Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.044149

ARTICLE

Machine Learning Security Defense Algorithms Based on Metadata Correlation Features

Ruchun Jia, Jianwei Zhang* and Yi Lin

College of Computer Science, Sichuan University, Chengdu, 610065, China

*Corresponding Author: Jianwei Zhang. Email: zhangjianwei@scu.edu.cn

Received: 22 July 2023 Accepted: 25 December 2023 Published: 27 February 2024

ABSTRACT

With the popularization of the Internet and the development of technology, cyber threats are increasing day by day. Threats such as malware, hacking, and data breaches have had a serious impact on cybersecurity. The network security environment in the era of big data presents the characteristics of large amounts of data, high diversity, and high real-time requirements. Traditional security defense methods and tools have been unable to cope with the complex and changing network security threats. This paper proposes a machine-learning security defense algorithm based on metadata association features. Emphasize control over unauthorized users through privacy, integrity, and availability. The user model is established and the mapping between the user model and the metadata of the data source is generated. By analyzing the user model and its corresponding mapping relationship, the query of the user model can be decomposed into the query of various heterogeneous data sources, and the integration of heterogeneous data sources based on the metadata association characteristics can be realized. Define and classify customer information, automatically identify and perceive sensitive data, build a behavior audit and analysis platform, analyze user behavior trajectories, and complete the construction of a machine learning customer information security defense system. The experimental results show that when the data volume is 5×10^3 bit, the data storage integrity of the proposed method is 92%. The data accuracy is 98%, and the success rate of data intrusion is only 2.6%. It can be concluded that the data storage method in this paper is safe, the data accuracy is always at a high level, and the data disaster recovery performance is good. This method can effectively resist data intrusion and has high air traffic control security. It can not only detect all viruses in user data storage, but also realize integrated virus processing, and further optimize the security defense effect of user big data.

KEYWORDS

Data-oriented architecture; metadata; correlation features; machine learning; security defense; data source integration

1 Introduction

User information security defense refers to the adoption of a series of measures and technologies to protect the confidentiality, integrity, and availability of user information, and prevent information leakage, tampering, or abuse. In today's digital age, the security of user information for individuals and organizations is facing unprecedented challenges and threats [1]. With the popularization of the

Internet and technologies progress, network attack methods are becoming increasingly complex and covert, posing serious risks to user privacy and data security. In this context, studying user information security defense has become particularly urgent and important. However, there are many difficulties in the field of user information security defense. The openness and anonymity of the Internet provide opportunities for security incidents such as hacker attacks and internal leaks. Personal information being stolen, and user data being tampered with or lost can cause irreparable losses to individuals and enterprises. In response to this situation, various Internet enterprises and relevant organizations should tend to safeguard the security of big data on the Internet. In addition, regulations and supervision worldwide have also put forward requirements for the security of big data on the Internet. For example, on May 25, 2018, the European Union (EU) implemented the General Data Protection Regulation (GDPR), which requires enterprises to protect the user data they collect and specifies the punishment measures for violating the regulations. For Internet enterprises involving EU citizen data, they may need to change their data storage and processing methods to comply with regulatory requirements. Similar regulatory measures have also been implemented in different countries and regions. In summary, the security of big data of Internet users has become an issue that cannot be ignored in the Internet field. How to safeguard the security of big data on the Internet, and prevent and respond to security attacks and leaks, is a major challenge facing the Internet industry today [2].

Regarding the task at hand, which is to protect the security of network user data, specific measures include:

(1) Collecting and analyzing network user data to identify and predict potential security risks.

(2) Developing and implementing a secure plan for encrypting, backing up, and recovering network user data.

(3) Establishing an efficient and secure network architecture, including physical and logical security measures.

(4) Implementing access control and permission management to limit user access and modification rights to data.

(5) Deploying various security tools such as firewalls, intrusion detection, and anti-virus software to detect and prevent security threats on time.

Currently, there are various solutions available for the security defense of network user data. For example: Zhou et al. [3] proposed a security defense system based on intelligent immunity, which can protect the security of the entire system and identify unauthorized mobile users. The system can achieve security protection through three functions: recognition, learning, and adjustment. However, this method may be limited by the coverage of existing attack samples, and its adaptability to new attacks may be insufficient. Kang et al. [4] proposed a network security data classification method based on multiple feature extraction, which solves the instability problem of nonlinear time series. This method adopts adaptive estimation technology and constructs a classification system based on the principle of phase interval reconstruction, achieving dynamic and autonomous adaptive estimation of network security threats, and ensuring the feasibility of network security information classification. However, when dealing with large-scale data, this approach may face the problem of low processing efficiency, especially in real-time environments. Reference [5] proposed a single modulation arbitrary vector beam encoding method, which uses non-interference methods to modulate a single polarization component for encoding, removing complex optical settings. This method is effective in the application of information security and has been experimentally proven to be practical. However, in complex network environments, this approach may face reliability problems, such as synchronization

errors and noise interference. Reference [6] explored the application of blockchain technology in public sector entity management, particularly examples of security and energy efficiency in cloud computing data processing. Blockchain technology can ensure data security, sharing protection, and automated development, which is of great significance for fields such as cloud solutions, big data, and the Internet of Things, and allows the creation of new programmable ecosystems. By utilizing blockchain technology, the carbon footprint of Shared Service Centers (SSCs) can be reduced, and the efficiency and development of SSCs can be improved. This technology can not only reduce costs but also achieve positive ecological effects. While blockchain can provide decentralized and immutable features to enhance the security of data, in practice, the technology has some performance and scalability limitations. Reference [7] proposed a network security defense Algorithm Based on supporting blockchain. By supporting blockchain technology to detect network vulnerabilities, a federal cloud computing framework is designed, and a network deployment scheme is optimized to enhance network security defense capability and reduce network risk. However, the efficiency and scalability of the algorithm in large-scale network environments still need to be further improved. Reference [8] proposed a data-driven machine learning security defense method. This method can effectively improve the security of network transmission data. However, the method may be challenged by sample imbalance and attacker antagonism, and more research is needed to improve robustness and accuracy.

For the defects in literature, by introducing correlation features and machine learning methods in this paper, we can improve the adaptability of the system and make it better able to cope with new attacks. The extraction of associated features and the integration of machine learning algorithms can provide more comprehensive and accurate analysis and prediction capabilities, thereby enhancing the system's ability to detect and defend against various attacks. For the defects in literature, this paper uses the extraction method of associated features to improve its feature extraction process and adopts more efficient machine learning algorithms to improve the accuracy and efficiency of classification, to solve the efficiency problems it may face when processing large-scale data. For the defects of literature, this paper introduces correlation features and machine learning methods to improve its robustness. By analyzing and learning a large amount of network data, we can build more accurate models to deal with synchronization errors and noise interference and improve the reliability of coding methods. For the defects of literature, this paper draws on its decentralized and immutable characteristics and combines correlation features and machine learning methods to solve the performance and scalability problems. By establishing a secure record and verification of user information on the blockchain, the security and privacy protection capabilities of the system can be improved. For the defects in literature, this paper introduces correlation features and machine learning methods, which can improve the efficiency and scalability of its algorithm, and further improve the system's ability to verify and record network events. For the defects in literature, this paper uses association features for data processing and adopts a more robust machine learning algorithm, which can reduce the impact of sample imbalance and attacker antagonism on the method and improve its accuracy and effect in practical applications.

To further improve the shortcomings of the existing methods mentioned above, this paper designs a machine learning security defense algorithm based on metadata association features. The aim is to use the descriptive information of metadata to analyze the correlation between data, thereby improving data security and effectively preventing malicious attacks. This algorithm innovatively utilizes machine learning algorithms to analyze metadata and discover connections between data. By associating features with metadata, the accuracy of machine learning models can be improved, thereby enhancing the efficiency and reliability of security defense algorithms. The key contributions of this algorithm include the following aspects:

(1) Firstly, it utilizes descriptive information from metadata, such as data type, data source, data attributes, etc., to help analyze the correlation of data. This method can provide a more comprehensive and in-depth understanding of data, reveal potential data associations, and help identify potential security threats.

(2) Secondly, security defense algorithms based on metadata association features can also achieve automated data classification and security monitoring. By analyzing the associated features of metadata, data can be automatically classified, and abnormal behavior can be monitored in real-time. This provides better protection mechanisms for enterprises and users, helping them quickly identify and respond to potential security threats [9,10].

Experimental results show that the algorithm has been widely applied in many fields, such as natural language processing, image processing, speech recognition, and credit scoring. The experimental results also show that the proposed method can solve the security issues in user data. In summary, this algorithm improves the effectiveness and reliability of network security protection. The innovation lies in the use of metadata-associated features to achieve rapid data classification, better interpretation of security risks exposed by data, and better support for manual labor in identifying malicious attacks. This will be very helpful and reduce the cost and manpower needed for enterprises in terms of information security.

The organizational structure of the machine learning security defense algorithm based on metadata association features is as follows:

(1) Information security defense goal setting and defense principles: Determine the overall and detailed goals of information security defense, and set defense principles for user information security, such as confidentiality, integrity, availability, etc.

(2) Integration of association feature extraction and machine learning algorithms: Use association features to extract and analyze user metadata, and integrate machine learning algorithms to handle heterogeneous data sources, improving model prediction accuracy.

