IoT Equipment Monitoring System Based on C5.0 Decision Tree and Time-Series Analysis

BIAOKAI ZHU, XINYI HOU, SANMAN LIU, WANLI MA, MEIYA DONG, HAIBIN WEN, QING WEI, SIXUAN DU, AND YUFENG ZHANG

1Network Security Department, Shanxi Police College, Taiyuan 030021, China
2State Grid Information & Telecommunication Company of SEPC, Taiyuan 030001, China
3College of Information and Computer, Taiyuan University of Technology, Jinzhong 030600, China
4Shanxi of Computer Emergency Response Team, Taiyuan 030006, China

Corresponding author: Wanli Ma (sgitco@outlook.com)

This work was supported in part by the Education Innovation Fund of Ministry under Grant 2019J02009, in part by the Applied Basic Research Program of Shanxi Province under Grant 201901D211551, in part by the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi under Grant 2019L1009, in part by the Key Research Projects at School Level under Grant 2019yzd002, in part by the Training Tender Research Project at School Level under Grant 2019yzb008, and in part by the Shanxi Provincial “1331 Project” Key Discipline Construction Project under Grant 1331KSC.

ABSTRACT

Abnormal traffic and vulnerability attack monitoring play an important role in today’s Internet of Things (IoT) applications. The existing solutions are usually based on machine learning for traffic, and its disadvantage is that a large number of manual operations are needed in the classification process, and the adaptability is poor. Moreover, for unknown attacks, the system cannot make a relative response in time. In this work, we design a monitoring system of IoT based on C5.0 decision tree and time-series analysis. The system transforms time-series into GAF graph, and uses CNN-LSTM combination model to monitor the traffic. The time-series model based on deep learning can also improve the inefficiency and manual intervention caused by data analysis. At the same time, the system introduces LSTM technology, which can solve a series of problems that may be caused during long sequence training. We select KDD Cup 99 data set for simulation experiments and comparison with traditional traffic monitoring methods. The results show that the average error rate of abnormal traffic attack types is 3.22%. The evaluations show that the system can effectively monitor unknown attacks with 96% accuracy. We further use whitelist matching technology to identify IoT device models. After comparison of experiments, it is proved that this method has its superiority in the monitoring of IoT devices.

INDEX TERMS

Time-series model, C5.0 decision tree, IoT device identification, traffic feature extraction.

I. INTRODUCTION

Driven by the rapid development of big data, artificial intelligence and information communication technology, the scale of the Internet of things is growing rapidly. Internet of things technology and its related applications have also made innovative breakthroughs. With the increasing number of IoT devices, the IoT industry ushers in the golden development stage. With the large-scale application of the IoT, attackers can also take advantage of it. The attacker will access the device with vulnerability into the target network to hide and launch an attack at any time. The architecture of IoT has its own characteristics, once attacked, it may lead to network paralysis, causing great threat and loss to the country and individuals. Compared with the traditional Internet, this situation is more complex and difficult [1].

In the field of network security, abnormal traffic monitoring and vulnerability scanning are always difficult to study. There are many existing methods, such as face recognition, fingerprint feature and characterization acquisition, RFID technology, car networking, smart grid, etc., [2], [3], [4]. Shangguan et al. broke the disadvantage of using special hardware for phase determination in the past and propose a new face image recognition method. The recognition image of face features is divided into several regions, and the LBP data distribution is obtained from them, so that the system can recognize faces more accurately [5]. With the continuous development of science and technology, mobile devices and smart phones are widely used. Various commercial applications came into being. With the increasing deployment of
large-scale sensors, unknown measurement errors are more likely to occur. Based on this, Xiang et al. propose to use a two-stage iterative algorithm to estimate the source existence, source parameters and sensor noise iteratively, so as to reduce the error [6], [7]. Wang et al. package and bundle tasks when collecting data, and use the incentive mechanism of personnel scheduling to solve the problem of unbalanced participation [8]. Taking joint learning as an example, mengkai song et al. systematically combines the working principle of privacy preserving learning technology. There are still many security threats in maintaining privacy preserving learning technology [9]. Fauri D et al. define roles through the semantic information of the BACnet protocol, realize the situational awareness of threats by constructing a knowledge graph of equipment information, and achieve the purpose of abnormal monitoring of equipment [10]. Wu et al. achieve the purpose of network security defense by optimizing the black and white list technology [11]. With the wide deployment of intelligent brain in cities, wireless monitoring system involves the security and privacy of unauthorized video. Cheng et al. design a detection system called dewicam. Arias O et al. take Google nest thermostat and Nike + fuelband as research cases to do in-depth research on the performance of wearable devices, in order to improve its security and privacy settings [12]. The system can detect the wireless camera based on the lightweight smartphone [13]. Xixi Li et al. propose a method of automatically extracting information based on time series technology [14]. Because traditional data monitoring methods can no longer reflect the value and change range of data in real time, some scholars propose adding time series to flow monitoring. People have added time-series technology to applications such as temperature, currency exchange rates, and stocks [15]–[18].

