Research on Attack Identification Method and Device Method Based on Random Forest Algorithm

Yang Yu1,2, Shengyuan Xiao3, Rui Chen3, Yu Du4, Kai Ding3, Xiaofan Zhu3, Guoqi Zhou3, Shengguo He3, Hongyan He3, Zhiwei Huang3, Xiaoquan Gao3, Hao Yao1,2, Wei Xi1,2, Qing Xian3, Yang Fu3 and Qing Ding4

1Electric Power Research Institute, China Southern Power Grid. Guangzhou 51000, China
2Digital Grid Research Institute, China Southern Power Grid. Guangzhou 51000, China
3CYG SUNRI CO., LTD, Shenzhen 518000, China
4Shenzhen Power Supply Bureau Co., Ltd, China Southern Power Grid. Shenzhen 518000, China

Corresponding author e-mail:xiaosy@sznari.com

Abstract. Because the energy storage system has the function of stabilizing the voltage and frequency of the grid, when the energy storage system is connected to the grid, it is necessary to obtain its operating status information in a timely and accurate manner. The operating status data of the energy storage system is uploaded to the higher-level dispatch center via the energy storage coordination control device. However, with the intelligence and informationization of the power grid, various cyber attack methods have emerged endlessly. There are various security risks in the way of information interaction based on the network. Attackers can analyze the communication protocols within them, or implement Eavesdropping attacks, Dos attacks, and tampering with sensitive data, etc., causing the energy storage coordination control device to obtain the wrong data sent by the dispatch center. The wrong data causes the energy storage system to output the wrong power, which may have the opposite effect when adjusting the voltage and frequency of the power grid, resulting in great economic losses, and even causing casualties due to equipment failure. Attackers can also gradually invade the upper-level dispatch center by using the energy storage coordination control device as a springboard. In view of the above problems, this paper uses a random forest-based network attack detection module is used to make real-time judgments on all data flows of the energy storage coordination control device. That is, when the energy storage coordination control device is subjected to a network attack, all data flows of the energy storage coordination control device pass a trained random forest attack detection model, thereby implementing attack detection on real-time data flows.

1. Introduction

Facing the energy crisis, financial crisis, and climate crisis, people recognize the importance of new energy development. Countries' investment in new energy has increased significantly, and the capacity of new energy has also expanded dramatically. Renewable energy power generation is the core of new energy development. However, due to the impact of seasonal, meteorological, and regional conditions, wind, solar, and marine energy generation has obvious discontinuities and instabilities. The generated power fluctuates greatly, the adjustability is poor, and the generation and consumption are sometimes delayed. When there is too much new energy generation capacity connected to the grid, the stability of...
the grid will also be affected. By supporting a large-scale energy storage system, the time difference between generation and consumption can be resolved, and intermittent renewable energy generation can be directly connected to the grid to impact the power grid to adjust the power quality[1-3].

Because the energy storage system also has the function of stabilizing the voltage and frequency of the grid, when the energy storage system is connected to the grid, it must obtain its operating status information in a timely and accurate manner. The operating status data of the energy storage system is uploaded to the higher-level dispatch center via the energy storage coordination control device. However, with the intelligence and informatization of the power grid, various cyber attack methods have emerged endlessly. There are various security risks in the way of information interaction on the network. Attackers can analyze the communication protocols within them, or implement Eavesdropping attacks, Dos attacks, and tampering with sensitive data, etc. Causing the energy storage coordination control device to obtain the wrong dispatch center to issue data. The wrong data leads to the wrong output power of the energy storage system, which may have the opposite effect when adjusting the voltage and frequency of the power grid, causing an imbalance in the voltage and frequency of the power grid, resulting in great economic losses, and even caused by the abnormal operation of the equipment. attackers can also gradually invade the upper-level dispatch center by using the energy storage coordination control device as a springboard[4-9].