(3) Establishment of machine learning user information security defense system: Based on association features and machine learning algorithms, automatic recognition of user information and autonomous perception of user sensitive data are achieved, and a machine learning user information security defense system is established.

(4) Audit analysis platform and user behavior trajectory analysis: Establish an audit analysis platform to explore abnormal user behavior. By analyzing user behavior trajectories, potential security threats can be identified and addressed on time.

2 Security Defense Objectives and Principles of User Information

Obtaining the security defense objectives and principles of user information is the foundation and prerequisite for researching the security defense of user big data. Firstly, it is necessary to clarify the security defense objectives of user information and determine the direction and goals of research, to effectively prevent security risks such as leakage, tampering, and misuse of user information. Secondly, it is necessary to formulate principles for the security defense of user information, such as confidentiality, integrity, availability, and the principle of least privilege. This allows the research to follow these principles and protect the information security of users. Therefore, obtaining the security defense objectives and principles of user information is a necessary step in researching the security defense of user big data.

The security defense of user information is to achieve the ideal goal, which includes privacy, integrity, availability, and controllability.

(1) Privacy

Privacy is an important attribute of information security, which means that unauthorized users have no right to access sensitive information. In cloud computing, it is necessary to prohibit unauthorized access to information and prevent permission holders from transmitting the information they access to the outside world, to avoid information leakage. To protect privacy, encryption algorithms can be used to encrypt sensitive information [11–13]. Among them, the Tiny Encryption Algorithm (TEA) micro encryption algorithm is a commonly used encryption algorithm, which groups the data that needs to be encrypted and achieves data encryption through iterative loops. By encrypting data, it can prevent unauthorized users from accessing sensitive information and improve privacy defense capabilities, the TEA encryption expression is:

$$s_n = \begin{cases} \underset{loop=f(e)}{\overset{31}{TEA}} \left(\underset{k_e=(e,n)}{RSA} \left(\underset{loop=0}{\overset{f(e)-1}{TEA}} (m_1) \right) \right) & \text{When n} = 1, \text{the key is } k\,(e) \\ \underset{loop=f(m_{n-1})}{\overset{31}{TEA}} \left(\underset{k_e=(e,n)}{RSA} \left(\underset{loop=0}{\overset{f(m_{n-1})-1}{TEA}} (m_n) \right) \right) & \text{When } n \succ 1, \text{the key is } h\,(m_n - 1) \end{cases}$$

$$m_n \begin{cases} \underset{loop=32-f(e)}{\overset{31}{TEA}} \left(\underset{k_d=(d,n)}{RSA} \left(\underset{loop=0}{\overset{f(e)-1}{TEA}} (s_1) \right) \right) & \text{When } n = 1, \text{the key is } k\,(e) \\ \underset{loop=32-f(m_{n-1})}{\overset{31}{TEA}} \left(\underset{k_d=(d,n)}{RSA} \left(\underset{loop=0}{\overset{f(m_{n-1})-1}{TEA}} (s_{n-1}) \right) \right) & \text{When } n \succ 1, \text{the key is } h\,(m_n - 1) \end{cases} \tag{1}$$

In the formula, e represents clear text; $f\,(m_{n-1})$ represents 64-bit encryption of the m_{n-1}-th data stream; $k_e = (e, n), k_d = (d, n)$ represent 128 bit encryption key and plaintext, respectively; s_n represents ciphertext.

(2) Integrity

Integrity refers to the protection of user information from unauthorized modification. The purpose of integrity is to protect the initial state of digital information and maintain the true face of user information. If the user information is deliberately tampered with, inserted, deleted, and other operations, the reformed information will no longer be valuable [14–17]. Using digital signature technology to verify the integrity of data, the steps for generating a digital signature are:

1. The sending end of A will use the SHA-1 algorithm to obtain the information summary M in plaintext k_d.

2. A uses the private key of the sender d_a to encrypt the information digest M and obtain the signature C.

3. Encrypt plaintext information k_d using AES algorithm and key $k\,(e)$ to obtain ciphertext C_a.

4. AES key C_h encrypted with recipient public key D.

5. Merge C, C_a, and C_h and send them to receiver B.

After generating the digital signature, it needs to be verified to determine the integrity of the data. The AES key C_h is decrypted using the private key d_b of receiver B to obtain key d. The AES algorithm is used to decrypt the ciphertext C_a to obtain plaintext $k_d^{'}$. The SHA-1 algorithm is used to obtain its information digest M'. The sender's public key D is used to decrypt the digest signature C to obtain digest result M.

By comparing M and M', if $M = M'$, the signature is correct, which verifies the integrity of the data.

(3) Availability

Availability refers to the right of digital information owners to use information on time when they need it. Availability is a new requirement for information security protection in the network era, and it is also an information security requirement that cloud computing must meet.

(4) Controllability

To achieve the goal of user information security, it should follow some basic principles to achieve the ideal protection effect [18,19]. There are three main principles: the principle of minimization, the principle of authority distribution, and the principle of security isolation.

(1) The principle of minimization requires limiting the management and access permissions of user information to the necessary range, and not granting excessive permissions to minimize security risks. This means that access and management permissions to sensitive user information can only be obtained when specific needs and authorizations are met.

(2) The principle of permission allocation requires the reasonable allocation of user information management permissions, ensuring that different management personnel form a mutually restrictive and supervisory situation, and jointly assume responsibility for information security. This can prevent the risk of authority abuse and information leakage [20,21], ensuring that information is managed legally and appropriately.

(3) The principle of security isolation is to isolate digital information from the outside world and within user information, ensuring that only authorized users can access and manage data. This includes appropriate access control and management of data to prevent unauthorized access and tampering. The principle of security isolation lays the foundation for information access control and management and helps to protect the security of user information.

Adhering to these principles can effectively protect the security of user information and reduce the risk of information leakage, abuse, and tampering. At the same time, it can also ensure that user information can be obtained and used by authorized users when needed, thereby achieving the goal of user information security defense.

To achieve the goal of user information security, it should follow some basic principles to achieve the ideal protection effect. There are three main principles: the principle of minimization, the principle of authority distribution, and the principle of security isolation.

(1) Principle of minimization

The protected user information is only allowed to be managed within the appropriate scope. On the premise of complying with the prescribed security regulations and laws, the appropriate management authority of user information obtained by the management personnel is called the principle of minimization. The access and management of user information must be limited, which is a kind of restrictive authority.

(2) Restriction principle of authority distribution

In the security management of user information, it should use reasonable and appropriate authority allocation for user information management. The purpose of this is to enable authorized users to have a part of authority, which is to form a situation of mutual restriction and supervision between information managers and to share the responsibility for digital information security. If the authorized subject has too much authority, it will lead to absolute control of information, and the

transitional control authority will form the abuse of digital information, which will inevitably bring considerable security risks.

(3) Principle of security isolation

Isolation and control of user information is a common technical means to protect information security, and isolation is the premise of good control. The principle of security isolation includes the isolation of digital information itself from the outside world and the isolation of user data within user information. The isolation of digital information lays a good foundation for access control, that is, it can realize the authorized access and management of information.

3 Integration of Heterogeneous Data Sources Based on Metadata Correlation Characteristics

Heterogeneous data source integration refers to the integration and unification of data from different data sources for better analysis and utilization. In the security defense research of user big data, adopting heterogeneous data source integration technology can solve the following problems:

(1) Diversity of data sources: User big data generally includes data from different channels, such as social network data, mobile application data, sensor data, etc. The use of heterogeneous data source integration technology can integrate data from these different channels, providing a more comprehensive information foundation for data analysis.

(2) Heterogeneity of data structure: The data structure of different data sources may vary greatly, such as the collected data format, fields, etc. By adopting heterogeneous data source integration technology, these data can be standardized and processed, making the structure of the data more unified, thus facilitating data analysis.

(3) Data quality issues: There are differences in data quality between different data sources, such as data accuracy, completeness, and accuracy. Heterogeneous data source integration technology can improve the quality of data through methods such as data cleaning and filtering, and reduce errors and biases in data analysis.

Therefore, adopting heterogeneous data source integration technology can integrate and unify data from different channels, improve the efficiency and accuracy of data analysis, and thereby enhance the security defense ability of users' big data.

Metadata is data that describes other data, or structured data used to provide information about a resource. In the storage field, metadata describes certain aspects or information of data, in other words, metadata is data that describes and indexes data. In file systems, metadata usually refers to data that describes basic attribute information of a file, such as file size, creation time, access time, modification time, and the file's owner, which is the metadata information of the file. Some file systems also provide some extended attributes, such as file backup date, conversion times, operation process number, etc., which can also be called metadata information of the file. Early definitions of metadata are: metadata is a concept in the database management field, that is related to the organization of data; metadata is a description of data and an explanation of data items in a data collection to enhance the value of data utilization; metadata is something between data and information, which can transmit both data and information. The concise and well-known definition of metadata is "data about data". This definition reveals the essential characteristics of metadata and is the core definition of metadata. If the metadata of each data source is extracted according to a unified standard stored in a meta base, and mapped to the user mode established according to the user's query requirements, the corresponding data source mode query can be obtained by parsing the user mode; the query results of each data source are connected, merged and output according to the user mode, so that data sharing and integration can be

realized [22,23]. Based on the above analysis, a metadata-based data resource sharing and integration scheme is proposed. By generating mapping relationships between user patterns and data source metadata, the transformation between user patterns and data source patterns is achieved, thereby eliminating obstacles to data integration. Next, by generating query statements and completing the query, the goal of integrating heterogeneous data according to user needs was achieved. In the process of generating query statements, it is necessary to utilize the mapping relationship between user patterns and metadata to transform user query requirements into query statements that can be understood by the data source, to complete the query and integration of heterogeneous data [24–26].