Taking RFID tag technology as an example, many methods are still limited by software and hardware conditions and experimental scenarios. First, the error of the results obtained in most experiments is still relatively large, and some are not suitable for the positioning of related objects [19]. Second, in order to achieve absolute target positioning, many methods need to use special hardware or pay attention to experimental deployment, and invest more tags and antennas [20]. In order to accurately track the object in all directions, Jiang et al. add a specific phase model to the RFID system to obtain its position in three-dimensional space. Finally, the system can accurately track the object in three-dimensional space [21]. Guo et al. propose to apply RFID in industry to detect liquid leakage. By combining coarse-grained RSSI and phase, the system can detect liquid leakage by using the characteristics of inductive coupling between adjacent tags [22].

Based on the above research, we propose a research method for monitoring of IoT devices based on C5.0 and timing analysis. On the basis of IoT monitoring, combined with decision tree C5.0 to build a feature classification framework. The decision tree C5.0 is used as a feature classification module to process the collected data information, which improves the efficiency of data processing and makes the feature classification more accurate. This feature-based time series analysis technique can optimize the real-time monitoring system. By establishing time-series model, the system analyzes and processes the attack information that has occurred. This makes the whole monitoring system more stable and real-time. We design a matching system between data and equipment. The system can match the processed data information with the device model one-to-one, and judge whether the data is abnormal and infer whether the device has loopholes. Our method is nested from data to equipment, which greatly improves the accuracy of IoT equipment monitoring (as shown in Fig. 1).

The design of the IoT equipment monitoring involves the following challenges:

(1) **High precision processing.** In machine learning classification, the classification results are easily affected by noise and data redundancy, which requires us to do a good job in data preprocessing and noise reduction.

(2) **The balance between local feature selection and computational complexity.** Traditional time series monitoring uses global known features, but this paper uses computer view technology to transform data features into GAF. At the same time, the CNN-LSTM combination model is added to this part for feature extraction and classification, which not only improves the accuracy and real-time performance of the system, but also increases the difficulty and complexity of monitoring.

(3) **Data self-adaptive collection.** This paper designs a data and equipment matching system. This link is to realize the conversion of data to equipment. The system constructs the data matching of IoT devices through white list technology. The system corresponds the processed data information with the equipment model. However, building a matching system requires a lot of equipment data. In terms of data collection and processing, we should consider how to achieve adaptation.
We summarize the main contributions of this article as follows:

- We propose a monitoring and forecasting system for IoT devices. At present, with the increasing number of IoT devices and large-scale applications, attackers take this opportunity to connect vulnerable devices to the target network and launch attacks at any time. In this paper, aiming at the current situation of network security, the Device Hive IoT virtual platform is used to collect traffic data of IoT devices. The system uses decision tree C5.0 and time series monitoring model as the information classification module and data module of the system respectively, which improves the classification accuracy and improves the real-time monitoring of traffic.

- We use the CNN-LSTM hybrid model for time-series monitoring. We add a time series model to the monitoring of IoT devices, convert the time series data into a GAF graph and use it as a data feature of the attack information. After that, the pictures are input into the CNN-LSTM combined model for training. Therefore, it is possible to obtain the monitoring results of IoT devices that may have vulnerabilities or attacks in the future, which is more stable and real-time for the entire monitoring system.

- We design a matching system between IoT devices and traffic. For the identification of Internet of Things devices, most of them use deep learning for classification, and the experimental process is relatively abstract and complicated. Moreover, the experiment can only detect whether the equipment is abnormal, and cannot identify the vulnerabilities and attack types of the equipment. This article introduces whitelist matching technology to filter out traffic that does not meet the rules. At the same time, the system can identify the models of IoT devices and realize the conversion of analysis results from data to devices.

The remainder of this paper is organized as follows. Section II describes the system overview. Section III introduces the system structure in detail. Section IV describes the experimental design and performance of the system. Section V reviews the related research works. Finally, Section VI concludes this work.

II. OVERVIEW

With the increasingly fierce network attacks, traffic intrusion and vulnerability attacks against the IoT emerge endlessly. This paper improves on the traditional abnormal traffic detection research, and proposes an IoT device monitoring system based on C5.0 and timing analysis. The main parts are as follows (as shown in Fig. 2):

A. DATA COLLECTION AND PREPROCESSING

The system uses the Device Hive IoT virtual platform to collect traffic data of IoT devices. At the same time, the system uses the KDD 99 data set for simulation experiments, and randomly selects 10% data to form the data set of this experiment. Data is preprocessed through data identification, data cleaning, data standardization and data normalization.

B. CLASSIFICATION DETECTION

The system inputs the preprocessed data into the C5.0 decision tree model to realize the classification and recognition of attack types. At the same time, boosting and pruning methods are used to improve the accuracy of the model, reduce the problem of over fitting of the decision tree, and improve the accuracy.

C. TIME-SERIES MONITORING MODEL

Using visualization technology, the time series data is transformed into GAF graph, which is input into CNN model as feature information. The key features obtained are combined with the selected features in the decision tree training process to form a new feature set, which is input into LSTM model, and finally the data monitoring information and abnormal trend are obtained.

