CIT. Journal of Computing and Information Technology, Vol. 31, No. 4, December 2023, 233–250
doi: 10.20532/cit.2023.1005776

Enhanced Network Security Protection through Data Analysis and Machine Learning: An Application of GraphSAGE for Anomaly Detection and Operational Intelligence

Yujing Lu

Shijiazhuang College of Applied Technology, Shijiazhuang, China

With the Internet's rapid expansion, network security challenges have become increasingly complex and prominent. Traditional protection methods, largely dependent on predefined rules and patterns, demonstrate limited effectiveness against sophisticated and unknown network attacks, failing to harness the full potential of extensive network data. This study addresses the challenges faced by modern cybersecurity, particularly the limitations of traditional defense methods in countering unknown and complex attacks, by proposing a solution that integrates data analysis and machine learning technologies. The focus of this research is placed on network security anomaly detection as well as on intelligent network operations and maintenance exception management based on graph network algorithms, aiming to enhance security defense capabilities and operational efficiency. Specifically, the main contributions and innovations of this paper include:

1. Innovations in sampling, aggregation, and loss functions within the Graph Sample and Aggregation (GraphSAGE) model to improve the accuracy and robustness of the model for network anomaly detection;

2. The introduction of a novel network anomaly root cause analysis and localization model, which, combined with an optimized root cause likelihood assessment method and search scheme, significantly enhances the speed and accuracy of anomaly localization;

3. The design of an integrated decision support system that can automatically adjust protection strategies as network conditions change, achieving a high level of automation and intelligence in cybersecurity management. This work not only provides effective technical support for network security protection but also opens new avenues for future cybersecurity research.

ACM CCS (2012) Classification: Security and privacy → Network security → Denial-of-service attacks

Keywords: network security, data analysis, machine learning, graph network algorithms, anomaly detection, operational intelligence, anomaly control, Graph Sample and Aggregation (GraphSAGE) model

1. Introduction

The advent and rapid proliferation of information technology alongside the expansive reach of the internet have rendered network security a critical concern. This issue is closely linked not only to the security of information assets but also to the stable functioning of network systems [1, 2]. Concurrently, the development of big data and machine learning technologies has catalyzed a paradigm shift in strategies for network security, incorporating both preventive measures and offense-defense mechanisms. Despite this progress, the application of such technologies in the realm of network se-

curity encounters numerous challenges [3–5], underscoring the necessity for effective data analysis and machine learning applications in enhancing network security protocols [6, 7].

Research in this domain has predominantly employed traditional statistical methods and empirical rules to detect network attacks, a strategy largely confined by its dependency on recognized attack patterns. This approach often falls short in combatting intricate and novel network attacks [8–12]. As attack methodologies evolve in tandem with technological advancements, the imperative for mastering sophisticated and potent techniques in network security becomes increasingly salient [13, 14]. In this context, the extraction and utilization of comprehensive network data for security purposes have emerged as a vital research trajectory [15].

Existing methodologies, however, display several shortcomings in effectively addressing network security issues [16–19]. Primarily, the reliance on pre-established rules and patterns constrains their efficacy against unfamiliar and complex network attacks. Furthermore, these methods frequently fail to capitalize on the comprehensive nature of network data, which encompasses both structural and temporal elements. The absence of effective mechanisms for pinpointing and managing network anomalies further compounds these challenges.

With the rapid development of information technology, network security and stability have become the cornerstone of social operation. However, new types of network attacks are emerging constantly, making traditional rule-based security defense measures seem inadequate. To improve the accuracy of network security anomaly detection, this paper makes innovative improvements to GraphSAGE. Specifically, adjustments were made to the sampling function of GraphSAGE, that is, when selecting local neighbor nodes for information aggregation from large-scale graphs, more efficient or more representative mechanisms were adopted; the aggregation function was optimized, improving the way of merging node information to better capture the relationships between nodes and the structural features of the graph; at the same time, the loss function was also refined to ensure that the model training process focuses more on the key features of the anomaly detection task. In terms of network intelligent operations and maintenance exception control, a model capable of accurately analyzing and locating the root causes of network anomalies was designed. This model, by optimizing the likelihood assessment method of root cause points, improves the accuracy of localization.The search scheme was also improved, enhancing the efficiency and effectiveness of the search process. Through the above improvements, not only the performance of graph network algorithms in the field of network security is enhanced, but also a more intelligent and efficient anomaly handling tool is provided for network operations and maintenance, promoting the development of network security and intelligent operations and maintenance technology.

By innovatively improving the GraphSAGE model, this paper can more effectively discover and locate network security anomalies, thereby enhancing the adaptive defense capability of network systems. Meanwhile, the network anomaly root cause analysis and localization model and optimized algorithms proposed in the research further improve the efficiency and accuracy of network intelligent operations and maintenance, which is of practical significance and application value for ensuring the stable operation and data security of network systems and maintaining the continuity and reliability of network services.

The forward-looking nature of this paper is reflected in applying graph network algorithms such as GraphSAGE to the fields of network security, as well asintelligent network operations and maintenance, pushing the graph network algorithms from theory to a broader range of practical applications, especially in dealing with complex network security issues, demonstrating the potential and value of the algorithm in real-world problems. In the field of network intelligent operations and maintenance, the optimization of the likelihood assessment method of root cause points and the search scheme directly improves the efficiency and accuracy of operations and maintenance, facilitating a rapid response and handling of network events.

2. Anomaly Detection in Network Security Utilizing Graph Network Algorithms

Network security data intrinsically comprises abundant structural information, such as IP addresses, port numbers, and protocol types. These elements form complex structural relationships within an extensive network connectivity graph. Network security data also encompasses temporal information, evident from significant variations in network traffic over different periods. In addressing these aspects of network security data, the GraphSAGE model has been employed for anomaly detection. This model capitalizes on the structural data of network security, applying deep learning to the network connectivity graph through graph neural networks, thus unraveling potential patterns hidden within network data. Additionally, it adeptly captures temporal information by processing time-series data, enhancing the precision of anomaly detection.

In the realm of network security protection, the significance of path information for anomaly detection is paramount. Attackers often exploit specific paths for their attacks, such as infiltrating a system via a particular port and subsequently executing internal attacks. Therefore, path information is instrumental in better comprehending and identifying attack behaviors. Moreover, the analysis of path information aids in pinpointing the origin and target of attacks, fostering more effective protection strategies. While the traditional GraphSAGE model utilizes node edge information, it overlooks path information. Consequently, this paper has integrated path factors into the algorithm, refining the sampling function, aggregation function, and loss function. Figure 1 depicts the structure of the model developed in this research.

In scenarios of network security protection, nodes with extensive path associations are likely to exert substantial influence on network security. They can be primary targets for attackers or pivotal nodes in the propagation of attacks. Hence, sampling these nodes more frequently can elevate the accuracy of network anomaly detection. The GraphSAGE model's sampling function, a crucial element, determines the neighboring nodes to be included in the analysis. Traditional sampling functions might predominantly consider node degree, but in network security scenarios, nodes with extensive path associations warrant greater emphasis. The study has therefore augmented the sampling function, increasing the sampling likelihood for neighbors with significant path associations. This enhancement is grounded in the principle of betweenness centrality. Figure 2 illustrates the betweenness centrality distribution within the sample set analyzed. Betweenness centrality, an index measuring a node's prominence in the network, accounts for the shortest paths between all node pairs, with the quantity of paths traversing a specific node denoting its betweenness centrality. Thus, nodes with elevated betweenness centrality are influential within the network, potentially serving as pivotal nodes or key pathways in network attacks. Incorporating this metric into the sampling function enables more effective capture of critical network information, thereby augmenting the efficacy of network security anomaly detection. The weight value is denoted by ϕSA. The number of paths through node n that also pass through node s is represented by $B^s_{PA}(n)$. To regulate the sampling scale, the process is controlled by the parameter B_{SA}. The proposed sampling model in this paper is expressed as follows:

$$\begin{aligned} O_n(s) &= \sum_{n,(s,n)\in R} \frac{\mu(s,n)}{\sum_s \mu(s,n)}, \mu(s,n) \\ &= 1 + \varphi_{SM} \cdot \frac{B^s_{PA}(n)}{B^s_{PA}} \end{aligned} \tag{1}$$

In the methodology of this research, an enhanced sampling function was developed that preferentially samples nodes with numerous path associations within the network. These nodes are likely to be influential, serving either as critical nodes or as key conduits in the propagation of network attacks. This refinement in the sampling function has led to improvements in the model's efficiency. The incorporation of the betweenness centrality metric has refined the sampling process, making it more focused and thereby reducing computational load and storage requirements.

