XGBoost-Based Intelligent Decision Making of HVDC System with Knowledge Graph

Qiang Li [1], Qian Chen [2], Jiyang Wu [2], Youqiang Qiu [1], Changhong Zhang [3], Yilong Huang [3], Jianbao Guo [3] and Bo Yang [1,*]

1   EHV Power Transmission Company of China Southern Power Grid Co., Ltd., Dali Bureau, Dali 671000, China
2   EHV Power Transmission Company of China Southern Power Grid Co., Ltd., Guangzhou 510000, China
3   Maintenance and Test Center of CSG, EHV Power Transmission Company of China Southern Power Grid Co., Ltd., Guangzhou 510000, China
*   Correspondence: yangbo_ac@outlook.com

Abstract: This study aims to achieve intelligent decision making in HVDC systems in the framework of knowledge graphs (KGs). First, the whole life cycle KG of an HVDC system was established by combining intelligent decision making. Then, fault diagnosis was studied as a typical case study, and an intelligent decision-making method for HVDC systems based on XGBoost that significantly improved the speed, accuracy, and robustness of fault diagnosis was designed. It is noteworthy that the dataset used in this study was extracted in the framework of KGs, and the intelligent decision making of KG and HVDC systems was accordingly combined. Four kinds of fault data extracted from KGs were firstly preprocessed, and their features were simultaneously trained. Then, sensitive weights were set, and the pre-computed sample weights were put into the XGBoost model for training. Finally, the trained test set was substituted into the XGBoost classification model after training to obtain the classification results, and the recognition accuracy was calculated by means of a comparison with the standard labels. To further verify the effectiveness of the proposed method, back propagation (BP) neural network, probabilistic neural network (PNN), and classification tree were adopted for validation on the same fault dataset. The experimental results show that the XGBoost used in this paper could achieve accuracy of over 87% in multiple groups of tests, with recognition accuracy and robustness being higher than those of its competitors. Therefore, the method proposed in this paper can effectively identify and diagnose faults in HVDC systems under different operation conditions.

Keywords: knowledge graph; intelligent decision making; HVDC; fault diagnosis; XGBoost

1. Introduction

As the main part of the power system that undertakes power transmission tasks, high-voltage direct-current (HVDC) systems can retain the independence of power grids at both ends of transmission and reception, which has distinctive advantages over AC power transmission, such as no inductance and no synchronization. It is an important guarantee for high-capacity grid interconnection and large-scale power exchange [1–3]. To solve the problem of increasing the distance between power production and load centers [4–6], the HVDC system has taken on the main power transmission task in many transmission projects. Currently, seven ±800 kV HVDC transmission projects have been completed and put into operation in China, effectively contributing to the realization of West-East Power Transmission Project [7–9]. However, with the increase in voltage level and scale of HVDC systems, the stability of their fault occurrence and the safety of personal equipment are serious problems, so it is particularly important to carry out efficient and comprehensive fault diagnosis [10,11]. Currently, fault diagnosis methods applied to HVDC transmission systems mainly include analytical model-based, expert system, neural network-based [12], support vector machine (SVM) [13], ensemble learning (EM) [14], and K-nearest neighbor

(KNN) methods [15]. The analytical model-based approach requires a mathematical model based on the factual input–output relationships of the power system and is thus easily limited by the complexity of the power system in practical engineering. Neural network is a kind of strong learner with the advantages of high accuracy and robustness, but it involves a complex design, which often requires a lot of effort to strengthen the server computing capacity for its construction. SVM is a novel small-sample learning method that avoids dimensional catastrophes in the computation process and has good robustness. However, it is difficult to implement on large-scale training samples and solve multi-classification problems of the power system. EM uses the cooperation of multiple weak learners to achieve the effect of a strong learner, largely reducing the gray area of traditional single learners. However, it can overfit classification or regression problems with high noise, long iteration times, and high data processing costs.

In recent years, HVDC fault diagnosis has attracted widespread academic attention, and various fault diagnosis methods have emerged. Reference [16] took the 500 kV HVDC transmission system from Yunnan to Guizhou as an example to study the impact of lightning current peak and grounding resistance on the change in shield failure flashover. Reference [17] analyzed the impact of converter transformers on the single-phase grounding fault current on the grid side and revealed the mechanism of high short-circuit current in the nearby DC area. In order to ensure the safe and stable operation of the HVDC system, the lightning over-voltage at the neutral point of the high-voltage DC converter transformer was studied in work [18]. To improve the safety of the ±800 kV HVDC system, the grounding mode of the neutral point of the converter transformer was considered and analyzed in work [19]. Reference [20] used convolutional neural network (CNN) to identify internal and external faults of HVDC transmission lines, but with long computation time. In view of the sudden changes in current in HVDC transmission line faults, the Teager energy operator was adopted to form a feature vector based on the energy ratio of positive to negative pole current sudden change variables in faults, and 1D-CNN was applied to train and test the feature vector set, thus realizing the effective discrimination of fault types and fault poles inside and outside the area [21]. For the collection and processing of transient quantity information during faults, reference [22] carried out variational mode decomposition (VMD) on the transient current signal during faults to obtain the intrinsic mode function (IMF) component of the transient current signal and then calculated multi-scale fuzzy entropy using the intrinsic mode function component of the transient current signal; finally, multi-scale fuzzy entropy was input to the Softmax classifier for HVDC transmission line fault identification. When a fault occurs on an HVDC transmission line, there are a lot of characteristics that can be collected. Reference [23] used SVM to classify 13 fault characteristics, such as the AC/DC voltage and current of the HVDC transmission line in case of a ground fault, to realize fault identification. Reference [24] used AdaBoost SVM optimized with the bird swarm algorithm to identify HVDC transmission line faults after extracting fault characteristics from the DC-side voltage signal with the wavelet packet transform. Reference [25] designed a fault diagnosis method for HVDC transmission systems based on the improved gray wolf algorithm (IGWO) optimized time convolution neural network (TCN) to solve the problems of HVDC transmission line fault diagnosis and pole selection under the condition of high transition resistance, focusing on the shortcomings of low sensitivity and low identification accuracy of existing HVDC fault diagnosis methods [26]. However, the aforementioned methods have some distinct problems, such as weak robustness, high modeling cost, and slow diagnosis speed, due to complex models [27].

This study proposes a fault diagnosis model for HVDC transmission systems based on extreme gradient boosting (XGBoost). XGBoost is an integrated learning method based on gradient lifting proposed based on gradient boosting decision tree (GBDT) that also supports column sampling, which can greatly improve the efficiency of the algorithm and reduce overfitting. Moreover, this study uses back propagation (BP) neural network and probabilistic neural network (PNN) as comparison methods to diagnose HVDC system

faults. The simulation results show that the proposed method has high accuracy and reliability in fault diagnosis in HVDC systems.

In addition, the main contributions of this study are as follows:

- The dataset used with several classification methods is obtained from the Tianshengqiao HVDC transmission project in Guangzhou, China, which is of great practical engineering significance;
- XGBoost is applied to HVDC fault diagnosis for the first time;
- This research combines KGs with fault diagnosis to realize the visualization of HVDC fault processing.

Moreover, the rest of the article is arranged as follows: Section 2 shows the application of KGs in HVDC systems; Section 3 presents fault classification in HVDC systems; Section 4 is an introduction to the XGBoost algorithm; Section 5 is an introduction to the XGBoost algorithm; Section 6 presents a case analysis of fault diagnosis; finally, Section 7 presents the summary of the whole paper.

2. Knowledge Graph Platform in HVDC System

With the rapid development of artificial intelligence, knowledge graph (KG) technology has become one of the core driving technologies to promote the development of cognitive intelligence. At the same time, machine learning technology has been widely used [28].

This research aims to study abnormal signal identification and auxiliary decision making in HVDC systems based on state information. The research content is mainly divided into three parts: sequence-of-events recorder (SER) data abnormal signal identification module, SER data abnormal signal oriented fault identification module, and typical fault auxiliary decision-making module based on KGs. The data used in this study were all measured fault data from the HVDC transmission system of China Southern Power Grid. The actual HVDC system is named the Tianshengqiao (Guangxi Province, China)–Guangzhou (Guangdong Province, China) transmission project. The voltage level of the project is ±500 kV; the total length is 960 km; and the rated power is 1800 MW. Figure 1 shows the frame diagram of abnormal signal identification and auxiliary decision-making technology of the Tianshengqiao HVDC system, which is divided into three parts, namely, the SER data signal abnormal identification module, the SER data abnormal signal oriented fault identification module, and the typical fault assistant decision-making module based on KGs.

