scientific reports

OPEN

Application of big data technology in enterprise information security management

Ping Li[1] & Limin Zhang[2]✉

This study aims to explore the application value of big data technology (BDT) in enterprise information security (EIS). Its goal is to develop a risk prediction model based on big data analysis to enhance the information security protection capability of enterprises. A big data analysis system that can monitor and intelligently identify potential security risks in real-time is constructed by designing complex network analysis algorithms and machine learning models. For different types of security threats, the system uses feature engineering and model training processes to extract key risk indicators and optimize model prediction performance. The experimental results show that the constructed risk prediction model has excellent performance on the test set, and its Area Under the Curve reaches 0.95, indicating that the model has good differentiation ability and high prediction accuracy. In addition, in the multi-class risk identification task, the model achieves an average precision of 0.87. Compared with the traditional method, it has remarkably improved the early warning accuracy and response speed of enterprises to various information security incidents. Therefore, this study confirms the effectiveness and feasibility of applying BDT to EIS risk management, and the successfully constructed prediction model provides strong technical support for EIS protection.

Keywords Big data technology, Enterprise information security, Risk prediction model, Information security

Research background and motivations

Today, with the deepening of globalization and informatization, all walks of life have been deeply embedded in the tide of digitalization[1–3]. To maintain competitiveness, enterprises have accelerated the pace of digital transformation, achieving seamless connection and intelligent upgrading of production, operation, sales, customer service, and other links[4,5]. In this process, a large number of business data, user information, and other key digital assets flow at high speed in cyberspace, forming an unprecedented data deluge[6–8].

At the same time, the accompanying problems of information security (IS) have become increasingly prominent[9]. The means of network attacks are changing with each passing day, seriously threatening the security of enterprise core information assets[10–12]. According to statistics, the amount of economic losses caused by IS incidents is staggering every year. Not only that, the incident can also trigger a series of chain reactions such as a crisis of trust, legal disputes, and brand reputation damage[13–15].

In the face of such a severe security situation, the traditional information security management system (ISMS) and technical means are unable to cope with security threats of massive data and dynamic changes[16–18]. This is the moment when big data technology (BDT) comes to the fore. With its ability to efficiently collect, store, process, and intelligently analyze large-scale datasets, BDT enables deep insight into security threat patterns, predicts risks ahead of time, and takes timely action[19–21].

Especially in the field of enterprise information security (EIS), BDT helps to improve security defense mechanisms. Moreover, BDT also realizes real-time monitoring, accurate positioning, and intelligent early warning of potential security threats through deep integration and analysis of various data sources such as network traffic, system logs, and user behavior. Thus, enterprises' traditional cognition and practice of IS prevention and control can be completely changed[22–24]. However, how to deeply integrate it with the existing ISMS of enterprises and construct a new information security management (ISM) architecture that not only adapts to the characteristics of the big data era but also takes into account the individual needs of enterprises is a task that combines theoretical challenges and practical innovations[25].

[1]School of Information and Mechatronic Engineering, Hunan International Economics University, Changsha 410205, China. [2]College of Electrical and Information Engineering, Hunan Institute of Traffic Engineering, HunanHengyang 421001, China. ✉email: zlmin2024@163.com

Research Objectives and Innovation Point

Building upon the current development frontier of BDT and the actual needs of EIS, this study is committed to deeply exploring and mining the innovative application path and implementation strategy of BDT in EIS management (EISM). The main objectives are divided into the following points. (1) Theoretical construction: It is necessary to systematically sort out the application scenarios of BDT in EISM, and try to construct a set of big data-driven ISM models suitable for enterprise environments. (2) Empirical analysis: Typical enterprise cases are selected, combined with actual business scenarios, and BDT is used to conduct empirical research on ISM to verify its effectiveness in risk identification and security situation awareness. (3) Technical solution design: The BDT-based EIS solution is proposed, including the selection of data sources, the design of data processing processes, and the establishment of security early warning systems.

In recent years, the importance of risk management in IS has become increasingly prominent, with several internationally recognized standards and frameworks providing references for enterprises' IS risk assessment and management. ISO/IEC 27,005 is a standard designed for IS risk management, covering the entire lifecycle of risk identification, assessment, treatment, and monitoring. However, its application process is highly dependent on manual decision-making and static assessments, which are limited in efficiency when dealing with massive and dynamically changing data. The NIST Cybersecurity Framework offers a security management framework based on the five stages of "Identify, Protect, Detect, Respond, Recover," emphasizing rapid response to risks through the detection and response modules. However, it still falls short in terms of large-scale real-time data processing and dynamic optimization capabilities. In response to the limitations of the aforementioned frameworks and research objectives, this study introduces BDT and ML models to design a dynamic analysis and real-time response system. Innovative improvements are achieved in the following areas:

1) Enhanced real-time monitoring and anomaly detection capabilities: The system adopts multi-source data fusion technology, integrating data sources such as system logs, network traffic, and user behavior. Moreover, it combines real-time stream processing technologies like Apache Spark to achieve dynamic analysis and anomaly detection of high-frequency data. Compared to the traditional framework's reliance on periodic analysis reports, the system markedly improves the identification efficiency of potential threats.

2) Closed-loop automated response mechanisms: With the aid of intelligent alarm engines and automated response modules, the system can quickly trigger isolation, traceback, and recovery operations after threat identification. Additionally, optimizing the manual response processes in traditional frameworks into a closed-loop automated process significantly reduces response time and labor costs.

3) Refined risk prediction capabilities: Through deep learning (DL) models and feature engineering, the system can accurately predict the occurrence probabilities of different types of security events. Meanwhile, it has combined quantitative indicators to verify its efficiency in multi-class risk identification tasks. This prediction capability further strengthens the system's early warning functions, giving enterprises more time for defense.

4) Broad applicability verified across industries: The effectiveness of the methods has been validated in actual cases in industries such as manufacturing, finance, and information technology, distilling universal security management strategies.

Literature review

In recent years, in terms of big data-driven IS, many studies have focused on the potential of big data analysis in detecting network abnormal behaviors and predicting security events[26–28]. For example, Alomari et al. (2023) used machine learning (ML) algorithms combined with big data platforms to conduct real-time analysis of network traffic, effectively improving the accuracy and response speed of identifying malicious activities[29]. In addition, Li et al. (2023) and Sánchez-Zas et al. (2023) improved enterprises' ability to understand and control the overall security environment by implementing a security situational awareness model based on big data[30,31].

However, although some achievements have been made in theoretical research and technical application, there are still some problems to be solved. On the one hand, when dealing with complex and mixed security threats, existing big data security solutions often face problems such as uneven data quality and large noise interference, which affect the accuracy of threat detection and prediction[32–34]. On the other hand, Dhirani et al. (2023) and Ke & Sudhir (2023) argued that with the increasing strictness of privacy protection regulations, ensuring user privacy and personal IS while utilizing big data to enhance IS was a dilemma[35,36]. For instance, Nguyen & Tran (2023) discussed the impact of jurisdictional differences in global cybersecurity rules, focusing on issues arising from international data flows and multinational governance[37]. Moreover, although many enterprises attempted to introduce BDT, in practical operation, the high difficulty of technology integration and the prominent problem of data silos were not fully realized the value of big data in the IS field[38].

For instance, Gonzalez-Granadillo et al. (2021) pointed out that security information and event management systems were widely deployed as powerful tools to prevent, detect and respond to cyber-attacks. These security system solutions evolved into comprehensive systems that provided broad visibility to identify high-risk areas[39]. Meanwhile, these systems focused on mitigation strategies designed to reduce the cost and time of incident response. Mirtsch et al. (2020) applied Web mining to explore the adoption of ISO/IEC 27,001. They estimated a probability model and found that larger, more innovative enterprises were more likely to obtain ISO/IEC 27,001 certification[40]. In addition, nearly half of the certified enterprises belonged to the information and communication technology services industry. In terms of cybersecurity incidents, Ahmad et al. (2020) drew on organizational learning theory to develop a conceptual framework explaining how to better integrate ISM and incident response functions. It was found that, in turn, a strong integration of ISM and incident response functions created learning opportunities, thereby providing organizations with a security advantage[41]. This advantage included increasing awareness of security risks, compiling threat intelligence, eliminating deficiencies in security defenses, assessing the logic of security defenses, and enhancing security responses.

Overall, the application of BDT in EISM has gradually gained recognition, with many studies indicating that big data can effectively improve the detection and prediction capabilities of security incidents. Especially, by combining ML algorithms with big data platforms, enterprises can conduct real-time analysis of massive data, promptly identifying anomalous behaviors and potential threats. Applications such as network traffic analysis and security situational awareness have achieved certain progress. However, despite technological advancements, real-world issues persist and require further resolution. Firstly, existing big data security solutions still face issues such as unstable data quality and significant noise interference when dealing with complex and variable security threats, directly affecting the precision of threat detection.