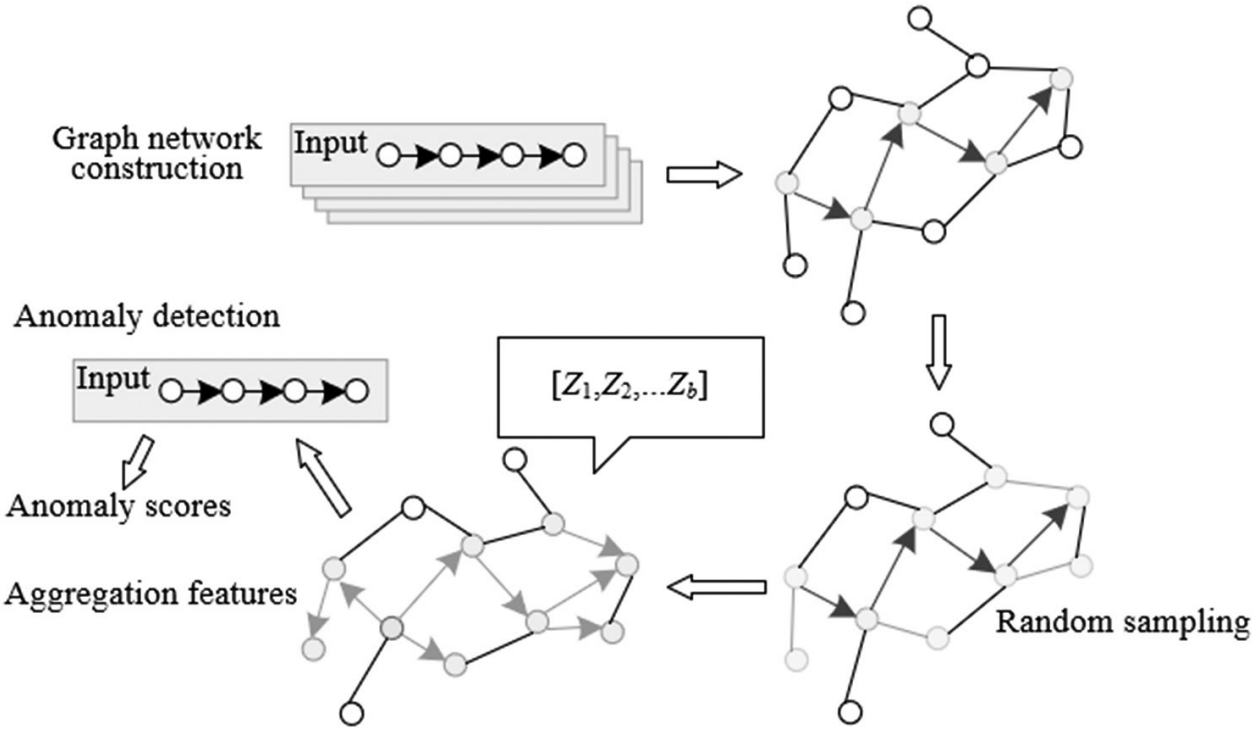

Figure 1. Model structure.

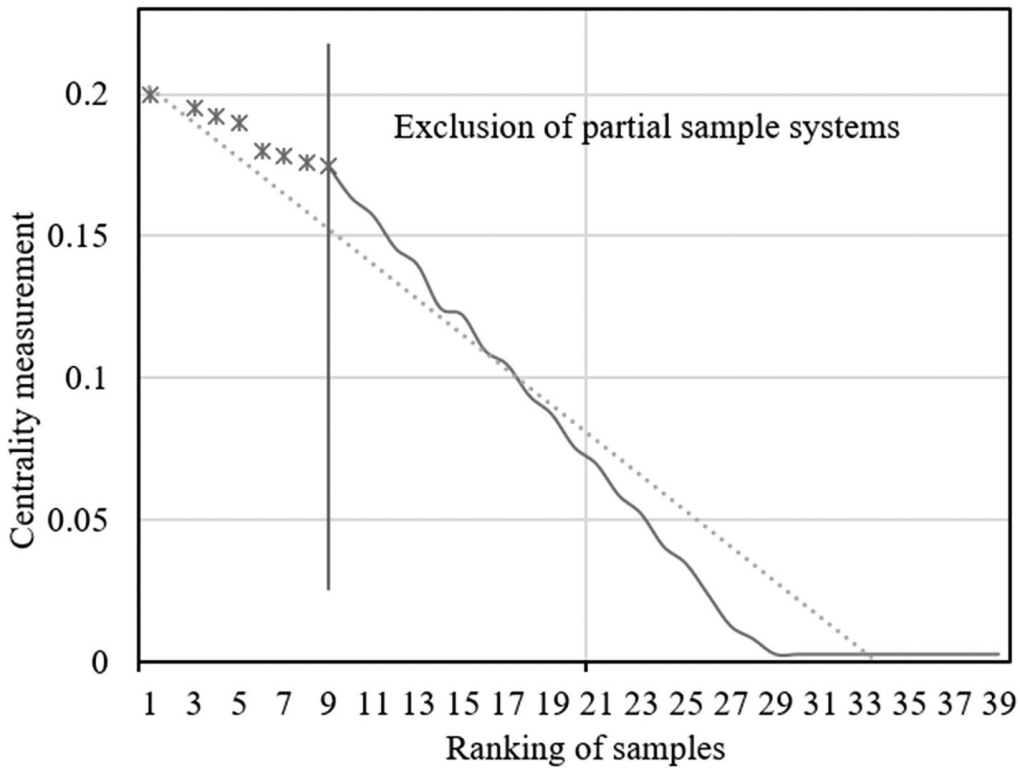

Figure 2. Distribution of betweenness centrality.

Path information is crucial in network security scenarios, as attackers often exploit specific paths, where nodes on these paths contain key attack information. Effective aggregation of adjacent node information, incorporating path characteristics, can heighten anomaly detection accuracy. The traditional GraphSAGE model, however, employs a mean aggregator, which averages the outputs of adjacent nodes and the node itself from the previous layer, inadequately utilizing path information. Hence, an enhancement in the aggregation function is necessitated. The implementation of this mean aggregator scheme is elaborated upon as follows:

$$g_c^j \leftarrow \delta\left(Q \cdot mean\left(\left\{g_c^{j-1}\right\} \cup \left\{g_i^{j-1}, \forall i \in B(c)\right\}\right)\right) \quad (2)$$

Furthermore, an innovative aggregation function is introduced, integrating parameters pertinent to path information. This function aggregates the features of neighboring nodes in accordance with path characteristics. Each path is assigned a specific weight, regulated by parameters related to the path. The features of neighboring nodes are then averaged based on these weights, culminating in an aggregated feature set. As a result, nodes situated on paths utilized in attacks receive higher weights, thus playing a more prominent role in the aggregation process. For instance, if all neighbors of a node Cu are represented by CNE, the aggregation function is defined as follows:

$$g_c^j \leftarrow \delta\left(Q \cdot mean\left(\left\{g_c^{j-1}\right\} \cup \left\{R\left\{o(i) \cdot g_i^{j-1}\right\}, \forall i \in B(c)\right\}\right)\right) \quad (3)$$

The introduction of path-related parameters and the consequent aggregation of neighbor features based on these paths enable the model to more effectively capture information from nodes on attack paths, thus boosting the precision of network security anomaly detection. Furthermore, the optimized aggregation function can resist path deception attacks by attackers, enhancing the model's robustness.

Nodes on the same network path often exhibit similar characteristics, aiding in more accurate anomaly detection. The traditional GraphSAGE model, however, primarily optimizes its loss function based on node-to-neighbor similarity, overlooking node-to-node path relationships. An improved loss function is therefore proposed, designed to aggregate embeddings of nodes on the same path and compare these with original node embeddings. Substantial discrepancies result in a higher loss function value, and vice versa. Figure 3 presents a comparative illustration of neighbor aggregation methods.