The distributed word vector representation of natural language words provides a new foundation for the in-depth application of different artificial intelligence methods in natural language processing. KG relational reasoning is an effective means to solve knowledge verification, prediction, and reasoning. Focusing on the problems of HVDC systems, i.e., the lack of massive data collection carriers and lack of intelligent means for fault anomaly analysis [29], this research aimed to build a technical framework for fault diagnosis in HVDC systems based on small-sample machine learning and multi-parameter fusion. The SER data and fault recording data were obtained by sending a request to the HVDC transmission system knowledge base; then, the obtained data were extracted from the key recording segments [30]. Finally, the processed fault data were input into the HVDC system risk analysis model for fault classification. In particular, Figure 2 shows the fault handling and risk analysis framework of the KG-based HVDC system. Due to the long route and high voltage level of the Guangzhou–Tianshengqiao HVDC system, fault diagnosis needs high accuracy and high security, and KGs can efficiently assist researchers in completing fault diagnosis, and especially, fault treatment can be quickly solved with KGs.
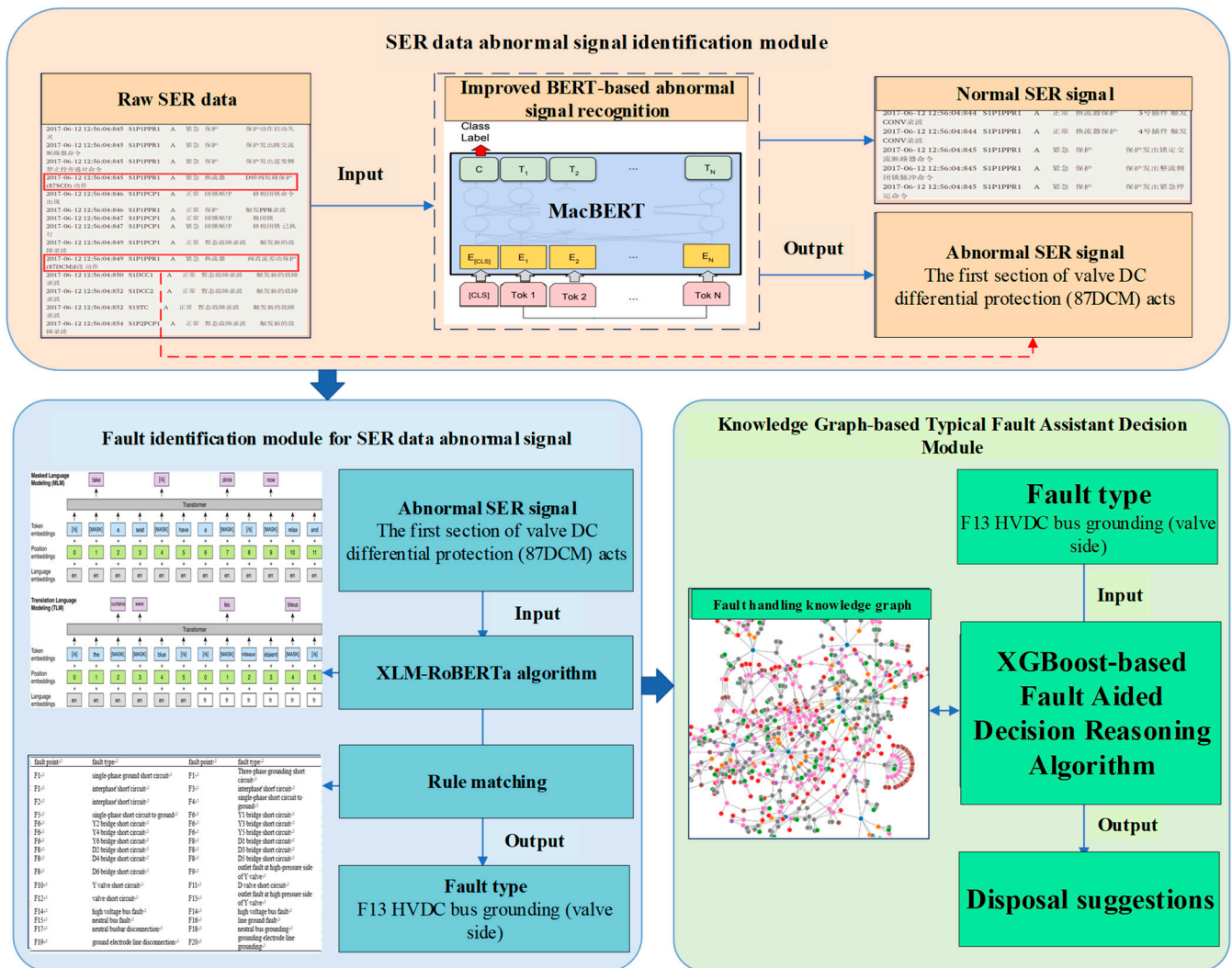
Figure 1. Flow chart of HVDC system abnormal signal identification and auxiliary decision making based on state information.
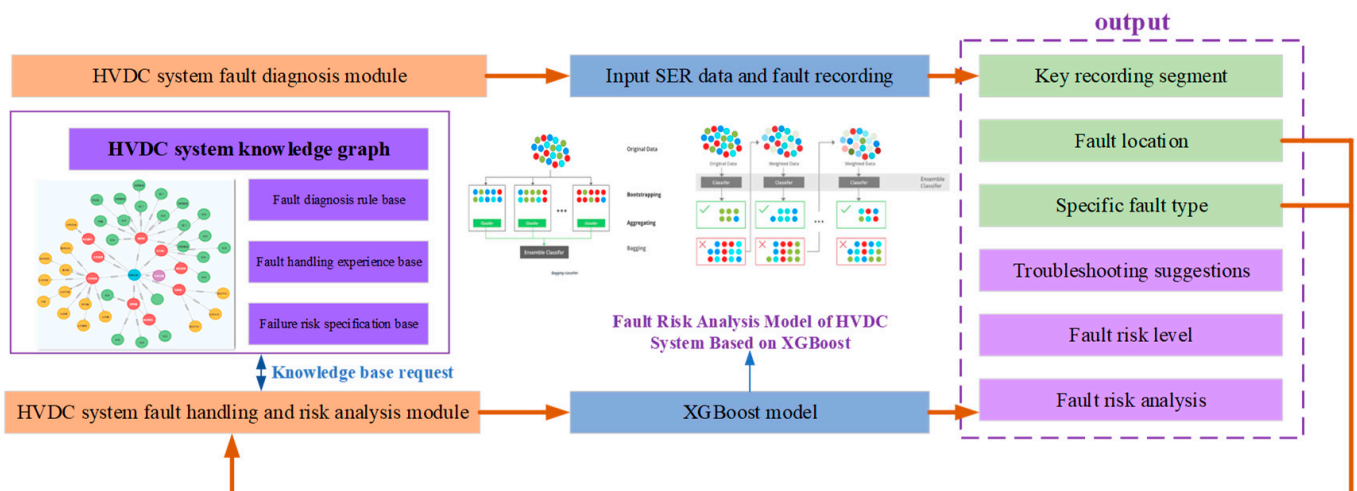

Figure 2. Flow chart of KG-based DC transmission system fault handling and risk analysis module.

3. Fault Classification of HVDC System

HVDC transmission systems are mainly composed of a converter station, a transmission line, and a grounding electrode system [31]. Among them, the converter station is one of the core components of HVDC transmission systems, has a complex structure, and often becomes a high-incidence area of faults [32]. HVDC transmission systems have many fault types, such as AC faults, DC faults, inverter commutation faults [33], converter valve faults [34], single-phase faults [35], interphase faults, and lightning stroke faults [36]. This study constructed a fault diagnosis model based on the XGBoost algorithm according to the measured data of four types of faults in a substation of a southwest power grid, and analyzed and diagnosed the four types of faults [37].

3.1. Grounding Fault on Converter Transformer Valve Side

Grounding faults on the valve side of the DC converter have spatiotemporal dispersion. A fault occurring at the high- and low-voltage bridges causes a huge difference in the working conditions of the valve bridge [38]. The relative positions of the fault point and the current transformer also lead to the difference in the measured current of the fault phase. The change in fault occurrence time causes a change in the fault characteristics. When single-phase grounding faults occur in HVDC systems, arc grounding occurs at the fault point, and over-voltage and resonant over-voltage form in the non-fault phase of the fault line bus. Compared with the normal distribution network operation, the over-voltage value becomes 1.732 times the original voltage level under the condition of complete grounding, or the resonance over-voltage formed exceeds the bearing range of the line and directly burns the line. The impact of single-phase grounding faults on distribution network lines is direct. If the line is in the state of voltage rise many times, it accelerates the aging of insulation weak links of line cables and equipment and causes short circuits due to the breakdown of insulation weak links. During the operation of the distribution network line with grounding, it is possible to cause relay-type short circuits and power failure due to the grounding fault. For over-voltage faults of small current grounding systems, there are difficulties in fault line selection, fault point location, and distance measurement. Researchers can solve the problem of reliable detection of small current grounding faults by studying the characteristics of single-phase grounding faults of medium- and low-voltage distribution networks, timely finding the grounding fault line, finding the fault point, and taking corresponding treatment measures.

3.2. Interphase Short-Circuit Fault on Converter Transformer Valve Side

Interphase short circuits refer to power supply short circuits caused by the connection between the end lines with no passing of the load. Interphase short circuits only have positive sequence current [39] and negative sequence current, and no zero sequence current. The device includes two-phase short circuits and three-phase short circuits. When interphase short circuits occur in HVDC systems, the harmonic component and its variation rule in the line are consistent with those of single-phase grounding faults, but the harmonic component content in interphase short circuits is higher, so the probability of 50 Hz protection maloperation during fault recovery is higher.

