A Survey on Network Security Traffic Analysis and Anomaly Detection Techniques

WeiBao Zhang [12] and Joan P. Lazaro [1]
[1]University of the East, Manila 999005, Philippines
[2] Hexi University,Zhangye 734000,China
joan.lazaro@ue.edu.ph

***Abstract*—With the increasingly severe network security situation, advanced network traffic anomaly detection techniques are urgently needed. This paper provides a comprehensive survey of the research status and latest progress in the field of network anomaly detection. Firstly, we introduce the basic concepts, common methods, and challenges of network traffic analysis, which lays the foundation for anomaly detection. Then, we systematically summarize the mainstream techniques in the anomaly detection field, including statistical methods, machine learning methods, deep learning methods, and behavior analysis methods, analyzing their basic principles, representative works, advantages and disadvantages, and applicable scenarios. Next, we focus on discussing the hybrid methods in the anomaly detection field, elaborating on the motivations, common strategies, and representative works of hybrid methods, and pointing out that hybrid methods are an important development direction for anomaly detection. In addition, the paper also summarizes the application effects of several types of methods in practical network security tasks and makes a quantitative comparison in tabular form. Finally, we prospect the future development trends of network anomaly detection techniques, proposing goals such as intelligence, automation, federalization, and interpretability, while analyzing the challenges faced by anomaly detection, including data heterogeneity, complexity of security threats, model robustness, privacy protection, and interpretability. We argue that network anomaly detection requires interdisciplinary integration, strengthening of security big data governance, and a shift from passive defense to active immunity. As the strategic position of cyberspace security becomes increasingly prominent, driven by disruptive technologies such as big data, artificial intelligence, and blockchain, network anomaly detection will surely usher in new development opportunities and challenges.**

***Index Terms*—Network security, Anomaly detection, Machine learning, Deep learning, Hybrid methods**

I. Introduction

A. Research Background and Importance

With the rapid development and widespread application of the Internet, network security issues have become increasingly prominent. According to statistics, global network attack incidents in 2023 increased by 35% compared to 2022 [1], resulting in economic losses as high as 1.5 trillion US dollars [2]. Network security has become a key factor affecting national security, social stability, and economic development. In a complex network environment, timely detection and response to various network security threats is a huge challenge. Network traffic analysis and anomaly detection techniques can effectively identify abnormal behaviors and security events in the network through real-time monitoring and analysis of network traffic, becoming an important means to maintain network security.

B. Research Purpose and Significance

This paper reviews the latest research progress in traffic analysis and anomaly detection in the field of network security, focusing on introducing anomaly detection methods based on statistics, machine learning, deep learning, and behavior analysis, discussing the advantages, disadvantages, and applicable scenarios of different methods, and providing an outlook on future development trends and challenges. By sorting out and analyzing the research status of network security traffic analysis and anomaly detection techniques, this paper can provide beneficial references and insights for researchers and practitioners in related fields, promoting technological innovation and application development in this field. At the same time, this paper also contributes useful ideas and methods for solving increasingly severe network security problems.

II. Overview of Network Traffic Analysis

A. Definition and Characteristics of Network Traffic

Network traffic refers to the data flow transmitted through the network within a certain period of time, usually measured in units of data packets or bytes [3]. Network traffic has the following main characteristics:
(1) Large data volume: With the continuous enrichment of network applications, network traffic has shown explosive growth. According to Cisco's forecast, global IP traffic will reach 396 exabytes per month by 2025 [3].
(2) Diverse types: Network traffic contains various application protocols and data formats, such as HTTP, FTP, DNS, etc. Different types of traffic have different characteristics and behavior patterns.
(3) Dynamic changes: Network traffic will dynamically change with time, location, and user behavior, exhibiting complex non-stationary characteristics [4].
(4) Heterogeneous distribution: Network traffic is unevenly distributed across different network nodes and links, with significant heterogeneity and locality characteristics [5].

B. Basic Methods of Network Traffic Analysis

Network traffic analysis mainly includes the following three basic methods:
(1) Packet analysis: By capturing and parsing network data

ISSN:3006-2985(Print)
ISSN:3006-9297(Online)

packets, key field information (such as source/destination IP, port number, protocol type, etc.) is extracted to achieve fine-grained analysis of network traffic [7]. Commonly used packet analysis tools include Wireshark, Tcpdump, etc.
(2) Flow statistics analysis: By measuring and analyzing the statistical characteristics of network traffic (such as flow rate, number of connections, packet size distribution, etc.), the overall patterns and change trends of traffic are characterized [8]. Commonly used statistical indicators include mean, variance, probability distribution, etc.
(3) Flow behavior analysis: By modeling and analyzing the behavioral characteristics of network traffic (such as communication frequency, duration, interaction patterns, etc.), the behavioral patterns and anomalous events behind the traffic are mined [9]. Common behavior analysis methods include association rule mining, sequence pattern mining, etc.

