Journal of Electronic Research and Application, 2025, Volume 9, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Agricultural IoT Security Risk Assessment Method Based on Random Forest

Xinzhe Liu*

College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China

***Corresponding author:** Xinzhe Liu, 1102744585lxz@gmail.com

Abstract: The agricultural Internet of Things (IoT) system is a critical component of modern smart agriculture, and its security risk assessment methods have garnered increasing attention from the industry. Current agricultural IoT security risk assessment methods primarily rely on expert judgment, introducing subjective factors that reduce the credibility of the assessment results. To address this issue, this study constructed a dataset for agricultural IoT security risk assessment based on real-world security reports. A PCARF algorithm, built on random forest principles, was proposed, incorporating ensemble learning strategies to enhance prediction accuracy. Compared to the second-best model, the proposed model demonstrated a 2.7% increase in accuracy, a 3.4% improvement in recall rate, a 3.1% rise in Area Under the Curve (AUC), and a 7.9% boost in Matthews Correlation Coefficient (MCC). Extensive comparative experiments showed that the proposed model outperforms others in prediction accuracy and robustness.

Keywords: Random forest; PCA; Agricultural Internet of Things; Security risk assessment

Online publication: February 14, 2025

1. Introduction

Rapid population growth brings increasing competition for land, water, and other natural resources [1]. These issues urgently require reducing the dependence of food systems on the environment [2]. At the same time, traditional agricultural production methods overuse resources such as water, electricity, fertilizers, and pesticides, resulting in a decrease in land and underground water power year by year [3]. Therefore, a more advanced agricultural model is needed to meet the growing demand for crop production to guarantee sustainable development.

Smart agriculture is made up of emerging technologies such as blockchain, artificial intelligence, and the Internet of Things (IoT) [4]. China has issued a series of policy documents to support the development of smart agriculture, including the Opinions of the Chinese Communist Party (CPC) Central Committee and The State Council on Implementing the Strategy of Rural Revitalization, which put forward the development of digital agriculture and promote the trial and demonstration of the Internet of Things [5]. To meet these needs, the number

of IoT devices used for agricultural purposes is also expected to increase significantly [6].

However, there are many security and privacy concerns at each layer of IoT architecture [7]. There are many cyber-attacks that attackers can launch, such as distributed denial-of-service (DDoS) attacks that make services unavailable and then inject fake data that can affect food safety, agricultural supply chain efficiency, and agricultural productivity [8]. Jawarneh *et al.* have studied the main issues facing agricultural IoT, which they believe include heterogeneous devices and communication, physical device integration, and data privacy concerns, among others [9].

Although the standard proposed by some scholars has a good evaluation range, it does not specify its scoring calculation method, there are no unified scoring calculation rules in the industry and academia, and the evaluation method relies heavily on the scoring of experts in the industry [10–13]. This will not only lead to the Internet of Things platform security assessment score calculation methods are not uniform but also lead to score evaluation because different companies have different calculation methods and evaluation experts have a large error, reducing the feasibility of the national standard. Therefore, government agencies and related enterprises urgently need an objective and accurate security assessment system to unify the scoring standards of agricultural Internet of Things systems.

The machine learning model has the characteristics of relatively objective evaluation scores, independent of prior expert knowledge in the calculation process, good generalization performance, and low application cost, so it has been widely used in different fields. For example, Wang *et al.* used the assessment model based on Random Forest (RF) to evaluate and forecast flood disaster risk, and the results showed that the model could provide a reference for the study of flood risk management and disaster prevention and reduction in river basins [14]. Cen *et al.* proposed a risk assessment method for the operation and maintenance of municipal pipe networks based on machine learning and built a model example based on the data of a pipe network base in a park in Suzhou City to evaluate the leakage risk of the pipe network [15].

Inspired by the above work, this study proposed a machine learning-based security risk assessment method for agricultural Internet of Things systems to quantify the risk factors, which can be used for security risk assessment of specific agricultural Internet of Things platforms, and solve problems such as the inconsistency between the subjective assessment of experts and the calculation standards in the traditional assessment methods.