3.3. Short-Circuit Fault of Converter Valve Arm

As core equipment of HVDC systems, converters undertake the energy conversion function in AC/DC systems, that is, they convert AC electric energy into DC electric energy at the power transmission end of the system and then transmit it to the AC power grid at the receiving end to complete the energy transmission process (from the sending end to the receiving end). Bridge arm short-circuit faults of the converter valve are common faults of the converter valve. After such fault occurs, the AC system alternately has two-phase short circuits and three-phase short circuits; then, the AC system power cannot be transmitted to the receiving converter station through the DC line, and the receiving power grid cannot receive the DC power normally, which has a serious impact on the AC systems on both

sides. Valve arm short-circuit faults can be divided into AC-side area valve arm short-circuit fault and DC-side area valve arm short-circuit fault. AC-side valve arm short-circuit faults mainly refer to interphase short-circuit faults caused by the reduction in the interphase insulation performance of the converter valve side. DC-side valve arm short-circuit faults include single-bridge valve arm short-circuit faults, single-phase valve short-circuit faults, and pole bus and neutral bus short-circuit faults [40].

3.4. Fault of Converter Valve Group

Due to the nonlinearity of the converter, a large number of harmonics are generated in the HVDC system during operation, resulting in the distortion of the voltage and current of the transmission system, thus polluting the power. The core of the converter is the converter valve group, which is the key piece of equipment of the converter. Therefore, the analysis of the harmonic characteristics of converter valve group faults (valve false opening and valve non-opening faults) is of great significance to the safe and stable operation of HVDC transmission systems.

4. HVDC Fault Identification Based on XGBoost

In classification and recognition, to train an algorithm model with excellent recognition effect, it is usually possible to build the model with the aid of integration ideas. Boosting is a supervised classification learning method that combines weak separators to form strong classifiers, such as Adaboost and GBDT. Each submodel or subtree tries to enhance the overall effect of the model by constantly iterating and updating sample point weights. The boosting method has excellent classification and recognition performance when the dataset is not complex. However, when the dataset is complex, the model is constantly iterated so that the number of iterations increases, which directly leads to a sharp increase in the amount of computation. It not only slows down the training speed of the model but also affects the final classification and recognition effect of the model. This is the biggest disadvantage of the boosting algorithm.

Inspiringly, an XGBoost model based on the boosting integration idea and the C++ parallel construction of the regression tree is constructed, which is consistent with the GBDT idea. Each iteration is trained based on the residual of the weak classifier generated by the previous iteration. Multiple weak classifiers are trained with multiple iterations and then combined into an accurate and efficient integrated learner. Both are improved in the negative gradient direction of the loss function. However, XGBoost has higher accuracy, better generalization ability, and higher efficiency than GBDT.

4.1. Principle of XGBoost

The training objective of XGBoost is constant prediction to minimize the error between the predicted value and the real value. XGBoost integrates many CART classification regression trees generated by iterations, and the new tree generated by each iteration is based on the training and prediction of the tree generated by the previous iteration, that is, it is optimized according to the negative gradient direction of the loss function. Each iteration generates a tree to fit the prediction residual of the spanning tree of the last iteration and continues to iterate until the residual can no longer be reduced, thus improving the performance of the model.

The generation of the XGBoost tree depends on the addition model of the decision tree and forward distribution algorithm. The features are continuously split to generate a tree. Each time a tree is generated, it is equivalent to a new function. To fit the residuals of the last prediction, a new prediction value and new residuals are obtained. This training process is repeated. This is the forward-step algorithm. When N trees are obtained after training, the sample features have a node and a corresponding node predictive value in each tree. Finally, the final predictive value of the sample is the sum of all node values, which is the addition model. The specific steps are shown in Figure 3.
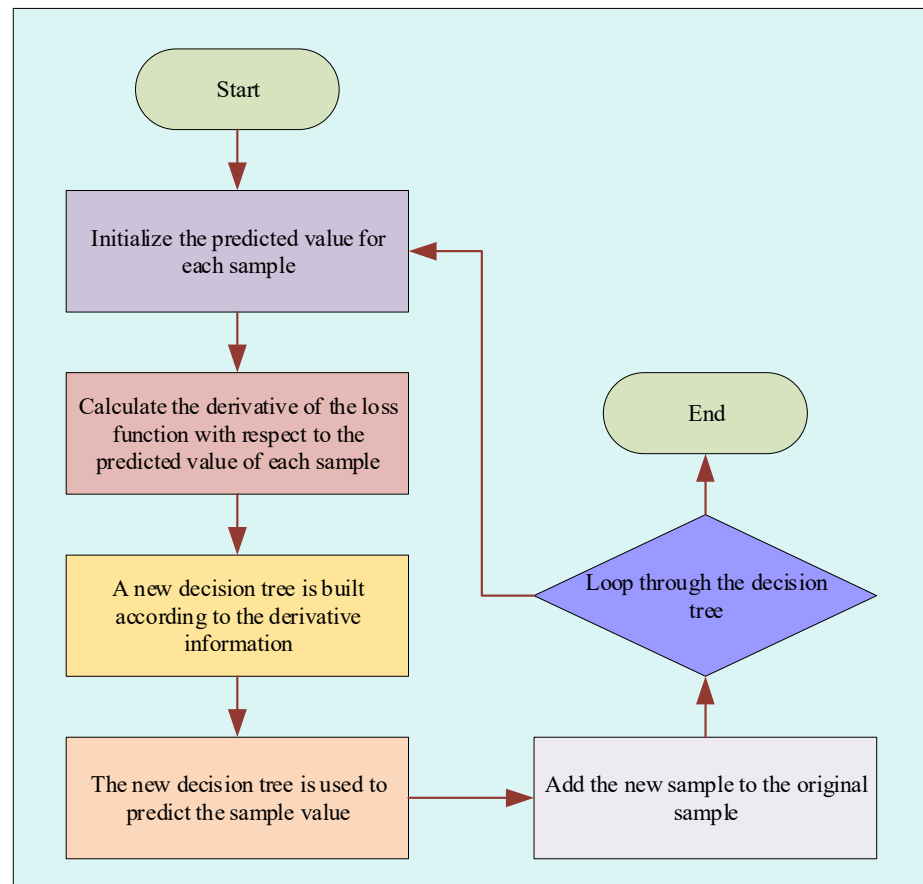
Figure 3. Overall flow chart of XGBoost.

XGBoost only uses the decision tree as the basic classifier, which essentially integrates multiple decision trees. Therefore, the model can be expressed as follows [41]:

$$f_N(X_i) = \sum_{n=1}^{N} T(X_i, \theta_n) \tag{1}$$

where $T(X_i, \theta_n)$ is the decision tree, X_i is the ith input sample, θ_n is the parameter of the corresponding decision tree, N is the number of decision trees, and $f_N(X_i)$ represents the predicted value of the model after the nth iteration.

The generation of the XGBoost tree needs to initialize the predictive value of each sample to be 0, that is, $f_0(X_i) = 0$; then, the model of the nth iteration is [42]

$$f_n(X_i) = f_{n-1}(X_i) + T(X_i, \theta_n) \tag{2}$$

The parameters can be obtained by minimizing the loss function of the algorithm; in particular, the solution formula is

$$\theta'_n = \arg\min \sum_{i=1}^{N} L(y_i, f_{n-1}(X_i) + T(X_i, \theta_n)) \tag{3}$$

The lifting tree model, $f_N(X_i)$, completed by the final iteration depends on the forward distribution algorithm and the addition model of XGBoost. The XGBoost model is obtained by adding the n class decision tree, $T(X_i, \theta_n)$, obtained by means of iteration.

4.2. Construction of Loss Function

For a dataset $D = \{(x_i, y_i)\}, x_i \in R^b, y_i \in R$ with a samples and b features, the final prediction output, $\hat{y}_i$, of M classification regression trees is [43]

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), f_m \in F \tag{4}$$

$$F = \left\{ f(x) = \omega_{q(x)} \right\}, q : R^b \to T, \omega \in R^T \tag{5}$$

where $\hat{y}_i$ is the final predictive value of the ith sample, which is obtained by summing up the scores of leaf nodes ω_i corresponding to each classification regression tree $f_m(x_i)$; T is the number of leaf nodes, that is, each tree f_m has a leaf tag q and a leaf weight ω corresponding to the current prediction sample; and $\omega_{q(x)}$ is the sum of the predicted scores of the weights of leaf nodes q corresponding to all classification regression trees, that is, the final predictive value of the XGBoost model for the sample.

To make the XGBoost model learn fully to achieve the best performance of classification and recognition, it is necessary to minimize the loss function of the XGBoost model, and at the same time, add regular items to prevent the model from being too complex. The loss function of the XGBoost model is [44]

$$L(\varnothing) = \sum_{i=1}^{a} l(\hat{y}_i - y_i) + \sum_{m=1}^{M} \Omega(f_m) \tag{6}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

where formula $L(\varnothing)$ is intended to calculate the loss function of the deviation between the predicted value of the sample and the true value, including deviation calculation term $l(\hat{y}_i - y_i)$ and regular term f_m to prevent overfitting; and γ and λ are used to control the regularization parameter of the model complexity. The larger the parameter value is, the more difficult the model is to overfit.