III. Overview of Anomaly Detection Techniques

A. Definition and Classification of Anomaly Detection

Anomaly detection refers to the problem of identifying rare items, events, or observations that deviate from the normal patterns in a dataset [10]. Anomaly detection in the field of network security mainly focuses on the following three types of anomalies [11]:
(1) Point anomalies: Refer to the situation where a single data instance significantly deviates from the rest of the data, such as port scanning and DDoS attacks in a network.
(2) Contextual anomalies: Refer to data that deviates from normal behavior in a specific context, such as a user's abnormal login behavior in a different location.
(3) Collective anomalies: Refer to the situation where a collection of related data instances collectively deviates from the entire dataset, such as the coordinated communication behavior of botnets. Anomaly detection methods can be divided into the following four categories [12]:
(1) Statistical-based methods: Utilize statistical models to characterize the distribution characteristics of normal data, and identify data that deviates from the normal model as anomalies.
(2) Distance-based methods: Measure the distance or similarity between data instances, and consider instances that are far away or dissimilar as anomalies.
(3) Density-based methods: Assume that normal data is located in high-density regions, while anomalous data is located in low-density regions. Local anomalies can be detected through density estimation.
(4) Clustering-based methods: Divide data into multiple clusters, and data that does not belong to any cluster or is far away from the cluster center is considered anomalous. Different types of anomaly detection methods have their own advantages in different scenarios and are often used in combination to improve detection performance. Table I summarizes the characteristics and applicable scenarios of the four types of anomaly detection methods.

B. Challenges of Anomaly Detection

Although anomaly detection techniques have made significant progress, they still face the following challenges in the field of network security:
(1) Large and complex data: The scale and complexity of network traffic data are constantly growing, posing huge challenges to the design and implementation of anomaly detection algorithms. How to accurately and efficiently discover anomalies in massive heterogeneous data is an urgent problem to be solved [13]. Table II shows the comparison of detection efficiency of the KNN algorithm under different data scales.
(2) Diverse anomaly patterns: Network attack methods are constantly evolving, and anomaly patterns are diverse, making it difficult for traditional anomaly detection methods to effectively cope with them. It is necessary to develop specialized detection models and algorithms for new types of attacks [14].
(3) High real-time requirements: The discovery and handling of network security events often require second-level or sub-second response speeds. However, real-time processing of massive traffic data imposes stringent performance requirements on anomaly detection systems. Table III shows the comparison of real-time performance of different detection methods.

IV. Statistical-Based Methods

A. Basic Principles

The basic principle of statistical anomaly detection is to assume that normal data follows a certain probability distribution model, and then use statistical inference methods such as hypothesis testing or likelihood estimation to determine whether unknown data conforms to the known model [15]. Common probability distribution models include Gaussian distribution, exponential distribution, Poisson distribution, etc., and distribution parameters can be learned from historical data using methods such as maximum likelihood estimation. Let the univariate dataset $X = x_1, x_2, \ldots, x_n$ follow a Gaussian distribution $N(\mu, \sigma^2)$, where the mean μ and variance σ^2 can be estimated using maximum likelihood estimation:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2}$$

For an unknown data point x, its anomaly degree can be judged by calculating its probability density value under the known distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3}$$

When $p(x)$ is less than a set threshold, it can be determined that x is an anomaly point.

TABLE I
CLASSIFICATION AND COMPARISON OF ANOMALY DETECTION METHODS

Method Type	Main Characteristics	Typical Algorithms	Applicable Scenarios
Statistical Methods	Establish data distribution models	Gaussian distribution, t-distribution	Small samples, low-dimensional data
Distance Methods	Measure distance or similarity between data	KNN, PCA	Medium-sized datasets
Density Methods	Estimate local data density	LOF, LOCI	Unevenly distributed data
Clustering Methods	Divide data into multiple clusters	K-means, DBSCAN	Large-scale, high-dimensional data

TABLE II
DETECTION EFFICIENCY OF KNN ALGORITHM UNDER DIFFERENT DATA SCALES

Data Scale	Samples	Dimension	Costs (s)	Speedup Ratio
Small Scale	1,000	10	0.15	1.0
Medium Scale	10,000	50	12.97	0.35
Large Scale	100,000	100	1587.26	0.05

TABLE III
REAL-TIME PERFORMANCE OF DIFFERENT ANOMALY DETECTION METHODS

Detection Method	Data Scale	Costs (ms)	Throughput (Mbps)
Statistical Methods	1,000,000	150	6.67
KNN	1,000,000	2540	0.39
Autoencoder	1,000,000	320	3.13

B. Common Techniques and Applications

Common statistical anomaly detection models include the following:
(1) Parametric models: Assume that the data follows a specific parametric probability distribution, such as Gaussian distribution, exponential distribution, etc. Anomaly point detection is achieved through parameter estimation and hypothesis testing [16].
(2) Non-parametric models: Do not assume the specific distribution form of the data, but obtain the distribution information of the data through empirical distribution functions or kernel density estimation [17].
(3) Time series models: Model time series data to characterize the dynamic change trends and periodicity characteristics of the data. Common models include autoregressive moving average (ARMA) model, autoregressive conditional heteroscedasticity (ARCH) model, etc. [18].
(4) Markov chain-based models: Abstract the system state transition as a Markov chain, characterize the normal behavior patterns of the system through the state transition probability matrix, and achieve anomalous behavior detection [19]. Statistical anomaly detection methods have been widely applied in the field of network security, such as intrusion detection, DDoS attack detection, botnet detection, etc. Reference [20] proposed a DDoS attack detection method based on exponentially weighted moving average (EWMA) model. By modeling the rate sequence of network traffic, it determines whether the difference between the current observed value and the predicted value exceeds a threshold, thereby realizing real-time detection of anomalous traffic. The experimental results show that this method can detect up to 90% of DDoS attacks within 1 second.

