Research Article

Jinkui He* and Weibin Su

Establishment of nonlinear network security situational awareness model based on random forest under the background of big data

https://doi.org/10.1515/nleng-2022-0265
received March 26, 2022; accepted December 3, 2022

Abstract: In order to explore the establishment of a nonlinear network security situational awareness model based on random forest in the context of big data, a multi-level network security knowledge system evaluation model based on random forest is proposed. This article proposes a multi-level CSSA analysis system and then uses random memory algorithm to create a CSSA evaluation model. Also, it proposes a CSSA multi-level analysis framework and then uses random forest algorithm to build a CSSA evaluation model. A random vector distribution of the same values is used for all forest trees. In this article, the interval [0,1] is used to quantitatively describe the weight of the security level. The training sample ratio of test samples is 110:40, in order to predict the security of the network, the prediction of knowledge is closer to the true value, and the complexity of multi-level security is predicted. Use unusual forests. The tree returns the most recommended part, which is a more realistic assessment of network security. The experimental results show that considering the network security situation, the prediction performance of this method is closer to the actual value, and the performance is better than the other two methods. Therefore, perception of multi-level security situations can be effectively predicted using random access memory. It is proved that random forest is faster and more efficient in network security.

Keywords: network security situational awareness, multi level CSSA, random forest

*** Corresponding author: Jinkui He,** School of Intelligent Science and Engineering, Yunnan Technology and Business University, Kunming 650217, China, e-mail: hejinkui2@163.com
Weibin Su: Edge Computing and Network Center, Yunnan Technology and Business University, Kunming 650217, China, e-mail: suweibin2@126.com

1 Introduction

Following the rapid development of information technology, applications, human societies, people's lives, clouds, and big data, the Internet of Things is evolving, and the structure, scale, data, and application of the network are becoming more and more complex. Network security has increasingly become the focus of attention [1]. Key technology research on the knowledge of network security situations in big data environments can help to effectively solve the problem of decomposing the nature of huge and rapidly changing network security data from a scientific theoretical point of view and can be useful for quantitative research on network security assessment. It helps with network attack randomness. It can analyze and filter network security situations in the face of uncertainty and uncertainty, measure network security perceptions in big data environments, and solve scientific problems of integrated decision-making [2]. A key technology study of knowledge about network security situations in big data environments provides data acquisition and data processing methods for network security monitoring. Studying the characteristics of network security data, especially network attacks, vulnerabilities and threats, harmful programs, *etc*., studying the mechanism of characteristics and the relationship between them, and mastering the laws of network users and group structure security behavior are of reference significance for the study of network security active defense [3,4]. Some studies combine small filters with convolutional neural networks and use nonlinear sliding method and N-gram language model to construct a classification model for processing short texts, which improves the classification effect of short texts. These one-dimensional data processing methods do not effectively take advantage of the spatial advantages of convolutional neural networks and do not take advantage of the weight sharing and local perception advantages of convolutional neural networks in two-dimensional matrices.

As the number of network terminals increases, the amount of network data generated increases, which puts forward higher requirements for the understanding and assessment of the network security situation, as well as the predicted response time and data processing speed. Scholars' research in the field of network security situational awareness is mostly based on traditional machine learning algorithms. In the case of multiple data features and large data samples, the processing speed of traditional machine learning algorithms is relatively weak, and most machine learning algorithms rely on artificial intelligence. Feature selection is not conducive to efficient response to the network. Therefore, it is particularly important to select a situational awareness method that can efficiently handle large sample sizes and multi-attribute features without relying on human factors.

On the basis of this research, this article proposes a nonlinear network security situation awareness model based on random forest under the background of big data. The CSSA evaluation model is constructed by using the random forest algorithm. The random vector distribution of all trees in the forest is the same value, and the CSSA evaluation model is constructed. The experiment shows that the model and evaluation method in this article are feasible, superior to other existing methods in prediction accuracy, time and other performance, and can provide options to evaluate network security.