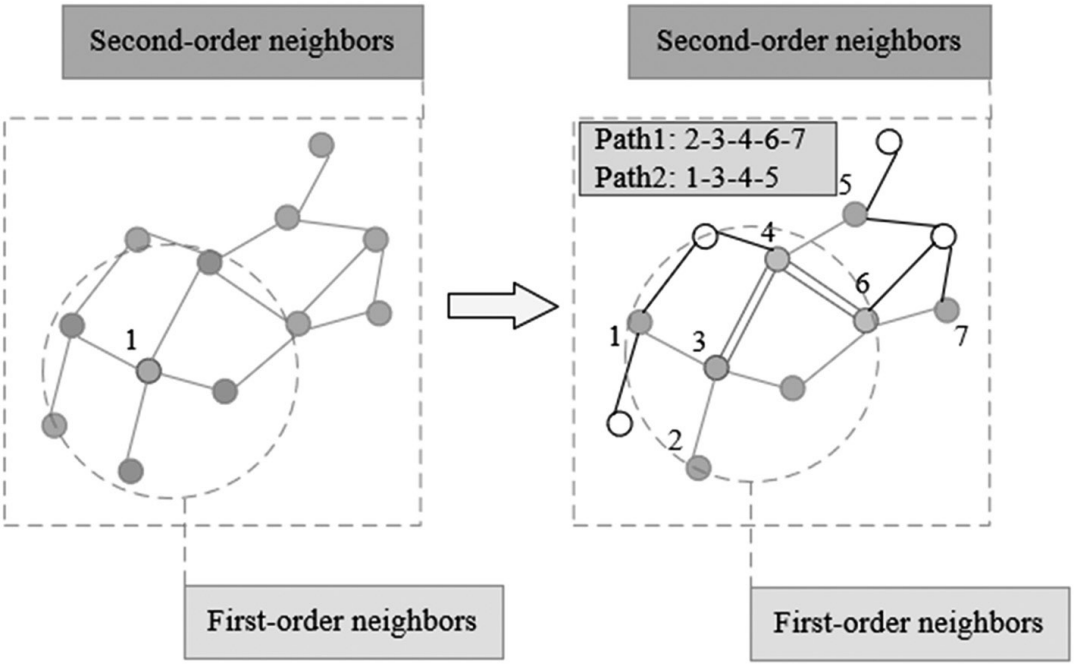

Figure 3. Comparative illustration of neighbor aggregation methods.

$$K_H(X_i) = -\log\left(\delta\left(x_i^Y x_c\right)\right) - W \cdot R_{c_b \sim O_{b(c)}} \log\left(\delta\left(-x_i^Y x_c\right)\right) \quad (4)$$

A revision to the graph-based unsupervised loss function is proposed, enhancing its capability to reflect the similarity of node embeddings on the same path. This revision enables the model to more effectively comprehend and capture path information within the network during the learning process, thereby increasing the accuracy of anomaly detection. The optimized loss function, considering path relationships between nodes, allows the model to utilize not only the local information of nodes but also the structural information of the entire network, thus enhancing its generalizability.

In network security protection scenarios, the distinction in path characteristics between neighboring nodes often serves as a crucial discriminator between attacks and normal behaviors. Overlooking these differences in the loss function calculation might impede accurate anomaly detection. Hence, the loss function optimization, as articulated in Equation 4, is premised on the similarity of node embeddings on the same path, yet it may overlook feature differences between neighbor nodes due to path variations. To address this, further optimization of the loss function is required, accommodating situations involving path characteristic differences. The loss function design, centered on path-based aggregation, ensures adequate similarity in embeddings of nodes on the same path. The function expression is provided as follows.

Nodes situated on the same path, represented by R_{cb}~$O_{IN}(c)$, and nodes not on the same path, denoted by R_{cb}~$O_{IN}(c)$, are considered in this approach. Loss due to neighborhood topology in the graph network, expressed as $K_H(X_i)$, and loss due to path topology, also represented as $K_H(X_i)$, are weighted and adjusted through the value of ϕ_{LO}.

$$K(X_i) = K_H(X_i) + \varphi_{LO} K_O(X_i) \quad (5)$$

$$K_O(X_i) = -W \cdot R_{\backslash c_b \sim O_{IN(c)}} \log\left(\delta\left(x_i^Y x_c\right)\right) - W \cdot R_{c_b \sim O_{IN(c)}} \log\left(\delta\left(-x_i^Y x_c\right)\right) \quad (6)$$

This optimized loss function considers both the similarity of node embeddings on the same path and the feature differences between neighboring nodes due to path variations. This dual consideration allows the model to more effectively understand and capture both path information and path differences during the learning process, thereby enhancing its accuracy in detecting network security anomalies.

The specific steps for network security anomaly detection utilizing graph network algorithms are detailed below.

Node embedding in the network is achieved through graph network algorithms. Node embedding is a technique that maps nodes to a low-dimensional vector space, allowing vectors in this space to reflect the nodes' positions and roles within the network. Once the corresponding node embeddings are obtained, graph network algorithms account for the network's topology and the nodes' attributes. For each node $C_u \in O_s \in E_s$ in the routes E_s requiring detection, an embedding representation r_u is generated. It is necessary to obtain the embeddings for each node on the path and calculate the sum of the embedding differences between adjacent nodes along the path. This process aims to capture the relationships and differences between nodes on the path, as attacks in network security often occur via specific routes. The following formula is used to calculate the anomaly score based on the negative log similarity formula:

$$F_s = \sum\left(\left\|r_u - e_{u-1}\right\|\right) \quad (7)$$

Upon the acquisition of path embeddings, a threshold-based approach is implemented for anomaly detection. It is posited that if the sum of embedding differences among nodes on a given path surpasses the predetermined threshold, ϕ_{TH}, an anomaly is likely present on that path. This assumption stems from the norm that node embeddings on a path should exhibit similarity under typical conditions. Therefore, pronounced deviations may be indicative of malicious activities or attack maneuvers. The formula for calculating anomaly scores that exceed ϕ_{TH} is articulated as follows:

$$O_{AN} = \left\{\sum\left(\left\|r_u - r_{u-1}\right\|\right), c_u \in O_{IN}\right\} \geq \varphi_{TH} \quad (8)$$

Subsequent to the anomaly detection phase, paths registering anomaly scores above the established threshold are selectively identified for further scrutiny and analysis. This methodology, harnessing graph network algorithms, proves effective in pinpointing anomalies within the network. Not only does this approach facilitate the efficient detection of network anomalies, but it also yields comprehensive insights into the nature of anomalous paths, thereby enhancing the understanding and defense mechanisms against network attacks.

This research significantly enhances the performance of network security anomaly detection by making multidimensional improvements to the GraphSAGE model. Specifically, in terms of the sampling function, a more efficient or anomaly detection-oriented neighbor selection strategy is introduced, enabling the model to more accurately select information from a large dataset that aids in anomaly determination. The improvements in the aggregation function, through the introduction of new aggregation strategies or optimization of existing ones, better integrate neighborhood information, allowing the model to more precisely capture anomaly patterns. As for the adjustment of the loss function, it is designed to make the model more focused on the characteristics of anomalous data during the training process. This is achieved by introducing specific loss items for anomaly detection or adjusting the loss weights, thereby improving the accuracy of the model in detecting anomalous behavior in real network environments.

3. Methodological Approach to Anomaly Control in Network Intelligent Operational Maintenance

The development of a conceptual model for root cause analysis in network intelligent operational maintenance is integral to enhancing network security protection and intelligent decision support. This model establishes a structured framework for analysis, enabling a systematic approach to understanding and addressing network anomalies for more effective security measures. Within this model, key indicators of network anomalies are identified, which guide the protection strategies. This framework facilitates the precise determination of network anomaly root causes, thereby augmenting the efficacy of network security measures and aiding in the creation of an intelligent decision support system. This system is capable of autonomously adapting its security strategies in response to variations in network conditions, thereby enhancing the overall effectiveness of network protection.