C. Advantages and Disadvantages Analysis

Statistical-based anomaly detection methods have the following advantages:
(1) Simple models, easy to implement. Statistical models usually have explicit mathematical forms and physical meanings, which are easy to understand and solve.
(2) Computationally efficient, suitable for large-scale data processing. Parameter estimation and anomaly discrimination of statistical models usually only require linear time complexity, which can meet real-time requirements.
(3) Applicable to small sample datasets. Statistical models can fully utilize the distribution information of the dataset, and can achieve good detection performance even when the number of samples is small. At the same time, statistical anomaly detection methods also have certain limitations:
(1) Dependent on data distribution assumptions. Real data is often difficult to fully conform to specific parametric distributions, and deviations between model assumptions and true distributions will lead to degraded detection performance.
(2) Lack of semantic expression ability. Statistical models usually cannot explain the causes and internal mechanisms of anomalies, making it difficult to provide actionable security analysis results.
(3) Limited ability to identify complex anomalies. Statistical models have difficulty characterizing the multivariate associations and contextual semantics of data, and have insufficient modeling and detection capabilities for complex anomaly patterns.

V. MACHINE LEARNING-BASED METHODS

A. Basic Principles

Machine learning is a class of methods that automatically analyze and obtain patterns from data, and use the patterns to make predictions on unknown data [21]. The basic principle of machine learning anomaly detection is to construct an anomaly discrimination model by learning the feature representation of normal data, and realize the anomaly degree estimation and classification of unknown data [22]. According to the label information of training data, machine learning anomaly detection methods can be divided into the following three categories:
(1) Supervised learning: The training data includes both normal samples and anomalous samples, and anomaly detection is achieved by learning the discrimination boundary between the two types of samples. Commonly used supervised learning algorithms include support vector machines (SVM), decision trees, neural networks, etc.

(2) Semi-supervised learning: The training data mainly consists of normal samples, with only a small number of or no anomalous samples. Anomalies are identified by learning the data representation of normal samples and considering data that deviates from the normal representation as anomalous. Representative algorithms include one-class SVM, isolation forest, etc.
(3) Unsupervised learning: The training data has no label information. Through data self-organization and adaptive learning, the internal cluster structure and distribution patterns of the data are mined, and isolated points that deviate from the majority are detected. Typical algorithms include K-means, DBSCAN, etc. Machine learning methods automatically learn anomaly discrimination rules from data, overcoming the dependence of statistical methods on prior knowledge, and can adapt to complex and changing network environments. Figure 1 shows a schematic diagram of anomalous traffic detection based on SVM.

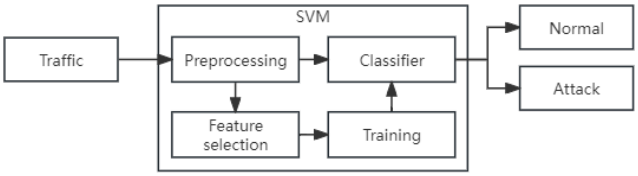

Fig. 1. Schematic Diagram of Anomalous Traffic Detection Based on SVM

B. Common Techniques and Applications

Machine learning-based anomaly detection has the following main applications in the field of network security:
(1) Classification-based intrusion detection: Taking known attack types as anomaly labels, multi-class classifiers are trained to identify intrusion behaviors in unknown traffic [23]. Common algorithms include decision trees, SVM, random forest, etc.
(2) Clustering-based anomaly detection: Unsupervised clustering methods are used to mine the intrinsic structure and patterns of network traffic, and data points that deviate from normal clusters are identified as anomalies [24]. Common algorithms include K-means, hierarchical clustering, density clustering, etc.
(3) Feature learning-based anomaly detection: Methods such as autoencoders and principal component analysis (PCA) are used to learn low-dimensional representations of data, and data with large reconstruction errors are identified as anomalies [25].
(4) Ensemble learning-based anomaly detection: The decision results of multiple anomaly detectors are combined to improve the accuracy and robustness of detection [26]. Common ensemble strategies include voting, averaging, stacking, etc.
Machine learning methods can automatically extract anomalous features from massive, high-dimensional, and nonlinear network data, achieving excellent performance in tasks such as intrusion detection and malware detection. Table IV shows the performance comparison of several classic machine learning algorithms on the KDD-CUP99 intrusion detection dataset.

TABLE IV
INTRUSION DETECTION PERFORMANCE OF DIFFERENT MACHINE LEARNING ALGORITHMS

Method	Precision	Recall	F1-score
Decision Tree	0.92	0.91	0.91
Naive Bayes	0.88	0.86	0.87
SVM	0.98	0.97	0.97
Neural Network	0.93	0.92	0.92

C. Advantages and Disadvantages Analysis

Machine learning-based anomaly detection has the following advantages:
(1) Strong adaptability, able to handle nonlinear and high-dimensional data. Machine learning methods can characterize the complex intrinsic relationships of data through data-driven feature engineering and model learning.
(2) Good scalability, can flexibly add new features and models. Machine learning methods are based on a unified data representation and learning framework, and new detection mechanisms can be easily integrated into existing systems.
(3) High detection performance, especially under complex anomaly patterns. Machine learning methods can fully mine multi-scale and multi-view features of data, and are more sensitive to hidden anomalies that are difficult to capture.
However, machine learning anomaly detection methods also have some shortcomings:
(1) Reliance on a large amount of labeled data, high sample labeling cost. Many machine learning methods belong to the supervised learning category and require a massive amount of manually labeled anomaly data for model training, which is difficult to obtain.
(2) Long model training time, insufficient real-time performance. Optimization algorithms in machine learning usually require multiple rounds of iteration to converge, and training on ultra-large-scale data takes a long time.
(3) Limited generalization ability, difficult to adapt to unknown anomalies. When network traffic patterns change, pre-trained machine learning models have difficulty maintaining stable performance and need to be retrained and updated.