Potential safety hazards of existing energy storage coordination control devices. Under normal circumstances, one or more security vulnerabilities will give the attacker a chance. The attacker can use the security breach as a breakthrough point, by creating a large amount of useless data or sending repeated requests to occupy the network resources of the attacked person or interfering with the normal communication of the attacked. In the energy storage coordination control device, hackers can easily launch various attacks on them based on known security holes. For example, an attacker can create a large amount of useless data, causing network congestion of the energy storage coordination control device, making it unable to communicate with the superiors and subordinates normally; the attacker can use the energy storage coordination control device transmission protocol to handle the defects of repeated connections, repeatedly The high-frequency sending of repeated connection and repeated requests makes the energy storage coordination control device unable to process other normal requests in a timely manner; an attacker can inject a Trojan into the energy storage coordination control device and use it as a springboard to gradually invade the upper-level dispatch center; the attacker can also root the energy storage coordination control device transmission protocol defects and repeatedly send malformed attack data, such as tampering with the output power of the energy storage system, causing grid voltage oscillations or the dispatch center to incorrectly allocate a large amount of system resources, which directly affects the safe and stable operation of the grid.

A new network attack detection module is added to the energy storage coordination controller. Its principle is to detect the data flow in the energy storage coordination controller based on a random forest algorithm to determine whether it has been attacked and what kind of network attack. The network attack detection model is first trained using data after various network attacks, so that the model can meet the required performance requirements.

2. Grid-Side Energy Storage System Model
Figure 1 shows a conventional energy storage coordination control device and system structure diagram. It can be seen from the figure that the energy storage coordination control device is a communication bridge that connects the upper-level dispatch center and the lower-level energy storage power station monitoring system and other intelligent devices. It collects the PCS working status, PCS charge and discharge power, SOC value, etc., and receives it. Relevant remote signaling data from the lower-level controller, as well as forward-relevant remote adjustment, start-stop commands to the lower-level controller; upper-level real-time transmission of the input and output of the energy storage system, active and reactive power, power consumption, and voltage, current, power and other data of the grid connection point.
Figure 1. Structure diagram of energy storage system

The energy storage coordination control device herein includes: a communication module, a CPU module, an HMI LCD screen, a network attack detection module, a measurement module, a log module, a time synchronization module, a printing module, and a wave recording module. The CPU module communicates with the communication module. Module, HMI LCD screen, network attack detection module, measurement module, log module, time synchronization module, printing module, and wave recording module are connected; as shown in Figure 2.

Figure 2. Structure block diagram of energy storage coordination control device

The communication module is used for communication with the data center of the dispatching center master system and PCS, and sends the data flow to the dispatching center; and sends the received data flow to the control unit; the data flow includes grid data and control instructions;

Among them, the grid data includes voltage, current, frequency, power factor, harmonics, three-phase voltage imbalance, voltage fluctuations and flicker, temporary overvoltage and transient overvoltage at the common connection point, and the uploaded data of PCS to output active power.
And reactive power, battery SOC, maximum chargeable power, maximum dischargeable power, rated power, working status;

The grid data also includes PT secondary rating, CT secondary rating, current and voltage change dead zone, frequency change dead zone, time synchronization method, device address, A \ B network IP address and subnet mask, and remote power Power, remotely set reactive power, system reactance value, grid-connected voltage value, reactive power adjustment compensation, input of hard pressure plate and opening and closing state of the signal;

The CPU module is used to receive the data stream sent by the communication unit, the measurement power source and the filtering unit and send it to the network attack detection module for real-time detection and classification, and determine whether there is an attack behavior in the data stream based on the detection and classification results; When the data has an attack behavior, an alarm is displayed on the HMI LCD screen to display the type of attack behavior and generate a log record; when the grid data in the detection result does not exist, the normal data flow is sent to the main center of the dispatch center through the communication module System; At the same time, the CPU module also sends log records to the log module for storage or sends attack alarms and log records to the dispatching center master station system through the communication module.

The measurement module is used to obtain the data stream in the energy storage power station system and upload it to the CPU module; the wave recorder module is used to send the data stream of the fault record and waveform to the CPU module; the print module is used for printing; the log module is used for the log Records, data streams and alarms are stored; the timing module is used for device timing; HMI LCD screen is used for display;

The network attack detection module is used to detect the data stream sent by the CPU module in real time through a random forest model, and output the detection classification result to the CPU module.