Use visualization technology to convert time series data into GAF graphs, input them as feature information into the CNN model, combine the key features obtained with the selected features in the decision tree training process to form a new feature set, and input the long and short-term memory network (LSTM) model to finally obtain data monitoring information and abnormal trends.

D. WHITELIST-BASED DATA AND DEVICE MATCHING SYSTEM

The monitoring information is filtered twice through whitelist technology, overlapping data is extracted, the range of early
warning retrieval is narrowed, and the final early warning data is obtained; then, the early warning data is matched with the IoT device, and the abnormal IoT device model is identified, and the result is from data to Conversion of equipment. After testing, identify the IoT devices that are vulnerable to attacks and potential vulnerabilities and give warnings to them, so that users can obtain the status information of the IoT devices in advance.

III. SYSTEM DESIGN

A. DATA EXTRACTION AND DATA DESCRIPTION

By comparing the characteristics of different data sets, we believe that the KDD Cup 1999 data set is more suitable for our simulation experiments, and it can prove the superiority of our method to a greater extent. The KDD Cup 1999 data set is a new data set formed by Professor Sal Stolfo and Professor Wenke Lee through data mining and data preprocessing of the DARPA 98 data set. It is more authoritative as the data set used in the KDD competition held in 1999.

The data set contains approximately 5 million pieces of training data and 300,000 pieces of test data. The data is divided into two categories: normal and abnormal. The abnormal types are divided into four major types of attacks: DoS attacks, U2R attacks, R2L attacks, and Probe. The attack types are further subdivided into 39 types of different types of attacks. There are 22 types in the training set, and the test set contains the remaining 17 unknown attacks. This paper randomly selected 10% from the KDD Cup 1999 data set as the training data sample for the experiment.

B. DATA PREPROCESSING

We use Device Hive platform to obtain IoT data information. At the same time, in order to ensure the accuracy of the experimental results, we conduct simulation experiments on the KDD Cup 99 data set.

In order to detect abnormal traffic, we need to mine the data. However, the data is incomplete and inconsistent. This also brings many problems to data mining. In order to obtain high-quality data samples and lay the foundation for the smooth development of the subsequent work of network traffic detection, the data preprocessing part is particularly important. Therefore, we perform data processing on KDD 99 data, the processing process is as follows.

1) DATA IDENTIFICATION

After opening the data set, a plain text file will be obtained. The function of data identification is to classify and identify the data according to the attack type.

2) DATA CLEANING

The data that does not conform to the specification is checked and eliminated. The redundant and repeated data are removed by using the text string processing method, and the useful information is extracted.

3) DATA STANDARDIZATION

These include centralization and normalization, that is, panning and zooming eliminate the differences between features and remove the data unit limit.

We need to calculate the average value of each attribute \( \bar{x}_k \) and standard deviation \( s_k \).

\[
\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \quad (1)
\]

\[
s_k = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} \quad (2)
\]

Among them, \( \bar{x}_k \) is the average value of \( k \) attributes, and \( S_k \) is the standard deviation of each attack type. The data in the data set after data cleaning is standardized, centered by the mean \( \bar{x}_k \), and then scaled by the standard deviation \( S_k \), and finally a normal distribution with a value range of \([0, 1]\) is obtained. The standardized formula is:

\[
x^* = \frac{x - \bar{x}_k}{S_k} \quad (3)
\]

We use KDD 99 dataset to preprocess the data through various steps. The simplified data set provides reliable data source for the subsequent training of classifiers and provides convenience for the follow-up work (as illustrated in Fig. 3).

C. BUILDING FEATURE CLASSIFICATION FRAMEWORK BASED ON DECISION TREE C5.0

The C5.0 decision tree algorithm is used to classify the collected IOT data, and at the same time, the C5.0 training set is formed. The working principle of C5.0 decision tree algorithm in this experiment: as a method of classifying decision tree model, C5.0 algorithm takes information entropy as the measurement to generate decision tree. Based on information gain degree splitting method, it iteratively splits until the sample subset cannot continue to be split, and finally obtains the optimal classification result. The faster the information entropy decreases, the faster the segmentation threshold will also change. In this paper, C5.0 decision tree is used to distinguish the characteristics of IOT data and identify the IoT data with malicious attacks and vulnerabilities.

