iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53146

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Machine Learning-Based Research on Network Security's Primary Technologies

Mr. J. Santhosh[1], Mr. N. Vijayakumar[2]
[1]*Assistant Professor, Department of Information Technology, Karpagam Institute of Technology, Coimbatore, Tamilnadu, India*
[2]*Assistant Professor, Department of Computer Science and Technology, Karpagam College of Engineering, Coimbatore, Tamilnadu, India*

Abstract: ***A large number of network devices, applications, and the explosive growth of network data have made the network environment extremely complex with significant potential hazards to network security. This is due to the ongoing development of information technology. Cyber attackers are now attacking network settings with connected histories rather than just common users; examples of these environments include businesses, governments, and nations. Massive amounts of Internet data have been produced by the diversification of network services, and traditional network security systems have had trouble keeping up with the demands of network security in terms of performance and self-adaptability. The growth of concepts in the field of network security has been greatly aided by research on machine learning-based network security that has produced numerous findings, demonstrating strong skills in processing large data, automatic learning, detection, and identification. In this paper, we integrate machine learning-related technologies to enhance the performance of intrusion detection and alarm correlation automation, and we investigate key technologies such as machine learning-based network security situational awareness methods and dynamic data stream classification methods based on judgment feedback, to enhance the detection performance, adaptive and generalization capabilities of machine learning-based network security.***
Keywords: network security; machine learning; intrusion detection; anomaly detection; key technologies

I. INTRODUCTION

The results of "Internet+" have helped people's life, economics, government, and other aspects, and have become a new driving force for national growth with the rapid development of information technology, owing to the mature application of communication, big data, and cloud computing. The number of users and financial advantages of online leisure, travel, and education have increased dramatically over the past year [1]. The public has benefited from internet technology and the advancement of network technology, but they have also raised a number of security concerns, leading to the development of specialised cyberattack methods for particular scenarios [2]. Cybercriminals can use harmful ads, password-stealing attacks against unsuspecting users, and other tactics to accomplish their goals [3]. These online criminals have begun to target networks with institutional and governmental roots. The Internet has made a lot of hidden threats public, and once a state or an individual is the victim of a cyberattack, this damage to their interests can be significant. As a result, governments, big businesses, banks, etc., are starting to see increased risks from outside cyberattacks. If these attackers are successful in their attacks, the state, government, etc., will suffer significant losses.

The current state of network security has created new demands for network security research because intrusion detection techniques based on fixed rule matching are unable to respond to the growing network traffic, changing network environment, and developing network technologies, as well as being unable to detect unidentified network attacks [4]. Machine learning has produced a significant number of significant results in data analysis, detection and identification, and artificial intelligence by extracting useful information from enormous amounts of big data, which offers new solutions to the current network security issues [5].

II. TECHNIQUES RELATING TO MACHINE LEARNING

A. Statistical Machine Learning Methods

A significant area of machine learning is statistical learning techniques. Single classification support vector machines are an extension of support vector machines in the unsupervised domain, which are a special application of statistical learning theory [6]. The ideas of structural risk minimization theory and VC dimension theory serve as the foundation for the discipline of machine learning, which is focused on the theory of finite sample statistical estimation and prediction. The VC dimension theory is presented to assess whether the learning method satisfies the consistency constraint. A higher VC dimension indicates more complex functions, correspondingly larger confidence bounds, and a greater difference between real and empirical risks, which affects the

generalisation ability of machine learning algorithms. VC dimension theory defines a measure of function set capacity and learning ability, reflecting the complexity of the function set. Machine learning methods are necessary to not only minimise the empirical risk, or the fundamental principle of structural risk minimization, in the case of limited samples in practise [7].

The primary idea behind the commonly used classification technology known as the support vector machine is to discover the largest edge hyperplane between classification samples in order to improve generalisation error performance. The type marker can be predicted by the following for a given set of data samples (X(,yi)| = i 1, 2,..., The linear decision boundary can be expressed as wx + b = 0.