Leaves in the conceptual model for root cause analysis are defined as a set of elements, each with uniquely determined attribute values constituting a network anomaly's root cause point. Each leaf represents a distinct root cause, comprising a specific combination of attribute values not found concurrently in other leaves. Thus, leaves serve as the most detailed representation of network anomaly root causes. For conducting root cause analysis in network intelligent operational maintenance, a comprehensive collection and analysis of leaves are undertaken to pinpoint the actual causes of network anomalies. Leaves are represented as $MRSD = \{r|r = (s, n, v, f), s \neq *, n \neq *, v \neq *, f \neq *\}$, and the descendant elements of r are denoted as $DE® = \{r'|r'$ is DE of $r\}$. The formula for calculating the *KPI* values for elements of coarser granularity is presented as follows:

$$c(r) = \sum_{r' \in DE(r)} c(r') \tag{9}$$

The model defines a data parallelepiped as a set formed by combinations of attributes *s*, *n*, and *v*, representing the root cause points of anomalies. Each combination pertains to one or more leaves or root causes. The data parallelepiped is envisioned as a multidimensional space, with each dimension corresponding to an attribute and the spatial position of each point determined by its attribute values. Figure 4 illustrates the schematic of the data cube. Assuming the parallelepiped consists of attribute combinations from *s*, *n*, *v* dimensions and comprises several smaller cubes that form a larger cube, represented by $V_{S, N, V}$, the definitions are as follows:

$$R(V_S) = \{r|r = (s,*,*,*), s \neq *\} \tag{10}$$

$$R\left(V_{S,N,V,F}\right)=\left\{r\middle|r=\left(s,n,v,*\right),s\neq *,n\neq *,v\neq *\right\} \quad (11)$$

$$MRSD=R\left(V_{S,N,V,F}\right) \quad (12)$$

In the proposed model for root cause analysis within network intelligent operational maintenance, the search within the data parallelepiped plays a pivotal role. It facilitates the identification of leaves corresponding to anomalous events, thereby pinpointing the anomaly's root cause. Observation of the leaves' distribution within the data parallelepiped uncovers patterns and regularities in anomalies, contributing to the prediction and mitigation of future anomalous events. When the parallelepiped comprises m distinct values across f unique attribute categories, it encapsulates m^f attribute combinations.

The response mechanism to anomalies in network intelligent operational maintenance adheres to a conceptual model encompassing a generalized ripple effect of anomaly propagation. Within a network environment, an anomaly typically initiates locally and swiftly proliferates, impacting other network segments and creating a ripple-like diffusion effect. This phenomenon manifests through hierarchical, cross-domain, and dynamic propagation modes. Anomalies generally extend from an affected node to its connected nodes and subsequently to succeeding hierarchical layers, facilitating rapid expansion of the anomaly's influence over extensive network areas. The complexity inherent in network systems implies that an anomaly can simultaneously affect multiple functional domains. For example, a hardware failure could impair both communication and data processing operations, illustrating the extensive and multifaceted nature of such anomalies. Moreover, as networks operate, the impact of an anomaly may fluctuate over time, indicative of dynamic propagation characteristics, where the severity of an anomaly could either escalate or diminish.

For an effective response to the generalized ripple effect in network intelligent operational maintenance, appropriate control measures are imperative to prevent further spread of anomalies and minimize their network impact. These measures may involve isolating affected nodes, severing problematic paths, and accurately identifying the sources of anomalies. Continuous network monitoring is essential for timely detection and management of new anomalies. A root cause is represented by A, direct network data by L_1 and L_2, derived data calculated from basic collected data by L_3, the actual value by $c(r)$, and the predicted value of the leaf element r by $d(r)$. It is postulated that for each anomalous root cause, there exists a constant j such that any attribute combination satisfies the following equation:

$$\begin{cases}\dfrac{c(r)-d(r)}{d(r)}=\dfrac{c(A)-d(A)}{d(A)}\\ L_3=\dfrac{L_1}{L_2}\\ \dfrac{c_{L_3}(r)}{d_{L_3}(r)}=j\end{cases} \quad (13)$$

The model defines a quantifiable index, termed the potential score, to represent the importance or impact level of a node or path in the network in the context of anomaly propagation. By calculating the potential score of each node or path, priority areas for attention and processing

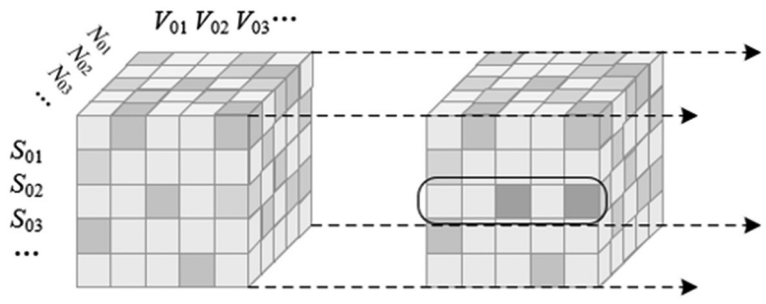

Figure 4. Schematic diagram of the data cube.

are identified, further assessing the risk distribution within the network and aiding in formulating more effective network security strategies. The L_2 norm is utilized as a measure of deviation, denoted by $\| \ \|_2$. Simultaneously, the expected value, a key component in the model's calculations, is represented by $s(A_1)$. The potential score is represented as follows:

$$PS = MAX\left(0{,}1 - \frac{\left\|\left(c(A_1), c(A_2)\right) - \left(s(A_1),\, d(A_2)\right)\right\|_2}{\left\|\left(c(A_1), c(A_2)\right) - \left(d(A_1),\, d(A_2)\right)\right\|_2}\right) \tag{14}$$

$$\begin{aligned} s_a(r) &== d(r) - \frac{d(r)}{d(A)}\left(d(A) - c(A)\right) \\ &= d(r)\frac{c(A)}{d(A)} \end{aligned} \tag{15}$$

The methodology for conducting root cause analysis in network intelligent operational maintenance comprises several key steps, outlined as follows.

Initially, an anomaly detection module is established for monitoring the network's operational state. This module incorporates a range of predefined rules, such as sudden shifts in network traffic or notable delays in response times of network nodes. Activation of these rules triggers the module to ascertain the presence of network anomalies using binary determinations (anomalous or non-anomalous). Subsequent to anomaly detection, relevant data, encompassing network traffic, node status, and system logs, is gathered and stored. This data forms the foundation for later stages of root cause analysis.

The collected data is then processed and analyzed, utilizing a MySQL database for data exportation and segmentation based on timestamps, thus facilitating data preprocessing. Selection and documentation of all leaf combinations occur from the Topostatics table. These leaf combinations represent potential root causes of anomalies, defined by specific attribute combinations. For example, combinations of network traffic and response time attributes are noted if they are identified as potential root causes.

For each leaf combination, specific metrics are chosen to describe its state, such as link rate, latency, jitter, and packet loss rate. These metrics provide the actual values for the leaf combinations, crucial for subsequent anomaly detection and root cause analysis. The Moving Average (MA) algorithm is employed to generate predictive values for each leaf combination, assisting in anticipating potential trends and preempting possible anomalies.

$$R_u = \left\{\left(RE, PR, SI, SP, DI, SP\right)\right\} \tag{16}$$

The final step involves conducting root cause analysis. Inputs such as switch ID numbers and port numbers are fed into the root cause analysis algorithm, which then performs an analysis across the entire network's link information. This process identifies the anomaly's root cause, laying the groundwork for developing effective network security strategies.

In the context of network security protection, disparities in the scale of monitoring metrics can potentially diminish the accuracy of the potential score in reflecting the anomaly degree of leaf elements. For example, a particularly high value in one metric might mask variations in others, leading to an imprecise assessment of the network's anomaly state. To address this, a redefined custom distance for the potential score is proposed for more accurate network anomaly evaluations.