3.1 Establishment of User Mode

Due to the deepening of research and the demand for data integration, there are some recognized norms in the relevant knowledge field, such as the Gene Ontology (GO) of bioinformatics, which makes the description of data semantics of each data source have a unified reference standard, to facilitate the access and understanding of data, and lay a software foundation for data integration. However, due to different application purposes and backgrounds, "the same concept can get completely different attribute information from different data sources". Therefore, for users, data integration is to clean up and integrate the data obtained from various data sources according to the query requirements, and use the GO and other specifications as the connection medium to convert it into the desired style [27,28]. The expression for data integration is:

$$D = \frac{\sum_{i=1}^{N} D_s W_t N}{\sum_{i=1}^{N} W_t W_p N} \tag{2}$$

In the formula, D represents the final integration result; W_t represents the correction coefficient given for data deviation in integration; D_s represents a natural resource data parameter within the system during the integration process; N represents data richness; W_p represents the support level of data during the integration process.

The data after integration still exists in the form of a "table" or "view" conceptually, and the data of each column comes from different heterogeneous data sources. Errors may occur due to different types during integration. Therefore, it is necessary to establish the corresponding user mode according to the needs of users and unify the data type and name of the query. We use DBMS's view management form for reference, so that users can define and manage "table name", "field name", "type" and "field length" and other related contents when establishing their mode, and can establish multiple virtual tables according to different query requirements, which can be adjusted at any time according to needs [29,30]. The "field" of the user mode is the attribute information required by the user. Its type and length reflect the user's requirements for query results. This way of creating a virtual table is equivalent to defining metadata of user mode, so it is the same form as metadata of data source in storage, which is the data in the database of integration scheme. This also prepares establishing the mapping between user patterns and data source metadata.

3.2 Generation of Mapping between User Patterns and Data Source Metadata

Only the user pattern is not enough. It must be associated with the metadata of the data source to get the required data from the data source. Therefore, the mapping between user patterns and data source metadata becomes an indispensable and important step. The mapping is based on the user

mode as a reference, the user selects the tables to be integrated, and associates the metadata of the data source with the same semantics to the related user mode "fields". At this time, different data sources will generate a certain correlation with each other through the user mode as a medium [31–33]. As long as the user mode and the corresponding mapping relationship are analyzed, each table can be obtained. The actual query statement of the data source realizes the query of the heterogeneous data source. The mapping expression between data source metadata is:

$$S = (O, D, P, P') \tag{3}$$

In the formula, O represents the dataset ontology; P represents the mapping between ontology and metadata; P' represents the mapping between metadata and ontology.

It should be noted that this kind of mapping is unique to the determined user mode and data source. It is not allowed that the same "field" of the user mode corresponds to multiple fields in the data source table or that the same field in the data source table corresponds to multiple "fields" of the user mode.

Extracting metadata correlation features from heterogeneous data first involves dimensionality reduction to effectively reduce data redundancy [34,35]. The gradient descent algorithm introduces a fixed momentum term to obtain a low dimensional data representation of heterogeneous data, completing dimensionality reduction of heterogeneous data. The formula is as follows:

$$V = \chi \frac{F}{u} + \delta(v)\, S\left(V^{(v-1)} - V^{(v)}\right) \tag{4}$$

In the formula, V represents the heterogeneous data after dimensionality reduction; δ represents the local minimum value of heterogeneous data; χ represents the momentum term; v represents the distribution interval of t.

Input the dimensionality of reduced heterogeneous data into the correlation model, and achieve feature extraction of heterogeneous data through the maximum information coefficient between the data [36,37]. Fit the maximum information coefficient using the stretching index distribution to obtain the attribute features of heterogeneous data, as follows:

$$\begin{cases} b = d^{-\left(\frac{1}{a_0}\right)\cdot k} \\ 0 \le MIC(V) \le 1 \\ k \in [0, 1] \end{cases} \tag{5}$$

In the formula, d represents the attribute feature; V represents a heterogeneous dataset; a_0 represents the scale parameter; MIC represents the maximum information coefficient; k represents the extended parameter.

3.3 Query Statements

By analyzing the user mode and its corresponding mapping relationship, it can decompose the query of user mode into the query of heterogeneous data sources, and then integrate and clean the query results of each data source according to the user mode, which is the desired result of user mode [38–41]. The specific steps are as follows:

(1) According to the needs of user mode reduction, generate a user view, and select the corresponding mapping relationship.

(2) Analyze the user view, determine the fields to be queried for each data source according to the "fields" of the user view and the corresponding mapping relationship, and use the "fields" of the user view as the alias of the fields in the data source table; if the "fields" of the user view have no mapping relationship in the data source, the data source fields are replaced by NULL or empty values.

(3) Generate the query statements of each data source according to the determined data source fields.

(4) According to the sequence of each data source in the mapping, externally connect the query statement from the beginning to the end with the key word of the field where the recognized standard content is stored. Externally connect each connection result with the next query statement as a new query statement and eliminate the duplicate word segments until a complete query statement is generated.

(5) Execute the generated query statement, complete the query, and output according to the style of user view.

According to the design concept, the schematic diagram of the integrated query system is shown in Fig. 1.

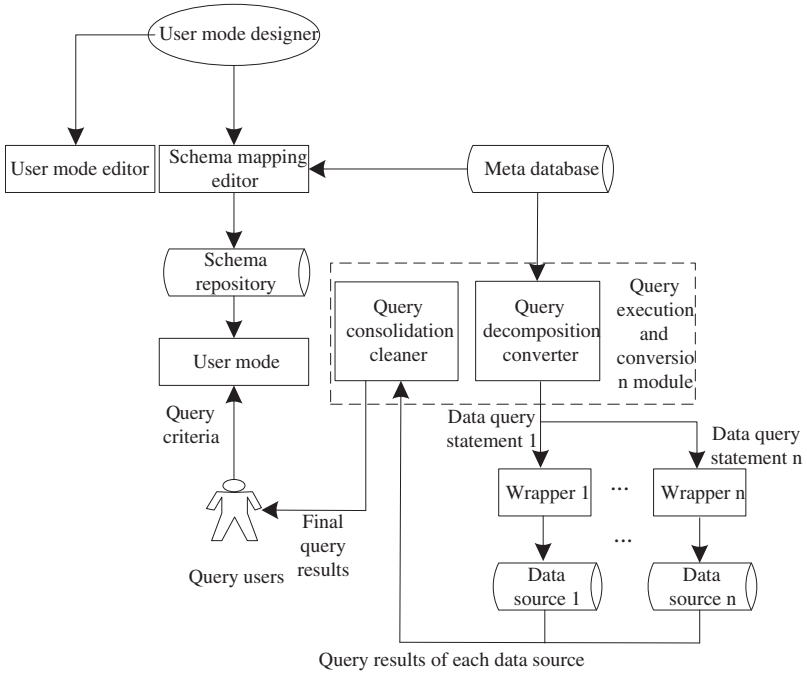

Figure 1: Schematic diagram of integrated query system

On this basis, the specific algorithm is given below:

```
input: query point q, Global parameters global_param, k
output: Query result set result_set
point_topk_query(query q, global_param, k)
qset=Ø;
result_set=Ø;
```

(Continued)

Algorithm (continued)

```
 vector temp[]=NULL;
 for each i=0 to param.k
   do
temp[i]=compute_lsh_hash(q, 0, i);
end for
 group_id=compute_group_hash(temp[]);
fetch meta from group=group_id;
 add meta to qset;
for each i=0 to qset.num
   do
 if query_type=POINT_QUERY
              do
                    get meta where meta=query;
      else if query_type=KNN_QUERY
              do
 get k metas  where metas  are k-nearest from query;
      add meta to result_set;
           end if
end for
return result_set;
end
```

4 Security Defense System of Customer Information Based on Machine Learning

4.1 Security Threat and Defense Technology Based on Machine Learning

4.1.1 Machine Learning and Common Security Threats

Machine learning utilizes experience to improve the performance of algorithms through computational means. Experience is data. Computer systems use existing data to learn, generate models, and then make decisions on future behavior [42–44]. The solution paradigm of the machine learning approach is shown in Fig. 2.

The process of machine learning can be divided into two stages: training and prediction. Machine learning can be divided into supervised learning and unsupervised learning. The training data set with supervised learning is labeled, and supervised learning is mostly used for classification and regression problems, such as image recognition, spam classification, etc. The unsupervised learning training data set is not labeled, and unsupervised learning is mostly used for clustering, and data dimensionality reduction, etc., such as pre-training, intrusion detection, etc. After the training stage, the target model is obtained, and people can use the target model to predict and make decisions [45].

In our daily lives, machine learning is more and more popular. The security threat of machine learning seriously impacts normal life and even health. Table 1 shows the common security threats, which are respectively for the training stage and prediction stage in the machine learning process, as shown in Fig. 3.

