.

2. Multi-source data fusion system design

The converter valve is an important device in DC transmission, and the converter valve includes gate unit assembly, thyristor press-fit structure, damping resistor assembly, saturation reactor assembly, damping capacitor assembly, cooling pipeline, DC equalizing resistor, and electrical connection structure, etc., and there exists multi-source data [23]. It is very easy to produce obvious unexplained scheduling coverage behavior among multi-source data, which leads to a rapid decline in the error-free fusion capability of fault information parameters. In order to enhance the efficiency of converter valve condition monitoring, the study designs a novel multi-source data fusion system with the support of random forest structurer.

2.1. Hardware design of multi-source data fusion system

The hardware execution environment of the multi-source data fusion system consists of two parts: the information management module and the multi-source fault analysis module.

2.1.1. Information management module. When the converter valve information is entered into the data fusion system, the execution element must directly open the scheduling transfer processing of the information parameter and transform the unstored data file into the form of fusion to be stored. In order to make the system have strong fusion processing capability, the information management module can maintain the query and inventory function of data parameters at the same time, and in the case of ensuring the complete transmission of multiple sources, each scheduling node can directly execute the agent application program that has been shaped and can establish the physical connection with the system database under the role of random forest decision tree organization, so as to realize the direct scheduling output of the stored information parameter, and reduce the number of The level of unexplained coverage between multiple sources of data is accounted for.

2.1.2. Multi-source fault analysis module. The multi-source fault analysis module and the system information management module belong to the level hardware structure. Under the action of the fault analysis host, the processor can divide all the input information into several equal structures equally, and then fuse the temporary data parameters in the converter valve network to form a bundle of information fluids, so as to ensure the smooth execution of the subsequent commands for the fusion of multi-source fault data. In order to ensure the error-free fusion of the input information parameters, the multi-source fault data in the converter valve network can be precisely adjusted to the generalization error coefficients with the allocation of the decision tree training subset.

2.2. Multi-source data fusion system software design

With the support of the relevant hardware execution structure, according to the processing flow of base classifier construction, decision tree training set allocation, generalization error calculation, to realize the improvement of the system software execution environment, combined with the hardware design, to complete the design of Random Forest-based converter valve multi-source data fusion system.

2.2.1. Base Classifiers for Random Forests. The base classifier of the fusion system based on random forest takes the form of decision tree connection as shown in Fig. 1, which can be used to realize the transmission and simplified processing of multi-source fault data of the converter valve in accordance with the recursive analysis, and ultimately generates the inverted tree structure to shorten the spacing of fusion implementation between neighboring nodes. The common base classifier decision tree organization consists of head node, middle node, and tail node together (e.g., 0, 1, and 2 in Fig. 1), in order to ensure the real-time scheduling and fusion of multi-source data, the 0 node usually represents the decision tree organization structure connected with the management module, the 1 node, as an intermediate transmission unit, is often connected with the multi-source analysis module, and in the case of the tail node maintaining a stable connection, the 2 node can directly reflect the converter valve The final fusion result of the fault information data, thus providing the multi-source parameter structure required for the construction of the decision tree training set samples.

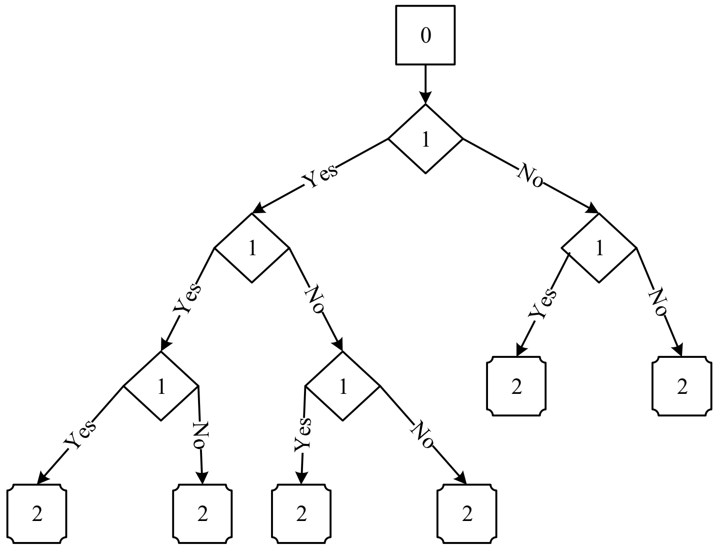


Fig. 1. Structure of random forest based classifier

2.2.2. Decision Tree Training Set. The decision tree training set is a collection of multi-source fault information data associated with a random forest, which is able to fully adapt to the basic execution needs of random dispatch instructions and filter the information parameter structure that best meets the system fusion criteria. If one system fusion cycle is taken as the established discriminative length of the information data, the theoretical length criterion of the decision tree training set is only affected by the commutator valve scheduling authority and the first connection of the multi-source fault information characterization quantity. Converter valve scheduling authority is often expressed as q', and throughout the data fusion cycle, it always maintains a relatively stable fluctuation trend. Multi-source fault information characterization quantity is often expressed as $\dot{y}$. Unlike other application indexes, this physical quantity is not affected by other information parameters except the fusion time, and has strong computational stability. Let $\bar{T}$ represent the standard length of one data fusion cycle in the system, the theoretical length of the decision tree training set can be defined as:

$$L = \frac{\dot{y} \cdot \log_q \left(\frac{w_{\max}}{w_{\min}}\right)}{\bar{T}^2} \tag{1}$$

where $w_{\max}$ represents the maximum data transmission parameter difference associated with multi-source fault information and $w_{\min}$ represents the minimum data transmission parameter difference associated with multi-source fault information.

