IJNS

International Journal of Network Security

ISSN 1816-353X (Print)
ISSN 1816-3548 (Online)

Vol. 25, No. 1 (January 2023)

INTERNATIONAL JOURNAL OF NETWORK SECURITY

PUBLISHING OFFICE

Min-Shiang Hwang
Department of Computer Science & Information Engineering, Asia University, Taichung 41354, Taiwan, R.O.C.
Email: mshwang@asia.edu.tw

International Journal of Network Security is published both in traditional paper form (ISSN 1816-353X) and in Internet (ISSN 1816-3548) at http://ijns.jalaxy.com.tw

PUBLISHER: Candy C. H. Lin

© Jalaxy Technology Co., Ltd., Taiwan 2005
23-75, P.O. Box, Taichung, Taiwan 40199, R.O.C.

Volume: 25, No: 1 (January 1, 2023)

International Journal of Network Security

A Mobile RFID Authentication Protocol Based on Self-assembling Cross-bit Algorithm

Sheng-Hua Xu[1], Dao-Wei Liu[2], and Wen-Tao Zuo[2]
(Corresponding author: Sheng-Hua Xu)

Network Information Center, Guangdong Polytechnic Normal University[1]
Tianhe District, Guangzhou 510640, China
Engineering College, Guangzhou College of Technology and Business[2]
Guangzhou 510006, China
Email: 995565519@qq.com

(Received Dec. 8, 2021; Revised and Accepted Oct. 11, 2022; First Online Dec. 10, 2022)

Abstract

The wired communication between the fixed reader and database in traditional RFID systems is considered a safe channel. However, the mobile reader and the database communicate wirelessly in mobile RFID systems, so the channel is no longer safe and reliable. Therefore, the traditional RFID authentication protocol cannot be applied to mobile RFID systems. An ultra-lightweight mobile-wireless bidirectional authentication protocol MAP-SKBO based on a shared private key and bitwise operation, is proposed to solve this problem. MAP-SKBO is based on the bitwise operation mechanism, which adopts ultra-lightweight bit replacement and self-combined cross-bit operation to encrypt the transmitted information and uses random numbers to maintain the transmission information's freshness and the shared information private key. During the communication process, the tag, the reader, and the database authenticate each other to resist sabotage by an attacker. Security analysis shows that MAP-SKBO can achieve tasks such as the dynamic update of the shared private key and desynchronization-attack resistance. The formal mathematical reasoning of MAP-SKBO by GNY logic proves the correctness of MAP-SKBO. A performance analysis indicates that MAP-SKBO has low computational complexity and is suitable for low-cost mobile RFID systems.

Keywords: Internet of Things; Mutual Authentication; RFID; Sac; Shared Key

1 Introduction

Radio frequency identification (RFID) is a non-physical contact using object recognition and data exchange technology. RFID arose in the last century, and large-scale applications were implemented in the late nineties [4,21].

Current RFID systems typically consist of three parts, the tag, the reader and the database. In a traditional RFID system, the reader is generally fixed, so the communication between the reader and the database is based on a wired channel, which is considered to be safe [3,10,12,24]. The reader is embedded in a mobile intelligent terminal to form a mobile RFID system. In a mobile RFID system, the reader is no longer fixed but mobile. Therefore, information transmission between the reader and the database can only be accomplished wirelessly. A wireless channel can easily be eavesdropped by attackers, which makes information transmission between the reader and database unreliable [9,17–20,22].

Based on the above description, the traditional RFID authentication protocol is clearly not suitable for mobile RFID systems. In view of these problems, this paper proposes a two-way authentication protocol MAP-SKBO based on shared private key and bitwise operation in a mobile RFID system. MAP-SKBO applies ultra-lightweight bit replacement and self-combination of cross-bit operation to encrypt the transmitted information, thereby reducing the calculations on the tag side. Each round of the authentication process uses random numbers to maintain the freshness of the communication information and then updates the shared private key's freshness after the authentication process. During the communication process, the authenticity of the side that sends the message is verified first; then, the response information is verified to realize authentication among the three parties of the tag, the reader and the database.

The first section of this article provides an introduction to describe the limitations of the traditional RFID authentication protocol and the security flaws in mobile RFID systems, leading to the focus of this paper. The second section introduces the authentication protocol proposed in recent years for mobile RFID systems. The third section introduces the mathematical knowledge and defines the computational symbols required in the MAP-

SKBO design process. The fourth section establishes a security model for the authentication protocol applicable to mobile RFID systems and gives an abstract description of the authentication protocol. The fifth section systematically describes the MAP-SKBO design steps. The sixth section analyzes the security of MAP-SKBO with respect to identity authentication, desynchronization attacks, track attacks and replay attacks. The seventh section adopts the formal logic of GNY to perform rigorous mathematical reasoning for MAP-SKBO. The eighth section analyzes the performance of MAP-SKBO in terms of the computational complexity, storage capacity, etc. of the tag, the reader and the database. The ninth section summarizes the full text and provides directions for future research.

2 Related Research Works

Reference [15] proposes an RFID one-way authentication protocol based on PRF, but the analysis finds that the protocol cannot completely resist denial-of-service (DoS) attacks. If an attacker constantly sends a message c to the tag, the tag continually updates the value of its own counter ctr, causing the reader to spend more time traversing the query. Reference [13] proposes a PFP protocol based on a hash function and pseudo-random generator, but the protocol has some drawbacks. If the internal state chain length w is too small, it cannot resist DoS attacks. If the value of w is too large, the reader will pay a large cost to calculate the internal state chain. The authentication scheme proposed in Reference [25] cannot resist desynchronization attacks. The attacker makes the shared private key stored between the tag and the reader inconsistent by means of replay attack and information forging and then destroys the subsequent authentication between the tag and the reader. The authentication scheme proposed in Reference [5] cannot resist active attack. An attacker can gradually derive the private key stored in the tag by continuously interrogating the tag and analyzing the reply information of the tag. Although the scheme proposed in Reference [11] is resistant to common attacks, the tag side needs to generate five random numbers during the authentication process, which makes the computational complexity of the tag excessive. All the above authentication protocols have a common feature that they are designed for traditional RFID systems. However, they are not applicable to mobile RFID systems.

Reference [23] proposes a one-way mobile authentication protocol but found that the agreement cannot resist man-in-the-middle attacks and replay attacks. A mobile authentication protocol based on elliptic curve is proposed in Reference [8]; however, this scheme cannot ensure the privacy of the reader and the computational complexity of the tag is also high. An ultra-lightweight mobile authentication protocol is proposed in Reference [2]. The analysis shows that the protocol cannot resist replay attacks on the tag. The mobile authentication protocol proposed in Reference [14], which is based on a hash function, cannot prevent tag forgery, man-in-the-middle attacks and replay attacks. Reference [7] proposes a tripartite authentication of the mobile protocol. However, the protocol burdens the database with a heavy workload and also cannot resist DoS attacks. The Edwards curve-based mobile protocol proposed in Reference [26] does not implement authentication from the reader side to the tag in the authentication process, which makes the protocol vulnerable to impersonation attack. Reference [16] proposes a mobile authentication protocol based on a shared private key. However, the protocol, in which the tag authenticates the reader, does not achieve full authentication, which makes the protocol vulnerable to impersonation attacks.

Considering the shortcomings of many existing schemes, this paper proposes a mobile authentication system, MAP-SKBO, based on shared private key and bitwise operation for a mobile RFID system. The MAP-SKBO authentication process involves first verifying the authenticity of the message source and then conducting follow-up operations, which can resist the deliberate destruction of the attacker. Encrypting information by bitwise operation enables MAP-SKBO to achieve an ultra-lightweight level, which can effectively reduce the computational load of the RFID system. From the perspective of safety and performance, MAP-SKBO is suitable for low-cost mobile RFID systems.

3 Related Knowledge

To facilitate the description, we use “Sac(Z)” to represent self-assembling cross-bit operation. Let X, Y, and Zbe three binary numbers of l bits, $X = x_1x_2 \cdots x_L$, $Y = y_1y_2 \cdots y_L$, and $Z = z_1z_2 \cdots z_L$, where $X \in \{0,1\}^l$, $Y \in \{0,1\}^l$, $Z \in \{0,1\}^l$. X undergoes bitwise XOR with Y to obtain Z. $Sac(Z)$ is a new binary number W with l bits formed by the combination of the high and low bits of Z, that is, $Sac(Z) = z_1z_L/2 + 1z_2z_L/2 + 2 \cdots z_L/2z_L$.

The self-assembling cross-bit operation can be implemented in the tag and reader as described below. Introduce two pointers, one for P_1 and one for P_2, where P_1 points to the head of binary number Z and P_2 points to the end of binary number Z. When P_1 traverses from the head of Z, P_2 simultaneously starts traversing from the end of Z. The numbers traversed by pointer P_1 are sequentially placed in the odd bits of the new binary number W, and the numbers traversed by pointer P_2 are sequentially placed in the even bits of the new binary number W. Finally, through combination we can obtain the new binary number W, that is, $Sac(Z)$ [6].

The self-assembling cross-bit operation requires only shift and bitwise OR operation and the final combination, thereby reducing system throughput and storage capacity to achieve an ultra-lightweight level. Different orders of pointer assignment will produce different values, thereby increasing the difficulty of cracking. For example, if $l = 8$, $X = 11011001$ and $Y = 01100101$, then $X \oplus Y = Z$ and

$Sac(Z) = 11011010$. The specific process is shown in Figure 1.

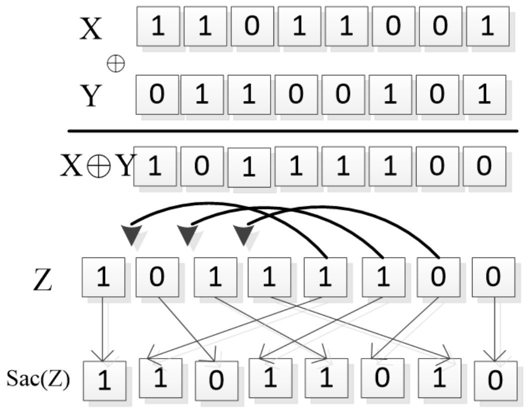


Figure 1: Self-assembling cross-bit operation flow chart

4 RFID Security Model

The goal of a mobile RFID authentication protocol is not only to ensure the safety of the tag's private information but also to ensure that both the tag and the mobile reader cannot be tracked. This paper uses MySQL query mode to model the attack ability of Attacker-A while establishing the non-traceable model of a mobile RFID system. T represents the tag, R represents the reader, DB represents the database, and P represents the protocol that the tag, the mobile reader, the database are involved in. The participants in the protocol can initiate several instances of P, where M_T represents the instance initiated by the tag, M_R represents the instance initiated by the mobile reader, and M_{DB} represents the instance initiated by the database. Attacker-A can perform the following query operation:

1) Execute (M_T, M_R, M_{DB}, n) Query Operation: This query operation describes an instance where Attacker-A executes protocol P. Simultaneously, all the two-way transferring communication information between tag T and mobile reader R or between mobile reader R and database DB is acquired during the n-th round of communication. The query operation modeling is equivalent to a static attack.

2) Send $(M_T, message, n)$ Query Operation: In this query operation, Attacker-A sends a message to tag T during the n-th round of communication, and the query operation is modeled as a dynamic attack. Through this query operation, tag T will return a value as a response message based on the protocol and the stored data.

3) Send $(M_R, message, n)$ Query Operation: In this query operation, Attacker-A sends a message to mobile reader R during the n-th round of communication, and the query operation modeling is equivalent to a dynamic attack. Through this inquiry operation, mobile reader R will return a value as a response message according to the protocol and the stored data.

4) Corrupt (M_T) Query Operation: This query operation describes the ability of Attacker-A to bribe tag T so that T will actively leak the private information stored by itself. This query operation modeling is equivalent to a dynamic attack.

5) Corrupt (M_R) Query Operation: The query operation describes the ability of Attacker-A to buy a mobile reader R so that R will actively leak the private information stored by itself. The query operation modeling corresponds to a dynamic attack.

5 Design of MAP-SKBO

5.1 Initinal Conditions and Symbols

Before MAP-SKBO is executed, all entities in the mobile RFID system must initialize the memory unit as follows.

The tag stores Key_L, Key_R and ID_T, forming the triple (Key_L, Key_R, ID_T). The mobile reader stores Key_L, Key_R and ID_R to form the triple (Key_L, Key_R, ID_R). The database stores Key_L, Key_R, ID_T and ID_R to form the four-tuple $(Key\_L, Key\_R, ID_T, ID_R)$. The definitions and descriptions of the symbols used in MAP-SKBO are shown in Table 1.

Table 1: Symbol definitions

Symbol	Description
T	The tag
R	The mobile reader
DB	The database
ID_T	Identifier ID of the tag
ID_R	Identifier ID of the reader
Key	The private key shared among the reader, the tag and the database
Key_L	The left half of the shared private key
Key_R	The right half of the shared private key
Key_old	The shared private key of the last round of authentication
Key_new	The shared private key of the current round of authentication
r_T	The random number generated by the tag
r_R	The random number generated by the reader
r_{DB}	The random number generated by the database
$\oplus$	Bitwise XOR operation
&	Bitwise AND operation
$Sac(X)$	Self-combined cross-bit operation

5.2 MAP-SKBO Authentication Process

The MAP-SKBO authentication process is shown in Figure 2. The following gives a description of the specific meanings of formulas M0 to M12 in Figure 2, as shown in Table 2. Then, in combination with Figure 2, a description of the specific steps of the MAP-SKBO authentication is given.

Table 2: Formula descriptions

Symbol	Description
M0	$r_R \oplus Key\_R$
M1	$r_R \oplus Key\_L$
M2	$r_R \oplus r_T$
M3	$r_T \oplus ID_T$
M4	$Sac(r_T \& Key\_L, r_T \oplus Key\_R)$
M5	$Sac((r_T \& r_R) \oplus (r_T \& Key\_L))$
M6	$r_R \oplus ID_R$
M7	$Sac(r_R \& ID_R \& Key\_L, r_R \& ID_R \& Key\_R)$
M8	$r_R \oplus r_{DB} \oplus ID_R$
M9	$r_T \oplus r_{DB} \oplus ID_T$
M10	$Sac(r_R, r_{DB})$
M11	$Sac(r_T, r_{DB})$
M12	$r_R \& r_{DB}$

The MAP-SKBO authentication process is shown in Figure 2.

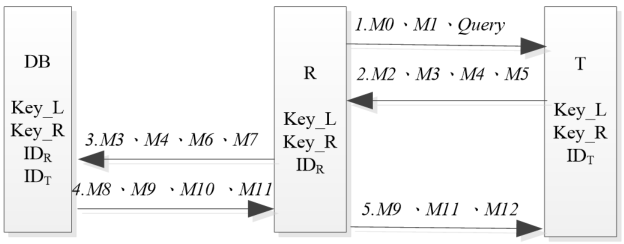


Figure 2: MAP-SKBO authentication flow chart

The detailed steps of the MAP-SKBO authentication process are as follows.

Step 1. The mobile reader generates a random number $r_R \in 0, 1^l$; then, the reader calculates the values of M0 and M1 and sends M0, M1 and the authentication request command query to the tag.

Step 2. After the tag receives the information, the values of $M0 \oplus Key\_R$ and $M1 \oplus Key\_L$ are calculated and compared to determine if they are the same.

If they are equal, the tag verifies that the mobile reader has passed and proceeds to step three; otherwise, the tag indicates that the mobile reader is forged and MAP-SKBO terminates immediately.

Step 3. The tag calculates random number r_R and generates a random number $r_T \in 0, 1^l$. Then, the tag calculates M2, M3, M4, and M5 and transmits (M2, M3, M4, M5) to the mobile reader.

Step 4. After the mobile reader receives the message, it first calculates the values of $M2 \oplus r_R$ and M5' and determines whether the values of M5' and M5 are equal. If they are equal, then the mobile reader verifies that the tag has passed and proceeds to Step 5; otherwise, the reader indicates that the tag is forged and MAP-SKBO terminates immediately, where $M5' = Sac(((M2 \oplus r_R) \& r_R) \oplus ((M2 \oplus r_R) \& Key\_L))$.

Step 5. The mobile reader calculates random number r_T and the values of M6 and M7 and transmits (M3, M4, M6 and M7) to the database.

Step 6. The database authenticates the mobile reader.

1) After the database receives the information, it first calculates the values of $M6 \oplus ID_R$ and M7‘ and determines whether the values of M7‘ and M7 are equal. If they are equal, the database verifies that the mobile reader has passed and proceeds to Step 3; otherwise, the process proceeds to Step 2, where $M7‘ = Sac((M6 \oplus ID_R) \& ID_R \& Key\_L, (M6 \oplus ID_R) \& ID_R \& Key\_R)$.

2) The database uses Key_old instead of Key_new to perform the calculation in Step 1. If they are equal, the database verifies that the reader has passed and goes to Step 3; otherwise, the mobile reader is forged and MAP-SKBO terminates immediately.

3) The database calculates random number r_R and goes to the seventh step.

Step 7. The database authenticates the tag.

1) After the database verifies that the mobile reader has passed, it calculates the values of $M3 \oplus ID_T$ and M4‘ and determines whether M4 ’ and M4 are equal. If they are equal, the database verifies that the tag has passed and Step 3 is performed; otherwise, Step 2 is performed, where $M4‘ = Sac((M3 \oplus ID_T) \& Key\_L, (M3 \oplus ID_T) \oplus Key\_R)$.

2) The database uses Key_old instead of Key_new to perform the calculation in Step 1. If they are equal, then the database verifies that the tag has passed and goes to Step 3; otherwise, the tag is forged and MAP-SKBO terminates immediately.

3) The database calculates random number r_T and proceeds to the eighth step.

Step 8. The database generates a random number $r_{DB} \in 0, 1^l$, calculates the values of M8, M9, M10 and M11, starts to update the information of the shared private key, that is, Key_old=Key and Key=Key_new, and transmits (M8, M9, M10, M11) to the mobile reader, where $Key\_new = Sac((r_{DB} \oplus r_T \oplus r_R) \oplus (r_{DB} \& r_T \& r_R))$.

Step 9. After the mobile reader receives the message, it calculates the values of $M8 \oplus r_R \oplus ID_R$ and M10‘ and compares the values of M10‘ with M10. If they are equal, the mobile reader verifies that the database is authentic and proceeds to Step 10; otherwise, it indicates that the database is forged and MAP-SKBO terminates immediately, where $M10‘ = Sac(r_R, (M8 \oplus r_R \oplus ID_R))$.

Step 10. The mobile reader calculates the value of M12, updates the information of the shared private key, that is, Key_old=Key and Key=Key_new, and transmits (M9, M11, M12) to the tag, where $Key\_new = Sac((r_{DB} \oplus r_T \oplus r_R) \oplus (r_{DB} \& r_T \& r_R))$.

Step 11. The tag authenticates the mobile reader.

1) After the tag receives the information, the tag calculates the values of $M9 \oplus r_T \oplus ID_T$ and M12‘ and compares M12‘ with M12 for equality. If they are equal, then the tag verifies the mobile reader and proceeds to Step 2; otherwise, it indicates that the mobile reader is forged and MAP-SKBO terminates immediately, where $M12‘ = r_R \& (M9 \oplus r_T \oplus ID_T)$. The tag authenticates the database.
2) The tag calculates the value of M11‘ and determines whether M11‘ and M11 are equal. If they are equal, then the tag verifies the database and proceeds to Step 3; otherwise, it indicates that the database is forged and MAP-SKBO immediately terminates, where $M11‘ = Sac(r_T, (M9 \oplus r_T \oplus ID_T))$.
3) The tag starts to update the information of the shared private key, that is, $Key = Sac((r_{DB} \oplus r_T \oplus r_R) \oplus (r_{DB} \& r_T \& r_R))$. The authentication process among the tag, the mobile reader and the database terminates.

6 Safety

1) Replay Attack: The tag generates a random number r_T in each authentication process, and the authentication message (M2, M3, M4, M5) contains r_T in each calculation. If the attacker adopts the old message, the tag will use the newly generated random number r_T when verifying the authentication message (M9, M11, M12). This will cause the tag to fail when validating the mobile reader and the database, and the MAP-SKBO will terminate immediately, preventing attackers from completing the follow-up authentication process. Therefore, MAP-SKBO can resist replay attacks.

2) Asynchronous Attack: In an asynchronous attack, during the authentication process, due to the deliberate sabotage of the attacker, the shared private key of the mobile reader (or the database) and the tag becomes asynchronous. An asynchronous attack is also known as a desynchronization attack. To resist asynchronous attacks, MAP-SKBO stores the shared private key, Key_old, used in the previous authentication process to recover synchronization with the tag. When the database authenticates the tag and the mobile reader through (M3, M4, M6, M7), it first calls Key_new. If the verification fails, Key_old is called to resist the attacker's desynchronization attack. Therefore, the existence of Key_new and Key_old makes MAP-SKBO resistant to asynchronous attacks.

3) Man-in-the-middle Attack: Replacing the message and tampering with the news are the most common forms of man-in-the-middle attacks. According to the protocol application scenario, an attacker can obtain all communication message sets of the tag, the mobile reader and the database among all three MS=M0, M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12. Because the above message is encrypted, even if the attacker acquires the above message, he cannot derive any useful information from it. Although an attacker can modify or tamper with one of the messages, MAP-SKBO will verify the message at each step and find that the message has been tampered with. Meanwhile, the calculation of the above messages is associated with random numbers r_R, r_T and r_{DB}, and the random numbers are randomly generated and unpredictable, making it harder for an attacker to modify the message. Thus, MAP-SKBO can resist man-in-the-middle attacks.

4) Forward Security: Due to the database storing the shared private key of the previous round of the authentication process, here only the forward security of the tag is discussed. If the attacker wants to obtain the current shared private key value of the tag, the attacker needs to decrypt the previous authentication message from the last received message. However, the attacker cannot succeed for the following reasons. First, the attacker cannot crack the message encrypted by the bitwise operation because at least two quantities in the ciphertext are unknown to the attacker. Second, after the authentication, MAP-SKBO immediately updates the value of the shared private key, and there is no correlation between the initial and updated values. Furthermore, the calculation of the value is dependent on three random numbers r_R, r_T and r_{DB}, which are impossible for the attacker to obtain. Therefore, MAP-SKBO can ensure the forward safety of the tag.

5) Bidirectional Authentication: In the mobile RFID system, because the communication between the

tag and the reader or between the reader and the database is performed through a wireless channel, which is not secure, each message transmission requires authentication.

The tag authenticates the reader. The reader sends the message to the tag for the first time, and the tag completes the first authentication of the reader in the second step. In the tenth step, the reader transmits a message to the tag a second time, and the tag completes the second authentication of the reader in the eleventh step. The reader authenticates the tag. The tag sends a message to the reader in the third step, and the reader completes the authenticity verification of the tag in the fourth step.

The reader authenticates the database. The database transmits the message to the reader in the eighth step, and the reader completes the authenticity verification of the database in the ninth step.

The database authenticates the reader and the tag. To ensure resistance to desynchronized attacks, the database simultaneously stores the values of Key_new and Key_old. In the fifth step, after the reader sends the message to the database, the database authenticates the reader in the sixth step and authenticates the tag in the seventh step. Through these processes, the tag, the reader, and the database can achieve mutual authentication, so MAP-SKBO can achieve bidirectional authentication.

7 GNY Logical Formal Proof

In this paper, the formal analysis and proof of the WKGA-BO protocol are performed by using GNY [1] formal logic analysis.

1) Formal Description of the Protocol: The following conventions are used to simplify the application of the GNY formal logic language description to the MAP-SKBO. R represents the mobile reader, T represents the tag, and DB represents the database. The flow of the MAP-SKBO protocol is as follows:

Msg1: $R \rightarrow T : \{M0, M1, Query\}$

Msg2: $T \rightarrow R : \{M2, M3, M4, M5\}$

Msg3: $R \rightarrow DB : \{M3, M4, M6, M7\}$

Msg4: $DB \rightarrow R : \{M8, M9, M10, M11\}$

Msg5: $R \rightarrow T : \{M9, M11, M12\}$

After using GNY formal logic language to standardize the above protocol, the process can be described as follows:

Msg1: $T < *\{M0, M1, Query\}$

Msg2: $R < *\{M2, M3, M4, M5\}$

Msg3: $DB < *\{M3, M4, M6, M7\}$

Msg4: $R < *\{M8, M9, M10, M11\}$

Msg5: $T < *\{M9, M11, M12\}$

2) The Initalization Assumption of the Protocol: The MAP-SKBO protocol assumptions are as follows: the combination of R, DB, T represent the body, where R represents the mobile reader, T represents the tag, and DB represents the database.