To build the final XGBoost model, one needs to calculate f_m of each tree. It is necessary to train the tth tree with the forward distribution algorithm. By setting the initial predictive value of the first tree to 0, that is, $f_0(x_i) = \hat{y}^{(0)} = 0$, the following model is obtained with t iterations [45]:

$$\begin{cases} \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \hat{y}_i^{(3)} = f_1(x_i) + f_2(x_i) + f_3(x_i) = \hat{y}_i^{(2)} + f_3(x_i) \\ \qquad\qquad\qquad \vdots \\ \hat{y}_i^{(t)} = \sum_{m=1}^{t} f_m(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{cases} \tag{8}$$

By summing the iteratively generated t trees with the addiction model, the objective function is

$$\mathrm{Obj}^t \approx \sum_{i=1}^{a} \left(l\left(y_i, \hat{y}_i^{(t-1)} \right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_i) + \mathrm{const} \tag{9}$$

The second-order Taylor expansion of each training objective function is obtained as follows:

$$\mathrm{Obj}^t = \sum_{i=1}^{a} \left(l\left(y_i, \hat{y}_i^{(t-1)} \right) + f_t(x_i) \right) + \Omega(f_i) + \mathrm{const} \tag{10}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)} \right), h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l\left(y_i, \hat{y}_i^{(t-1)} \right) \tag{11}$$

where, g_i and h_i are the first and second step degrees of the loss function, respectively. By removing the constant term, we can obtain the simplified objective function of the *t*th training as follows:

$$\text{Obj}^t \approx \sum_{i=1}^{a} \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_i) \tag{12}$$

5. Multi-Classification Fault Diagnosis Model Based on XGBoost

In this section, based on the measured fault data of a substation of a southwest power grid, the specific electrical diagram of the fault points and fault types of the transmission system is shown in Figure 4. In particular, Table 1 summarizes the fault types represented by the number of each fault point. From the original dataset, the recording data of 15 cycles before and after the fault were extracted, that is, the extraction duration of the recording data was 0.5 s. In the extraction of the recording data, 11 representative signal channels were sorted out. The specific signal meaning is described in Table 2. Among them, the elements in the data samples of single-phase ground faults, interphase faults, converter valve arm short-circuit faults, and converter valve group faults were $N_1 = 56$, $N_2 = 42$, $N_3 = 44$, and $N_4 = 96$, respectively. Based on the original dataset, the XGBoost algorithm was used for fault diagnosis, and the effectiveness of this method was verified. The algorithm implementation process is shown in Figure 5, and the meaning of the six classifiers therein is elaborated in Table 3; the first brace indicates the four labels of the XGBoost algorithm, and the second brace shows the specific classification of the six classifiers.

Table 1. Fault types of fault points of a substation in a southwest power grid in China.

| Fault Point	Fault Type	Fault Point	Fault Type
F1	Single-phase ground short circuit	F1	Three-phase grounding short circuit
F1	Interphase short circuit	F3	Interphase short circuit
F2	Interphase short circuit	F4	Single-phase short circuit to ground
F5	Single-phase short circuit to ground	F6	Y1-bridge short circuit
F6	Y2-bridge short circuit	F6	Y3-bridge short circuit
F6	Y4-bridge short circuit	F6	Y5-bridge short circuit
F6	Y6-bridge short circuit	F8	D1-bridge short circuit
F8	D2-bridge short circuit	F8	D3-bridge short circuit
F8	D4-bridge short circuit	F8	D5-bridge short circuit
F8	D6-bridge short circuit	F9	Outlet fault at high-pressure side of Y-valve
F10	Y-valve short circuit	F11	D-valve short-circuit
F12	Valve short circuit	F13	Outlet fault at high-pressure side of Y-valve
F14	High-voltage bus fault	F14	high-voltage Bus fault
F15	Neutral bus fault	F16	Line ground fault
F17	Neutral busbar disconnection	F18	Neutral bus grounding
F19	Ground electrode line disconnection	F20	Grounding electrode line grounding
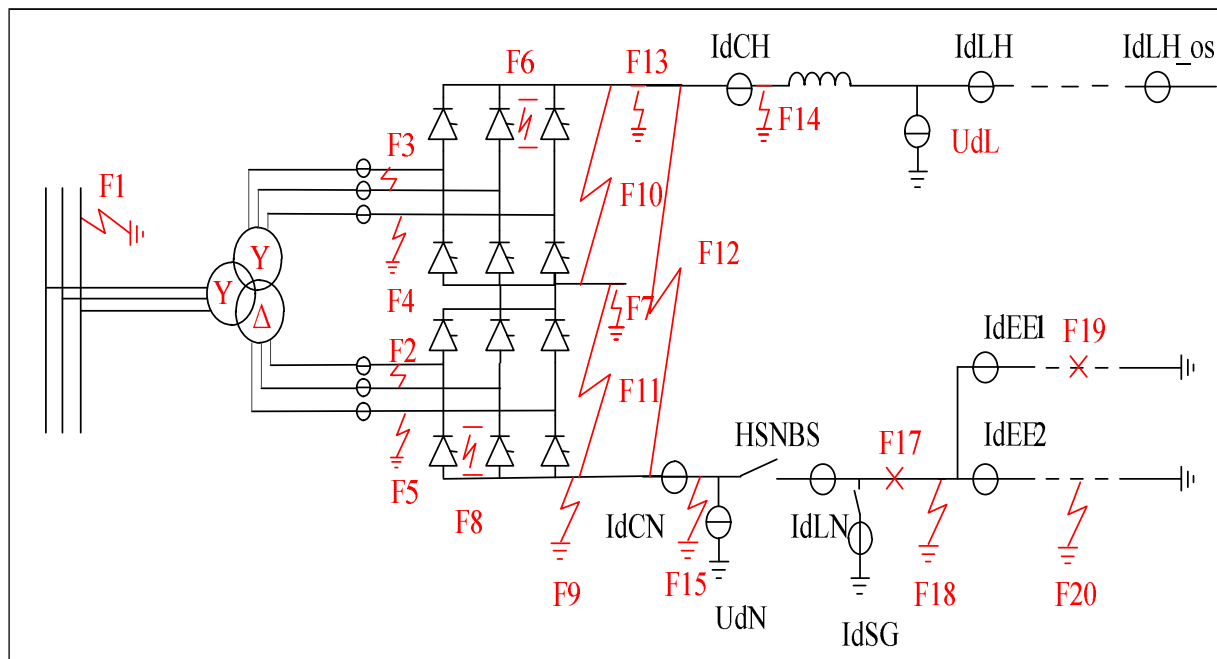
Figure 4. Main fault points of a substation in a southwest power grid in China.

Table 2. Signal names and meaning descriptions.

Signal Name	Signal Meaning Description	Signal Name	Signal Meaning Description
UAC_IN_L1(V)	A-phase AC voltage	IVD_L1(A)	AC on valve side A of bridge D
UAC_IN_L2(V)	B-phase AC voltage	IVD_L2(A)	AC on valve side B of bridge D
UAC_IN_L3(V)	C-phase AC voltage	IVD_L3(A)	AC on valve side B of bridge D
IVY_L1(A)	A-phase AC on Y-bridge valve side	UDL(V)	DC line voltage
IVY_L2(A)	B-phase AC on Y-bridge valve side	UDN(V)	Neutral bus voltage
IVY_L3(A)	C-phase AC on Y-bridge valve side		

Table 3. HVDC system fault classification modes and significance.