Secondly, privacy protection and data security have become a prominent contradiction in the application of big data. With increasingly strict data privacy regulations in various countries and internationalized data governance requirements, how to protect user privacy while effectively using BDT to enhance IS has become a significant challenge for enterprises. Particularly in multinational corporations and global operations, the legal conflicts between data flow and privacy protection are more complex, and designing a security architecture that meets global standards remains an unresolved issue. Furthermore, although BDT has been introduced into ISM systems, many enterprises have not fully exploited the potential of BDT in actual operations due to the complexity of technology integration and data silo problems.

In view of the above problems, this study aims to fill the following gaps. On the one hand, by combining big data and other ML algorithms, the existing security threat detection model is optimized to improve its identification ability and accuracy of complex and diverse security threats. On the other hand, based on the contradiction between privacy protection and data utilization, how to design a compliant data processing process is discussed. This can ensure that the dual goals of privacy protection and IS are achieved in big data applications. In addition, given the problem of technology integration and data island, the ISM system based on big data architecture is studied. Also, how to break the data island and improve the cross-system and cross-department data sharing and collaboration efficiency is explored, to provide more intelligent solutions for the security management of enterprises. Through these innovations and practices, this study hopes to provide enterprises with feasible improvement schemes in the big data security field and promote the technological innovation of ISM.

Research methodology

The BDT framework

The BDT framework is a collection of tools and techniques for processing large-scale data[42,43]. In this study, a comprehensive BDT framework based on the Hadoop ecosystem is carefully constructed, which consists of four interrelated and progressive core components, aiming to achieve efficient management and intelligent analysis of large-scale security-related data. The specific architecture is displayed in Fig. 1:

Figure 1presents the deployment of a distributed log collection system in the data collection phase. They act as a real-time data pipeline to continuously capture and transmit raw security event data from diverse internal information systems, network devices, server logs, and other security devices within the enterprise. Following this, to meet the storage requirements, the Hadoop Distributed File System (HDFS) is chosen as the underlying large-scale distributed storage infrastructure[44,45]. HDFS not only accommodates massive data but also ensures

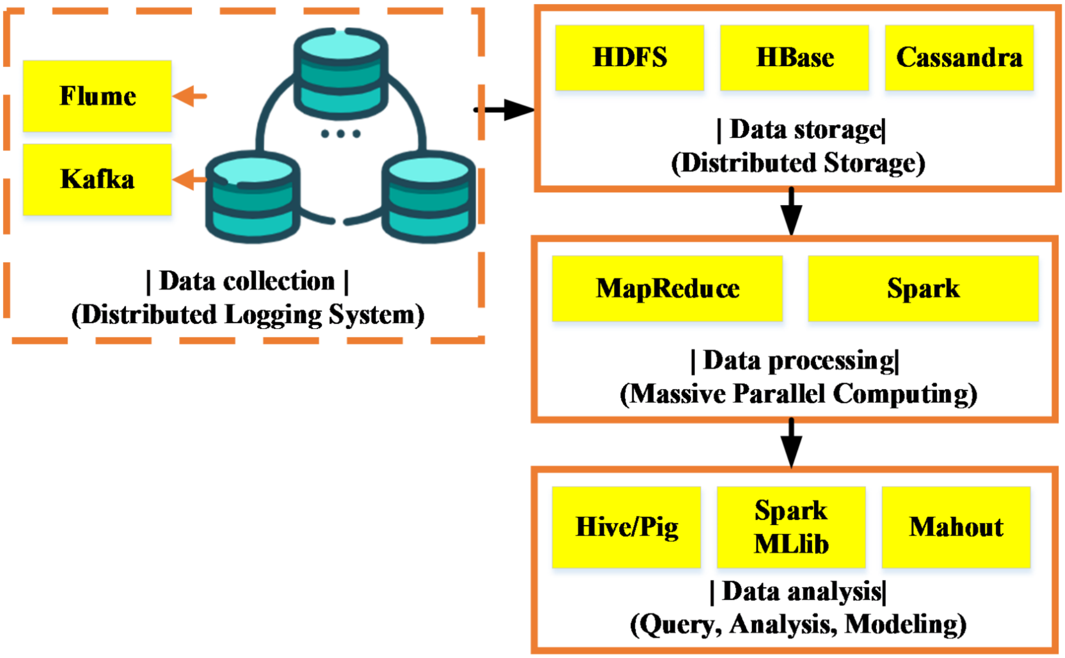

Fig. 1. The BDT architecture.

data's high availability and reliability through redundant replication. Additionally, for complex-structured non-structured and semi-structured data, NoSQL databases are further utilized. Leveraging their flexible data model and efficient column-family storage characteristics, they meet various data management needs in different scenarios.

In the data processing phase, two complementary technical paths are employed to address diverse business scenarios. On one hand, the classical MapReduce programming model is used for batch processing jobs, excelling in partitioning large datasets into independent subtasks and parallel execution of data cleaning, transformation, and preliminary analysis on cluster nodes[46]. On the other hand, considering scenarios with high demands for real-time and low latency, the Apache Spark framework is introduced to enable real-time response and dynamic analysis of security events.

Finally, in the data analysis phase, a range of advanced tools and technologies are fully utilized to uncover potential patterns of security threats. For instance, Apache Hive provides a SQL-like query interface for structured querying and statistical analysis of processed data, revealing patterns and trends hidden within vast amounts of data. Simultaneously, Apache Pig offers a high-level data flow language, simplifying the writing and execution of large-scale data processing scripts. Crucially, Spark's built-in MLlib library and mature tools like Mahout are employed for deep learning and ML modeling of pre-processed data. Thus, it develops precise threat detection and prediction models, effectively enhancing the recognition and prevention capabilities against network security threats. Through the BDT framework, enterprises can achieve efficient processing and analysis of massive data, enabling real-time monitoring, threat warning, and rapid response in ISM.

The architecture of the ISM model

To explore in depth how the BDT framework supports EISM, the following is a detailed analysis of the ISM model architecture. This model architecture employs BDT to improve the IS level of enterprises, mainly through real-time monitoring, abnormal behavior detection, and risk prediction. The specific model architecture is presented in Fig. 2. In the ISM model architecture, risk identification and feature extraction are the core links, which run through the data and processing layers. They can provide strong support for subsequent security management through the integration and intelligent analysis of multi-source data.

In Fig. 2, a closed-loop is formed between the layers through data flow and feedback, where the upper layers depend on the data and analysis results from the lower layers to drive the entire security management process.

This study utilizes data sources such as "system logs," "network traffic records," "user behavior," and "external threat intelligence" to support threat detection. (1) System logs leverage log information generated by operating systems, applications, databases, etc., to monitor and analyze system activities in real-time, identifying potential anomalous behaviors and security vulnerabilities. System logs can provide early signs of attack behaviors, such as illegal login attempts, unauthorized access, and more. (2) Network traffic records analyze the incoming and outgoing data packets through real-time monitoring of network traffic. Special attention is paid to abnormal traffic and data transmission patterns, identifying potential DDoS attacks, network scanning, and other malicious activities. Traffic records assist in detecting security threats at the network layer through feature analysis and behavior recognition. (3) User behavior data, such as user login activities, access permission changes, and file operation records are used to analyze account anomalies. These data help identify whether users exhibit abuse of permissions, abnormal logins, and other behaviors, and through behavioral analysis models, determine if there are potential internal threats. (4) External threat intelligence data provides information on the latest security vulnerabilities, malicious attack activities, and their patterns, aiding in cross-domain threat detection. By integrating this intelligence with internal system data for comprehensive threat analysis, the accuracy and timeliness of detection are enhanced. These multidimensional data are converged through a distributed

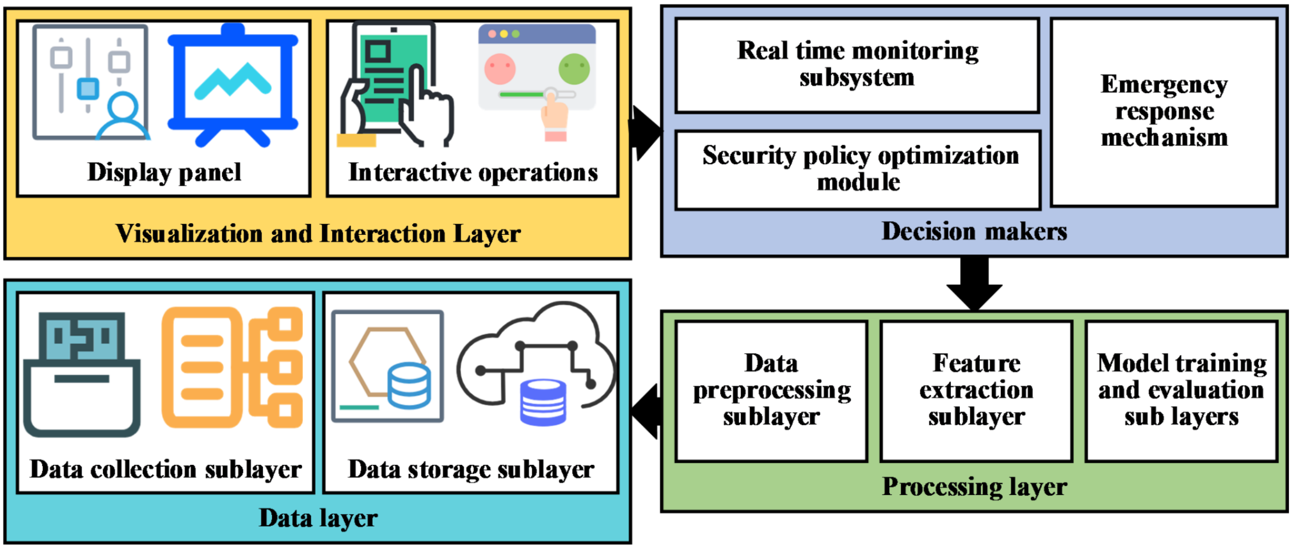
Display panel
Interactive operations
Visualization and Interaction Layer
Real time monitoring subsystem
Security policy optimization module
Emergency response mechanism
Decision makers
Data collection sublayer
Data storage sublayer
Data layer
Data preprocessing sublayer
Feature extraction sublayer
Model training and evaluation sub layers
Processing layer