2. Research methods

2.1. Data set construction

The goal of this study is to improve the quality and applicability of the dataset through a series of methods to provide a solid foundation for model development and further analysis. As there is no open-source data set available for agricultural Internet of Things security risk assessment at present, this study screened 200 agricultural Internet of Things security risk assessment reports from a company with information security level assessment qualification and produced agricultural Internet of Things security risk assessment data set according to 87 security risk items stipulated in the national standard.

2.2. Conditional GAN data enhancement scheme

The dataset contains 200 agricultural Internet of Things security risk assessment data, of which 184 items passed the security risk assessment accounted for the majority, while only 16 items failed the assessment. This leads to

an obvious category imbalance in the data set. Therefore, when constructing the prediction model, corresponding strategies should be taken to deal with the imbalance of the data set to ensure the predictive performance and reliability of the model.

Considering that the characteristic variables of the agricultural Internet of Things security risk assessment data set are all sparse high-dimensional attributes, the traditional interpolation data enhancement method cannot effectively generate similar samples when sampling sparse high-dimensional features. In this study, the conditions to be followed when generating synthetic data were regulated by modeling attributive features and introducing condition vectors represented by masks in the training process of Generative Adversarial Network (GAN) models. When dealing with attributive features, the goal of this study is to constrain the generator results $\{\widehat{d_1}, ..., \widehat{d_{N_d}}\}$ to satisfy as much as possible.

$$\mathbb{P}_G\,(row|D_{i^*} = k^*) \approx \mathbb{P}\,(row|D_{i^*} = k^*) \quad (1)$$

Specifically, for attribute class features in the i row, D_i mask vector is generated m_i, where k represents the mask of the class k; For each category k, according to the number of times that category appears in the column D_in (the total number of categories is N), calculate the probability function $P(D_i = k) = n/N$, select the category according to the probability function k^*, and update the corresponding element k^* in the mask vector m_i to 1; All the generated attribute class feature masks are spliced to form the final discrete quantity generation result $\{d_{1,i},...,d_{N_d,i}\}$. The virtual evaluation unqualified sample generated by the final generator can be expressed as **Equation 2**.

$$\widehat{r_j} = \alpha_{1,j} \oplus \beta_{1,j} \oplus \cdots \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus d_{1,j} \oplus \cdots \oplus d_{N_d,j} \quad (2)$$

Where: $\widehat{r_j}$ is the feature of the final generated security risk assessment sample j, c is the dimension of the numerical feature, α_{N_c} is the numerical feature of the generated sample, is the mode quantity of the numerical feature of the generated sample, and $\beta_{i,j} = [0,0,1]$ is the unique thermal code. If the calculated value belongs to the learned pattern, it can be obtained.

In this study, the general GAN normal form is selected as the generator and discriminator, and the cross entropy of the virtual feature $\widehat{r_j}$ and the real feature generated $\widehat{r_j}$ by the generator is taken as the loss function.

2.3. PCARF

In addition, there is some correlation between the data set indicators. Taking physical location selection as an example, physical location selection often affects other characteristic variables such as anti-theft, waterproofing, physical access control, etc. The nonlinear classification capability of the random forest model enables it to be applied to the agricultural Internet of Things security risk assessment scenario with complex high-dimensional sparse features. However, some noise factors that have little impact on classification results may be incorrectly learned by the random forest model. Therefore, Principal Component Analysis (PCA) technology should be used to find appropriate dimensionality reduction embedding space to unify the measurement of various risk factors. To reduce the influence of noise factors on the prediction results.

The standardization method of traditional PCA is a simple linear average, and the difference in degree information between parameters in various dimensions is eliminated while dimensionality is reduced. Therefore, improvements are made in the standardization method, such as log-centric PCA, which is log-centric processing when data is standardized, and balanced PCA, which is weighted normalized mean processing when data is standardized. The nonlinear PCA method proposed in this section is exponential centralized processing when data is standardized.

2.4. Construction of agricultural Internet of Things security risk assessment model based on random forest

To further improve the prediction accuracy, this study uses the ensemble learning method of "Boosting + Stacking" to enhance the prediction effect by combining multiple base learners, forming an ensemble model with better performance by combining homogeneity or heterogeneity. This process creates multiple base learners, integrates, and outputs results by combining modules.