The decision tree algorithm constructs the decision tree and uses the method of approaching the discrete function value to achieve the purpose of data classification. From ID3 algorithm, C4.5 algorithm to C5.0 algorithm, after continuous optimization and improvement, the classification efficiency and accuracy have been greatly improved [23]. Therefore, this paper finally selects C5.0 algorithm as the classification model. In this paper, a feature classification model based on decision tree C5.0 is designed. After preprocessing the obtained data, it is input into the training set of decision tree C5.0 model, which is used as the feature classification module to classify the data information. This provides a good data base for the subsequent time-series model to monitor.
Gain \((S, A)\) is the information gain divided by attribute \(A\), which means that the information gain obtained by dividing the data set \(S\) by attribute \(A\) represents the degree of information eliminating random uncertainty under one condition. \(S^i_a\) represents a subset of the sample set obtained when the attribute \(A\) takes the value \(a^i\) in the data set \(S\). When the attribute \(A\) takes the value \(a^i\), the larger the characteristic intrinsic value \(\text{SpliInformation}\). The formula for the characteristic intrinsic value is:

\[
\text{SpliInformation}(S^i_a, A) = -\sum_{i=1}^{n} \frac{|S^i_a|}{S} \log_2 \frac{|S^i_a|}{S}
\]  

Information gain rate calculation method:

\[
\text{GainRatio}(S^i_a, A) = \frac{\text{Gain}(S, A)}{\text{SpliInformation}(S^i_a, A)}
\]  

The pruning method is used in the C5.0 decision tree model to improve the accuracy of the model. As a regularization technique, pruning can be calculated to make the model close to the optimal structure, prevent overfitting, and improve the accuracy index.

Based on the data set \(S\) and its information gain rate, build a decision tree \(T\), where \(t\) is the leaf node, where the number of leaf nodes is \(|T|\). \(P_t\) and \(H_t\) are the attributes of the \(t-th\) node on the decision tree, and \(\alpha\) can be regarded as one. Parameters are used to adjust the size of the tree and the trade-off between the fit of the tree and the data. Then \(\alpha|T|\) is the complexity of the tree, and pruning is regarded as the calculation of the model loss value.

Then the loss function can be defined as:

\[
C_{\alpha T} = \sum_{i=1}^{|T|} P_t H_t T + \alpha |T|
\]  

D. MONITORING MODEL BASED ON CHARACTERISTIC TIME SERIES

With the rapid development of the network, the data gradually presents the characteristics of quantification, diversification and fluctuation at any time. Traditional data monitoring methods have been unable to reflect the range of data values in real time, so people have added time series technology to the daily temperature, currency exchange rate, stock and other applications. In this paper, time-series model is added to the safety monitoring of Internet of things. Firstly, the data is transformed into GAF graph by visualization technology, and it is used as the data feature of attack information. Then, one-dimensional convolutional neural network (CNN) is used for training to obtain the key features of detection. After that, a new feature set is formed by combining the original features processed manually and the hidden features generated based on convolutional neural network. The abnormal trend of Internet of things device traffic is obtained by inputting long-term memory network (LSTM). Finally, the system combines whitelist matching technology to filter out traffic characteristics that do not meet the rules. (as depicted in Fig. 4).
After preprocessing the time series data of 39 types of attacks, GAF images are obtained. We randomly selected 6, as shown in Fig. 5.

It is difficult to build a monitoring model by directly adding deep learning into time series. Therefore, we use Python library to convert one-dimensional image into two-dimensional image GAF, (Gram Angular Field, GAF). Its working principle is to transform one-dimensional time series into polar coordinate system in Cartesian coordinate system, and then generate GAF matrix by trigonometric function.

We assume that \( u \) and \( v \) are vectors in two-dimensional space, then the inner product is defined as:

\[
\langle u, v \rangle = u_1 \cdot v_1 + u_2 \cdot v_2
\]  

(7)

The matrix is formed by the dot product of each set of vectors, which is expressed as:

\[
G = \begin{pmatrix}
\langle u_1, v_1 \rangle & \langle u_1, v_2 \rangle & \ldots & \langle u_1, v_n \rangle \\
\langle u_2, v_1 \rangle & \langle u_2, v_2 \rangle & \ldots & \langle u_2, v_n \rangle \\
\ldots & \ldots & \ldots & \ldots \\
\langle u_n, v_1 \rangle & \langle u_n, v_2 \rangle & \ldots & \langle u_n, v_n \rangle
\end{pmatrix}
\]  

(8)

Assuming that the time series is \( Y = x_1, x_2, \ldots, x_n \), we need to scale it to \([-1, 1]\) to make the inner product more accurate:

\[
\langle \cdot, \cdot \rangle : [-1, 1] \times [-1, 1] \rightarrow [-1, 1] \\
(x, y) \rightarrow x \cdot y
\]  

(9)

The combination model of CNN-LSTM is used to predict the devices that may produce vulnerabilities or be attacked in the future. This can make the whole monitoring system more stable, real-time and adaptive. Long short-term memory network can solve a series of problems that may arise in long-term prediction. At the same time, the visualization technology is introduced into the system, which makes the global features localized and improves the accuracy of the model (as shown in Fig. 6).

Neural network is a network structure composed of single neurons. Convolution neural network can train and learn pixels, audio and other objects through the connection of each layer, and finally achieve the function of imitating biological vision and perception. Generally speaking, convolutional neural network is given an unknown picture or audio, by building a model for feature extraction, and the final output results can determine its type, shape, etc. Moreover, with the decrease of the level, the accuracy of the extracted features will also decrease. The closer to the bottom, the more detailed classification features will be extract.