2.2.3. Fusion generalization error calculation. Generalization error is an application metric related to the generalization response capability of random forests in a converter valve multi-source data fusion system. In general, the smaller the real value of the generalization error, the better the learning performance of the system, and vice versa, the worse the performance. Under the premise that the distribution samples of the decision tree training set have been determined, the value of the system fusion generalization error is directly affected by two physical quantities, namely, the number of decision tree nodes and the multi-source distribution coefficient. The number of decision tree nodes can be denoted as $\tilde{k}$, and this physical quantity always fluctuates between the upper and lower extremes under the effect of the random forest decision tree. The multi-source distribution coefficient can be expressed as β. Usually, the actual value of this physical quantity is always less than 1, but the physical ability to contribute to the final generalization error calculation results always remains positive. With the support of the above physical quantity, the associative equation (1). The result of the system fusion generalization error calculation can be expressed as:

$$s = \int_{e_0}^{e_1} \frac{(1+\beta^2)\cdot L}{\chi \cdot \tilde{k}^2} d\tilde{k} \tag{2}$$

In the formula, e_0 represents the lower limit fluctuation extreme value of the converter valve scheduling information, e_1 represents the upper limit fluctuation extreme value of the converter valve scheduling information, and χ represents the quantization coefficient of the power limit disposition of the multi-source fault information data. So far, the establishment of various software and hardware execution conditions is completed, and the smooth application of the converter valve multi-source fault information data fusion system is realized under the support of random forest decision tree.

3. Converter valve electrical characteristics time-varying law extraction and selection

Based on the multi-source data fusion system constructed above, this chapter utilizes the system data combined with the Random Forest algorithm to extract the time-varying laws of the electrical characteristics of the converter valve.

3.1. Converter valve fault harmonic theory characteristics

Due to the nonlinearity of the converter, a large number of harmonics are generated during the operation of the UHV DC transmission system, which causes the distortion of voltage and current in the transmission system, and this distortion is a kind of "pollution" to the electric energy. The converter is a harmonic current source for AC system, and it is also a harmonic voltage source for DC line. The core of the converter is the converter valve, and the converter valve is the key equipment of the converter, therefore, through the converter valve fault (valve error opening and valve does not open the fault) harmonic characteristics of the converter valve electrical characteristics of the time-varying law.

The voltage and current waveforms of the AC system are sinusoidal under ideal conditions and satisfy the following expressions:

$$u_{(t)} = \sqrt{2}\sin(\omega t + \alpha) \tag{3}$$

In the above equation: U is the rms value of the voltage. ω is the angular frequency. α is the initial phase angle, where $\omega = 2\pi f = \frac{2\pi}{T}$. However, in the actual operation of the grid, the voltage will change due to the influence of the converter and other factors. The non-sinusoidal voltage $u(\omega t)$ with period $T = 2\pi/\omega$ is transformed using Fourier and the results obtained are shown in expressions (4) and (6):

$$f(\omega t) = A_{(0)} + \sum_{n=1}^{\infty}(A_{(n)}\cos(n\omega t) + B_{(n)}\sin(n\omega t)) \tag{4}$$

In the formula:

$$\begin{cases} A_{(0)} = \frac{1}{2\pi}\int_0^{2\pi} f(\omega t)d(\omega t) \\ A_{(n)} = \frac{1}{\pi}\int_0^{2\pi} f(\omega t)\cos(n\omega t)d(\omega t) \\ B_{(n)} = \frac{1}{\pi}\int_0^{2\pi} f(\omega t)\sin(n\omega t)d(\omega t) \\ n = 1,2,3\cdots \end{cases} \tag{5}$$

Or:

$$f(\omega t) = A_{(0)} + \sum_{n=1}^{\infty} C_{(n)} \sin(n\omega t + \varphi_{(n)}) \tag{6}$$

Comparing equations (4) and (6) together, the coefficient relationship between the two is obtained as in equation (7):

$$\begin{cases} C_{(n)} = \sqrt{A_{(n)}^2 + B_{(n)}^2} \\ \varphi_{(n)} = \tan^{-1}\left(\frac{A_{(n)}}{B_{(n)}}\right) \\ A_{(n)} = C_{(n)} \sin\varphi_{(n)} \\ B_{(n)} = C_{(n)} \sin\varphi_{(n)} \end{cases} \tag{7}$$

Or in the plural as:

$$B_{(n)} + jA_{(n)} = C_{(n)} e^{i\varphi(n)} \tag{8}$$

In the Fourier series of Eqs. (4) and (6), the frequency of the fundamental wave is $1/T$. Making $\omega t = \omega t + \theta_0$ in Eq. (6) yields the following equation:

$$f(\omega t + \theta_0) = A_{(0)} + \sum_{n=1}^{\infty} C_{(n)} \sin(n\omega t + (n\theta_0 + \varphi_{(n)})) \tag{9}$$

Comparing Eq. (6) with Eq. (9), it can be concluded that the amplitude of the dc component in $f(\omega t + \theta_0)$ and $f(\omega t)$, the amplitude of the fundamental wave and the amplitude of each of their harmonics are equal in currents or voltages with frequency f, respectively. While the initial phase of the fundamental and each harmonic of $f(\omega t + \theta_0)$ and $f(\omega t)$ are increased by $n\theta_0 (n = 1,2,3\cdots)$ respectively.

3.2. Feature extraction

1) Based on time domain feature extraction. In this paper, 11 time-domain features are extracted separately as inputs to the subsequently designed condition monitoring model. The extracted time-domain features are peak-to-peak value, average value, mean square value, standard deviation, RMS value, variance, peak factor, impulse factor, margin factor, crag factor, and skew factor of the current [24].

The peak-to-peak value (x_{pp}) is the difference between the maximum and minimum values of the current signal, which indicates the range of signal variation, and the formula for the peak-to-peak value is shown in equation (10):

$$x_{pp} = x_{\max} - x_{\min} \tag{10}$$

The average value ($\bar{x}$) is the average value of the current signal in one cycle, and the formula for the average value is shown in equation (11):

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{11}$$

The mean square value (x_e) is the average value of the square of the current signal in one cycle, and the formula for calculating the mean square value is shown in equation (12):

$$x_e = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{12}$$

The RMS value (x_{rms}) is the open square of the rms value of the current signal, also called the root-mean-square value, and the formula for calculating the RMS value is shown in equation (13):

$$x_{ms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{13}$$

The standard deviation (x_{mse}) is the square of the difference between the current signal and its average value in a cycle of the open square, the standard deviation of the formula shown in equation (14):

$$x_{mse} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{14}$$

The variance (x_{var}) describes the fluctuation range of the current signal and indicates the strength of the AC component of the signal, i.e., the average power of the AC signal, and the formula for calculating the variance is shown in equation (15):

$$x_{var} = \frac{\sum_{i=1}^{N} (x - \bar{x})^2}{N-1} \tag{15}$$

The peak factor (C_f) is the ratio of the maximum value of the current signal to the RMS value, which indicates the extreme degree of the peak value of the current signal compared with the overall waveform, and the formula for calculating the peak factor is shown in equation (16):

$$Cf = \frac{x_{\max}}{x_{rns}} = \frac{\max(x_i)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}} \tag{16}$$