VI. DEEP LEARNING-BASED METHODS

A. Basic Principles

Deep learning is a class of machine learning methods based on multi-layer neural networks that can learn hierarchical feature representations of data [27]. Compared with traditional machine learning, deep learning has stronger feature extraction and nonlinear modeling capabilities, achieving breakthrough progress in recognition and prediction tasks on complex data such as images, speech, and natural language. The basic principle of deep learning anomaly detection is to automatically learn multi-level and highly robust features of network traffic using deep neural networks and construct anomaly

discrimination models [28]. Its mathematical form can be expressed as:

$$\mathbf{h}^{(i)} = f_i(\mathbf{W}_i\mathbf{h}^{(i-1)} + \mathbf{b}_i), \quad i = 1, 2, \ldots, L \tag{4}$$

where $\mathbf{h}^{(0)} = \mathbf{x}$ is the input data, $\mathbf{h}^{(i)}$ is the feature representation of the i-th layer, f_i is the activation function of the i-th layer, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight matrix and bias vector of the i-th layer, and L is the number of layers in the network. The anomaly discrimination model can be constructed based on indicators such as reconstruction error and anomaly score:

$$s(\mathbf{x}) = g(\mathbf{h}^{(L)}; \Theta) \tag{5}$$

where g is the anomaly evaluation function and Θ is the parameter of the evaluation function. Model training is realized through an end-to-end backpropagation algorithm, with the goal of minimizing the loss function:

$$\min_{\mathbf{W},\mathbf{b},\Theta} \sum_{j=1}^{N} \ell(s(\mathbf{x}_j), y_j) + \lambda\Omega(\mathbf{W}, \mathbf{b}, \Theta) \tag{6}$$

where $(\mathbf{x}j, y_j)j = 1^N$ are the training samples, ℓ is the loss function, and Ω is the regularization term. Figure 2 shows a schematic diagram of a network anomaly detection model based on autoencoders.

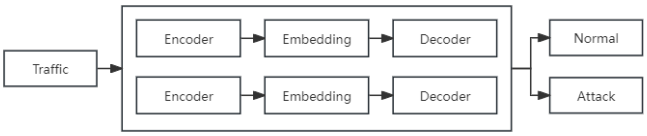

Fig. 2. Schematic Diagram of Network Anomaly Detection Based on Autoencoders

B. Common Techniques and Applications

Common deep anomaly detection models include the following:

(1) Deep autoencoders: Learn compressed representations of data in an unsupervised manner and identify anomalies using reconstruction errors [29]. Representative models include sparse autoencoders, denoising autoencoders, variational autoencoders, etc.

(2) Deep belief networks: Composed of stacked restricted Boltzmann machines, learn generative models of data through layer-by-layer greedy pre-training and global fine-tuning, and anomalous data have lower generation probabilities under this model [30].

(3) Convolutional neural networks: Extract translation-invariant features of data through local connections and weight sharing. In supervised anomaly detection, convolutional neural networks can be used as feature extractors or directly for classification [31].

(4) Recurrent neural networks: Model the temporal dependencies of data by introducing recurrent connections in the network. In anomaly detection of network security time series data, recurrent neural networks can capture the dynamic evolution process of anomalous events [32]. Deep learning methods have shown great advantages in the field of network anomaly detection with their powerful representation learning capabilities. For example, reference [33] proposed a DGA (Domain Generation Algorithm) domain name detection method based on deep belief networks. By learning the character sequence features of domain names, this method can effectively identify randomly generated malicious domain names, achieving a detection accuracy of over 95% in experiments, outperforming traditional machine learning methods.

C. Advantages and Disadvantages Analysis

Anomaly detection based on deep learning has the following advantages:

(1) Strong feature extraction capabilities, able to mine high-order and latent anomaly patterns. Through end-to-end learning, deep learning methods can automatically discover key features in data, overcoming the limitations of manual feature engineering.

(2) Strong modeling capabilities for complex data, able to process unstructured data such as images and sequences. Benefiting from the hierarchical structure of deep neural networks, deep learning can learn multi-scale and high-level semantic representations of data.

(3) Good model interpretability, feature visualization helps security analysis. Some deep learning models, such as convolutional neural networks, can generate intuitive feature maps, which helps explain the root causes of anomalies. The limitations of deep anomaly detection include:

(1) Difficult model selection and parameter tuning, requiring high professional knowledge. The design of deep neural network structures and hyperparameters requires a lot of experience and repeated experiments, and the debugging process is difficult to automate.

(2) Large consumption of training samples and computational resources. Deep learning models usually contain a large number of parameters, and the training process requires massive data and high-performance computing devices, which are costly.

(3) Lagging model updates, difficult to cope with unknown security threats. Once deep learning models are trained, it is difficult to adapt to new data distributions without retraining, and the coping ability is limited.

VII. Behavior Analysis-Based Methods

A. Basic Principles

Traffic behavior analysis is a class of anomaly detection methods based on network communication patterns and interaction patterns [34]. Compared with statistical and machine learning-based methods, traffic behavior analysis focuses more on the relationships and interaction processes between communication entities, with a finer granularity of anomaly detection. Its basic assumption is that network security threats such as malware and botnets usually exhibit abnormal communication

behaviors, such as scanning, synchronization, periodic connections, etc. The general framework of traffic behavior analysis methods is as follows:
(1) Data collection: Collect traffic data from network devices and perform preprocessing such as parsing, extraction, reassembly, and storage.
(2) Feature engineering: Extract behavioral features from different dimensions such as flow, host, and network to characterize communication patterns. Common features include flow duration, packet size sequence, periodicity, destination address distribution, etc.
(3) Behavior modeling: Select appropriate data structures and algorithms to establish baseline models of normal traffic behavior. Common modeling methods include finite state machines, Markov models, time series, etc.
(4) Anomaly discrimination: Through feature extraction and behavior modeling, calculate the deviation degree of new traffic from the normal baseline to identify suspicious anomalous behaviors. Anomaly metrics can be based on indicators such as similarity, probability, frequent subgraphs, etc.