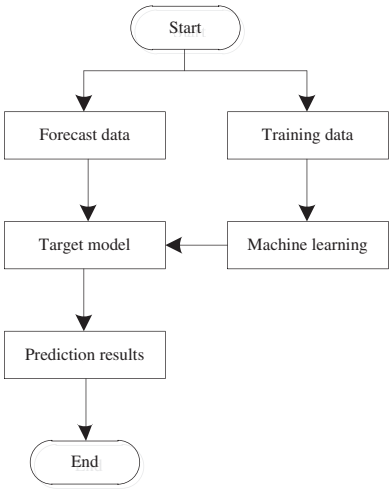

Figure 2: Machine learning problem-solving process

Table 1: Common security threats of machine learning

Stage	Training phase	Prediction stage
Adversary strategy	Poisoning attack	Counter attack
Enemy target	Integrity/availability	Integrity/availability
Adversary ability	Modify training data	Make a counter sample
Adversary knowledge	Limited knowledge	Black box/white box

4.1.2 Security Defense Algorithm Based on Machine Learning

The security problem of machine learning threatens the availability and integrity of the algorithms, and the adversary usually relies on the different distribution of test data and training data to launch attacks. Improving the robustness of the target model can have good adaptability to unknown samples in the prediction stage, and the model can also predict normally when malicious samples appear [46].

When defending, the defender plays a game with the opponent. The defender makes a defense strategy and the opponent makes the best response. The defender cost function is:

$$E_{xy} \sim D\left[l_h(\Delta(X), y)\right] \tag{6}$$

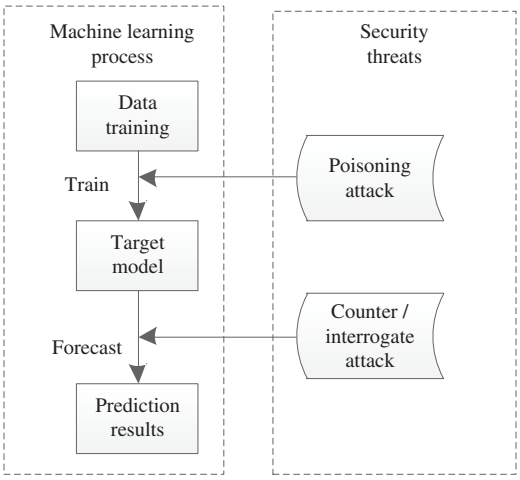

Figure 3: Common security threats of machine learning

The adversary cost function can be:

$$E_{xy} \sim D\left[l_h^{\ a}\left(\Delta\left(X\right), y\right) + c\left(x, \Delta\left(X\right)\right)\right] \tag{7}$$

where, $\Delta\left(X\right)$ represents the function of the opponent modifying the sample, and $c\left(x, \Delta\left(X\right)\right)$ is the cost of modifying the sample. The process of the game between the defender and the opponent increases the robustness of the target model. The game process between the defender and the opponent can be expressed as the opponent minimizing the cost of manufacturing samples, and the defender minimizing the cost function in the presence of the opponent samples, as shown in Eq. (8):

$$\begin{aligned} &\min \sum_{i=1}^{n} E_{x,y} \sim D\left[l_h\left(\Delta\left(X\right), y\right)\right] \\ &\Delta \in \operatorname{argmin}_{\Delta} \sum_{i=1}^{n} E_{x,y} \sim D\left[l^a{}_h\left(\Delta\left(X\right), y\right) + c\left(x, \Delta\left(x\right)\right)\right] \end{aligned} \tag{8}$$

In addition to improving the robustness of the model, direct rejection of malicious samples is also an important means to improve the security of the model. Data cleaning can be used to directly remove the poison data to prevent a poison attack. In the same way, it can also directly discard the data which is determined as the counter sample to resist the counter attack. If the detection and prediction samples are legal samples, the prediction will be carried out, and if they are counter samples, they will be discarded directly, achieving a certain effect in resisting counter attack.

4.2 Construction of Security Defense for Customers Information Based on Machine Learning

To do a better job in customer information security defense, solve the problems that operators have not been able to effectively solve for a long time, such as low efficiency and not comprehensive in customer sensitive information identification and personnel operation audit, based on many years of research and practice in building the security defense capacity of customer information, and according to the overall structure of customer information access of operators (Fig. 4), the former background personnel operation, user behavior tracking, and other application scenarios are the entry points, and sensitive information is quickly identified by using automation tools to visually display the situation awareness of sensitive data. At the same time, flow processing, big data, and other technologies are

used to build an audit and analysis platform for the operation behavior of front and back office personnel. Through clustering analysis, decision trees, and other machine learning algorithms, it can monitor the data flow, conduct real-time analysis of user operation, detect and warn abnormal behaviors, and verify and confirm suspected abnormal behaviors to form a closed-loop solution [47–49].

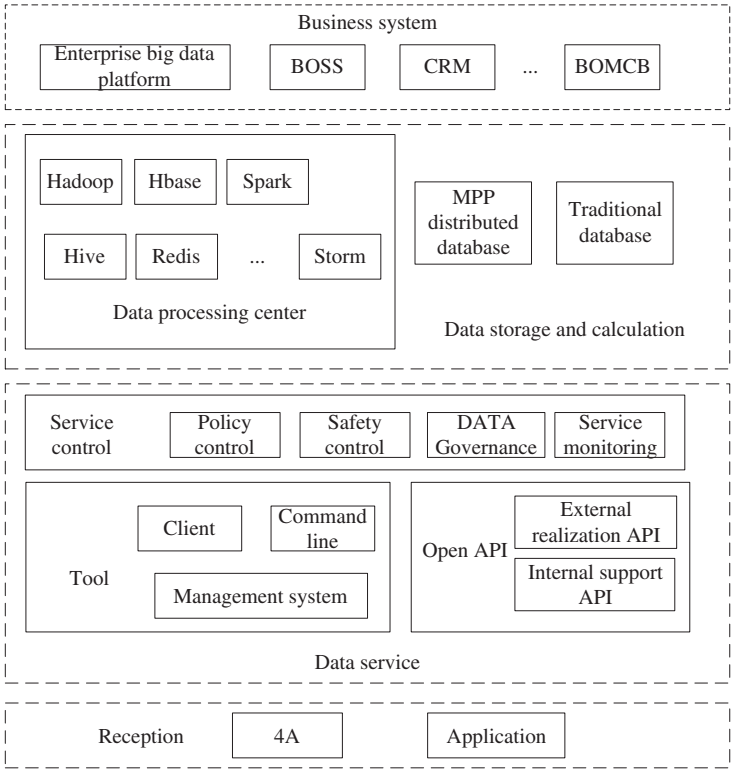

Figure 4: Architecture of machine learning customer information security

This scheme realizes the whole process of security control of automatic discovery and grading of customer sensitive information in advance, real-time monitoring and early warning of sensitive data in the event, intelligent and efficient security audit analysis after the event, and verification and confirmation of abnormal behaviors, to protect the sensitive data to the maximum extent and ensure the security of customer information.

4.2.1 Definition and Classification of Customer Information

The definition and classification of customer information is the foundation of the construction of a customer information security defense algorithm. Customer information refers to the information collected by telecommunication operators in the process of providing services, which can identify the personal identity of customers and involve the personal privacy of customers independently or in combination with other information, including but not limited to the personal identity information such as customer name, date of birth, ID number, address, etc., as well as the information such as number, account number, password, time, place of using services. Specifically, it includes three categories: user identity and authentication information, user data and service content information, and user service-related information.

User identity and authentication information: includes but is not limited to the user's natural person identity and identity information, user virtual identity, user authentication information, etc.

User data and service content information: includes but is not limited to the user's service content information, contact information, user's private data, private social content, etc.

User service related information: includes but is not limited to business ordering relationship, service record and log, consumption information and bill, location data, violation record data, terminal equipment information, etc.

Based on the "Guidelines for Grading the Protection of Personal Information of Telecom and Internet Service Users" and combined with the business characteristics of the telecommunication industry, the typical definition and grading standards of customer information for telecommunication operators are listed in Table 2. According to the sensitivity of customer information, it is divided into four levels: extremely sensitive, sensitive, more sensitive, and low sensitive. Based on the principle of classified and graded control, the security control requirements for different sensitive data and the corresponding range of sensitive personnel are determined.

Table 2: Definition and classification standard of customer information

Classification of sensitive information	Subclass information	Related data	Principle of security and defense
Low sensitive level	Business ordering relationship	Basic business ordering relationship: brand, package customization, etc.; value-added business ordering relationship: registration, modification, cancellation of value-added business such as address book, Laixian, RBT, etc.	Implement basic technology and management measures to ensure data life cycle security
	Violation record	User violation records, such as spam messages, harassment calls, and other records; Records of business violations, such as port abuse, illegal channels, bad website domain names, etc.	
Sensitive level	Consumption information and bill	Consumption information: such as shutdown, access time, points, credit rating, payment, balance, and transaction history; Bill information: communication expense, overdue information, data expense, collection expense, etc.	Implement necessary technical and management measures to ensure data life cycle security and establish data security management specifications

(Continued)

Table 2 (continued)

Classification of sensitive information	Subclass information	Related data	Principle of security and defense
	Tag terminal equipment	Use the unique mobile device identification code IMEI, device MAC address, and SIM card information to accurately locate the device information.	
	Terminal equipment information	Terminal model, brand, manufacturer, use time, preset/installed software application, etc.	
High sensitive level	Identification of the natural person	Customer name, certificate number, driver's license number, bank account, customer entity number, etc., can accurately locate the personal data of the entity customer.	Strictly implement technical and management measures, ensure data confidentiality and integrity, and data access control security, establish data security management specifications and data quasi real-time monitoring mechanism
	Network identity	Contact number, email address, network customer number, instant messaging account, network social user account, etc., can accurately identify the information of network/communication users.	
	User profile	Customer occupation, work unit, age, gender, native place, hobbies, etc.; province, industry, group signing time and agreement expiration time of group customer, basic data of unit member personal and user social life entity number, etc.	
	Service content data	Telecommunication network service content data: SMS, MMS, voice and other communication content; mobile Internet service content information: including call content, intra group publishing content, data file, email content, user online access content, etc., involved in mobile Internet services such as WeChat, QQ, mailbox; private text and multimedia stored or cached by user cloud storage, SDN, IDC, etc., Body data information	
	Contact information	User data such as user address book, friend list, group list, etc.	