The following constrained optimisation problem can be used to represent the issue of finding the maximum edge hyperplane.

Therefore mentioned issue is a convex quadratic optimisation problem with linear constraints, and the issue can be resolved by adding Lagrange multipliers to convert the issue to a quadratic programming problem that satisfies the Karuch-Kuhn-Toucher requirements [8].

B. Feature Dimensionality Reduction Methods In Machine Learning

Feature reduction is a technique for lowering the amount of features without lowering the data's expressiveness. Many machine learning techniques have been used to reduce the number of features, including rough set theory, which is frequently used for attribute selection in data classification issues, and principal component analysis, which extracts features by transforming the feature space [9].

The theory of optimum linear and planar fitting to a number of points in space is called principle component analysis, or PCA. Using a linear combination of the features of the few variables to explain the variance-covariance of the random variable structure as thoroughly as possible, PCA aims to represent the original high-dimensional features in terms of a few uncorrelated composite variables. The composite variable with the largest variance among all combinations is known as the first principal component, and the second and third principal components can then be calculated.

C. Artificial Neural Network Method

The MP model and the perceptron model served as the foundation for the development of artificial neural networks. The perceptron model serves as the fundamental processing unit in many neural network models, which are combinations or deformations of the perceptron model. Below is an illustration of a typical perceptron computational result.

The theory of optimum linear and planar fitting to a number of points in space is called principle component analysis, or PCA. Using a linear combination of the features of the few variables to explain the variance-covariance of the random variable structure as thoroughly as possible, PCA aims to represent the original high-dimensional features in terms of a few uncorrelated composite variables. The composite variable with the largest variance among all combinations is known as the first principal component, and the second and third principal components can then be calculated.

Where W = W1, W2,..., Wn is the weight, Wt is the parameter to be trained in the perceptron model, and X = X1, X2, —, Xn is the input vector. Wt determines the contribution of Xt to the output.

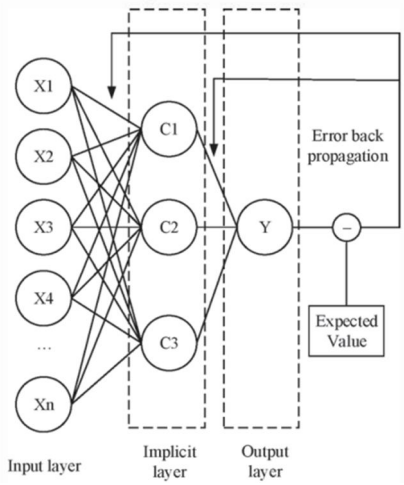


Figure 1. Typical BP neural network structure with three layers

Neural networks are frequently employed to address classification issues because they have high classification capabilities for nonlinear situations [11]. The most popular neural network models are called feed-forward neural networks. In these networks, neurons are arranged in layers, with multiple neurons in each layer. The neurons in a layer are not connected to one another, and the neurons in the layer below serve as the input to the neurons in the layer above. Back propagation neural network, or BP, is the most popular feedforward neural network. The input, implicit, and output layers make up the architecture of the BP algorithm, a supervised learning technique that employs gradient descent to reduce the sum of squared errors between the network's output value and expected value [12]. Figure 1 depicts a typical three-layer BP neural network configuration.

III. AN APPROACH TO NETWORK SECURITY SITUATIONAL AWARENESS BASED ON MACHINE LEARNING

A. Situational Awareness for Cybersecurity as a Whole

Based on the detection of security events, this study develops a mechanism for evaluating and forecasting the network security posture. The major body of this approach is divided into two sections: the first section develops the detection methods for malicious script security events, DGA malicious domain name security events, and SQL injection attack security events [13]. The impact of these security events on the present network security state is given particular weight values in the second section by extracting the components of these security events, and the current network security posture is then assessed and projected. Figure 2 depicts the whole network security situational awareness procedure.