The pulse factor (I_f) is the ratio of the maximum value of the current signal to the absolute mean value, which indicates the impact nature of the current signal, and the formula for calculating the pulse factor is shown in equation (17):

$$I_f = \frac{x_{\max}}{|\bar{x}|} = \frac{\max(x_i)}{\frac{1}{N}\sum_{i=1}^{N} |x_i|} \tag{17}$$

The margin factor (CL_f) is the ratio of the peak value of the current signal to the square root amplitude, which indicates the fullness of the current signal, and the calculation formula of the margin factor is shown in equation (18):

$$CL_f = \frac{x_{\max}}{x_r} = \frac{X_{\max}}{(\frac{1}{N}\sum_{i=1}^{N}\sqrt{|x_i|})^2} \tag{18}$$

The crag factor (K_v) is the ratio of the signal crag value and the fourth power of the root mean square value, which indicates the degree of smoothness of the current waveform, and the formula for calculating the crag factor is shown in equation (19):

$$K_v = \frac{\beta}{x_{rms}^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}x_i^4}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2})^4} \tag{19}$$

The skewness factor (S_k) is the ratio of the normalized 3rd order center distance and standard deviation 3rd power, which indicates the crest distribution of the current waveform, and the formula distribution of the skewness factor is shown in equation (20):

$$S_k = \frac{\frac{1}{N}\sum_{i=1}^{N}x_i^3}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2})^3} \tag{20}$$

where, β is the signal cliff value, x_{rms} is the root mean square value of the signal, N is the number of discrete signals, x_i is the amplitude of the i th discrete signal, $x_{\max}$ is the peak value of the current signal, $x_{\min}$ is the minimum value of the current signal, $|\bar{x}|$ is the absolute mean value of the signal, x_r is the square root amplitude of the signal, and $x_\tau(\tau)$ is the sequence of the discrete signals processed.

2) Based on wavelet packet energy feature extraction. Wavelet packet analysis can better analyze the time-frequency localization of signals containing a large amount of low-frequency and high-frequency information [25]. Therefore, in this paper, the wavelet packet transform is used to process the current signal to realize the calculation of harmonic energy. Orthogonal wavelet packet bases are generated from standard orthogonal scale parameters by means of two-scale difference equations. The orthogonal wavelet packet basis is calculated as in equation (21) and the bi-orthogonal difference equation is shown in equation (22):

$$w_{n,j,k}(t) = 2^{-j/2}w_n(2^{-j}t-k, n\in Z/Z^{-1}, j,k\in Z) \tag{21}$$

$$\begin{cases} w_{2n}(t) = \sqrt{2}\sum_{k\in Z}h_k w_n(2t-k) \\ w_{2n+1}(t) = \sqrt{2}\sum_{k\in Z}g_k w_n(2t-k) \end{cases} \tag{22}$$

where n,j,k denotes the position of the wavelet packet in the hierarchy, h_k is the low-pass filter and g_k is the high-pass filter, respectively. The wavelet packet decomposition coefficients are obtained from the projection of signal $S(t)$ in the orthogonal wavelet basis space. The calculation is shown in equation (23):

$$c_{n,j,k} = \int_{-\infty}^{+\infty}S(t)w_{n,j,k}(t)dt \tag{23}$$

When the wavelet packet basis functions are a set of orthogonal bases, the wavelet packet transform has the property of energy conservation. The wavelet packet energy at a single scale is the sum of the squares of the wavelet packet coefficients at that scale. The calculation is shown in equation (24):

$$E_j = \sum_k c_{n,j,k}^2 \tag{24}$$

3.3. Principal component analysis based feature dimensionality reduction

Principal Component Analysis (PCA) can be used to reduce the dimensionality of the features to reduce the dimensionality of the network inputs, reduce data redundancy and improve the training efficiency and speed of the model while retaining the classification accuracy of the features that can distinguish between current signals of converter valve faults and those of normal operating conditions [26].

Setting the first m principal component in the orthogonal space with the eigenvalues ranked from largest to smallest is $y_1, y_2, \dots, y_m$, then its cumulative variance contribution is obtained as shown in equation (25):

$$\phi(m) = \frac{\sum_{m}^{k+1} \lambda_k}{\sum_{n}^{i+1} \lambda_n} \times 100\% \tag{25}$$

where, $i = 1, 2, 3, \dots, n$, n is the number of eigenvalues before dimensionality reduction. $k = 1, 2, 3, \dots, m$, m is the number of eigenvalues after dimensionality reduction After calculating the cumulative variance contribution rate, it is required that the weight of the cumulative variance contribution rate of several main components is large enough, and in this topic, the cumulative variance contribution rate is set to 98%. That is, when $\phi(m) > 98\%$, it indicates that the matrix after dimensionality reduction contains 98% of the original information. The first m principal components are chosen to represent the information contained in the original signal and the new feature matrix is constructed.

3.4. Feature selection based on random forest importance ranking

Features used in machine learning are rarely completely independent, making it difficult to interpret the importance of features. Random Forest (RF) is a classifier that utilizes multiple decision trees to train and predict samples. At each node of the tree, a decision tree is generated by randomly selecting a few features from all the original features, calculating a splitting criterion based on the selected features, and the feature with the best value is used to split the node [27]. Splitting the data in the decision tree produces leaf nodes, which represent a class label and the path from the root node to the leaf node represents a defined decision process. RF can calculate the relative importance of features using out-of-bag error rate (OBB) and rank and filter the features to evaluate the importance of each feature.