Sup1: $T \ni (Key\_R, Key\_L, ID_T r_T)$

Sup2: $R \ni (Key\_R, Key\_L, ID_R, r_R)$

Sup3: $DB \ni (Key\_R, Key\_L, ID_R, ID_T, r_{DB})$

Sup4: $R| \equiv \#(r_R, r_T, r_{DB})$

Sup5: $T| \equiv \#(r_R, r_T, r_{DB})$

Sup6: $DB| \equiv \#(r_R, r_T, r_{DB})$

Sup7: $T |\equiv R \overset{Key\_R, Key\_L}{\longleftrightarrow} T$

Sup8: $R |\equiv T \overset{Key\_R, Key\_L}{\longleftrightarrow} R$

Sup9: $DB |\equiv R \overset{Key\_R, Key\_L, ID_R}{\longleftrightarrow} DB$

Sup10: $R |\equiv DB \overset{Key\_R, Key\_L, ID_R}{\longleftrightarrow} R$

Sup11: $DB |\equiv T \overset{Key\_R, Key\_L, ID_T}{\longleftrightarrow} DB$

Sup12: $T |\equiv DB \overset{Key\_R, Key\_L, ID_T}{\longleftrightarrow} T$

3) The Proof Target of the Protocol: There are five main proof targets of the MAP-SKBO protocol, namely, mutual trust of the freshness of the information exchanged among the tag, the mobile reader and the database. The proof formulas of the target are as follow:

Goal1: $T |\equiv R| \sim \#(M0, M1)$

Goal2: $R |\equiv T \sim \#(M2, M3, M4, M5)$

Goal3: $DB |\equiv R| \sim \#(M3, M4, M6, M7)$

Goal4: $R |\equiv DB| \sim \#(M8, M9, M10, M11)$

Goal5: $T |\equiv R| \sim \#(M9, M11, M12)$

4) The Protocol Proving Process: The proof of the MAP-SKBO protocol is based on the initialization hypothesis, which proves that the process follows the rules of logical reasoning in Reference [1], and the notification rules, fresh rules, procession rule and the rules of message interpretation follow the GNY logic inference rules in Reference [1], which are, respectively, represented as T, P, F and I.

Because the protocol proves that the processes of proving Goal 2: $R |\equiv T| \sim \#(M2, M3, M4, M5)$, Goal 3: $DB |\equiv R| \sim \#(M3, M4, M6, M7)$, Goal 4: $R |\equiv DB \sim \#(M8, M9, M10, M11)$, Goal 5: $T |\equiv R| \sim \#(M9, M11, M12)$ are similar to the proof process of Goal 1: $T |\equiv R| \sim \#(M0, M1)$, this section proves only Goal 1: $T |\equiv R| \sim \#(M0, M1)$ as an example. The proof process is given below.

Proof.
∵ Rule P_1: $\frac{P<X}{P\ni X}$ and Msg 1: $T < *\{M0, M1\}$,
∴ $T \ni \{M0, M1\}$.

$\because$ Rule: F1: $\frac{P \mid\equiv (X)}{P \mid\equiv (x,y), P \mid\equiv \# F(X)}$ and Sup 4: $R \mid\equiv \#(r_R, r_T, r_{DB})$,

$\therefore T = \#\{M0, M1\}$.

$\because$ Rule P_2: Sup 1: $T \ni (Key\_R, Key\_L, ID_T, r_T)$ and Sup 2: $R \ni (Key\_R, Key\_L, ID_R, r_R)$,

$\therefore T \ni \{M0, M1\}$.

$\because$ Rule F10: $\frac{P \mid\equiv (X), P \ni X}{P \mid\equiv \#(H(X))}$ and the formula derived: $T = \#\{M0, M1\}$, $T \ni \{M0, M1\}$,

$\therefore T \mid\equiv \#\{M0, M1\}$.

$\because$ Rule I3: $\frac{P < H(X, <S>) >, P \ni (X,S), P \mid\equiv P \longleftrightarrow Q, P \mid\equiv \#(X,S)}{P \mid\equiv Q | \sim (X,S), P \mid\equiv Q \sim H(X, <S>)}$

Then, $\because$ Sup 7: $T \mid\equiv R \overset{Key\_R, Key\_L}{\longleftrightarrow} T$, Sup 8: $R \mid\equiv T \overset{Key\_R, Key\_L}{\longleftrightarrow} R$ and Msg 1: $T < *\{M0, M1\}$,

$\therefore T \models R \sim \{M0, M1\}$.

$\because$ freshness definition and its derivation: $T = \#\{M0, M1\}, T \models R \sim \{M0, M1\}$,

$\therefore$ Goal 1: $T \mid\equiv R | \sim \#(M0, M1)$ has been proved. □

8 Performance Analysis

A mobile RFID system includes a tag, mobile reader, and database. Because the latter two have strong computing power and large storage capacity, they have little effect on the performance of the protocol. Therefore, the computation power and storage capacity of only the tag are analyzed. The performance analysis of the RFID authentication protocol is conducted from four main perspectives: the computational load of the tag, the storage of the tag, the number of conversations, and protocol traffic. Table 3 shows the performance comparison results of MAP-SKBO and other authentication protocols.

In Table 3, H represents a hash function operation, M represents scalar multiplication, S represents a random number calculation, and Sac represents a combination of self-cross-bit operation. As described in the third section of the article, H, M and S are lightweight operations, whereas Sac is ultra-lightweight operations. That is, the former require much more computation than the latter. Because the bitwise XOR operation and bitwise AND operation require less computation, their computation is ignored in the performance analysis. The lengths of the shared private key Key, the identifier ID and the result of each operation (i.e., H, M, S, Sac) are set to l.

Table 3: Performance comparison of authentication protocols

Reference	Computational Load	Storage Capacity	Protocol Traffic
Reference [23]	H	3l	10l
Reference [8]	4M+ H	2l	22l
Reference [14]	3H	3l	14l
Reference [7]	2H	2l	13l
Reference [26]	3M+2H	6l	17l
Reference [16]	3S+H	2l	12l
This protocol	S+4Sac	2l	17l

1) Storage and Computation Load of the Tag: In this paper, the MAP-SKBO tag needs to store only two values, the shared privates key Key and the identifier of the tag ID_T. According to the previous convention, the required storage capacity of the tag is 2l. Compared with the References [14, 23, 26], the storage capacity of the tag of this protocol is reduced; compared with the References [7, 8, 16], the storage capacity of the tag in this paper is equivalent.

   In terms of the computation load of the tag, the bitwise XOR operation and bitwise AND operation have small computational cost, so their costs are not considered. Therefore, the computational cost is much less than that of other studies. In this paper, we do not encrypt the message using a hash function or scalar multiplication, which are computationally intensive. Instead, we encrypt the message by ultra-lightweight bitwise operation to reduce the computational load. In summary, the protocol in this paper has some improvements compared to other protocols in terms of the storage and computational load of the tag.

2) Communication and the Number of Conversations: The communication traffic of this protocol is slightly larger than those in References [7,14,16,23], but there are some security risks in the previous protocols. The proposed protocol overcomes the defects of previous protocols. This protocol is equivalent to those in References [8, 26] in terms of communication traffic and solves their security problems.

   The number of conversations in most of the protocols is five. The proposed protocol has no advantage in terms of the number of conversations. In summary, this protocol achieves little improvement with respect to overall communication traffic and the number of conversations but solves the security flaws in other protocols. Therefore, this protocol still has some practical value.

9 Conclusion

This paper describes the differences between traditional RFID systems and mobile RFID systems and notes that the traditional RFID system authentication protocol cannot be applied to mobile RFID systems. Therefore, an MAP-SKBO authentication protocol is proposed for mobile RFID systems. The paper expounds the defects and deficiencies in some current authentication protocols applicable to mobile RFID systems and then proposes an improved authentication scheme. The proposed MAP-

SKBO protocol abandons the hash function encryption method and instead uses bitwise operations to encrypt the information, making the protocol achieve an ultra-lightweight level. The use of bit replacement operations and self-combined cross-bit operations increases the difficulty of the attacker in cracking the protocol. A security analysis and performance analysis illustrate the security and advantages of the protocol. GNY logic formally proves the correctness of MAP-SKBO; MAP-SKBO is not only suitable for mobile RFID systems but also for traditional RFID systems. Future potential research directions include the following: Optimize the MAP-SKBO protocol to reasonably reduce the traffic of the whole communication; and implement a MAP-SKBO mobile RFID system prototype to determine the total number of gates, the time needed to achieve complete communication and other issues to achieve the combination of theory and practice.

Acknowledgments

This paper is supported by the 2021 Guangdong Education Science Planning Project (Special Project for Higher Education) (China) (2021GXJK); the 2021 School-level Quality Engineering Project of Guangzhong College Of Technology and Busincess "Flipped Teaching Practice of Network Security Based on SPOC" (ZL20211140); and the 2021 Research on Software-defined Key Technology of Industrial Control Real-time Network (China) (2021FNA0210).

References

[1] M. Burrows, M, Abadi, R. M. Needham, "A logic of authentication," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, pp. 233-271, 1989.

[2] Z. Cao and O. Markowitch, "Analysis of Shim's attacks against some certificateless signature schemes," *International Journal of Network Security*, vol. 23, no. 3, pp. pp. 545-548, 2021.

[3] Y. C. Chen, W. L. Wang, M. S. Hwang, "RFID authentication protocol for anti-counterfeiting and privacy protection", in *The 9th International Conference on Advanced Communication Technology*, pp. 255–259, 2007.

[4] S. Y. Chiou, "An efficient RFID authentication protocol using dynamic identity," *International Journal of Network Security*, vol. 21, no. 5, pp. 728-734, 2019.

[5] S. Y. Chiou, W. T. Ko, E. H. Lu, "A secure ECC-based mobile RFID mutual authentication protocol and its application," *International Journal of Network Security*, vol. 20, no. 2, pp. 396-402, 2018.

[6] J. F. Chong, Z. Zhuo. , "Constructions of balanced quaternary sequences of even length," *International Journal of Network Security*, vol. 22, no. 6, pp. 911-915, 2020.

[7] Y. P. Duan, "Lightweight RFID group tag generation protocol," *Control Engineering of China*, vol. 27, no. 4, pp. 751-757, 2020.

[8] K. Fan, W. Jiang, H. Li, *et al.*, "Lightweight RFID protocol for medical privacy protection in IoT," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1656-1665, 2018.

[9] M. S. Hwang, E. F. Cahyadi, S. F. Chiou, C. Y. Yang. , "Reviews and analyses the privacy-protection system for multi-server," *Journal of Physics: Conference Series*, vol. 1237, pp. 022091, June 2019.

[10] M. S. Hwang, C. H. Wei, C. Y. Lee, "Privacy and security requirements for RFID applications", *Journal of Computers*, vol. 20, no. 3, pp. 55–60, Oct. 2009.

[11] K. Li and R. Huang, "A CKKS-based privacy preserving extreme learning machine," *International Journal of Network Security*, vol. 24, no. 1, pp. 166-175, 2022.

[12] D. W. Liu, J. Ling, "An improved RFID authentication protocol with backward privacy," *Computer Science*, vol. 43, no. 8, pp. 128-130, 2016.

[13] L. H. Liu and X. Y. Cao, "A note on one privacy-preserving centralized dynamic spectrum access system," *International Journal of Network Security*, vol. 23, no. 6, pp. 1074-1077, 2021.

[14] G. F. Shen, S. M. Gu, and D. W. Liu, "An anti-counterfeit complete RFID tag grouping proof generation protocol," *International Journal of Network Security*, vol. 21, no. 6, pp. 889-896, 2019.

[15] S. Sundaresan, R. Doss, S. Piramuthu, *et al.*, "A secure search protocol for low cost passive RFID tags," *Computer Networks*, vol. 122, pp. 70-82, 2017.

[16] J. Q. Wang, Y. F. Zhang, D. W. Liu, "Provable secure for the ultra-lightweight RFID tag ownership transfer protocol in the context of IoT commerce," *International Journal of Network Security*, vol. 22, no. 1, pp. 12-23, 2020.

[17] C. H. Wei, M. S. Hwang, A. Y. H. Chin, "A mutual authentication protocol for RFID", *IEEE IT Professional*, vol. 13, no. 2, pp. 20–24, 2011.

[18] C. H. Wei, M. S. Hwang, Augustin Y. H. Chin, "A secure privacy and authentication protocol for passive RFID tags," *International Journal of Mobile Communications*, vol. 15, no. 3, pp. 266–277, 2017.

[19] C. H. Wei, M. S. Hwang, Augustin Y. H. Chin, "An authentication protocol for low-cost RFID tags", *International Journal of Mobile Communications*, vol. 9, no. 2, pp. 208–223, 2011.

[20] C. H. Wei, M. S. Hwang, Augustin Y. H. Chin, "An improved authentication protocol for mobile agent device in RFID," *International Journal of Mobile Communications*, vol. 10, no. 5, pp. 508–520, 2012.

[21] Y. Wei, J. Chen, "Tripartite authentication protocol RFID/NFC based on ECC," *International Journal of Network Security*, vol. 22, no. 4, pp. 664-671, 2020.

[22] H. Xia and W. Yang, "ID-authentication based on PTPM and certificateless public-key cryptography in cloud," *International Journal of Network Security*, vol. 23, no. 6, pp. 952-961, 2021.

[23] R. Xie, B. Y. Jian, D. W. Liu, "An improved ownership transfer for RFID protocol," *International Journal of Network Security*, vol. 20, no. 1, pp. 149-156, 2018.

[24] R. Xie, J. Ling, D. W. Liu, "A wireless key generation algorithm for RFID system based on bit operation," *International Journal of Network Security*, vol. 20, no. 5, pp. 938-950, 2018.

[25] Y. Xue, "Research on network security intrusion detection with an extreme learning machine algorithm," *International Journal of Network Security*, vol. 24, no. 1, pp. 29-35, 2022.

[26] X. Zhao, "Attack-defense game model: research on dynamic defense mechanism of network security," *International Journal of Network Security*, vol. 22, no. 6, pp. 1037-1042, 2020.

Biography

Dao-wei Liu received a master's degree in School of Computers from Guangdong University of Technology (China) in June 2016. His current research interest fields include information security.

Sheng-hua Xu received a master's degree in School of Computers from Guangdong University of Technology (China) in June 2009. He is now a lecturer, working in Guangdong Polytechnic Normal University. At present, his research interests mainly include information security.

Wen-tao Zuo received a master's degree in School of Computers from South China Agricultural University (China) in June 2010. He is now a lecturer, working in Guangzhou College of Technology and Business. His current research interest fields include information security.

Digital Image Copyright Protection Method Based on Blockchain and Perceptual Hashing

Qiu-Yu Zhang and Guo-Rui Wu
(Corresponding author: Qiu-Yu Zhang)

School of Computer and communication, Lanzhou University of Technology
No. 287, Lan-Gong-Ping Road, Lanzhou 730050, China
Email: zhangqylz@163.com, vigdis_r@163.com

(Received May 7, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 10, 2022)

Abstract

To enhance the transparency and credibility of the copyright information of digital images of grotto murals and to solve the problems of low sampling rate in the frequency domain, easy tampering, and unclear ownership of copyright information, a digital image copyright protection method based on blockchain and perceptual hashing was proposed by using the features of blockchain, such as tamper resistance, decentralization, and traceability. Firstly, the copyright owner uses the improved perceptual hashing function to convert the mural image into a 256-bit hash value and uploads it to the blockchain along with other copyright information; Then, the intelligent contract detects the infringement of the image by calculating the similarity of the hash value, and the detected copyright information and encrypted image are uploaded to the InterPlanetary File System (IPFS). Finally, the consumer downloads the image and copyright information initiates a copyright transaction, and calls the double smart contract to obtain the key and decryption the image. The experimental results show that the proposed method solves the problem of unclear ownership of copyright owners and consumers and ensures the security of asset transfer in the process of copyright transaction. The improved perceptual hashing function also improves the accuracy of determining similarity in digital image infringement detection.

Keywords: Blockchain; Copyright Protection; Grotto Murals; IPFS; Perceptual Hashing; Smart Contract

1 Introduction

With the rapid development of cloud storage, internet and multimedia technology, multimedia files show an explosive growth, which embodies the characteristics of digitalization of content, digitalization of product form, digitalization of management process and network of communication channels [35]. As a world cultural heritage [34], the dissemination of grotto murals in the form of digital images has become one of the feasible ways to popularize grotto murals. However, due to the downloading and copying of grotto mural images stored in the cloud, it is difficult to ensure the security of digital copyright, resulting in an increasing number of copyright problems. In recent years, the widely used blockchain has the characteristics of decentralization, de-trust, non-tampering and traceability [6, 18]. It can be applied in the process of image copyright protection and copyright transaction to solve the problems of digital copyright information being easily tampered with, high cost and non-traceability.

At present, the centralized copyright management system adopted in the traditional way of copyright protection is inefficient and costly, which cannot effectively solve the dilemma faced by digital copyright in the digital communication environment. Although there are many existing digital rights management (DRM) technologies such as information hiding [31, 33] and encryption, there are still some deficiencies in image copyright protection. Under the network platform, the existing technologies such as [24] digital watermarking have been unable to fully protect the rights of copyright owners [3, 4]. Users can destroy the digital watermark by removing attacks, geometric attacks, encryption attacks and protocol attacks, or eliminate, destroy and tamper with watermark through image cropping and image restoration software [2,11]. For different digital images, the existing encryption methods also have some problems such as loss of encryption and decryption data, weak security, and difficulty in retrieval. Most of the existing copyright protection methods first process the image data by image encryption based on hash function, and then compare it with the image data stored in the database, so as to determine whether the infringement occurs. Traditional cryptographic hash algorithms such as MD5 and SHA-256 are simple to implement, but due to the avalanche effect, even small changes in the image will cause drastic changes in generated hash, so the infringement detection effect of the image after adding noise and rotating operation is not good. Image perceptual hashing has a certain correlation between the original input data and the input data after slight tampering, and has a good tamper detection effect on the image data that has not changed significantly. Therefore, perceptual hashing is used to process the image data and calculate the similarity of generated hash value to determine whether the image is infringing. The blockchain-based digital image copyright protection method realizes automatic infringement detection by writing smart contracts, but there is still the problem of unclear user ownership. The mainstream blockchain is expensive to store data, and the image copyright file is relatively large, which is not suitable for direct storage on the blockchain. IPFS is a point-to-point distributed file storage system, which has the characteristics of distributed storage and openness. It can dynamically expand the storage capacity, thus effectively solving the problem of insufficient storage capacity on the blockchain caused by the large amount of grotto mural image data. However, the openness of IPFS also makes the stored image data face the risk of misappropriation. The security of copyright information and transaction process cannot be guaranteed.

To overcome such drawbacks, this paper adopts the digital images of grotto murals as the research carriers, and presents a digital image copyright protection method based on blockchain and perceptual hashing, which completes the whole process of copyright information registration, consumer information registration, and copyright transaction by writing and calling double smart contract. The summary of contributions of our work is given below:

1) By improving the spatial sampling rate of perceptual hashing frequency domain, the problem of missing image details in feature capture is solved, and the accuracy of determining similarity in digital image infringement detection is improved.

2) To combine the image perceptual hashing and MD5 algorithm, generate the unique corresponding encryption and decryption key related to the image, and encrypting the successfully registered image to be uploaded by chaotic sequence, which effectively solves the data security problem after the image is uploaded to IPFS.

3) The copyright owner and consumers use smart contract to register information respectively, which solves the problem of unclear rights between copyright owner and consumers. Meanwhile, double smart contract is called for triple security verification, which effectively solves the security problem of asset transfer during copyright transactions.

The rest of this paper is organized as follows: Related work is discussed in details in Section 2. Section 3 introduces the proposed system model and specific implementation scheme in details. Section 4 carries out the performance analysis of the proposed scheme, and compared with existing digital image copyright protection scheme. Section 5 gives example simulation results. Finally, Section 6 concludes this paper.

2 Related Works

The DRM is widely used in computer software, audio-on-demand and download, video-on-demand and download, digital library, digital image, mobile payment and other fields [27]. Scholars at home and abroad put forward different DRM schemes from different angles (such as rights sharing DRM [14], privacy protection DRM [7, 8], enterprise DRM [25] and DRM in cloud computing [16, 21]). However, most of the above schemes need centralized license servers and third-party financial platforms to assist in the issuance of licenses and the smooth execution of transactions, while the centralized server is vulnerable to attacks, resulting in services termination, and the rights transaction information and license information are opaque [17].

In recent years, many scholars have tried to apply blockchain to DRM system to solve the problems existing in traditional DRM system. Mehta *et al.* [20] proposed a decentralized peer-to-peer photo sharing marketplace built on top of Ethereum test chain, which effectively avoided the avalanche effect, but the image hashing could not deal with the 90 ° rotation of image. Shi *et al.* [28] proposed a DRM system based on blockchain and SIFT local feature extraction algorithm, which realized automatic similar infringement detection, decentralized storage, tamper-proof and traceability of copyright information. Guo *et al.* [10] proposed a blockchain-enabled DRM system, which includes an entirely new network architecture for sharing and managing multimedia resources of online education on the basis of the combination of the public and private blockchains, as well as three specific smart contract schemes for the realization of the recording of multimedia digital rights, the secure storage and the unmediated verification of digital certificates, respectively. Li *et al.* [15] proposed an image copyright protection system based on the fusion of deep neuron network and blockchain, and design a scalable DNN accelerator and SHA-256 using field-programmable gate array (FPGA). Experimental results show that the whole acceleration system can achieve up to 40x speedup comparing to software implementation on CPU. Dobre *et al.* [5] proposed a digital image copyright protection system based on blockchain, which uses as the picture identifier a joint photographic experts group (JPEG) resistant image signature. Through testing in the process of image compression, it is proved that the image signature extraction algorithm can effectively resist JPEG compression. Wang *et al.* [30] proposed a secure image copyright protection framework combines blockchain and zero-watermark technology, which realize the copyright traceability of the image and solve the problem of lack of trusted third parties. Sultana *et al.* [29] proposed a secure medical image sharing system based on zero trust principles and blockchain, which effectively improves the security of medical images in the complete transmission stage through the audit tracking of blockchain reserved data transmission, but there are still some limitations in network speed. Kr *et* *al.* [12] proposed a social media DRM system based on secret sharing, which effectively realized decentralized social media copyright management. However, when implement on traditional social media could suffer latency and low transaction rate. Abba *et al.* [9] proposed a distributed media transaction framework for DRM, which is based on the digital watermarking and a scalable blockchain model, and built a scalable blockchain model using an overlay network, which solved the scalability and security problems of DRM system. Ren *et al.* [26] proposed a robust zero-watermarking algorithm based on the angular features, and introduced blockchain to solve the problem that zero-watermarking relies on third-party copyright organizations. This framework is robust against common watermark attacks, and realizes the copyright protection for the lossless vector map. Abrar *et al.* [1] proposed to use blockchain and digital watermark for image security authentication. By using SHA-256 encryption algorithm to extract the hash value of the generated watermark, it was stored in the blockchain to realize identity authentication independent of the third-party platform. Kumar *et al.* [13] proposed a secure distributed industrial image and video data security detection system based on IPFS and blockchain, which effectively avoided a single point of failure with the help of blockchain characteristics. Liu *et al.* [19] proposed a blockchain copyright protection system combined with fabric's smart contract, which realized the automatic management of the complete digital rights life cycle of digital copyright. Nan *et al.* [22] proposed a code copyright management system based on blockchain, which verified the originality of code based on abstract syntax tree, and achieved good response speed and storage efficiency. Wang *et al.* [32] proposed an image copyright protection model based on blockchain, which ensured that the image information would not be tampered with through consensus nodes, and added digital watermark to prevent the leakage of image information. Natgunanathan *et al.* [23] proposed a multimedia copyright protection audio watermarking technology based on blockchain, which kept the perceived quality of audio signals to the greatest extent.

In summary, most of the existing DRM framework is divided into three parts: image processing, infringement detection and image storage. Considering the problem that the digital watermark is easy to be tampered with, this paper chooses perceptual hashing instead of digital watermark, and calls smart contract to calculate the similarity of hash value to complete the image infringement detection. Most of the existing smart contracts are single-chain codes, and the ownership is unclear. At the same time, the encryption algorithm of data files uploaded to IPFS and the method of obtaining keys by calling contracts are complex, which is not suitable for image information storage in big data environment. Therefore, this paper presents a digital image copyright protection method combining blockchain, improved perceptual hashing, image encryption and IPFS. The proposed method not only improves the accuracy of similarity determination in image infringement detection, but also encrypts the simple chaotic sequence by using the key related to image hashing, which enhances the security of grotto mural image data stored in IPFS, and enhances the security of copyright transaction process through double contract call and triple verification.

3 The Proposed Scheme

3.1 System Model

Figure 1 shows the system model of digital image copyright protection method based on blockchain and perceptual hashing. This model is divided into three different parts: blockchain, user and IPFS, and realizes the functions of copyright information registration, consumer information registration and copyright transaction through smart contracts.

As shown in Figure 1, the users in the system model are divided into copyright owner and consumers. The copyright owner uploads the digital images of grotto murals and related copyright information to the blockchain and IPFS, using the perceptual hashing generated by the images for infringement detection, and encrypting the images with the relevance key. Anyone who obtain grotto mural images and corresponding copyright information through legal channels as a consumer. They can download encrypted mural images and part of copyright information from IPFS as needed, which uploaded by copyright owner, and calls the double smart contract to obtain the decryption key according to the obtained copyright information. If the image file is not invalid, consumers can get the decryption key after completing asset evaluation and payment.