B. Method for Detecting SQL Injection Attacks

The main procedure of the injection attack detection method entails: filtering or setting the criteria of the data to be returned or inserting new and updated records by examining the injection points of typical SQL injection attacks; preprocessing the text of existing SQL statements and segmenting the text after matching data to SQL statements using

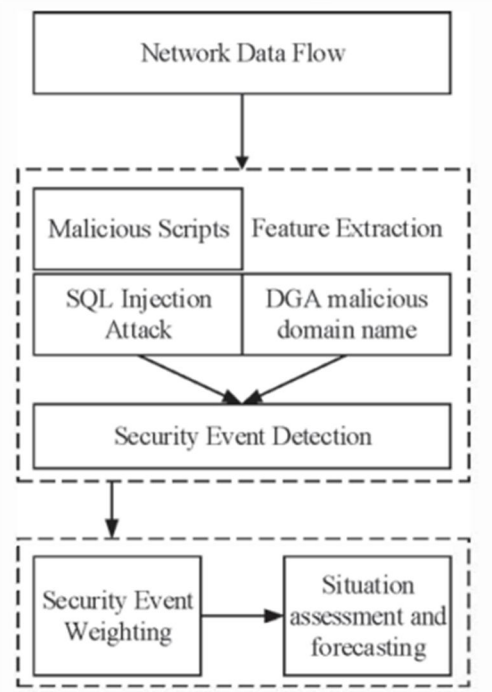


Figure 2. The entire network security situational awareness procedure

Regular expressions; hashing the text that has been matched; SVM is used to model the acquired data in order to develop a model for identifying SQL injection attacks [14]. The SQL injection attack detection technique is created using the SVM algorithm, and the framework diagram of the method is displayed in Figure 3.

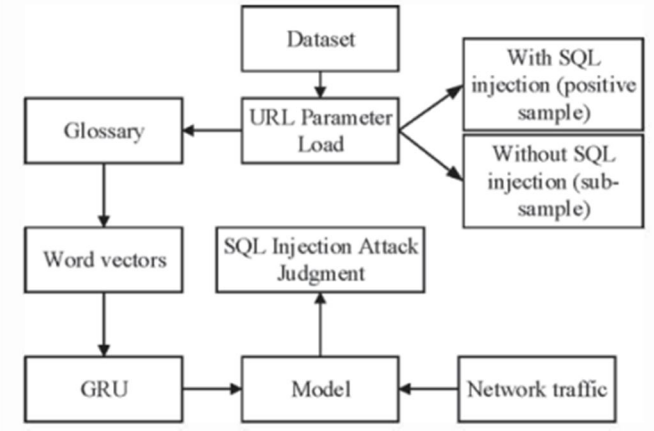


Figure 3. Foundation for an SVM-based injection attack detection technique

C. Method for Detecting Malicious Domain Name Security Events (DGA)

Data-based methodologies are being used by more and more academics and developments to identify harmful domains. In addition to detecting dangerous domains that already exist, new detection DGA models should be able to identify malicious domains created by malware authors utilising novel techniques. The identification of malicious domains produced by DGA can be improved by using auxiliary information as pertinent feature values.

A well-known integrated learning technique based on decision trees is gradient boosting decision tree (GBDT). The GBDT trains negative gradients to create a decision tree throughout each iteration. Finding the best cut spots when learning the decision tree is where the majority of the overhead occurs.

The cut spots are typically discovered using pre-ranking algorithms and histogram-based methods. Large-scale datasets are frequently sparse in actual applications, and GBDT based on preprocessing methods can minimise training loss by eliminating zeros [15]. Figure 4 depicts the strategy for detecting malicious domain security events using the DGA.