2 Literature review

Over the past few years, the speed of the terminal gathering information is increasing, linked to cloud computing, Internet, and stuff technologies that have evolved rapidly. An early age of data and a rapid increase in data can be challenges and opportunities for many fields [5]. Cybersecurity knowledge is the ability to identify, understand, and anticipate the elements that are relevant to the network environment within a given space and time. A model understanding of the network security situation is shown in Figure 1.

Data aggregation is the core of a network security situation knowledge model because it is the core of network security situation knowledge. This is a data aggregation model [6]. Ahmed *et al.* offer a more general framework based on this, combined with the real environment of the network. As a result, we offer the CyberSA model, which includes configuration, logging, attack, and other elements that fully reflect the characteristics of the network [7]. Rongrong *et al.* studied and summarized the latest measurement models and methods for establishing and enhancing situational awareness in aviation environment and put forward the future research direction of situational awareness evaluation method [8]. Pokharel proposed a prediction framework that can infer the driver's maneuver intention for assisted driving and automatic driving, predict the future travel path of traffic participants by solving and reinforcing optimal control problems, and determine the parameters of the optimal management formula [9]. Zeadally *et al.* used blockchain technology to support multiple scalable architectures of the Internet of things to improve smart city security and situational awareness according to the objectives and needs of emerging smart city applications. The neural network-based network security situational knowledge model mainly uses the advantages of neural networks

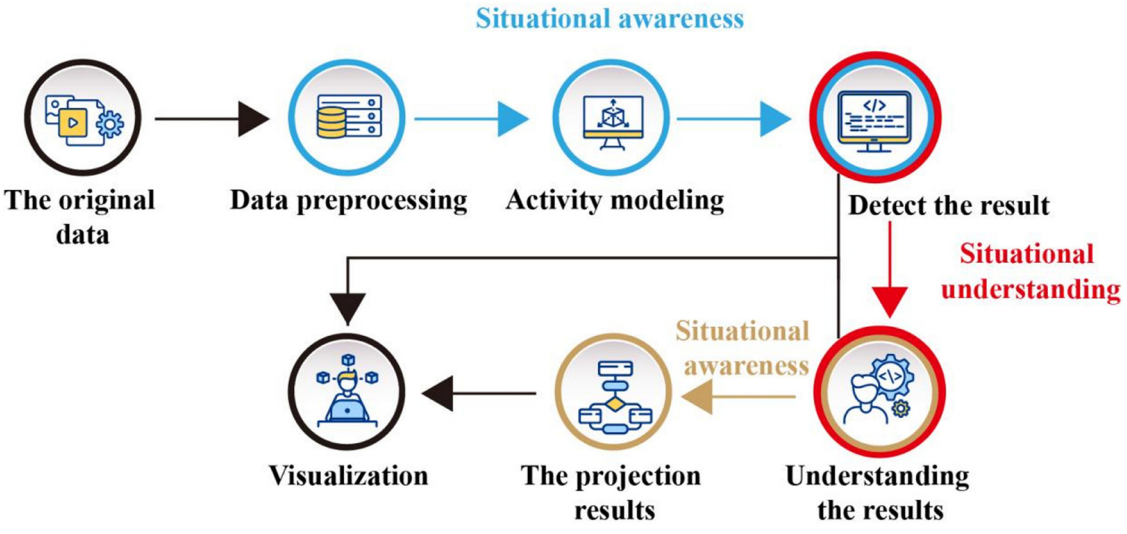

Figure 1: Network security situation awareness model.