3.2 Algorithm Design

3.2.1 Improved Perceptual Hashing Algorithm

In order to tolerate the deformation of the image, the traditional perceptual hashing function only selects the low-frequency part of the image when it is created, which leads to the lack of image details during feature capture, reduces the accuracy of determining similarity in digital image infringement detection, and makes the result of infringement detection unsatisfactory. After cropping, noise addition, blur, sharpen, rotation and re-size similar but different images, we compared with the accuracy in the realization of infringement detection of three different hashing algorithms, Average hash (Ahash), Difference hash (Dhash) and Perception hash (Phash), and it is found that perceptual hashing algorithm is stable for the similarity measurement of images after different degrees of noise processing. Combining with the characteristics of digital images of grotto murals, the perceptual hashing algorithm is improved, and the specific processing steps are as follows:

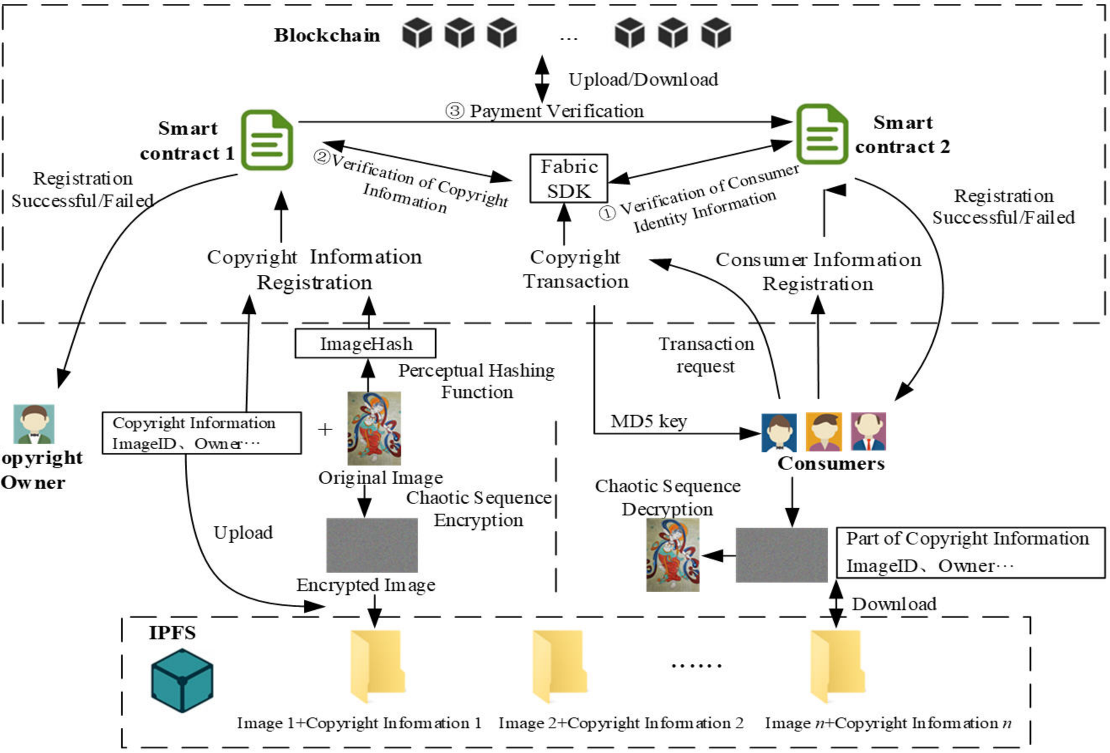

Blockchain
...
Upload/Download
③ Payment Verification
Smart contract 1
Smart contract 2
②Verification of Copyright Information
Fabric SDK
① Verification of Consumer Identity Information
Registration Successful/Failed
Registration Successful/Failed
Copyright Information Registration
Copyright Transaction
Consumer Information Registration
ImageHash
Perceptual Hashing Function
Transaction request
MD5 key
Copyright Owner
Copyright Information ImageID、Owner···
+
Original Image
Chaotic Sequence Encryption
Consumers
Chaotic Sequence Decryption
Part of Copyright Information ImageID、Owner···
Upload
Encrypted Image
Download
IPFS
......
Image 1+Copyright Information 1
Image 2+Copyright Information 2
Image n+Copyright Information n

Figure 1: The system model

Step 1. Reduces the image to get a three-dimensional array of $32 \times 32 \times 3$.

Step 2. Convert the image into grayscale.

Step 3. Carry out discrete cosine transform (DCT) on the image and convert it into frequency domain.

Step 4. Select the part with frequency domain of 16×16 to calculate the average value.

Step 5. Generate a 2D array according to the binarization of the average value (if less than the average value then assign 1 otherwise 0), it constructs the 256-bit perceptual hashing by expanding the 2D array into a 1D array.

The improved perceptual hashing algorithm obtains the outline details of the image by moving the value part in the frequency domain, and at the same time, it doubles the sampling matrix in the frequency domain, so that the length of the generated corresponding hash value is expanded by 4 times, the sampling efficiency is high in the approximate sampling time, and the accuracy of infringement detection of hash value is greatly improved. Table 1 shows the time required to generate the hash values and calculate the similarity measurement of the images 1.jpg and 2.jpg before and after improving the perceptual hashing algorithm.

As shown in Table 1, the accuracy of the improved perceptual hashing algorithm for judging the similarity of different images has nearly doubled and the verification time has only increased by 0.000005 seconds (basically negligible).

3.2.2 Image Encryption Algorithm

IPFS is designed for all users, allowing all nodes to access it at will. The nodes connected to IPFS network can find file content and download it. At the same time, the file upload of IPFS does not need to verify the identity of the sender, and generates a unique hash value according to the file content. When the file content changes, the hash value is completely different. The grotto mural data files stored in IPFS face the following two possible threats: 1) Once a user obtains the image data through illegal means and publishes the data to IPFS or other distributed file storage systems in advance before the copyright owner completes the copyright registration of the image data, the infringing party cannot be verified according to the generation time; 2) Users can easily obtain the data stored in IPFS. After obtaining the data, they can also tamper with the image data and copyright files and publish them to other platforms.

After considering the above situation, before the copyright owner uploads the grotto mural images and copyright information to IPFS, the copyright owner must encrypt the chaotic sequence of the images by using 1D Logistic chaotic map. The equation of 1D Logistic chaotic map is shown in Equation (1).

$$X_{k+1} = U * X_k * [1 - X_k] \tag{1}$$

Table 1: The hash value generation and similarity measurement time of Images 1.jpg, 2.jpg

		Traditional Phash	The improved Phash
	Hash value (bit)	1000001010101100 0100000001000010 1000011001000010 0000000000010000	11101100001101010110010101100100000000011001001 01000000110100101111001010100001011101110011 1 0110010110001110001101101100111100001101110 00 1111011110001010010011100011010100011001011 10 0101011100011001101110100001010000110001101 01 1000111000111001110101010010100
1.jpg	Generate time (s)	0.000979	0.000997
	Hash value (bit)	1010000010001000 0100100000010001 0100000000100000 0000001000000001	00001111101001001100111001100111101010110011 0 0111001001111110000110011001111001011011101 00 1100111001110110001100101001100110110100110 00 1100010010010110110000110101111010011001010 11 1100111001101010011101111110101011011011100 00 0110110011000100011001010000011
2.jpg	Generate time (s)	0.000978	0.000997
Similarity	Similarity (%)	70.31250	47.65625
measurement	Execution time (s)	0.000996	0.001001

where $k = 0, 1, ..., n$.

When the initial value $X_0 \in (0,1)$ and the control parameter $U \in (3.569945,4]$, the Logistic mapping reaches chaotic state. As the value of the U approaches 4, the iterative generated values are pseudo-random distribution, and the better encryption effect of chaotic sequences. In a chaotic system, even if the initial value X_0 changes slightly, the structure of the obtained data is completely different. In case of a fixed set of parameters, the algorithm can be easily cracked even if it is iterated to achieve a fully chaotic state. In order to solve this problem, U=4, and the key generation method is improved from a fixed X_0 to a different random key $Q(0<Q<1)$ uniquely corresponding to the image as the initial value, so as to improve the security of chaotic sequence encryption. The specific processing steps of the image encryption algorithm are as follows:

Step 1. Calculate the perceived hash value of the original image, and perform MD5 encryption on the generated 256-bit binary hash value to generate a unique $Md5Key$ (32-bit hexadecimal string).

Step 2. Partition $Md5Key$ and convert each digit from hexadecimal to decimal, and find the maximum value. In order to ensure that the chaotic function is in a chaotic state, the generated Q value must meet the condition that it is greater than 0 and less than 1. Therefore, the maximum value obtained is treated as R. After $Md5Key$ segmentation and conversion, each decimal number is divided by R (Ten decimal places are reserved). The median of the generated 32 numbers is Q.

Step 3. This 1D chaotic sequence **A** is generated by iterating Q for the same number of times according to the pixel size of the original image, then another 1D chaotic sequence **B** is obtained by normalizing the 1D sequence **A** to (0, 255), and transform **B** into a 2D matrix **G** of $M \times N$.

Step 4. The **G** and the image are bitwise XOR to obtain chaotic sequence encrypted image.

The realization of chaotic sequence encryption algorithm is shown in Algorithm 1.

Algorithm 1 Logistic chaotic sequence encryption

```
Input: $Md5Key$, original image
Output: encrypted image
initialization $R$, array $a$, array $b$, 1D sequence A, 2D matrix G
for $i \rightarrow Md5Key$ do
    $a \leftarrow$ int($i$,16)
end for
$R \leftarrow$ max($a$)*(3/2)
for $j \rightarrow a$ do
    $b \leftarrow$ round($j/R$,10)
end for
$Q \leftarrow$ median($b$)
for $s \rightarrow 100$ do
    $X_0 \leftarrow Q$
end for
Generate chaotic sequence A by applying Equation (1)
Generate matrix G by applying Step 3
Get the encrypted image according to Step 4
return encrypted image
```

3.2.3 Construction of Double Smart Contract

In order to realize the copyright protection of digital images of grotto murals, by setting up Hyperledger Fabric network, creating channels and nodes (Orderer, Peer) and issuing corresponding certificates, smart contract 1 and smart contract 2 (smart contract 1 realizes the registration of copyright information and smart contract 2 realizes the registration of consumer information) are written for copyright owner and consumers respectively. The contracts are deployed to the same channel of the same network, and written by Typescript language. When a consumer initiates a copyright transaction application, it interacts with peer-to-peer service nodes through Fabric-SDK to realize the call of double smart contract, and improves the algorithm of copyright information creation and copyright transaction process.

1) Registration of copyright information

a. Copyright information creation module
Before writing the copyright information into the blockchain, the copyright owner generates a 256-bit binary string from the image data by using the improved perceptual hashing function, and then uploads it to the blockchain network with other copyright information. Then, it calls the smart contract 1 to verify whether the $imageID$ in the copyright information to be created exists, if it exists, the creation fails. If not, the uploaded image data is subject to infringement detection. By automatically calculating the Hamming distance between the $imageHash$ of copyright information to be created and $HashAll$ already stored in the blockchain, and then calculate similarity. When the similarity is less than or equal to the pre-set threshold of 47.65625%, it is proved that the image is different from others stored in the blockchain, and no infringement is involved. At this time, the copyright information is created successfully, and the unique corresponding $Md5Key$ is generated according to the hash value of the image. Otherwise, the image is judged as an infringing image, and the input is refused.

b. Copyright information inquiry module
The copyright owner can read the copyright information according to the $imageID$, and compare the $imageID$ of the copyright information to be queried with the $imageID$ existing in the blockchain. If it exists, the data will be returned to the copyright owner, otherwise the query will fail.

c. Copyright information update module
Firstly, the copyright owner inputs the $imageID$, and queries the corresponding storage position of copyright information in the blockchain according to the $imageID$, so as to update all data including $imageID$, $imageName$, $imageHash$, $Md5Key$, $imageResolution$, $CopyrightOwnerEmailAddress$, $imageCopyrightOwnerName$ and $phocopyrightValue$.

d. Copyright information deletion module
Updating copyright information in real time helps to save the storage space of blockchain, so it is necessary to delete copyright information for lost image data or expired copyright information (the copyright owner or individual of the image is changed).

The process of copyright information registration is shown in Algorithm 2.

Algorithm 2 Registration of copyright information

```
Input: $imageID, imageName, phocopyrightValue, imageResolution, imageCopyrightOwnerName, imageHash, CopyrightOwnerEmailAddress$
Output: results of copyright information creation, inquiry, update and deletion
initialization $Similarity$, $Hamming\ distance$, $HashAll \in \{0,1\}^{256}$
// calculate similarity
for $i \rightarrow 256$ do
    if $imageHash[i] != HashAll[i]$ then
        $Hamming\ distance$ += 1
    end if
end for
$Similarity = (1 - Hamming\ distance/256) * 100\%$
if $imageID$ exists then
    The copyright has been registered, so you can inquire, update and delete it
else if $Similarity > 47.65625$ then
    The $imageHash$ already exists, and the copyright creation failed
else
    Input image copyright information and generate $Md5Key$ according to $imageHash$
end if
return results of copyright information creation, inquiry, update and deletion
```

2) Registration of consumer information

a. Consumer information creation module
Consumers must have a legal identity before conducting copyright transactions, so they need to create consumer information and complete the audit through smart contract 2. After the approval, consumers can recharge certain assets to their wallets for payment in the copyright transactions.

b. Consumer information inquiry module
Consumers can query the asset value according to their personal account $OwnerName$, and determine whether they meet the conditions for purchasing digital images of grotto murals.

c. Consumer information update module
Considering the popularity of the sharing economy, current consumers may sell their accounts.

At this time, they need to change $OwnerName$ of their personal account so that they can continue to use the remaining assets under the original account. When consumers want to buy copyright information, but the current asset value is not enough, they can recharge their accounts and update the asset value part of the information.

d. Consumer information deletion module
Generally, the consumer may register multiple accounts and own corresponding assets in real applications. Once the consumer forgets the account, the account and its assets become "inactive assets" similar to banks, which not only causes the waste of blockchain storage, but also causes the loss of consumers. Therefore, consumers need to delete useless $OwnerName$ in time.

The process of consumers information registration is shown in Algorithm 3.

Algorithm 3 Registration of consumer information

```
Input: $OwnerName$, $OwnerValue$
Output: results of consumer information creation, inquiry, update and deletion
if $OwnerName$ ! exists then
    Create consumer information($OwnerName$, $OwnerValue$)
else
    The consumers has been registered, so you can inquire, update and delete it
end if
return results of consumer information creation, inquiry, update and deletion
```

3) Copyright transaction
Any node can download data from IPFS. Thus, the grotto mural images and copyright information files uploaded to IPFS will be tampered with. Therefore, the copyright owner encrypts the image and deletes some copyright information before uploading the data. When consumers download images and copyright information, the obtained images are encrypted images of grotto murals encrypted by chaotic sequences, and the obtained copyright information files are other copyright data excluding $imageHash$ and $Md5Key$. Consumers can obtain encryption and decryption keys by purchasing image copyright to decrypt images, and this process is realized by calls double smart contract. Firstly, the consumer calls smart contract 2 to verify the identity information according to their personal accounts. If consumer information exists, they calls smart contract 1 to verify the existence of image copyright information according to $imageID$. If copyright information exists, they calls smart contract 2 again for asset evaluation, and determine whether the current assets of consumers meet the payment conditions. If so, complete payment, change the balance of consumer assets and return $Md5Key$. If not, the transaction fails. After consumers get the key, they can decrypt the image, thus obtaining the digital image of grotto murals. The copyright transaction process is shown in Algorithm 4.

Algorithm 4 Copyright transaction

```
Input: $imageID$, $OwnerName$
Output: results of copyright transaction
//verify the identity information according to Algorithm 3
//verify the existence of image copyright information according to Algorithm 2
//asset evaluation according to Algorithm 2 and Algorithm 3
if $OwnerName$ ! exists then
    return Consumer identity verification failed, copyright transaction failed
else if $imageID$ ! exists then
    return Image copyright information does not exist, copyright transaction failed
else if $OwnerValue < phocopyrightValue$ then
    return Payment failed, copyright transaction failed
else
    Complete and update $OwnerValue$
    return $Md5Key$
end if
```

4 Experimental Results and Performance Analysis

The system builds Fabric 2.3 network based on Hyperledger Fabric under CentOS7, writes chain code with Typescript and completes deployment, and completes image hash generation and image chaotic sequence encryption with Python. The proposed method is tested by using a self-defined data set of grotto murals, which is about 20GB in size and includes 16 kinds of images.

4.1 Performance Analysis for Digital Image Copyright Infringement Detection

At first, perform various conversion operations on any image in the data set, such as cropping, noise addition, blur (using mean blur, adjusting blur radius to 2px, 4px, 6px, 8px, 10px), sharpen (enhancing sharpness from 1 to 6), rotation(counterclockwise) and re-size, etc. The image "2.jpg" in the experimental data set is selected for these transformations, and each original image generates 37 corresponding transformed images, as shown in Figure 2.

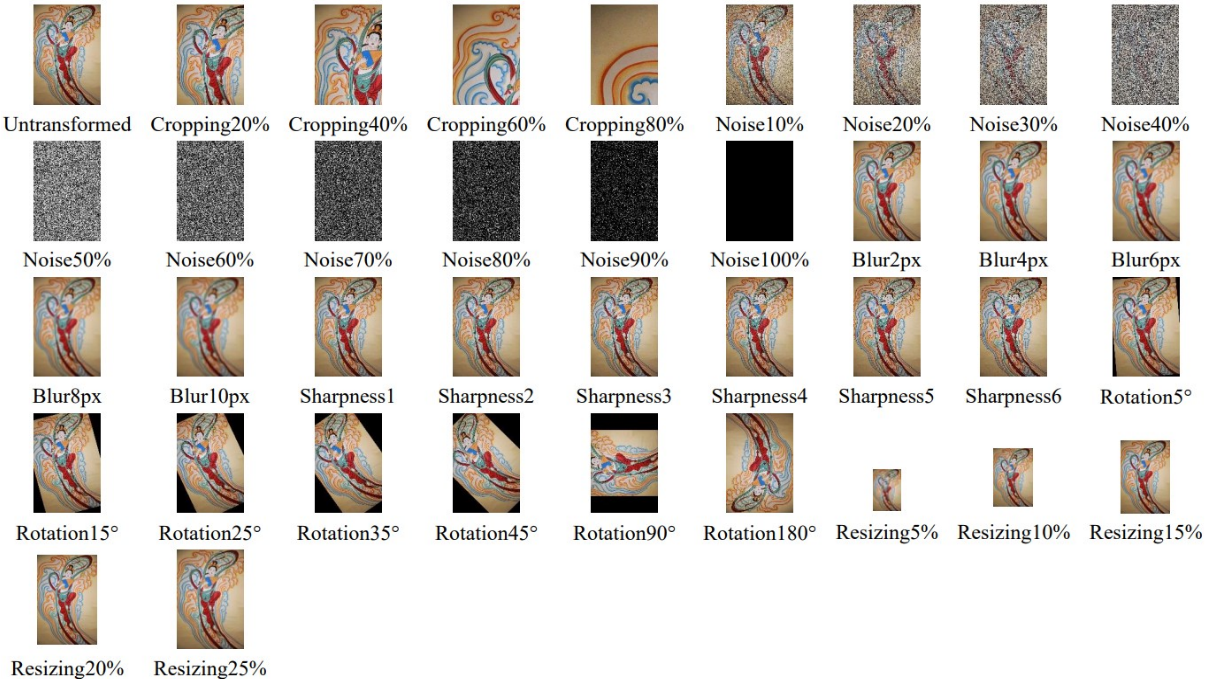


Figure 2: Operations on image 2.jpg with different modification

In order to evaluate the effectiveness of perceptual hashing in copyright infringement detection, we compare Phash, which we have used in the proposed frame work with two other techniques, Ahash and Dhash, and uses *Similarity* to measure the similarity between the original image and the modified image. The three hashing algorithms of Ahash, Dhash and Phash all return 64 bits hashes. The calculation method of *Similarity* is shown in Equation (2).

$$Similarity = (1 - \frac{Hamming\ distance}{Hashlength}) \times 100\% \quad (2)$$

where *Hamming distance* is the Hamming distance between the hash values of the two images, *Hashlength* is the length in bits of the hash value.

Considering that the image data of grotto frescoes have bright colors, many portraits, damaged, similar and noisy, two similar but different the original images (1.jpg and 2.jpg) are selected to calculate their similarity under different hash algorithms. Figure 3 shows the similarity between the image 1.jpg and 2.jpg under different hash algorithms. Figure 3(a) shows the image 1.jpg and 2.jpg. Figure 3(b) shows the similarity between the two images under three different hash algorithms.

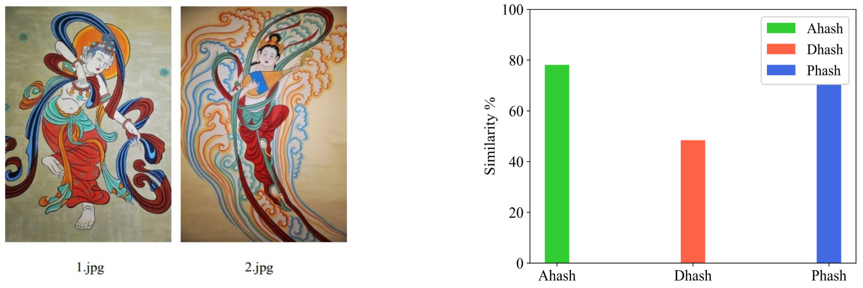
The similarity of different images is lower, the better the implementation effect of perceptual hashing algorithm. As shown in Figure 3(b), the similarity of similar but different images calculate by the Ahash, Dhash and Phash are 78.12500%, 48.43750% and 70.31250% respectively. The difference hash effect is relatively good.

In order to further evaluate the effectiveness of the three hashing algorithms for infringement detection, the original image 2.jpg is subjected to a series of image processing operations, and then the similarity between the processed images of 1.jpg and 2.jpg under different hashing algorithms is calculated again. Figure 4 shows the similarity measurement between the original image 1.jpg and the original image 2.jpg after clipping, adding noise, blurring, sharpening, rotating and resizing. Except for the noise addition, the performance of the three hashing algorithms under other image processing is: Dhash >Phash >Ahash.

Compared with the copied images and published edition in the market, the murals have higher noise which obtained by shooting, and the image effect tends to the image after noise addition processing, so it is more necessary to consider the selection of hash algorithm under noise addition processing. As shown in Figure 4(b), the similarity between Ahash and Dhash for different degrees of noise processing is quite different. Relatively speaking, Phash is more stable in the case of image noise addition processing, so Phash is used to measure the similarity of images. Although the image similarity value of Phash is relatively stable, the accuracy of similarity judgment is low. Therefore, this scheme effectively solves the problem of low accuracy by improving the perceptual hashing algorithm. Figure 5 shows the similarity measurement results of the image 1.jpg and similar image 2.jpg using improved perceptual hashing algorithm, Ahash and Dhash.

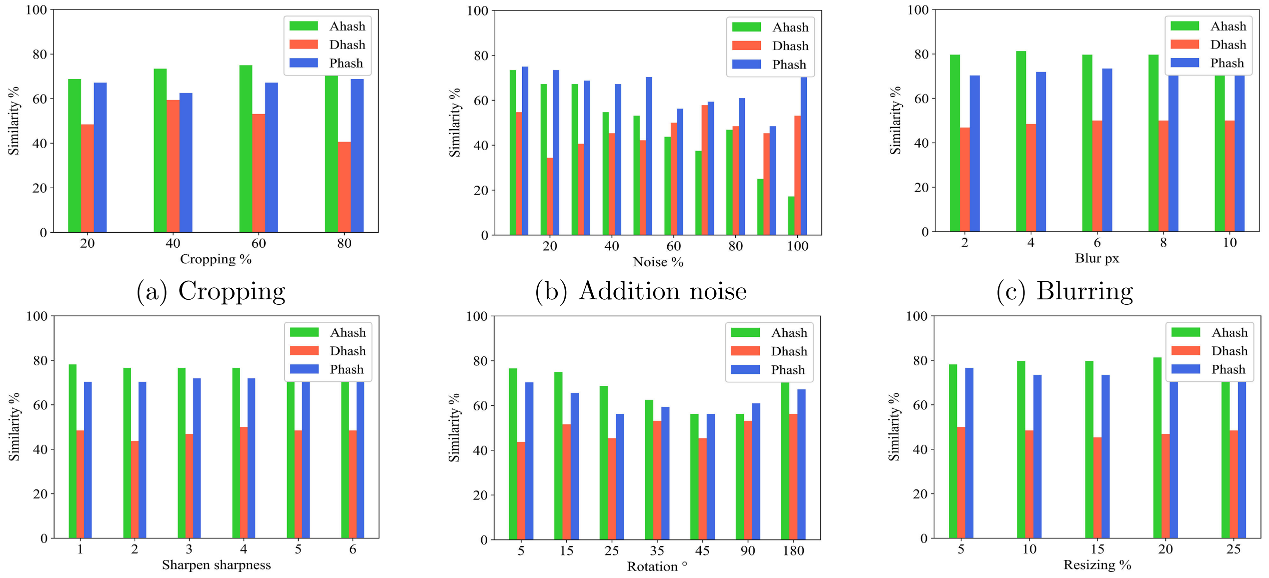
As shown in Figure 5, for similar but different images, the similarity of the improved perceptual hashing algorithm is 47.65625%. Compared with 70.31250% of the unimproved perceptual hashing algorithm and 48.43750% of the Dhash algorithm, the accuracy of similarity measurement has been greatly improved.

After clipping, noising, blurring, sharpening, rotating



(a) The image1.jpg and 2.jpg (b) The similarity between the two images

Figure 3: The similarity under different hash algorithms



(a) Cropping (b) Addition noise (c) Blurring

(d) Sharpening (e) Rotation (f) Resizing

Figure 4: Before the improvement of perceptual hashing algorithm, the image 2.jpg after various transformation processing and the image 1.jpg similarity measurement.