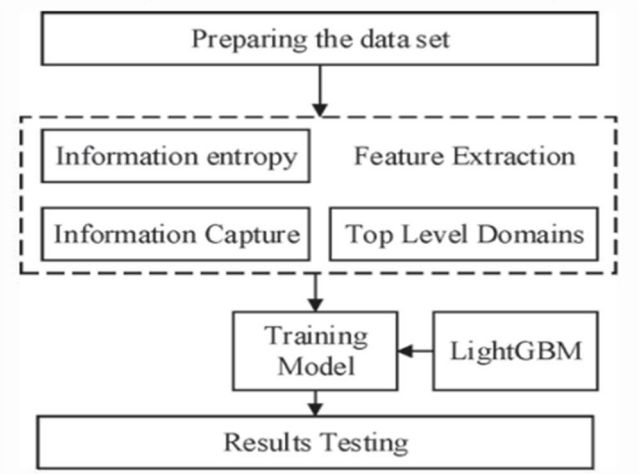


Figure 4. Flow of the DGA method for detecting malicious domain names in communications

IV. JUDGMENT-BASED DYNAMIC DATA STREAM CLASSIFICATION METHOD

Single classifier and integrated classifier are the two basic types of data stream classification techniques. Data streams frequently show dynamic properties like concept drift, which refers to the statistical characteristics of data changing in an unpredictable way over time. Integrated learning is another method for classifying data streams. The single classifier can be seen as a continuation of the traditional data classification model for data stream classification, which is based on the sliding window method to construct data blocks.

Integrative learning is a different approach to data stream classification that has been found to be easier to update, more resistant to idea drift, and more accurate in classifying data. As a result, research on data stream classification algorithms has shifted its focus in recent years. Semi-supervised learning techniques have been used in intrusion detection research because it is challenging to get labelled data in this field.

Semi-supervised learning enhances classification performance from the sample perspective by maximising the consistency of the characteristics between labeled-like labelled samples concurrently to complete the data classification. Integration learning can improve performance from the classifier perspective by combining multiple classifiers with some strategy to improve the generalisation ability of the classifier. Semi-supervised learning builds an expanded integrated classifier using the information in unlabeled data, and the variety of classification outcomes introduced by unlabeled samples improves the learner's performance. The primary goal of the semi-supervised approach to integrated classification models is to divide the data stream into finite-sized data blocks, train the classifier, and compose the initial integrated learning model using the existing labelled data blocks; when the newly arrived data blocks are labelled, train the new classifier, and update the integrated learning model; if the newly arrived data blocks are unlabeled, use the existing model to coarsely judge the data. In the event that the newly incoming data block is unlabeled, the existing model's coarse judgement data is used to pre-classify known data types, store the outliers, and then pass them along to the new type detection module for processing. Figure 5 depicts the semi-supervised integrated learning model's data processing flow.

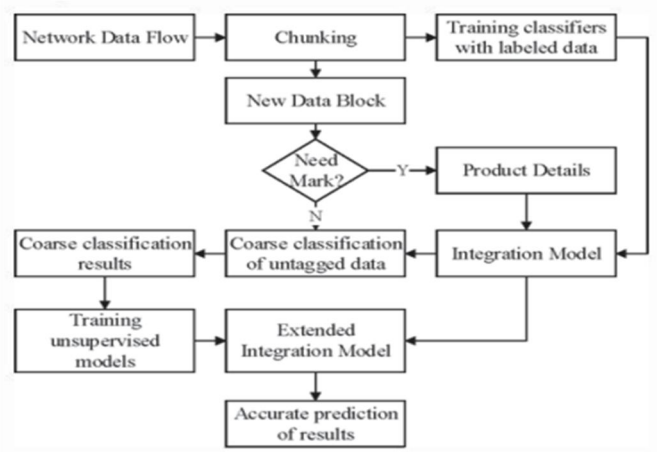


Figure 4. Semi-supervised integrated learning model's data processing flow

The variety and complexity of cyberattacks cause the network data that is gathered to have dynamic data flow characteristics, meaning that both new and existing data kinds may simultaneously appear and disappear. The initial integrated model's classification can identify both known and unidentified categories. When new data is captured, it is classified according to the initial integration model if it falls within the decision boundary of existing types; if not, it is stored as a new outlier, and when the number of outliers reaches a certain level and has similarity, it is considered to be a new type of data.