to process high-complexity network environments and nonlinear data to create a network security situation awareness model [10]. Kumar *et al.* used deep neural networks to study network security situation assessments and network security assumptions and proposed the optimal blurred network security assessment model, and the improved long-term short-term memory network-based situation prediction model [11]. Hang *et al.* combined deep learning with power grid security situational awareness and proposed a power grid security situational awareness model based on deep learning [12]. Chen and Miao proposed an improved LSTM network security situational awareness framework based on cross entropy and linear rectification function [13]. Kstle *et al.* proposed a security situational awareness evaluation and prediction model of network security events based on machine learning and deep learning [14]. Kstle *et al.* offered, for big data, three models of network security knowledge. Each template contains three modules: In order to achieve different functions, the key technologies for recognizing the network situation include method of extracting features, classification of network attacks, detection methods, methods of assessing network security, and methods of predicting network security [15]. Aiming at the traditional network security situation awareness technology, in recent years, Russell *et al.* studied the problem that the high-dimensional characteristics of intrusion data will reduce the performance of intrusion detection, proposed a KPCA (PKPCA) feature extraction method based on particle swarm optimization, and verified the effectiveness of this method [16]. Peng *et al.* studied apt attack feature detection technology under the Salda model [17]. Gomes *et al.* studied a hospital network illegal intrusion detection model based on the combination of an ant colony optimization algorithm and support vector machine, and experiments show that it has good application value [18]. Antoniou *et al.* proposed a k-nearest neighbor high-speed matching algorithm, and the experimental results show that the improved k-nearest neighbor high-speed matching algorithm has high efficiency in processing high-dimensional and massive sample data [19]. Based on the results of a study on the knowledge of the network security situation, a method for evaluating a non-linear multi-level network security knowledge model based on a random forest is proposed. Compared to the Bayesian network and the BP neural network, the random memory algorithm is more accurate and efficient. In addition, for unbalanced data, our method can balance errors without creating too many adjustments.

3 Research methods

Big data environment refers to the network environment for collecting, storing, analyzing, calculating, sharing, and using big data. It belongs to a huge nonlinear complex system. In a large data environment, a large amount of network security data is combined into a large network security data. These data are of different types and large amounts. They access and create network security data such as status log data, business security data, abnormal traffic data, intrusion data, vulnerability data, and property data. (i) Data Diversity: There are many types of network security data in a large data environment, and the data attributes are diverse, including category attributes, sequence attributes, and association attributes. the data formats are also different. (ii) Dynamic complexity: the network security status of big data environment changes complexly with time, and the network security data are obtained in real-time. (iii) Complex data structure: the data structure is complex and changeable, showing a variety of different characteristics, and the characteristic attributes change over time. (iv) Multi-repetition and heterozygosity fusion: the characteristics of network security data affect and interact with each other, as well as noise interference [20,21].

3.1 Multi-level CSSA based on random forest

3.1.1 Multi-level CSSA model

The CSSA process basically corresponds to the life cycle of security data. In the life cycle, the data take different forms, starting from the original sensor data, involving data cleaning, data fusion, and event perception, and ending with a scenario. The upstream security data life cycle mainly involves data preprocessing, distributed data storage, data aggregation, and event processing, while the downstream lifecycle of security data includes situational assessment, modeling, sequential modeling and model analysis, baseline, management, and situational imaging. In order to extract high levels of value from security data, CSSA conducts a multi-layered analysis process. Figure 2 shows the information flow of the CSSA multi-level analysis system. The flow of information from security sensors to configuration files forms a valuable information network to implement CSSA. At the bottom, sensors receive operational, configuration, and

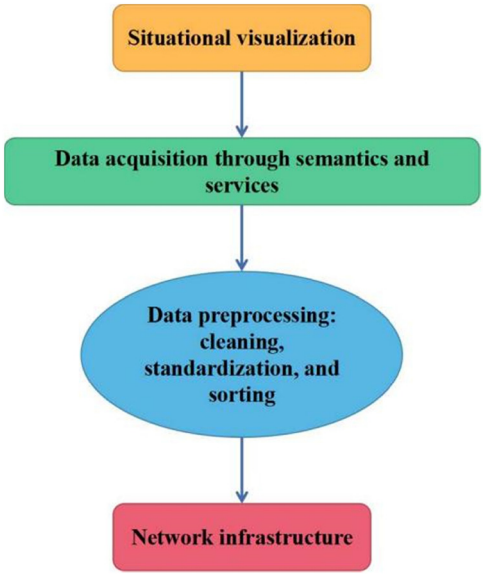

Figure 2: Multi-level CSSA analysis framework.