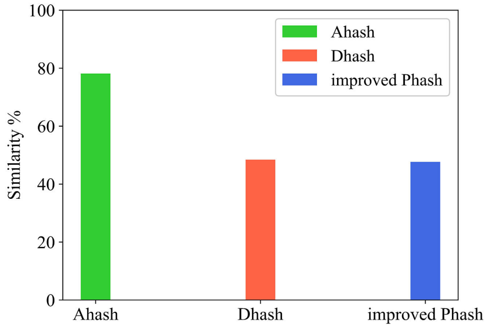


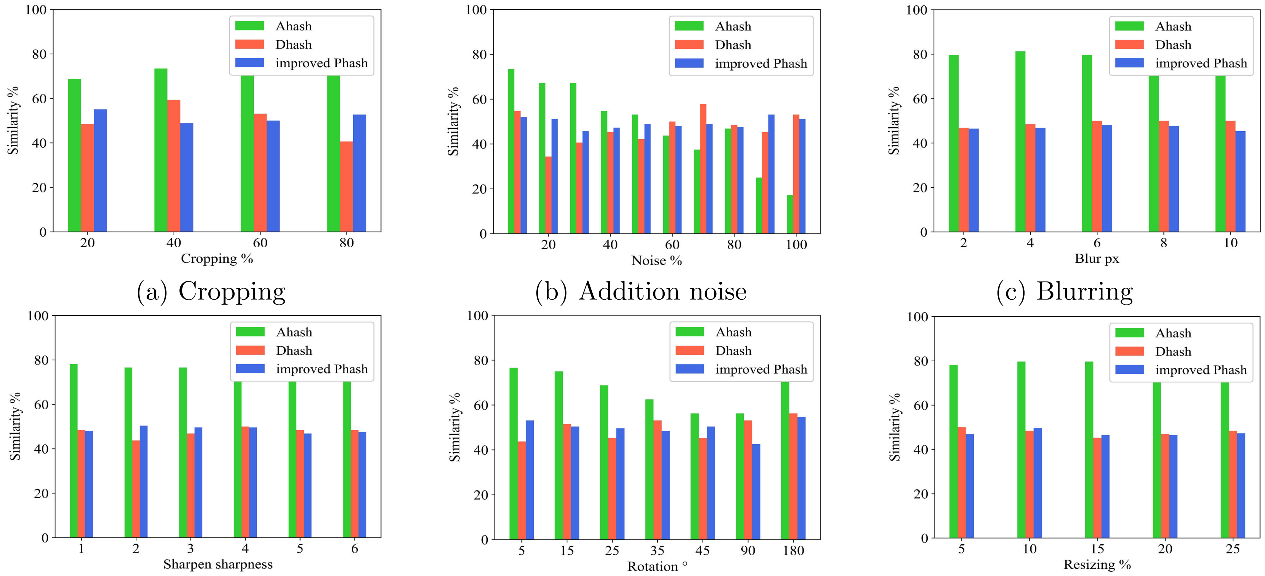
Figure 5: Similarity measurement between image 1.jpg and similar image 2.jpg under three hash algorithms

and resizing the original image 2.jpg, the similarity measurement with the original image 1.jpg is done by using the improved perceptual hash algorithm. Figure 6 shows the similarity measurement of infringement detection under different hash algorithms.

As shown in Figure 6, the accuracy of similarity measurement using the improved perceptual hashing algorithm has been greatly improved. The performance of the improved perceptual hashing algorithm is relatively stable in different degrees of noise addition processing. Except for noise addition, other image processing methods use the improved perceptual hashing algorithm to measure image similarity, which is better than Ahash and Dhash. Therefore, this scheme sets the similarity threshold to 47.65625%, selects the improved perceptual hashing algorithm to process the image data, uploads the generated hash value to the blockchain, and measures the similarity in the smart contract to realize the infringement detection of the image copyright.

4.2 Performance Analysis for Image Encryption Algorithm

In order to enhance the security of the images uploaded to IPFS, a unique chaotic sequence encryption initial value X_0 is determined for each image by MD5 algorithm, and the images are encrypted. Figure 7 shows the compari-



(a) Cropping (b) Addition noise (c) Blurring

(d) Sharpening (e) Rotation (f) Resizing

Figure 6: After improved perceptual hashing algorithm, the image 2.jpg after various transformation processing and the image 1.jpg similarity measurement of infringement detection.

son of the original image, encrypted image and decrypted image of the image 1.jpg.

(a) Original (b) Encrypted (c) Decrypted

Figure 7: The comparison of the original image, encrypted image and decrypted image of the image 1.jpg

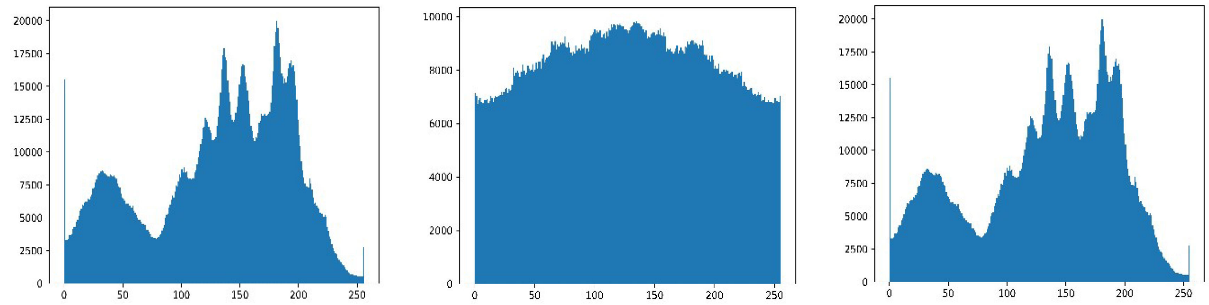
The performance of image encryption algorithm is analyzed by histogram analysis in statistical analysis, and the generated gray histogram of the image 1.jpg is shown in Figure 8.

As shown in Figure 8, the gray histogram of the original image shows obvious statistical laws, and is vulnerable to statistical analysis attacks. The distribution of gray histogram of the encrypted image is uniform, which is completely different from the original image. It is proved that the encryption scheme in the proposed method is secure, and can resist statistical analysis attacks well.

4.3 Performance Comparison with Existing Copyright Protection Schemes

Table 2 shows the comprehensive performance comparison results between proposed scheme and the existing multimedia content copyright protection schemes in [10, 13, 19, 20, 23, 32].

As shown in Table 2, the existing digital copyright protection schemes based on blockchain mostly realize automatic infringement detection through smart contracts, thus ensuring the privacy and security of digital copyright. The DRM system proposed in scheme [10] takes into account the characteristics of online educational multimedia resources, uses three smart contracts to realize copyright protection, and conducts infringement detection through watermark extraction and hash digest. Although the accuracy rate is improved, the verification is complicated and the system is not scalable. Scheme [13, 20] detects infringement by perceptual hashing, which ensures the verification effect and reduces the verification complexity, but the ownership of users is not treated. Scheme [19, 32] uses the consensus mechanism of blockchain to realize full-cycle copyright protection by smart contract, but does not consider the infringement detection and data security of digital copyright content itself. Scheme [23] ensures the security of multimedia copyright protection through improved watermarking algorithm, but there are still security and ownership problems due to the restriction of digital watermarking algorithm and single contract. By analyzing the problems existing in the above copyright protection technologies, this scheme improves the traditional perceptual hashing algorithm, and improves the judgment accuracy while ensuring the low verification complexity. Meanwhile, using the scheme [10] for reference, the double smart contract is used to solve the problem that the user's rights in a single contract are unclear. In the smart contract 1, the infringement detection module is combined with the copyright information registration process to realize automatic infringement detection. At the same time, the correlation key and chaotic sequence encryption are combined to encrypt the image data to be uploaded to



(a) Original (b) Encrypted (c) Decrypted

Figure 8: The gray histogram of the image 1.jpg

Table 2: Performance comparison with existing copyright protection schemes

Authors	Scalability	Privacy	Security	Double smart contract	Infringement detection
Ref. [20]	✓	✓	✓	×	wavelet hash
Ref. [10]	×	✓	✓	×	watermark/hash digest
Ref. [13]	✓	✓	✓	×	perceptual hashing
Ref. [19]	×	✓	✓	×	none
Ref. [32]	×	✓	✓	×	none
Ref. [23]	✓	✓	✓	×	watermark
Proposed	✓	✓	✓	✓	perceptual hashing

IPFS, which improves the security of digital images of grotto murals.

5 Example Simulation Results

5.1 Digital Image Copyright Infringement Detection

In order to verify the realization of the infringement detection part in the smart contract 1, after the copyright information creation of the image 2.jpg is successfully completed, the images processed with different degrees of noise are uploaded, Table 3 shows the creation results of corresponding copyright information of processed images.

Table 3: The copyright information creation results of the image 2.jpg

Noise addition (%)	Similarity with image 1.jpg (%)	Acceptance result
10	51.95312	Rejected
20	51.17188	Rejected
30	45.70312	**Accepted**
40	47.26562	**Accepted**
50	48.82812	Rejected
60	48.04688	Rejected
70	48.82812	Rejected
80	47.65655	Rejected
90	53.12500	Rejected
100	51.17188	Rejected

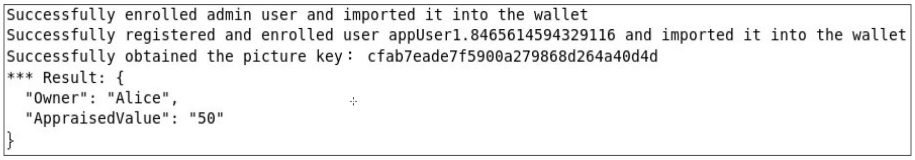
As shown in Table 3, the images with different proportions of noise addition (except for 30% and 40% of noise addition) failed to pass the infringement detection created by copyright information, which proves that using the improved perceptual hashing algorithm to judge the similarity can effectively identify the infringing images and realize the copyright protection of grotto mural images.

5.2 Image Copyright Transaction

In order to obtain the required image, consumers need to purchase the image copyright and obtain the $Md5Key$ to decrypt the encrypted image downloaded by IPFS. Copyright transaction mainly includes three parts: verification of consumer identity information, verification of the existence of image copyright information and asset evaluation. The copyright transaction process is completed by calling double smart contracts. The copyright transaction success interface is shown in Figure 9.

5.3 Upload and Download of Files in IPFS

The copyright owner completes the uploading of encrypted images and part of copyright information on IPFS. Before uploading, put the image and copyright information to be uploaded into a folder (the folder is named after the $imageID$ and contains encrypted image data and copyright information), and then upload the folder through IPFS network. After success, the encrypted image data hash value, copyright information hash value and folder hash value will be generated at the same time. Ac-



```
Successfully enrolled admin user and imported it into the wallet
Successfully registered and enrolled user appUser1.8465614594329116 and imported it into the wallet
Successfully obtained the picture key: cfab7eade7f5900a279868d264a40d4d
*** Result: {
  "Owner": "Alice",
  "AppraisedValue": "50"
}
```

Figure 9: The copyright transaction success

cording to the configured IPFS URL, port number and generated folder hash value, you can directly view the contents of phocopyright1 folder in the browser, and all nodes in IPFS can download the uploaded files according to the hash value generated when uploading the folder. The process of uploading and downloading IPFS files is shown in Figure 10.

As shown in Figure 10(a), the copyright owner stores the grotto murals encrypted image (1-encrypt.jpg) and part of the copyright information text file (phocopyright-info.txt) in the folder (phocopyright1). As shown in Figure 10(b), the copyright owner uses the IPFS file upload command to complete the upload of phocopyright1. After successful upload, IPFS hash of phocopyright1, 1-encrypt.jpg and phocopyroght-info.txt are generated respectively. At this time, the uploaded phocopyright1 can be viewed in the browser, and the results displayed in the browser are shown in Figure 10(c), (d) and (e). The consumer uses the IPFS file download command to get the folder named after the IPFS hash value of phocopyright1, and the successful download interface is shown in Figure 10(f).

6 Conclusions

In order to realize the copyright protection of digital images of grotto murals, and solve the problems that copyright information is easy to tamper and cannot be traced, a digital image copyright protection method based on blockchain and perceptual hashing was proposed by combining perceptual hashing, image encryption, blockchain and IPFS. The proposed method improves the existing perceptual hashing algorithm to solve the problems of missing image details in feature capture and low accuracy in image infringement detection, and redefines the frequency domain spatial sampling in image data hashing, which greatly improves the accuracy of image similarity threshold. In addition, combining image hash and MD5 to generate a unique and random key for chaotic encryption reduces the possibility of users stealing copyright information. Through smart contracts, copyright owner and consumers can register first and then upload/download, and the security of copyright trading process is enhanced through triple judgment when double smart contract are called. Copyright owners only upload encrypted images and part of copyright information to IPFS, which also improves the security of copyright data to a certain extent, effectively expands the storage space of blockchain and reduces the storage cost. The simulation results show that the proposed method can effectively protect the copyright of digital images of grotto murals, and can be further applied to other multimedia data copyright protection fields.

In this paper, the chaotic sequence encryption key generation method is simple, and the key space is small, which ensures the security and improves the encryption efficiency. However, there is a risk of being cracked, and there are still shortcomings in fine-grained access control of copyright transactions. The further research plan is to further improve the security of DRM by improving the key generation method, combining zero trust mechanism, searchable encryption and other technologies.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61862041). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] A. Abrar, W. Abdul and S. Ghouzali, "Secure image authentication using watermarking and blockchain," *Intelligent Automation and Soft Computing*, vol. 28, no. 2, pp. 577–591, 2021.

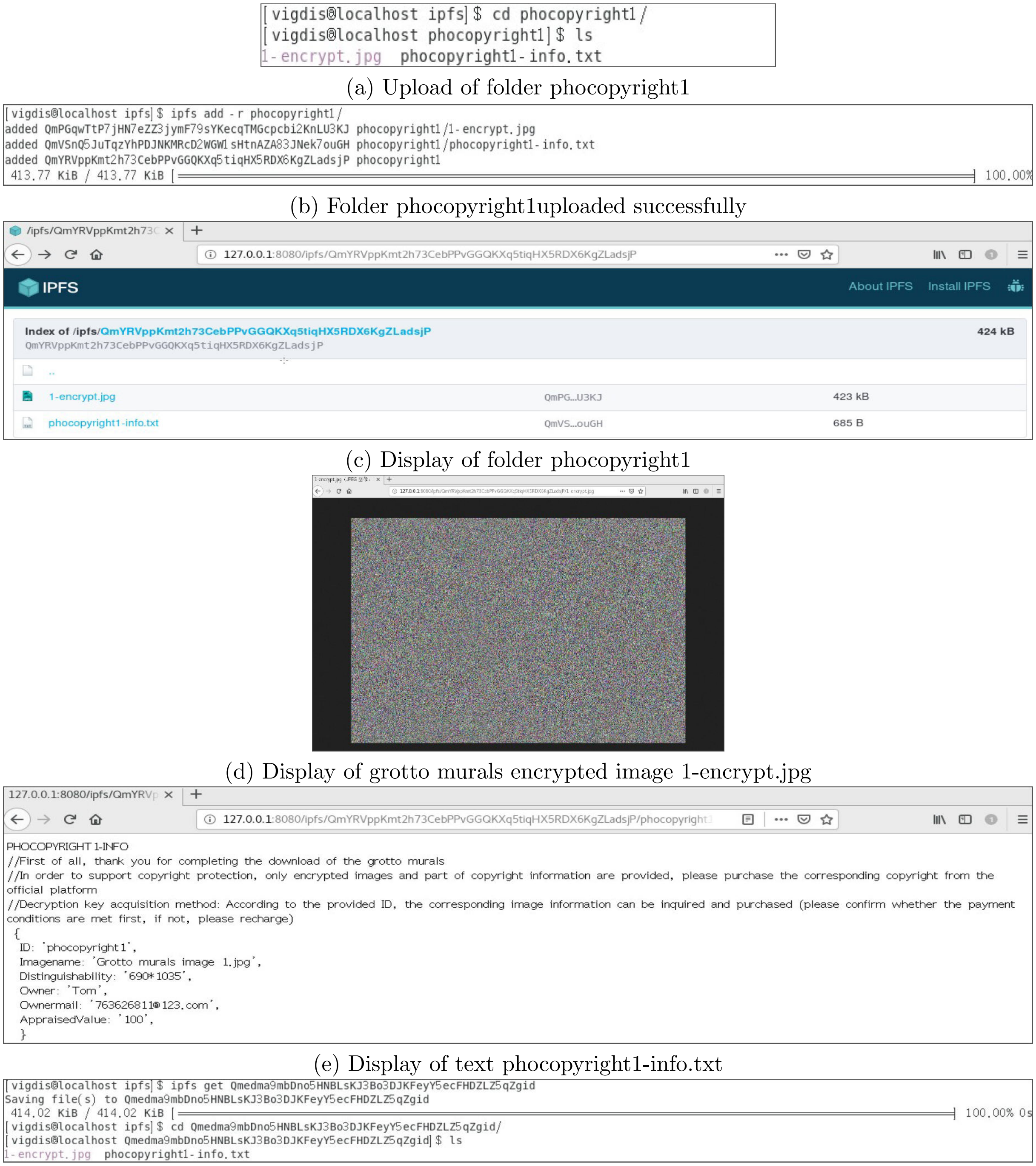
[2] C. C. Chang, K. F. Hwang, M. S. Hwang, "A digital watermarking scheme using human visual effects", *Informatics*, vol. 24, no. 4, pp. 505–511, Dec. 2000.

[3] C. C. Chang, K. F. Hwang, M. S. Hwang, "A block based digital watermarks for copy protection of images," in *Fifth Asia-Pacific Conference on Communications*, vol. 2, pp. 977-980, 1999.

[4] C. C. Chang, K. F. Hwang, M. S. Hwang, "Robust authentication scheme for protecting copyrights of images and graphics", *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 43-50, 2002.

[5] R. A. Dobre, R. O. Preda, R. A. Badea, *et al.*, "Blockchain-based image copyright protection system using jpeg resistant digital signature," in *2020 IEEE 26th International Symposium for Design and Technology in Electronic Packaging(SIITME)*, IEEE, Pitesti, Romania, pp. 206–210, Oct. 2020.

[6] T. Feng, R. Y. Yang and R. B. Gong, "Digital copyright protection system for oil and gas knowledge achievements based on blockchain," *Interna-*



```
[vigdis@localhost ipfs]$ cd phocopyright1/
[vigdis@localhost phocopyright1]$ ls
1-encrypt.jpg  phocopyright1-info.txt
```

(a) Upload of folder phocopyright1

```
[vigdis@localhost ipfs]$ ipfs add -r phocopyright1/
added QmPGqwTtP7jHN7eZZ3jymF79sYKecqTMGcpcbi2KnLU3KJ phocopyright1/1-encrypt.jpg
added QmVSnQ5JuTqzYhPDJNKMRcD2WGW1sHtnAZA83JNek7ouGH phocopyright1/phocopyright1-info.txt
added QmYRVppKmt2h73CebPPvGGQKXq5tiqHX5RDX6KgZLadsjP phocopyright1
 413.77 KiB / 413.77 KiB [==========================================================] 100.00%
```

(b) Folder phocopyright1uploaded successfully

/ipfs/QmYRVppKmt2h73C
127.0.0.1:8080/ipfs/QmYRVppKmt2h73CebPPvGGQKXq5tiqHX5RDX6KgZLadsjP
IPFS
About IPFS
Install IPFS

```
Index of /ipfs/QmYRVppKmt2h73CebPPvGGQKXq5tiqHX5RDX6KgZLadsjP                424 kB
QmYRVppKmt2h73CebPPvGGQKXq5tiqHX5RDX6KgZLadsjP
..
1-encrypt.jpg                          QmPG...U3KJ                            423 kB
phocopyright1-info.txt                 QmVS...ouGH                            685 B
```

(c) Display of folder phocopyright1

(d) Display of grotto murals encrypted image 1-encrypt.jpg

127.0.0.1:8080/ipfs/QmYRVp
127.0.0.1:8080/ipfs/QmYRVppKmt2h73CebPPvGGQKXq5tiqHX5RDX6KgZLadsjP/phocopyright1

```
PHOCOPYRIGHT 1-INFO
//First of all, thank you for completing the download of the grotto murals
//In order to support copyright protection, only encrypted images and part of copyright information are provided, please purchase the corresponding copyright from the official platform
//Decryption key acquisition method: According to the provided ID, the corresponding image information can be inquired and purchased (please confirm whether the payment conditions are met first, if not, please recharge)
 {
  ID: 'phocopyright1',
  Imagename: 'Grotto murals image 1.jpg',
  Distinguishability: '690*1035',
  Owner: 'Tom',
  Ownermail: '763626811@123.com',
  AppraisedValue: '100',
  }
```

(e) Display of text phocopyright1-info.txt

```
[vigdis@localhost ipfs]$ ipfs get Qmedma9mbDno5HNBLsKJ3Bo3DJKFeyY5ecFHDZLZ5qZgid
Saving file(s) to Qmedma9mbDno5HNBLsKJ3Bo3DJKFeyY5ecFHDZLZ5qZgid
 414.02 KiB / 414.02 KiB [========================================================] 100.00% 0s
[vigdis@localhost ipfs]$ cd Qmedma9mbDno5HNBLsKJ3Bo3DJKFeyY5ecFHDZLZ5qZgid/
[vigdis@localhost Qmedma9mbDno5HNBLsKJ3Bo3DJKFeyY5ecFHDZLZ5qZgid]$ ls
1-encrypt.jpg  phocopyright1-info.txt
```

(f) Folder phocopyright1 download successfully

Figure 10: Upload and download of files in IPFS

tional Journal of Network Security, vol. 23, no. 4, pp. 631–641, 2021.

[7] T. Gaber, A. Ahmed and A Mostafa, "Privdrm: A privacy-preserving secure digital right management system," in*Proceedings of the Evaluation and Assessment in Software Engineering*, ACM, Trondheim, Norway, pp. 481–486, Apr. 2020.

[8] J. T. Gao, H. Y. Yu, X. Q. Zhu, *et al.*, "Blockchain-based digital rights management scheme via multi-authority ciphertext-policy attribute-based encryption and proxy re-encryption," *IEEE Systems Journal*, vol. 15, no. 4, pp. 5233-5244, 2021.

[9] A. Garba, A. D. Dwivedi, M. Kamal, *et al.*, "A digital rights management system based on a scalable blockchain," *Peer-to-Peer Networking and Applications*, vol. 14, no. 5, pp. 2665–2680, 2021.

[10] J. Q. Guo, C. Y. Li, G. Z. Zhang, *et al.*, "Blockchain-enabled digital rights management for multimedia resources of online education," *Multimedia Tools and Applications*, vol. 79, no. 15, pp. 9735–9755, 2020.

[11] M. S. Hwang, C. C. Chang, K. F. Hwang, "Digital watermarking of images using neural networks", *Journal of Electronic Imaging*, vol. 9, no. 4, pp. 548–555, Jan. 2000.

[12] M. Kripa, A. Nidhin Mahesh, R. Ramaguru, *et al.*, "Blockchain framework for social media drm based on secret sharing," in *2020 International Conference on Information and Communication Technology for Intelligent Systems*, Springer, Singapore, pp. 451–458, Oct. 2020.

[13] R. Kumar, R. Tripathi, N. Marchang, *et al.*, "A secured distributed detection system based on ipfs and blockchain for industrial image and video data security," *Journal of Parallel and Distributed Computing*, vol. 152, pp. 128–143, 2021.

[14] T. Li, H. Wang, D. B. He, *et al.*, "Blockchain-based privacy-preserving and rewarding private data sharing for iot," *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 15138–15149, 2022.

[15] W. Li, Y. Zhu, L. Tian, *et al.*, "FPGA-based hardware acceleration for image copyright protection system based on blockchain," in *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, IEEE, New York, NY, USA, pp. 234–239, Aug. 2020.

[16] X. B. Li, M. Darwich, M. A. Salehi,*et al.*, "A survey on cloud-based video streaming services," *Advances in Computers*, vol. 123, pp. 193–244, 2021.

[17] L. Liu, W. Shang, W. Lin, *et al.*, "A decentralized copyright protection, transaction and content distribution system based on blockchain 3.0," in *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, IEEE, Ho Chi Minh City, Vietnam, pp. 45–50, Jan. 2021.

[18] L. Liu, W. Zhang and C. Han, "A survey for the application of blockchain technology in the media," *Peer-to-Peer Networking and Applications*, vol. 14, no. 5, pp. 3143–3165, 2021.

[19] Y. Liu, J. Zhang, S. Wu, *et al.*, "Research on digital copyright protection based on the hyperledger fabric blockchain network technology," *PeerJ Computer Science*, vol. 7, p. e709, 2021.

[20] R. Mehta, N. Kapoor, S. Sourav, *et al.*, "Decentralised image sharing and copyright protection using blockchain and perceptual hashes," in *2019 11th International Conference on Communication Systems Networks(COMSNETS)*, IEEE, Bengaluru, India, pp. 1–6, Jan. 2019.

[21] D. Mishra, A. Kasi, M. S. Obaidat, *et al.*, "Construction of lightweight content key distribution framework for drm systems," in *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, Arad, Romania, pp. 863–868, Dec. 2021.

[22] J. Nan, Q. Liu and V. Sugumaran, "A blockchain-based code copyright management system," *Information Processing & Management*, vol. 58, no. 3, p. 102518, pp. 1-17, 2021.

[23] I. Natgunanathan, P. Praitheeshan and L. X. Gao, "Blockchain-based audio watermarking technique for multimedia copyright protection in distribution networks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 3, pp. 1–23, 2022.

[24] J. S. Pan, X. X. Sun, S. C. Chu, *et al.*, "Digital watermarking with improved sms applied for qr code," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104049, 2021.

[25] S. Rana and D. Mishra, "Provably secure authenticated content key distribution framework for iot-enabled enterprise digital rights management systems," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 36, no. 3, pp. 131–140, 2021.

[26] N. Ren, Y. Z. Zhao, C. Q. Zhu, *et al.*, "Copyright protection based on zero watermarking and blockchain for vector maps," *ISPRS International Journal of Geo-Information*, vol. 10, p. 294, pp. 1–20, 2021.