Fig. 2. The ISM model architecture.

collection framework, achieving comprehensive awareness of the system's operational status and ensuring the timeliness and integrity of the data.

In the data preprocessing sub-layer of the processing layer, the initial step involves cleansing and optimizing raw data to ensure its suitability for subsequent analysis requirements. Data preprocessing includes multiple stages, beginning with filtering the data to remove redundant information unrelated to security events, thereby reducing the interference of noise on the model's analytical results. Subsequently, denoising techniques are applied to eliminate unnecessary fluctuations in the data, ensuring stable data quality and enhancing the accuracy of subsequent analysis. Additionally, the preprocessing process involves data format conversion and standardization to better integrate data from different sources. Standardized data can be unified into specific structures or formats, making it more suitable for use in various models and algorithms, ensuring the compatibility of multi-source data.

After preprocessing, the data enters the feature extraction sub-layer. At this stage, the system extracts features from the processed data that are meaningful for the detection of security threats. These features include, but are not limited to, anomalies in user login patterns, changes in access frequency, and anomalies in resource access patterns. These features can reveal potential threats in system operations. For instance, unusual login times or locations may indicate misuse of user accounts, and frequent access requests may be signs of system scanning or penetration testing. Feature extraction helps enhance the model's precision in identifying potential threats by recognizing high-value security information from large volumes of raw data.

Once feature extraction is complete, the process moves to the model training and evaluation sub-layer. In this sub-layer, ML and data mining techniques are utilized to construct a security threat detection model. Two ML algorithms. Support Vector Machine (SVM) and Random Forest (RF), are employed. SVM is a supervised learning algorithm particularly adept at classifying high-dimensional data. In this study, SVM is used as the primary classifier. The fundamental concept is to find an optimal hyperplane (or, in non-linear cases, map data into high-dimensional space using kernel functions) that can effectively distinguish between normal and abnormal behaviors (i.e., potential security threats). The advantage of SVM lies in its ability to handle high-dimensional data and its strong generalization abilities, making it especially suitable for situations with high feature dimensions. During the training phase, SVM utilizes historical data (such as system logs, network traffic records, etc.) to learn the characteristics of security events and implement a classification model. The trained SVM model can promptly identify potential security threats with new data inputs.

RF is an ensemble learning method that builds multiple decision trees and determines the final classification result through a voting mechanism. RF is used in this study to improve the model's robustness and to select the most discriminative feature among multiple security event features. RF algorithms can effectively handle large datasets, especially in the face of noisy or redundant characteristics, and can reduce overfitting by integrating the results of multiple trees. Through continuous random sampling and feature selection, RF can extract information useful for threat identification from high-dimensional complex security event data.

In the process of model training and evaluation, firstly, historical data (such as system logs, user behavior data, network traffic records, etc.) are used to train the two models. The data preprocessing and feature extraction sub-layers ensure the quality and validity of the input data, after which the threat detection model is trained using SVM and RF algorithms. The training process continuously optimizes the model's hyperparameters (e.g. kernel type of SVM, number of trees of RF, etc.) to ensure that the model can adapt to different security threat scenarios. In the model evaluation phase, many performance indicators (such as accuracy, recall, F1 score, Area Under the Curve (AUC), etc.) are employed to evaluate the model's detection effectiveness. For example, by comparing the performance of SVM and RF models on different datasets, these two algorithms' advantages and disadvantages can be analyzed in specific scenarios, thus adjusting and optimizing the model. In the process of continuously optimizing the algorithm and adjusting the parameters, the model gradually improves the detection ability of security threats. Furthermore, the model can make timely and accurate responses in the face of new and complex security events.

By integrating both SVM and RF algorithms, this study not only achieves stable performance across various security threat detection scenarios but also effectively addresses the evolving security threats. The results of model training and evaluation are passed to the decision layer, where they enter the real-time monitoring subsystem for immediate analysis. Within this subsystem, the system conducts statistical analysis of the output results from the processing layer to monitor anomalies in networks and systems in real-time. The intelligent alert engine can detect and trigger security threat alarms promptly based on predefined thresholds. Once abnormal behavior or attack activities are detected, the system immediately sends alerts to security administrators to take corresponding defensive measures. During this process, the system can also dynamically adjust security policies. For example, based on historical analysis results and real-time security intelligence, the system can automatically adjust firewall rules, access control lists, etc., to counter current or potential threats. In addition to these preset security measures, the system enhances overall defense capabilities through optimized security policies.

Furthermore, the decision layer includes an emergency response mechanism to handle security incidents that occur. Once a security threat is confirmed, the system can automatically initiate response procedures based on the type and severity of the security event. These procedures include isolating affected systems, implementing repair measures, and even tracking attack paths to determine the source of the attack and its propagation methods. This mechanism can be activated automatically or manually, ensuring that in the event of a security incident, it can be handled swiftly and effectively, minimizing potential losses and impacts. Through this dynamic response mechanism, enterprises can maintain system security and stability in the face of changing security threats, ensuring timely responses to various complex security challenges. Through the organic combination and seamless connection of these links, the system can realize real-time response, efficient defense, and monitoring of IS threats. Thus, it can enhance overall security management capabilities and efficiency in dealing with complex threats.

Ultimately, the visualization and interaction layers are responsible for presenting the final results. The visualization display subsystem uses forms such as data dashboards, heatmaps, and network topologies to transform complex analysis results into a visual interface that is easy to understand and operate. The interaction operation subsystem allows security administrators to customize queries, filter, and drill down for in-depth analysis through the interface, and formulate, execute, and adjust security protection strategies based on visual results. Through this model architecture, enterprises can more effectively utilize BDT to enhance the efficiency and effectiveness of ISM, enabling rapid identification, accurate assessment, and effective response to IS threats.

Threat prediction and management optimization of Big Data-driven EIS

This study adopts a diversified strategy in enterprise management optimization methods, cleverly integrating both quantitative and qualitative analysis research paradigms to comprehensively explore and analyze the practical effectiveness and challenges of BDT in EISM. The quantitative analysis applies statistical principles and ML algorithms to deeply mine large-scale security-related data. By constructing a security threat prediction model, specifically a logistic regression model[47–49], this study predicts the probability of security events occurring. The equation for the prediction model is as follows:

$$y = f(x; \theta) \tag{1}$$

y represents the probability of the predicted occurrence of a security event. The feature vector *x* includes a series of attributes related to security threats, such as user login behavior, network traffic features, system state parameters, etc. θ is the learning parameter of the model, obtained through the training process to achieve the optimal solution, reflecting the mapping relationship between features and predicted results. The function *f* is the ML model's specific expression. The ML model's training process is illustrated in Fig. 3.

Qualitative analysis focuses on gaining a deep understanding of the causes, mechanisms, and contexts behind phenomena. In this study, industry experts and frontline security management personnel are invited for in-depth interviews to understand how they apply BDT in their practical work, the challenges they face, and their evaluations of existing solutions along with improvement suggestions. In the selection of experts, this study pays special attention to the industry background and practical experience of the experts. Experts are selected based on the following criteria. First, senior security executives with extensive experience in the ISM field are selected, including security architects, system security engineers, and heads of IS departments with more than a decade of industry experience; Second, academic experts who have made outstanding contributions to the research and development of safety technology are invited. These experts have in-depth research in the application of BDT technology and the construction of safety management system. To ensure that the selected experts can provide representative feedback, all experts come from industries with highly complex security requirements, including finance, manufacturing, and information technology. This diverse selection of experts ensures a comprehensive understanding of the application needs and challenges of BDT in different industries.

In the process of understanding the insights of experts, a semi-structured interview questionnaire is designed, and the interview content revolves around the following aspects. Firstly, experts evaluate the application of BDT technology in practical ISM, and explore how they combine BDT for security threat detection, risk assessment, and decision support in their work; Secondly, experts share specific challenges they have encountered in applying BDT, such as the complexity of data processing, the difficulty of real-time analysis, and compatibility issues with existing security management systems. Finally, the experts propose suggestions to improve the existing BDT technology solutions, encompassing how to improve the efficiency of data processing, enhance the ability of model prediction, and customize the solutions in different industries. To ensure the depth and breadth of the interview, the interview time of each expert is limited to about one hour. This can ensure that the experts are free to express their views, and provide specific cases and personal experience in real work.

Additionally, in terms of case analysis, representative enterprises from various industry sectors are selected for detailed study. The selected cases span multiple domains including manufacturing, financial services, and information technology, where these enterprises have high demand and practical experience in ISM. By analyzing these enterprise cases, the study examines how they apply BDT technology in actual operations. Concurrently, it focuses on the difficulties encountered, resolution strategies, and outcomes during the implementation process, thereby distilling general security management strategies applicable to a wide range of industries. The insights from these experts and the analysis of enterprise cases provide not only practical insights for this study but also help further optimize the BDT model. Thus, it ensures it better meets the security management needs of different fields. Firstly, the experts' suggestions help identify practical obstacles in applying BDT in the security management process, providing valuable references for model optimization and subsequent analysis; Secondly, the improvements suggested by the experts play a direct role in adjusting and expanding the model's functions, ensuring the research outcomes are closely aligned with industry demands.

Experimental design and performance evaluation

Datasets Collection

The dataset for this study is sourced from five enterprises, covering manufacturing, financial services, information technology, and other critical infrastructure sectors. Each enterprise signs a confidentiality agreement and performs de-identification processing on the raw data to protect the privacy of the enterprises and individuals. The selection criteria for these enterprises include the following. The selected enterprises must span important industries such as manufacturing, financial services, information technology, and critical infrastructure to ensure the model's broad applicability in EIS. Enterprises are required to provide ample historical data, especially log data regarding security events, such as transaction records, network logs, employee communications, etc., which are crucial for model training and testing. The selected enterprises are willing to share anonymized data

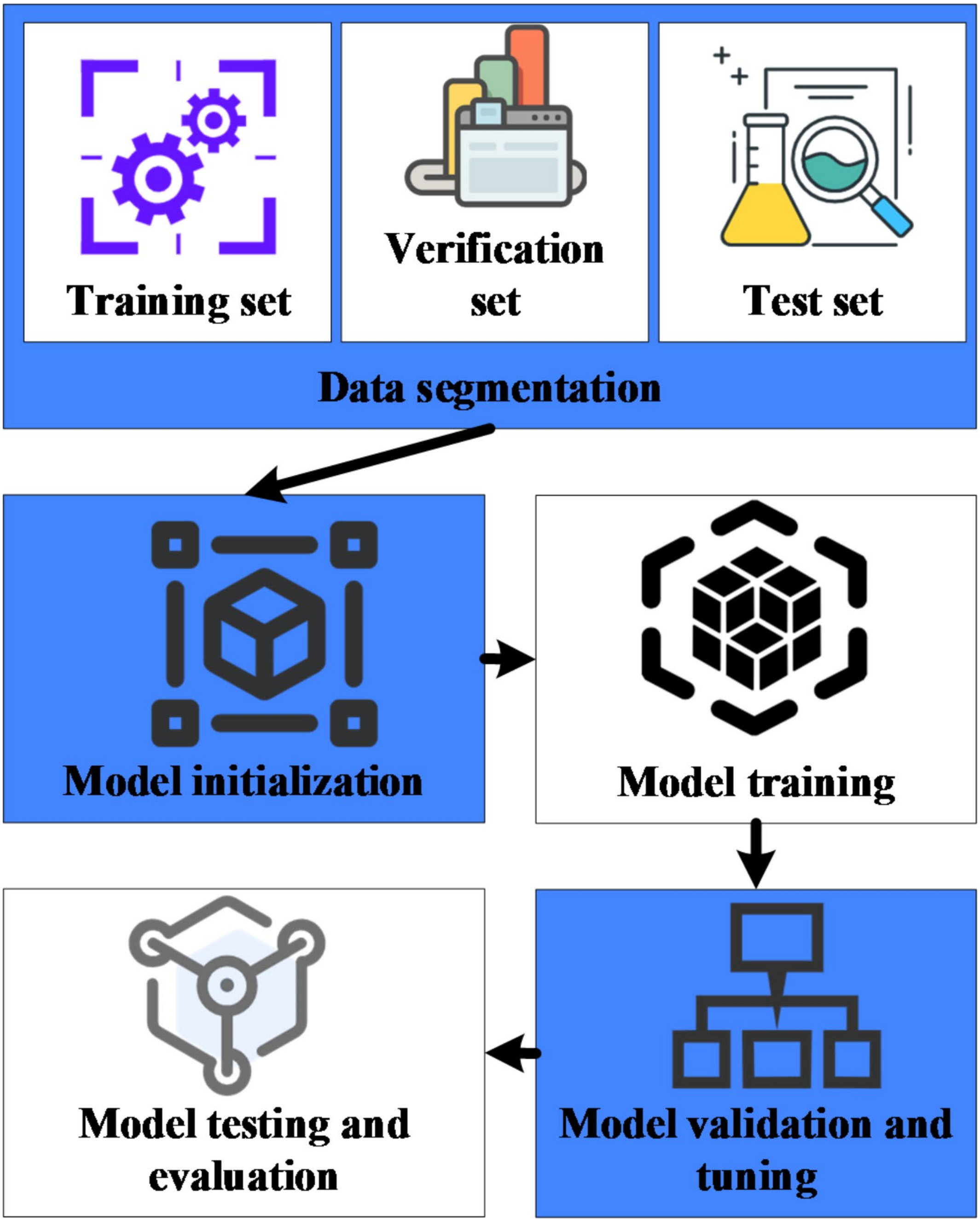

Fig. 3. Training process of the ML model.

under strict confidentiality agreements, ensuring the privacy and security of the data. The chosen enterprises should have different security needs and challenges to comprehensively assess the model's performance in various operational environments. These criteria ensure that the dataset covers security events from different domains. These allow for a comprehensive evaluation of the proposed big data model's performance across multiple industry contexts, thereby ensuring the broad applicability of the research findings. Information on the five selected enterprises is presented in Table 1.

In terms of data types, this study constructs a rich and multidimensional dataset; It integrates structured data (such as transaction records from Enterprise Resource Planning (ERP) systems), semi-structured data (such as logs from web servers and firewalls), and unstructured data (such as employee email communications, social media texts, and web camera images). The total volume of the dataset is approximately 100GB, containing over 500,000 transaction records, millions of network logs, and thousands of unstructured text data. The amount of data provided by each enterprise varies, but the overall dataset includes approximately 5,000 to 10,000 security-related event records from each enterprise. These events cover various security events such as abnormal logins, permission changes, data leaks, and attack detections.

Corporate name	Industry field	Number of employees	Data type	Remark
Manufacturing Enterprise A	Manufacturing	8,000	Transaction records, production data, network logs	Transnational enterprise, involving multiple national markets
Financial Services Enterprise B	Financial service	2,000	Transaction records, customer behavior data, logs	Regional financial service providers
Information Technology Enterprise C	information technology	1,500	Network security logs, employee communication data	Focusing on cloud computing services
Energy Enterprise D	Critical infrastructure	4,000	Operation records, equipment monitoring data	Leading enterprises in the domestic energy sector
Medical Startup Enterprise E	medical technology	150	Medical data, network communication data	Focusing on the field of intelligent medical technology

Table 1. Detailed information on research subjects.

Technical configuration items	Specific parameters or configurations
Platform hardware	Intel Xeon processors and NVIDIA Tesla GPU acceleration card
Storage system	A distributed storage system, with a total capacity of PB level
Operating system	CentOS 7.9, 64-bit
Big data framework	Apache Hadoop v3.2.1, Spark v3.1.2
Database system	MySQL 8.0 is applied for relational data storage, and MongoDB 4.4 is used for non-relational data
Actual deployment situation	All components are implemented for cluster deployment on the cloud platform

Table 2. The technical configuration of the experimental environment.

Parameter name	Optional range or set value
Learning rate	0.001, 0.005, 0.01, 0.05
The maximum depth of the tree	5, 10, 15, 20
Regularization intensity	0.001, 0.01, 0.1, 1.0
Number of feature selections	Top 10, Top 20, Top 30

Table 3. Parameter settings.

Regarding the feature dimensions, the dataset includes about 50 to 100 features, involving user behavior, access patterns, system errors, transaction records, communication content, and more. Each event sample typically contains information from multiple dimensions, enabling the model to identify potential security risks in different contexts. Additionally, the data collection spans 12 to 18 months, covering different seasons and business cycles, ensuring that the model can recognize both long-term and short-term security trends. Through this diverse and rich data, the study aims to evaluate the application effects of big data models in security management across different types of enterprises, ensuring their good performance across various data scales. The multidimensionality and complexity of the dataset are crucial for enhancing the model's robustness, improving the ability to identify security threats, and addressing complex security challenges.

Experimental environment

The technical configuration of the experimental environment is outlined in Table 2:

Parameters setting

Parameter settings are exhibited in Table 3:

Performance evaluation

This study designs a set of evaluation indicators, including accuracy, recall, F1 score, AUC, and Average Precision (AP), to measure the efficiency of the IS protection model. Accuracy assesses the proportion of all predictions the model correctly predicts; Recall reflects the model's ability to identify positive (threatening events); F1 score is the harmonic average of precision and recall, comprehensively measuring the model's performance; AUC evaluates the model's overall capability by calculating the classification performance of the model under different thresholds; AP provides a more stable evaluation indicator in the unbalanced dataset. Figure 4depicts different ISM models' performance comparisons. The comparison models include rule-based security detection (such as signature-based detection techniques)[50], AI-driven (SVM)[51], collaborative defense[52], Dl-based (CNN)[53], and hybrid security models[54].

The comparison results indicate that the proposed big data-based model performs best on several evaluation indicators, especially in the accuracy (90.3%) and AUC (0.92) advantages, showing its superior ability in dealing with complex security threats. This advantage is due to the ability of big data models to delve into hidden features

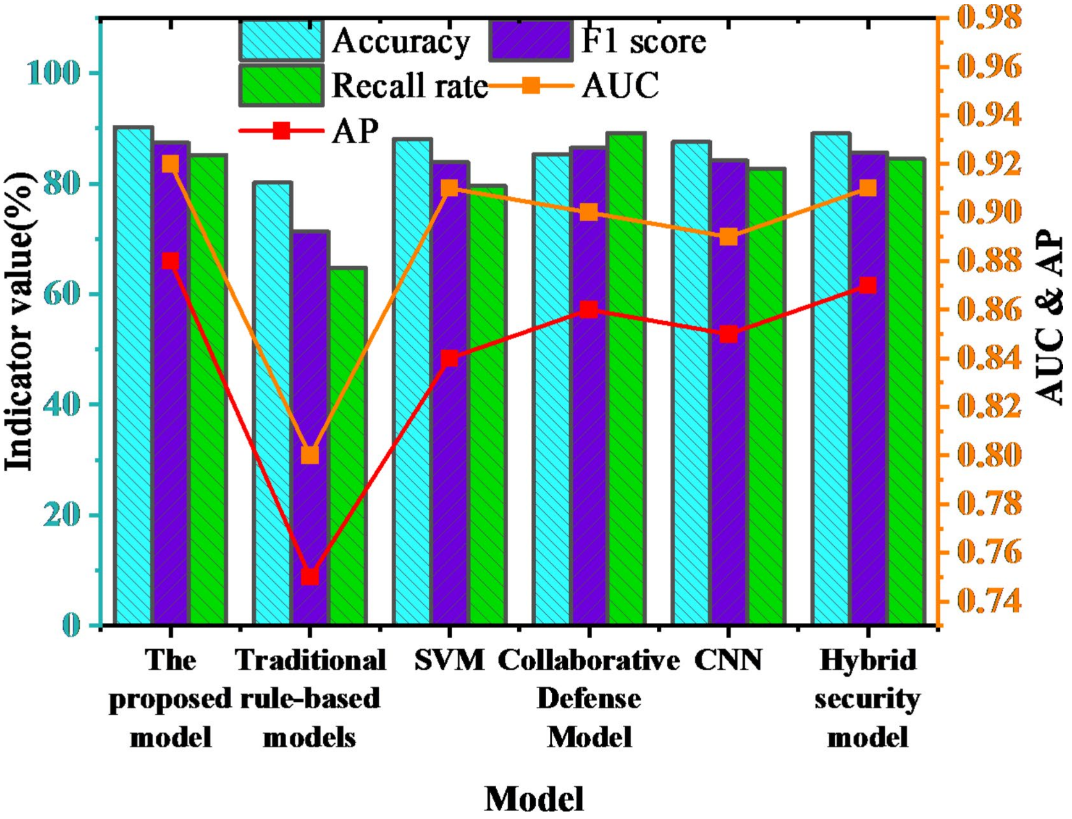

Fig. 4. Performance comparison of diverse ISM models.

in data through DL algorithms, overcoming the limitations of traditional models and single AI-driven models. It can be found that using BDT can provide remarkable performance gains when handling large amounts of data. The reasons are as follows. Firstly, the distributed storage and computing capabilities provided by Hadoop and Spark greatly improve the efficiency of data processing; Secondly, ML algorithms based on big data platforms can efficiently process multi-dimensional and large-scale data, which traditional technology stacks often struggle to handle. Finally, HiveQL and Pig, as big data analysis tools, can simplify the data query and processing process, thus further improving the efficiency and effect of model training.

In contrast, the traditional rule-based model lacks flexibility and expansibility, its accuracy and recall are low, and it is difficult to deal with complex network attacks. Although the SVM model has strong classification ability, especially in feature space, it has limited processing ability for high-dimensional data and is susceptible to noise. As a result, its application effect in a complex security environment is not good. The Convolutional Neural Network (CNN), as a representative model of DL, has advantages in extracting spatial features. CNN can effectively identify potential threats from logs, traffic, and other data. The model's performance is relatively stable, with an accuracy of 87.6%, an AUC of 0.89, and an F1 score of 84.2%. However, the processing of time series data is relatively weak, making it difficult to fully meet the needs of real-time security detection.

Collaborative defense models enhance defense capabilities by integrating different security mechanisms, with the advantage of being able to handle multiple types of threats simultaneously. However, its accuracy (85.3%) and F1 score (86.5%) are relatively low, reflecting that the interplay between multiple defense layers may introduce unnecessary complexity, thus affecting precision. Although its recall is high (89.2%), it still faces the issue of false positives. The hybrid security model achieves a better balance by fusing different technologies, but there is still room for improvement in terms of real-time performance and scalability. Overall, the proposed big data-based DL model, through efficient feature extraction and intelligent analysis, has remarkable advantages in terms of accuracy and real-time response capabilities, being better equipped to handle dynamic and complex security threats.

The key differences, advantages, and limitations of different security models are further compared, as detailed in Table 4:

The performance of big data models under different configurations is then assessed based on a series of empirical tests and theoretical foundations. The criteria for parameter grouping primarily consider key factors affecting model efficiency, including dataset size, the time window length for event detection, and the type of ML algorithms used. Parameters are grouped according to dataset size to evaluate the model's performance when dealing with data of different scales. This factor is crucial for understanding how the model copes with large-scale data in practical applications. The length of the time window is an important factor in time series analysis when predicting security events. Different time window lengths are selected to observe the model's sensitivity to temporal changes and how it adapts to different data input frequencies. Parameters related to ML algorithms

Characteristic	Big data model	Traditional rule-based Model	SVM	Collaborative defense model	CNN	Hybrid security model
Model type	Based on big data (ML)	Based on rule	Supervised learning	External threat sharing	DL	Integrating multiple models
Data processing	Multiple types of data (structured and unstructured)	Main structured data	Structured data	External intelligence and logs	Image, video, and other data	Multiple data types
Threat detection methods	Big Data Analysis and ML	Rule-driven detection	Based on ML classification	Combining external with internal testing	Pattern recognition	Integration of multiple methods
Scalability	High, supporting distributed	Restricted by rules	Secondary	Medium, dependent on collaboration	High	High
Adaptability	High, supports real-time learning	Poor adaptability	Secondary	secondary	High	High
Real-time processing capability	Supporting real-time analysis and alerts	More restrictions	More restrictions	High latency	Strong	Strong
Privacy protection	Data sensitization	Many privacy issues	Limited privacy protection	Shared data risk	Support privacy protection	Privacy protection needs attention

Table 4. Key differences between different security models.

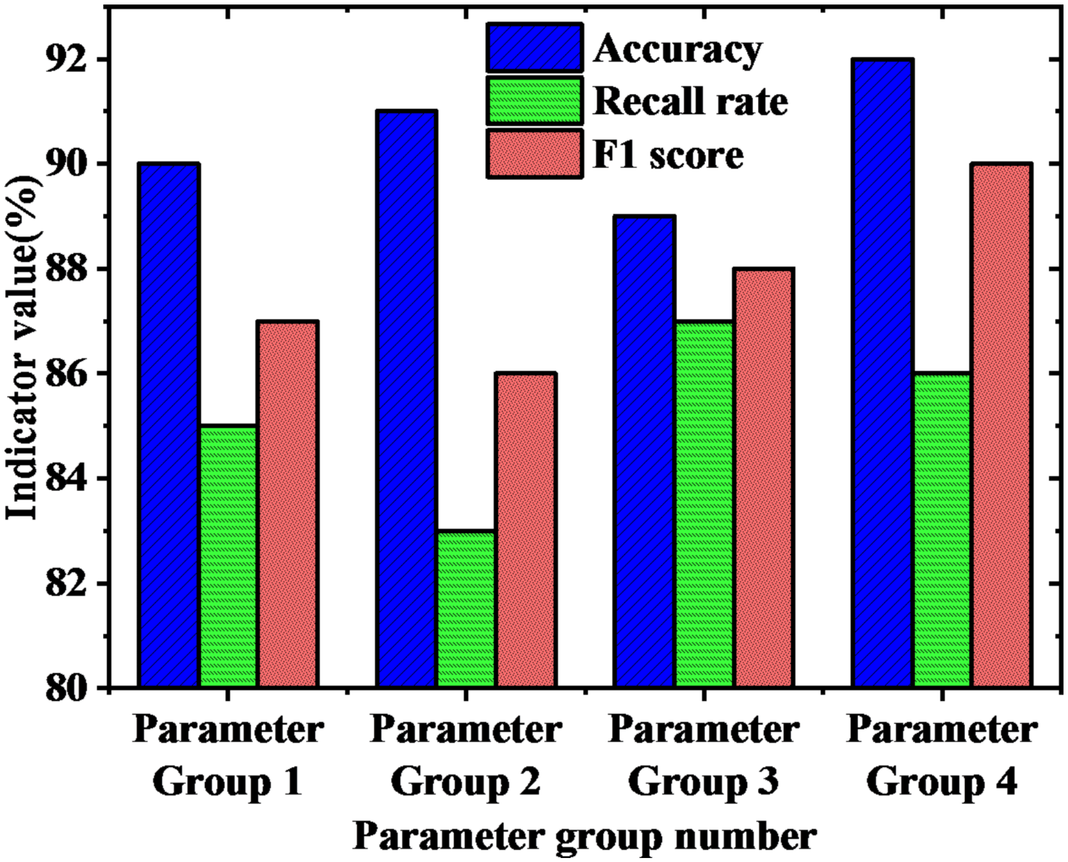

Fig. 5. Performance of the big data model with various parameter groups.

(such as learning rates or the number of layers and other hyperparameters) are also included in the grouping to assess the impact of diverse configurations on the model's prediction accuracy and efficiency. Selecting only four groups of parameters is to maintain appropriate simplicity in the analysis while ensuring coverage of the most critical variables affecting model performance. Such grouping can fully reflect the model's performance under various configurations without adding excessive complexity. The big data model's performance under different parameter groups is plotted in Fig. 5:

In Fig. 5, the performance differences of the big data model under diverse parameter combinations are observed. For instance, parameter group 1 gets the highest F1 score of 0.87 when the learning rate is 0.01, the depth of the tree is 10, the regularization intensity is 0.1, and the feature selection is Top 20. It indicates that this parameter group may be one of the best configurations for the big data model. However, although the accuracy of parameter group 4 is 92%, the recall is slightly lower, at 86%. It suggests that while pursuing high accuracy, it should also pay attention to the balanced performance of the model in terms of recall. The impact of the time window on the big data model's performance is drawn in Fig. 6:

Figure 6 denotes that model performance fluctuates with the change in time window length. For example, with a 14-day time window, the model reaches 90% accuracy and an 85% recall, illustrating that the 14-day time window may be the most beneficial time scale for the model to catch and predict security threats. The contribution of different data sources to the big data model's performance is suggested in Fig. 7:

Figure 7 illustrates the contribution of different data sources to the accuracy of big data models. Among them, system logs play the most significant role in enhancing the accuracy of big data models, with an improvement of 5%. System logs provide the model with a wealth of operational historical data, including records of system startups, user logins, file accesses, and more. By thoroughly analyzing this log information, anomalies within the system can be identified, such as frequent failed login attempts or unauthorized privilege changes. These insights offer crucial clues for the early detection of security threats, thereby effectively enhancing the model's predictive abilities. Network traffic records contribute the most to the model's enhancement, with an accuracy increase of 8%. These data sources offer detailed information about the flow of packets within the network, encompassing the size, type, source, and destination of incoming and outgoing packets. Through in-depth analysis of network traffic, the model can detect network-level security threats such as DDoS attacks, malicious scanning, or data breaches. The contribution of this data source indicates that monitoring and analysis of network traffic are indispensable in threat detection and are key factors in improving overall security protection effectiveness.

External threat intelligence improves the model's accuracy by 4%. These external data sources include real-time information about emerging attacks, vulnerabilities, and malware signatures, typically published by professional threat intelligence providers. This intelligence aids in identifying the strategies, techniques, and tools of external attackers and cross-referencing this information with behaviors within the internal network, thereby enhancing the model's cross-domain threat detection capabilities. The contribution of these data sources reflects the importance of multi-source data in threat detection. Different types of data sources offer a multi-dimensional perspective, ranging from system logs to external threat intelligence. By conducting an in-depth analysis and integration of the characteristics of these data sources, the model's detection capabilities and prediction accuracy can be comprehensively enhanced. The varying contributions of different data sources in security threat prediction indicate that in practical applications, data fusion and analysis strategies must be formulated based on specific business scenarios and data characteristics to achieve more precise security protection.

Discussion

The above has delved into the application effectiveness of big data models in the IS domain and the extent to which they are influenced by several key variables, as evidenced by a series of empirical data analyses. The big data-driven security management model, compared to traditional rule-based models and other AI models, demonstrates superiority in core indicators such as recall, accuracy, and F1 score. Although traditional rule-based security models have been widely used in past applications, they have obvious limitations. Rule-based models rely on manually defined security rules and known threat characteristics, which makes them slow and less accurate in the face of unknown, complex security threats. For example, traditional signature-based intrusion detection systems often cannot cope with new types of attacks, especially zero-day attacks and advanced persistent threats. In addition, rule-based models also have problems in real-time and scalability, particularly in large-scale data environments, where they often struggle to handle large amounts of complex data. Therefore, although these

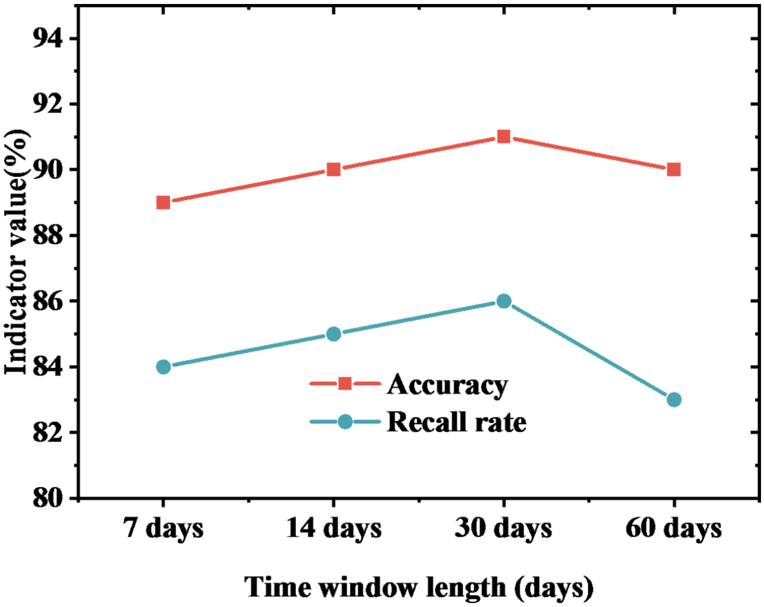

Fig. 6. The effect of the time window on the performance of the big data model.

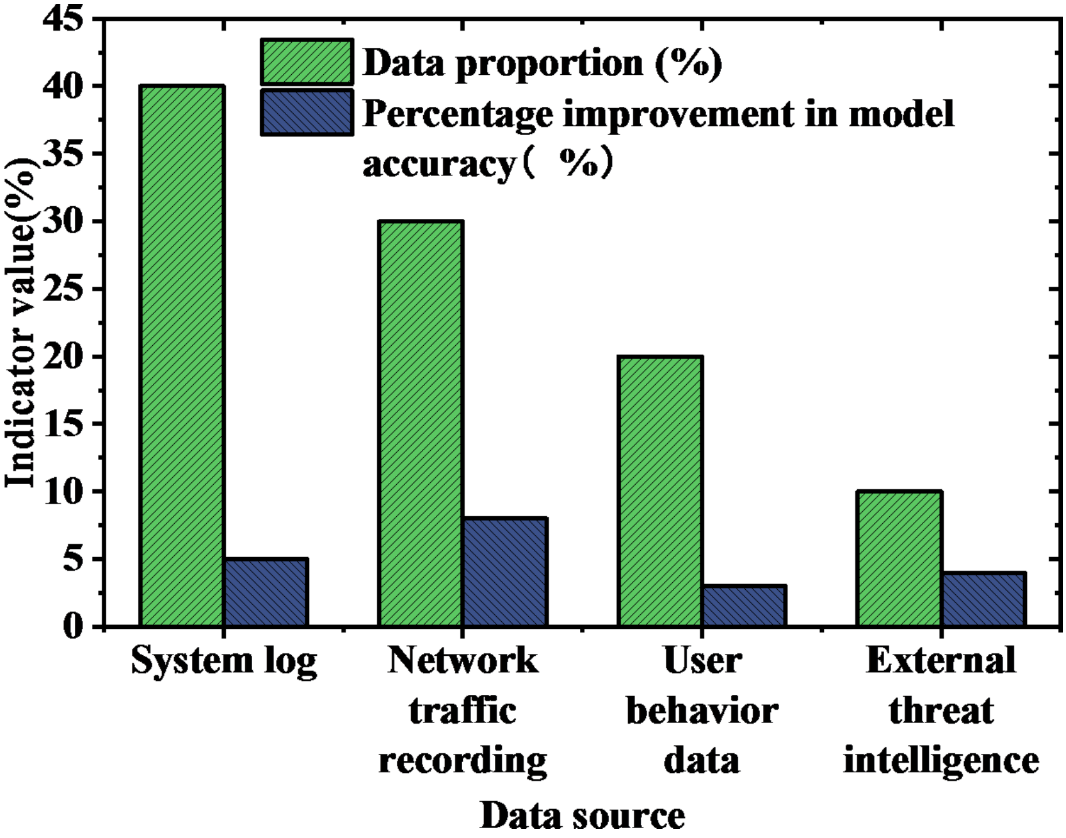

Fig. 7. Analysis of the contribution of diverse data sources to the performance of the big data model.

methods have advantages in detecting historical data and known threats, their effectiveness and adaptability are limited in a dynamically changing network environment.

In contrast, the big data-driven security management model demonstrates significant advantages, especially when dealing with massive, complex, and diverse security data. By utilizing the distributed storage and computing power of BDT, the model can process more security data in a shorter time, improving the accuracy and efficiency of detection. ML algorithms such as SVM, RF, and DL can self-learn from large amounts of historical data and identify potential security threats. This gives big data models unique advantages in detecting unknown threats and identifying new attack patterns. This aligns with the viewpoint emphasized in previous research literature by Himeur et al. (2023), stating that the data processing capabilities of big data contribute to enhancing threat detection efficiency[55]. It is noteworthy that the model performance reaches its optimum under specific time window lengths, a conclusion echoing the critical role of time series analysis in security event prediction as indicated in existing research by Kaffash et al. (2021)[56]. System logs have been confirmed as the most influential data source, aligning with recent studies by George (2023) and Vadhil et al. (2021), emphasizing the irreplaceable value of internal system logs in identifying potential security threats[57,58]. Simultaneously, this also serves as a reminder for researchers and practitioners to comprehensively consider the integration of various data sources to maximize the effectiveness and accuracy of threat detection.

However, big data models also face several challenges in the application process. First, the data quality and noise interference remain significant issues, especially when dealing with security data from diverse sources and non-uniform formats; Ensuring the accuracy and consistency of data is key to improving model performance. Second, privacy protection issues are becoming increasingly severe. With the strictness of global privacy protection regulations, how to protect user privacy and avoid data breaches while utilizing big data for threat detection has become an urgent problem. Additionally, technical integration and data silo issues limit the comprehensive application of big data models in practical operations. Although many enterprises have begun to introduce BDT, data silo problems still prevail in the technical implementation process, preventing the full potential of big data from being realized.

From a management and practical perspective, big data-driven security management models have important application value and practical significance. By effectively integrating and analyzing various data sources, enterprises can achieve early warning and real-time response to security threats, greatly improving the efficiency and effectiveness of ISM. Moreover, big data-based threat detection models can dynamically adjust and optimize to adapt to the constantly changing security environment, thereby better addressing various complex cybersecurity challenges. This flexibility and adaptability enable big data models to be widely applied in enterprises of different scales and types, providing stronger support for ISM. In summary, big data-driven security management models have significant advantages in enhancing the accuracy, real-time responsiveness, and adaptability of security threat detection. However, they also face challenges such as data quality, privacy protection, and technical integration. Future research should focus on how to address these issues to further improve the application value of big data models in the IS field and promote their widespread application in practical operations.

Conclusion

This study proposes a big data analysis-based risk prediction model for EIS, which improves the detection accuracy and coverage of security threats by mining various data resources. Compared to traditional methods, the model performs better in the AUC value, demonstrating stronger discriminative power. The study also showcases the advantages of BDT in ISM, effectively integrating and analyzing enterprises' historical and real-time data, especially in early warning, anomaly detection, and risk quantification. Experimental results indicate that optimizing model parameters and feature engineering markedly enhances model performance, improving enterprises' response speed and decision-making effectiveness to potential risks. Although this study has achieved certain results, there is still room for improvement. Future research could explore the impact of different industries and company sizes on model performance and design more adaptive security models. Additionally, integrating emerging DL technologies, such as graph neural networks and self-attention mechanisms, may enhance the model's performance. This study has limitations in data quality and dataset representativeness. Future work should expand the dataset and optimize verification and testing methods to ensure the model's reliability in practical scenarios.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author Limin Zhang on reasonable request via e-mail zlmin2024@163.com.

Received: 17 September 2024; Accepted: 2 January 2025
Published online: 06 January 2025

References

1. Attaran, M. The impact of 5G on the evolution of intelligent automation and industry digitization[J]. *J. Ambient Intell. Humaniz. Comput.* **14** (5), 5977–5993 (2023).
2. Fang, S., Moreno Brenes, A. & Brusoni, S. Technology Intelligence and Digitalization in the Manufacturing Industry[J]. *Research-Technology Manage.* **66** (5), 22–33 (2023).
3. Peres, R. S. et al. Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook[J]. *IEEE Access.* **8**, 220121–220139 (2020).
4. Singh, J. et al. Sales profession and professionals in the age of digitization and artificial intelligence technologies: concepts, priorities, and questions[J]. *J. Personal Sell. Sales Manage.* **39** (1), 2–22 (2019).
5. Jan, Z. et al. Artificial intelligence for industry 4.0: systematic review of applications, challenges, and opportunities[J]. *Expert Syst. Appl.* **216**, 119456 (2023).
6. Tavera Romero, C. A. et al. Business intelligence: business evolution after industry 4.0[J]. *Sustainability* **13** (18), 10026 (2021).
7. Mijwil, M. M. & Aljanabi, M. From Analog to Digitization: Rethinking Management and Operations through eHealth Integration in Industry 4.0[J]. Mesopotamian Journal of Artificial Intelligence in Healthcare, 2023: 27–30. (2023).
8. Subeesh, A. & Mehta, C. R. Automation and digitization of agriculture using artificial intelligence and internet of things[J]. *Artif. Intell. Agric.* **5**, 278–291 (2021).
9. Singh, N., Krishnaswamy, V. & Zhang, J. Z. Intellectual structure of cybersecurity research in enterprise information systems[J]. *Enterp. Inform. Syst.* **17** (6), 2025545 (2023).
10. Soltani Delgosha, M., Hajiheydari, N. & Fahimi, S. M. Elucidation of big data analytics in banking: a four-stage Delphi study[J]. *J. Enterp. Inform. Manage.* **34** (6), 1577–1596 (2021).
11. Gaurav, A., Gupta, B. B. & Panigrahi, P. K. A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system[J]. *Enterp. Inform. Syst.* **17** (3), 2023764 (2023).
12. Aguboshim, F. C., Obiokafor, I. N. & Emenike, A. O. Sustainable data governance in the era of global data security challenges in Nigeria: a narrative review[J]. *World J. Adv. Res. Reviews.* **17** (2), 378–385 (2023).
13. Gebremeskel, B. K., Jonathan, G. M. & Yalew, S. D. Information security challenges during digital transformation[J]. *Procedia Comput. Sci.* **219**, 44–51 (2023).
14. Khan, M. A. et al. The role of post-implementation strategies for projects of enterprise information systems in enhancing management system: a case study approach[J]. *Hum. Syst. Manage.* **42** (2), 247–256 (2023).
15. Lachenmaier, J., Weber, P. & Lasi, H. Enterprise Information Systems vs. Digital Twins–A Case Study on the Properties, Purpose, and Future Relationship in the Logistics Sector[J]. (2023).
16. Musti, K. S. S. & Baporikar, N. Industry 4.0-based enterprise information system for P2P lending[J]. *J. Sci. Technol. Policy Manage.* **14** (1), 6–24 (2023).
17. Herath, T. C., Herath, H. S. B. & Cullum, D. An information security performance measurement tool for senior managers: balanced scorecard integration for security governance and control frameworks[J]. *Inform. Syst. Front.* **25** (2), 681–721 (2023).
18. Tejay, G. P. S. & Mohammed, Z. A. Cultivating security culture for information security success: a mixed-methods study based on anthropological perspective[J]. *Inf. Manag.* **60** (3), 103751 (2023).
19. Zuo, Y. Big data and big risk: a four-factor framework for big data security and privacy[J]. *Int. J. Bus. Inform. Syst.*, **42**(2): 224–242. (2023).
20. Tang, W. & Yang, S. Enterprise digital management efficiency under cloud computing and big data[J]. *Sustainability* **15** (17), 13063 (2023).
21. Pejić-Bach, M., Jajić, I. & Kamenjarska, T. A bibliometric analysis of phishing in the Big Data era: high focus on algorithms and low focus on people[J]. *Procedia Comput. Sci.* **219**, 91–98 (2023).
22. Wang, Y. et al. The trickle-down effect of big data use to predict organization innovation: the roles of business strategy alignment and information sharing[J]. *J. Enterp. Inform. Manage.* **36** (1), 323–346 (2023).
23. He, W., Hung, J. L. & Liu, L. Impact of big data analytics on banking: a case study[J]. *J. Enterp. Inform. Manage.* **36** (2), 459–479 (2023).
24. Ragazou, K. et al. Big data analytics applications in information management driving operational efficiencies and decision-making: mapping the field of knowledge with bibliometric analysis using R[J]. *Big Data Cogn. Comput.* **7** (1), 13 (2023).

25. Kumar, N., Kasbekar, G. S. & Manjunath, D. Application of data collected by endpoint detection and response systems for implementation of a network security system based on zero trust principles and the eigentrust algorithm[J]. *ACM SIGMETRICS Perform. Evaluation Rev.* **50** (4), 5–7 (2023).
26. Cao, Y. & Yang, M. Legal regulation of big data killing:--From the perspective of Personal Information Protection Law[J]. *J. Educ. Humanit. Social Sci.* **7**, 233–241 (2023).
27. Ferreira, D. J., Mateus-Coelho, N. & Mamede, H. S. Methodology for predictive cyber security risk assessment (PCSRA)[J]. *Procedia Comput. Sci.* **219**, 1555–1563 (2023).
28. Valmohammadi, C. & Varaee, F. Analyzing the interaction of the challenges of big data usage in a cloud computing environment[J]. *Bus. Inform. Rev.* **40** (1), 21–32 (2023).
29. Alomari, E., Katib, I., Mehmood, R. & Iktishaf A big data road-traffic event detection tool using Twitter and spark machine learning[J]. *Mob. Networks Appl.* **28** (2), 603–618 (2023).
30. Li, L. et al. Big data and big disaster: a mechanism of supply chain risk management in global logistics industry[J]. *Int. J. Oper. Prod. Manage.* **43** (2), 274–307 (2022).
31. Sánchez-Zas, C. et al. Ontology-based approach to real-time risk management and cyber-situational awareness[J]. *Future Generation Comput. Syst.* **141**, 462–472 (2023).
32. Bandari, V. Enterprise data security measures: a comparative review of effectiveness and risks across different industries and organization types[J]. *Int. J. Bus. Intell. Big Data Analytics*. **6** (1), 1–11 (2023).
33. Singh, R. K. et al. Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda[J]. *TQM J.* **35** (1), 262–291 (2023).
34. Chen, X. & Metawa, N. Enterprise financial management information system based on cloud computing in big data environment[J]. *J. Intell. Fuzzy Syst.* **39** (4), 5223–5232 (2020).
35. Dhirani, L. L. et al. Ethical dilemmas and privacy issues in emerging technologies: a review[J]. *Sensors* **23** (3), 1151 (2023).
36. Ke, T. T. & Sudhir, K. Privacy rights and data security: GDPR and personal data markets[J]. *Manage. Sci.* **69** (8), 4389–4412 (2023).
37. Nguyen, M. T. & Tran, M. Q. Balancing security and privacy in the digital age: an in-depth analysis of legal and regulatory frameworks impacting cybersecurity practices[J]. *Int. J. Intell. Autom. Comput.* **6** (5), 1–12 (2023).
38. Debbarma, R. The changing landscape of privacy laws in the age of big data and surveillance[J]. *Rivista Italiana Di Filosofia Analitica Junior*. **14** (2), 1740–1752 (2023).
39. González-Granadillo, G., González-Zarzosa, S. & Diaz, R. Security information and event management (SIEM): analysis, trends, and usage in critical infrastructures[J]. *Sensors* **21** (14), 4759 (2021).
40. Mirtsch, M., Kinne, J. & Blind, K. Exploring the adoption of the international information security management system standard ISO/IEC 27001: a web mining-based analysis[J]. *IEEE Trans. Eng. Manage.* **68** (1), 87–100 (2020).
41. Ahmad, A. et al. How integration of cyber security management and incident response enables organizational learning[J]. *J. Association Inform. Sci. Technol.* **71** (8), 939–953 (2020).
42. Miao, F. et al. Wearable sensing, big data technology for cardiovascular healthcare: current status and future prospective[J]. *Chin. Med. J.* **136** (9), 1015–1025 (2023).
43. Pathak, S., Krishnaswamy, V. & Sharma, M. Big data analytics capabilities: a novel integrated fitness framework based on a tool-based content analysis[J]. *Enterp. Inform. Syst.* **17** (1), 1939427 (2023).
44. Kumar, D. S. et al. A novel distributed file system using Blockchain Metadata[J]. *Wireless Pers. Commun.* **129** (1), 501–520 (2023).
45. He, Q. et al. Design and optimization of a distributed file system based on RDMA[J]. *Appl. Sci.* **13** (15), 8670 (2023).
46. Ibtisum, S., Rahman, S. M. A. & Hossain, S. M. S. Comparative analysis of MapReduce and Apache Tez performance in Multinode clusters with data compression[J]. *World J. Adv. Res. Reviews*. **20** (3), 519–526 (2023).
47. Supsermpol, P., Thajchayapong, S. & Chiadamrong, N. Predicting financial performance for listed companies in Thailand during the transition period: a class-based approach using logistic regression and random forest algorithm[J]. *J. Open. Innovation: Technol. Market Complex.* **9** (3), 100130 (2023).
48. Awad, F. H., Hamad, M. M. & Alzubaidi, L. Robust classification and detection of big medical data using advanced parallel K-means clustering, YOLOv4, and logistic regression[J]. *Life* **13** (3), 691 (2023).
49. Runchi, Z., Liguo, X. & Qin, W. An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects[J]. *Expert Syst. Appl.* **212**, 118732 (2023).
50. Sommestad, T., Holm, H. & Steinvall, D. Variables influencing the effectiveness of signature-based network intrusion detection systems[J]. *Inform. Secur. Journal: Global Perspective*. **31** (6), 711–728 (2022).
51. Jiang, R., Ma, Z. & Yang, J. An assessment model for cloud service security risk based on entropy and support vector machine[J]. *Concurrency Computation: Pract. Experience*. **33** (21), e6423 (2021).
52. Jingle, I. D. J. & Paul, P. M. A collaborative defense protocol against collaborative attacks in wireless mesh networks[J]. *Int. J. Enterp. Netw. Manage.* **12** (3), 199–220 (2021).
53. Gupta, C. et al. A systematic review on machine learning and deep learning models for electronic information security in mobile networks[J]. Sensors, 22(5): 2017. (2022).
54. Javid, T., Gupta, M. K. & Gupta, A. A hybrid-security model for privacy-enhanced distributed data mining[J]. *J. King Saud University-Computer Inform. Sci.* **34** (6), 3602–3614 (2022).
55. Himeur, Y. et al. AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives[J]. *Artif. Intell. Rev.* **56** (6), 4929–5021 (2023).
56. Kaffash, S., Nguyen, A. T. & Zhu, J. Big data algorithms and applications in intelligent transportation system: a review and bibliometric analysis[J]. *Int. J. Prod. Econ.* **231**, 107868 (2021).
57. George, A. S. Securing the future of finance: how AI, Blockchain, and machine learning safeguard emerging Neobank technology against evolving cyber threats[J]. *Partners Univers. Innovative Res. Publication*. **1** (1), 54–66 (2023).
58. Vadhil, F. A., Nanne, M. F. & Salihi, M. L. Importance of machine learning techniques to improve the open source intrusion detection systems[J]. *Indonesian J. Electr. Eng. Inf. (IJEEI)*. **9** (3), 774–783 (2021).

Acknowledgements

The work was supported in part by Scientific Research Project of Hunan Provincial Department of Education (Grant No.23A0659). This work was also supported by Research on the gain effect of school-enterprise collaborative education in the mixed teaching mode ——Using “Computers” Taking the course as an example (Grant No.ND233163). The work was also supported by Construction of Employment Internship Base for Software Engineering Major (Grant No.2023122952026).

Author contributions

Ping Li: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparationLimin Zhang: writing—review and editing, visualization, supervision, project administration, funding acquisition.

Declarations

Competing interests
The authors declare no competing interests.

Ethics statement
This article does not contain any studies with human participants or animals performed by any of the authors. All methods were performed in accordance with relevant guidelines and regulations.

Additional information
Correspondence and requests for materials should be addressed to L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.