In one-dimensional convolution, the function of convolution is to extract data and obtain translation features in a certain direction (as shown in Fig. 7). We use one-dimensional convolution to predict time series and extract sequence features. We use one-dimensional convolution to predict time series, which is expressed as follows:

\[
a = \{a_0, a_1, a_2, \ldots, a_m\}, L(a) = m + 1
\]  

(10)

\[
b = \{b_0, b_1, b_2, \ldots, b_n\}, L(b) = n + 1
\]  

(11)

\( a \ast b \) represents a new sequence obtained from convolution of \( a \) and \( b \). Assuming that the \( k \) element in \( a \ast b \) is \( (a \ast b)_k \), then the convolution is as follows:

\[
(a \ast b)_k = a_0b_k + a_1b_{k-1} + \ldots + a_{k-1}b_1 + a_kb_0
\]  

(12)

\[
L(a \ast b) = L(a) + L(b) - 1
\]  

(13)

It can be seen that the essential principle of convolution operation is to extract features accurately from cyclic product and addition operation. It can be expressed as:

\[
y(c) = d(c) \ast e(c) = \sum_{j=0}^{l} d(c-j)e(i)
\]  

(14)

where \( c \) is the convolution degree, \( d, e, y \) are time series, and \( l \) is the length of \( e \).

LSTM (Long Short-term Memory) is a kind of long short-term memory network, which belongs to a type of time recurrent neural network (RNN). By training the data features, it can solve a series of problems such as gradient...
disappearance and gradient explosion in the process of RNN training. The core problem of LSTM lies in its information transmission path and the selection and processing of information to be preserved and forgotten in the training process, that is, the problem of cell state and “gate”.

LSTM has a total of three gates: “input gate”, “output gate” and “forgotten gate”, which are expressed as:

\[ i_t = \sigma (w_i \cdot [h_{t-1}^{(i)}, x_t^{(i)}] + b_i) \]  
\[ f_t = \sigma (w_f \cdot [h_{t-1}^{(f)}, x_t^{(f)}] + b_f) \]  
\[ o_t = \sigma (w_o \cdot [h_{t-1}^{(o)}, x_t^{(o)}] + b_o) \]

\[ X_{\text{doppler}} = \frac{\nu}{\lambda} = \frac{1}{4\pi} \cdot \frac{\psi_{i+1} - \psi_i}{2\pi} \]  

**IV. IMPLEMENTATION AND EVALUATION**

**A. EXPERIMENTAL SETUP**

In this paper, the experimental environment is as follows: the operating system is windows 10, the memory is 32 GB, the Intel (R) core (TM) i7 – 8550u CPU is 1.80 GHz, and the software environment is Pycham 2020.

**B. DATA EXTRACTION**

We use KDD 99 data set for simulation experiment. Before classifying the data, it is still necessary to process the KDD Cup 1999 dataset. The tag item is the statistical diagram of the total traffic of attack type, as shown in Figure 8. Through research and experimental analysis, we put forward a research method of Internet of things equipment monitoring based on C5.0 and time series analysis, which provides a new idea for the monitoring of IoT equipment. We use KDD 99 data set for experimental test (as shown in Table 1).

| Attack category | Training set | Test set |
|-----------------|--------------|----------|
| DOS             | 79.24%       | 73.90%   |
| U2R             | 1.10%        | 7.30%    |
| R2L             | 2.28%        | 5.20%    |
| Probe           | 8.81%        | 1.33%    |
In the following, we extract 7 graphs from 19 statistical distribution charts of attack traffic and normal traffic, as shown in Fig. 9 to Fig. 15. These include the following attack types src_bytes, srrror_rate, dst_bytes, cont, access control file times, login failure times, wrong segments and continuous.

C. CLASSIFICATION DETECTION
Because decision tree C5.0 has the advantages of low data preparation requirements, it can analyze large data sources in a relatively short time and obtain relatively accurate results. Therefore, this paper uses decision tree C5.0 method to establish a classification model for traffic information. The system reduces the dimension of the data input from feature extraction, so as to build a model and identify the attack type. At the same time, the system introduces boosting technology, which can effectively improve the accuracy and achieve the purpose of model data pruning and optimization.

D. EXPERIMENTAL RESULTS
In this paper, open data set KDD 99 is used for simulation experiment. The system combines C5.0 decision tree and timing series technology to monitor the data of IoT. The results show that the proposed method has some advantages over the traditional methods in monitoring accuracy. Therefore, the application of this system opens up a new idea for the traffic analysis of IOT devices.
In the IoT monitoring, we can collect data through the device hive IoT virtual platform. Most of the devices used in the monitoring platform are representative and commonly used in the market, including 43 kinds of intelligent temperature management system, household appliances, etc. In order to make all the tested devices generate enough training data, we repeat the process 10 times. The experimental architecture is shown in Fig. 17.

The experimental operation process is as follows:

After getting the data of IoT devices through Device Hive and training, the initialization settings are set to activate the device. At the same time, with the help of the application software provided by the supplier, the device is connected to WiFi or Ethernet, and the credentials will be synchronously transmitted to the user network. After a series of operation, the equipment will be forced to restore the factory settings, which is convenient for repeated testing of equipment data.