[27] J. Y. Shen, "Blockchain technology and its applications in digital content copyright protection," in *Proceedings of the 4th International Conference on Economic Management and Green Development*, Springer, Singapore, pp. 18–25, Jan. 2021.

[28] J. Shi, D. Yi and J. Kuang, "A blockchain and sift based system for image copyright protection," in *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications*, ACM, Xi'an, China, pp. 1–6, Dec. 2019.

[29] M. Sultana, A. Hossain and F. Laila, "Towards developing a secure medical image sharing system based on zero trust principles and blockchain technology," *BMC Medical Informatics and Decision Making*, vol. 20, p. 256, pp. 1–10, 2020.

[30] B. Wang, S. Jiawei, W. Wang, *et al.*, "A blockchain-based system for secure image protection using zero-watermark," in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, IEEE, Delhi, India, pp. 62–70, Dec. 2020.

[31] C. Wang, B. Ma, Z. Xia, *et al.*, "Geometric resistant polar quaternion discrete fourier transform and its application in color image zero-hiding," *ISA transactions*, vol. 125, pp. 665–680, 2022.

[32] Z. Wang and T. Li, "Research on image copyright confirmation and protection model based on blockchain," in *2021 2nd International Conference on Control, Robotics and Intelligent System*, ACM, Qingdao, China, pp. 230–234, Aug. 2021.

[33] N. I. Wu, M. S. Hwang, "A novel LSB data hiding scheme with the lowest distortion", *The Imaging Science Journal*, vol. 65, no. 6, pp. 371–378, 2017.

[34] Z. Yan, "The color and artistic features of murals in dunhuang cave 465 in mogao grottoe," in *The 6th International Conference on Arts, Design and*

Contemporary Education (ICADCE2020), Atlantis Press, Moscow, Russia, pp. 56–62, Jan. 2021.

[35] D. Zhang, X. J. Wu, T. Xu, *et al.*, "Watch: Two-stage discrete cross-media hashing," *IEEE Transactions on Knowledge and Data Engineering*, https://doi.org/10.1109/TKDE.2022.3159131, pp. 1–13, 2022.

Biography

Qiu-yu Zhang Researcher/Ph.D. supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is vice dean of Gansu manufacturing information engineering research center, a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, multimedia communication technology.

Guo-rui Wu is currently a master student of the School of Computer and Communication, Lanzhou University of Technology, China. She received the BS degrees in network engineering from Lanzhou Institute of Technology, Gansu, China, in 2020. Her research interests include network and information security, multimedia data security, blockchain.

Research on Coverless Image Steganography

Kurnia Anggriani[1,2], Nan-I Wu[3], and Min-Shiang Hwang[1,4]

(Corresponding author: Min-Shiang Hwang)

Department of Computer Science & Information Engineering, Asia University[1]
500, Lioufeng Rd., Wufeng, Taichung 41354, Taichung, Taiwan, ROC
Faculty of Engineering, University of Bengkulu, Indonesia[2]
Department of Digital Multimedia, Lee-Ming Institute of Technology[3]
No.2-2, Lijhuan Rd., Taishan Township, Taipei County 243, Taiwan, ROC
Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan[4]
Email: mshwang@asia.edu.tw

(Received May 12, 2022; Revised and Accepted Dec. 8, 2022; First Online Dec. 10, 2022)

Abstract

In recent years, Coverless Image Steganography (CIS) has become an essential research topic for many scholars. This research topic was initiated in 2015 under steganography without embedding. Accordingly, CIS was presented to withstand the inadequacies of the general steganography method against the risks of steganalysis tools. This is because information concealed in CIS has no modification left on the stego image. Conversely, CIS performs a mapping operation to select a stego image containing secret information. This paper attempts to examine the development of CIS over the past five years, summarizing and analyzing the benefits and challenges of the current methods used by each survey paper. Moreover, we present the experimental results in tables and graphs to provide a precise performance comparison. This is done to outline future research requirements and opportunities in the CIS research topic. For example, in the future, developing a CIS method that is resistant to steganalysis tools, has a large hiding capacity, and can be used in real-time applications would be ideal.

Keywords: Coverless Data Hiding; Coverless Image Steganography; Mapping Operation

1 Introduction

In an internet-based world, the necessity for secure communication routes is unavoidable. One of the solutions to achieve safe communication is to use data hiding, commonly known as steganography. Steganography is described as a secret mechanism that conceals confidential data in other mediums while causing undetectable alterations in human perception [21, 26, 27]. Aside from the benefits of data hiding in establishing a safe environment, one of the most significant obstacles in the traditional steganographic approach is the risk of steganalysis tools. It is due to the stego image's alteration traces [16, 28].

In the traditional steganographic approach, image modification can occur in the spatial, transform, and compressed domains. In the spatial domain, the modification is performed directly in the pixel value of the image [2]. The standard method of spatial domain data hiding is utilizing the least significant bit (LSB) [10, 23–25, 29] and pixel value differencing (PVD) [14, 22]. As a result, spatial domain data hiding achieved high hiding capacity and, as a trade-off, was very vulnerable to steganalysis attacks [8]. In the transform domain and compressed domain data hiding, the modification deals with the transform coefficients [3, 20, 31] and compressed code [9, 11] of the images, respectively. As a result, transform domain, and compressed domain data hiding perform more robust against steganalysis attacks. However, the traditional steganographic approach has become increasingly susceptible and insufficiently safe due to the rapid development of steganographic tools in the previous year [6, 19]. Especially with the potential of image processing attacks like Additive Gaussian White Noise, Salt & Pepper noise, low-pass filtering, and JPEG compression. Because of the difficulties above, traditional data hiding must continue to evolve to improve the robustness of the data hiding method.

To address the challenges mentioned above, in 2015, Zhou *et al.* [33] proposed a concept of steganography without embedding, often known as coverless data hiding. Instead of embedding secret information by modifying the images' attributes, this technique matches the image to the secret information. The mapping operation's unique key is a hash sequence comprising secret information and an image. The picture will automatically incorporate the secret information when they have the same hash sequence. In the last five years, abundant methods have been introduced in the CIS. As a result, CIS can be divided into image-mapping-based CIS [4, 13, 30, 32, 34] and image-generation-based CIS [1, 5, 7, 12, 15, 17, 18]. In the image-mapping-based CIS, the main character is the mapping operation between hashing sequences of secret information to find the most similar image in the image dataset. On the other hand, image-generation-based CIS utilizes the capabilities of deep learning to produce an image representing secret information.

In this paper, we presented and summarized some articles in credible journal papers to analyze existing methods in coverless image steganography (CIS). Furthermore, this survey paper aims to identify a research gap to develop new strategies for future research.

In this survey paper, we highlighted five of the most relevant papers, "Towards a High Capacity Coverless Information Hiding Approach" [1] identified as survey paper 1 (Abdulsattar's scheme), "A Novel Coverless Information Hiding Method Based on the Most Significant Bit (MSB) of the Cover Image" identified as survey paper 2 (Yang *et al.*'s scheme), "Robust Coverless Image Steganography Based on DCT and LDA Topic Classification" [32] identified as survey paper 3 (Zhang *et al.*'s scheme), "A novel coverless information hiding method based on the average pixel value of the sub-images" [34] identified as survey paper 4 (Zou *et al.*'s scheme) and "Coverless Information Hiding Based on the Molecular Structure Images of Material" [4] identified as survey paper 5 (Cao *et al.*'s scheme).

The remaining paper is managed as follows: Section 2 presents the related works in detail. Section 3 discusses the comparison of survey papers' performance. Then, in Section 4, future research is provided. Lastly, the paper is concluded in Section 5.

2 Related Works

The fifth survey paper shared the same phase of coverless information hiding. The first is image hashing generation. The second is database or lookup table production, and the last is mapping operation. The main difference is in the used algorithm and image selection. The summarization of the fifth survey paper characteristic is shown in Table 1.

2.1 Survey Abdulsattar's Scheme

In 2021, Abdulsattar [1] introduced a coverless data hiding by utilizing the eigenvalues decomposition in a block of sub-images. This schema employs a single image, subsequently segmented into several block images. Each block image has its eigenvalues computed, which are then utilized to construct a hash sequence. The hash sequence is then saved in an ASCII code lookup table. When a secret message is provided, it is transformed into ASCII code format. Following that, mapping the hash sequence in the lookup table will begin. The lookup table contains the hash sequence's x and y coordinates to ensure an efficient mapping operation. Parlier public key algorithm encrypts the location information of the block image, which is shared between the sender and receiver in a public transmission channel. The flowchart of Abdulsattar's scheme is shown in Figure 1.

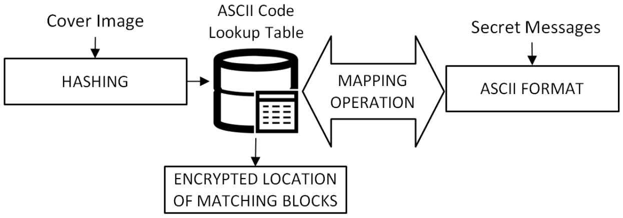


Figure 1: The flowchart of Abdulsattar's scheme [1]

In this scheme, the longer the hash sequence, the higher the hiding capacity. Therefore, Abdulsattar [1] investigated three parameters, namely block arrangement, block size, and overlapping block on the number of hash sequences. Figure 2 depicts four alternative arrangements. Several experiments revealed that arrangement two could generate more hash sequences than the other arrangements. Abdulsattar [1] adjusts six different block sizes. As a result, block sizes in the range of 3×3 and 6×6 can generate more hash sequences. The last parameter is the overlapping block. After the in-depth experiment, it can be concluded that overlapping blocks will produce more hash sequences.

To assess the robustness of Abdulsattar's scheme, seven types of image attacks were employed, including Gaussian noise, Salt & Pepper attack, speckle noise, median filtering, mean filtering, gaussian filtering, and histogram equalization. The test results are presented in Table 2.

2.2 Survey Yang *et al.*'s Scheme

In 2020, Yang *et al.* [30] introduced a coverless data-hiding scheme based on the MSB of the cover image. This method utilizes the average value μ of the fragment and maps the binary form of secret bits with the MSB of μ under pre-defined critical K. This approach achieves good image quality and is robust against steganalysis tools. However, Yang *et al.*'s scheme have a lower hiding capacity since one fragment only hides one secret bit. The flowchart of the embedding procedure is presented in Figure 3.

Four image attacks were employed to assess the robustness of Yang *et al.*'s scheme, including Gaussian noise, Salt & Pepper attack, low-pass filtering, and JPEG compression. Table 3 present the robustness analysis of Yang *et al.*'s scheme.

2.3 Survey Zhang *et al.*'s Scheme

In 2018, Zhang *et al.* [32] proposed a robust coverless image steganography based on DCT and LDA Topic Classification. The main idea of this scheme is to find the most relevant image according to secret messages. The purpose is to avoid the susceptibility of an irrelevant picture in a

Table 1: Summarization of survey paper

Schemes	Block Used Properties	Mapping Approach	Key Sharing Approch	Block Division Approch	Robustness Analysis
Abdulsattar's Scheme	Eigen decompistion	ASCII code	Pailier public key encryption algorithm	Partially overlapping	Under 7 kinds of attacks
Yang *et al.*'s Scheme	Average pixel value	Binary code (MSB)	Pseudo-random serial numbers	Non-overlapping	Under 4 kinds of attacks
Zhang *et al.*'s Scheme	Discrete cosine transform	Latent dirichlet allocation	Pseudo-random serial numbers	Non-overlapping	Under 14 kinds of attacks
Zou *et al.*'s Scheme	Average pixel value	Label sequence	Pseudo-random serial numbers	Non-overlapping	Not specified
Cao *et al.*'s Scheme	Average pixel value	Label sequence	Pseudo-random serial numbers	Non-overlapping	Not specified

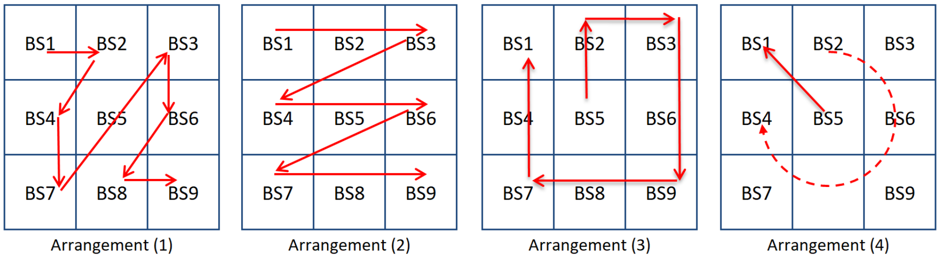


Figure 2: The block arrangement of Abdulsattar's scheme [1]

Table 2: Robustness analysis of Abdulsattar's scheme

Attacks	Parameter	Abdulsattar's Scheme [1]
Gaussian Noise Attack	v = 0.001	27.09
	v = 0.005	35.69
Salt & Pepper Noise Attack	r = 0.001	0.79
	r = 0.005	3.57
Speckle Noise Attack	v = 0.01	31.99
	v = 0.05	26.47
Median Filtering Attack	w = 3×3	16.54
	w = 5×5	25.01
Mean Filtering Attack	w = 3×3	13.18
	w = 5×5	22.64
Gaussian Filtering Attack	w = 3×3	4.01
	w = 5×5	13.13
Histogram Equalization	-	4.78

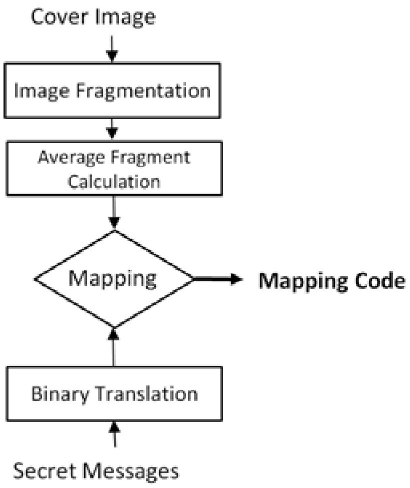


Figure 3: The flowchart of Yang *et al.*'s scheme

Table 3: Robustness analysis of Yang *et al.*'s scheme

Methods	Parameter	Yang *et al.*'s Scheme [30]
Gaussian Noise	v = 0.1	6.75
	v = 0.2	9.13
	v = 0.5	15.13
	v = 0.6	17.37
	v = 0.9	20.38
	v = 1.0	21.25
Salt & Pepper Noise	v = 0.001	0
	v = 0.003	0.13
	v = 0.005	0.25
	v = 0.007	0.63
	v = 0.009	0.88
	v = 1.0	1.38
Low Pass Filtering	w = 3×3	1.25
	w = 5×5	1.25
	w = 7×7	1.25
	w = 9×9	1.87
JPEG Compression	q=90	0.125

specific topic. The flowchart of the embedding procedure of Zhang *et al.*'s method is shown in Figure 4. Table 4 present the robustness analysis of Zhang *et al.*'s scheme.

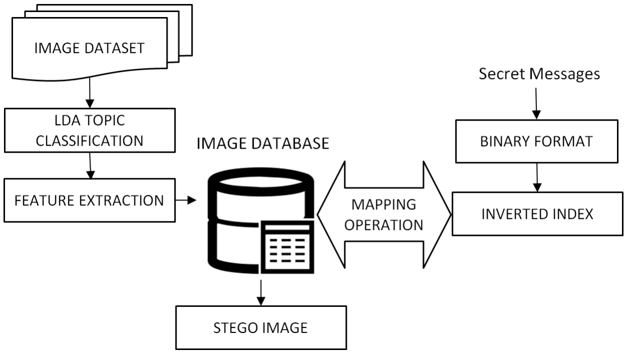


Figure 4: The Flowchart of Zhang's method [32]

2.4 Survey Zou *et al.*'s Scheme

In 2018, Zou *et al.* proposed a coverless information-hiding method based on the average pixel value of the sub-images. First, a Chinese-based dictionary is generated to manage the secret messages. Next, a hash sequence is generated by a hashing algorithm. Then a mapping relationship between the secret messages and a hashing sequence is operated to obtain the most appropriate image. The flowchart of Zou *et al.*'s scheme [34] is shown in Figure 5.

Table 4: Robustness analysis of Zhang *et al.*'s scheme

Attacks	Parameter	Zhang *et al.*'s Scheme [32]
Gaussian Noise	v = 0.001	3,01
	v = 0.005	1,72
	v = 0.1	0,86
Salt & Pepper Noise	r = 0.001	0
	r = 0.005	0
	r = 0.1	2,15
Speckle Noise	v = 0.01	0,86
	v = 0.05	0,86
	v = 0.1	2,15
Median Filtering	w = 3×3	0
	w = 5×5	0,86
	w = 7×7	1,72
Mean Filtering	w = 3×3	0
	w = 5×5	0
	w = 7×7	0
Gaussian Filtering	w = 3×3	0
	w = 5×5	0
	w = 7×7	0
Histogram Equalization	-	26,61

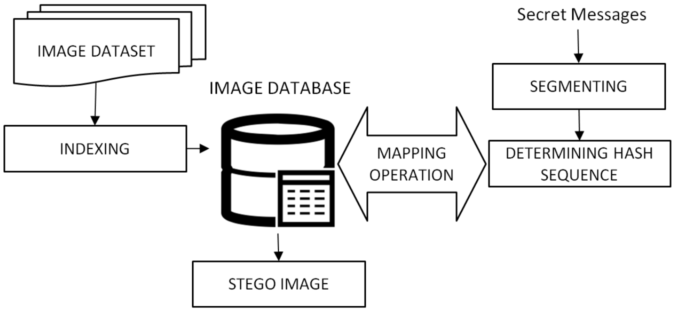


Figure 5: The flowchart of Zou *et al.*'s scheme

2.5 Survey Cao *et al.*'s Scheme

In 2018, Cao *et al.* [4] presented a coverless information hiding based on the molecular structure images of material. This scheme utilizes the average value of the sub-image pixels to represent the secret information, according to the mapping between pixel value intervals and secret information, as shown in Table 5. In addition, a pseudo-random label sequence was used to establish the sub-image location to strengthen the method's security. The Bag of Words Model (BOW) histogram calculates the number of sub-images in a picture that reveals secret information. A multi-level inverted index structure was also created to boost retrieval performance. The flowchart of Cao *et al.*'s scheme [34] is shown in Figure 6.

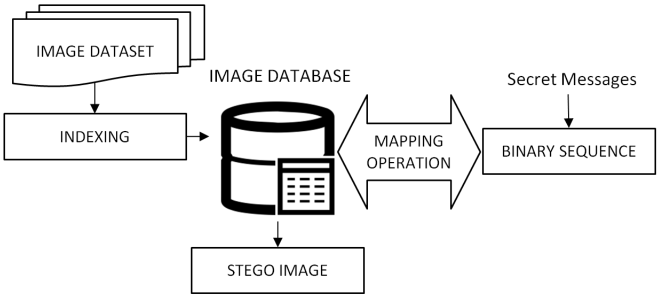


Figure 6: The flowchart of Cao *et al.*'s scheme

Table 5: The mapping relationship of secret information [4]

Pixel value intervals	Binary sequence code
0 ∼ 15	0000
16 ∼ 31	0001
32 ∼ 47	0010
48 ∼ 63	0011
64 ∼ 79	0100
80 ∼ 95	0101
96 ∼ 111	0110
112 ∼ 127	0111
128 ∼ 143	1000
144 ∼ 159	1001
160 ∼ 175	1010
176 ∼ 191	1011
192 ∼ 207	1100
208 ∼ 223	1101
224 ∼ 239	1110
240 ∼ 255	1111

3 Comparisons

In this section, the hiding capacity of five survey papers is summarized in Table 6 and compares the hiding capacity of the survey papers in the number of bits parameter. The hiding capacity is depicted in Figure 7.

Table 6: Hiding capacity comparison

Scheme	Hiding Capacity
Abdulsattar's Scheme [1]	32.736
Yang *et al.*'s Scheme [30]	16.384
Zhang *et al.*'s Scheme [32]	8.193
Zou *et al.*'s Scheme [34]	16.368
Cao *et al.*'s Scheme [4]	4.092

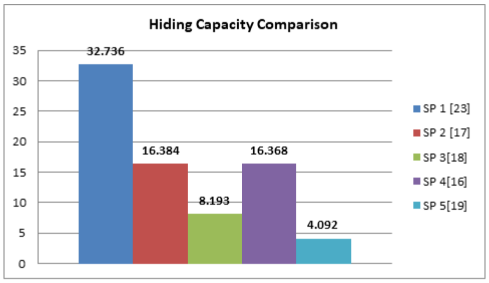


Figure 7: Hiding capacity representation

4 Future Research

According to the studies conducted, the five survey papers are mapping-based CIS. That is, the length of the hash sequence is a crucial component that should be prioritized to improve hiding capacity. Survey paper 1 investigated block size, block arrangement, and overlapping blocks. This will be future research into other properties of an image in expanding the length of the hash sequence. The second option for future CIS study is to devise a hashing algorithm that ensures image dependability against steganalysis tools and image attacks.

In terms of computational difficulty, as the size of the picture database grows, so does the complexity. It is possible to use either an image database or a lookup table to accommodate the secret messages in the basic computation. The secret messages are specified as survey paper 4 implemented a Chinese-based dictionary. It means we should develop specific language-based dictionaries when the secret messages are in the form of other languages. It will be the limit of this scheme and could be a research question for future work.

5 Conclusions

This paper provides a thorough examination of mapping-based coverless image steganography. Because the cover image is not modified, the image quality of the stego image will be the same as the cover, so the quality is optimal. The main concern in mapping-based CIS is hiding capacity. Current mapping-based CIS has limited hiding capacity. Therefore, it is necessary to do further research on how to maximize the image property in secret mapping messages. Overall, the current mapping-based CIS has a high level of robustness against steganalysis tools and image attacks.

Acknowledgments

The Ministry of Science and Technology partially supported this research, Taiwan (ROC), under contract no.: MOST 109-2221-E-468-011-MY3, MOST 108-2410-H-468-023, and MOST 108-2622-8-468-001-TM1.

References

[1] F. S. Abdulsattar, "Towards a high capacity coverless information hiding approach," *Multimedia Tools and Applications*, vol. 80, pp. 18821–18837, 2021.

[2] A. M. Alhomoud, "Image steganography in spatial domain: Current status, techniques, and trends," *Intelligent Automation and Soft Computing*, vol. 27, no. 1, pp. 69–88, 2021.

[3] S. Arunkumar, V. Subramaniyaswamy, V. Vijayakumar, N. Chilamkurti, and R. Logesh, "SVD-based robust image steganographic scheme using RIWT and DCT for secure transmission of medical images," *Measurement*, vol. 139, pp. 426–437, 2019.

[4] Y. Cao, Z. Zhou, X. Sun, and C. Gao, "Coverless information hiding based on the molecular structure images of material," *Computers, Materials and Continua*, vol. 54, no. 2, pp. 197–207, 2018.

[5] X. Chen, Z. Zhang, A. Qiu, Z. Xia, and N. N. Xiong, "Novel coverless steganography method based on image selection and StarGAN," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 1, pp. 219–230, 2022.

[6] I. Gustavo, "Deep learning applied to steganalysis of digital images: A systematic review," *IEEE Access*, vol. 7, 2019.

[7] D. Hu, L. Wang, W. Jiang, S. Zheng, and B. Li, "A novel image steganography method via deep convolutional generative adversarial networks," *IEEE Access*, vol. 6, pp. 38303–38314, 2018.

[8] L. C. Huang, L. Y. Tseng, M. S. Hwang, "The study on data hiding in medical images", *International Journal of Network Security*, vol. 14, no. 6, pp. 301–309, 2012.

[9] D. Kapoor and A. J. Kulkarni, "Improved cohort intelligence — A high capacity , swift and secure approach on JPEG image steganography," *Journal of Information Security and Applications*, vol. 45, pp. 90–106, 2019.

[10] C. Kavitha and K. Sakthivel, "Enhanced least significant bit replacement algorithm in spatial domain of steganography using character sequence optimization," *IEEE Access*, vol. 8, pp. 136537–136545, 2020.

[11] X. Liao, J. Yin, S. Guo, X. Li, and A. Kumar, "Medical JPEG image steganography based on preserving inter-block dependencies," *Computers & Electrical Engineering*, vol. 67, pp. 320–329, 2018.

[12] Q. Liu, X. Xiang, J. Qin, Y. Tan, and Q. Zhang, "A rpbust coverless steganography using camouflage image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 4038–4051, 2022.

[13] X. Liu, Z. Li, J. Ma, W. Zhang, J. Zhang, and Y. Ding, "Robust coverless steganography using limited mapping images," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4472–4482, 2022.

[14] H. C. Lu, Y. P. Chu, M. S. Hwang, "A new steganographic method of the pixel-value differencing", *The Journal of Imaging Science and Technology*, vol. 50, no. 5, pp. 424–426, 2006.

[15] Y. Luo, J. Qin, X. Xiang, and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2779–2791, 2021.

[16] J. Qin, Y. Luo, X. Xiang, Y. Tan, and H. Huang, "Coverless image steganography: A survey," *IEEE Access*, vol. 7, pp. 171372–171394, 2019.

[17] F. Peng, G. Chen, and M. Long, "A robust coverless steganography based on generative adversarial networks and gradient descent approximation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5817–5829, 2022.

[18] A. H. S. Saad, M. S. Mohamed, and E. H. Hafez, "Coverless image steganography based on optical mark recognition and machine learning," *IEEE Access*, vol. 9, pp. 16522–16531, 2021.

[19] S. Q. Saleh, "Digital image steganalysis: Current methodologies and future challenges," *IEEE Access*, vol. 10, no. August, pp. 92321–92336, 2022.