The adoption of a variable distance, replacing the traditional L_2 norm distance, is a key innovation of this study. The variable distance's advantage lies in its ability to automatically adjust weighting in distance calculations based on data characteristics, thereby avoiding the suppression effect of large-scale metrics on the potential score. The process involves calculating the discrepancies between the predicted and actual values for each leaf element. Subsequently, weights are assigned to each leaf element based on the magnitude of these discrepancies, with larger discrepancies warranting greater weights. This approach ensures that metrics significantly impacting network anomalies receive appropriate attention without being overshadowed by large-scale metrics. The distance calculation between vectors $\vec{z}$ and $\vec{t}$, denoted by $m(\vec{z}, \vec{t})$, the anomalous and normal leaf

combinations distinguished by the subscripts sn and b, and the custom adjustment for the variable distance represented by η, are defined as follows:

$$PS = MAX\left(0{,}1 - \frac{m_{sn}\left(\vec{c},\vec{s}\right) + m_b\left(\vec{c},\vec{d}\right)}{m_{sn}\left(\vec{c},\vec{d}\right) + m_b\left(\vec{c},\vec{d}\right)}\right) \quad (17)$$

$$m\left(\vec{z},\vec{t}\right) = \sum_u \sqrt{\left|z_u - t_u\right|^{\eta}} \quad (18)$$

In the methodology for network intelligent operational maintenance anomaly control, the complexity and dynamism of network states often lead to an exponentially growing search space for root cause analysis. To address this, a strategy is proposed whereby the candidate set of root causes is assessed in advance. The assessment and subsequent search are guided by the hierarchical structure of the data parallelepiped, potentially terminating the search early to enhance efficiency. It is posited that the real and predicted *KPI* values of the candidate set A are denoted by $c(A)$ and $d(A)$, respectively, as shown in the equation:

$$IM - PS(A) = \frac{c(A) - d(A)}{\sum_{r \in MRSD} c(r) - d(r)} \quad (19)$$

Figure 5 illustrates the proposed influence score-based search scheme for anomaly root causes in network intelligent operational maintenance. This scheme is restructured to augment efficiency in identifying root causes of network anomalies, as delineated below.

Candidate sets of root causes are initially evaluated based on the generalized ripple effect, prioritizing those sets likely to trigger such an effect. The candidate sets are then ranked according to their influence scores, with higher scores warranting earlier consideration. The search for the root cause candidate set is conducted sequentially, following the hierarchical structure of the data parallelepiped, thus circumventing inefficacious searches and augmenting efficiency. Upon identifying a plausible root cause, the search can be concluded prematurely, conserving both time and computational resources, further elevating efficiency.

This redesign of the search scheme fundamentally aims to expedite the discovery of root causes in network anomalies, thereby facilitating more efficacious network security protection.

In summary, in the proposed model, the decision-making process begins with the collection and processing of massive data in complex network environments based on a structured analysis framework, utilizing the capability of graph network algorithms to capture the multidimensional characteristics and potential correlations of network anomalies. The model analyzes network behavior through intelligent algorithms, identifies anomaly patterns, and matches these

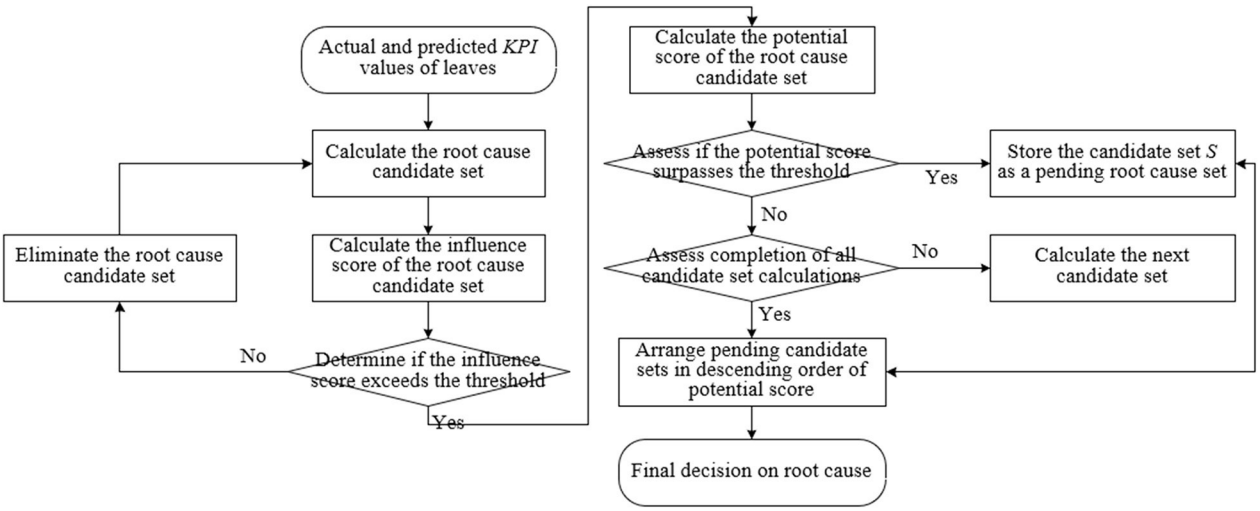

Figure 5. Influence score-based network intelligent operational maintenance anomaly root cause search scheme.

patterns with known security incidents or anomalous behaviors to determine key indicators of anomalies. Then, using these key indicators, the model conducts in-depth root cause analysis by calculating and comparing various potential root cause points and their likelihood scores to identify the most probable root cause. Finally, the results of the root cause analysis are input into the decision support system, which can automatically recommend or adjust security protection strategies based on the analysis results and the current network state, ensuring a rapid response and effective prevention or mitigation of network anomalies, thereby enhancing the overall efficacy of network security protection. This intelligent decision support process demonstrates a high degree of automation and intelligence, greatly improving the efficiency and accuracy of network security operations and maintenance.

4. Results and Analysis

In this research, to evaluate the network security anomaly detection model based on graph network algorithms, we designed the following experimental setup and parameter configuration. Specifically, the experiment uses two sets of data. The first set is the network traffic data from the day before an anomaly occurred, serving as the reference input for the normal state, used to train the model. The second set of data consists of streaming routing update data at the time of the anomaly, used as input for real-time anomaly detection. In the graph sampling process, to balance the importance of graph structure and path information, we assign equal weight to these two factors. The sampling scale is set to 6, meaning that in constructing the graph, we sample 6 neighbor nodes from each node to ensure the graph representation is rich enough while reducing computational burden. The model is first trained using the data from the day before the anomaly is tagged. After training, the model will be applied to data within the time frame of the anomaly tagging for real-time anomaly detection. Precision and F-1 score are used as the main performance evaluation metrics. Precision measures the model's ability to correctly mark anomalies, while the F-1 score is the harmonic mean of precision and recall, providing an assessment of overall performance. The threshold for the anomaly score is set to 0.5, with network behavior exceeding this score considered anomalous. This threshold is the boundary the model uses to differentiate between normal and anomalous behavior.

In the domain of network security anomaly detection, enhancements in the GraphSAGE model have been introduced, encompassing modifications to its sampling, aggregation, and loss functions. Table 1 presents a detailed evaluation of this model's effectiveness. Enhanced accuracy and F-1 score are consistently noted across Networks 1, 2, and 3, when compared to the decision tree and GraphGAGE models. Such results distinctly highlight the superior efficacy of the refined model in detecting anomalies within network security, outperforming the aforementioned models. Furthermore, an analysis of runtime efficiency reveals that, while the enhanced GraphSAGE model requires a longer runtime than the GraphGAGE model, it remains significantly more time-efficient than the decision tree model. This balance of high performance with efficient operation is a notable feature of the proposed model. The comprehensive experimental data thus underscores the robustness of the improved GraphSAGE model in the context of network security anomaly detection, excelling in both predictive accuracy and operational efficiency. The advancements in the GraphSAGE model, evidenced by these comparative analyses, suggest its effectiveness as a method for anomaly detection in network security.

Data presented in Figure 6 illustrates the influence of various parameters on model accuracy, contingent upon changes in sampling parameter values. It is observed that an increase in sampling parameters initially leads to a substantial rise in accuracy for loss calculation parameters, culminating in a peak, beyond which the accuracy declines as the parameter values continue to escalate. This peak signifies the presence of an optimal value for loss calculation parameters, where the model attains its highest level of accuracy. Correspondingly, the trend lines for sampling parameters exhibit a similar pattern, where accuracy enhances with the rise in sampling values, reaches an apex, and then moderately decreases. This trend suggests that the model demonstrates improved performance within the range of optimal sampling parameter values. Furthermore, a direct relationship between sam-

pling scale and accuracy is noted, with accuracy amplifying until reaching a maximum value, followed by a gradual decline. This trend implies that while modulating the sampling scale can optimize model performance, an excessively extensive sampling scale may not necessarily lead to superior outcomes. The graphical data vividly portrays the substantial impact these parameters exert on model accuracy, emphasizing that each parameter has a specific optimal range conducive to achieving heightened accuracy. This observation underscores the importance of meticulous parameter tuning in substantially elevating the efficacy of network security anomaly detection methodologies.