topology information from facilities that contain system infrastructure and information assets. Sensor data must be cleaned and normalized as it enters the distributed data store. Stored in popular databases are key documents on the history and current state of network infrastructure. It is difficult to combine data from different sensors into different formats to obtain information. From network traffic logs to the use of statistics and topology maps, converting data from disparate data sources into a common representation format at the appropriate syntax level is difficult and expensive. A more practical approach should be to move to semantic or service-level data integration, such as data communication, data virtualization, and the use of data-as-a-service. These generic descriptive elements connect to a distributed database. After processing the data, the main process of CSSA is to assess and forecast scenarios, generate an understanding and representation of the current situation, and predict how the situation will develop in the near future [22].

3.1.2 Random forest nonlinear assessment model

Random forest is the theory of learning statistics. The unimaginable forest is a solution tree. The schemes of the tree are often mismanaged and mismanaged. When grouping a tree, the idea is to use the "Merge" method, and the main idea is to recreate the bootstrap using the first step to design the bootstrap tops. Many nonlinear decision trees can then be combined to make predictions, and the final hypothesis can be made by voting [23]. All decision tree classifiers are defined as $h(X,\theta_k)$, and the classification result of each decision ternary model is defined as shown in formula (1):

$$H(x) = \max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y). \tag{1}$$

The advantages of the random forest model are convergence and upper bound of generalized error. For group classifiers, the margin function is used to measure whether the average number of correct classifiers exceeds the average number of incorrect classifiers. The higher the margin value, the more reliable the classification assumption. The margin value function is shown in formula (2):

$$\begin{aligned} \mathrm{mg}(X, Y) = {} & av_k I(h(X, \theta_k) = Y) \\ & - \max_{j \neq Y} av_k I(h(X, \theta_k) = j). \end{aligned} \tag{2}$$

Errors in general conclusions should be determined according to the characteristics of ordinary votes, as shown in formula (3):

$$\mathrm{PE}^* = \mathrm{PX}, P(\mathrm{mg}(X, Y) < 0). \tag{3}$$

As the number of decision trees in random forest increases, all sequences θ 1,... θ_k. PE* converges almost everywhere, and the convergence of generalization error is shown in formula (4):

$$\lim_{k \to \infty} \mathrm{PE}^* = P_{X,Y}\left(\begin{array}{c} P_\theta(h(X, \theta_k) = Y) \\ - \max\limits_{j \neq Y} P_\theta(h(X, \theta_k) = j) < 0 \end{array} \right). \tag{4}$$

This indicates that the model is not suitable for augmenting decision trees. At the peak of the general error, Chebyshev's inequality is shown in formula (5):

$$\mathrm{PE}^* \leq \frac{\mathrm{var}_{X,Y}(\mathrm{mg}(X, Y))}{E_{X,Y}\mathrm{mg}(X, Y)^2}. \tag{5}$$

According to Eq. (5), we define the classification strength of a single decision tree, as shown in formula (6).

$$s = E_{X,Y}\mathrm{mg}(X, Y). \tag{6}$$

Then, a function of the general upper error limit, as shown in formula (7):

$$\mathrm{PE}^* \leq \frac{\rho(1 - s^2)}{s^2}. \tag{7}$$

4 Result analysis

4.1 Analysis

This article selects the detection data from the 140–20 week report from 2018 to 2021, including data from the

Table 1: Network security situation awareness level

Level	Threatening	Vulnerability	Disaster tolerance	Stability
Security	Low	Low	High	High
Mild hazard	Medium	Low/medium	High/medium	High
General hazard	High/medium	High/medium	Medium	Medium
Moderate risk	High	Medium/high	Medium/low	Low
Highly dangerous	High	High	Low	Low

Table 2: Corresponding table of network security situation

Level	Threat index	Vulnerability index	Disaster recovery index	Stability index
Low	[0,0.4)	[0,0.3)	[0,0.4)	[0,0.3)
Medium	[0.4,0.7)	[0.3,0.6)	[0.5,0.8)	[0.3,0.7)
High	[0.7,1.0]	[0.6,1.0]	[0.8,1.0]	[0.7,1.0]

China National Internet Emergency Response Center, which contains five characteristics of the virus.

The number of virus-infected hosts, the number of stolen and modified websites, the number of hijacked cross-border websites, the number of phishing pages on cross-border websites, and the number of new information security vulnerabilities [24,25] are listed in Table 1.

The weight of safety level in Table 1 is quantitatively described by interval [0,1], as shown in Table 2, which is convenient for the evaluation and analysis of the model.

For the evaluation of security situation indicators, the ratio of test samples to training samples in this experiment is 110:40. The comparison results of network output and expected output of some test samples are in Table 3.

The above experiments show that 110 training sets and 40 test sets are close to the actual output when using the random forest assessment model for network security situation assessment. The model is evaluated, as shown in formula (8):

Table 3: Actual and expected outputs of model evaluation

Sample number	Actual output	Expected output	Actual threat level	Expected threat level
1	0.9	0.9	High	High
2	0.9	0.9	High	High
3	0.8	0.8	Medium	Medium
4	0.7	0.7	Medium	Medium
5	0.6	0.6	Medium	Medium
6	0.3	0.3	Low	Low
7	0.5	0.4	Low	Low

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}). \tag{8}$$

Among them, TP assumes the positive class as the positive class, and FN assumes the positive class as the negative class. The data in Table 4 show that the random forest-based multi-level CSSA model evaluation method described in this article is superior to Bayesian network and BP neural network in terms of accuracy, absolute error, and prediction speed.

This article and other methods are used to predict the network security situation. The specific results are shown in Figure 3.

As can be seen from the figure, considering the network security situation, the prediction performance of this method is closer to the actual value and outperforms the other two methods. Therefore, the use of random access memory is effective in predicting the perception of multi-level security situations.

Table 4: Comparison of CSSA evaluation results under different algorithms

Performance index	Bayesian network	BP neural network	Random forest
TPR/%	75.0	87.5	97.50
Kappa statistical mean	0.23	0.40	0.93
absolute error	0.15	0.09	0.06
Root mean square error	0.30	0.21	0.14
Time taken/s	0.05	0.05	0.03

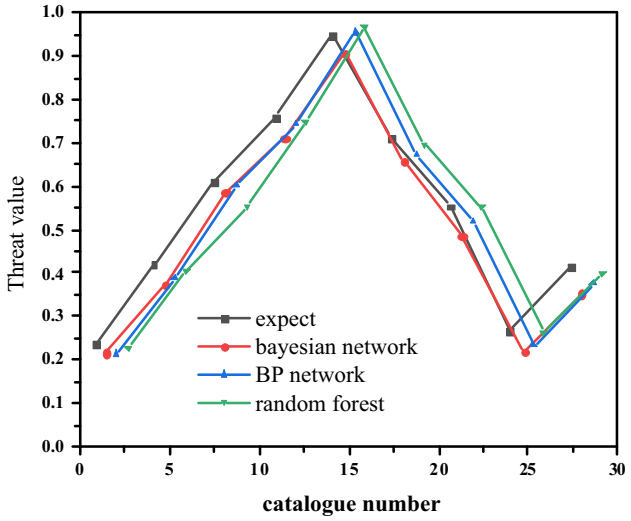

Figure 3: Prediction results of network security situational awareness.