[20] J. Sharafi, Y. Khedmati, and M. M. Shabani, "Image steganography based on a new hybrid chaos map and discrete transforms," *Optik*, vol. 226, no. P2, p. 165492, 2021.

[21] N. Subramanian and O. Elharrouss, "Image steganography: A review of the recent advances," *IEEE Access*, vol. 9, pp. 23409–23423, 2021.

[22] G. Swain, "Adaptive and non-adaptive PVD steganography using overlapped pixel blocks," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 7549–7562, 2018.

[23] A. O. Vyas and S. V Dudul, "A novel approach of object oriented image steganography using LSB," in *Proceedings of the 1st International Conference on*

Data Science, Machine Learning and Applicationsin (ICDSMLA'19), pp. 144–151, 2019.

[24] Y. L. Wang, J. J. Shen, M. S. Hwang, "An improved dual image-based reversible hiding technique using LSB matching", *International Journal of Network Security*, vol. 19, no. 5, pp. 858–862, 2017.

[25] Y. L. Wang, J. J. Shen, M. S. Hwang, "A novel dual image-based high payload reversible hiding technique using LSB matching", *International Journal of Network Security*, vol. 20, no. 4, pp. 801–804, 2018.

[26] C. C. Wu, M. S. Hwang, S. J. Kao, "A new approach to the secret image sharing with steganography and authentication", *The Imaging Science Journal*, vol. 57, no. 3, pp. 140–151, 2009.

[27] C. C. Wu, S. J. Kao, and M. S. Hwang, "A high quality image sharing with steganography and adaptive authentication scheme", *Journal of Systems and Software*, vol. 84, no. 12, pp. 2196–2207, 2011.

[28] C. C. Wu, S. J. Kao, W. C. Kuo, M. S. Hwang, "Enhance the image sharing with steganography and authentication," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 1177-1181, 2008.

[29] N. I. Wu, M. S. Hwang, "A novel LSB data hiding scheme with the lowest distortion", *The Imaging Science Journal*, vol. 65, no. 6, pp. 371–378, 2017.

[30] L. Yang, H. Deng, and X. Dang, "A novel coverless information hiding method based on the most significant bit of the cover image," *IEEE Access*, vol. 8, pp. 108579–108591, 2020.

[31] H. Zhang and L. Hu, "A data hiding scheme based on multidirectional line encoding and integer wavelet transform," *Signal Processing: Image Communication*, vol. 78, pp. 331–344, 2019.

[32] X. Zhang, F. Peng, and M. Long, "Robust coverless image steganography based on DCT and LDA topic classification," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3223–3238, 2018.

[33] Z. Zhou, H. Sun, R. Harit, X. Chen, and X. Sun, "Coverless image steganography without embedding," in *International Conference on Cloud Computing and Security (ICCCS'15)*, vol. 1, pp. 123–132, 2015.

[34] L. Zou, J. Sun, M. Gao, W. Wan, and B. B. Gupta, "A novel coverless information hiding method based on the average pixel value of the sub-images," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 7965–7980, 2019.

Biography

Kurnia Anggriani received BS degree in Informatics from University of Bengkulu, Indonesia in 2011, and the MS degree in Informatics from Bandung Institute of Technology, Indonesia in 2014. Currently she is taking Ph.D degree in Asia University, Taiwan. Her current research interests include steganography and image processing.

Nan-I Wu received a Ph.D. degree in the Institute of Computer Science and Engineering from Nation Chung Hsing University (NCHU), Taichung, Taiwan, in 2009. From 2010 to 2011, he was a post-doctoral research fellow at the Academia Sinica Institute of information science. He was an assistant professor at the Department of Animation and Game Design, TOKO University (Taiwan), during 2011-2018 and an associate professor during 2018-2019. Now he is an associate professor at the Department of Digital Multimedia, Lee-Ming Institute of Technology (Taiwan) since 2019 and also the Director of the eSports Training Centre since 2020. His current research interests include game design, eSports training/magagement, multimedia processing, multimedia security, data hiding, and privacy-preserving. He published more than 10 international journal papers (SCI) and conference papers.

Min-Shiang Hwang received M.S. in industrial engineering from National Tsing Hua University, Taiwan in 1988, and a Ph.D. degree in computer and information science from National Chiao Tung University, Taiwan, in 1995. He was a distinguished professor and Chairman of the Department of Management Information Systems, NCHU, during 2003-2011. He obtained 1997, 1998, 1999, 2000, and 2001 Excellent Research Awards from the National Science Council (Taiwan). Dr. Hwang was a dean of the College of Computer Science, Asia University (AU), Taichung, Taiwan. He is currently a chair professor with the Department of Computer Science and Information Engineering, AU. His current research interests include information security, electronic commerce, database, data security, cryptography, image compression, and mobile computing. Dr. Hwang has published over 300+ articles on the above research fields in international journals.

Research on Data Mining Detection Algorithms for Abnormal Data

Yanying Yang

(Corresponding author: Yanying Yang)

Information Technology College, Nanjing Forest Police College
No. 28, Wenlan Road, Xianlin University Town, Qixia District, Nanjing, Jiangsu 210023, China
Email: yi59204@163.com

(Received Sept. 4, 2019; Revised and Accepted Dec. 5, 2022; First Online Dec. 10, 2022)

Abstract

Adequate protection against malicious attacks is required to enhance the security of the Internet. This paper briefly introduced a data mining algorithm for network anomaly data detection, i.e., the K-means clustering algorithm. The detection performance of the K-means algorithm was improved by introducing the density-based spatial clustering of applications with noise (DBSCAN) algorithm and adjusting the number of clustering centers autonomously with standard deviation and cross-entropy. Simulation experiments compared the optimized K-means algorithm with support vector machine (SVM) and traditional K-means algorithms in MATLAB software. It was found that the optimized K-means algorithm had the best detection performance and the least detection time among the three abnormal data detection algorithms; the performance of the optimized K-means algorithm decreased as the proportion of abnormal data in the detected data increased, but it remained to be the best.

Keywords: Abnormal Data; Data Mining; Internet; K-Means

1 Introduction

The emergence of the Internet has greatly facilitated people's lives, and people can get the information they need anytime and anywhere through the Internet [8]. Due to the Internet's openness, criminals can also launch malicious attacks on normal users through the Internet. In this process, they can obtain users' private information or cause damage to the Internet structure, directly affecting the Internet experience of normal users [12].

In the face of malicious attacks from criminals, traditional Internet protection means is the firewall, whose protection principle is isolation, i.e., in the face of data from outside the wall, whether the data have the firewall pass is determined rather than whether the data are abnormal. However, as the size of the Internet expands, the amount of data traffic in the network increases dramatically, and the rigid judgment of the firewall greatly limits the speed of data transmission [11].

An intrusion detection system is a kind of Internet protection means to detect abnormal data actively. Compared with a firewall, this protection means does not directly intercept data but actively detects the transmitted data to find abnormal data and intercept it, so it has less impact on the normal data transmission speed and can intercept the malicious data more precisely.

Li *et al.* [7] proposed a combined data structure based on the principle of cognitive computing and various pruning strategies based on cognitive computing and proposed a new intrusion detection model based on the above theory. They found through experiments that the algorithm model had good recognition performance.

Yin *et al.* [14] proposed a Hadoop distributed file system (HDFS) and deep neural network-based MapReduce anomaly data mining and detection algorithm and verified the effectiveness of the algorithm through experiments.

Cheng *et al.* [2] designed an independent component analysis-based anomaly data mining algorithm and found that the ICA algorithm had more prominent advantages and better detection performance compared with the birds swarm algorithm and the KLE algorithm. This paper briefly introduced the K-means clustering algorithm, a data mining algorithm for network anomaly data detection.

The detection performance of the K-means algorithm was improved by introducing the density-based spatial clustering of applications with noise (DBSCAN) algorithm and adjusting the number of clustering centers autonomously with standard deviation and cross-entropy. Finally, the improved K-means algorithm was compared with support vector machine (SVM) and traditional K-means algorithms in MATLAB software.

2 Detection of Abnormal Traffic Data Based on Data Mining

The K-means algorithm determines the cluster center by the mean value of the data in the cluster and divides the data by comparing the distance between the data and the cluster center [10]. It has a simple calculation principle and a high efficiency, which is suitable for processing big data on the network; however, the K-means algorithm needs to determine the K value first. The choice of the K value will directly affect the clustering results. In addition, isolated points [1] and noise in the data will also lead to serious bias when the algorithm reselects the clustering centers [9]. In order to solve the above problems, the traditional K-means algorithm was improved, and the flow of the improved K-means algorithm is shown in Figure 1.

1) The input data are preprocessed.

2) Isolated points in the dataset are scanned [13] and classified as abnormal data.

3) After eliminating the isolated points, K clustering centers are selected from the remaining data [5].

4) The data are assigned according to the proximity principle [4], and the distance is calculated by the following formula:

$$d_{a,b} = \sqrt{\sum_{i=1}^{o} (a_i - b_i)^2}$$

where $d_{a,b}$ is the distance between data a and b; a_i and b_i are the i-th dimensional feature vector of data a and b, respectively, and o is the dimensionality of the feature vector.

5) The standard deviation of every cluster is calculated. If the standard deviation of all clusters exceeds the preset threshold, it goes to Step 6; if the standard deviation of all clusters does not all exceed the preset threshold, the clusters whose standard deviation exceeds the preset threshold will generate two new clustering centers [6], and the selection criteria for the new clustering centers are the two data within the cluster closest to the original clustering center. Then, it returns to Step 4.

6) The cross-entropy value of any two clusters is calculated using the following formula:

$$\begin{aligned} D_R &= -\log(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} G(x_i - y_j, \sigma^2)) \\ G(x_i - y_j, \sigma^2) &= e^{(-|x-y|^2)/2\sigma^2} \end{aligned}$$

where D_R is the cross-entropy; $G(\cdot)$ is the Gaussian kernel function; m is the number of all data in one class; n is the number of all data in another class; x is the set of data of one class; y is the set of data of another class, and σ^2 is the variance of the Gaussian function.

7) Whether the cross-entropy values of any two clusters exceed the preset threshold is determined. If both exceed the preset threshold, the cluster center of every cluster is output, and it goes to Step 8. If there are cross-entropy values of any two clusters within the preset threshold, the two clusters are merged to recalculate the new cluster center before returning to Step 6.

8) Based on the calculated clustering centers, traditional K-means clustering is performed on the data to be detected. The iteration stops until the termination condition is reached. The percentage of every cluster in the total data volume is calculated. The clusters whose percentage is less than the preset threshold are identified as abnormal data.

3 Simulation Experiments

3.1 Experimental Data

The simulation experiments were conducted in a laboratory server using MATLAB software [15]. The data used in the simulation experiments were the KDD99 dataset [3], and 4000 normal data, 500 denial-of-service (DOS) data, 500 remote-to-local (R2L) data, 500 user-to-root (U2R) data, and 500 probe intrusion data were randomly selected from the KDD99 dataset.

3.2 Experimental Setup

Before using the dataset for simulation experiments, the data in the dataset were preprocessed first. The data in the KDD99 dataset was 42-dimensional, and the first 41 dimensions were the feature dimensions; however, the span between the feature values in different dimensions was large. Three-dimensional features were character features, which needed to be transformed into digital features, and then the feature sparsity caused by the large span of feature values was reduced by normalization. The normalization formula is:

$$z' = \frac{z - \bar{z}}{z_{var}}$$

where z' is the normalized data; z is the data to be normalized; $\bar{z}$ is the mean of the data; and z_{var} is the variance of the data.

The relevant parameters of the improved K-means clustering algorithm are as follows. The initial K value was 6. The standard deviation threshold was 2.1. The cross-entropy threshold was 0.4. The maximum number of clustering iterations was 1000.

In order to verify the performance of the improved K-means clustering algorithm, the simulation experiment was performed, and the traditional K-means clustering

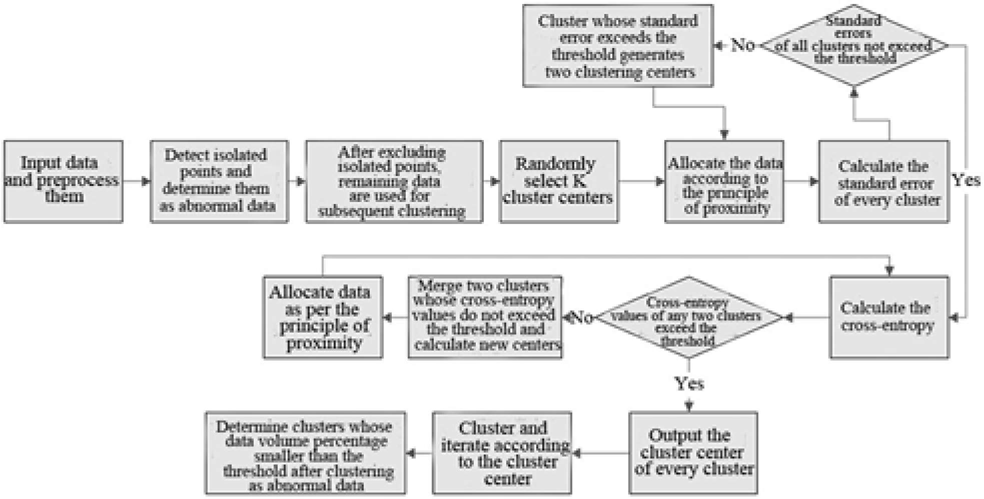

Cluster whose standard error exceeds the threshold generates two clustering centers
No
Standard errors of all clusters not exceed the threshold
Input data and preprocess them
Detect isolated points and determine them as abnormal data
After excluding isolated points, remaining data are used for subsequent clustering
Randomly select K cluster centers
Allocate the data according to the principle of proximity
Calculate the standard error of every cluster
Yes
Allocate data as per the principle of proximity
Merge two clusters whose cross-entropy values do not exceed the threshold and calculate new centers
No
Cross-entropy values of any two clusters exceed the threshold
Calculate the cross-entropy
Yes
Determine clusters whose data volume percentage smaller than the threshold after clustering as abnormal data
Cluster and iterate according to the cluster center
Output the cluster center of every cluster

Figure 1: Basic flow of the improved K-means algorithm

algorithm and the SVM algorithm were also simulated. This paper also tested the three abnormal data detection algorithms by the testing data set with 2% 20% of abnormal data in addition to the normal and abnormal data sets containing constant numbers of data.

3.3 Evaluation Criteria

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

$$FAR = \frac{FN}{TP+FN}$$

$$DR = \frac{TP}{TP+FP}$$

where ACC, FAR, and DR are the accuracy, false alarm rate, and precision, respectively. TP is the number of attacks classified as attacks; TN is the number of normal data classified as normal data; FP is the number of attacks classified as normal data; and FN is the number of normal data classified as attacks.

In addition to the above evaluation criteria, this paper also used the area under the curve (AUC) to measure the detection performance of the algorithm. AUC is the area under the receiver operating characteristic (ROC) curve. Here, the ROC curve is the curve of an abnormal detection algorithm whose horizontal and vertical coordinates are the false and true positive rates under different abnormal data recognition thresholds, and a point on the ROC curve represents the true and false positive rates under a recognition threshold. Generally, we can compare the algorithms by comparing whose curve is higher. However, if the ROC curves of different algorithms cross, it is difficult to judge them, so AUC under the ROC curve is used as the basis of comparison. The calculation formula of AUC is:

$$AUC = \frac{\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})}{2}$$

where m is the number of thresholds selected for plotting the ROC curve; x_i is the false positive rate of the algorithm under the i-th threshold; and y_i is the true positive rate of the algorithm under the i-th threshold.

3.4 Experimental Results

The recognition performance of the three abnormal data detection algorithms for the normal and abnormal data sets containing fixed numbers of data is shown in Figure 2. It was seen from Figure 2 that the ACC, FAR, and DR of the SVM algorithm was 76.35%, 12.35%, and 72.36%, respectively; the ACC, FAR, and DR of the traditional K-means algorithm were 88.74%, 9.86%, and 87.41%, respectively; the ACC, FAR, and DR of the improved K-means algorithm were 97.15%, 5.26%, and 94.55%, respectively. The change of the broken line in Figure 2 intuitively showed that the improved K-means algorithm had the best performance in identifying abnormal data in the dataset, the traditional K-means algorithm was the second best, and the SVM algorithm was the worst.

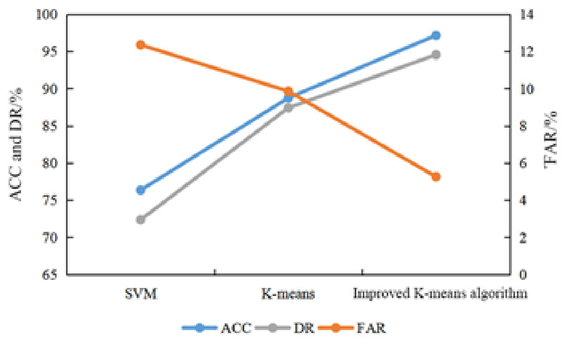


Figure 2: Recognition performance of three anomalous data detection algorithms

The previous section examined the detection performance of the three abnormal data detection algorithms when facing data sets with fixed sizes, but in the actual application, the proportion of abnormal data in the network will change at any time, so it is necessary to ensure that the abnormal data detection algorithms can also have relatively good detection performance when facing data sets with different proportions of abnormal data. Figure 3 shows the variation of the AUC of three abnormal data detection algorithms when facing data sets with different proportions of abnormal data.

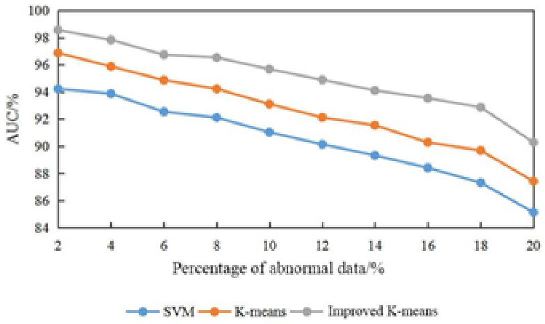


Figure 3: The AUC of three abnormal data detection algorithms in the face of different proportions of abnormal data

It was noticed in Figure 3 the AUC of all three abnormal data detection algorithms tended to decrease as the proportion of abnormal data in the test dataset increased, but they were all above 84%, which was within the usable range. Under the same percentage of abnormal data, the AUC of the improved K-means algorithm was the largest, the traditional K-means algorithm was the second, and the SVM algorithm was the smallest.

Figure 4 shows the time required for data mining of abnormal data by the three abnormal detection algorithms, 5.73 s by the SVM algorithm, 4.36 s by the traditional K-means algorithm, and 3.39 s by the improved K-means algorithm. It was seen from Figure 2 that when mining abnormal data in the network, the SVM algorithm took the longest time, the traditional K-means algorithm took the second longest time, and the improved K-means algorithm took the least time.

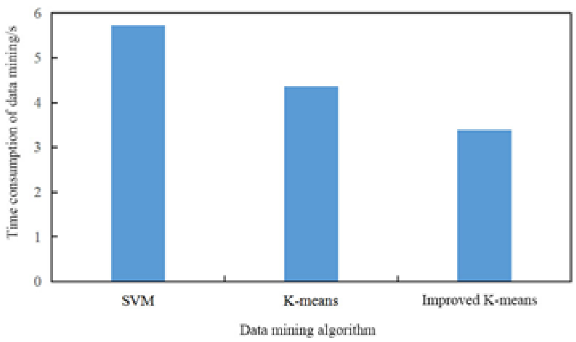


Figure 4: Time consumption of three abnormal data detection algorithms for data mining

4 Discussion

While the Internet provides convenience for ordinary users, it also provides an open platform for lawbreakers, who launch malicious attacks on ordinary users to obtain private and confidential information. In order to protect the private information of users on the Internet and improve the experience of using the Internet, it is necessary to make effective protection against malicious attack data. A firewall is a traditional means of protection, and its principle is to brutally block all data and allow only specific data to pass through. However, with the expansion of the Internet, network traffic increases, so the passive protection of firewalls is no longer effective. In addition, its brutal interception will affect the efficiency of network data transmission. When an intrusion detection system protects the network, it does not intercept all data but actively detects network data to identify abnormal data and intercept it.

The K-means algorithm studied in this paper is an intrusion detection algorithm. The K-means algorithm aggregates similar data into clusters based on the characteristics of network data and identifies abnormal data based on the proportion of the data in the cluster to the total data volume. In order to further strengthen the detection performance of the K-means algorithm, standard deviation and cross-entropy were introduced to split and aggregate clusters so that the clustering algorithm could set the appropriate number of cluster centers independently, and the DBSCAN algorithm was used to remove the interference of isolated points. Finally, SVM, traditional K-means, and improved K-means algorithms were compared in simulation experiments, and the results are shown above. The detection performance of the improved K-means algorithm was the best, the traditional K-means was the second best, and the SVM algorithm was the worst, both for fixed-size data sets and for data sets with varying proportions of abnormal data. The time spent by the improved K-means algorithm in mining abnormal data was the shortest, and the time spent by the SVM algorithm was the longest. The reason for the above results is as follows. Although the SVM algorithm is a supervised classification algorithm, it was difficult to fit the nonlinear law in the training process effectively. The traditional K-means algorithm classified data based on the similarity of features between data, but isolated points in the data set caused a serious shift of cluster centers in the iteration process, and the setting of the number of cluster centers depended on experience. The improved K-means algorithm used the DBSCAN algorithm to remove the isolated points first and then split and aggregated the clusters by standard deviation and cross-entropy to obtain the appropriate number of cluster centers, so it performed better in recognition.

5 Conclusion

This paper briefly introduced the K-means clustering algorithm, a data mining algorithm, for abnormal network data detection and improved the detection performance of the K-means algorithm by introducing the DBSCAN algorithm and using standard deviation and cross-entropy to adjust the number of clustering centers autonomously. The improved K-means algorithm was compared with SVM and traditional K-means algorithms in MATLAB software. The following results are obtained. Facing datasets containing fixed numbers of data, the improved K-means algorithm had the best performance in identifying abnormal data in the dataset, the traditional K-means algorithm was the second-best, and the SVM algorithm was the worst. As the proportion of abnormal data in the dataset increased, the recognition performance of all the data detection algorithms tended to decline, but under the same proportion of abnormal data, the improved K-means algorithm had the best recognition performance, the traditional K-means algorithm was the second-best, and the SVM algorithm was the worst. When mining abnormal data, the SVM algorithm took the longest time, the traditional K-means algorithm took the second-longest, and the improved K-means algorithm took the least time.

References

[1] N. Chaabene, A. Bouzeghoub, R. Guetari, H. B. Ghezala, "Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data: A survey," *Multimedia Systems*, vol. 28, pp. 2133–2143, 2022.

[2] J. Cheng, X. Mai, S. Wang, "Research on abnormal data mining algorithm based on ICA," *Cluster Computing*, vol. 22, no. 4, pp. 3613-3619, 2019.

[3] N. O. F. Elssied, O. Ibrahim, A. H. Osman, "Enhancement of spam detection mechanism based on hybrid k-mean clustering and support vector machine," *Soft Computing*, vol. 19, no. 11, pp. 3237-3248, 2015.

[4] X. Hao, X. Zhang, "Research on abnormal detection based on improved combination of K - means and SVDD," in *IOP Conference Series: Earth & Environmental Science*, pp. 1-6, 2018.

[5] S. H. Kang, K. J. Kim, "A feature selection approach to find optimal feature subsets for the network intrusion detection system," *Cluster Computing*, vol. 19, no. 1, pp. 325-333, 2016.

[6] W. Laftah Alyaseen, Z. Ali Othman, M. Z. Ahmad Nazri, "Hybrid modified K-means with C4.5 for intrusion detection systems in multiagent systems," *Scientific World Journal*, vol. 2015, no. 2, pp. 1-14, 2015.

[7] J. Li, W. Cao, J. Huang, "An intrusion detection algorithm based on data streams mining and cognitive computing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2020, no. 9, pp. 1-14, 2020.

[8] M. A. Lili, J. Liu, "Research on abnormal data detection of optical fiber communication network based on data mining," *Journal of Applied Optics*, vol. 41, no. 6, pp. 1305-1310, 2020.

[9] A. Lishchytovych, V. Pavlenko, A. Shmatok, Y. Finenko, "Comparative analysis of system logs and streaming data anomaly detection algorithms," *Information Systems and Technologies Security*, vol. 1, no. 2, pp. 5-7, 2020.

[10] V. Saranya, R. Umagandhi, "A comparative study of outlier detection in large-scale data using data mining algorithms," *International Journal of Data Mining and Emerging Technologies*, vol. 7, no. 1, pp. 10-15, 2017.

[11] S. M. Shareef, S. H. Hashim, "Proposed hybrid classifier to improve network intrusion detection system using data mining techniques," *Engineering and Technology Journal*, vol. 38, no. 1, pp. 6-14, 2020.

[12] X. Sun, "Similarity detection method of abnormal data in network based on data mining," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 4, pp. 1-8, 2019.

[13] H. Wen, "A new algorithm based on artificial intelligence to realize rapid detection and recognition of mass data abnormal point," *Boletin Tecnico/Technical Bulletin*, vol. 55, no. 6, pp. 364-371, 2017.

[14] C. Yin, C. Pan, P. Zhang, "Deep neural network combined with MapReduce for abnormal data mining and detection in cloud storage," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2020, no. 5, pp. 1-12, 2020.

[15] Y. Zhang, K. Wang, M. Gao, Z. Quyang, S. Chen, "LKM: a LDA-based K-means clustering algorithm for data analysis of intrusion detection in mobile sensor networks," *International Journal of Distributed Sensor Networks*, no. 3, pp. 1-11, 2015.