(Continued)

Table 2 (continued)

Classification of sensitive information	Subclass information	Related data	Principle of security and defense
	Service records and logs	Service details and signaling: including voice, SMS, MMS and GPRS details, user calling number, calling home, called number, communication time, time, traffic, and other information; Mobile Internet service record: including cookie content, online log, APP connection, etc., including calling number, website, online shopping record, etc.	
	Location data	Accurate location information, approximate location information	
Extremely sensitive level	Entity identification	Photocopy of ID card, passport, driver's license, business license, etc.; fingerprint, voice print, iris, etc.	Implement strict technical and management measures, protect the confidentiality and integrity of data, ensure the security of data access control, and establish strict data security management specifications and real-time data monitoring mechanism
	User privacy information	Disclose the user's private information related to personal race, family member information, residential address, religious belief, gene, personal health, private life, and other information prohibited from the public by laws and administrative regulations such as the regulations on the administration of credit investigation industry	
	User password and related information	User's network identity password and associated information, such as mobile customer service password, email password, WeChat password, mobile WLAN password, and password protection answers associated with these passwords, etc.	

4.2.2 Automatic Recognition and Perception of Sensitive Data

It is the first step of customer information's security defense to identify sensitive data in complex and comprehensive massive data. Only knowing the storage location, storage form, and application scenario of sensitive data can help administrators configure appropriate authorization policies and protection measures.

The traditional method based on manual identification of customer-sensitive information mainly depends on the personal experience of the risk assessor. In this way, firstly, when facing a large number of data, the carding speed cycle is long, the recognition speed is slow and not comprehensive; secondly, it mainly depends on the subjective judgment of people, and the evaluation criteria are not uniform. Based on the definition and classification criteria of sensitive information, automatic tools and interfaces are developed. Through rule configuration, fuzzy matching, natural language analysis, and other means, the automatic, rapid, and comprehensive recognition and classification of customer-sensitive data are realized, and the situation awareness of sensitive data is displayed in a visual panorama.

(1) Sensitive data identification and classification

Sensitive data automation tools are mainly based on keyword matching of metadata, regular expression matching of data content, and natural language analysis to realize the automatic discovery and classification of customer-sensitive data.

(a) Sensitive data recognition based on Metadata

First of all, the keyword matching of sensitive data is defined. Through accurate or fuzzy matching table field names, annotation, and other information, metadata information is used to match the database tables and files one by one. When the fields meet the keyword matching, sensitive data is judged and automatically graded. This matching method has the advantages of low cost and quick effect and can identify more than 50% of the sensitive data of customers in the whole network.

(b) Sensitive data identification based on data content

Some temporary tables or sensitive tables developed in history that are not established according to the specification cannot be judged as sensitive data according to the metadata, which is more determined by analyzing the data content. The automation tool obtains these tables by scanning and matching a large number of sensitive information (such as mobile phone number, ID card number, mailbox, etc.) of numerical type and English type in the system through pre-defined regular expression to make sensitive data and its level judgment.

(2) Situation awareness of sensitive data

Identifying the sensitive data can help the administrator to have a global understanding of the sensitive data of the whole system through situation awareness and blood relationship tracking of the sensitive data, and realize the real-time monitoring and early warning function of user/application behavior employing machine learning and other technologies, to reduce the risk of customer information leakage and malicious destruction of core data.

According to the sensitive level of customer information, the number of sensitive fields, the number of records, the frequency of visits, and other parameters, a data-sensitive level evaluation model is constructed to evaluate the overall sensitivity level of the host file directory, host and database tables, and database, and to display them in the security situation awareness platform in a centralized way. Through icon size, it can show the number of sensitive information of different hosts and their directories, different databases, and tables, to realize the panoramic situation awareness of sensitive data of the system.

4.2.3 Construction of Behavior Audit Analysis Platform

By analyzing the logs of user operations, we can find different abnormal behavior operations, such as the user's non-working time login, abnormal IP login, and so on. Based on big data technology,

a behavior audit analysis platform is built. Through the method of log data tagging, a flexible and customized audit model can be built to quickly support the analysis of application scenarios such as personnel operation behavior portrait and behavior trace analysis.

The architecture of the audit analysis platform for user behavior is shown in Fig. 5. Based on the Hadoop distributed architecture, the platform matches the user operation logs of each system with the data model based on the real-time data flow and realizes the centralized audit analysis of the user operation logs using centralized storage of logs, big data modeling analysis, and fast retrieval.

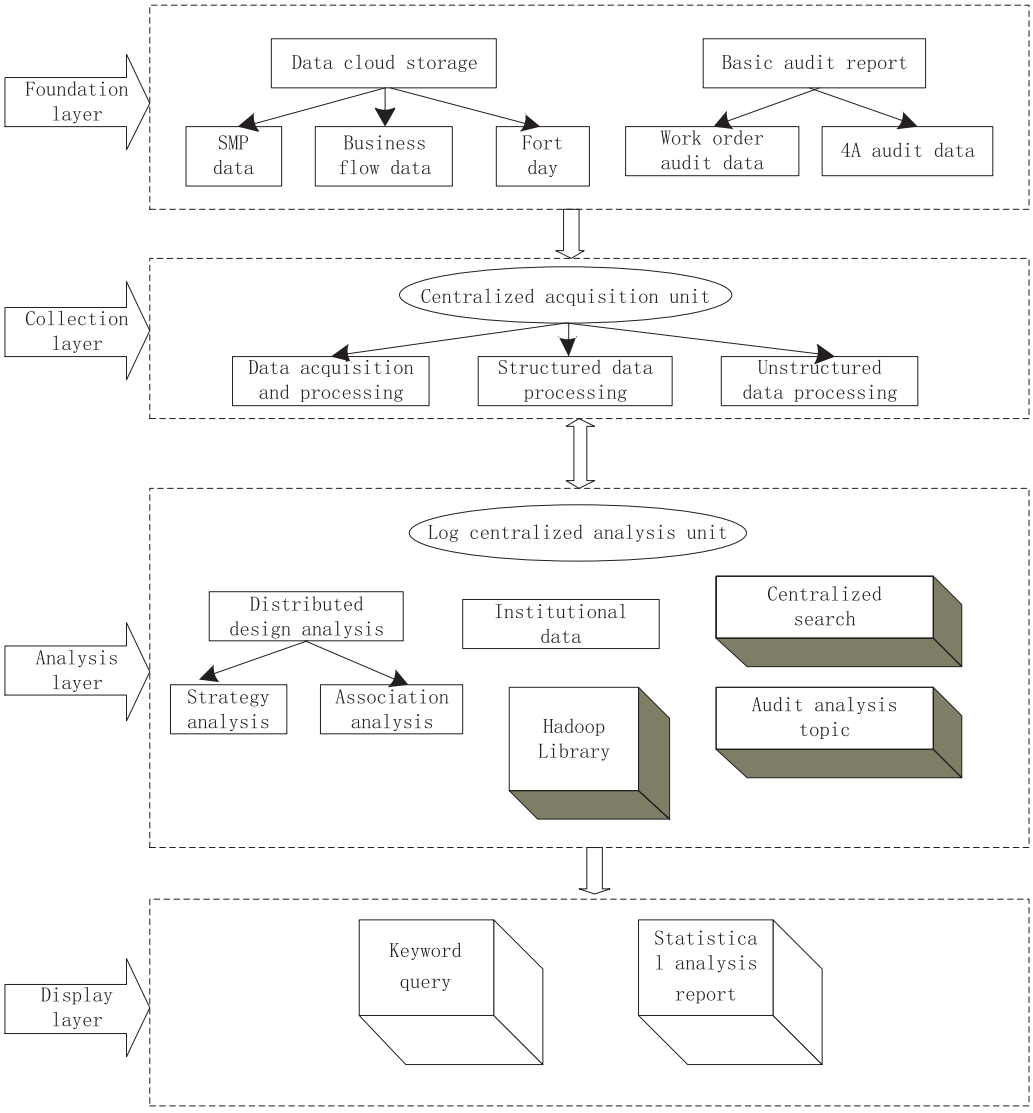

Figure 5: Architecture of audit analysis platform for user behavior

4.2.4 Analysis of User Behavior Trace

Based on the system resource log data, application resource data, and other information, the deep mining of data is carried out to traverse the paths frequently used by users from the user access

behavior. Through quasi real-time analysis of the user's behavior track, some operation changes of users are found in advance, and through these changes and the standard access path in the existing knowledge base, the possibility of risk occurrence is predicted.