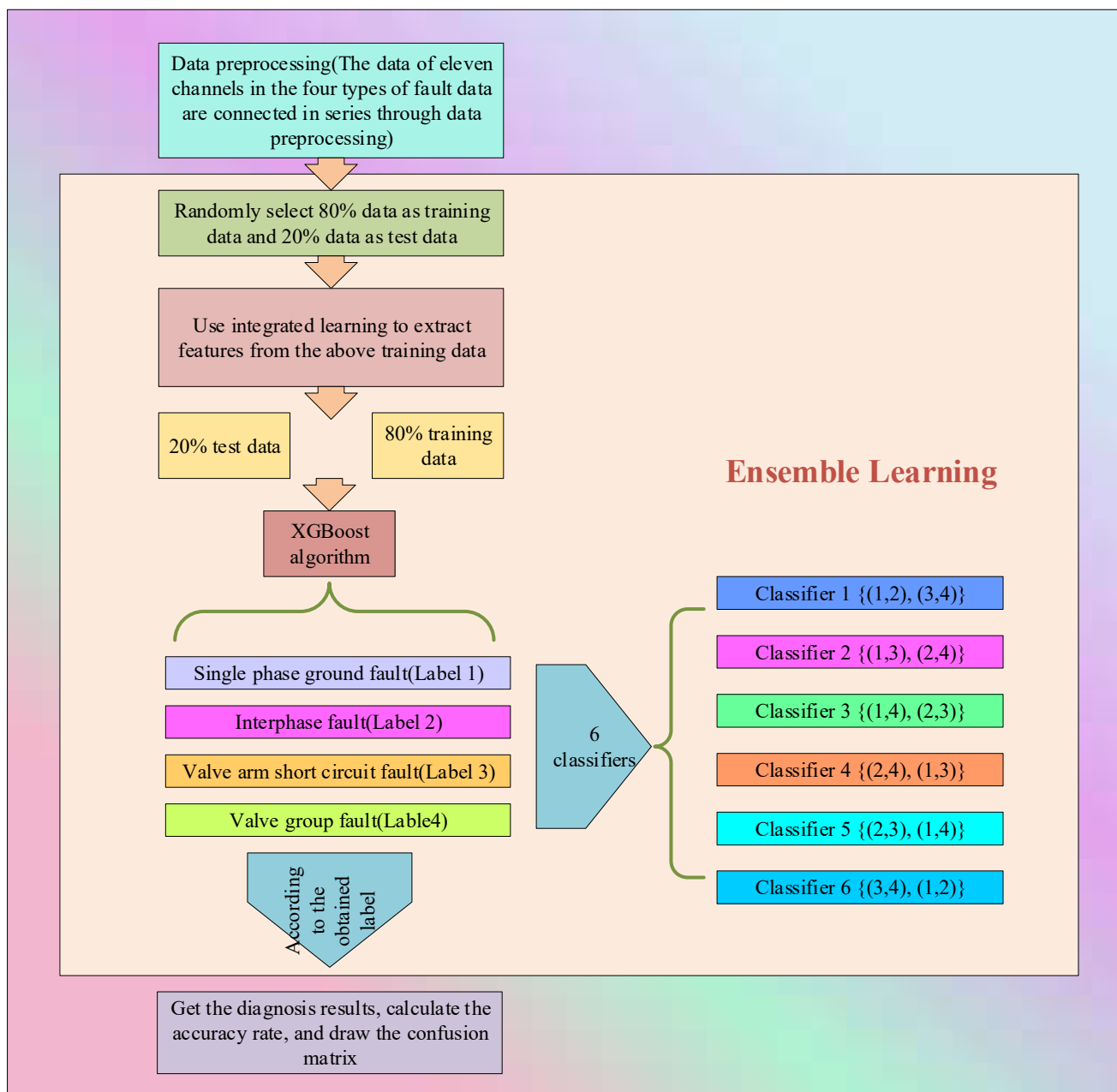| Classifier	Classification	Specific Significance
1	({1, 2}, {3, 4})	If the test data belong to a single-phase or interphase fault, and the output is 1; otherwise, the output is −1.
2	({1, 3}, {2, 4})	If the test data belong to an interphase or converter valve arm fault, the output is 1; otherwise, the output is −1.
3	({1, 4}, {2, 3})	If the test data belong to an interphase or converter valve group fault, output 1; otherwise, the output is −1.
4	({2, 3}, {1, 4})	If the test data belong to a single-phase or converter valve arm fault, the output is 1; otherwise, the output is −1.
5	({2, 4}, {1, 3})	If the test data belong to a single-phase or converter valve group fault, the output is 1; otherwise, the output is −1.
6	({3, 4}, {1, 2})	If the test data belong to the failure of converter valve arm or converter valve group, and the output is 1; otherwise, the output is −1.

Figure 5. Flow chart of HVDC fault classification based on XGBoost.

The specific steps are as follows: Firstly, 11 channel data of each sample in each type of fault data were connected in series to conduct data preprocessing and then stacked according to the number of samples to form a full fault dataset. Then, 80% of the total fault dataset was randomly selected as the training dataset, and 20% was selected as the test dataset. Secondly, integrated learning was used to extract the features of fault data, and 80% of the data were intensively trained. Multi-classification XGBoost was used to determine the number of classifiers and labels. According to the introduction of the multi-classification XGBoost method, the number of samples was four, so six classifiers were required. Among them, the labels of single-phase ground fault, interphase fault, converter arm short-circuit fault, and converter valve group fault were 1, 2, 3, and 4, The specific classification method and explanation are shown in Table 3. Figure 6 shows the data waveforms of 11 channels corresponding to the four HVDC faults.
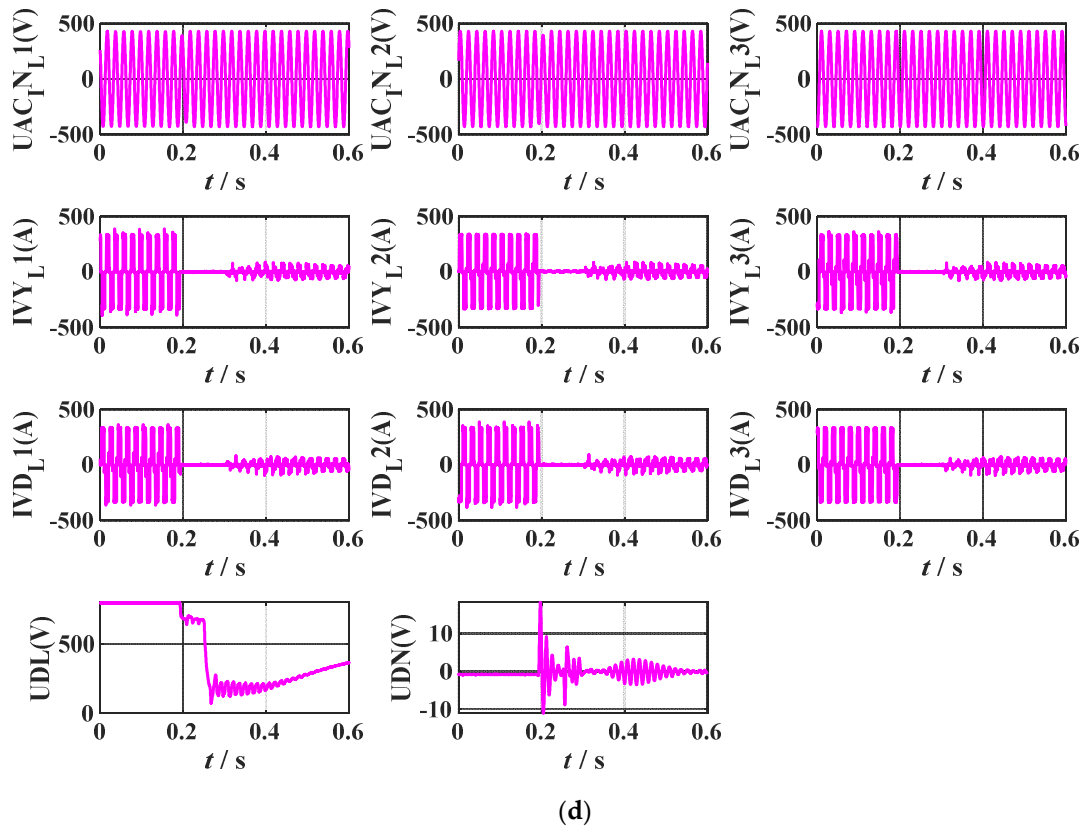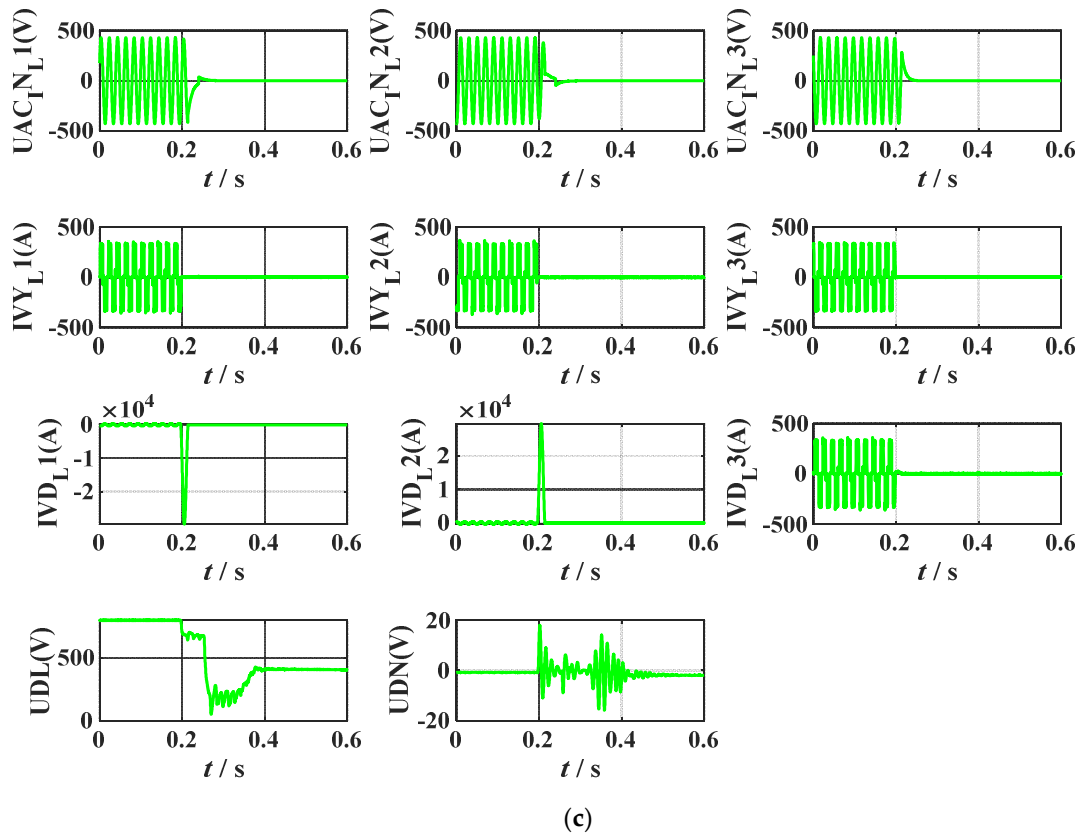
(**a**)