Biography

Yanying Yang, born in October 1973, received the master's degree of engineering from Shandong Normal University in July 1996. She is an associate professor in Nanjing Forest Police College. She is interested in database, data mining, and big data application.

An Approach for Security Assessment of the Internet of Things in Healthcare for the Disabled

Reem Almasoudi, Mohammad Arafah, Waleed Alghanem, and Saad Bakry
(Corresponding author: Reem Almasoudi)

Computer Engineering, College of Computer and Information Sciences, King Saud University
Riyadh, Kingdom of Saudi Arabia
Email: 436204287@student.ksu.edu.sa

(Received Jan. 19, 2022; Revised and Accepted Dec. 11, 2022; First Online Dec. 13, 2022)

Abstract

The Internet of Things (IoT) is increasingly important for applications in many fields, including healthcare. One application in this field is concerned with using IoT to protect the disabled. This paper is concerned with developing an approach for assessing IoT security of information for the disabled. The approach is based on identifying key related security issues and describing their interrelationships for the assessment. The issues involved are concerned with: IoT assets; threats to these assets; protection controls that reduce the impact of threats; and security performance, including confidentiality, integrity, and availability. For flexibility and updatability in applying the approach to case studies, the approach constructs the issues according to the five-domain broad-scope structured view of "Strategy, Technology, Organization, People, and Environment: STOPE". The use of the approach is illustrated through an assessment of a related case study. The outcome of the work provides a base for assessing other case studies and future useful research extensions.

Keywords: Healthcare; Internet of Things (IoT); Information Security;; Structured Approach

1 Introduction

The International Telecommunication Union (ITU) defines the Internet of Things (IoT) as follows: "a global infrastructure for the information society, enabling advanced services by interconnecting physical and virtual things based on existing and evolving interoperable Information and Communication Technology (ICT)" [22]. IoT attracted many fields of application. Eight major fields of IoT application have been identified by the Institute of Electrical and Electronic Engineers (IEEE), and these are listed in Table 1 [10]. An important field in this respect is "healthcare"; and this paper is concerned with IoT information security in healthcare, specifically with regards to the disabled. For introducing the work presented in this paper, the IoT architecture is briefly described, and a literature review associated with the work is briefly addressed. This is followed by identifying the work presented in this paper.

Table 1: The major fields of IoT applications

FIELD	USE
Healthcare	Hospitals, Doctors & Patients (Disabled)
Homes/Buildings	Appliances Providers/Facility Management
Retails	Retails Stores/Application Developers
Energy	Utilities
Manufacturing	Manufacturing Industries/Automation, Equipment Providers
Transportation	Public Transportation/City Authorities
Logistics	Regulators/Logistics Companies
Media	Information and Communication Technology Infrastructure Providers

1.1 IoT Architecture

The IoT architecture has three main layers, as shown in Figure 1, and these are identified in the following [39].

1) "Perception layer", which is also known as the physical layer. It is associated with different devices including sensors; Radio Frequency Identification (RFID) Tags; and various other smart devices. It gathers data and send it to the "network layer".

2) The "network layer" receives the data; process it; and communicates with the application layer.

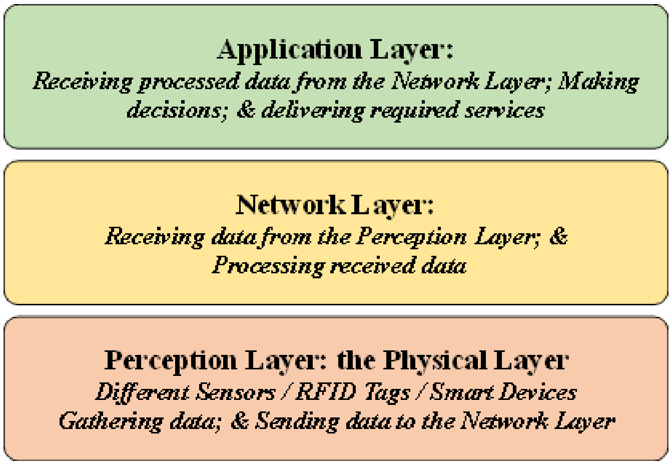

Application Layer:
Receiving processed data from the Network Layer; Making decisions; & delivering required services
Network Layer:
Receiving data from the Perception Layer; & Processing received data
Perception Layer: the Physical Layer
Different Sensors / RFID Tags / Smart Devices
Gathering data; & Sending data to the Network Layer

Figure 1: IoT layered architecture

3) The "application layer" is where the decisions are made; and the required services are delivered.

1.2 Literature Review

The available literature associated with the addressed problem can be viewed as concerned with the following two main sources: information security standards published by specialized international organizations; and work published by researchers. These sources are addressed in the following.

International standards on information security include generic documents that can be of general use in various information security applications, such as: International Standards Organization (ISO) ISO 27002, and ISO 27005 [18, 19]. They also include documents concerned with information security in health care, such as ISO 27799 [20]. Standard documents specifically on IoT are mainly of tutorial nature, such as the documents of ITU [22] and ISO-IEC [21]. While all these documents do not directly apply to the addressed problem, they provide useful source of information that supports its general requirements.

Research papers related to the addressed problem can be viewed as associated with the following four main dimensions.

1) The first is concerned with healthcare systems servicing patients. An important contribution in this respect is the paper by [30], which provides a state of the art review of different techniques proposed for systems aiding disabled people. The paper is useful in identifying the needed assets for the disabled.

2) The second is associated with healthcare security and privacy. In this regard, one paper [7] addresses classification of mobile-health issues concerned with: usability, security and privacy; while targeting: confidentiality, integrity and availability.

3) The third is specifically concerned with IOT security and privacy, which is addressed by paper [17]. The paper reviews and compares various IoT security frameworks considering the requirements of the international standard ISO 27001.

4) The fourth is related to IoT security and privacy in health care, which is addressed by the thesis [2]. This thesis identifies different types of privacy and security risks and emphasizes various measures for their protection.

The literature review given above provides useful sources of information for IoT security and privacy protection in health care for the disabled. One missing gap in this respect is the existence of a sound recognized approach for the assessment of IoT security protection in health care for the disabled that can integrate and assess the various issues involved, with flexibility toward future development.

1.3 The Given Work

The work presented in this paper is concerned with closing the above gap and developing the targeted sound assessment approach. For this purpose, the work involves the following.

1) Developing the targeted approach for the assessment of IoT protection in healthcare for the disabled.

2) Using the developed approach to address a specific case study concerned with IoT protection in health for the disabled.

3) Evaluating the outcome and concluding recommendation.

2 Approach Development

The development of the assessment approach is addressed here. Four main steps are considered. The first is concerned with introducing the principles upon which the assessment is based. The second is associated with introducing key "assets" concerned with IoT in healthcare for the disabled. The third is related to identifying key "threats" to these assets; and the fourth is concerned with the "protection controls" that eliminate or mitigate the impact of these threats on the key assets.

2.1 Development Principles

The assessment approach is based on the following principles.

1) The first principle is concerned with the need to identify the main issues involved in the problem. These include: "assets" that needs to be protected; "threats" to assets; and protection "controls" that reduce the impact of threats.

2) The second principle is related to constructing the above according to a well-structured framework for the purpose of providing flexible assessment of various case-studies and enabling easy updates. In this respect, the chosen structure is that of the (STOPE) framework, which provides a wide-scope view of the problem considered through its five broad domains of "strategy, technology, organization, people, and environment: (STOPE)". This framework has been previously used for the investigation of various technology and society problems, including problems concerned with information security [3].

3) The third principle considers the "strategy (S)" domain to target the achievement of "confidentiality, integrity and availability of information (CIA)". Confidentiality is concerned with protecting information from unauthorized users. Integrity ensures the safety of information from any loss or contamination; while "availability", is associated with the protection of timely access to information.

4) The fourth principle is related to the central issues of the problem that is the "assets". They will be distributed among the "(TOPE) four domains". Of course, the threats and the protection controls will be related to the "assets"; and their impacts will affect (CIA). Therefore, all issues of the problem will be organized around the assets.

5) The fifth principle considers the "assets" concerned with the "technology (T)" domain to be structured according to the IoT layers of "perception, network and application" shown in Figure 1, and this will lead to further enhancement of flexibility.

6) The sixth principle is concerned with providing assessment "scales". These scales are needed for assessing: the level of "importance of assets"; the level of "impact of threats" on assets; the level of "impact reduction" provided by protection controls; and the level of the state of "confidentiality, integrity and availability". The principle considers that a "unified scale should be given to these four types of assessment levels.

Figure 2 illustrates the structure of the main issues of the approach, and Table 2 shows the "unified assessment scale.

2.2 IoT Healthcare Assets for the Disabled

The key assets of the IoT healthcare applications to the disabled are given in Table 3. They include assets concerned with the "technology" domain and associated with the three IoT layers shown in Figure 1. Assets concerned with "organization, people and environment" domains are also given. All given assets are of "essential" or of "high importance" depending on the case under consideration.

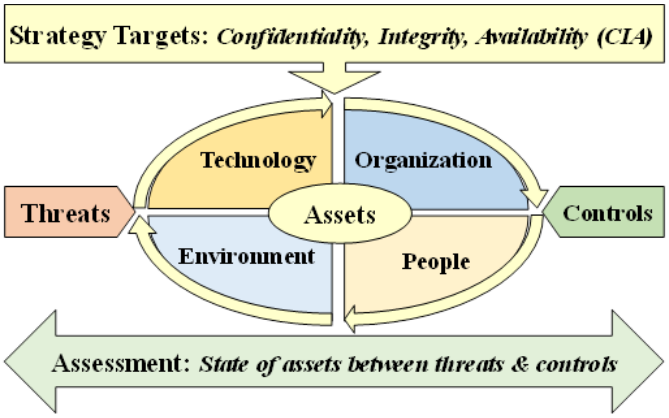

Strategy Targets: Confidentiality, Integrity, Availability (CIA)
Technology
Organization
Threats
Assets
Controls
Environment
People
Assessment: State of assets between threats & controls

Figure 2: The main issues of the assessment approach

2.3 Threats to IoT Assets

The key IoT assets of Table 3 face different threats. Table 4 gives key threats challenging these assets. In addition, the table also provides a brief identification of these threats, considering their "cause and effect". Each threat would have impacts, of different levels, on one or more assets; and this will reduce the level of the state of "confidentiality, integrity, and availability" of the IoT. This is addressed later in the "assessment of the case study".

2.4 Protection Controls

For eliminating or mitigating the impact of the threats of Table 4, on the assets of Table 3, Table 5 presents a set of protection controls that can be used for this purpose. Like the fact that each threat can have different levels of impact on one or more assets, each control here can lead to different reduction levels of impact caused by one or more threats. The outcome of this will result in promoting the state of "confidentiality, integrity, and availability", of course at the cost of these controls.

The "assets, threats, and protection controls" presented in this section for the IoT healthcare applications to the disabled provide a "base" that enables the assessment of different case-studies according to the assessment principles given above. A case-study that illustrates this is addressed in the coming section. The case considers specific "healthcare applications concerned with disabled".

3 A Case–Study

IoT on Android for Neuromotor Disabled Patients The case study addressed here considers an "IoT application to Android concerned with interacting with neuromotor disabled patients" [28]. Neuromotor disabilities are conditions of the nervous system, with motor deficits being their defining feature. Android is used here as a "medical end-device". In this section, the case-study is first introduced. This is followed by assessing security, according

Table 2: The unified assessment scale

Scale	Asset Importance	Impact of Threats	Impact Reduction of Threats	State of C, I & A
3	Essential	High	Prevention	No Loss
2	High importance	Moderate	Mitigation	Low Loss
1	Medium importance	Limited	Detection	Moderate Loss
0	Low importance	None	None	Complete Loss

Table 3: Key “assets” of IoT healthcare applications to the disabled

Symbol	Asset	Description
ATP (1)	Servers	Technology: Perception Layer
ATP (2)	End devices: Smart phones/Laptops/Tablets	
ATP (3)	Sensors: Wearable/Non-Wearable	
ATP (4)	Cameras	
ATP (5)	Medical devices: Pacemaker/Others	
ATP (6)	Radio Frequency Identification (RFID) Tags (attached to objects for tracking)	
ATP (7)	Special glasses	
ATN (1)	Networks: Wire/Wireless	Technology: Network Layer
ATA (1)	Applications	Technology: Application Layer
ATA (2)	Databases	
ATA (3)	Directories: Contacts	
ATA (4)	Images: X ray/others	
ATA (5)	Vital signs: Temperature/Heart rate/Respiratory rate/Blood pressure/others	
ATA (6)	Signals: Electrocardiogram (ECG)/others	
AO (1)	Organization’ data	Organization
AO (2)	Organization’ reputation	
AP (1)	Patient’s data	People
AP (2)	Patient’s trust	
AE (1)	Power supply	Environment
AE (2)	Air conditioning	
AE (3)	Signal propagation environment	

Table 4: Key “threats” to IoT “assets” concerned with healthcare applications to the disabled

Symbol	Threat	Description
TTP (1)	Node Capture	Causing fatal problems to the network. Capturing a targeted node by different ways including sending many requests to that node
TTP (2)	Malicious node injection	Gaining an unauthorized access to an IoT network by inserting a malicious node.
TTP (3)	Jamming	Transmitting signals that disrupt communications by decreasing the Signal-to-Inference ratio.
TTN (4)	Wormholes	Disrupting network function by creating virtual tunnel between nodes.
TTN (5)	Sybil	Disrupting network function by creating multiple identities in the network.
TTN (6)	Man-In-The-Middle (MITM)	Intercepting and contaminating data flow in different ways.
TTN (7)	Sinkhole	Misdirecting packets through a malicious node.
TTN (8)	Hello flood	Harming network function by sending “hello packets” to nodes
TTN (9)	Traffic analysis	Gathering network information & identifying important nodes using traffic analyzers (Sniffers)
TTN (10)	Desynchronization	Damaging communication protocol by changing the sequen of packets.
TTA (1)	Denial of Service (DoS)	Preventing users from access. Caused by flooding with traffic
TTA (2)	Structured Query Language (SQL) injection	Accessing database illegally. Caused by injecting malicious code into SQL.
TTA (3)	Cross-Site Scripting (XSS)	Controlling web applications by injecting malicious script the web browsers
TTA (4)	Malware	Harming system function by contaminating software.
TTA (5)	Phishing	Stealing & misusing confidential information (login/credit card/others) through social engineering & e-mail.
TO (1)	Legal issues	Violating laws on data security & privacy.
TE (1)	Power outage	May happen accidently or maliciously

Table 5: Key “protection controls” for IoT “assets” concerned with healthcare applications to the disabled

Symbol	Protection Measures	Description
P (1)	Cryptography	Data encryption-decryption.
P (2)	Authentication	Data certification & validation.
P (3)	Intrusion Detection System (IDS)	Monitoring & detecting attacks.
P (4)	Firewalls	Monitoring and controlling traffic.
P (5)	Protection Software	Any software that provides protection to devices against threats.
PO (1)	Security policies	Rules & practices for the confidentiality, integrity & availability of information.
PE (1)	Uninterruptible Power Supply (UPS)	To ensure continuous operation.

to the above approach. Each asset in the case-study is assessed individually considering threats and protection controls. In addition, collective consideration of all assets is also addressed.

3.1 Introducing the Case-Study

The case considered is described in the following in terms of its key "assets; threats and protection controls".

1) The case study has "six" key "assets" representing a subset of the assets given in Table 3. They involve: a network (wireless body area network): ATN (1); a server: ATP (1); an end-device: ATP (2); sensors: ATP (3); applications (Android): ATA (1); and database: ATA (2).

2) It has "eleven" threats to the key assets; and these are: TTN (4), TTN (5), TTN (6), TTN (7), TTA (1), TTA (2), TTA (3), TTA (4), TTA (5), TTP (1), TTP (2) given in Table 4 involving: "Wormhole; Sybil; MITM; Sinkhole; DoS; SQL injection; XSS; Malware; Phishing; Node Capture; and Malicious node injection". All these threats can cause "complete loss of CIA".

3) It considers "five" protection controls; and these are: P (1) to P (5) in Table 5 involving: "Cryptography; Authentication; IDS; Firewall; and Protection Software". Each control can reduce the impact of "threats" leading to the improvement of the (CIA) state. For the application here, both "Authentication and Cryptography" are considered together as one control.

It should be mentioned here that the assessment outcome concerned with the reduction of impact of the various threats caused by the different protection controls on (CIA) is based on results produced by different researchers. The references concerned are cited below in their appropriate place.

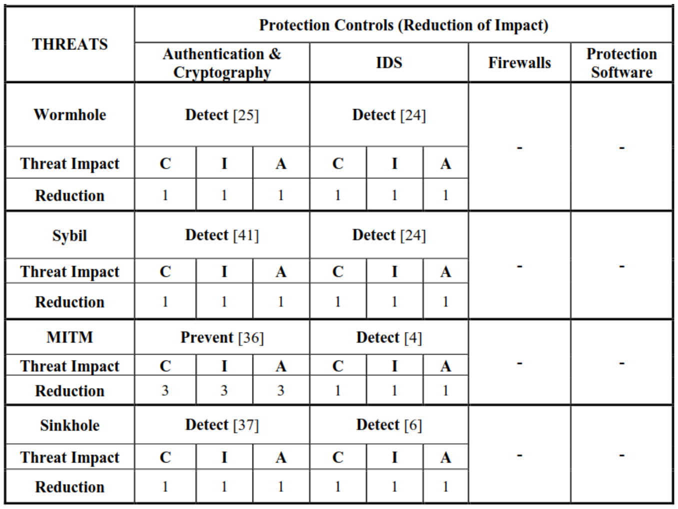
3.2 Assessment of Asset (1): The Network (Wireless Body Area Network: WBAN)

This asset faces "four threats (Wormhole; Sybil; MITM; and Sinkhole)" and considers "two protection controls (Authentication & Cryptography; and IDS) to limit their impact. Table 6 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can prevent the impact of MITM threat, and can also detect Wormhole, Sybil, and Sinkhole for (CIA).

2) The IDS control can detect all considered threats for (CIA).

Table 6: Threats versus protection measures for the "network: WBNA"



THREATS	Protection Controls (Reduction of Impact)							
	Authentication & Cryptography			IDS			Firewalls	Protection Software
Wormhole	Detect [25]			Detect [24]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	1	1	1	1	1	1		
Sybil	Detect [41]			Detect [24]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	1	1	1	1	1	1		
MITM	Prevent [36]			Detect [4]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	3	3	3	1	1	1		
Sinkhole	Detect [37]			Detect [6]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	1	1	1	1	1	1		

3.3 Assessment of Asset (2): The Server

This asset faces three threats (DoS, SQL injection, and XSS) and considers the four protection controls (Authentication & Cryptography, IDS, Firewalls, and Protection Software) to limit their impact. Table 7 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can prevent the impact of SQL injection threat and can also mitigate DoS threat for (CIA) and XSS threat, with regards to (C & I), but not (A).

2) The IDS control can detect all considered threats for (CIA).

3) Firewalls control can mitigate DoS and XSS; but can only detect SQL injection for (CIA).

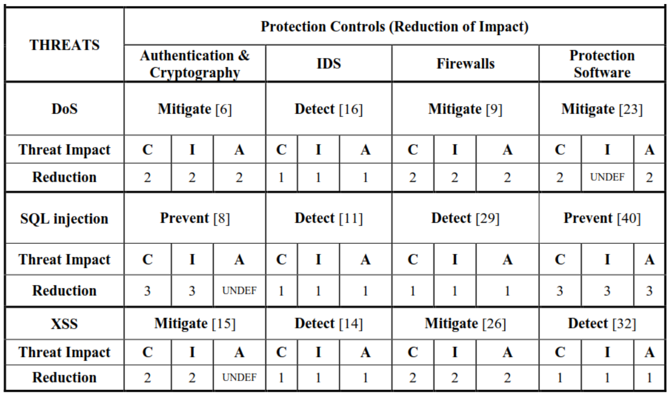
4) The Protection Software control can mitigate DoS for (C & A), but not (I); prevent SQL injection for (CIA); and can also detect XSS for (CIA).

3.4 Assessment of Asset (3): The End-Device

This asset faces two threats (Malware; and Phishing) and considers four protection controls (Authentication & Cryptography, IDS, Firewalls and Protection Software) to limit their impact. Table 8 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can mitigate the impact of Malware and Phishing threats for (CIA).

Table 7: Threats versus protection measures for the "server: PC"

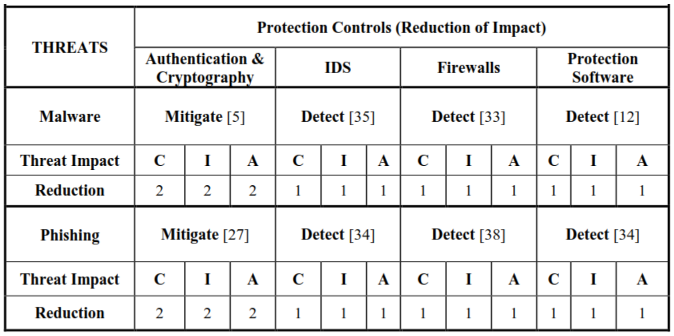


THREATS	Protection Controls (Reduction of Impact)											
	Authentication & Cryptography			IDS			Firewalls			Protection Software		
DoS	Mitigate [6]			Detect [16]			Mitigate [9]			Mitigate [23]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	2	2	2	2	UNDEF	2
SQL injection	Prevent [8]			Detect [11]			Detect [29]			Prevent [40]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	3	3	UNDEF	1	1	1	1	1	1	3	3	3
XSS	Mitigate [15]			Detect [14]			Mitigate [26]			Detect [32]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	UNDEF	1	1	1	2	2	2	1	1	1

2) The IDS control can detect the two considered threats for (CIA).

3) Firewalls control can detect Malware and Phishing for (CIA); and the Protection Software control can do the same.

Table 8: Threats versus protection measures for the "end-device"



THREATS	Protection Controls (Reduction of Impact)											
	Authentication & Cryptography			IDS			Firewalls			Protection Software		
Malware	Mitigate [5]			Detect [35]			Detect [33]			Detect [12]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	1	1	1	1	1	1
Phishing	Mitigate [27]			Detect [34]			Detect [38]			Detect [34]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	1	1	1	1	1	1

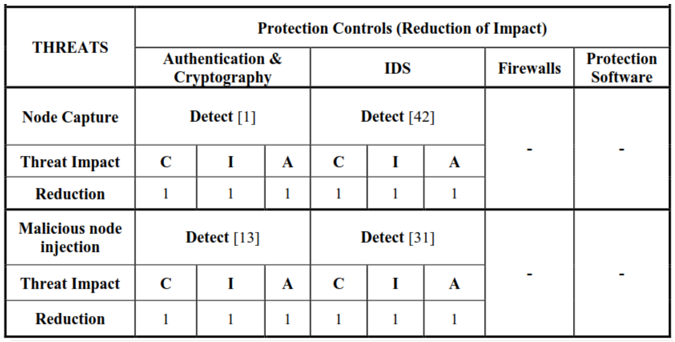
3.5 Assessment of Asset (4): Sensors

This asset faces two threats (Node Capture and Malicious Node Injection) and considers two protection controls (Authentication & Cryptography and IDS) to limit their impact. Table 9 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can detect Node Capture and Malicious Node Injection for (CIA).

2) The IDS control can detect the two considered threats for (CIA).

Table 9: Threats versus protection measures for "sensors"



THREATS	Protection Controls (Reduction of Impact)							
	Authentication & Cryptography			IDS			Firewalls	Protection Software
Node Capture	Detect [1]			Detect [42]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	1	1	1	1	1	1		
Malicious node injection	Detect [13]			Detect [31]			-	-
Threat Impact	C	I	A	C	I	A		
Reduction	1	1	1	1	1	1		

3.6 Assessment of Asset (5): The Application

This asset faces five threats (DoS, SQL injection, XSS, Malware and Phishing) and considers four protection controls (Authentication & Cryptography, IDS, Firewalls and Protection Software) to limit their impact. Table 10 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can prevent the impact of SQL injection threat and can also mitigate DoS and Malware and Phishing threats for all (CIA). It can also mitigate XSS for (CI).

2) The IDS control can detect all considered threats for (CIA).

3) Firewalls control can mitigate DoS and XSS; detect SQL injection, Malware and Phishing for (CIA).

4) The Protection Software control can mitigate DoS for (C & A), but not (I); prevent SQL injection for (CIA); detect Phishing, Malware and XSS for (CIA).

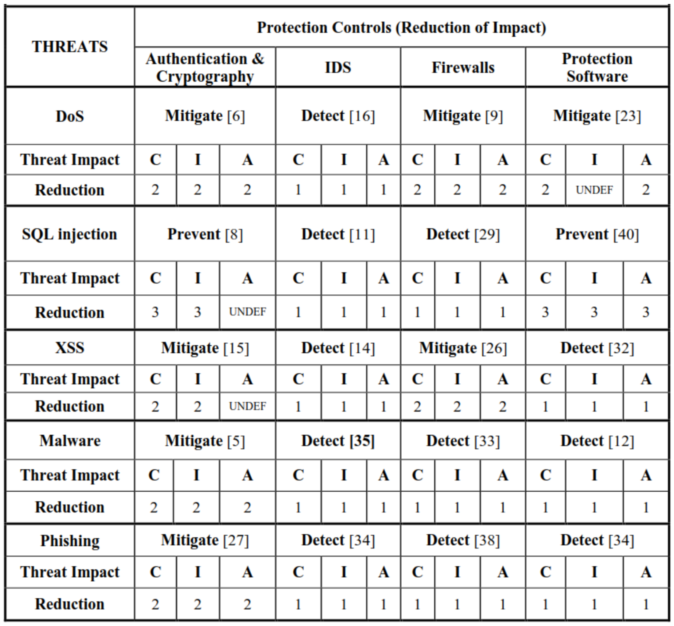
3.7 Assessment of Asset (6): The Database

This asset faces three threats (DoS; SQL injection; and Malware) and considers four protection controls (Authentication & Cryptography, IDS Firewalls and Protection Software) to limit their impact. Table 11 shows these threats and controls and provides the level of impact reduction caused by each control on every threat considering (CIA). The results show the following:

1) The Authentication and Cryptography control can prevent the impact of SQL injection threat and can also mitigate DoS and Malware threats for (CIA).