3. Network Attack Detection Model
This model is a network attack detection model based on the energy storage coordination controller. Figure 3 is a flowchart of network attack detection.

![Flow chart of network attack detection](image)

**Figure 3.** Flow chart of network attack detection
The network attack detection model based on random forest is obtained by the following methods:

1. Using the data stream as training samples, establish N training sample sets and M features, where N is the number of training sample sets and M is the number of features, and construct at least one tree to train the training sample set; from The training sample set is randomly sampled from the training sample set. The 2N / 3 training samples in the randomly selected training sample set are used as the training set (each sample will be put back into a sample). The remaining training samples in the set are taken as out-of-bag samples (Out of bag test set), and then m features are randomly selected from the training set as the basis for the branching of this tree, where m < M. The way of selecting features builds a tree, as shown in Figure 4[10-14].

2. Determine the feature values. By judging the importance of m features in a single tree and calculating the importance of m features in all trees, then sort the importance of m features and remove some features with low importance. To get a new feature set, as shown in Figure 5;

3. Update and iterate the random forest model, and find the optimal random forest model in the random forest model as the final random forest model[15-18].

Among them, the determined characteristic values include:

1. Determine whether a feature plays a role in this tree or whether the feature is invalid in this tree by randomly changing the attribute value (attribute value of the feature) of a feature;
2. Compare the error rate of the test set before and after the change, and take the difference of the
error rate of the test set as the importance of the feature in the tree. Calculate once for \( m \) features in a
tree to obtain \( m \) The importance of the feature in the tree;

The error rate is obtained using formula (1):

\[
\text{OOBerr}_i = \frac{FP + FN}{TP + TN + FP + FN}
\]  

(1)

Among them, \( i \) refers to the number of features, TP: the number of positive examples correctly
predicted, TN: the number of positive examples correctly predicted, FP: the number of positive
examples predicted as negative, FN: the number of positive examples predicted as negative
The number of positive examples is yes and negative examples are not.

3. Calculate the importance of each feature in the forest, take the mean value of the importance
of this feature in multiple trees as the importance of the feature in the forest, and use the formula (2) to
obtain the feature value:

\[
\text{MDA}(A_i) = \frac{1}{n_{\text{sum}}} \sum_{t=1}^{n_{\text{sum}}} (\text{OOBerr}_{A_i} - \text{OOBerr}_t)
\]  

(2)

Among them, MDA represents Mean decrease accuracy; A in \( A_i \) represents a feature, and \( i \)
represents the number of the feature: \( n_{\text{sum}} \) represents the number of times feature \( A_i \) appears in
the forest, and OOBerr\(_{A_i}\) represents the value of the \( A_i \) attribute in the \( t \)-th tree Out-of-bag error rate after
change, OOBerr\(_t\) represents the out-of-bag error rate of normal \( A_i \) value in the \( t \)-th tree.

The error rate outside the bag is obtained by the formula (3):

\[
\text{OOBerr}_i = \frac{FP + FN}{TP + TN + FP + FN}
\]  

(3)

Among them, \( i \) refers to the number of features, TP: the number of positive examples correctly
predicted, TN: the number of positive examples correctly predicted, FP: the number of positive
examples predicted as negative, FN: the number of positive examples predicted as negative
quantity[19-23].

4. After obtaining the importance of all the features in the forest, sort all the features according to
the importance, remove some of the features with low importance in the forest, and get a new feature
set to complete an iteration.

Determining feature values also includes: repeating 1-4, gradually removing relatively poor
features, each time a new random forest model is generated, until the number of remaining features is
\( m \), and then the best among these random forest models is found The random forest model is used as
the final random forest model for real-time detection of data streams. Figure 6 is a specific flowchart
for determining the characteristic value.
Randomly change the attributes of a feature

Compare test set error rates before and after change

Calculate the importance of each feature in the forest

Sort all features in order of importance, remove some features of low importance in the forest, and complete an iteration

Until the number of remaining features is m, find the optimal random forest model as the final random forest model

Figure 6. Specific flowchart for determining characteristic values

4. Conclusions
Compared with the prior art, this paper uses all data streams in the energy storage coordination control device as training samples and uses a random forest model as a detection engine to detect and classify the input data streams. When the energy storage coordination control device is under attack, issue an alarm, isolate suspicious data while generating log records, thereby improving the information security of the energy storage coordination control device.