(**b**)

Figure 6. *Cont.*

(**c**)

(**d**)
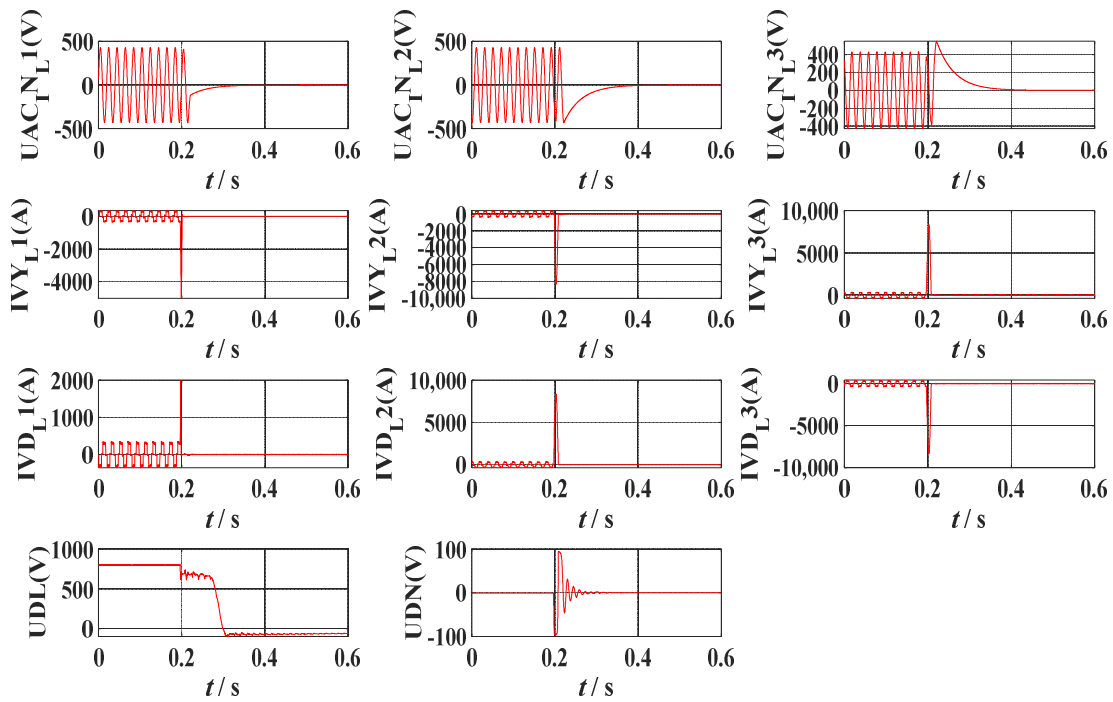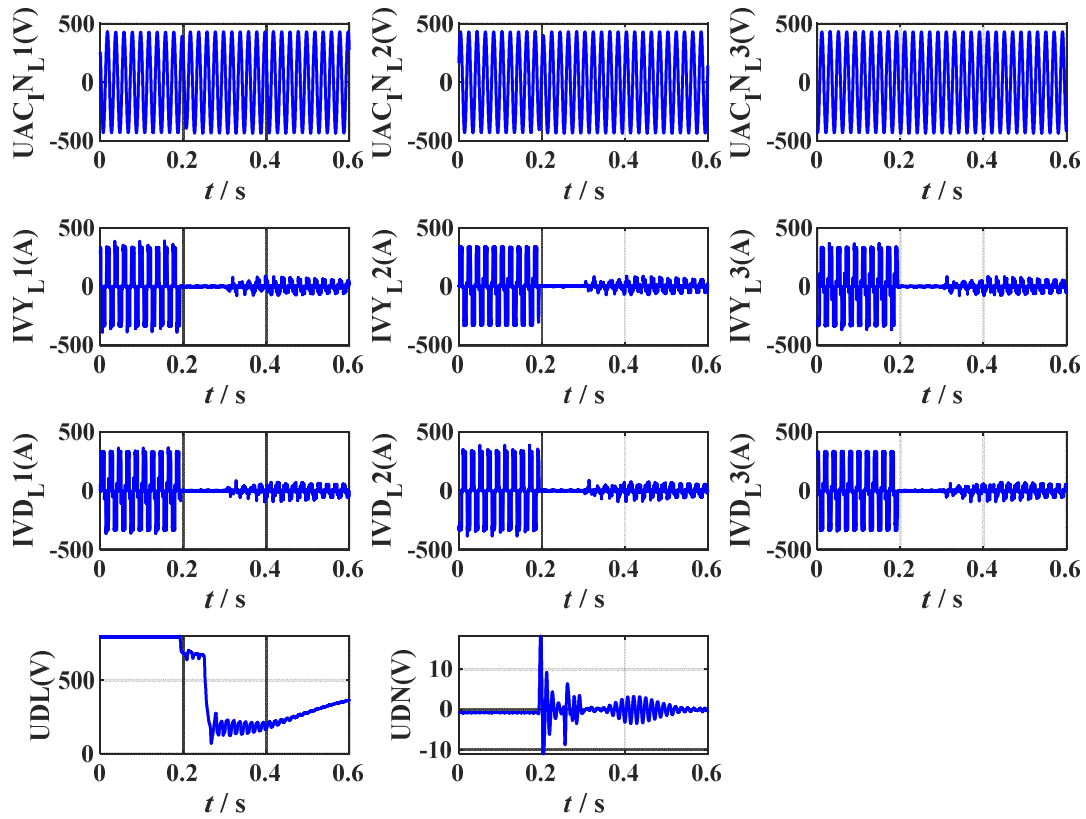
Figure 6. Diagnosis waveforms of four types of HVDC faults: (**a**) DC−side transmission line single-phase fault data, (**b**) DC−side transmission line interphase fault data, (**c**) fault data of converter valve arm, and (**d**) fault data of converter valve group.

As reported in the next section, after determining the number of data classifiers and training data, the remaining 20% of data were used as test samples for fault diagnosis and classification, and the test results were compared with the standard fault category threshold. In addition, the Euclidean distance between the test results and the standard fault threshold was used to determine the fault type. Finally, the accuracy of fault data diagnosis using this method was discussed.

6. Case Study

In this section, we report on the remaining 20% of all datasets being used as test data to verify the accuracy of XGBoost. Note that test datasets were randomly selected from all datasets. In particular, BP neural network and PNN neural network were used as comparison methods to verify the progressiveness and effectiveness of XGBoost. We input the test samples into XGBoost, BP neural network, and PNN, respectively. The test set data of XGBoost were $N_1 = 10$, $N_2 = 5$, $N_3 = 27$, and $N_4 = 5$; the test set data of BP neural network were $N_1 = 16$, $N_2 = 9$, $N_3 = 5$, and $N_4 = 17$; the test set data of PNN were $N_1 = 21$, $N_2 = 5$, $N_3 = 5$, and $N_4 = 16$; and the test set data of Classification learner were $N_1 = 14$, $N_2 = 11$, $N_3 = 9$, and $N_4 = 13$. The parameter settings of the six methods are shown in Table 4. We compared the accuracy of fault diagnosis of the six methods using the same number of test sets.

Table 4. Parameter settings of three methods.

Method	Parameter Name	Parameter Setting
XGBoost	gamma	0.4
	max_depth	8
	reg_lambda	2
	subsample	0.7
	colsample_bytree	0.7
	min_child_weight	3
	eta	0.1
	seed	1000
	N-thread	4
BP neural network	hidden neurons	30
PNN	spread	1.5
Classification learner	Max_split tree	100
SVM	C	1
	degree	3
KNN	radius	1
	leaf size	30

To intuitively observe the accuracy of fault diagnosis of the three methods, confusion matrices were drawn for data statistics and analysis. After the three methods trained their fault diagnosis models, the fault diagnosis results of XGBoost, BP neural network, PNN, Classification learner, SVM, and KNN were obtained and are shown in Figure 7a, Figure 7b, Figure 7c, Figure 7d, Figure 7e, and Figure 7f, respectively. It is not difficult to see that the three methods produced errors in the diagnosis results of the four types of faults. However, the diagnostic error rates of BP neural network, PNN, Classification learner, SVM, and KNN were significantly higher than that of XGBoost. In particular, XGBoost had 100% accuracy in identifying single-phase ground faults, converter arm faults, and converter valve group faults. The BP neural network only had 100% accuracy in identifying single-phase ground faults, while PNN and Classification learner showed different degrees of misdiagnosis of the four faults, which effectively shows that XGBoost could accurately and efficiently extract and identify the characteristics of fault data.