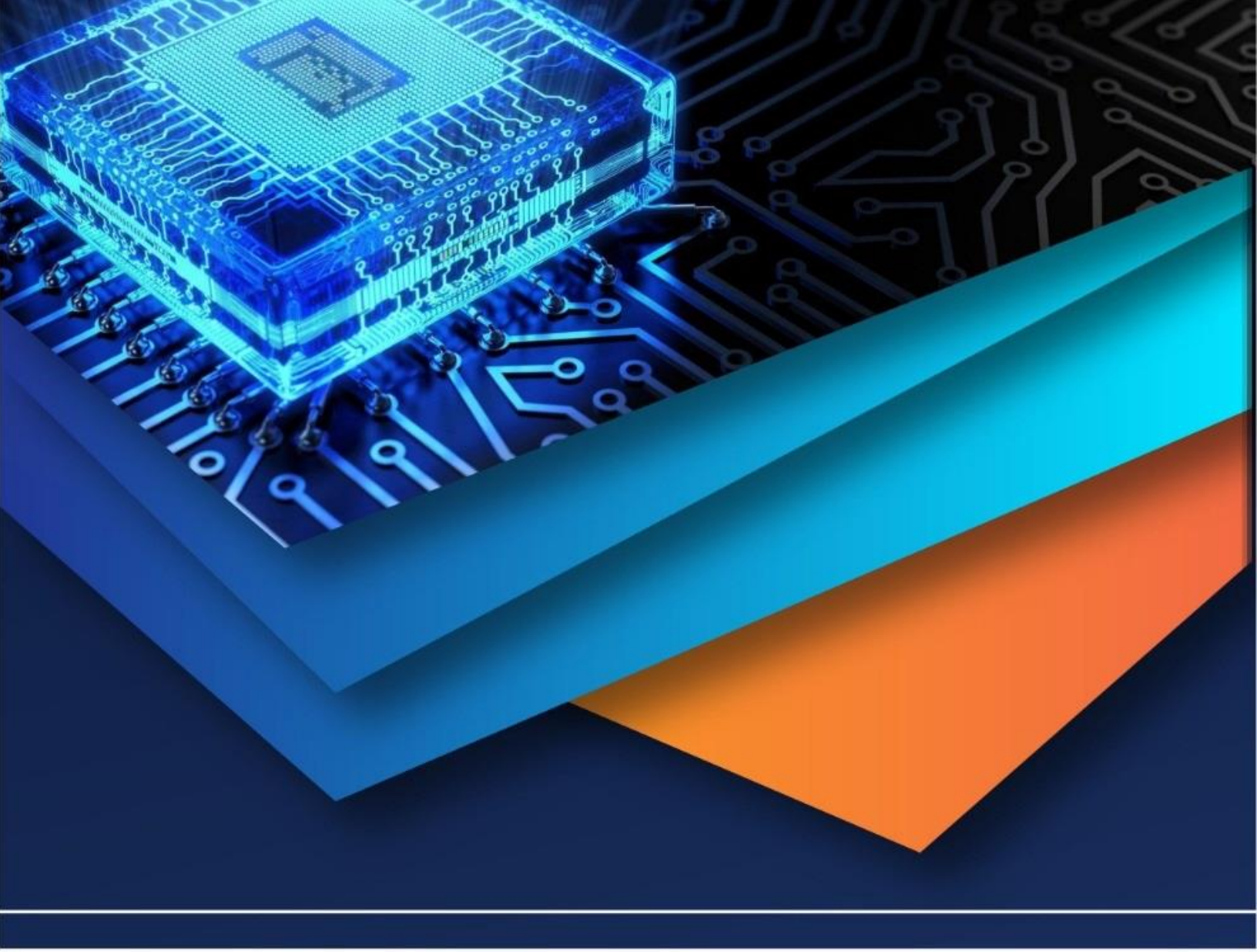
Flow of the DGA method for detecting malicious domain names in communications