Decision tree C5.0 is adopted as a training model, and the system inputs processed data for classification training. At the same time, the system uses methods such as pruning to improve the accuracy of the model. This can solve the problem of overfitting of the decision tree, thereby improving the accuracy index. Then, the system builds a time series model to analyze the classification results and obtain data monitoring and analysis results.

Compared with the traditional Internet of things equipment identification and monitoring methods, the accuracy and stability of this paper are improved. The system optimizes the processing of data duplication and redundancy. At the same time, we introduce C5.0 decision tree to split the sample
more difficult. This means that the ability to detect intruders comprehensive monitoring of information flow will become FAR attack.

DR is regarded as the ratio of C5 ACC action evaluation.

FAR is the proportion of false positive rate. After testing and evaluating the classification effect, Table 2 shows the accuracy rate, detection rate and false alarm rate of the four attacks in the overall attack classification. The experimental results show that the overall detection rate and accuracy of the data samples are improved.

V. RELATED WORK

TABLE 2. Comparison of detection rates of different attack types.

| Type  | ACC(%) | DR(%) | FAR(%) |
|-------|--------|-------|--------|
| DOS   | 95.68  | 96.24 | 4.77   |
| U2R   | 97.55  | 98.03 | 2.91   |
| R2L   | 97.06  | 98.01 | 3.87   |
| Probe | 96.55  | 97.98 | 4.81   |
| Total | 96.71  | 97.54 | 4.09   |

subset until the sample subset cannot be split. In this way, the optimal classification results can be obtained. Moreover, time series analysis will affect the real-time performance of feature classification and data monitoring. The system adds time series technology to predict abnormal traffic, which makes traffic monitoring more timely, and can avoid many unnecessary losses caused by malicious attacks.

We take the accuracy ACC, detection rate DR and false positive FAR as the measurement of decision tree C5.0 classification evaluation. ACC is used to evaluate system performance. DR is regarded as the ratio of C5.0 model to detect intrusion attack. FAR is the proportion of false positive rate. After testing and evaluating the classification effect, Table 2 shows the accuracy rate, detection rate and false alarm rate of the four attacks in the overall attack classification. The experimental results show that the overall detection rate and accuracy of the data samples are improved.

V. RELATED WORK

We review the related works in three aspects.

A. IoT EQUIPMENT MONITORING METHOD BASED ON FLOW ANALYSIS

Doshi R et al. carry out a distributed denial of service (DDoS) attack on the Internet infrastructure to avoid a series of security risks that may be caused by its connection with insecure IoT devices. This method promotes innovative research on the direction of IoT attack traffic under automatic detection [24]. Gendreau A et al. believe that in a heterogeneous network, comprehensive monitoring of information flow will become more difficult. This means that the ability to detect intruders in each type of device will directly determine the smooth spread of system information [25]. Doshi et al. take ios devices as the research object, and conduct in-depth research on data filtering of IoT devices [24]. Trilles S and others borrow IoT platforms or software to coordinate device connections and program applications in the network. Among them, the microservice architecture and serverless computing are used as examples to avoid potential vulnerabilities of the Internet of Things that may cause a series of hidden security threats [26].

B. INFORMATION CLASSIFICATION OF IoT BASED ON MACHINE LEARNING

Machine learning can realize data mining and information acquisition, as a method widely used in classification detection and data processing. It can provide a new idea to solve the security problems of the IoT. Majumdar et al. deeply analyze a variety of machine learning technologies and conclude that machine learning can be used not only in intelligent monitoring, traffic analysis, model training, and classification, but also in cloud computing, big data, and artificial intelligence [27]. Based on machine learning, Doshi and others build the overall framework of the IoT security system, and realize a reasonable security management method [24]. Tarek Salah et al. design and implement a classifier based on C5.0 decision tree, build a hardware-based intrusion detection system model. This method detects most network attacks with FPGA technology, and greatly improves experimental detection rate [28]. In the field of machine learning classification and detection, the application of various data sets in experiments makes the detection results more accurate. Shone N et al. propose a new deep learning technology for intrusion detection. This technique utilizes an unsupervised feature learning asymmetric depth automatic encoder. The deep learning classification model is evaluated by KDD 99 and NSL-KDD data sets [29]. Devi et al. classify and compare the KDD 99 and NSL KDD datasets. This detection system can classify and process different data [30].

C. MONITORING METHOD BASED ON TIME-SERIES TECHNOLOGY

With the overall situation of network attack becoming more and more serious, the research of intrusion and traffic monitoring is particularly important. Based on this, the method adopting time series analysis technology is widely used in the field of network security maintenance. Li et al. use deep learning technology to obtain features from time series, convert time series information into images, and improve self-system adaptability and reduce human intervention [14].