(**a**)

(**b**)

Figure 7. *Cont.*

	Single-phase ground fault (52.38%)	Interphase fault (62.5%)	Converter valve arm fault (55.56%)	Converter valve group fault (84.21%)
Single-phase ground fault (52.38%)	11 (23.4%)	0 (0%)	0 (0%)	0 (0%)
Interphase fault (62.5%)	3 (6.4%)	5 (10.64%)	0 (0%)	0 (0%)
Converter valve arm fault (55.56%)	4 (8.5%)	0 (0%)	5 (10.64%)	0 (0%)
Converter valve group fault (84.21%)	3 (6.4%)	0 (0%)	0 (0%)	16 (34.04%)

(c)

	Single-phase ground fault (66.67%)	Interphase fault (50%)	Converter valve arm fault (41.67%)	Converter valve group fault (69.23%)
Single-phase ground fault (66.67%)	10 (21.28%)	0 (0%)	1 (2.1%)	0 (0%)
Interphase fault (50%)	2 (4.3%)	10 (21.28%)	3 (6.4%)	4 (0%)
Converter valve arm fault (41.67%)	2 (4.3%)	1 (2.1%)	5 (10.64%)	0 (0%)
Converter valve group fault (69.23%)	0 (0%)	0 (0%)	0 (0%)	9 (19.15%)

Diagnostic label

(d)

Figure 7. *Cont.*

(**e**)

(**f**)

Figure 7. Test result confusion matrices of three methods: (**a**) XGBoost model test results, (**b**) BP neural network model test results, (**c**) PNN model test results, (**d**) Classification learner, (**e**) SVM, and (**f**) KNN.

Finally, according to the confusion matrix, the accuracy of the four methods in diagnosing the four types of faults in the HVDC system was obtained, as shown in Table 5. In addition, the five parameters of F1-score, precision score, recall score, AUC score, and test time obtained with the six methods are shown in Table 6. Note that all parameters are the average values obtained after cross validation, of which the fold is 5. It is not difficult to find that the accuracy of XGBoost model fault diagnosis in the full dataset was as high as 87.23%, while the fault diagnosis accuracy rates of BP neural network, PNN, Classification learner, SVM, and KNN were only 74.47%, 78.72%, 72.30%, 55.32%, and 65.96%, respectively, which fully proves the progressiveness of XGBoost in HVDC system fault diagnosis. All simulation experiments were run in the Python PyCharm Community Edition 2022 environment on a computer configured with 2.90 GHz Intel (R) Core (TM) i5-9400 CPU, 32.0 GB RAM, and 64-bit Windows 10.

Table 5. Diagnostic accuracy of six methods in four types of faults.

XGBoost		BP Neural Network		PNN	
Fault Type	**Diagnostic Accuracy**	**Fault Type**	**Diagnostic Accuracy**	**Fault Type**	**Diagnostic Accuracy**
Single-phase ground fault	100%	Single-phase ground fault	100%	Single-phase ground fault	52.38%
Interphase short-circuit fault	100%	Interphase short-circuit fault	33.3%	Interphase short-circuit fault	100%
Converter valve arm failure	81.48%	Converter valve arm failure	20%	Converter valve arm failure	100%
Converter valve group failure	100%	Converter valve group failure	88.24%	Converter valve group failure	100%
Total	87.23%	Total	74.47%	Total	78.72%
Classification Learner		**SVM**		**KNN**	
Fault Type	**Diagnostic Accuracy**	**Fault Type**	**Diagnostic Accuracy**	**Fault Type**	**Diagnostic Accuracy**
Single-phase ground fault	71.43%	Single-phase ground fault	53.85%	Single-phase ground fault	46.67%
Interphase short-circuit fault	90.91%	Interphase short-circuit fault	8.3%	Interphase short-circuit fault	33.3%
Converter valve arm failure	62.5%	Converter valve arm failure	53.13%	Converter valve arm failure	58.62%
Converter valve group failure	69.23%	Converter valve group failure	9.09%	Converter valve group failure	42.86%
Total	72.30%	Total	55.32%	Total	65.96%

Table 6. Comparison of results of six methods.

Method / Parameter	XGBoost	BP Neural Network	PNN	Classification Learner	SVM	KNN
F1-score	0.85	0.81	0.79	0.71	0.51	0.63
AUC score	0.91	0.8	0.72	0.70	0.53	0.65
Precision score	0.93	0.89	0.81	0.79	0.58	0.61
Recall score	0.81	0.75	0.69	0.71	0.56	0.66
Test time	72.03 s	84.75 s	114.13 s	363.33 s	76.95 s	97.56 s

7. Discussion

This study took the measured data of the Guangzhou–Tianshengqiao HVDC system as an example and classified and diagnosed four types of faults using XGBoost as the main algorithm. In addition, BP neural network, PNN, classification learning, SVM, and KNN were used as comparison algorithms to compare and verify the results of XGBoost. The results show that XGBoost had the shortest running time, only 72.03 s, and its accuracy rate was up to 87.23%, 8.51% higher than PNN, ranking second. In particular, all the results are the average values obtained after cross validation, with a fold of 5, which also verifies the robustness of the XGBoost algorithm. In addition, parameters such as the AUC are shown and compared in Table 6, i.e., an AUC score of XGBoost of 0.91, a precision score of 0.93, a recall score of 0.81, and an F1-score of 0.85, which are the best scores among those of the six methods.

8. Conclusions

This paper proposes a new HVDC system fault diagnosis model—the XGBoost-based fault diagnosis model—which can effectively extract the characteristics of various faults and accurately identify various faults in a transmission system. Its main contributions can be summarized as follows:

(1) Firstly, four classical faults of HVDC systems were analyzed, namely, single-phase grounding faults, interphase short-circuit faults, converter arm short-circuit faults, and converter valve group faults. Meanwhile, the data from 11 channels with representative characteristics within 15 cycles (0.5 s) of these four types of faults were collected and sorted out as the original sample dataset;

(2) The XGBoost HVDC system fault diagnosis strategy, which does not consider the distribution of data and can deal with the problem of linear inseparability, is proposed. Meanwhile, XGBoost can accurately extract the characteristics of each type of fault, which is suitable for the data classification of small samples. In addition, BP neural network and PNN were used as comparison methods to analyze the differences among the three methods in HVDC fault diagnosis results;

(3) The simulation results show that XGBoost can efficiently extract features of datasets and train a large number of datasets. The diagnosis results of the XGBoost, BP neural network, PNN, and Classification learner models are shown using three confusion matrices. The overall fault diagnosis accuracy rate of XGBoost was as high as 87.23%, significantly higher than those of BP neural network (74.47%), PNN (78.72%), Classification learner (72.30%), SVM (55.32%), and KNN (65.96%), respectively.

Fault diagnosis is an important research direction to ensure the safe and reliable operation of power systems. Whether it is important equipment such as transmission lines or substations, when a fault occurs, it is accompanied by a large amount of fault data. Thus, the prediction of the probability of future faults in the power system based on a large number of data when the fault occurs and the reduction in losses caused by faults are research focuses. Future fault diagnosis research should focus on efficient processing of fault data, promote the combination of fault diagnosis technology and artificial intelligence, and improve the efficiency and accuracy of fault diagnosis technology as much as possible.

Author Contributions: Conceptualization, Q.L.; methodology, Q.C.; software, J.W.; validation, Y.Q.; formal analysis, C.Z.; investigation, Y.H.; resources, J.G.; data curation, B.Y. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: All research data are confidential data of the enterprise. Please contact this email if necessary: 2410924502@qq.com.

Conflicts of Interest: Authors Qiang Li, Qian Chen, and Youqiang Qiu are employed by EHV Power Transmission Company of China Southern Power Grid Co., Ltd., Dali Bureau. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nomenclature

BP	Back propagation
CNN	Convolutional neural network
EM	Ensemble learning
GBDT	Gradient boosting decision tree
IGWO	Improved gray wolf algorithm
KG	Knowledge graph
KNN	K-nearest neighbor
HVDC	High-voltage direct-current

IMF	Intrinsic mode function
PNN	Probabilistic neural network
SER	Sequence-of-events recorder
SVM	Support vector machine
TCN	Time convolution neural network
VMD	Variational mode decomposition
XGBoost	Extreme gradient boosting