4.2 Future research direction of network security situational awareness technology in big data environment

In this age of big data, research on the security situation of the network in a big data environment is highly valued, improving knowledge of the situation and assessing the security situation of the network. Predictability, timeliness, and accuracy can better ensure network security in a large data environment. In the future, network security knowledge technology in big data environments can be studied in the following ways: (i) Conducting research on big data environments, systematized, standardized, and adapted to different application options in the network security knowledge model, that is, adaptive research, Highly scalable, sustainable network security situation knowledge system breaks down key technologies for extracting network security situations, situation assessments, and predicting situations in big data environments; (ii) establishing an effective, complete, and measurable comprehensive system of network security situation assessment indexes in a standardized big data environment, and accurately, comprehensively, and accurately describing the network security situation in a big data environment, and providing a scientific basis for situation assessment provide. assumptions. (iii) In-depth study of security data characteristics, analysis of the relationship between security data characteristics, and analysis of methods for extracting and analyzing elements of dynamic heterogeneous security data from multiple sources. (iv) In-depth study of network security measurement models in large data environments, measurement methods, measurement functions, analysis models, in-depth research on measurement security object decision criteria, achievements network security measurement system, comprehensive and large decision-making criteria in the information environment. in-depth knowledge; (v) to absorb and learn the advantages of various industries, to study new methods and ideas for quantitatively assessing the security of the network in a big data environment, to study the nature of a huge amount of security information in a big data environment, to be comprehensive and efficient; and to make effective method for accurately assessing the security situation of a network in a large data environment.

(vi) To study the technology of predicting the security situation of the network based on the new generation of artificial intelligence science and technology, thus adapting to the security requirements of large information network environment with large data volume, dynamic changes, high real-time requirements and high coordination, and support the proper management of network security. Decide on guidelines for active and dynamic protection of cyber security.

5 Conclusion

The situation awareness model evaluation method based on random forest can realize fast and accurate situation prediction. The multi-level CSSA model constructed in this study aligns the CSSA process with the security data life cycle. Experiments show that this model and evaluation method are feasible, superior to other existing methods in prediction accuracy and time, and can provide options to evaluate network security. The future research direction is to optimize and improve this method, so that this method can be suitable for the model evaluation of large-scale complex networks. Although the situational awareness research method based on random forest proposed in this article improves the effect of network security situation understanding, assessment, and prediction to a certain extent, and has certain advantages in model establishment time and accuracy, there is still much room for improvement. Although the imbalanced data are adjusted and optimized in this article, the data imbalance still has a great impact on the experimental results, so the model framework will be optimized to minimize the impact of imbalanced data on the accuracy of the model. The experiments in this article are all simulation experiments, and the situation understanding and evaluation and prediction methods proposed in this article have not been applied to the real network for practice loss.

Funding information: This work is supported by the Yunnan Provincial Education Department, Science Foundation of China under Grant 2021J0893 and construction of the seventh batch of key engineering research centers in colleges and universities of Yunnan Province.

Author contributions: Each author made significant individual contributions to this manuscript. Jinkui He: writing and performing surgeries; Weibin Su: data analysis and performing surgeries, article review, and intellectual concept of the article.

Conflict of interest: The authors declare that they have no competing interests.