The ablation study detailed in Table 2 critically evaluates the performance of different network security anomaly detection models across three distinct network scenarios. In all three network scenarios, the proposed model consistently exhibits superior accuracy and F1-score, affirming its optimized nature and robust generalizability across varied datasets. Models employing random sampling showed a marked decrease in performance, suggesting that such an approach might result in the loss of vital structural information crucial for effective model functioning. In contrast, models with path-based sampling demonstrated enhanced performance, indicating that the refinements in sampling methods incorporated in the proposed model more efficiently capture critical network structural data, thus boosting anomaly detection accuracy.

Regarding aggregation functions, models relying solely on mean aggregation underperformed when compared to the proposed model. This underscores the inadequacy of relying solely on mean values for aggregating neighboring node information. An improvement is noted when path-based mean aggregation is employed, suggesting that the modifications in aggregation functions in the proposed model are more apt for handling information aggregation in complex network scenarios. Models that focused only on minimizing neighbor embedding were found to be the least effective, likely due to their failure to sufficiently account for the path relationships between nodes. This was substantiated by the observed enhancement in performance with path-based embedding minimization, validating the importance of improvements made in the loss function in the proposed model for augmenting overall model performance. The findings of the ablation study are conclusive in demonstrating the effectiveness of the enhancements made to the sampling functions, aggregation functions, and loss functions in the proposed model. These modifications play a pivotal role in the heightened accuracy and efficiency of the proposed model in network security anomaly detection. The model not only accurately identifies network anomalies but also consistently displays high accuracy and F1-score across different network environments, emphasizing its practical applicability and effectiveness in network security domains.

Table 1. Comparative experimental results of different network security anomaly detection models.

Model		**Decision tree**	**GraphGAGE**	**The proposed model**
Network 1	Accuracy	0.812	0.785	0.832
	F1-score	0.824	0.758	0.827
	Run time	715 s	123 s	158 s
Network 2	Accuracy	0.852	0.785	0.814
	F1-score	0.814	0.789	0.821
	Run time	758 s	117 s	163 s
Network 3	Accuracy	0.788	0.745	0.781
	F1-score	0.768	0.712	0.787
	Run time	24 s	13 s	18 s

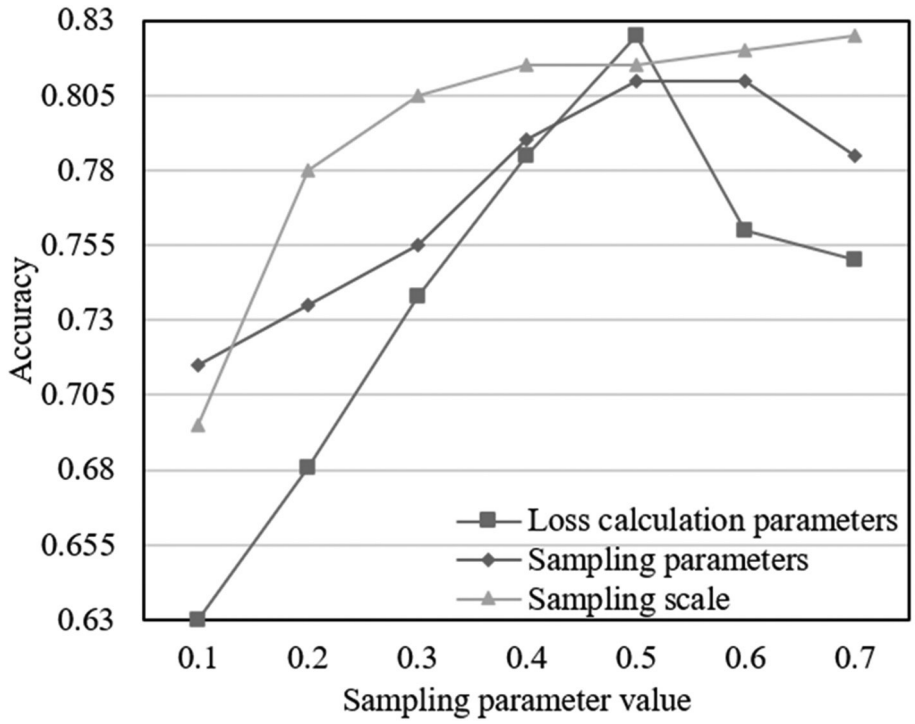

Figure 6. Parameter analysis experimental results.

Table 2. Ablation experimental results of network security anomaly detection models.

Model		**Network 1**		**Network 2**		**Network 3**	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
The proposed model		0.839	0.827	0.835	0.819	0.814	0.785
Neighbor Sampling	Random sampling	0.756	0.745	0.784	0.761	0.745	0.741
	Path-based sampling	0.784	0.776	0.784	0.782	0.782	0.753
Aggregation function	Mean aggregation function	0.738	0.725	0.748	0.748	0.721	0.725
	Path-based mean aggregation function	0.739	0.736	0.739	0.739	0.725	0.736
Loss function	Neighbor embedding minimization	0.718	0.725	0.718	0.725	0.726	0.721
	Path-based embedding minimization	0.789	0.784	0.779	0.771	0.774	0.756

In the realm of network intelligent operational maintenance for anomaly control, a novel model for network anomaly root cause analysis and localization was developed. This model enhances the efficiency of operational processes through the optimization of root cause likelihood assessment methods and search strategies. As delineated in Table 3, three distinct network security anomaly scenarios were examined, each exhibiting unique characteristics. The variations in traffic patterns across these scenarios typify different forms of network security anomalies: unauthorized access anomalies are marked by abrupt traffic surges, DDoS attack anomalies exhibit short-term traffic fluctuations, and zero-day attack anomalies are characterized by persistent abnormal traffic changes. The selection of these varied scenarios facilitates the precise analysis and identification of network anomalies' root causes within the framework of this study's model.

Figure 7 presents the anomaly degree scores as evaluated by three disparate network intelligent operational maintenance methods across network anomaly scenarios. An elevated anomaly degree score signifies a more severe anomaly and a method's enhanced proficiency in anomaly detection. The figure reveals that with an increasing number of computations, the local outlier factor (LOF) method manifests comparatively lower scores with minimal fluctuations. Conversely, the isolation forest method scores moderately with more pronounced fluctuations. The method proposed in this study consistently yields higher scores with moderate fluctuations, underscoring its effectiveness in network security anomaly detection. Relative to the LOF and isolation forest algorithms, the proposed method not only demonstrates superior accuracy in anomaly detection across various scenarios but also maintains stable and consistent scoring. This stability and consistency are paramount in operational environments aimed at the real-time detection and response to network anomalies, further accentuating the practical utility and efficacy of this study's model in network security protection.

Figure 8 depicts the accuracy of control responses by various algorithms in different anomaly scenarios. It has been observed that the algorithm proposed in this study exhibits superior control response accuracy when addressing network security anomalies, including unauthorized access, DDoS attacks, and zero-day attacks, relative to other algorithms. This superiority is consistently evident across various anomaly scenarios, indicating the algorithm's effectiveness. The consistently high accuracy in each scenario further highlights the comprehensive nature of the proposed algorithm, suggesting its widespread applicability to a range of network security challenges. The performance of the algorithm elucidated in this study accentuates its efficacy in accurately identifying and managing diverse types of network attacks in real-world environments, thereby establishing its reliability and preeminence as an anomaly detection and response tool.

Table 3. Network security anomaly scenarios.