In this study, the extreme gradient lifting algorithm (XGBoost) in Boosting is used for pre-training as one of the base learners in the Stacking algorithm. Then, Principal Component Analysis Based Random Forest (PCARF), statistical learning model Logistic Regression (LR), and XGBoost are selected to be used. The improved BP neural network, four kinds of models that are intrinsically different from each other, are used for model fusion. In terms of the model fusion strategy, RF, LR and XGBoost are used as first-level learners to be responsible for preliminary feature learning and prediction. The improved BP neural network acts as a second-level learner to further synthesize and fine-adjust the output of the first-level learner. Through this hierarchical fusion method, the advantages of different models are integrated to improve the accuracy and generalization ability of the overall prediction.

In terms of the final score of the agricultural Internet of Things security risk assessment, this study takes the probability that the model output-specific case results are qualified as the agricultural Internet of Things security risk assessment score and determines whether the agricultural Internet of Things security risk assessment is qualified according to the existing grade assessment conclusions.

3. Experiment and analysis

3.1. Performance analysis of agricultural Internet of Things security risk assessment model

In this study, the performance of data enhancement methods was evaluated using t-Distributed Stochastic Neighbor Embedding (t-SNE) combined with Kernel Density Estimation (KDE) dimensionality reduction graphs, as shown in **Figure 1**.

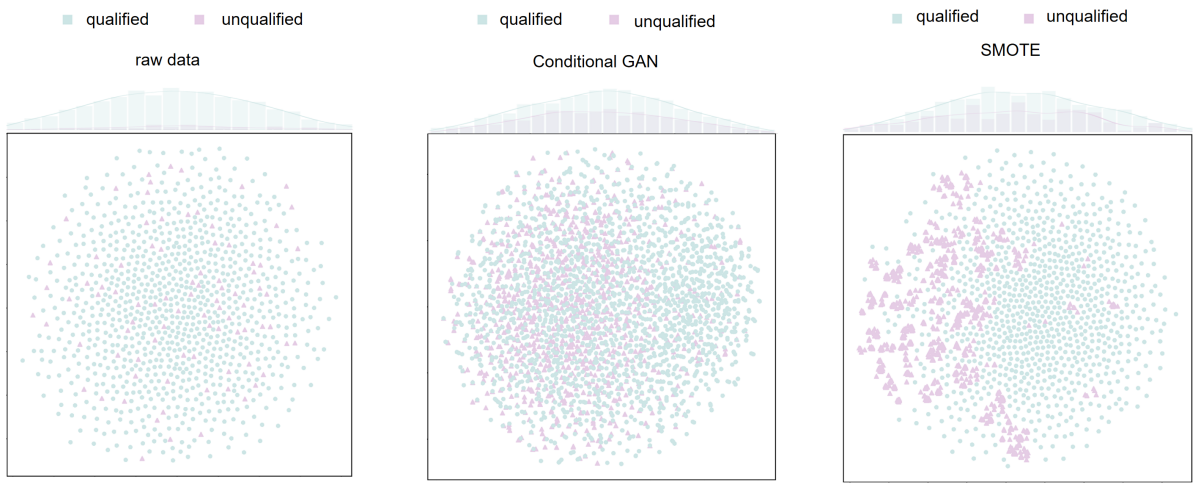

Figure 1. Evaluate unqualified sample data to enhance sample balance results

The Synthetic Minority Over-sampling Technique (SMOTE) method used in this case generated a more uniform sample distribution than the existing interpolation method, which indicates the richness of the generated data. In addition, the closer the distance between the generated sample and the original sample in the t-SNE dimensionality reduction graph, the higher the similarity is. These samples with small differences carry more information, and can better refine the model decision boundary in training and improve the training quality of the model.

The comparative experimental results are shown in **Table 1**. The author compared the accuracy of different models in agricultural Internet of Things security risk assessment when different balanced data participated in model training. The experimental results show that the model used in this study and the conditional GAN achieved the best performance in the test. Compared with the commonly used data-driven model, the recall rate of the agricultural Internet of Things system security risk assessment increased by 3.4%, the AUC increased by 3.1%, and the MCC of the model reached 94.6%. It shows that the model has a good balance in the classification of positive and negative samples.