The input feature matrix is first computed for its Gini impurity, which is a measure of the likelihood of misclassifying a new instance of a random variable, and is calculated as shown in (26):

$$I_{\mathrm{G}}(f) = \sum_1^m f_i(1-f_i) = \sum_1^m f_i - \sum_1^m f_i^2 = 1 - \sum_1^m f_i^2 \tag{26}$$

In the formula, f_i indicates the proportion of samples belonging to a certain category. Taking the converter valve detection as an example, there are two types of normal and faulty, so $m = 2$. At this time, f_1 indicates the proportion of normal samples and f_2 indicates the proportion of faulty samples.

Through the formula (27), the Gini impurity of a feature on the decision tree $I_G - E(f)$ and the Gini impurity of the decision tree without the feature $I_{G\sim NE}(f)$ are calculated as the reduction of the Gini impurity of all the decision trees, and n is the number of decision trees in the RF:

$$\Delta IG^n(f) = IG\_NE(f) - IG\_E(f) \tag{27}$$

The MDI is the average of the Gini impurity reduction for all decision trees, and this value quantitatively describes the effect of each feature on the random forest. N is the feature dimension. The larger the MDI, the higher the classification accuracy of the features and the more stable the RF model:

$$MDI = \frac{\sum_{1}^{N} \Delta IG^{n}(f)}{N} \tag{28}$$

4. Condition monitoring modeling

Currently widely used in fault detection BP neural network has shortcomings such as slow convergence speed and poor accuracy, so in order to solve the problem, this paper is based on the random forest algorithm, and combined with the particle swarm algorithm to optimize the parameters of the random forest, constructed random forest particle swarm optimization algorithm identifies the status monitoring model of the faults of the converter valve.

4.1. Random Forest Particle Swarm Optimization Model for Fault Detection

Random Forest is a Bagging algorithm that uses decision trees as estimators, and he combines multiple decision trees with the advantages of fewer parameters, less likely to be overfitted, and better generalization ability. Unlike traditional decision trees, each division of the decision tree in Random Forest does not use the feature set of the current node (assuming M feature), but randomly selects a certain number (assuming m) of feature subsets from the feature set of the node, and then selects an optimal feature from the subset for division. In general, m is more appropriate to take $\log_2 M$ or $\sqrt{M}$. Like the Bagging algorithm, the Random Forest algorithm also requires the use of a voting strategy. The voting strategy is a method that follows the principle of majority rule, i.e., the results of multiple weak classifiers are aggregated, and then the plurality of their classification categories is taken as the final classification result to improve the model's generalization ability and classification accuracy. The expression of the output result of random forest is (29):

$$R(x) = \frac{1}{L}\sum^{L} y_i \tag{29}$$

where: L is the number of decision tree trees in the random forest, γ_i is the output result of each decision tree, and $R(x)$ is the output result of the random forest.

When constructing the random forest algorithm model, the number of decision trees and the minimum number of leaves are two important parameters to be determined. Generally speaking, the more decision trees and the smaller minimum number of leaves can significantly improve the classification accuracy of the random forest, but it also increases the complexity of the random forest, leading to overfitting of the random forest and reducing its performance. In this paper, the particle swarm optimization algorithm is used to optimize the number of decision trees and the minimum number of leaves of the random forest to reduce the complexity of the random forest. The particle swarm optimization algorithm originates from the observation of bird foraging behavior, through the search for the optimal population convergence to the global optimum, can be very good for the random forest decision tree number and the minimum number of leaves to optimize, so as to improve the accuracy of the classification of the random forest model. The main steps of the random forest particle swarm optimization algorithm used in this paper are:

1) Determine the initial parameters of the algorithm, randomly set the number of decision trees, the minimum number of leaves, and the default number of random features m is $\sqrt{M}$.

2) Using the classification accuracy as the fitness value, use the particle swarm optimization algorithm to search for the optimal number of decision trees L and the optimal number of leaves MINLS.

3) Divide the sample set into training set and test set.

4) Bootstrap is used to generate L sample sets for the training set, and each sample set is used to generate a decision tree.

5) When each node of the decision tree is split, the feature set of the node is sampled without putback, m features are selected, and only the data corresponding to the m features are retained as the training samples, and if the number of samples contained in the split node is less than the minimum number of leaves MINLS, the decision tree stops splitting.

6) The training generates L decision trees, and the final classification result of the model is obtained by Eq. (16).

4.2. Converter valve fault identification process

First, the current waveforms during normal operation and when a converter valve fault occurs are collected to extract the time-domain features of the current, and the time-domain features are normalized. Then, the feature selection method in Section 3.4 is utilized to select the obtained commutator valve feature indexes to construct the commutator valve feature index set, which is used as the input features of the classification model. Finally, the classification model is trained and tested using the converter valve feature indicator set.

5. Simulation testing

In this paper, ±800kV UHV DC transmission system is built based on PSCAD/EMTDC, and simulation experiments of converter valve state monitoring are carried out under this system.

5.1. Simulation Model and Fault Types

The bipolar operation of ±800kV UHV DC transmission system is rated at 8000MW, the DC current is rated at 5kA, the total length of DC transmission line is 1720km, and there are 300mH and 75mH dry reactors installed on the DC line at the rectifier side, the inverter side and the neutral bus side. The transformers of the converter station are of double winding type and the equipment parameters of the transformers are shown in Table 1. According to the given equipment parameters, the simulation model of the UHV DC transmission system established by using PSCAD/EMTDC software is shown in Fig. 2.

Table 1. The parameters of the UHVDC system changer for the rheological pressure

	Single capacity (MVA)	Network side winding rated voltage (kV)	Valve side winding rated voltage (kV)	Transformer ratio (%)	Percentage of short circuit impedance (%)	Flow impedance (Pu)
Rectifier side	403.74	441.66	174.9	1.162	19.6	0.28
Inverse side	386.92	524.98	154.53	1.091	17.6	0.28