Based on the above analysis, a complete user operation behavior trace can be established, to realize the complete event description of "person", "source terminal", "access channel", "access resource" and "what operation", effectively support the abnormal behavior monitoring such as bypass access to sensitive data, and improve the level of customer information's security prevention and control. At the same time, the user's behavior track analysis changes the previous audit to find security risks from some pre-set abnormal scenarios, and links up the normal operation events of users to find possible abnormal problems, which greatly improves the predictability of the information security defense system.

5 Experimental Analysis

To verify the user's information security defense effect of the proposed method, this paper will carry out experimental verification from three directions: data storage security, data disaster recovery security, and industrial control security. The test data comes from the NSL-KDD dataset. The initial learning rate of the experimental model is set to 0.001, and the number of iterations is 64. The simulation software MATLAB 2020b is used for testing. Data integrity, data accuracy, and intrusion success rate are selected as the test indicators, and literature [7] and literature [8] methods are used as comparison methods for experimental testing. The experiments are as follows:

5.1 Security Detection of Data Storage

The integrity of data storage is tested by the methods of reference [7], reference [8] and this paper, and the results are as follows:

Analysis of Fig. 6 shows that the data storage integrity is different under different methods. When the data amount is 5×10^3 bit, the data integrity of the method in reference [7] is 85%, the data storage integrity of the method in reference [8] is 78%, and the data storage integrity of the proposed method is 98%, showing that the data storage integrity of the proposed method is the highest. When the data amount is 50×10^3 bit, the data integrity of the method in reference [7] is 72%, the data storage integrity of the method in reference [8] is 71%, the data storage integrity of the proposed method is 92%, and the data storage integrity of the method in this paper is always at a high level, which shows that the data storage of the proposed method is relatively safe.

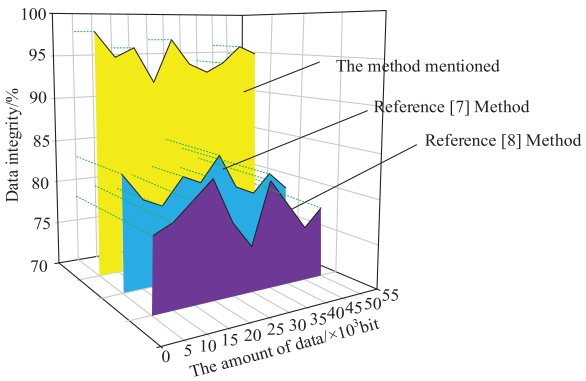

Figure 6: Comparison of data integrity

5.2 Security Detection of Data Disaster Recovery

To verify the data disaster recovery performance of the proposed method, the recovery situation after data failure is simulated, and the data accuracy results under different methods are as follows:

Analysis of Fig. 7 shows that there are differences in data accuracy under different methods. When the amount of data is 15 $\times$ 10^3 bit, the data accuracy of the method in reference [7] is 73%, the data accuracy of the method in reference [8] is 70%, and the data accuracy of the proposed method is 97%, showing that the data accuracy of the proposed method is the highest. When the data volume is 40 $\times$ 10^3 bit, the data accuracy of the method in reference [7] is 82%, the data accuracy of the method in reference [8] is 71%, and the data accuracy of the proposed method is 98%, showing that the data accuracy of the proposed method has always been at a high level, with better data disaster recovery performance.

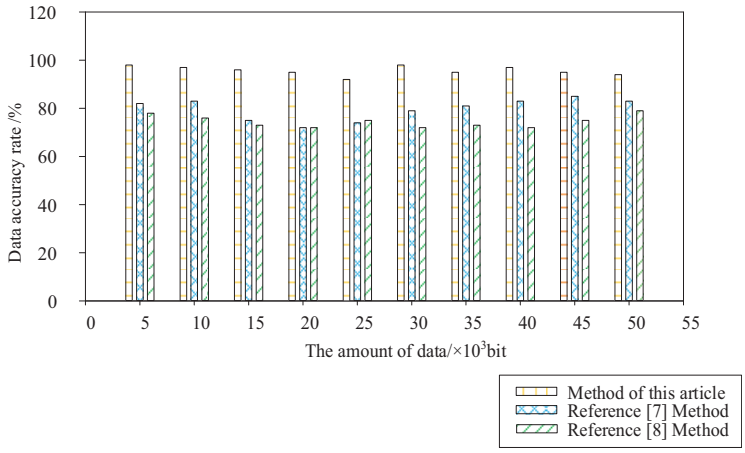

Figure 7: Data accuracy under different methods

5.3 Air Traffic Control Safety Inspection

To verify the security performance of the research method in this paper, we compared the failure rate of data intrusion under different methods using the methods from reference [7] and reference [8]. The results are shown below.

According to Fig. 8, when the data volume is 100 bit, the data intrusion success rate of the method in reference [7] is 38%, that of the reference [8] method is 67%, and that of the proposed method is only 1%. With the increase of data volume, when the data volume is 1000 bit, the data intrusion success rate of the method in reference [7] is 36%, the data intrusion success rate of the method in reference [8] is 52%, and the data intrusion success rate of the proposed method is only 2.6%, which shows that the proposed method can effectively resist data intrusion and has high industrial control security.

5.4 Visual Verification of Security Defense Performance of Research Methods

To make the simulation experiment more convincing, various types of viruses were introduced in the experiment, which are commonly seen in user data storage, as follows.

Port scanning for viruses: Send virus packets to all ports of the client host to determine whether the port is open. If it is open, inject viruses into it.

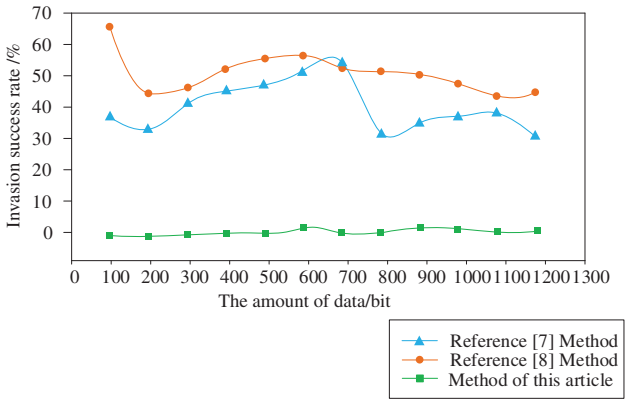

Figure 8: Comparison of intrusion success rate

IP scanning for viruses (circled in Fig. 9): Find the target network and send virus packets to it to determine which target can be attacked.

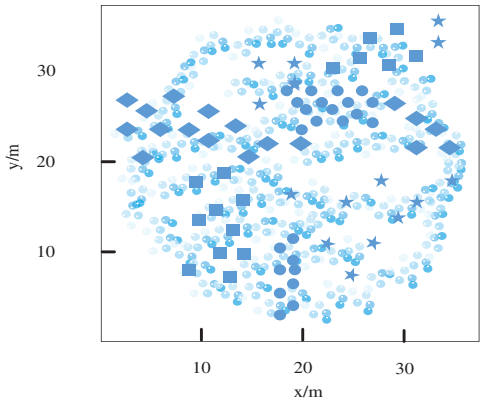

Figure 9: Distribution of various virus data

Apache 2 virus (marked with a square in Fig. 9): The server receives a large number of requests, slows down server processing speed, and even consumes all resources.

Virus detection is the top priority of virus defense, and the detection results directly affect the effectiveness of virus defense.

Smurf virus (marked with a diamond in Fig. 9): Sends a forged address to the target host, which requires the host to respond to the fake source host, resulting in the inability to process legitimate data packets promptly.

Neptune virus (marked with a pentagram in Fig. 9): Sends false messages to the client, and the host needs to confirm each message, causing a large number of ports to be illegally occupied.

According to Fig. 10, The results of this experiment indicate that this research method not only detected all viruses in user data storage but also centrally clustered each virus. Based on the aggregation results, it can clearly distinguish which viruses were invaded during data storage, further improving the security defense effect of user big data. This is achieved by using correlation feature extraction of user metadata and integrating heterogeneous data sources based on machine learning

algorithms. A machine learning-based user information security defense system was built, which defines and classifies user information into different levels, and automatically identifies and perceives sensitive data. The system achieves autonomous detection of sensitive data and classification of user information.

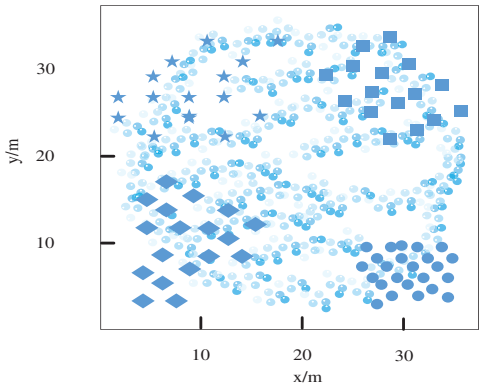

Figure 10: Virus integration detection results of the research method

6 Countermeasures for User's Information Security Defense Management

The security defense of user information is a systematic project, involving management, process, technology, and other aspects. The information security defense of user information depends on technology and management. If technology represents the strength of information security defense, then management represents the attitude, will, and determination for information security. Management is the management of internal organizational structure, including personnel organization and system constitution. In addition, management should also include external management, such as the IT Audit Association established by the government and industry associations, education, and publicity of security policies.