Taking stock selection as an example, Zhang et al. construct a new stock selection model by combining decision tree C5.0 algorithm with factor analysis. Experimental results show that the model can effectively help investors avoid risks [31]. Different from the traditional way, Shang et al. study nonlinear time series analysis and process traffic management data through simulation experiments. The experiments show
that it is feasible to apply time series technology to traffic data monitoring and processing, and improve the accuracy and real-time of data [32]. Taking joint learning as an example, mengkai song et al. systematically combines the working principle of privacy preserving learning technology. There are still many security threats in maintaining privacy preserving learning technology [9]. Nanda et al. use sensor data for smart city data monitoring. They propose a method to build a recurrent neural network model by using two attention mechanisms to establish a connection between intelligent buildings and intelligent transportation [33].

VI. CONCLUSION

In this work, we introduce a monitoring system for IoT devices based on decision tree C5.0 and timing analysis. The system is based on the KDD Cup 99 data set for simulation experiments. The experimental results can achieve an average error rate of 3.22%, and can monitor unknown attacks with 96% accuracy. By comparing with traditional flow monitoring methods, excellent performance can be obtained. It is proved that this method has its superiority in the monitoring of IoT devices.

REFERENCES

[1] J. A. Stankovic, “Research directions for the Internet of Things,” IEEE Internet Things J., vol. 1, no. 1, pp. 3–9, Feb. 2014.
[2] M. El-hajj, M. Chamoun, A. Fadlallah, and A. Serhrouchni, “Analysis of authentication techniques in Internet of Things (IoT),” in Proc. 1st Cyber Secur. Netw. Conf. (CSNet), Oct. 2017, pp. 1–3.
[3] M. El-hajj, A. Fadlallah, M. Chamoun, and A. Serhrouchni, “A survey of Internet of Things (IoT) authentication schemes,” Sensors, vol. 19, no. 5, p. 1141, Mar. 2019.
[4] K. Fan, Y. Gong, C. Liang, H. Li, and Y. Yang, “Lightweight and ultra-lightweight RFID mutual authentication protocol with cache in the reader for IoT in 5G,” Secur. Commun. Netw., vol. 9, no. 16, pp. 3095–3104, Nov. 2016.
[5] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[6] C. Xiang, P. Yang, C. Tian, H. Cai, and Y. Liu, “Calibrate without calibrating: An iterative approach in participatory sensing network,” IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 2, pp. 351–361, Feb. 2015.
[7] C. J. D’Orazio, K.-K.-R. Choo, and I. T. Yang, “Data exfiltration from Internet of Things devices: IOS devices as case studies,” IEEE Internet Things J., vol. 4, no. 2, pp. 524–535, Apr. 2017.
[8] Z. Wang, J. Hu, Q. Wang, R. Lv, J. Wei, H. Chen, and X. Niu, “Task-Bundling-Based incentive for location-dependent mobile crowdsourcing,” IEEE Commun. Mag., vol. 57, no. 2, pp. 54–59, Feb. 2019.
[9] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, “Analyzing user-level privacy attack against federated learning,” IEEE J. Sel. Areas Commun., vol. 38, no. 10, pp. 2430–2444, Oct. 2020.
[10] D. Fauri, M. Kapsalakis, D. R. dos Santos, E. Costante, J. den Hartog, and S. Etalle, “Role inference anomaly detection situational awareness in bacnet networks,” in Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment. Springer, 2019, pp. 461–481.
[11] C. J. Wu, Y. Tie, S. Haru, K. Tamiya, A. Fujita, K. Yoshioka, and T. Matsumoto, “IoTPrevent: Highly deployable whitelist-based protection for low-cost Internet-of-Things devices,” J. Inf. Process., vol. 26, pp. 662–672, Oct. 2018.
[12] O. Arias, J. Wurm, K. Hoang, and Y. Jin, “Privacy and security in Internet of Things and wearable devices,” IEEE Trans. Multi-Scale Comput. Syst., vol. 1, no. 2, pp. 99–109, Apr. 2015.
[13] Y. Cheng, X. Ji, T. Lu, and W. Xu, “DeWiCam: Detecting hidden wireless cameras via smartphones,” in Proc. Asia Conf. Comput. Commun. Secur., May 2018, pp. 1–13.