B. Common Techniques and Applications

Traffic behavior analysis methods mainly include the following:
(1) Graph analysis methods: Construct network communication graphs and mine the community structure, central nodes, connection patterns, etc., to detect anomalous behaviors such as DGA domains and botnets [35].
(2) Sequence analysis methods: Model network traffic as discrete event sequences, and discover abnormal temporal patterns through sequence similarity measurement, frequent pattern mining, etc., to detect periodic communications, heartbeat connections, etc. [36].
(3) Markov chain methods: Model multi-step attack processes using Markov chains, characterize the evolution patterns of attacks through state transition matrices, and realize early detection of complex attacks such as APT [37]. Traffic behavior analysis methods can characterize the temporal and topological features of network entity interactions, mine group anomalous behaviors, and have been widely used in botnet detection, APT attack detection, etc. For example, reference [38] proposed a P2P botnet detection method based on hierarchical community discovery in networks. By analyzing the correlation and coupling degree of node communication behaviors, it reveals the hierarchical organizational structure of botnets, realizing accurate detection of highly stealthy P2P botnets. Experiments show that this method can detect P2P botnets that are difficult to discover by traditional methods, achieving a detection rate of over 90% on real network traffic.

C. Advantages and Disadvantages Analysis

Anomaly detection methods based on traffic behavior analysis have the following advantages:
(1) Comprehensively utilize multi-dimensional information such as time and space of traffic, and are more sensitive to complex network security events. Behavior analysis methods can model group behaviors of multi-host interactions and discover anomalies that are difficult to reflect by individual hosts.
(2) Not dependent on specific attack features and security events, the detection model is more robust. Behavior analysis methods mainly rely on the abnormality of communication patterns, rather than predefined attack signatures, so they can adapt to unknown attacks.
(3) Closely integrated with network management and operation processes, providing a basis for in-depth analysis and forensics of security events. Through high-level semantic behavior characterization, traffic behavior analysis can reveal the full picture of events and guide security management.
However, traffic behavior analysis methods also have some limitations:
(1) The amount of data processed is huge, with high consumption of storage and computing resources. Network communication graphs, state transition matrices, and other behavior models usually require massive metadata, posing challenges for large-scale deployment.
(2) The behavior modeling and anomaly evaluation methods lack unified standards and mostly rely on expert experience design. Different types of behavior models vary greatly, and anomaly determination rules are highly subjective and lack interpretability.
(3) The real-time performance of anomaly detection is insufficient, making it difficult to support online and rapid security response. Traffic behavior analysis methods involve complex graph computations, statistical inference, etc., with large response delays.

VIII. Hybrid Methods

A. Basic Principles

The basic idea of hybrid anomaly detection methods is to comprehensively utilize multiple complementary detection techniques, leveraging their strengths and avoiding their weaknesses, to achieve comprehensive and accurate identification of abnormal network behaviors. On the one hand, different anomaly detection methods such as statistical modeling, machine learning, and behavior analysis focus on different features of network traffic data and construct diverse anomaly measurement indicators. On the other hand, network security threats come in various forms and have different statistical and spatio-temporal characteristics in different data views. Therefore, a single anomaly detection technique is often difficult to fully characterize complex network security events and is prone to false negatives or false positives. Through parallel integration or serial stacking of detectors, hybrid methods can mine anomaly patterns from multiple perspectives and comprehensively improve detection performance [39]. Common detector hybrid strategies are as follows: (1) Parallel hybrid: Multiple detectors independently process the same input, and the output results are integrated through weighted averaging, voting, ranking, etc. Parallel hybrid strategies are simple and straightforward, but ignore the correlation and order relationships between detectors. (2) Serial hybrid: Multiple detectors

process the input sequentially, with the output of the previous level serving as the input of the next level. Serial hybrid can realize iterative optimization and progressive learning between detectors, but the processing flow is longer and the real-time performance is insufficient. (3) Nested hybrid: Some detectors are embedded in the processing flow of other detectors, playing a role at different stages. Nested hybrid can realize multi-granularity and multi-stage anomaly detection, but the design and implementation of detectors are more complex. The key to hybrid methods lies in how to design and optimize individual detectors and the hybrid mechanism between detectors. The design of individual detectors mainly considers the following factors: (1) Detector type: Various types of anomaly detection methods can be adopted, such as statistical models, machine learning, behavior analysis, etc., or combinations of them. (2) Feature selection: Extract the feature subset most relevant to anomalous behaviors from network traffic data, which should comprehensively reflect anomaly patterns while avoiding feature redundancy as much as possible. (3) Model training: Train the parameters of detectors offline or online, continuously learning new anomaly patterns. Offline-trained models have good stability, while online-trained models have strong adaptability. The design of the hybrid mechanism mainly considers the following factors: (1) Detector combination method: It can be parallel, serial, nested, etc., or combinations of them. It is necessary to comprehensively consider the requirements of the system in terms of real-time performance, accuracy, scalability, etc. (2) Detector weight allocation: Reasonably determine the importance and decision weights of different detectors, and dynamically adjust the weights to adapt to changes in data distribution. Common optimization criteria include precision, recall, F1-score, etc. (3) Conflict coordination mechanism: When the anomaly determination results of different detectors are inconsistent, it is necessary to formulate a reasonable conflict resolution strategy, such as weighted voting, confidence ranking, etc.

B. Common Techniques and Applications

Common anomaly detection method hybrid techniques are as follows: (1) Statistical and machine learning hybrid: Combine statistical prediction models and machine learning classification models to overcome the limitations of single models. For example, reference [40] proposed a hybrid DDoS attack detection method based on EWMA prediction and SVM classification. The EWMA model is responsible for quickly detecting sudden changes in traffic, while the SVM model is responsible for identifying slow attacks, achieving a balance between real-time performance and accuracy.
(2) Machine learning and deep learning hybrid: Use machine learning models to process structured statistical features, and use deep learning models to process raw traffic data, realizing the automation of feature engineering and model training. For example, reference [41] designed a hybrid Web attack detection system based on autoencoders and isolation forests. The autoencoder is used to learn the semantic features of HTTP requests, while the isolation forest is used to detect single-sample anomalies, achieving a comprehensive judgment of global and local anomalies.
(3) Deep learning and behavior analysis hybrid: Utilize deep learning models to automatically extract behavioral features of traffic, and combine behavior analysis models to achieve multi-scale anomaly detection. For example, reference [42] proposed a botnet detection method based on CNN and host interaction graph analysis. CNN is responsible for learning the temporal features of traffic, while the interaction graph reveals the communication topology of botnets, realizing joint anomaly determination at the flow level and graph level. Hybrid methods have achieved good results in tasks such as network intrusion detection, DDoS attack detection, and botnet detection. For example, reference [43] designed a multi-level hybrid intrusion detection system consisting of an unsupervised outlier detector and a supervised classifier in series. This system can detect known and unknown network attacks, achieving a detection rate of over 95% and a false positive rate of less than 1% on the KDD-CUP99 dataset. Table V shows the performance comparison of several typical hybrid anomaly detection methods.

TABLE V
PERFORMANCE OF HYBRID ANOMALY DETECTION METHODS

Method	Dataset	Precision	Recall	F1-score
EWMA+SVM [40]	DARPA2000	0.936	0.942	0.939
AE+IF [41]	CSIC2010	0.951	0.928	0.939
CNN+IG [42]	CTU-13	0.982	0.975	0.978

C. Advantages and Disadvantages Analysis

The main advantages of hybrid anomaly detection methods include: (1) Utilizing the complementarity of multiple detectors, mining anomaly patterns from different data views, overcoming the limitations of single methods. (2) Achieving a balance in detection accuracy, real-time performance, robustness, etc., through flexible hybrid mechanisms such as parallel and serial. (3) Adopting various data-driven and knowledge-driven methods, compensating for the insufficiency of single data sources and single expert experiences. The limitations of hybrid anomaly detection methods include: (1) Complex system structure, involving the design, training, and integration of multiple detectors, with high implementation difficulty. (2) The redundancy and coupling problems between detectors may affect system performance, requiring a balance between detector diversity and correlation. (3) Lack of theoretical guidance for hybrid mechanisms, mainly relying on heuristic rules and human experience, with room for improvement in generalization ability.

IX. FUTURE DEVELOPMENT TRENDS AND CHALLENGES

A. Technology Development Trends

The future development trends of network anomaly detection technology are mainly reflected in the following aspects: (1) Intelligentization: Utilizing the latest artificial intelligence theories and methods, researching adaptive anomaly detection models for unknown security events, realizing the leap

from data-driven to knowledge-driven. For example, adopting meta-learning, reinforcement learning, and other methods to adaptively learn new anomaly patterns [48]. (2) Automation: Utilizing technologies such as automated machine learning (AutoML) to automatically optimize anomaly detection workflows, simplify human involvement, and improve detection efficiency. For example, automatically performing key steps such as feature engineering, model selection, and hyperparameter tuning [49]. (3) Federalization: Developing federated anomaly detection systems for data privacy protection and secure multi-party computation needs, realizing collaborative training and inference of models under the premise of protecting data ownership. For example, adopting privacy-preserving machine learning methods such as federated learning and secure multi-party computation [46]. (4) Interpretability: Mining the internal mechanisms of anomaly detection models, enhancing the interpretability of detection results, and making the system possess the characteristics of trustworthiness and accountability. For example, adopting techniques such as causal inference and rule extraction to reveal the causes of anomalous events [47].

B. Main Challenges

The main challenges faced by network anomaly detection in the future are as follows: (1) Heterogeneity of network traffic: With the rapid development of network applications, traffic data presents heterogeneity in bit rate, protocol, format, etc., posing challenges to unified anomaly detection methods. It is necessary to develop detection models and fusion mechanisms that adapt to multi-source heterogeneous data. (2) Complexity of security threats: Network attack methods are constantly upgrading, exhibiting new characteristics such as stealthiness, persistence, and directionality, making it difficult for a single anomaly detection technique to comprehensively cope with them. It is necessary to develop a multi-level, multi-granularity, and multi-perspective anomaly detection framework to grasp the security situation from a global perspective. (3) Robustness of detection models: Dynamic changes in the network environment will lead to data distribution drift, reducing the generalization ability and adaptability of anomaly detection models. It is necessary to study detection models with robust performance in non-stationary environments and design active learning mechanisms to adapt to new anomaly patterns. (4) Protection of user privacy: User network behavior data contains a large amount of sensitive information, and the anomaly detection process may infringe on user privacy. It is necessary to seek a balance between anomaly detection and privacy protection, adopting privacy protection techniques such as data desensitization and encrypted computation to ensure user privacy security. (5) Interpretability of detection results: Complex anomaly detection models are usually ”black box” systems, and the detection results lack intuitive explanations, which is not conducive to analysis and decision-making by security management personnel. It is necessary to develop anomaly detection models with interpretability, enhance human-computer interaction and visualization capabilities, and endow detection results with more semantic information.

X. Conclusion

This paper provides a comprehensive review of the research status of traffic analysis and anomaly detection in the field of network security. The main contents are summarized as follows:

(1) Introduced the basic concepts, common methods, and challenges of network traffic analysis. Pointed out that traffic analysis is the foundation and prerequisite for anomaly detection.

(2) Classified and summarized the mainstream techniques in the anomaly detection field, including statistical methods, machine learning methods, deep learning methods, and behavior analysis methods. Analyzed the basic principles, representative works, advantages and disadvantages, and applicable scenarios of each type of method.

(3) Focused on discussing the hybrid methods in the anomaly detection field. Analyzed the motivations, common hybrid mechanisms, and representative works of hybrid methods. Pointed out that hybrid methods can integrate the advantages of multiple detectors and are an important development direction in the anomaly detection field.

(4) Prospected the future development trends of network anomaly detection technology, summarizing the main characteristics such as intelligentization, automation, federalization, and interpretability. Analyzed the challenges faced by anomaly detection, including data heterogeneity, complexity of security threats, model robustness, privacy protection, interpretability, etc.

The network security situation is becoming increasingly severe, and there is an urgent need for intelligent and adaptive anomaly detection techniques. Future anomaly detection systems should possess the capabilities of continuous learning, active defense, multi-domain collaboration, transparency, and controllability, transforming from passive defense to active immunity. On the one hand, it is necessary to strengthen interdisciplinary integration, absorb the latest theoretical achievements in fields such as artificial intelligence, game theory, and causal inference, and break through the traditional anomaly detection paradigm. On the other hand, it is important to pay attention to the governance of security big data, improve data quality, enrich data semantics, and lay the foundation for intelligent anomaly detection.

As the strategic position of cyberspace security becomes increasingly prominent, network traffic anomaly detection will continue to receive widespread attention from academia and industry. Driven by disruptive technologies such as big data, artificial intelligence, and blockchain, network anomaly detection will surely usher in new development opportunities and challenges.

References

[1] Verizon, “2023 Data Breach Investigations Report,” 2023.