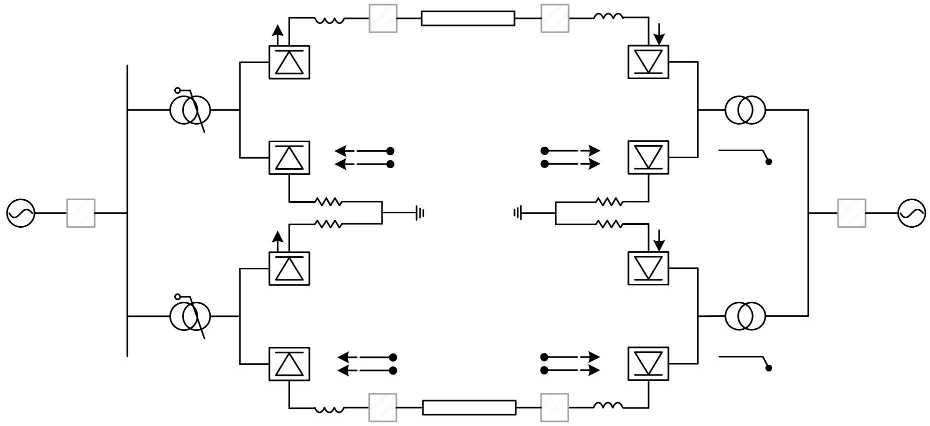


Fig. 2. UHVDC system model

Based on the harmonic theory characteristics can be converter valve faults can be categorized into the following three types:

1) Valve does not open fault. Converter valve conduction conditions for receiving the trigger pulse and positive voltage, if the valve is only in the positive voltage under the action of no trigger pulse (trigger pulse loss), the valve will produce a non-conduction fault. Valve in the non-conduction fault under each valve on-off waveforms shown in Figure 3. To 6 pulsating converter in the V1 does not open fault as an example, at this time, in a cycle V1 does not conduct, so that V5 has been subjected to the forward voltage and is in the state of conduction, when the trigger pulse of V3 arrives, V5 is closed and does not conduct, the other valves of the conduction sequence and the conduction interval with the same as the normal operating state, the rectifier side and inverter side of the on/off waveforms are the same.

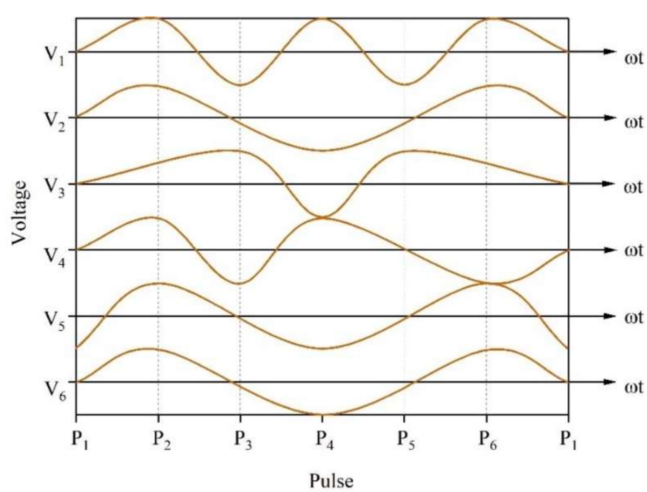


Fig. 3. Valve pass off the waveform under the failure of the valve

2) Rectifier-side converter valve misopening failure. Rectifier valve misopening fault valve on-off waveform shown in Figure 4. Taking the V1 misopening fault in the 6-pulsation converter as an example, when the trigger pulse reaches V1, V1 is subjected to a positive voltage and immediately conducts, which leads to the early shutdown of V6, and the other valves' conduction sequence and conduction interval remain unchanged.

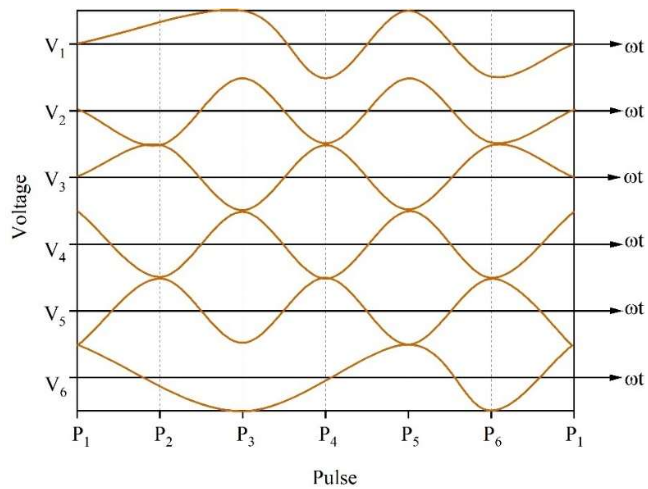


Fig. 4. The rectifier valve is wrong to open the valve through the waveform

3) Inverter-side converter valve misopening failure. Taking the V1 misopening fault in the 6-pulsation converter as an example, when the trigger pulse of V5 arrives, V3 fails to commutate to V5 because V5 is still in reverse voltage, and at this time V1 is in the forward voltage state, so the mis-triggered pulse of V1 causes V3 to commutate to V1 directly, which makes the on-time interval of V1 larger, and the other valves have unchanged on-time sequence and on-time interval, and the fault occurred at this time is equivalent to a commutation failure of the system. This fault is equivalent to a phase change failure of the system, and the waveform of valve turn-on and turn-off under the fault of inverter valve wrong turn-on is shown in Fig. 5. When the valve does not conduct or continues to conduct the fault, the three-phase commutation current is not symmetrical, but the three-phase commutation voltage is considered to be still symmetrical, which leads to the generation of 100Hz harmonic components in the dc current and dc voltage, and the control of the valves causes fundamental frequency oscillation.

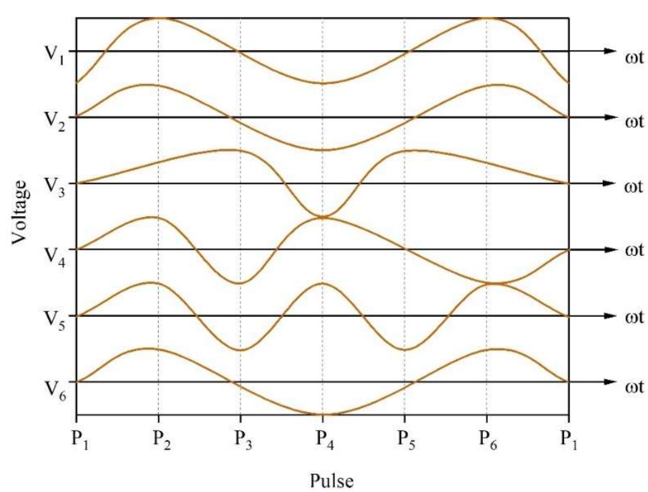


Fig. 5. The inverter valve fails to open the valve through the waveform

5.2. Feature Extraction and Selection

According to the feature extraction method designed in the previous section to extract the harmonic characteristics of the converter valve from the multi-source data fusion system, the time domain statistical features and wavelet packet energy features are extracted from the converter valve under each operating state, and 100 sets of feature samples are extracted under each operating state.