Anomaly type	Source	Destination	Business path	Normal traffic	Anomalous traffic
Unauthorized access	*X*11	*X*15	*D*5-*D*6-*D*7-*D*8 or *D*3-*D*1-*D*5-*D*6-*D*7-*D*8	10~50 Mbps	Spike to 80~100 Mbps
DDoS attack	*X*8	*X*14	*D*3-*D*5-*D*1-*D*6-*D*7-*D*8 or *D*3-*D*5-*D*1-*D*6-*D*7-*D*8	20~80 Mbps	Fluctuate to 50~60 Mbps
Zero-day attack	*X*14	*X*7	*D*14-*D*6-*D*5-*D*3	50~90 Mbps	Long-term fluctuation to 20~40 Mbps

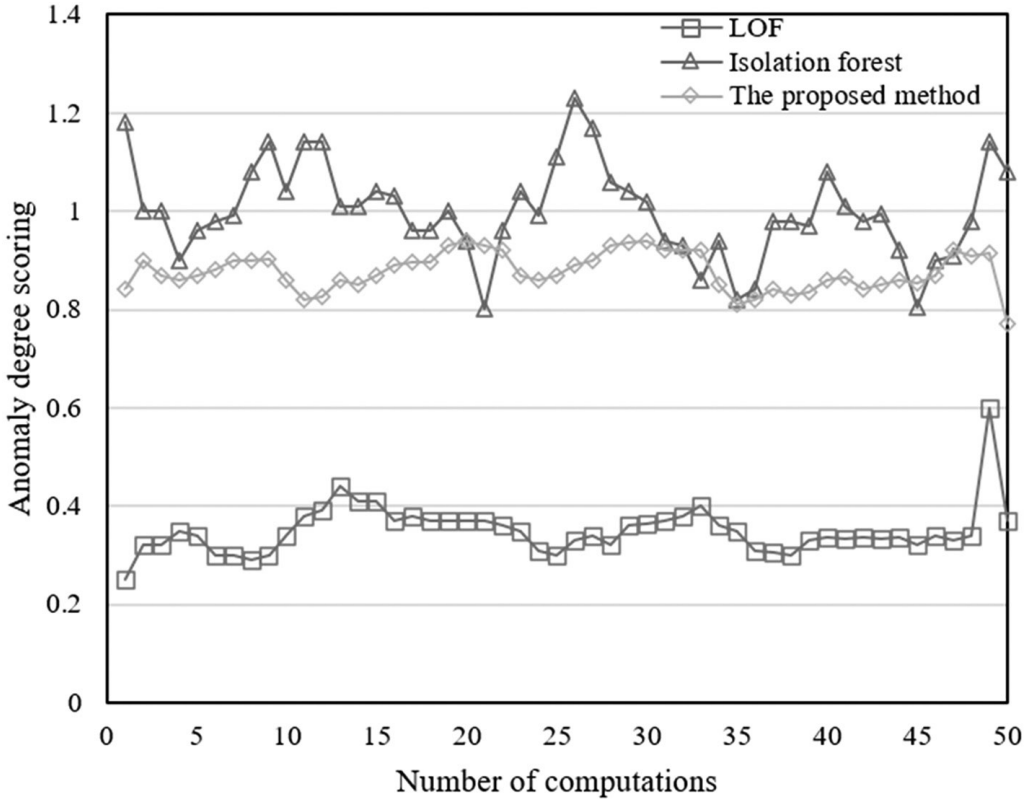

Figure 7. Anomaly degree scoring for different network intelligent operational maintenance methods.

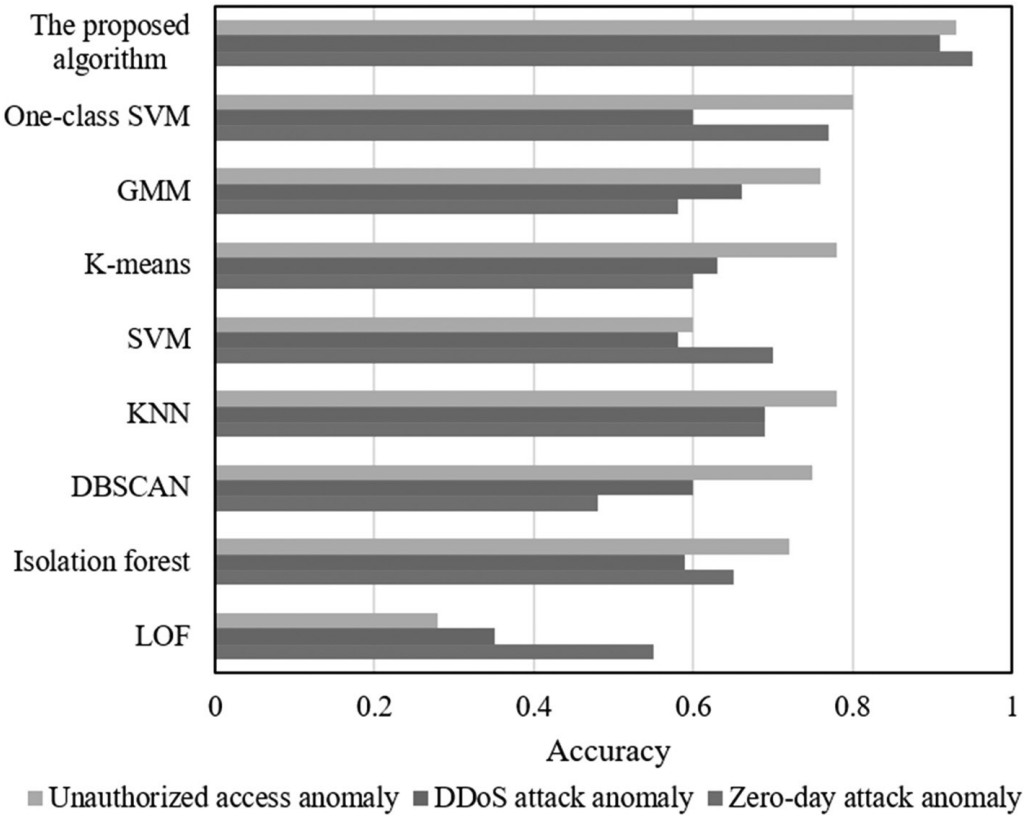

Figure 8. Control response accuracy of different algorithms in various anomaly scenarios.

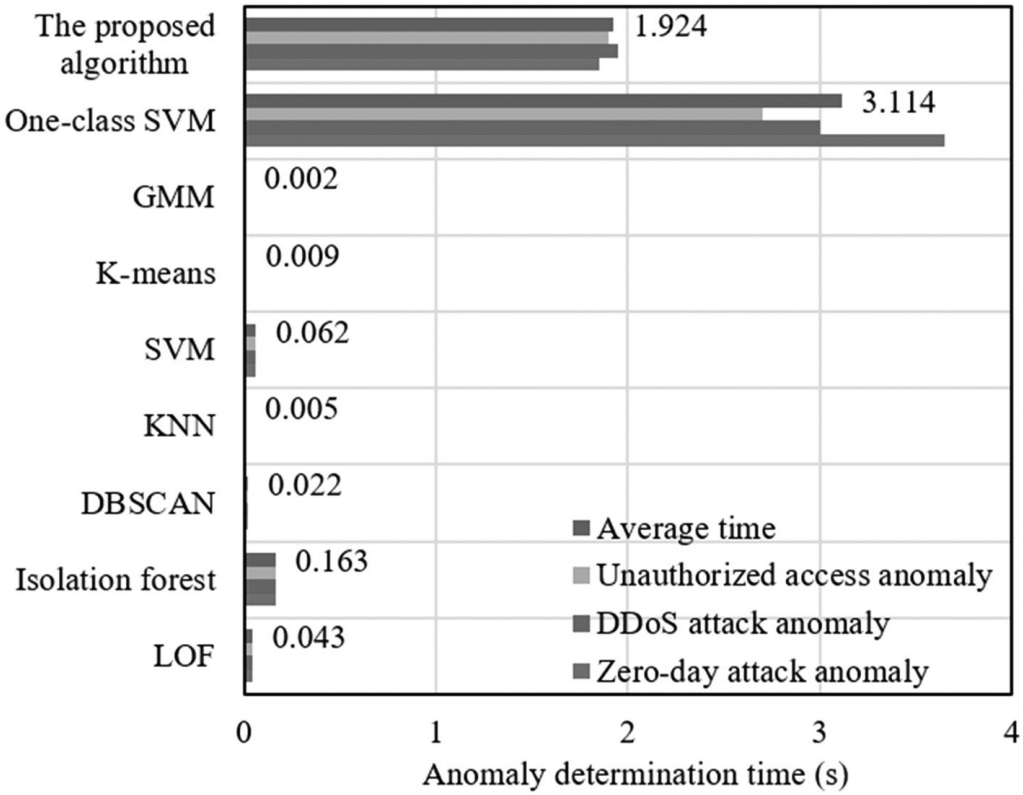

Figure 9. Anomaly detection time consumption of different algorithms in various anomaly scenarios.