References

[1] Nikoloudakis Y, Kefaloukos I, Klados S, Panagiotakis S, Markakis EK. Towards a machine learning based situational awareness framework for cybersecurity: An SDN implementation. Sensors. 2021;21(14):4939.
[2] Zhou Q, Shahidehpour M, Alabdulwahab A, Abusorrah A. A cyber-attack resilient distributed control strategy in islanded microgrids. IEEE Trans Smart Grid. 2020;11(5):3690–701.
[3] Li L, He W, Xu L, Ash I, Anwar M, Yuan X. Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. Int J Inf Manag. 2019;45(APR):13–24.
[4] Ukwandu E, Farah M, Hindy H, Brosset D, Bellekens X. A review of cyber-ranges and test-beds: current and future trends. Sensors. 2020;20(24):7148.
[5] Aldawood H, Skinner G. Reviewing cyber security social engineering training and awareness programs – pitfalls and ongoing issues. Future Internet. 2019;11(3):73–3.
[6] Nachin N. How to increase cybersecurity awareness. ISACA J. 2019;2:45–50.
[7] Ahmed A, Krishnan V, Foroutan SA, Touhiduzzaman M, Suresh S. Cyber physical security analytics for anomalies in transmission protection systems. IEEE Trans Ind Appl. 2019;55(99):6313–23.
[8] Rongrong X, Xiaochun Y, Zhiyu H. Framework for risk assessment in cyber situational awareness. IET Inf Secur. 2019;13(2):149–56.
[9] Pokharel HP. Will we see mch care in social security network. BMJ. 2021;331(7525):1107–10.
[10] Zeadally S, Adi E, Baig Z, Khan I. Harnessing artificial intelligence capabilities to improve cybersecurity. IEEE Access. 2020;8(99):23817–37.
[11] Kumar N, Kasbekar GS, Manjunath D. Application of data collected by endpoint detection and response systems for implementation of a network security system based on zero trust principles and the eigentrust algorithm. arXiv preprint. 2022;15(4):155014771984017.
[12] Hang F, Xie L, Guo W, Lv Y, Ou W, Shanthini A. Pervasive hybrid two-stage fusion model of intelligent wireless network security threat perception. Int J High Perform Syst Architecture. 2021;10(3/4):128–39.
[13] Chen J, Miao Y. Study on network security intrusion target detection method in big data environment. Int J Internet Protoc Technol. 2021;14(4):235–49.
[14] Kstle JL, Anvari B, Krol J, Wurdemann HA. Correlation between situational awareness and EEG signals. Neurocomputing. 2021;432(1):70–9.
[15] Huang Y. Retracted: research on the application of network security defense based on artificial intelligence. J Physics Conf Ser. 2021;1992(2):022077 (5pp).
[16] Russell L, Goubran R, Kwamena F, Knoefel F. Agile iot for critical infrastructure resilience: cross-modal sensing as part of a situational awareness approach. IEEE Internet Things J. 2019;5(6):4454–65.
[17] Peng H, Zhang Y, Yang S, Song B. Battlefield image situational awareness application based on deep learning. IEEE Intell Syst. 2019;35(1):36–43.
[18] Gomes HM, Bifet A, Read J, Barddal JP, Enembreck F, Pfahringer B, et al. Correction to: adaptive random forests for evolving data stream classification. Mach Learn. 2019;108(10):1877–8.
[19] Antoniou E, Bozios P, Christou V, Tzimourta KD, Tzallas AT. Eeg-based eye movement recognition using the brain–computer interface and random forests. Sensors. 2021;21(7):2339–9.
[20] Lei Y, Jiang W, Niu H, Shi X, Yang X. Fault diagnosis of axial piston pump based on extreme-point symmetric mode decomposition and random forests. Shock Vib. 2021;2021(4):1–16.
[21] Afiantara A, Mahawan B, Budiarto E. Predicting of banking stability using machine learning technique of random forests. ACMIT Proc. 2021;6(1):1–8.
[22] Sallam RA, El-Sheikh MMA, El-Saedy EI. On the oscillation of second order nonlinear neutral delay differential equations. Math Slov. 2021;71(4):859–70.
[23] Polyanin AD, Sorokin VG. A method for constructing exact solutions of nonlinear delay PDEs. J Math Anal Appl. 2021;494(2):124619.
[24] Ijaz M, Nadeem S, Ayub M, Mansoor S. Simulation of magnetic dipole on gyrotactic ferromagnetic fluid flow with nonlinear thermal radiation. J Therm Anal Calorim. 2021;143(3):2053–67.
[25] Kang J, Chunqing LI, Gao XH, Huibin XU, Chen C, Luo DQ. Establishment of a type 2 diabetes-tumor mouse model and analysis of its intestinal flora. SCI SIN Vitae. 2021;51(9):1308–18.