Table 1. Comparative experimental results

Model name	**Enhanced** data participation rates	**AUC**	**Recall**	**MCC**
Model for this study	0%	59.7%	56.7%	45.3. %
	40%	79.3%	77.6%	65.5%
	70%	93.7%	91.6%	94.6%
Random Forest (RF)	0%	57.7%	56.6%	44.6%
	40%	75.1%	75.0%	60.7%
	70%	90.6%	88.2%	86.7%
AdaBoost	0%	60.8%	59.1%	50.9%
	40%	74.4%	73.1%	62.8%
	70%	89.2%	88.0%	83.9%
MLPClassifier	0%	55.8%	56.2%	39.2%
	40%	75.9%	78.6%	69.6%
	70%	74.1%	69.8%	62.5%

4. Conclusion

Aiming at the problem of the influence of expert subjective factors in the agricultural Internet of Things security risk assessment, this study uses 200 real agricultural Internet of Things security risk assessment reports as data sources, constructs the agricultural Internet of Things security risk assessment data set, and selects the optimized Conditional Tabular Generative Adversarial Network (CTGAN) algorithm as the data balancing method. The random forest algorithm PCARF optimized based on PCA was used to construct the agricultural Internet of Things security risk assessment model by adopting the integrated model method. The accuracy and reliability of the model were proved through experiments. The prediction results of the data-driven model were used as the reference basis for the agricultural Internet of Things security risk assessment, without relying on expert judgment or other artificial risk grade labels. It is more objective and stable than the traditional method. Overall, this study plays an

important role in promoting the improvement of IoT security risk assessment algorithms and obtaining scientific, reliable, and objective agricultural IoT security risk assessment results.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Karlov AA, 2017, Cybersecurity of Internet of Things—Risks and Opportunities. Proceedings of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017), 2017: 182–187.

[2] Stafford JV, 2019, Precision Agriculture '19. The Netherlands: Academic, Wageningen.

[3] Ahmed N, De D, Hussain I, 2018, Internet of Things (IoT) for Smart Precision Agriculture and Farming in Rural Areas. IEEE Internet of Things Journal, 5(6): 4890–4899.

[4] Ferrag MA, Shu L, Friha O, et al., 2021, Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions. IEEE/CAA Journal of Automatica Sinica, 9(3): 407–436.

[5] Zhao C, 2019, Research on the Development Status and Strategic Goals of Smart Agriculture. Smart Agriculture, 1(01): 1–7.

[6] Malavade VN, Akulwar PK, 2016, Role of IoT in agriculture. IOSR J. Comput. Eng., 2016: 56–57.

[7] Tewari A, Gupta B, 2020, Security, Privacy and Trust of Different Layers in Internet-of-Things (IoTs) Framework. Future Gener. Comput. Syst., 108: 909–920.

[8] Zhu WJ, Deng ML, Zhou QL, 2018, An Intrusion Detection Algorithm for Wireless Networks based on ASDL. IEEE/CAA J. Autom. Sinica, 5(1): 92–107.

[9] Kuthadi VM, Selvaraj R, Rao YV, et al., 2023, Towards Security and Privacy Concerns in the Internet of Things in the Agriculture Sector. Turkish Journal of Physiotherapy and Rehabilitation, 32(3).

[10] General Office, Standing Committee of the National People's Congress, 2016, Cybersecurity Law of the People's Republic of China. China Democracy and Legal Press, Beijing, 2016.11.

[11] National Standard of the People's Republic of China, 2022, GB/T 20984-2022 Information Security Technology—Risk Assessment Method for Information Security.

[12] National Standard of the People's Republic of China, 2019, GB/T 22239—2019 Information Security Technology — Baseline for Classified Protection of Cybersecurity.

[13] National Standard of the People's Republic of China, 2019, GB/T 28448—2019 Information Security Technology — Evaluation Requirement for Classified Protection of Cybersecurity.

[14] Wang Z, Lai C, Chen X, et al., 2015, Flood Hazard Risk Assessment Model Based on Random Forest. Journal of Hydrology, 527: 1130–1141.

[15] Cen H, Huang D, Liu Q, et al., 2023, Application, Research on Risk Assessment of Municipal Pipeline Network Based on Random Forest Machine Learning Algorithm. Water, 15(10): 1964.