Figure 9 shows the time taken for anomaly detection by different algorithms across various scenarios. The response time of the algorithm introduced in this study is not the shortest for all anomaly types when compared to other algorithms, yet it remains within an acceptable range. Importantly, considering its high accuracy in response, the algorithm's utility in network security is deemed invaluable. Furthermore, its precise detection and responsive capabilities are instrumental in significantly reducing potential losses, underscoring the algorithm's practical significance in the realm of network security.

5. Conclusion

In this study, two primary areas were addressed: network security anomaly detection and the control of anomalies within intelligent operational maintenance. The study successfully incorporated and enhanced the GraphSAGE model for network security anomaly detection. Emphasis was placed on refining the model's sampling function, aggregation function, and loss function. The revised sampling function accentuates the significance of network structural information and inter-node connections, aiding in the identification of potential anomalies. The aggregation function has been optimized to effectively merge neighboring node data, capitalizing on graph structural characteristics to increase anomaly detection precision. Additionally, the loss function now accounts for path similarities between nodes, enhancing the model's sensitivity to anomalous data points. Experimental results indicate that the model, in comparison to conventional decision trees and the GraphGAGE graph algorithm, consistently achieves superior accuracy and F1-score across various network datasets, exhibiting notably stable and enhanced performance during parameter analysis.

For the aspect of intelligent operational maintenance, a novel model for network anomaly root cause analysis and localization was developed. This model refines the method of assessing the likelihood of root causes and introduces the concept of potential scores for quantitative

evaluation of anomaly impact and severity. Furthermore, the search strategy has been meticulously refined, considering the generalized ripple effect to efficiently and systematically identify anomaly root causes. Experimental findings demonstrate the model's exceptional performance, surpassing traditional models in ablation studies and exhibiting higher control response accuracy and quicker anomaly detection response times in diverse network security anomaly scenarios, such as unauthorized access, DDoS attacks, and zero-day attacks.

In conclusion, the conducted research and experiments underscore the effectiveness of the enhanced GraphSAGE model in network security anomaly detection. The study also confirms that the developed root cause analysis and localization model, along with the root cause point assessment method, significantly boost the efficiency of network intelligent operational maintenance. This study provides network operations personnel with a powerful decision support tool, enabling more intelligent and automated network security protection. It allows for the rapid response and handling of security incidents, significantly improving the timeliness and effectiveness of network security defense, reducing network security risks, and enhancing the stability and reliability of the network.

References

[1] P. A. W. Putro and D. I. Sensuse, "Review of Security Principles and Security Functions in Critical Information Infrastructure Protection", *International Journal of Safety and Security Engineering*, vol. 12, no. 4, pp. 459–465, 2022. https://doi.org/10.18280/ijsse.120406

[2] R. Vatambeti and G. Mamidisetti, "Routing Attack Detection Using Ensemble Deep Learning Model for IioT", *Information Dynamics and Applications*, vol. 2, no. 1, pp. 31–41, 2023. https://doi.org/10.56578/ida020104

[3] M. Jin, "Computer Network Information Security and Protection Strategy Based on Big Data Environment", *International Journal of Information Technologies and Systems Approach*, vol. 16, no. 2, p. 319722, 2023. https://doi.org/10.4018/IJITSA.319722

[4] T. Ayanwola *et al.*, "Enhancing Face Spoofing Attack Detection: Performance Evaluation of a VGG-19 CNN Model", *Acadlore Transactions on AI and Machine Learning*, vol. 2, no. 2, pp. 84–98, 2023. https://doi.org/10.56578/ataiml020204

[5] L. Meng, "Internet of Things Information Network Security Situational Awareness Based on Machine Learning Algorithms", *Mobile Information Systems*, vol. 2022, p. 4146042, 2022. https://doi.org/10.1155/2022/4146042

[6] O. B. Ohwo *et al.*, "Advancing DNS Performance through an Adaptive Transport Layer Security Model (ad-TLSM)", *Ingénierie des Systèmes d'Information*, vol. 28, no. 3, pp. 777–790, 2023. https://doi.org/10.18280/isi.280329

[7] K. Yun *et al.*, "A Network Security Approach based on Machine Learning", in *Proc. of the 2023 IEEE International Conference on Integrated Circuits and Communication Systems, Raichur, India*, 2023, pp. 1–5. https://doi.org/10.1109/ICICACS57338.2023.10100204

[8] T. Saha *et al.*, "Machine Learning Assisted Security Analysis of 5G-network-connected Systems", *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 2006–2024, 2022. https://doi.org/10.1109/T*Etc.*2022.3147192

[9] D. O. Alao *et al.*, "The Need to Improve DNS Security Architecture: An Adaptive Security Approach", *Information Dynamics and Applications*, vol. 2, no. 1, pp. 19–30, 2023. https://doi.org/10.56578/ida020103

[10] B. Prasad and S. Ramachandram, "Prevention and Detection Mechanisms for Re-entrancy Attack and King of Ether Throne Attack for Ethereum Smart Contracts", *Ingénierie des Systèmes d'Information*, vol. 27, no. 5, pp. 725–735, 2022. https://doi.org/10.18280/isi.270505

[11] K. B. Dasari and N. Devarakonda, "Detection of TCP-based DDoS Attacks with SVM Classification with Different Kernel Functions using Common Uncorrelated Feature Subsets", *International Journal of Safety and Security Engineering*, vol. 12, no. 2, pp. 239–249, 2022. https://doi.org/10.18280/ijsse.120213

[12] S. Bhosale and H. Patil, "Zigbee-based Intrusion Detection System for Wormhole Attack in Internet of Things", *Mathematical Modelling of Engineering Problems*, vol. 10, no. 2, pp. 663–670, 2023. https://doi.org/10.18280/mmep.100237

[13] K. T. Nitesh *et al.*, "Network Security Threat Detection: Leveraging Machine Learning Algorithms for Effective Prediction", in *Proc. of the 2023 12th International Conference on Advanced Computing, Chennai, India*, 2023, pp. 1–5. https://doi.org/10.1109/ICoAC59537.2023.10249943

[14] S. Choudhary and S. Dorle, "A Quality of Service-aware High-security Architecture Design for Software-defined Network Powered Vehicular ad-hoc Networks Using Machine Learning-based Blockchain Routing", *Concurrency and Computation: Practice and Experience*, vol. 34, no. 17, p. e6993. https://doi.org/10.1002/cpe.6993

[15] Q. Li *et al.*, "An Overview of Cybersecurity based on Network Security Situational Awareness and Machine Learning", in *Proc. of the 2023 8th International Conference on Intelligent Computing and Signal Processing, Xi'an, China*, 2023, pp. 279–285. https://doi.org/10.1109/ICSP58490.2023.10248496

[16] Z. Todorov *et al.*, "FPGA Implementation of Computer Network Security Protection with Machine Learning", in *Proc. of the 2021 IEEE 32nd International Conference on Microelectronics, Nis, Serbia*, 2021, pp. 263–266. https://doi.org/10.1109/MIEL52794.2021.9569201

[17] G. Chen *et al.*, "Mobile Cellular Network Security Vulnerability Detection using Machine Learning", *International Journal of Information and Communication Technology*, vol. 22, no. 3, pp. 327–341. https://doi.org/10.1504/IJICT.2023.10038425

[18] J. Z. Liu, "Research on Network Big Data Security Integration Algorithm Based on Machine Learning", in *Proc. of the 2021 International Conference of Social Computing and Digital Economy, Chongqing, China*, 2021, pp. 264–267. https://doi.org/10.1109/ICSCDE54196.2021.00067

[19] M. Furdek and C. Natalino, "Machine Learning for Optical Network Security Management", in *Proc. of the 2020 Optical Fiber Communications Conference and Exhibition, San Diego, CA, USA*, 2020, pp. 1–3.

Received: December 2023
Revised: March 2024
Accepted: March 2024

Contact addresses:
Yujing Lu
Shijiazhuang College of Applied Technology
Shijiazhuang
China
e-mail: 2016010804@sjzpt.edu.cn

Yujing Lu graduated from Hebei Normal University with a Master's degree in Engineering in 2016. Currently he is employed at Shijiazhuang College of Applied Technology, specializing in research areas of artificial intelligence and network security.