The normalized values of wavelet packet energy for the 11 time-domain characteristics of the converter valve under different operating conditions are shown in Figure 6. The numbers "1~11" represent 11 time-domain characteristics, including peak-to-peak, mean, mean square, standard deviation, rms, variance, peak factor, pulse factor, margin factor, kurtosis factor, and skewness factor. Type1, Type3, and Type3 respectively indicate that the valve is not opened fault, the rectifier side converter valve is opened by mistake, and the converter valve on the inverter side is opened by mistake.

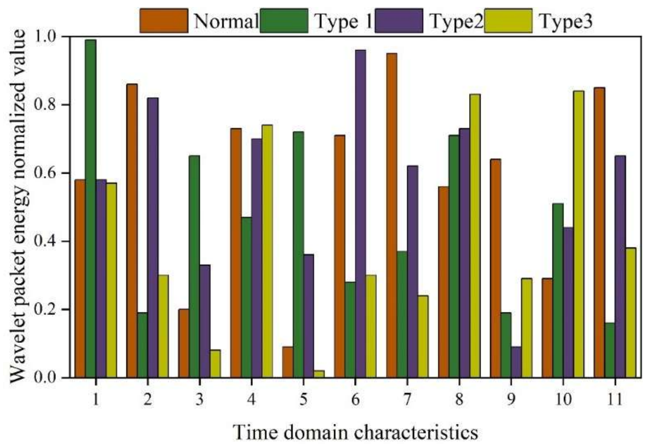


Fig. 6. The wavelet packet energy normalized value

The visualization effect of feature dimensionality reduction based on principal component analysis is shown in Fig. 7. After the harmonic feature extraction by integrated time domain and wavelet analysis, the samples corresponding to the four operating states have preliminary differentiation.

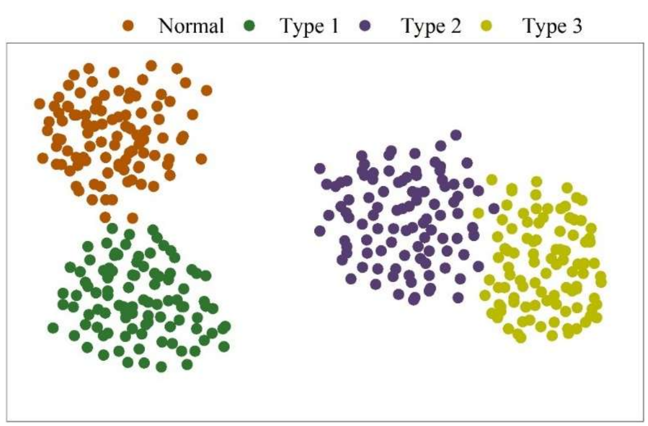


Fig. 7. Feature degradation visual effect

A feature selection method based on random forest importance ranking is introduced to optimize the selection of commutator valve feature parameters. The original 11 time-domain features are input into the random forest model, and the importance of each feature parameter for fault classification is shown in Figure 8. Except for the importance of “mean square value (0.572)” which does not exceed 0.8, the importance of the rest of the time-domain features are all above 0.8, so the remaining 10 time-domain features are selected to form the converter valve feature index set, which is applied to the condition monitoring model as the time-varying law of the converter valve electrical characteristics.

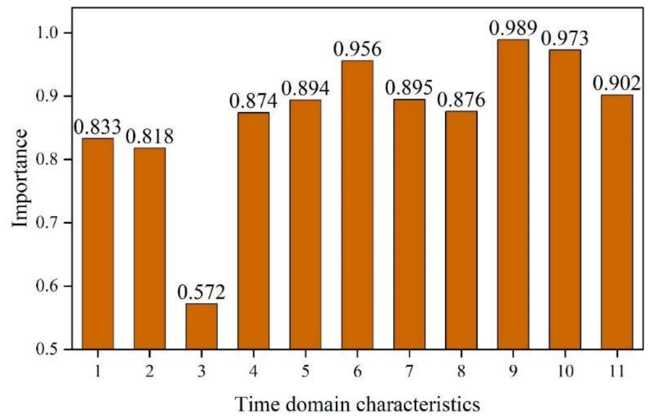


Fig. 8. The importance of each characteristic parameter to fault classification

5.3. *Analysis of condition monitoring results*

60% of the samples in each operation state are randomly selected as the training set, and the remaining 40% is used as the test set. Based on the training set, the decision rules of the random forest model are trained, and the hyperparameters of the model must be set before training, which are the framework parameters of the random forest model and the decision tree parameters. The particle swarm optimization algorithm searches for the optimal policy, and determines the number of decision trees, the maximum depth of the root node of the decision tree, and other key parameters of the random forest model. The fault diagnosis results of the converter valve based on the random forest particle swarm optimization model are shown in Figure 9. The number of correctly diagnosed samples reaches 156, and the accuracy rate is 97.5%.