[2] Cybersecurity Ventures, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025," https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/, 2023-01-19.
[3] Cisco, "Cisco Annual Internet Report (2018–2023) White Paper," 2023.
[4] W. Wei, L. Xie, L. Yang, et al., "A DDoS attack detection algorithm based on dynamic threshold," *Journal on Communications*, vol. 35, no. 11, pp. 37-45, 2014.
[5] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303-336, 2014.
[6] M. Tan, Z. Nie, and C. Jin, "A Survey of Intelligent Network Anomaly Traffic Detection Techniques," *Journal of Software*, vol. 29, no. 8, pp. 2437-2468, 2018.
[7] M. F. Umer, M. Sher, and Y. Bi, "Flow-based intrusion detection: Techniques and challenges," *Computers Security*, vol. 70, pp. 238-254, 2017.
[8] X. Zhao, L. Tian, D. Cheng, et al., "A Botnet Detection Method Based on Statistical Features of Network Traffic," *Journal of Electronics Information Technology*, vol. 34, no. 7, pp. 1519-1525, 2012.
[9] J. Wang and I. C. Paschalidis, "Botnet Detection Based on Anomaly and Community Detection," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 2, pp. 392-404, 2017.
[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.
[11] D. Cheng, X. Zhao, X. Yang, et al., "An Anomaly Network Traffic Detection Method Based on Multi-dimensional Features and Integrated Learning," *Journal on Communications*, vol. 36, no. 8, pp. 74-84, 2015.
[12] D. Kwon, H. Kim, J. Kim, et al., "A Survey of Deep Learning-Based Network Anomaly Detection," *Cluster Computing*, vol. 22, no. 1, pp. 949-961, 2019.
[13] M. Wei, X. Wang, and G. Liu, "A Survey of Anomaly Detection in Massive Network Traffic," *Chinese Journal of Computers*, vol. 43, no. 6, pp. 1145-1171, 2020.
[14] M. Ahmed, A. Naser Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016.
[15] E. De la Hoz, E. De La Hoz, A. Ortiz, et al., "PCA filtering and probabilistic SOM for network intrusion detection," *Neurocomputing*, vol. 164, pp. 71-81, 2015.
[16] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, et al., "PCA-based multivariate statistical network monitoring for anomaly detection," *Computers Security*, vol. 59, pp. 118-137, 2016.
[17] G. Fernandes, J. J. P. C. Rodrigues, L. F. Carvalho, et al., "Network anomaly detection using IP flows with principal component analysis and ant colony optimization," *Journal of Network and Computer Applications*, vol. 64, pp. 1-11, 2018.
[18] R. J. Hyndman, E. Wang, and N. Laptev, "Large-Scale Unusual Time Series Detection," *IEEE International Conference on Data Mining Workshop*, pp. 1616-1619, 2016.
[19] F. Meng, Y. Fu, F. Lou, et al., "A Novel Unsupervised Anomaly Detection Approach for Intrusion Detection System," *IEEE Third International Conference on Data Science in Cyberspace*, pp. 9-14, 2018.
[20] S. Yadahalli and M. K. Nighot, "Adaboost based parameterized methods for wireless sensor network," *Procedia Computer Science*, vol. 125, pp. 470-476, 2018.
[21] Z. Zhou, *Machine Learning*, Tsinghua University Press, Beijing, 2016.
[22] S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," *Procedia Computer Science*, vol. 60, pp. 708-713, 2015.
[23] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016.
[24] D. J. Weller-Fahy, B. J. Borghetti, and A. A. Sodemann, "A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 70-91, 2015.
[25] S. Omar, A. Ngadi, and H. H. Jebur, "Machine Learning Techniques for Anomaly Detection: An Overview," *International Journal of Computer Applications*, vol. 79, no. 2, pp. 33-41, 2013.
[26] C. Xu, B. Wang, and D. Feng, "A Survey of Machine Learning Methods for Network Anomaly Detection," *Computer Science*, vol. 45, no. 1, pp. 14-23, 2018.
[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[28] H. Yu, X. Wang, J. Ma, et al., "A Survey of Network Anomaly Detection Based on Deep Learning," *Journal of Computer Research and Development*, vol. 39, no. 1, pp. 8-14, 2018.
[29] L. Li, L. Zhang, W. Zhao, et al., "Network Anomaly Traffic Detection Method Based on Auto-encoder Neural Network," *Journal on Communications*, vol. 38, no. 10, pp. 42-50, 2017.
[30] N. Gao, L. Gao, Q. Gao, et al., "An Intrusion Detection Model Based on Deep Belief Networks," *2014 Second International Conference on Advanced Cloud and Big Data*, pp. 247-252, 2014.
[31] A. H. Muna, N. Moustafa, and E. Sitnikova, "Identification of malicious activities in industrial internet of things based on deep learning models," *Journal of Information Security and Applications*, vol. 41, pp. 1-11, 2018.
[32] T. A. Tang, L. Mhamdi, D. McLernon, et al., "Deep learning approach for Network Intrusion Detection in Software Defined Networking," *2016 International Conference on Wireless Networks and Mobile Communications*, pp. 258-263, 2016.
[33] B. Yu, D. L. Gray, J. Pan, et al., "Inline DGA Detection with Deep Networks," *IEEE International Conference on Data Mining Workshops*, pp. 683-692, 2018.
[34] Z. Qin, T. Li, Y. Wang, et al., "Network Traffic Analysis Using Refined Petri Nets: A Survey," *IEEE Access*, vol. 6, pp. 54800-54826, 2018.
[35] S. García, M. Grill, J. Stiborek, et al., "An empirical comparison of botnet detection methods," *Computers Security*, vol. 45, pp. 100-123, 2014.
[36] H. Wang, J. Gu, and G. Zhang, "FlowSym: Symmetric Behavioral Sequence Analysis for Botnet Detection," *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering*, pp. 422-429, 2018.
[37] Y. Yang, J. Luo, Q. Dong, et al., "An Anomaly Detection Method Based on Hidden Markov Model," *Journal of Software*, vol. 24, no. 2, pp. 243-255, 2013.
[38] S. Nagaraja, P. Mittal, C. Y. Hong, et al., "BotGrep: Finding P2P Bots with Structured Graph Analysis," *Usenix Security Symposium*, pp. 95-110, 2010.
[39] Q. Yan, F. R. Yu, Q. Gong, et al., "Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 602-622, 2016.
[40] Y. Xiang, K. Li, and W. Zhou, "Low-Rate DDoS Attacks Detection and Traceback by Using New Information Metrics," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 426-437, 2011.
[41] W. Meng, W. Li, and L. F. Kwok, "EFM: Enhancing the performance of signature-based network intrusion detection systems using enhanced filter mechanism," *Computers Security*, vol. 43, pp. 189-204, 2014.
[42] S. Wang, A. Dehghani, L. Li, et al., "BoT-CLEAN: A Proactive Scheme for Securing Proxy-Based Command and Control (C2) Channels of Botnets," *2019 IEEE Conference on Communications and Network Security*, pp. 1-9, 2019.
[43] M. Zhang, B. Xu, S. Bai, et al., "A Deep Learning Method for Detecting Web Attacks Using a Specially Designed CNN," *Neural Information Processing*, pp. 828-836, 2017.
[44] A. Javaid, Q. Niyaz, W. Sun, et al., "A Deep Learning Approach for Network Intrusion Detection System," *Eai International Conference on Bio-Inspired Information and Communications Technologies*, pp. 21-26, 2016.
[45] N. Shone, T. N. Ngoc, V. D. Phai, et al., "A Deep Learning Approach to Network Intrusion Detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41-50, 2018.
[46] Q. Yang, Y. Liu, T. Chen, et al., "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1-19, 2019.
[47] W. Samek, T. Wiegand, and K. R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *ITU Journal: ICT Discoveries*, vol. 1, no. 1, pp. 39-48, 2017.
[48] Q. Niyaz, W. Sun, and A. Y. Javaid, "A Deep Learning Based DDoS Detection System in Software-Defined Networking (SDN)," *EAI Endorsed Transactions on Security and Safety*, vol. 4, no. 12, pp. 1-12, 2017.
[49] J. Li, B. Zhao, and C. Zhang, "Fuzzing: a survey," *Cybersecurity*, vol. 1, no. 1, pp. 1-13, 2018.
[50] M. Ring, S. Wunderlich, D. Grüdl, et al., "Flow-based benchmark data sets for intrusion detection," *Proceedings of the 16th European Conference on Cyber Warfare and Security*, pp. 361-369, 2017.