[14] X. Li, Y. Kang, and F. Li, “Forecasting with time series imaging,” Expert Syst. Appl., vol. 160, Dec. 2020, Art. no. 113680.
[15] J. Arroyo, R. Espinola, and C. Mate, “Different approaches to forecast interval time series: A comparison in finance,” Comput. Econ., vol. 37, no. 2, pp. 169–191, Feb. 2011.
[16] X. Wang, S. Chen, and W. Zheng, “Traffic incident duration prediction based on partial least squares regression,” Procedia Soc. Behav. Sci., vol. 96, pp. 425–432, Nov. 2013.
[17] H. Liu, T. Taniguchi, Y. Tanaka, K. Takenaka, and T. Bando, “Visualization of driving behavior based on hidden feature extraction by using deep learning,” IEEE Trans. Intell. Transp. Syst., vol. 18, no. 9, pp. 2477–2489, Sep. 2017.
[18] L. Harris and A. Amato, “Illiquidity and stock returns: Cross-sectional and time-series effects: A replication,” Crit. Finance Rev., vol. 8, nos. 1–2, pp. 29–71, 2019.
[19] T. Liu, L. Yang, Q. Lin, Y. Guo, and Y. Liu, “Anchor-free backscatter positioning for RFID tags with high accuracy,” in Proc. IEEE Conf. Comput. Commun., Apr. 2014, pp. 379–387.
[20] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, “Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices,” in Proc. 20th Annu. Int. Conf. Mobile Comput. Netw., Sep. 2014, pp. 237–248.
[21] C. Jiang, Y. He, S. Yang, J. Guo, and Y. Liu, “3D-OmniTrack: 3D tracking with COTS RFID systems,” in Proc. 18th Int. Conf. Inf. Process. Sensor Netw., Apr. 2019, pp. 25–36.
[22] J. Guo, T. Wang, Y. He, M. Jin, C. Jiang, and Y. Liu, “TwinLeak: RFID-based liquid leakage detection in industrial environments,” in Proc. IEEE Conf. Comput. Commun., Apr. 2019, pp. 883–891.
[23] N. Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification,” ACM SIGCOMM Comput. Commun. Rev., vol. 36, no. 5, pp. 5–16, Oct. 2006.
[24] R. Doshi, N. Athanopoulou, and P. Feamster, “Machine learning DDoS detection for consumer Internet of Things devices,” in Proc. IEEE Secur. Privacy Workshops (SPW), May 2018, pp. 29–35.
[25] A. A. Gendreau and M. Moorman, “Survey of intrusion detection systems towards an end to end secure Internet of Things,” in Proc. IEEE 4th Int. Conf. Future Internet Things Cloud (FiCloud), Aug. 2016, pp. 84–90.
[26] S. Trilles, A. González-Pérez, and J. Huerta, “An iot platform based on microservices and serverless paradigms for smart farming purposes,” Sensors, vol. 20, no. 8, p. 2418, 2020.
[27] N. Majumdar, S. Shukla, and A. Bhatnagar, “Survey on applications of Internet of Things using machine learning,” in Proc. 9th Int. Conf. Cloud Comput., Data Sci. Eng., Jan. 2019, pp. 562–566.
[28] S. Sohbi and M. I. Amer, “Fpga-based network traffic security: Design and implementation using c5.0 decision tree classifier,” J. Electron. Sci. Technol., vol. 11, no. 4, pp. 393–403, 2013.
[29] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu, “Relative localization of RFID tags using spatial-temporal phase profiling,” in Proc. 12th Symp. Netw. Syst. Design Implement., 2015, pp. 251–263.
[30] R. D. Ravipati and M. Abukalibah, “Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets—A review paper,” SSRN Electron. J., 2019.
[31] Q. Zhang, J. Zhang, Z. Chen, M. Zhang, and S. Li, “A new stock selection model based on decision tree C5.0 algorithm,” J. Investment Manage., vol. 7, no. 4, pp. 117–124, 2018.
[32] P. Shang, X. Li, and S. Kamae, “Chaotic analysis of traffic time series,” Chaos, Solitons Fractals, vol. 25, no. 1, pp. 121–128, Jul. 2005.
[33] X. Fan, C. Xiang, C. Chen, P. Yang, L. Gong, X. Song, P. Nanda, and X. He, “BuildSenSys: Reusing building sensing data for traffic prediction with cross-domain learning,” IEEE Trans. Mobile Comput., early access, Feb. 28, 2020, doi: 10.1109/TMC.2020.2976936.

BIAOKAI ZHU is currently working with the Shanxi Police College, Taiyuan, China. His current research fields include information security, network traffic analysis, wireless identification, and sensing platform.
**Xinyi Hou** is currently pursuing with the Shanxi Police Academy, Taiyuan, China. Her research interests include information security, traffic analysis, and data analysis.

**Sanman Liu** is currently working with the Shanxi Police College, Taiyuan, China. Her current research fields include cyber security, information security, and data analysis.

**Wanli Ma** received the B.S. degree in communication engineering and the B.B.A. degree in MBA from North China Electric Power University, in 1987 and 2002, respectively. Since 2018, he has been the General Manager with State Grid Information & Telecommunication Company of SEPC.

**Meiya Dong** is currently pursuing the Ph.D. degree in electronic science and technology with the Taiyuan University of Technology, Taiyuan, China. His current research interests include wireless sensor networks, the Internet of Things, and data processing.

**Haibin Wen** received the master’s degree from the Taiyuan University of Technology, in 2014. He is currently an Engineer with the National Computer Emergency Technology Processing Coordination Center, Shanxi Branch, mainly engaged in computer network and information security research.

**Qing Wei** is currently pursuing the degree in information security with the Shanxi Police College. Her research interests include information security, traffic analysis, and deep learning.

**Sixuan Du** is currently pursuing the degree in information security with the Shanxi Police College. Her research interests mainly include SDN and traffic analysis.

**Yufeng Zhang** is currently working with the Shanxi Police College, Taiyuan, China. Her research interests include linguistics, second language acquisition, and the Internet of Things.