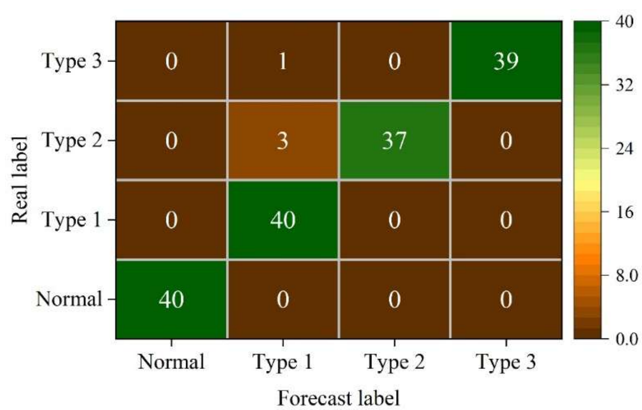
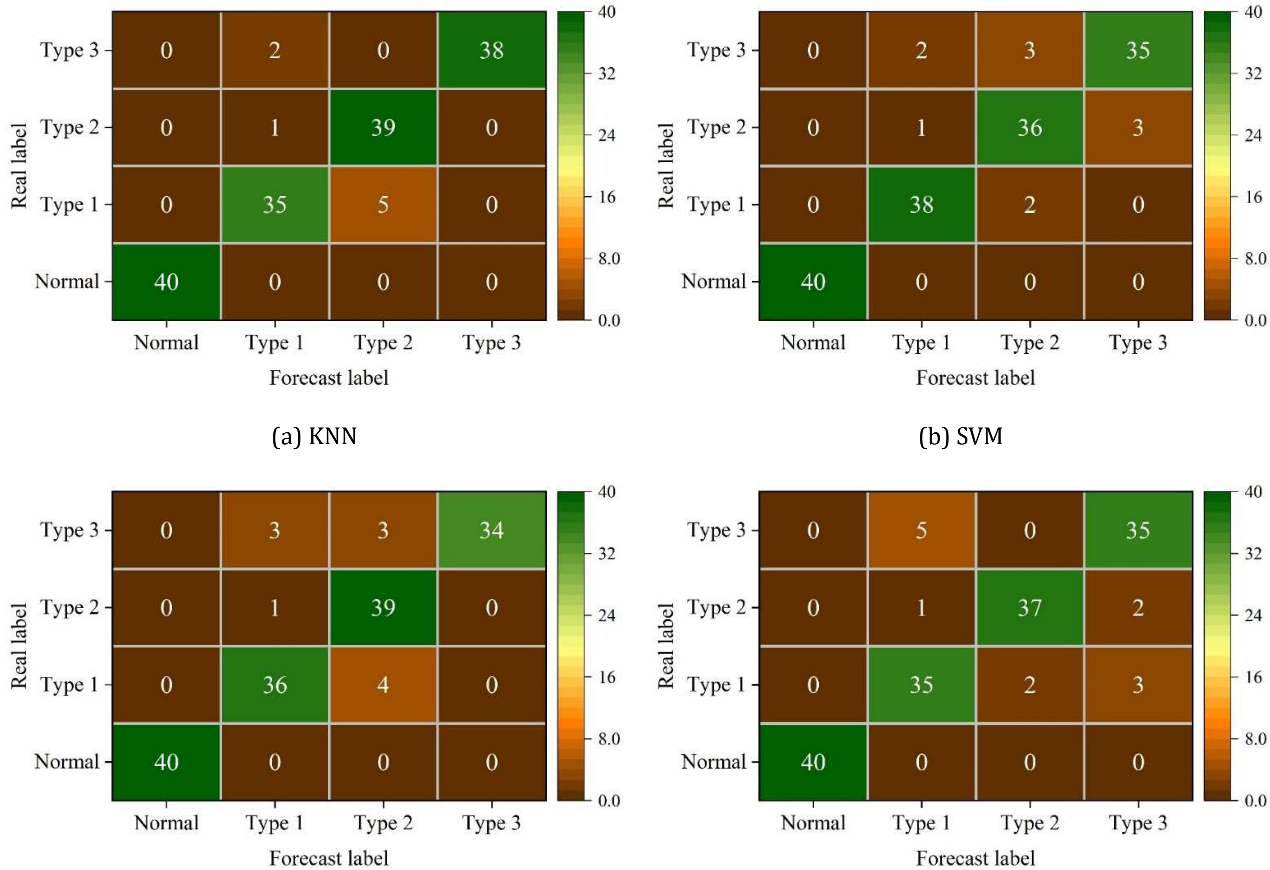


Fig. 9. Error of flow valve fault diagnosis

The constructed Random Forest Particle Swarm Optimization model is compared with machine learning models such as KNN, SVM, Logistic Steele Regression and Gaussian Simple Bayes. Under the condition that the training set and test set are the same, the classification effect of the four compared models is shown in Fig. 10. (a)~(d) represent the classification results of KNN, SVM, Logistic Steele Regression, and Gaussian Simple Bayes models, respectively. 95%, 93.13%, 93.13%, and 91.88% of the classification accuracies of the four comparison models, respectively.



(a) KNN (b) SVM

(c) Logistic regression (d) Gauss plain Beth

Fig. 10. Comparison model classification effect

In summary, it can be seen that all fault diagnostic models predict the normal state of the converter valve accurately, different fault diagnostic models produce different inductive preferences under the condition that the training set is the same, and the accuracy rate of the random forest particle swarm optimization model for fault detection constructed in this paper is 97.5%, and the state monitoring effect is better than that of other comparative models.

6. Conclusion

The research constructs a multi-source data fusion system to realize the error-free fusion of fault information parameters. Under this research result, Random Forest algorithm is combined to extract the time-varying law of electrical characteristics of the converter valve, and Random Forest Particle Swarm Optimization model is constructed for the state monitoring of the converter valve. Based on the harmonic theory characteristics, the converter valve faults are categorized into valve non-opening faults, rectifier-side converter valve misopening faults and inverter-side converter valve misopening faults. The normal state and the three fault states are categorized using the model in this paper and machine learning models such as KNN, SVM, logistic Steele regression, Gaussian plain Bayes, etc. The classification accuracy of this model is higher than that of the model in this paper. The classification accuracy of this paper's model is 2.5%, 4.37%, 4.37%, and 5.62% higher than the remaining four models, respectively. It shows that using the method of this paper for converter valve state monitoring can ensure the fault identification effect.

Funding

This work was supported by the Science and Technology Project of China Southern Power Grid Company Limited under (CGYKJXM20220335).

References

[1] Liu, C., Han, K., Yao, W., Hu, S., & Zheng, X. (2018, November). Electric-field Simulation of Insulation Type Test of±420kV HVDC Flexible Converter Valve. In 2018 IEEE International Power Electronics and Application Conference and Exposition (PEAC) (pp. 1-4). IEEE.

[2] Watson, N. R., & Watson, J. D. (2020). An overview of HVDC technology. Energies, 13(17), 4342.

[3] Wang, X., Li, J., Lu, Y., Yang, Y., Li, L., & Wang, W. (2024, August). Engineering Application of Valve Monitoring System for Flexible HVDC. In 2024 International Conference on HVDC (HVDC) (pp. 985-990). IEEE.

[4] Paez, J. D., Frey, D., Maneiro, J., Bacha, S., & Dworakowski, P. (2018). Overview of DC–DC converters dedicated to HVdc grids. IEEE Transactions on Power Delivery, 34(1), 119-128.

[5] Jianjian, G. A. O., Zhuokun, L. I. U., Guoqiang, L. I., Kai, Z. H. O. U., Ruicheng, D. A. I., Bi, Z. H. A. O., ... & Jianrui, S. O. N. G. (2024). A Review of Flexible DC Converter Valves and Valve Base Controller in China. Power System Technology, 48(9), 3890-3902.

[6] Pan, H., Wang, X., Jia, X., Hua, H., & Xu, J. (2022). Redundancy configuration strategy and condition maintenance arrangement of converter valves for mmc-hvdc transmission systems considering both reliability and economy. Energy Reports, 8, 452-460.

[7] Li, G., Li, C., & Van Hertem, D. (2016). HVDC technology overview. HVDC grids: for offshore and supergrid of the future, 45-78.

[8] Weidong, Z., Qian, W., Lei, Q., Donglai, Z., Guoliang, Z., & Linhai, C. (2016, September). The prediction of radio frequency interference from HVDC-flexible converter valve. In 2016 International Symposium on Electromagnetic Compatibility-EMC EUROPE (pp. 790-793). IEEE.

[9] Lu, J., Li, X., Tang, R., Wang, C., Wu, S., He, Z., & Hu, R. (2022, October). Research on Flexible HVDC Converter Valve and VBC Redundancy Reliability Improvement. In Journal of Physics: Conference Series (Vol. 2310, No. 1, p. 012005). IOP Publishing.

[10] Stan, A., Costinaș, S., & Ion, G. (2022). Overview and assessment of HVDC current applications and future trends. Energies, 15(3), 1193.

[11] Yang, Y., Li, Q., Li, J., Cheng, C., Zhou, J., & Cao, K. (2024, August). Research on Control and Protection Technology for Equivalent Operation Test of Adaptive Flexible Valve Module. In 2024 International Conference on HVDC (HVDC) (pp. 956-962). IEEE.

[12] Chen, N., Zha, K., Qu, H., Li, F., Xue, Y., & Zhang, X. P. (2022). Economy analysis of flexible LCC-HVDC systems with controllable capacitors. CSEE Journal of Power and Energy Systems, 8(6), 1708-1719.

[13] Sun, Q., Song, Q., Meng, J., Cui, B., Li, G., Cheng, F., & Wang, K. (2023). Flex-LCC: A new grid-forming HVDC rectifier for collecting large-scale renewable energy. IEEE Transactions on Industrial Electronics, 71(8), 8808-8818.

[14] Korompili, A., Wu, Q., & Zhao, H. (2016). Review of VSC HVDC connection for offshore wind power integration. Renewable and Sustainable Energy Reviews, 59, 1405-1414.

[15] Wang, L., Dai, M., Li, Q., & Pang, L. (2020, September). Research and application of wide-band modeling method for flexible HVDC converter valve. In 2020 IEEE International Conference on High Voltage Engineering and Application (ICHVE) (pp. 1-4). IEEE.

[16] Hannan, M. A., Hussin, I., Ker, P. J., Hoque, M. M., Lipu, M. H., Hussain, A., ... & Blaabjerg, F. (2018). Advanced control strategies of VSC based HVDC transmission system: Issues and potential recommendations. Ieee Access, 6, 78352-78369.

[17] Guo, J., Liu, H., Feng, L., Zu, L., Ma, T., & Mu, X. (2024). A Deep Learning Based Fault Diagnosis Method for Flexible Converter Valve Equipment. IEEE Access.

[18] An, T., Tang, G., & Wang, W. (2017). Research and application on multi-terminal and DC grids based on VSC-HVDC technology in China. High voltage, 2(1), 1-10.

[19] Li, C., Li, K., Ma, C., Zhang, P., Tao, Q., Sun, Y., & Wang, X. (2021). Flexible control strategy for HVDC transmission system adapted to intermittent new energy delivery. Global Energy Interconnection, 4(4), 425-433.

[20] Liu, Q., Chen, P., Liu, S., Wang, C., & Liu, J. (2024). A Topology and Control Method for Operational Testing of±800kV/8GW Flexible DC Transmission Modular Multilevel Converter Valve. IEEE Access.

[21] Qi, L., Wang, M., Zhang, X., Shen, H., Wang, H., & Jiao, C. (2023). Analysis for magnetic field disturbance of modular multilevel converter based high voltage direct current (MMC-HVDC) converter valve. High Voltage, 8(1), 91-101.

[22] Barnes, M., Van Hertem, D., Teeuwsen, S. P., & Callavik, M. (2017). HVDC systems in smart grids. Proceedings of the IEEE, 105(11), 2082-2098.

[23] Cao Wen,Liang Song,Yu Huang,Dong Peng,Peng Zhang,Jianquan Liao... & Shilin Gao. (2024). Research on Thyristor Reverse Recovery Behavior in High-Voltage Direct Current Transmission Converter Valves and Its Application in Integrated Protection Systems. Energies(24),6472-6472.

[24] Na Qu,Wenlong Wei,Congqiang Hu,Shang Shi & Han Zhang. (2024). Series arc fault detection based on multi-domain depth feature association. Journal of Power Electronics(11),1809-1819.

[25] Wang Wei,Song Liangjun,You Quanwei & Zheng Xudong. (2023). Research on Energy Characteristics of Shaft Blasting Vibration Based on Wavelet Packet. Geotechnical and Geological Engineering(1),307-319.

[26] Sushma Jaiswal & Manoj Kumar Pandey. (2023). Linear Discriminate Analysis based Robust Watermarking in DWT and LWT Domain with PCA based Statistical Feature Reduction. International Journal of Image, Graphics and Signal Processing(IJIGSP)(2),73-88.

[27] Xiaoguang Yuan,Shiruo Liu,Wei Feng & Gabriel Dauphin. (2023). Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm. Remote Sensing(21).