6.1 To Strengthen the Self-Discipline of the Network and the Protection Awareness of the Personal Information of the Network Users

Strengthening the awareness of protecting personal information, should not register some unknown websites wantonly, most of the time, it is the network users themselves who inadvertently disclose their personal information, so we must strengthen the awareness of protecting their information because they are the first person responsible for personal information. At the same time, it should strengthen the awareness of personal computer defense, and install genuine anti-virus software and firewall, to prevent hackers from intruding, leading to personal privacy data leakage and property theft; when shopping online, we must check whether the web address has a tick mark, so that the web address can be safe, develop good network use habits, and effectively prevent the leakage of personal information.

6.2 To Strengthen Network Moral Construction and Industry Self-Discipline, Establish and Improve Personal Information Security Protection and Prevention Mechanism

The protection of personal information security not only depends on the law but also needs the moral consciousness and self-discipline consciousness of the practitioners. Therefore, we should

strengthen the construction of network morality, restrict people's behaviors on the network with moral standards, and make network morality a standard when people implement behaviors on the network. We must strengthen self-restraint, improve the awareness of protecting customer information, the technical means, and the relevant information management system, to sort out and improve the user's personal information security management system and process, and establish and improve various management systems and norms that infringe on user's personal information. The violation of the personal information of users should be fast, accurate, and complete. Fast: it should ask the relevant person in charge to understand the specific situation the first time, form a preliminary treatment plan, and put it into practice. Accurate: it requires that decision-making and handling of problems should be traced back to the source, grasp the essence of the problem, and take effective actions. Complete: the internal staff of the enterprise should work together to solve problems. Its core is to adhere to the principle of comprehensive coordination, classified management, and hierarchical responsibility under the unified leadership of the leaders of network operation enterprises. For those who violate criminal law by divulging the personal information of users, the political and legal organs shall resolutely investigate and deal with them according to law.

6.3 To Strengthen the Role of Government in Information Security Management

As the main body of information security management, the government plays a leading role in information security management. On the one hand, the government, as its administrator, formulates relevant laws and regulations for information security management. On the other hand, the government also takes the lead in establishing policies for the development of the information security industry to ensure the healthy development of the information industry. In addition, the government also plays a leading role in organization supervision, handling illegal behavior of information security, and so on. For China, it is very important to strengthen and perfect the role of government in information security management. Firstly, information security of information security networks should take the national interest as the most important consideration, and it is the responsibility of every citizen and group to safeguard the national interest. It is very important to establish and maintain the awareness of network information crisis. It helps us to have a clear understanding of the impact of network development from the strategic height of national politics. We should be prepared for the future when we are in peace. On the one hand, we should master the network, and use the network, and on the other hand, we should actively look for countermeasures. Secondly, the government-led information security management mode should be implemented. Network information security is a kind of basic security. With the deepening of social informatization, no matter people's various activities of social production and life, or state organs, various enterprises and institutions perform social management and provide social services and their normal operation, they are more and more closely combined with computers and information networks; Whether it is economic and social development, or national political diplomacy, national defense and military activities, it is increasingly dependent on a large and vulnerable computer network information system. Therefore, the management of information security should be led by the state, and the macro management and control of information security should be carried out as a whole.

7 Conclusions

This paper proposes a machine-learning security defense algorithm based on metadata association features. The algorithm realizes the integrated query of heterogeneous data sources by analyzing the mapping relationship between the user model and data source metadata. At the same time, through the monitoring of user behavior and the automatic identification of sensitive data, the control of

unauthorized users and the protection of sensitive data are realized. This research has the following practical advantages: The algorithm can improve the accuracy and efficiency of user information security defense. By establishing a user model and mapping relationship, you can identify anomalies and threats in user behaviors more accurately, and take security measures on time. At the same time, the automatic identification and perception of sensitive data reduces the need for manual intervention and improves the efficiency of defense. This algorithm can provide more comprehensive user information security protection. By emphasizing privacy, integrity, and availability, the algorithm can precisely control unauthorized users, protect the privacy of user information and data integrity, and improve the availability of information security. The algorithm also has some limitations. First, the accuracy and effectiveness of the algorithm are affected by the limitations of data quality and data sources. If the data quality is low or the data source is incomplete, the recognition effect of the algorithm may be affected. In addition, the algorithm may consume large resources, and the performance of the algorithm may need to be further optimized and improved for large data sets and real-time security defense requirements.

Future research can further explore the following directions: First, more advanced machine learning algorithms such as deep learning can be considered to improve the accuracy and efficiency of defenses. Secondly, it can be combined with other security technologies, such as blockchain, cryptography, etc., to further enhance the security and protection of user information. Finally, the laws of user behavior and threat evolution can be studied to build more intelligent and adaptive defense systems to cope with ever-changing security challenges.

In conclusion, this research provides a new way of thinking and method for user information security defense in theory and practice. Emphasizing privacy, integrity, and availability through a machine learning algorithm based on metadata association features that provide comprehensive user information security while improving accuracy and efficiency. However, the research also faces some limitations, and future research can continue to explore more advanced algorithms and security techniques to improve the capability and intelligence of defense.

Acknowledgement: I would like to express my gratitude to all those who helped me during the writing of this thesis. My deepest gratitude goes first and foremost to Professor Jianwei Zhang, my supervisor, for his constant encouragement and guidance. He has walked me through all the stages of the writing of this thesis. Without his consistent and illuminating instruction, this thesis could not have reached its present form.

Funding Statement: This work was supported by the National Natural Science Foundation of China (U2133208, U20A20161).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Ruchun Jia; data collection, analysis and interpretation of results: Yi Lin, Ruchun Jia; draft manuscript preparation: Ruchun Jia. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Aebissa, G. Dhillon, and M. Meshesha, "The direct and indirect effect of organizational justice on employee intention to comply with information security policy: The case of Ethiopian banks," *Comput. Secur.*, vol. 130, pp. 103248, 2023. doi: 10.1016/j.cose.2023.103248.

[2] J. Guo and L. Wang, "Learning to upgrade internet information security and protection strategy in big data era," *Comput. Commun.*, vol. 160, pp. 150–157, 2020. doi: 10.1016/j.comcom.2020.05.043.

[3] C. Zhou, Y. Yu, S. Yang, and H. Xu, "'Intelligent immunity based security defense system for multi-access edge computing network," *China Commun.*, vol. 18, no. 1, pp. 100–107, 2021. doi: 10.23919/JCC.2021.01.009.

[4] Y. Kang, J. Zhong, R. Li, Y. Liang, and N. Zhang, "Classification method for network security data based on multi-featured extraction," *Int. J. Artif. Intell. Tools*, vol. 30, no. 1, pp. 2140006, 2021. doi: 10.1142/S0218213021400066.

[5] P. Kumar, A. Fatima, and N. K. Nishchal, "Arbitrary vector beam encoding using single modulation for information security applications," *IEEE Photonics Technol. Lett.*, vol. 33, no. 5, pp. 243–246, 2021. doi: 10.1109/LPT.2021.3052571.

[6] R. Karaszewski, P. Modrzyński, and J. Modrzyńska, "The use of blockchain technology in public sector entities management: An example of security and energy efficiency in cloud computing data processing," *Energ.*, vol. 14, no. 7, pp. 1873, 2021. doi: 10.3390/en14071873.

[7] Y. Lin *et al.*, "A deep learning framework of autonomous pilot agent for air traffic controller training," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 5, pp. 442–450, 2021. doi: 10.1109/THMS.2021.3102827.

[8] Y. Li, L. Tian, H. Qiu, and C. Zhang, "Research notes: Distributed shadow for router security defense," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 28, no. 2, pp. 193–206, 2018. doi: 10.1142/S021819401840003X.

[9] A. Meghea, "Graphene for defense and security," *MRS Bull.*, vol. 43, no. 7, pp. 556–556, 2018. doi: 10.1557/mrs.2018.171.

[10] D. K. Hsiao, "Federated databases and systems: Part I—A tutorial on their data sharing," *VLDB J*, vol. 1, no. 1, pp. 127–179, 1992. doi: 10.1007/BF01228709.

[11] A. Yao, G. Li, X. Li, F. Jiang, J. Xu and X. Liu, "Differential privacy in edge computing-based smart city applications: Security issues, solutions and future directions," *Array*, vol. 19, pp. 100293, 2023. doi: 10.1016/j.array.2023.100293.

[12] G. Sathish Kumar, K. Premalatha, G. Uma Maheshwari, and P. Rajesh Kanna, "No more privacy concern: A privacy-chain based homomorphic encryption scheme and statistical method for privacy preservation of user's private and sensitive data," *Expert Syst. Appl.*, vol. 234, pp. 121071, 2023. doi: 10.1016/j.eswa.2023.121071.

[13] K. Xu, B. H. M. Tan, L. P. Wang, K. M. M. Aung, and H. Wang, "Privacy-preserving outsourcing decision tree evaluation from homomorphic encryption," *J. Inf. Secur. Appl.*, vol. 77, no. 4, pp. 103582, 2023. doi: 10.1016/j.jisa.2023.103582.

[14] Y. Li, Y. Li, K. Zhang, and Y. Ding, "Public integrity auditing for dynamic group cooperation files with efficient user revocation," *Comp. Stand. Inter.*, vol. 83, pp. 103641, 2023. doi: 10.1016/j.csi.2022.103641.