V. CONCLUSION

The development of the nation's economy and social fabric as well as how individuals work are all impacted by the Internet. Due to the extensive use of the Internet in many industries, network attacks frequently result in network security issues. Network intrusion detection and alarm correlation technology is becoming more recognised as a critical component of the network security architecture. Machine learning-based network security research has advanced significantly and expanded the possibilities for network security development. Machine learning techniques have constraints in the real network data collecting, message feature extraction, and detection model design links because they rely on publically labelled data sets and empirical expertise. This paper compares machine learning-related methods and the state of machine learning-based network security in order to better understand the application of machine learning-based network security research in real environments. It then investigates the situational awareness method for machine learning-based network security and the dynamic data stream classification method based on judgement feedback. Due to time restrictions, this study does not undertake simulation analysis or further develop the design of many features; these topics will be covered in greater detail in the upcoming work.

REFERENCES

[1] M. G. Li, Y. Xiao, J. F. Chen. et al. A framework for security event mining based on big data. Communication Technology, 2015, 48(03):346-350.
[2] C. Shao, F. Z. Zhang. Research on the application of deep learning in public network security management. Network Security Technology and Applications, 2015(06):89-90.
[3] S. Y. Wang. Research on intrusion detection methods based on machine learning. Journal of Chaohu College, 2015,17(06):25-27.
[4] L. N. Jiang. Machine learning, deep learning and network security technology. China Information Security, 2016(05):94.
[5] K. J. Zhao, L. S. Ge, Y. Liu. et al. Building a scalable network security analysis platform based on Hadoop and Spark. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2016, 44(S1):25-28.
[6] K. Zhu, Q. Zhang. Application of machine learning in network intrusion detection. Data Acquisition and Processing, 2017,32(03):479-488.
[7] X. Zhang. Network intrusion detection based on machine learning algorithms. Modern Electronic Technology, 2018,41(03):124-127.
[8] L. Zhang, Y. Cui, J. Liu. et al. Application of machine learning in cyberspace security research. Journal of Computer Science, 2018,41(09):1943-1975.
[9] J. P. Liu. Network security protection based on machine learning technology. Cyberspace security, 2018,9(09):96-102.
[10] K. R. Liu, D. Li, M. D. Pei. et al. A review on the application of machine learning algorithms in network intrusion detection. Journal of Chifeng College (Natural Science Edition), 2018,34(12):44-46.
[11] D. R. Si, C. Hua, H. G. Yang. et al. A machine learning-based security threat analysis system. Information technology and network security, 2019,38(04):37-41.
[12] Sundaresan K, Dineshkumar T and Pugazhenthi A "An Intrusion Detection System to Prevent the K-Zero Day Attack Propagation" International Journal of Modern Trends in Engineering and Science, ISSN: 2348-3121, Volume 04, issue 03, Feb 2017.
[13] D. L. Zeng, S. Q. Zhang, Q. L. Meng. et al. Research on network intrusion detection based on improved BP neural network. Journal of Shijiazhuang College, 2019,21(03):23-30.
[14] P. Z. Zhu. Real-time network intrusion detection method based on deep learning. Journal of Anyang Institute of Technology, 2019,18(04):48-51.
[15] W. F. Wu, R. F. Li, G. Zeng. et al. A review of cybersecurity research on intelligent networked vehicles. Journal of Communication, 2020,41(06):161-174.
[16] Z. D. Wang, L. Zhang, H. H. Li. A review of machine learning-based intrusion detection system for the Internet of Things. Computer Engineering and Applications, 2021,57(04):18-27.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY