Advancing Cybersecurity through the Development of a Semantic Knowledge Base: Novel Methods and Schemes for Multi-source Data Integration

Yue Zhao [1,2,3], Gang Xiong [3], Jincheng Guo [4,*], Yi Ni [5], Bin Yang [2], and Qi Zhao [1,3]

[1] National Key Laboratory of Security Communication, Chengdu, China
[2] School of Computer and Information Engineering, Chuzhou University, Chuzhou, China
[3] No. 30 Research Institute of China Electronics Technology Group Corporation, Chengdu, China
[4] China Communications Construction Company and Dalian University of Technology Institute of Communication Technology CO., LTD, Dalian, China
[5] Sichuan Aerospace 706 Information Technology Co Ltd, Chengdu, China
Email: yuezhao@foxmail.com (Y.Z.); cetc30xg18@163.com (G.X.); guojincheng@pdiwt.com.cn (J.G.); ny_cdcn@163.com (Y.N.); ybcup@chzu.edu.cn (B.Y.); zq8484@yeah.net (Q.Z.)
*Corresponding author

***Abstract*—The paper presents a novel method for constructing a multi-domain, multi-layer cybersecurity semantic knowledge base. Utilizing diverse cybersecurity data sources, the framework seamlessly integrates entity extraction, event extraction, and semantic relation extraction through sophisticated semantic parsing techniques, resulting in a highly accurate and comprehensive knowledge base. By employing Bi-directional Long Short-Term Memory (Bi-LSTM) models with attention mechanisms for precise entity extraction, a hierarchical strategy network for detailed event extraction, and pattern rule matching for elaborate semantic relation extraction, the proposed method demonstrates exceptional efficacy. Experimental results show entity extraction with 95.4% accuracy and 92.5% recall, event extraction with 93.3% accuracy and 90.8% recall, and semantic relation extraction with 90.7% accuracy and 88.6% recall. The constructed knowledge base achieves an average query response time of 0.5 s and a query accuracy of 92%. This innovative approach enhances the processing and understanding of complex cybersecurity data, providing reliable and precise semantic knowledge query and reasoning, crucial for dynamic threat response.**

***Keywords*—semantic knowledge base, cybersecurity, semantic parsing, attention mechanism, hierarchical policy networks**

I. Introduction

The semantics is the knowledge used in natural language processing to understand and infer semantic relationships, assisting computers in linking and reasoning among knowledge to more accurately understand the true meaning of text. A semantic knowledge base is a structured knowledge base that provides representation and parsing of relevant semantic knowledge and describes semantic relationships. The paper uses attack and defense data, security events, security intelligence, and security knowledge as data sources to propose a multi domain and multi-level cybersecurity semantic knowledge base construction method and system. It constructs a semantic extraction framework that covers entity knowledge extraction, event knowledge extraction, and semantic relationship extraction. Through semantic mapping technology, the interrelationships among entity knowledge, event knowledge, and semantic relationships are established, thereby achieving more accurate and rich associations and parsing among semantic knowledge. The core objective of the paper is to construct and develop a semantic knowledge base tailored specifically for cybersecurity, addressing the unique challenges within this domain. By focusing on multi-modal, multi-type, and multi-format data integration, the proposed framework offers a unified semantic representation that is crucial for effective cybersecurity measures. The novel methods and schemes ensure that the semantic knowledge base can handle the intricate nature of cybersecurity data, thereby providing a robust foundation for intelligent cybersecurity applications.

Building a semantic knowledge base for cybersecurity is an indispensable but extremely challenging task in the construction and development of intelligent cybersecurity [1, 2]. The existing common methods mainly extract entity relationships from individual sentences in text, but due to the special nature of cybersecurity data, it increases the difficulty of constructing a cybersecurity semantic knowledge base. Firstly, cybersecurity data comes from different networks,

Manuscript received August 28, 2024; revised September 9, 2024; accepted October 9, 2024; published February 10, 2025.

applications, and entities, resulting in varying data structures and formats, leading to data consistency issues. It is necessary to merge, clean, and standardize data from different sources for effective semantic analysis. Secondly, cybersecurity data includes various specific concepts, technical terms, and contextual information. Many entities are often present in multiple sentences or even files within the context. The main challenge is how to extract entity knowledge in more complex contexts. Finally, cybersecurity data cannot guarantee the complete recording of all events and situations, which may lead to missing or incomplete data [3]. Therefore, semantic parsing algorithms need to use appropriate methods to infer possible complete information from fragmented and localized data, to better support cybersecurity defense actions, the systematic application of elements, and collaborative linkage.

At present, there are relatively few literatures related to the semantic knowledge base of cybersecurity. Lin *et al.* [4] proposed a method to calculate the similarity between two-character entities by integrating their attribute information and relationship information, to determine whether they are the same entity. Although the method belongs to the field of cybersecurity application technology, it only involves the construction of character entity models and the comparison of similarity between character entities. Lu *et al.* [5] presented a method for constructing a diabetes knowledge graph using Bidirectional Encoder Representations from Transformers (BERT)-based models, significantly improving entity and relation extraction performance. However, there is a lack of discussion on the scalability and generalizability of the proposed method to cybersecurity domains. Yao *et al.* [6] proposed an ontology-based method for constructing an information security knowledge base, described the construction process in detail, and designed the knowledge schemas. However, the practical validation of this method is limited to specific development cases and heavily relies on manual operations, without addressing its broader applicability and usability issues. Compared with the above methods, the paper proposes a semantic knowledge base construction method for multi-source heterogeneous cybersecurity data such as attack and defense data, security events, security intelligence, and security knowledge. Based on a comprehensive review of related literatures, the paper is the first in the field of cybersecurity to propose a semantic parsing method that encompasses entity knowledge, event knowledge, and semantic relationships.

These advancements underline the substantial progress made in developing a semantic knowledge base specifically designed to meet the needs of the cybersecurity domain. By directly addressing the complexities inherent in cybersecurity data, the novel methods and schemes pave the way for more effective and intelligent cyber defense mechanisms. This paper offers significant technical contributions to the efficient query and associative reasoning of cybersecurity semantic knowledge, demonstrating high practical value and underlining the importance of our specialized focus within the broader context of cybersecurity. The cybersecurity data is diverse and complex, comprising attack and defense data, security events, security intelligence, and security knowledge. By unifying these diverse data types into a coherent semantic structure, the novel methods and schemes enable more robust and intelligent analysis of cybersecurity threats and events. The paper offers a detailed example of extracting causal relationships from cybersecurity events, illustrating how a comprehensive semantic framework can enhance the understanding of cyber threats. By focusing on multi-source data integration and causal relationship extraction, the semantic knowledge base supports sophisticated cybersecurity strategies and intelligent threat analysis.

On the theoretical nature side, we propose the novel methods and schemes to constructing a semantic knowledge base, enhancing the accuracy and relevance of the knowledge base through the integration and semantic analysis of multi-source data. We also explore innovative schemes for data integration, leveraging machine learning and natural language processing techniques to achieve efficient data consolidation and matching. These theoretical methodologies offer new avenues for knowledge management and threat intelligence analysis in the cybersecurity domain.

On the practical application side, the novel methods and schemes can be directly applied to cybersecurity systems, enhancing the accuracy and speed of threat detection and response. For example, by integrating multi-source data, we can aggregate and analyze security events and threat intelligence from various channels in real-time, providing more comprehensive threat intelligence reports. Additionally, the application of the semantic knowledge base can optimize the automated response processes within cybersecurity systems, reducing the need for human intervention and increasing the level of automation and intelligence in the systems.

The remainder of this paper is organized as follows. Section II describes the system model of the cybersecurity semantic knowledge base. Section III presents the integrated semantic extraction method for entities, events, and relations, as well as the Resource Description Framework (RDF) construction process based on the aggregated mapping method. Section IV provides the simulation environment, the numerical results, and some discussions, and Section V concludes the paper.

II. System Model

The paper proposes a method for constructing a cybersecurity semantic knowledge base that includes four stages: preliminary preparation, semantic extraction, semantic mapping, and semantic knowledge base construction. The detailed process of constructing the cybersecurity semantic knowledge base is illustrated in Fig. 1.

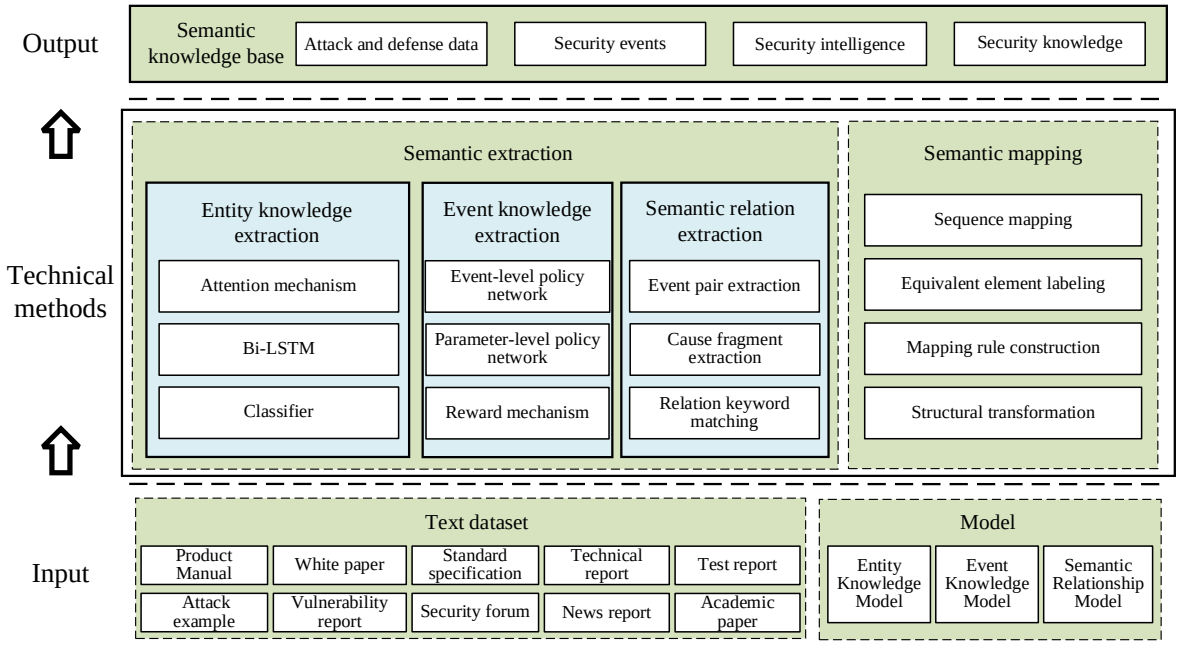

Fig. 1. The construction process of cybersecurity semantic knowledge base.

In the preliminary preparation stage, cybersecurity text data such as security product manuals, white papers, standards, technical reports, test reports, attack-defense cases, vulnerability reports, security forums, news reports, and academic papers are input. Simultaneously, entity knowledge models, event knowledge models, and semantic relationship models are set up to lay the foundation for subsequent semantic extraction and semantic mapping, thereby establishing cybersecurity semantic knowledge bases across multiple domains.

To construct the entity knowledge model, key entities in the cybersecurity domain must be identified first, followed by collecting specific attributes related to these entities, such as user identities [7], permission levels [8], confidentiality levels of files, and access permissions [9]. Based on application requirements, relevant entity relationships are set, such as affiliations of users and organizations, and associations among devices and files.

Constructing the event knowledge model involves classifying events according to the cybersecurity domain, determining the types of events that need to be collected and recorded, and gathering important information related to these events, such as the time, place, description, and impact level of events [10]. Based on application needs, relationships between events and related entities are established, such as entities triggering the events, entities affected by the events, and entities describing event responses and countermeasures.

The semantic relationship model is built by first identifying common semantic relationships in the cybersecurity domain, such as temporal relationships, hierarchical relationships, and logical relationships. These are then further subdivided into possible types of semantic relationships, such as concurrent, sequential, and reverse order relationships for temporal relationships; superior-subordinate and part-whole relationships for hierarchical relationships; and causal, conditional, parallel, and adversative relationships for logical relationships.

Semantic extraction involves extracting cybersecurity entities, events, and relationships. These elements are indispensable and mutually complementary in the cybersecurity domain, supporting the construction and application of the cybersecurity semantic knowledge base.

Cybersecurity entities include attackers, targets, vulnerabilities, malware, and security tools [11]. Through entity extraction, cybersecurity-related semantic information can be structured, facilitating subsequent semantic association and inference analysis.

Cybersecurity events include network attacks, malware incidents, information disruption events, content security events, and equipment failures. Event extraction allows for the automated processing and classification of large amounts of cybersecurity semantic knowledge, identifying different types of events, which helps in real-time monitoring of the occurrence and evolution of cybersecurity incidents. This enables timely detection of anomalies and the adoption of appropriate countermeasures, enhancing emergency response capabilities for cybersecurity incidents. Additionally, event extraction facilitates the mining and analysis of cybersecurity event data, uncovering relationships between events and potential attack patterns [12], thus better informing the formulation and implementation of security strategies and measures, improving the efficiency and accuracy of cybersecurity incident handling.

In cybersecurity semantic extraction, extracting relationships between entities or events can help identify their interdependencies and correlations. For example, the relationship between attackers and attack methods, and

the relationship between vulnerabilities and affected systems, are crucial for identifying potential threat behaviors and vulnerability exploitation methods [13]. Relationship extraction allows for the construction of complex association relationships among cybersecurity entities, providing support for security event analysis and threat warning.

III. The Entity-Event-Relation Integrated Semantic Extraction Method

The section provides a semantic extraction framework that encompasses entity knowledge extraction, event knowledge extraction, and semantic relation extraction, employing an entity-event-relation integrated semantic extraction approach. The integrated semantic extraction framework is illustrated in Fig. 2. The section presents a series of new methods and schemes to effectively address the increasingly complex nature of cyber threats. By developing novel algorithms for these processes, we enhance the extraction and integration of relevant data, ensuring semantic coherence across different domains in cybersecurity. Through these innovative methods and schemes, our research contributes to the creation of a robust semantic knowledge base that facilitates better understanding and analysis of cybersecurity phenomena.

A. Cybersecurity Entity Extraction

The paper employs Bi-directional Long Short-Term Memory (Bi-LSTM) networks based on the attention mechanism for extracting cybersecurity entities and concepts. The attention mechanism allows the machine learning model to focus on the most relevant cybersecurity entities or concepts for the current task, detecting the most critical content in the input data. Bi-LSTM assists the machine learning model in understanding the context in which cybersecurity entities and concepts are situated, enabling more accurate recognition of the entities and concepts based on their surrounding context. The machine learning model for entity knowledge extraction specifically includes an input layer, a Bi-LSTM layer, an attention mechanism layer, and an output layer.

In the input layer, cybersecurity text data is first preprocessed, including text cleaning, standardizing text formats, tokenization, stemming extraction, and part-of-speech label. The preprocessed text sequences are then fed into the Bi-LSTM layer for further processing.

The Bi-LSTM layer comprises a forward LSTM and a backward LSTM. The forward LSTM traverses the text sequence from start to finish, constructing a forward contextual understanding. The backward LSTM traverses the text sequence from end to start, constructing a backward contextual understanding. Therefore, for each word in the text sequence, regardless of its position, the model can effectively recognize and understand its context. In the task of cybersecurity entity extraction, the output processed by the Bi-LSTM is fed into a classifier, e.g., softmax, to identify whether each word belongs to a specific entity category, e.g., attacker, attack type, and victim asset. Since each word incorporates contextual information, the model can more accurately determine each word's part of speech and its association with other cybersecurity entities.

The output of the Bi-LSTM layer is a sequence of vectors containing surrounding contextual information. These vector sequences capture not only the information of the words themselves but also integrate the information of preceding and following words, thereby providing a comprehensive semantic representation. The vector sequence is denoted as $\boldsymbol{h}=(h_1, h_2, \ldots, h_T)$, where h_T represents the hidden state vectors of both the forward and backward LSTM at time step T.

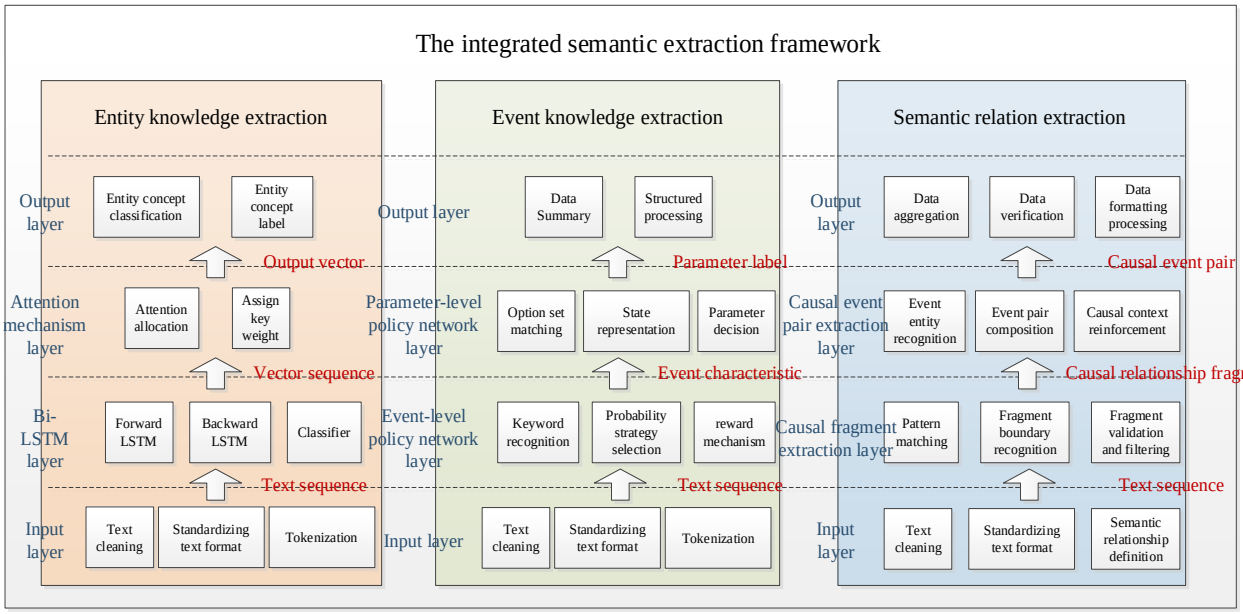

Fig. 2. The integrated semantic extraction framework.

The output of the Bi-LSTM layer is passed to the attention mechanism layer, which mimics a key characteristic of human attention, i.e., focusing on certain pieces of information while ignoring others when processing large amounts of data. The attention mechanism layer first needs to decide how to allocate attention. The query is a vector representing the current focus or the specific target of the task. For a given query, the attention mechanism assigns a weight to each key in the sequence by calculating the similarity between the query and each key. Similarity can be computed using the dot-product as score(q,k_i)=q·k_i. After obtaining the weight for each key, these weights are normalized using a SoftMax function to ensure that the sum of all weights equals 1, i.e.,

$$\alpha_i = \frac{\exp(\text{score}(q, k_i))}{\sum_{j=1}^{T} \exp(\text{score}(q, k_i))} \quad (1)$$

After the weight normalization operation, the weights are transformed into probabilities. Each value is weighted according to its corresponding weight, meaning that each element of the vector sequence input from the Bi-LSTM layer will have its output determined based on its assigned importance. The higher the weight, the greater the corresponding output value of the respective element [14]. Finally, all the weighted values are summed to form the output vector, which represents the significant information for the current task. The output vector (v) is represented as

$$v = \sum_{i=1}^{T} \alpha_i v_i \ .\backslash \quad (2)$$

Each element in the vector sequence undergoes this operation, resulting in an output vector sequence of the same length as the input vector sequence.

The primary task of the output layer is to classify and label entities and concepts based on the preceding processing results. The output layer converts the output vectors from the attention mechanism layer into final classification decisions, typically selecting the output vector with the highest value as the label for the entity or concept. In certain cases, additional post-processing operations are performed after the output layer to optimize the model's final output. This may include adjusting or optimizing classification results based on specific rules, such as using a knowledge base in the field of cybersecurity to correct entity recognition or applying heuristic methods to improve the recognition of specific entities or concepts by the machine learning model. The final output is structured data labeled with entity and concept types and their related information.

B. Cybersecurity Event Extraction

In this paper, we propose hierarchical policy networks to extract knowledge from cybersecurity events. The hierarchical policy networks are designed for layered detection of cybersecurity events and their parameters, and the event knowledge extraction process consists of a two-tier network, i.e., the event-level policy networks and the parameter-level policy networks. The machine learning model for event knowledge extraction specifically includes an input layer, an event-level policy network layer, a parameter-level policy network layer, and an output layer.

At the input layer, the textual data related to cybersecurity events are fed into the model, and preprocessing operations are performed. The preprocessed data serve as the input for the next layer to ensure standardization and consistency, like the preprocessing operations for entity knowledge extraction.

In the event-level policy network layer, the model performs a comprehensive analysis of the preprocessed cybersecurity event text data, scanning the text from start to finish character by character. When the model identifies a keyword for a cybersecurity event, it uses a probability-based random strategy option and a reward mechanism to determine the event type and detailed information associated with that keyword. This mechanism accurately identifies and classifies cybersecurity events. The option set, predefined based on previously collected cybersecurity event data, includes various types of cybersecurity events and related non-keywords. This option set clearly identifies the specific categories of cybersecurity events associated with the current key word.

Based on the identified key word, the model selects the most likely cybersecurity event type from the option set through probabilistic analysis. Once the model identifies a specific event type, it provides an immediate reward, estimated based on the short-term impact of the event [15, 16].

$$P(E \mid t_i) = \frac{\exp(f(t_i, E))}{\sum_j \exp(f(t_i, E_j))} \quad (3)$$

where $f(t_i, E)$ is a scoring function that measures the relevance of key word t_i to event type E.

The immediate reward $R(E)$ for identifying an event type E is based on the short-term impact as

$$R(E) = \sum_k \gamma^k r_k(E) \quad (4)$$

where $r_k(E)$ is the reward at step k and γ is the discount factor.

If the model identifies a new event type at the current step, it remains in the event-level policy state. If it recognizes a specific event from the option set, it initiates a new subtask and transitions to the parameter-level policy network layer for detailed analysis of the parameters involved in the event. This approach allows the event-level policy network to effectively identify and classify cybersecurity events and enhance the specificity and effectiveness of security event handling by establishing connections between events and related entities.

Upon detecting a specific cybersecurity event matching the option set, the model transitions from the event-level policy network layer to the parameter-level policy network layer. The goal of the parameter-level

policy network layer is to assign appropriate labels to the model parameters extracted for cybersecurity events, clarifying the roles and functions of each parameter. The model uses time steps as a benchmark, providing more detailed event information to make more precise parameter decisions. The parameter-level processing includes the options and state representations from the event-level policy network layer as auxiliary inputs.

The state input of the parameter-level processing also includes descriptive texts of historical events, recognized initial events, and specific cybersecurity events the model is currently processing. This includes detailed information about specific events, such as attack types, times, affected systems, and potential impacts of the attacks. In the parameter-level policy network layer, this event information helps the model understand the context, severity, and other relevant factors of the event, assisting the model in correctly classifying and labeling each parameter, thereby considering both global and detailed event information in the extraction and induction of security event knowledge.

The probability of assigning a label L to parameter θ_i given the event type E is represented as

$$P(L \mid \theta_i, E) = \frac{\exp(g(\theta_i, E, L))}{\sum_j g(\theta_i, E, L_j))} \tag{5}$$

where $g(\theta_i, E, L)$ is a scoring function that measures the relevance of parameter θ_i and event type E to label L.

The output layer aggregates all identified cybersecurity event types, related parameters, and specific labels of these parameters, including but not limited to attack methods, affected entities, potential consequences, and other event information. The aggregated information is then presented as structured data.

C. Cybersecurity Entity Extraction Cybersecurity Semantic Relations Extraction

The paper adopts a pattern rule matching method to extract semantic relations of cybersecurity events. Firstly, pattern matching is used to match keywords and extract fragments of cybersecurity event semantic relations. Then, a rule constraint method is used to extract event pairs from these fragments. Finally, the event pairs are combined with keywords to achieve semantic relation extraction. The paper focuses on the extraction of causal semantic relations of cybersecurity events. The machine learning model for semantic relation extraction specifically includes an input layer, a causal fragment extraction layer, a causal event pair extraction layer, and an output layer. The causal relation extraction process based on template rule matching is shown in Fig. 3.

The main functions of the input layer include two aspects. One is to perform preprocessing operations on the textual data of cybersecurity events, like the preprocessing operations in entity knowledge extraction and event knowledge extraction. The other is to define five matching patterns suitable for extracting causal relations of cybersecurity events by analyzing the expression features of cybersecurity textual data. Based on the positions of the cause-and-effect clauses, the pattern types are classified into cause-to-effect or effect-to-cause. The templates for causal relation pattern matching are shown in Table I.

The goal of the causal fragment extraction layer is to identify the boundaries of causal fragments. For a sentence containing a causal cybersecurity event, the model first performs word segmentation and part-of-speech tagging. Then, it combines the tagging results with the explicit causal relation matching patterns described in the table to construct regular expressions, extracting the cause-and-effect fragments from the original text sentence. In Fig. 3, for a sentence with causal keywords, such as On July 23, 2024, to steal sensitive data, the attacker implanted malware X into the computer Y containing sensitive data, the cause-to-effect preposition <prep> template matching results in the cause fragment the attacker implanted malware X into the computer Y containing sensitive data, the keyword to, and the effect fragment steal sensitive data.

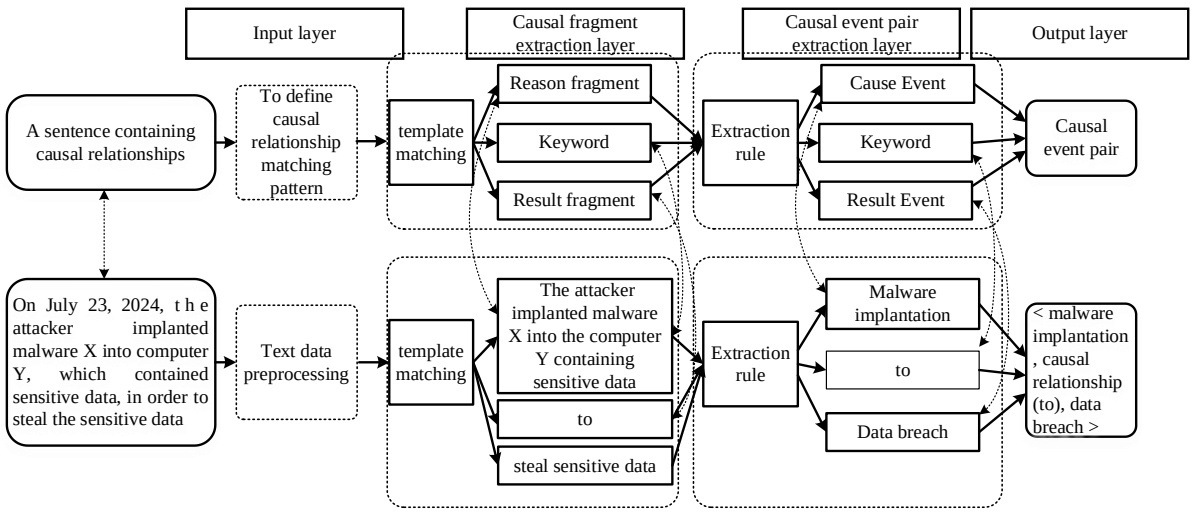

Fig. 3. The causal relation extraction process based on template rule matching.

TABLE I. THE TEMPLATES FOR CAUSAL RELATION PATTERN MATCHING

Template type	Template matching	Keywords for causal relationships
The complementary style from effect to cause	<conj>{effect}, <conj>{cause}	<the reason why>, <for this purpose>, <for this reason>, …
The complementary style from cause to effect	<conj>{cause}, <conj>{effect}	<because, therefore>, <because, for this reason>, <unless, only>, <because, causing>, …
The central style from cause to effect	{cause}, <conj…>{effect}	so, thus, causing, therefore, …
	{cause}, <verb|adverb…>{effect}	before, lest, in order to avoid, …
	{cause}, <prep…>{effect}	depend on, be decided by, hinge on, …
The frontend style from cause to effect	<prep…>{cause}, {effect}	for, based on, in accordance with, due to, …
	<conj…>{cause}, {effect}	if, as long as, assuming, …
The central style from effect to cause	<effect><prep…>{cause}	originate from, derived from, …

The cause-and-effect fragments extracted by the causal fragment extraction layer are still unstructured text and need to be further processed into structurally represented events. Although the relation expression template can identify the boundaries of cause-and-effect fragments in the text sentences, certain issues may arise. For instance, the cause-and-effect fragments may not contain keywords for cybersecurity events, making it impossible to extract valid causal event pairs and thus rendering the fragments invalid. Additionally, a sentence may be successfully matched by multiple complementary or combined cause-to-effect expression templates, yet the event pairs may not form valid causal relations.

Since the event keywords for cybersecurity events have already been obtained during the event knowledge extraction stage, the paper addresses the issues in the causal fragment extraction stage by designing extraction rules and rule constraints in the causal event pair extraction layer to achieve the extraction of causal event pairs. The specific process is as follows.

(1) To enforce the event keyword rule constraint, each set of cause-and-effect fragments obtained from a sentence through relation expression template matching is traversed to search for event keywords. If the cause or effect fragment does not contain any event keywords, the set of fragments is discarded.

(2) If a sentence involves multiple relation patterns and multiple sets of cause-and-effect fragments are obtained from previous steps, the set of cause-and-effect fragments with the closest distance between two relation keywords in the relation expression template is selected according to the relation keyword rule constraint.

(3) It is necessary to select the relation keyword nearest to the relation keyword in the cause fragment and designate the corresponding cybersecurity event as the cause event. Similarly, the event keyword closest to the position of the relation keyword in the effect fragment should be selected, and the cybersecurity event should be designated as the effect event. Based on the data obtained from the causal fragment extraction stage, the cause fragment extraction results in the cause event malware implantation, and the effect fragment extraction results in the effect event data breach, thereby achieving semantic relation extraction.

The output layer needs to aggregate and validate the information and data collected and processed by the input layer, causal fragment extraction layer, and causal event pair extraction layer to ensure data consistency and completeness. This includes confirming that the data format of each causal event pair meets the requirements and ensuring the correctness of the causal logic. Additionally, the output layer needs to format the aggregated data, such as converting it into JavaScript Object Notation (JSON) or Extensible Markup Language (XML) format, to improve data quality and enhance the efficiency of subsequent processing.

D. Semantic Mapping

The paper adopts a method for constructing RDF based on aggregation mapping. The specific software modules include the data input module, the mapping module, and the data output module. The construction process, as illustrated in Fig. 4, describes the workflow of importing XML documents through the data input module and transforming them into RDF documents in the data output module via the mapping module.

The data input module is responsible for importing XML documents and presenting them as Document Object Model (DOM) trees. XML documents are categorized into two types, which are standalone XML documents and documents containing XML Schema.

The mapping module is divided into the RDF Schema vocabulary construction module and the RDF sequence construction module. The RDF Schema vocabulary construction module uses specific mapping rules and algorithms, along with an abstract structural model of aggregation classes, to construct RDF Schema terms. For standalone XML documents, the module directly parses the documents themselves. For documents containing XML and XML Schema, it parses the XML Schema documents referenced. The design of the mapping rules and algorithms comprises two levels. The first level involves the classification and aggregation of XML elements to construct the RDF Schema, while the second level achieves the direct mapping from XML to RDF based on the first level. The RDF sequence construction module marks equivalent elements and constructs RDF sequences to reduce redundancy in the generated RDF data. Based on the obtained aggregation classes, constructed RDF Schema terms, and mapping rules, the module processes the marked XML documents to generate the corresponding RDF sequences. The data output module merges the RDF Schema and RDF sequences to generate RDF documents, which serve as the unified data structure for the semantic knowledge base.

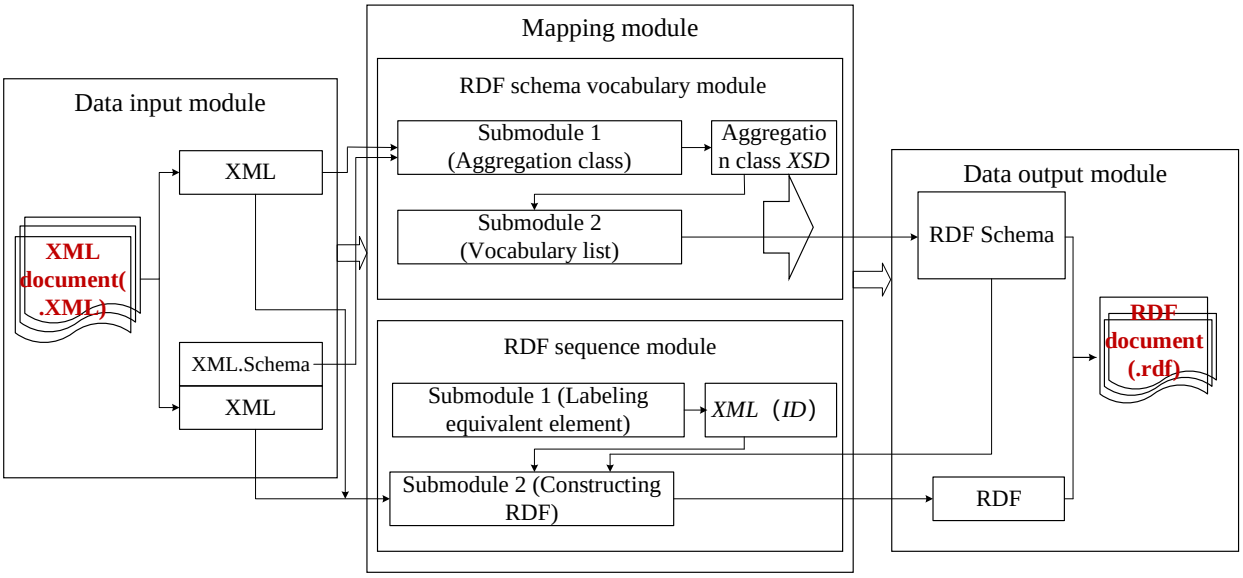

Fig. 4. The RDF construction process based on aggregated mapping.

E. Construction of Semantic Knowledge Base

Recently, the landscape of cyber threats has grown increasingly complex, necessitating more advanced and versatile approaches to cybersecurity. The proposed methods and schemes represent significant advancements in this domain. By developing a unified and comprehensive semantic knowledge base, these novel solutions not only improve the detection and analysis of complex cyber threats but also enhance the capability to respond proactively. Traditional methods often operate in isolation, focusing on specific domains or types of threats. In contrast, our approach leverages cross-domain correlations and semantic consistency, providing a more holistic and effective defense mechanism against evolving cyber threats. This innovative integration of semantically rich data and advanced analytical techniques marks a substantial leap forward in cybersecurity practices, particularly in the efficient integration of multi-source data.

The semantic knowledge integration and optimization involve integrating data collected at different stages into a unified RDF document data structure. Data quality and storage performance are optimized through deduplication operations. Additionally, the classification and indexing strategies are implemented to ensure quick and accurate query of required information, thereby maintaining the integrity and accuracy of the semantic knowledge base.

The fusion process of semantic extraction methods involves effectively integrating extracted entities, concepts, event knowledge, and semantic relationships within the field of cybersecurity. The Bi-LSTM network enhances context understanding by focusing on key information about entities and concepts. The hierarchical strategy network accurately identifies relationships between cybersecurity events and their parameters through its event-level and parameter-level structures. The pattern rule matching method strengthens the extraction of semantic relationships such as event causality. Finally, this extracted information is integrated into a multi-layered, highly interrelated structure through a series of well-defined fusion algorithms and techniques, such as weight allocation and context fusion, thereby enhancing the semantic depth and practical value of the knowledge base.

The process of continuous expansion and interoperability of the semantic knowledge base involves integration with external data sources and services, including connections with other cybersecurity data pools and knowledge bases, as well as the development of Application Programming Interfaces (APIs) and Software Development Kits (SDKs) to enable third-party developers to conveniently utilize the semantic knowledge base to build applications. Additionally, APIs for updating and maintenance are established to continuously import semantic information from new cybersecurity reports and databases, while implementing feedback mechanisms to optimize the services and quality of the knowledge base.

The semantic knowledge base can improve its processing capability and response time for large amounts of real-time data by adopting corresponding extensible frameworks.

1) Larger data sets processing

The semantic knowledge base utilizes distributed computing technology to enhance its ability to process large amounts of data. By adopting Apache Hadoop and Spark frameworks [17], it can significantly accelerate data processing speed and be more efficient compared to traditional single threaded methods. It ensures that the framework remains efficient as the amount of data expands.

2) Real-time data source processing

For real-time data processing, the semantic knowledge base integrates Apache Kafka to enhance stream processing capabilities [18]. The semantic knowledge base can maintain real-time analysis and threat detection capabilities, making it suitable for dynamic cybersecurity environments that require timely data processing.

By integrating these scalable technologies, our proposed framework can efficiently manage larger data sets and adapt to real-time data sources in the field of cybersecurity. This ensures its robustness and reliability across various operational environments. Through the above design and implementation steps, the semantic knowledge base possesses dynamic updating and efficient interoperability capabilities, providing reliable and accurate semantic information resources.

In constructing a semantic knowledge base, it is essential to implement various security measures to effectively manage privacy and security risks. Firstly, data encryption is a critical component. Utilizing advanced encryption standards like AES-256 can effectively protect data during transmission and storage. By implementing end-to-end encryption, you ensure that data remains encrypted throughout its journey from source to destination, thus preventing unauthorized access whether from network attacks or internal leaks.

Moreover, secure access control is fundamental to maintaining data security. Role-Based Access Control (RBAC) enables precise allocation of access rights based on user roles [19], allowing only relevant personnel to access necessary data and functional modules. Coupled with Multi-Factor Authentication (MFA) technologies [20], such as passwords, dynamic verification codes, and biometric authentication, the security of the access layer is further enhanced. Additionally, strictly adhering to the Principle of Least Privilege (PoLP) ensures that users possess only the minimum permissions required to perform their duties [21], thereby effectively limiting unnecessary data access.

To ensure the operational transparency and security of semantic knowledge base, log auditing and monitoring are crucial. Granular log records can track all data access, modifications, and transmissions. Employing machine learning algorithms for real-time analysis can proactively identify and flag abnormal behaviors, facilitating quick responses.

In terms of privacy protection, differential privacy techniques introduce minute noise into the data, preventing exposure of network security semantic knowledge while maintaining its efficacy in statistical analysis. By managing and controlling the noise level in differential privacy through privacy budget control, an optimal balance between data effectiveness and privacy protection is achieved.

These measures constitute a robust privacy and security guarantee system in the construction of semantic knowledge base.

Section III introduces our integrated semantic extraction framework, covering entity knowledge extraction, event knowledge extraction, and semantic relation extraction. These Key technologies include named entity recognition, event extraction models, and relation extraction models. Additionally, we discussed the processes of semantic mapping and the construction of the semantic knowledge base. By utilizing these specific technologies, we efficiently integrated multi-source data, enhancing the semantic consistency and accuracy of the knowledge base and demonstrating novel methods and schemes for multi-source data integration.

IV. Result and Discussion

To validate the effectiveness of the proposed method for constructing a cybersecurity semantic knowledge base, we designed a series of experimental tests focusing on evaluating the accuracy of semantic extraction and the query efficiency of the knowledge base.

To ensure the diversity and comprehensiveness of our experiment, we utilized a range of publicly available cybersecurity text datasets. Specifically, the datasets used include MITRE ATT&CK [22], which documents adversary tactics and techniques based on real-world observations; CICIDS 2017 [23], from the Canadian Institute for Cybersecurity providing data for intrusion detection and network analysis; UNSW-NB15 [24], created by the University of New South Wales, consisting of both normal and malicious network traffic; and various security blogs and reports from reputable sources.

Table II shows the results of the semantic extraction experiment testing, reflecting the performance of the proposed method in extracting entities, events, and semantic relationships. The experiments utilized a diverse text dataset rich in cybersecurity information, including security reports, vulnerability databases, and technical white papers. Our semantic extraction method was applied to these data to extract entities, events, and semantic relationships.

The accuracy and recall rates for entity extraction were 95.4% and 92.5%, respectively, indicating that the method effectively identifies key entities in the cybersecurity domain, such as attackers, victims, vulnerabilities, and malware. A high accuracy rate suggests that most identified entities are correct, while a high recall rate indicates that most existing entities were successfully identified. The accuracy and recall rates for event extraction were 93.3% and 90.8%, respectively, demonstrating the method's ability to accurately identify and classify cybersecurity events, such as cyberattacks, data breaches, and system failures [25]. The accuracy and recall rates for semantic relationship extraction were 90.7% and 88.6%, respectively, showing the method's strong performance in identifying semantic relationships between entities and events, such as causal relationships, dependency relationships, and so on.

TABLE II. The Results of the Semantic Extraction Experiment Testing

Method	Entity knowledge extraction		Event knowledge extraction		Semantic relationship extraction	
	Accuracy rate	Recall rate	Accuracy rate	Recall rate	Accuracy rate	Recall rate
Lu's method [5]	92.9%	90.2%	90.5%	88.6%	87.8%	85.3%
Yao's method [6]	93.1%	91.4%	91.0%	89.2%	88.5%	86.9%
Our Method	95.4%	92.5%	93.3%	90.8%	90.7%	88.6%

Compared with existing semantic parsing methods, our proposed method exhibits greater flexibility and accuracy in handling diverse and complex cybersecurity texts. Traditional methods often rely on predefined rules and templates, which struggle to adapt to new threats and event types. In contrast, our method, using deep learning models and integration strategies, dynamically adapts to new semantic information, particularly excelling in dealing with cybersecurity-specific terms and events. The benefits are primarily attributed to our proposed multi-layered semantic extraction architecture and optimized algorithms. The experimental results indicate that implementing more effective and intelligent defense mechanisms can significantly mitigate cybersecurity threats by addressing the multifaceted nature of cybersecurity data through semantic extraction and semantic mapping. The novel methods and schemes enhance the ability to detect, analyze, and respond to complex cyber threats by providing a unified and comprehensive semantic framework.

Fig. 5 presents the test results for the query efficiency of the cybersecurity semantic knowledge base, reflecting the query speed and accuracy of the semantic knowledge base. The constructed semantic knowledge base achieves the average query response time of 0.5 s and a query accuracy of 92.6%, primarily due to the optimized classification and indexing strategies. With a rigorously designed semantic knowledge classification system and indexing mechanism, the semantic knowledge base can efficiently organize and retrieve semantic information, significantly enhancing query speed and query accuracy.

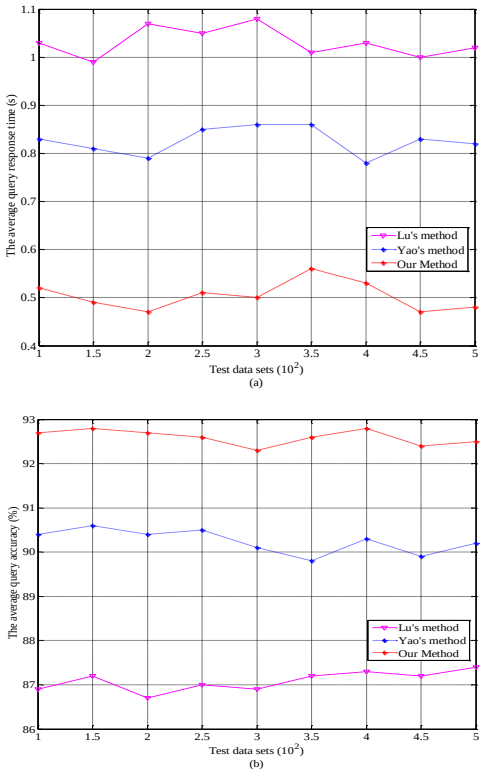

Fig. 5. The results for the query efficiency of the semantic knowledge base.

The method proposed in the paper demonstrates exceptional processing speed and system stability, maintaining high efficiency in high-load environments. This is particularly crucial for applications in the cybersecurity domain, where timely access to accurate information is essential in the face of rapidly changing threats and incidents. The high stability and low latency of the semantic knowledge base further enhance its value in emergency situations, ensuring rapid response and support. Through continuous optimization and feedback mechanisms, the semantic knowledge base constructed in this paper exhibits superior processing speed and system stability in practical semantic knowledge query and associative reasoning, providing reliable technical support for efficient query and associative reasoning of cybersecurity semantic knowledge.

V. Conclusion

The paper proposes a novel method for constructing a semantic knowledge base in the field of cybersecurity, addressing the challenges of parsing multi-modal, multi-type, and multi-format cybersecurity data to achieve unified semantic representation. The innovation of this framework lies in its comprehensive integration of entity knowledge extraction, event knowledge extraction, and semantic relation extraction techniques, all structured using an RDF-based data model. To enhance the accuracy and efficiency of semantic extraction, the framework employs an attention mechanism for entity extraction, utilizes a layered strategy network for event knowledge extraction, and applies pattern rule matching for efficient semantic relation extraction. The integration of these techniques significantly advances the comprehensive understanding and processing capabilities of semantic knowledge, enhancing the adaptability and robustness of the model to complex cybersecurity texts. The constructed semantic knowledge base not only supports dynamic updates and efficient interoperability but also provides precise semantic knowledge extraction and timely semantic knowledge query for the cybersecurity domain. The paper offers significant technical contributions to the efficient query and associative reasoning of cybersecurity semantic knowledge, demonstrating high practical value.

The paper advances the field by introducing novel methods and schemes for integrating multi-source data into a cybersecurity semantic knowledge base. Unlike existing work that typically relies on static and isolated data sets, the novel methods and schemes facilitate dynamic data interactions and adaptive learning, addressing key challenges in real-time threat detection and situational awareness. These contributions distinctively differentiate the cybersecurity semantic knowledge base from traditional knowledge bases. In developing the semantic knowledge base, we have employed advanced artificial intelligence and machine learning techniques to enhance the integration of multi-source data. The interdisciplinary methods and schemes leverage the strengths of both artificial intelligence and cybersecurity domains, providing a comprehensive

solution to contemporary cybersecurity challenges. Our methods highlight the collaborative nature of combining these diverse fields to achieve significant advancements. The potential impact and significance of the semantic knowledge base developed in this work are manifold. By integrating multi-source data through innovative semantic methods, the semantic knowledge base enables more precise threat detection and response, significantly reducing the time needed to identify and mitigate security breaches. Furthermore, it provides a comprehensive framework that enhances the interoperability among disparate security systems, thereby fostering a more collaborative and resilient cybersecurity ecosystem. The paper is relevant not only to cybersecurity researchers but also to professionals in the field. Additionally, the experts in related areas such as big data analysis and artificial intelligence may find it beneficial.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Yue Zhao, Gang Xiong, and Jincheng Guo conducted the research; Yi Ni and Bin Yang analyzed the data; Yue Zhao and Qi Zhao wrote the paper; all authors had approved the final version.

FUNDING

The work was partially supported by the Sichuan Science and Technology Innovation Talent Program under Grant No. 2024JDRC0007, the National Natural Science Foundation of China under Grant No. U20B2049, U20B2048, U20B2046, 62372076, and the National Key Laboratory of Security Communication Foundation under Grant No. 6142103022209.

ACKNOWLEDGMENT

The authors wish to thank Bo Tian, Fei Teng, Hao Zhang, and Xiaolong Chen.

REFERENCES

[1] Y. Zhao, B. Yang, F. Teng, X. Niu, N. Hu, and B. Tian, "A review of intelligent configuration and its security for complex networks," *Chinese Journal of Electronics*, vol. 33, no. 4, pp. 920–947, 2024.

[2] J. Yang, T. Li, G. Liang, W. He, and Y. Zhao, "A hierarchy distributed-agents model for network risk evaluation based on deep learning," *Computer Modeling in Engineering & Sciences*, vol. 120, no. 1, pp. 1–23, 2019.

[3] Y. Zhao, N. Hu, Y. Zhao, and Z. Zhu, "A Secure and flexible edge computing scheme for AI-driven industrial IoT," *Cluster Computing*, vol. 26, no. 1, pp. 283–301, 2023.

[4] L. Lin, Y. Zhao, J. Meng, and Q. Zhao, "A federated attention-based multimodal biometric recognition approach in IoT," *Sensors*, vol. 23, no. 13, pp. 1–14, 2023.

[5] Y. Lu, R. Zhao, S. Huang, and R. Liu, "Construction of diabetes knowledge graph based on deep learning," in *Proc. International Conference on Network and Information Systems for Computers*, Guiyang, China, 2021, pp. 966–970.

[6] Y. Yao, X. Ma, H. Liu, J. Yi, X. Zhao, and L. Liu, "A Semantic knowledge base construction method for information security," in *Proc. International Conference on Trust, Security and Privacy in Computing and Communications*, Beijing, China, 2014, pp. 803–808.

[7] N. Hu, Y. Teng, Y. Zhao, S. Yin, and Y. Zhao, "IDV: Internet domain name verification based on blockchain," *Computer Modeling in Engineering & Sciences*, vol. 129, no. 1, pp. 299–322, 2021.

[8] D. Yang, Y. Zhao, K. Wu, X. Guo, and H. Peng, "An efficient authentication scheme based on zero trust for UAV swarm," in *Proc. International Conference on Networking and Network Applications*, Lijiang, China, 2021, pp. 356–360.

[9] Y. Zhao, N. Hu, Y. Zhao, and Z. Zhu, "A secure and flexible edge computing scheme for ai-driven industrial IoT," *Cluster Computing*, vol. 26, no. 1, pp. 283–301, 2023.

[10] Y. Zhao, X. Fang, B. Huang, and Y. Chen, "Resource allocation scheme based on load balancing for ofdma two-hop relay networks," *Journal of Southwest Jiaotong University*, vol. 26, no. 1, pp. 94–101, 2013.

[11] Y. Liu, Q. Zeng, Y. Zhao, k. Wu, and y. Hao, "novel channel-hopping pattern-based wireless IoT networks in smart cities for reducing multi-access interference and jamming attacks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, pp. 1–19, 2021.

[12] Y. Zhao, Y. Yang, B. Tian, and T. Zhang, "An invocation chain test and evaluation method for fog computing," *Wireless Communications and Mobile Computing*, vol. 2020, no.1, pp. 1–11, 2020.

[13] J. Zhao, Z. Li, Y. Wang, Z. Wu, X. Ma, and Y. Zhao, "An analytical framework for reliability evaluation of d-dimensional IEEE 802.11 broadcast wireless networks," *Wireless Networks*, vol. 26, pp. 3373–3394, 2020.

[14] Y. Zhao, X. Fang, and Z. Zhao, "Interference coordination in compact frequency reuse for multihop cellular networks," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 93, no. 11, pp. 2312–2319, 2010.

[15] Y. Zhao, B. Tian, Z. Chen, J. Yang, and S. Li, "A relay-assisted secure handover mechanism for high-speed trains," *KSII Transactions on Internet and Information Systems*, vol. 13, no.2, pp. 582–296, 2019.

[16] Y. He, Y. Gu, P. Su, K. Sun, Y. Zhou, Z. Wang, and Q. Li, "A systematic study of Android Non-SDK (hidden) service API security," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 2, pp. 1609–1623, 2023.

[17] C. S. K. Sahith, S. Muppidi, and S. Merugula, "Apache spark big data analysis, performance tuning, and spark application optimization," in *Proc. International Conference on Evolutionary Algorithms and Soft Computing Techniques*, Bengaluru, India, 2023, pp. 1–8.

[18] P. L. Noac'h, A. Costan, and L. Bougé, "A performance evaluation of Apache Kafka in support of big data streaming applications," in *Proc. International Conference on Big Data*, Boston, MA, USA, 2017, pp. 4803–4806.

[19] T. Baumer, M. Müller, and G. Pernul, "System for Cross-domain Identity Management (SCIM): Survey and enhancement With RBAC," *IEEE Access*, vol. 11, pp. 86872–86894, 2023.

[20] B. O. ALSaleem and A. I. Alshoshan, "Multi-factor authentication to systems login," in *Proc. National Computing Colleges Conference*, Taif, Saudi Arabia, 2021, pp. 1–4.

[21] E. Billoir, R. Laborde, A. S. Wazan, Y. Rütschlé, and A. Benzekri, "Implementing the principle of least privilege using linux capabilities: Challenges and perspectives," in *Proc. Cyber Security in Networking Conference*, Montreal, QC, Canada, 2023, pp. 130–136.

[22] B. Al-Sada, A. Sadighian, and G. Oligeri, "Analysis and characterization of cyber threats leveraging the MITRE ATT&CK database," *IEEE Access*, vol. 12, pp. 1217–1234, 2024.

[23] L. Rakesh, L. Upadhyay, and P. M. Reddy, "Evaluation of network intrusion detection with machine learning and deep learning using ensemble Methods on CICIDS-2017 dataset," in *Proc. International Conference on Advances in Computing, Communication Control and Networking*, Greater Noida, India, 2023, pp. 1429–1433.

[24] K. G. Kumar, R. Chauhan, and D. Upadhyay, "Advancing intrusion detection with machine learning: Insights from the UNSW-NB15 dataset," in *Proc. International Conference on Information Technology, Electronics and Intelligent Communication Systems*, Bangalore, India, 2024, pp. 1–5.

[25] Y. Chen, X. Fang, and Y. Zhao, "Energy efficient adaptive power allocation in OFDM-based decode-and-forward relay link," *Journal of Electronics Information and Technology*, vol. 35, no. 2, pp. 285–290, 2013.