[15] Z. Liu, S. Wang, and Y. Liu, "Blockchain-based integrity auditing for shared data in cloud storage with file prediction," *Comput. Netw.*, vol. 236, pp. 110040, 2023. doi: 10.1016/j.comnet.2023.110040.

[16] J. Tian, H. Wang, and M. Wang, "Data integrity auditing for secure cloud storage using user behavior prediction," *Comput. Secur.*, vol. 105, no. 2, pp. 102245, 2021. doi: 10.1016/j.cose.2021.102245.

[17] Y. Zhang *et al.*, "OWL: A data sharing scheme with controllable anonymity and integrity for group users," *Comput. Commun.*, vol. 209, no. 6, pp. 455–468, 2023. doi: 10.1016/j.comcom.2023.07.022.

[18] M. Waniek, T. P. Michalak, and A. Alshamsi, "Strategic attack & defense in security diffusion games," *ACM Trans. Intel. Syst. Tec.*, vol. 11, no. 1, pp. 1–35, 2019. doi: 10.1145/3357605.

[19] A. Ho, "Roles of three lines of defense for information security and governance," *ISACA J.*, vol. 1, no. 4, pp. 38–42, 2018.

[20] K. Zhang, Z. Jiang, J. Ning, and X. Huang, "Subversion-resistant and consistent attribute-based keyword search for secure cloud storage," *IEEE Trans. Inf. Foren Sec.*, vol. 17, pp. 1771–1784, 2022. doi: 10.1109/TIFS.2022.3172627.
[21] K. Lounis, "Stochastic-based semantics of attack-defense trees for security assessment," *Electron Notes. Theor. Comput. Sci.*, vol. 337, pp. 135–154, 2018. doi: 10.1016/j.entcs.2018.03.038.
[22] J. I. Guerrero, A. García, E. Personal, J. Lugue, and C. León, "Heterogeneous data source integration for smart grid ecosystems based on metadata mining," *Expert. Syst. Appl.*, vol. 79, no. 3, pp. 254–268, 2017. doi: 10.1016/j.eswa.2017.03.007.
[23] O. O. Malomo, D. B. Rawat, and M. Garuba, "Next-generation cybersecurity through a blockchain-enabled federated cloud framework," *J. Supercomput.*, vol. 74, no. 10, pp. 5099–5126, 2018. doi: 10.1007/s11227-018-2385-7.
[24] M. Chaloupka, and M. Nečaský, "Efficient SPARQL to SQL translation with user defined mapping," in *Knowledge Engineering and Semantic Web*, Cham: Springer International Publishing, 2016, pp. 215–229.
[25] M. Bouzeghoub, B. F. Lóscio, Z. Kedad, and A. Soukane, "Heterogeneous data source integration and evolution," in *Database and Expert Systems Applications*, Berlin, Heidelberg: Springer, 2002.
[26] A. Torres, R. Galante, M. S. Pimenta, and A. J. B. Martins, "Twenty years of object-relational mapping: A survey on patterns, solutions, and their implications on application design," *Inform Software Tech.*, vol. 82, pp. 1–18, 2017. doi: 10.1016/j.infsof.2016.09.009.
[27] N. W. Paton, K. Belhajjame, S. M. Embury, A. A. A. Fernandes, and R. Maskat, "Pay-as-you-go data integration: Experiences and recurring themes," in *SOFSEM 2016: Theory and Practice of Computer Science*, Berlin, Heidelberg: Springer, 2016.
[28] E. Uprichard and L. Dawney, "Data diffraction: Challenging data integration in mixed methods research," *J. Mix. Method. Res.*, vol. 13, no. 1, pp. 19–32, 2016. doi: 10.1177/1558689816674650.
[29] B. El Idrissi, S. Baïna, A. Mamouny, and M. Elmaallam, "RDF/OWL storage and management in relational database management systems: A comparative study," *J. King. Saud. Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7604–7620, 2022. doi: 10.1016/j.jksuci.2021.08.018.
[30] K. Kaur, V. Mangat, and K. Kumar, "A review on virtualized infrastructure managers with management and orchestration features in NFV architecture," *Comput. Netw.*, vol. 217, no. 4, pp. 109281, 2022. doi: 10.1016/j.comnet.2022.109281.
[31] J. Hui, L. Li, and Z. Zhang, "Integration of big data: A survey," in *Data Science*, Singapore: Springer, 2018.
[32] R. Wang *et al.*, "Review on mining data from multiple data sources," *Pattern Recogn. Lett.*, vol. 109, no. 9, pp. 120–128, 2018. doi: 10.1016/j.patrec.2018.01.013.
[33] M. Obali and B. Dursun, "A model for dynamic integration of data sources," in T. Özyer, Z. Erdem, J. Rokne, S. Khoury (Eds.), *Mining Social Networks and Security Informatics*, Dordrecht, Netherlands: Springer, 2013, pp. 1–14.
[34] B. Ojokoh, M. Zhang, and J. Tang, "A trigram hidden Markov model for metadata extraction from heterogeneous references," *Inf. Sci.*, vol. 181, no. 9, pp. 1538–1551, 2011. doi: 10.1016/j.ins.2011.01.014.
[35] J. Shen and M. Chi, "A novel multiview topic model to compute correlation of heterogeneous data," *Ann. Data. Sci.*, vol. 5, no. 1, pp. 9–19, 2018. doi: 10.1007/s40745-017-0135-y.
[36] N. Kushwaha and M. Pant, "Textual data dimensionality reduction—A deep learning approach," *Multimed. Tools Appl.*, vol. 79, no. 15, pp. 11039–11050, 2020. doi: 10.1007/s11042-018-6900-x.
[37] C. Chen, W. Xu, and L. Zhu, "Distributed estimation in heterogeneous reduced rank regression: With application to order determination in sufficient dimension reduction," *J. Multivariate Anal.*, vol. 190, no. 3, pp. 104991, 2022. doi: 10.1016/j.jmva.2022.104991.
[38] S. Wei, X. Zhou, X. An, X. Yang, and Y. Xiao, "A heterogeneous E-commerce user alignment model based on data enhancement and data representation," *Expert. Syst. Appl.*, vol. 228, no. 9, pp. 120258, 2023. doi: 10.1016/j.eswa.2023.120258.
[39] B. Wu, W. Zhao, H. Hu, Y. Liu, and J. Lv, "Conceptual design of intelligent manufacturing equipment based on a multi-source heterogeneous requirement mapping method," *IFAC-PapersOnLine*, vol. 55, no. 2, pp. 475–480, 2022. doi: 10.1016/j.ifacol.2022.04.239.

[40] D. Yuan, "Intelligent innovative knowledge management integration method based on user generated content," *Cluster. Comput.*, vol. 22, no. 2, pp. 4793–4803, 2019. doi: 10.1007/s10586-018-2389-3.

[41] P. P. Verbeek and A. Slob, "Analyzing the relations between technologies and user behavior," in P. P. Verbeek, A. Slob (Eds.), *User Behavior and technology Development: Shaping Sustainable Relations between Consumers and Technology*, Dordrecht, Netherlands: Springer, 2006, pp. 385–399.

[42] J. Zhong, W. Cai, L. Luo, and M. Zhao, "Learning behavior patterns from video for agent-based crowd modeling and simulation," *Auton. Agents Multi.-Agent. Syst.*, vol. 30, no. 5, pp. 990–1019, 2016. doi: 10.1007/s10458-016-9334-8.

[43] Z. Obermeyer and E. J. Emanuel, "Predicting the future-big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, 2016. doi: 10.1056/NEJMp1606181.

[44] J. Kang, M. Liu, and W. Qu, "Using gameplay data to examine learning behavior patterns in a serious game," *Comput. Hum. Behav.*, vol. 72, pp. 757–770, 2017. doi: 10.1016/j.chb.2016.09.062.

[45] S. Cobo-López, A. Godoy-Lorite, J. Duch, M. Sales-Pardo, and R. Guimerà, "Optimal prediction of decisions and model selection in social dilemmas using block models," *EPJ Data. Sci.*, vol. 7, no. 1, pp. 48, 2018. doi: 10.1140/epjds/s13688-018-0175-3.

[46] M. Chen, Z. Xu, J. Zhao, Y. Zhu, Z. Shao and X. Li, "Closed-loop robust steady-state target calculation for model predictive control," *Comput. Chem. Eng.*, vol. 168, no. 6, pp. 108045, 2022. doi: 10.1016/j.compchemeng.2022.108045.

[47] H. Jahanshahi and M. G. Baydogan, "nTreeClus: A tree-based sequence encoder for clustering categorical series," *Neurocomputing*, vol. 494, no. 1989, pp. 224–241, 2022. doi: 10.1016/j.neucom.2022.04.076.

[48] M. B. Dale, P. E. R. Dale, and P. Tan, "Supervised clustering using decision trees and decision graphs: An ecological comparison," *Ecol. Model.*, vol. 204, no. 1, pp. 70–78, 2007. doi: 10.1016/j.ecolmodel.2006.12.021.

[49] Z. Chai, A. Nwachukwu, Y. Zagayevskiy, S. Amini, and S. Madasu, "An integrated closed-loop solution to assisted history matching and field optimization with machine learning techniques," *J. Petrol. Sci. Eng.*, vol. 198, no. 6, pp. 108204, 2021. doi: 10.1016/j.petrol.2020.108204.