5. Acknowledgments
This work was supported by the National Key R&D Program of China (2018YFB0904900, 2018YFB0904903).

6. References
[1] Yao D L, Choi S S, Tseng K J, et al. A Statistical Approach to the Design of a Dispatchable Wind Power-Battery Energy Storage System. IEEE Transactions on Energy Conversion, 2009, 24(4):916-925.
[2] Diaz-González, Francisco, Sumper, Andreas, Gomis-Bellmunt, Oriol, et al. Energy management of flywheel-based energy storage device for wind power smoothing. Applied Energy, 110(COMPLETE):207-219.
[3] N.S. Wade, P.C. Taylor, P.D. Lang, et al. Evaluating the benefits of an electrical energy storage system in a future smart grid. Energy Policy, 38(11):7180-7188.
[4] Jajodia S, Noel S, O’Berry B. Keynote 3: Topological Analysis of Network Attack Vulnerability// 2009.
[5] Hug, Gabriela, Giampapa, Joseph Andrew. Vulnerability Assessment of AC State Estimation With Respect to False Data Injection Cyber-Attacks. IEEE Transactions on Smart Grid, 3(3):1362-1370.
[6] Dawkins J, Hale J. A systematic approach to multi-stage network attack analysis. // 2004.
[7] Dong Wei, Yan Lu, Jafari, M, et al. Protecting Smart Grid Automation Systems Against Cyberattacks. *IEEE Transactions on Smart Grid*, 2(4):782-795.
[8] Andre Teixeira, Henrik Sandberg, K.H. Johansson. Networked Control Systems under Cyber Attacks with Applications to Power Networks. // *American Control Conference* (ACC), 2010.
[9] Gao Z. Tracing Cyber Attacks from the Practical Perspective. 2005, 43(5):123-131.
[10] Petit, Jonathan, Shladover, Steven E. Potential Cyberattacks on Automated Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):1-11.
[11] Mariana Belgiu, Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *Isprs Journal of Photogrammetry & Remote Sensing*, 2016, 114:24-31.
[12] Dong Wei, Yan Lu, Mohsen Jafari, et al. An integrated security system of protecting Smart Grid against cyber attacks. // *Innovative Smart Grid Technologies (ISGT)*, 2010. IEEE, 2010.
[13] Deng, Houtao, Runger, George. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483-3489.
[14] K. Hartmann, C. Steup. The vulnerability of UAVs to cyber attacks - An approach to the risk assessment. // Cyber Conflict (CyCon), 2013 5th International Conference on. IEEE, 2013.
[15] Teixeira A, Sandberg H, Johansson K H. Networked Control Systems under Cyber Attacks with Applications to Power Networks. 2010, 58(8):3690-3696.
[16] Piero Bonissone, José M. Cadenas, M. Carmen Garrido, et al. A fuzzy random forest. *International Journal of Approximate Reasoning*, 51(7):729-747.
[17] Pal, M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217-222.
[18] Claudia Lindner, Paul A. Bromiley, Mircea C. Ionita, et al. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9):1862-1874.
[19] Ham, J, Yangchi Chen, Crawford, M.M, et al. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience & Remote Sensing*, 43(3):492-501.
[20] Lindner C, Bromiley P A, Ionita M C, et al. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. // *Computer Vision – ECCV* 2012. 2012.
[21] Lindner, C, Thiagarajah, S, Wilkinson, J. M, et al. Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting. *IEEE Transactions on Medical Imaging*, 2013.
[22] Khosgoftaar T M, Golawala M, Hulse J V. An Empirical Study of Learning from Imbalanced Data Using Random Forest. // 2008.