References

1. Yang, B.; Sang, Y.Y.; Shi, K.; Yao, W.; Jiang, L.; Yu, T. Design and real-time implementation of perturbation observer based sliding-mode control for VSC-HVDC systems. *Control Eng. Pract.* **2016**, *56*, 13–26. [CrossRef]
2. Guan, Z. Real type test and research of $\pm$ 800kV UHVDC transmission tower. *Shanxi Electr. Power* **2022**, *7*, 19–22.
3. Yang, B.; Liu, B.; Zhou, H.; Wang, J.; Yao, W.; Wu, S.; Shu, H.; Ren, Y. A critical survey of technologies of large offshore wind farm integration: Summarization, advances, and perspectives. *Prot. Control Mod. Power Syst.* **2022**, *7*, 17. [CrossRef]
4. Yang, B.; Li, Y.; Li, J.; Shu, H.; Zhao, X.; Ren, Y.; Li, Q. Comprehensive summarization of solid oxide fuel cell: Control: A state-of-the-art review. *Prot. Control Mod. Power Syst.* **2022**, *7*, 36. [CrossRef]
5. Chen, Y.; Yang, B.; Guo, Z.; Wang, J.; Zhu, M.; Li, Z.; Yu, T. Dynamic reconfiguration for TEG systems under heterogeneous temperature distribution via adaptive coordinated seeker. *Prot. Control Mod. Power Syst.* **2022**, *7*, 38. [CrossRef]
6. Yang, B.; Wu, S.; Li, Q.; Yan, Y.; Li, D.; Luo, E.; Zeng, C.; Chen, Y.; Guo, Z.; Shu, H.; et al. Jellyfish search algorithm based optimal thermoelectric generation array reconfiguration under non-uniform temperature distribution condition. *Renew. Energy* **2023**, *204*, 197–217. [CrossRef]
7. Yang, B.; Wu, S.; Zhang, H.; Liu, B.; Shu, H.; Shan, J.; Ren, Y.; Yao, W. Wave energy converter array layout optimization: A critical and comprehensive overview. *Renew. Sustain. Energy Rev.* **2022**, *167*, 112668. [CrossRef]
8. Yang, B.; Jiang, L.; Yao, W.; Wu, Q.H. Perturbation observer based adaptive passive control for damping improvement of multi-terminal voltage source converter-based high voltage direct current systems. *Trans. Inst. Meas. Control* **2017**, *39*, 1409–1420. [CrossRef]
9. Yang, B.; Yu, T.; Zhang, X.; Huang, L.; Shu, H.; Jiang, L. Interactive teaching-learning optimizer for parameter tuning of VSC-HVDC systems with offshore wind farm integration. *IET Gener. Transm. Distrib.* **2018**, *12*, 678–687. [CrossRef]
10. Liu, S.; Zhao, L.; Li, J.; Hou, Z. Dynamic fault recovery method of a photovoltaic distribution network considering switch state set adjustment. *Power Syst. Prot. Control* **2021**, *49*, 24–31.
11. Yang, B.; Jiang, L.; Yu, T.; Shu, H.C.; Zhang, C.K.; Yao, W.; Wu, Q.H. Passive control design for multi-terminal VSC-HVDC systems via energy shaping. *Int. J. Electr. Power Energy Syst.* **2018**, *98*, 496–508. [CrossRef]
12. Qi, J.; Yang, C.; Zhao, H. Fault location of HVDC transmission lines based on neural network. *Electr. Power Sci. Eng.* **2014**, *30*, 45–49.
13. Li, F.; Wang, X.; Zhang, C.; Zhou, M. Boundary point based support vector machine classification algorithm. *J. Shaanxi Univ. Technol. (Nat. Sci. Ed.)* **2022**, *38*, 30–38.
14. Chen, L.; Wu, H.; Li, D.; Yang, Y. HVDC transmission line lightning fault identification method based on integrated learning. *J. Electr. Power Syst. Autom.* **2022**, *34*, 102–110.
15. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
16. Zalhaf, A.S.; Zhao, E.; Han, Y.; Yang, P.; Almaliki, A.H.; Aly, R.M. Evaluation of the transient overvoltages of HVDC transmission lines caused by lightning strikes. *Energies* **2022**, *15*, 1452. [CrossRef]
17. Zhao, E.; Han, Y.; Liu, Y.; Zalhaf, A.S.; Wang, C.; Yang, P. Feasibility analysis of neutral grounding by small reactor of HVDC converter transformer. *Energy Rep.* **2022**, *8*, 392–399. [CrossRef]
18. Hu, Z.; Han, Y.; Yang, P.; Wang, C.; Zalhaf, A.S. Evaluation of lightning overvoltage at neutral point of HVDC converter transformer based on EMTP. *Energy Rep.* **2022**, *8*, 274–283. [CrossRef]
19. Fu, J.; Han, Y.; Yang, P.; Wang, C.; Zalhaf, A.S. Analysis of fault current and overvoltage at the neutral point of $\pm$800 kV High-Voltage DC converter transformer. *Energy Rep.* **2022**, *8*, 292–300. [CrossRef]
20. Wei, D.; Gong, Q.; Lai, W. Research on fault diagnosis and fault phase selection of transmission line inside and outside the area based on convolutional neural network. *Chin. J. Electr. Eng.* **2016**, *36* (Suppl. S1), 21–28.
21. Wang, Q.; Wu, H.; Yang, J.; Li, D.; Liu, Y. HVDC transmission line fault identification method based on Teager energy operator and 1DCNN. *Smart Power* **2021**, *49*, 93–100.
22. Wang, Q.; Wu, H.; Hu, X.; Gu, X.; Chen, J. HVDC transmission line fault identification method based on VMD multi-scale fuzzy entropy. *J. Electr. Power Syst. Autom.* **2021**, *33*, 134–144.
23. Hu, W.; Shen, Y.; Liu, Y. Fault identification of MMCHVDC system based on improved support vector machine. *Smart Power* **2019**, *47*, 91–97.
24. Zheng, X.; Peng, P. Fault diagnosis of flexible DC transmission converter based on optimized wavelet packet and AdaBoost SVM. *J. Electr. Power Syst. Autom.* **2019**, *31*, 42–49.

25. Liu, H.; Li, Y.; Zhang, M.; Liu, W. HVDC transmission line fault diagnosis based on GWO-TCN network. *Electron. Meas. Technol.* **2021**, *44*, 168–174.

26. Muniappan, M. A comprehensive review of DC fault protection methods in HVDC transmission systems. *Prot. Control Mod. Power Syst.* **2021**, *6*, 1. [CrossRef]

27. Mitra, B.; Chowdhury, B.; Manjrekar, M. HVDC transmission for access to off-shore renewable energy: A review of technology and fault detection techniques. *IET Renew. Power Gener.* **2018**, *12*, 1563–1571. [CrossRef]

28. Wu, J.; Li, Q.; Chen, Q.; Peng, G.; Wang, J.; Fu, Q.; Yang, B. Evaluation, Analysis and Diagnosis for HVDC Transmission System Faults via Knowledge Graph under New Energy Systems Construction: A Critical Review. *Energies* **2022**, *15*, 8031. [CrossRef]

29. Chen, X.; Jia, S.; Xiang, Y. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.* **2020**, *141*, 112948. [CrossRef]

30. Zeng, X.; Tu, X.; Liu, Y.; Fu, X.; Su, Y. Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* **2022**, *72*, 114–126. [CrossRef] [PubMed]

31. Zamini, M.; Reza, H.; Rabiei, M. A Review of Knowledge Graph Completion. *Information* **2022**, *13*, 396. [CrossRef]

32. Yao, Z.; Zhang, Q.; Chen, P.; Zhao, Q. Research on fault diagnosis for MMC-HVDC systems. *Prot. Control Mod. Power Syst.* **2016**, *1*, 8. [CrossRef]

33. Nanayakkara, O.; Rajapakse, A.D.; Wachal, R. Traveling-wave-based line fault location in star-connected multiterminal HVDC systems. *IEEE Trans. Power Deliv.* **2012**, *27*, 2286–2294. [CrossRef]

34. Narendra, K.G.; Sood, V.K.; Khorasani, K.; Patel, R. Application of a radial basis function (RBF) neural network for fault diagnosis in a HVDC system. *IEEE Trans. Power Syst.* **1998**, *13*, 177–183. [CrossRef]

35. Vidal-Albalate, R.; Beltran, H.; Rolán, A.; Belenguer, E.; Peña, R.; Blasco-Gimenez, R. Analysis of the performance of MMC under fault conditions in HVDC-based offshore wind farms. *IEEE Trans. Power Deliv.* **2015**, *31*, 839–847. [CrossRef]

36. Pauli, B.; Mauthe, G.; Ruoss, E.; Ecklin, G.; Porter, J.; Vithayathil, J. Development of a high current HVDC circuit breaker with fast fault clearing capability. *IEEE Trans. Power Deliv.* **1988**, *3*, 2072–2080. [CrossRef]

37. Livani, H.; Evrenosoglu, C.Y. A single-ended fault location method for segmented HVDC transmission line. *Electr. Power Syst. Res.* **2014**, *107*, 190–198. [CrossRef]

38. Liu, J.; Tai, N.; Fan, C.; Chen, S. A hybrid current-limiting circuit for DC line fault in multiterminal VSC-HVDC system. *IEEE Trans. Ind. Electron.* **2017**, *64*, 5595–5607. [CrossRef]

39. Yeap, Y.M.; Geddada, N.; Ukil, A. Analysis and validation of wavelet transform based DC fault detection in HVDC system. *Appl. Soft Comput.* **2017**, *61*, 17–29. [CrossRef]

40. Tang, G.; Xu, Z.; Zhou, Y. Impacts of three MMC-HVDC configurations on AC system stability under DC line faults. *IEEE Trans. Power Syst.* **2014**, *29*, 3030–3040. [CrossRef]

41. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T. *Xgboost: Extreme Gradient Boosting*; R Package Version 0.4-2; Xgboost Developers: Beijing, China, 2015; Volume 1, pp. 1–4.

42. Ogunleye, A.; Wang, Q.G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *17*, 2131–2140. [CrossRef] [PubMed]

43. Brownlee, J. Xgboost with python. In *Machine Learning Mastery*; Jason Brownlee: Austin, TX, USA, 2019.

44. Ramraj, S.; Uzir, N.; Sunil, R.; Banerjee, S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *Int. J. Control Theory Appl.* **2016**, *9*, 651–662.

45. Zhang, D.; Qian, L.; Mao, B.; Huang, C.; Huang, B.; Si, Y. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access* **2018**, *6*, 21020–21031. [CrossRef]