2) The IDS control can detect all considered threats for (CIA).

Table 10: Threats versus protection measures for the "application"

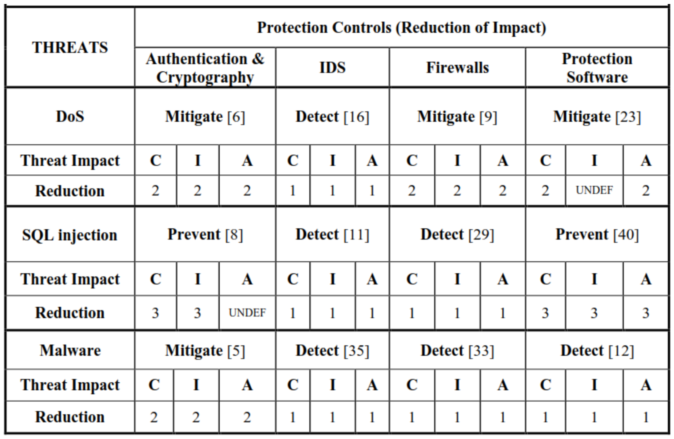


THREATS	Protection Controls (Reduction of Impact)											
	Authentication & Cryptography			IDS			Firewalls			Protection Software		
DoS	Mitigate [6]			Detect [16]			Mitigate [9]			Mitigate [23]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	2	2	2	2	UNDEF	2
SQL injection	Prevent [8]			Detect [11]			Detect [29]			Prevent [40]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	3	3	UNDEF	1	1	1	1	1	1	3	3	3
XSS	Mitigate [15]			Detect [14]			Mitigate [26]			Detect [32]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	UNDEF	1	1	1	2	2	2	1	1	1
Malware	Mitigate [5]			Detect [35]			Detect [33]			Detect [12]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	1	1	1	1	1	1
Phishing	Mitigate [27]			Detect [34]			Detect [38]			Detect [34]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	1	1	1	1	1	1

3) Firewalls control can mitigate DoS for (CIA); detect SQL injection and Malware for (CIA).

4) The Protection Software control can mitigate DoS for (C & A) but not for (I); prevent SQL injection for (CIA) and detect Malware for (CIA).

Table 11: Threats versus protection measures for the "database"



THREATS	Protection Controls (Reduction of Impact)											
	Authentication & Cryptography			IDS			Firewalls			Protection Software		
DoS	Mitigate [6]			Detect [16]			Mitigate [9]			Mitigate [23]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	2	2	2	2	UNDEF	2
SQL injection	Prevent [8]			Detect [11]			Detect [29]			Prevent [40]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	3	3	UNDEF	1	1	1	1	1	1	3	3	3
Malware	Mitigate [5]			Detect [35]			Detect [33]			Detect [12]		
Threat Impact	C	I	A	C	I	A	C	I	A	C	I	A
Reduction	2	2	2	1	1	1	1	1	1	1	1	1

3.8 Overall Assessment

The investigation approach followed in the investigation above will be useful in understanding the security problem on the one hand, and in making decision on what to do about it. For the case-study considered of "IoT application to Android concerned with interacting with neuromotor disabled patients", the following observations can be concluded.

1) IDS control impacts all "eleven threats" considered at the level of "detection", but it cannot "mitigate or prevent" any of them.

2) Authentication & Cryptography control impacts "six threats" at the levels of "prevention and mitigation". It can prevent "two threats" (MITM and SQL injection), and can mitigate "three others" (DoS, Phishing, and Malware) for (CIA) and it can also mitigate "one" XSS with regards to (C & I);

3) Firewalls control impacts "two threats (DoS and XSS) at the level of "mitigation".

4) Protection Software control impacts one threat (SQL injection) at the prevention level for (CIA); and another (DoS) at the mitigation level for (C & A).

According to the above, it can be viewed that with the above controls: two threats can be prevented (SQL injection and MITM); three others can be mitigated (DoS; Phishing; Malware); one more (XSS can be partly mitigated with (C&I).

All the other five (Wormhole; Sybil; Sinkhole; Node Capture; and Malicious node injection) can only be detected for further action.

4 Conclusions and Future Work

This paper delivered a distinguished approach for the security assessment of IoT in healthcare for the disabled. The approach provides a broad five-domain (STOPE: strategy; technology; organization; people and environment) structured view of the assessment problem. This enables flexibility and updatability to the application of the approach to the assessment of various related case-studies. The strategy domain directs the approach toward the achievement of (CIA: confidentiality; integrity and availability). The "assets" requiring protection are structured according to the (TOPE) domains, with the "technology" related assets are structured again according the IoT layers of (PNA: perception or physical; network; and application). For threats attacking assets, the approach considers that each threat can impact one or more assets leading to the degradation of the (CIA) state. The protection "controls" are associated with the threats; and each control work toward eliminating or mitigating the impact of one or more threats on assets leading to the promotion of the (CIA) state. The approach involves using a unified assessment scale for the levels of "asset importance; threat impacts on assets; control reduction of threat impacts; in addition to the (CIA) state".

The use of the approach has been illustrated through its application to a case study concerned with the security of "IoT on Android for Neuromotor disabled patients". The application illustrated how the assessment approach can be applied individually, per asset with various threats and protection controls; and how it can then be applied collectively considering all assets with their related threats and protection controls. This helps emphasizing individual assessments of specific "essential assets" whenever needed; without ignoring other less important assets when collective assessments are required.

Three directions can be considered for future work; and these are identified in the following.

1) The first and most obvious direction would be the use of the approach for future assessment of other related case-studies. This type of work would be close to professional work performed by hospital staff concerned, or by educational training work performed by students under supervision.

2) The second direction would be concerned with extending the approach, within the limit of its healthcare scope for the disabled, toward considering not only the (CIA) state in the strategy direction, but also the "cost" of the impacts of threats versus the cost of the protection controls that eliminates or mitigates them. Such research work would be useful to security protection decision makers.

3) The third direction is based on the second direction above, but it is more ambitious. It considers extending the approach, not only within the limit of its scope, but also beyond that toward providing a generic security assessment approach that can be of benefits to all cybersecurity applications in all fields.

It is hoped that the paper would be useful to all professionals and researchers in the field.

References

[1] S. Agrawal, M. L. Das, and J. Lopez, "Detection of node capture attack in wireless sensor networks," *IEEE Systems Journal*, vol. 13, no. 1, pp. 238-247, 2019.

[2] W. Al-Mawee, *Privacy and Security Issues in IOT Healthcare Applications for the Disabled Users*, MSc Thesis, Computer Science, Western Michigan University, 2015.

[3] B. S. Alghamdi, M. Elnamaky, M. A. Arafah, M. Alsabaan, S. H. Bakry, "A context establishment framework for cloud computing information security risk management based on the STOPE view," *International Journal of Network Security*, vol. 21, no. 1, pp. 166-176, Jan. 2019.

[4] F. Aliyu, T. Sheltami, and E. Shakshuki, "A detection and prevention technique for man in the middle attack in fog computing," *Procedia Computer Science*, vol. 141, pp. 24-31, 2018.

[5] O. Ami, Y. Elovici, and D. Hendler, "Ransomware prevention using application authentication-based file access control," in *the 33rd Annual ACM Symposium on Applied Computing*, pp. 1610-1619, 2018.

[6] O. Arazi, H. Qi, and D. Rose, "A public key cryptographic method for denial-of-service mitigation in wireless sensor networks," in *IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON'07)*, pp. 51-59, 2007.

[7] N. Asaddok and M. Ghazali, "Exploring the usability, security and privacy taxonomy for mobile health applications," in *International Conference on Research and Innovation in Information Systems (ICRIIS'17)*, pp. 1-6, 2017.

[8] I. Balasundaram and E. Ramaraj, "An authentication mechanism to prevent SQL injection attacks," *International Journal of Computer Applications in Technology*, vol. 19, no. 1, pp. 30-33, 2011

[9] A. Bonguet and M. Bellaiche, "A survey of denial-of-service and distributed denial of service attacks and defenses in cloud computing," *Future Internet*, vol. 9, no. 3, pp. 1-19, 2017.

[10] A. B. Chebudie, R. Minerva, D. Rotondi, "Towards a definition of the internet of things (IoT)," *IEEE Internet Initiative*, vol. 1, no. 1, pp. 1-86, 2015.

[11] A. S. Desai and D. P. Gaikwad, "Real time hybrid intrusion detection system using signature matching algorithm and fuzzy-GA," in *IEEE international conference on advances in electronics, communication and computer technology (ICAECCT'16)*, pp. 291-294, 2016.

[12] J. Gu, B. Sun, X. Du, J. Wang, Y. Zhuang, and Z. Wang, "Consortium blockchain-based malware detection in mobile devices," *IEEE Access*, vol. 6, pp. 12118-12128, 2018.

[13] A. Gupta, P. Mohit, A. Karati, R. Amin, and G. P. Biswas, "Malicious node detection using ID-based authentication technique," in *3rd International Conference on Recent Advances in Information Technology (RAIT'16)*, pp. 398-403, 2016.

[14] K. Gupta, R. R. Singh, and M. Dixit, "Cross site scripting (XSS) attack detection using intrusion detection system," in *International Conference on Intelligent Computing and Control Systems (ICICCS'17)*, pp. 199-203, 2017.

[15] S. Gupta and L. Sharma, "Prevention of Cross-Site Scripting Vulnerabilities using Dynamic Hash Generation Technique on the Server Side," *International Journal of Advanced Computer and Research*, vol. 2, no. 5, pp. 49-54, 2012.

[16] E. Hodo et al., "Threat analysis of IoT networks using artificial neural network intrusion detection system," in *International Symposium on Networks, Computers and Communications (ISNCC16)*, pp. 1-6, 2016.

[17] M. Irshad, "A Systematic Review of Information Security Frameworks in the Internet of Things (IoT)," in *IEEE International Conference on High Performance Computing and Communications (HPCC'16)*, pp. 1270-1275, 2016.

[18] ISO/IEC, *Information technology – Security techniques – Information security risk management*, ISO/IEC 27005, 2018.

[19] ISO/IEC, *Information Technology, Information Security, Cybersecurity and Privacy Protection – Information Security Controls*, ISO/IEC 27002, 2022.

[20] ISO/IEC, *Health informatics – Information security management in health International Organization for Standardization, Health Informatics-Information security management in health using ISO/IEC 27002*, ISO/IEC 27799, 2016.

[21] ISO/IEC, *Internet of Things (IoT): Preliminary Report*, International Standards Organization & International Electrotechnical Commission, Switzerland, 2014.

[22] ITU-T, "Series Y: Global information infrastructure, internet protocol aspects and next-generation networks, Next Generation Networks – frameworks and functional architecture models, Overview of the Internet of things Y.2060," ITU-T, Report, Switzerland, 2012.

[23] Kaspersky Lab., *DDOS Protection - Discover how Kaspersky Lab defends businesses against DDoS attacks*, Jan. 1, 2021. (`https://media.kaspersky.com/kaspersky-ddos-protection-data-sheet.pdf`)

[24] Z. A. Khan and P. Herrmann, "Recent advancements in intrusion detection systems for the internet of things," *Security and Communication Networks*, vol. 2019, Article ID 4301409, 2019.

[25] P. Khandare and N. Kulkarni, "Public key encryption and 2Ack based approach to defend wormhole attack," *International Journal of Computer Trends and Technology*, vol. 4, no. 3, pp. 247-252, 2013.

[26] E. Kirda, C. Krügel, G. Vigna, and N. Jovanovic, "Noxes: A client-side solution for mitigating cross-site scripting attacks," in *ACM Symposium on Applied Computing*, pp. 330-337, 2006.

[27] P. Kuacharoen, "An anti-phishing password authentication protocol," *International Journal of Network Security*, vol. 19, no. 5, pp. 711-719, 2017.

[28] R. G. Lupu et al., "Medical professional end-device applications on Android for interacting with neuromotor disabled patients," in *E-Health and Bioengineering Conference (EHB'15)*, pp. 1-4, 2015.

[29] Y. Manikanta and A. Sardana, "Protecting web applications from SQL injection attacks by using framework and database firewall," in *International Conference on Advances in Computing, Communications and Informatics*, 2012.

[30] I. Mohanraj and B. R. Raakesh, "ICT interventions on aiding people with disabilities - A state of art survey," in *International Conference on Inventive Communication and Computational Technologies (ICICCT'17)*, pp. 189-194, 2017.

[31] M. M. Ozcelik, E. Irmak, and S. Ozdemir, "A hybrid trust-based intrusion detection system for wireless sensor networks," in *International Symposium on Networks, Computers and Communications (ISNCC'17)*, pp. 1-6, 2017.

[32] R. M. Pandurang and D. C. Karia, "A mapping-based podel for preventing cross site scripting and sql injection attacks on web application and its impact analysis," in *International Conference on Next Generation Computing Technologies (NGCT'15)*, pp. 414-418, 2015.

[33] S. Raje, S. Vaderia, N. Wilson, and R. Panigrahi, "Decentralised firewall for malware detection," in *International Conference on Advances in Computing, Communication and Control (ICAC3'17)*, pp. 1-5, 2017.

[34] R. S. Rao and S. T. Ali, "Phish-Shield: A desktop application to detect phishing webpages through heuristic approach," *Procedia Computer Science*, vol. 54, pp. 147-156, 2015.

[35] I. A. Saeed, A. Selamat, and A. M. A. Abuagoub, "A survey on malware and malware detection systems," *International Journal of Computer Applications*, vol. 67, no. 16, pp. 25-31, 2013.

[36] A. Sahi and D. Lai, "Preventing man-in-the-middle attack in Diffie-Hellman key exchange protocol," in *International Conference on Telecommunications (ICT'15)*, Australia, 2015.

[37] S. Sharmila and G. Umamaheswari, "Detection of Sinkhole attack in wireless sensor networks using message digest algorithms," in *International Conference on Process Automation, Control and Computing*, pp. 1-6, 2011.

[38] B. Soewito and C. E. Andhika, "Next generation firewall for improving security in company and IoT network," in *International Seminar on Intelligent Technology and Its Applications (ISITIA'19)*, pp. 205-209, 2019.

[39] S. N. Swamy, D. Jadhav, and N. Kulkarni, "Security threats in the application layer in IOT applications," in *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC'17)*, pp. 477-480, 2017.

[40] O. P. Voitovych, O. S. Yuvkovetskyi, and L. M. Kupershtein, "SQL injection prevention system," in *International Conference Radio Electronics & Info Communications (UkrMiCo'16)*, pp. 1-4, 2016.

[41] J. Wadii, H. Rim, and B. Ridha, "Detecting and preventing Sybil attacks in wireless sensor networks," in *IEEE 19th Mediterranean Microwave Symposium (MMS'19)*, pp. 1-5, 2019.

[42] L. Watkins, S. Aggarwal, O. Akeredolu, W. Robinson, and A. Rubin, "Tattle tale security: an Intrusion Detection System for Medical Body Area Networks (MBAN)," in *Workshop on Decentralized IoT Systems and Security (DISS'19)*, 2019.

Biography

Reem Almasoudi is a Computer Engineer; she received her MSc degree in Computer Engineering from King Saud

University.

Mohammad Arafah is Professor of Computer Engineering at King Saud University and is the Dean of the College of Applied Studies at the University.

Waleed Alghanem is Assistant Professor of Computer Engineering and is a Consultant at the Saudi Ministry of Education.

Saad Bakry is Professor of Computer Engineering at King Saud University and is a weekly columnist, on Knowledge and Development, in the Saudi daily Aleqtisadia.

CSMTP: An RL-based Adversarial Examples Generation Method for Chinese Social Media Texts Classification Models

Xin Tong, Jingya Wang, Binjun Wang, Hanming Zhai, and Kaidi Zhang

(Corresponding author: Jingya Wang)

People's Public Security University of China
Beijing 100038, China
Email: wangjingya@ppsuc.edu.cn

(Received May 25, 2022; Revised and Accepted Dec. 2, 2022; First Online Dec. 13, 2022)

Abstract

Natural language processing models based on deep learning have been widely used to analyze Chinese social media texts, such as Sina Weibo sentiment analysis, rumor detection, and news topic classification. However, while these analysis models have made breakthroughs in performance, there are also potential adversarial attack risks. To test the robustness of these models, an adversarial examples generation method CSMTP is proposed in this paper. The algorithm includes two main parts: a key tokens searcher and an adversarial noise selector. The key tokens searcher is used to locate the key tokens that significantly affect the classification results when the internal details of the model are unknown, and the visual perception weight is introduced to measure human attention to these keywords. The adversarial noise selector provides a variety of adversarial noise generation strategies. It uses the policy network based on deep reinforcement learning training to select the appropriate type of noise according to the features of key tokens and target sentences. Experiments based on four public datasets show that CSMTP can launch effective adversarial attacks on Chinese social media text classification models, which has specific enlightening significance for the follow-up research on the reliability of these models.

Keywords: Adversarial Examples; Deep Q-learning; Social Media; Texts Classification

1 Introduction

Due to the development of the Internet and the emergence of Web 2.0 technology, all kinds of intelligent social media platforms have gradually become the primary way for people to obtain information, express emotions and communicate. In China, social media platforms such as Sina Weibo, Headline and Baidu news have become a part of everyone's life. Taking Weibo as an example, the 2021 Sina Financial Report shows that as of December 2021, the number of monthly active users of Sina Weibo has reached 573 million, and the number of daily active users has reached 249 million. Text data is one of the most common carriers for people to share information on these platforms. Therefore, the mining and analysis of these massive social media texts have essential research and commercial value.

From the perspective of models and methods, the current text classification models based on neural networks have greatly surpassed the early methods based on pattern matching and statistical machine learning. They have become the mainstream technology in Chinese social media texts mining. Various deep neural networks based on CNN [12], LSTM [7], and attention mechanism [24] have been widely used in natural language processing tasks such as sentiment analysis [15], hotspot discovery [10], intention recognition [3], malicious speech detection [2] and fake news detection [26]. However, most researchers pay more attention to the analysis performance of the models, such as accuracy and inference speed, but ignore their robustness. Work [25] showed that when faced with adversarial examples carefully constructed by attackers, even the deep neural network model with excellent performance also shows great vulnerability, that is, attackers can modify the characters, words and sentences of the text to generate adversarial examples, to significantly change the prediction results of these models without affecting human understanding of semantics.

To solve this problem, we focus on the "shadow under the sun" and mainly discuss the security of Chinese social media texts classification models. An adversarial examples generation method CSMTP (**C**hinese **S**ocial **M**edia **T**exts **P**erturbator) for Chinese social media texts classification is proposed in this paper. This method can launch character-level, word-level, targeted and untargeted adversarial attacks on Chinese social media text classification systems when obtaining a small amount of model information. Specifically, the main contributions of this paper include:

1) **A key tokens searcher that can comprehensively measure text elements' criticality and visual concealment is designed.** It can accurately locate the key characters and words that contribute significantly to the classification results while maintaining semantic and visual similarity as much as possible and using dynamic hyperparameters to limit the disturbance intensity. In addition, this location method can make the algorithm have the ability to realize targeted attacks and attack multi-label classification models.

2) **A variety of transformation strategies capable of generating adversarial noise are adopted.** On the one hand, it can help the algorithm generate adversarial text with highly consistent semantics with the original sentence. On the other hand, it can effectively expand the search space of adversarial noise and help to improve the attack success rate of the algorithm.

3) **An adversarial noise selector based on deep reinforcement learning is realized.** It can analyze the key tokens and the characteristics of their sentences and select the type of perturbation noise with a high probability of attack success.

Using social media texts data, including Sina Weibo sentiment analysis dataset, Weibo rumor dataset, E-commerce review dataset and news classification dataset, and classic models such as TextCNN, LSTM and self-attention mechanism as targets to test CSMTP. The results show that this algorithm can greatly reduce the accuracy of the target model by using less perturbation noise, which is of positive significance to studying the robustness of the existing Chinese social media texts classification models.

2 Related Work

2.1 Adversarial Examples

Szegedy *et al.* [18] first found that adding subtle noise to the input picture of CNN can lead to a wrong output of the models with high confidence to achieve the purpose of cheating them and put forward the concept of adversarial examples. The adversarial examples refer to the sample x' generated by adding a small disturbance to the input data x that is difficult for the human eye to detect. When x' is input into the trained deep learning model, the model will predict the output results differently from the original label with high confidence, as in Equation (1). Where f represents the forward propagation process of neural network, ε is used to constrain the intensity of perturbation.

$$f(x) \neq f(x') s.t. \|x' - x\| < \varepsilon. \tag{1}$$

Although adversarial attacks were first found in image classification tasks, they have extended to the field of natural language processing. Affected by the high dimensionality of text vectors, the closeness of the training process and the complexity of models, it is still difficult for both shallow neural networks and large-scale pre-training language models [11, 20] to learn the high-level semantic features of text, and there are defects in robustness. A lot of work [22] has proved that this vulnerability can be used to generate highly hidden and deceptive adversarial examples for text classification systems.

2.2 Adversarial Examples Generation Algorithms for Text Data

Text data has discrete features, complex syntax rules and abstract semantics, significantly different from image data. As a result, the adversarial examples generation methods based on gradient or optimization in computer vision are difficult to transfer to the field of text. The adversarial examples generation technology in the text field mainly focuses on two key points: one is how to search the key characters, words, and even sentences with high contribution to the model prediction results from the document; the other is what kind of noise is added to these tokens to achieve the purpose of changing the model prediction results. The related work can be divided into character-level, word-level, sentence-level and multiple levels according to the level of perturbation noise.

2.2.1 Adversarial Examples in English

Belinkov [1] first tried to add, delete and exchange characters in English text, which led to the decline of the performance of neural machine translation system, proving the existence of text adversarial examples. On this basis, DeepWordBug [5] algorithm used the temporal score mechanism to measure the importance of characters, which can accurately locate the key tokens when the gradient information of the model is unknown and inspired a series of subsequent adversarial examples generation algorithms based on score query. Papernot [17] used the FGSM (Fast Gradient Sign Method) [6] adversarial attack method commonly used in the image field to modify the keywords that have the most significant impact on the classification results. On obtaining the gradient information inside the model, this method had a success rate of 100% in the experiment of attacking the emotion classification system based on RNN models. Jia-Liang's method [8] can significantly reduce the effect of the question and answer system by inserting perturbation sentences that do not affect human reading into the text and realize the adversarial attack at the sentence level. The latest black box method TEXTFOOLER [9] further proved that although BERT and other models have good performance, they are still difficult to defend against the attack of adversarial examples. In addition, work [4, 13, 14] proposed a series of adversarial examples generation methods that can generate different levels of adversarial text simultaneously or use multiple levels of noise to cooperate with each other, which have a broader application scenario than the previous single-level method.

2.2.2 Adversarial Examples in Chinese

However, the above research was mainly aimed at the adversarial attacks methods of English, which is difficult to apply to the hieroglyphs represented by Chinese characters. Therefore, the research on the adversarial examples in the Chinese field is still in its infancy. Wang *et al.* [21] proposed a character-level adversarial attack method WordHandling that can attack the Chinese sentiment classification system, by misleading LSTM and CNN models to make wrong predictions and can reduce the accuracy of the model by an average of 29% (LSTM) and 22% (CNN). But this method was mainly used in untargeted attack scenarios. The generated adversarial examples cannot mislead the prediction results of neural networks into the label specified by the attacker. At the same time, this method only used the homophonic character replacement in the selection of noise, resulting in the limited diversity of generated examples. Tong *et al.* [19] proposed a word-level method CWordAttacker based on traditional Chinese character rewriting, Pinyin replacement and word order exchange, which can complete targeted and untargeted attacks. Experiments on the THUCNews dataset showed that the algorithm can reduce the accuracy of the CNN and LSTM models by 30.22% and 43.29%, respectively. The CWordAttacker assumed that all noise transformation strategies can map keywords into words not recorded in the word embedding space. Therefore, the algorithm randomly selected transformation strategies to modify the target words to ensure the diversity of generated adversarial examples. But this assumption does not apply to social media texts. Because these texts often contain Pinyin, special symbols and homonyms, the text replaced by these strategies may still be the samples stored in the word embedding space, which makes the perturbation effect of these noises on the target sentence uncertain and may harm the success rate of the attack. Xu *et al.* [23] added an aggregation module of shape-similar characters to expand further the search range of adversarial noise based on the above research. And violent search was used to determine the optimal perturbation strategy. Finally, the accuracy of the topic classification models trained on the THUCNews dataset can be reduced by 37.49% (CNN) and 37.85% (LSTM), respectively. But this further increased the algorithm's time complexity, which made the algorithm generate examples for large-scale datasets with higher computational cost and more time. At the same time, the algorithm only carried out the experiment of untargeted attacks.

The above algorithms can only achieve single-level adversarial attacks, and most experimental datasets are not social media texts. The performance of attacking Chinese social media text classification models still needs further verification. At the same time, when choosing the type of perturbation noise, these methods often only focus on the attack success rate but ignore the visual significance of the noise, resulting in the generated examples being easy to be found by readers. Intuitively, compared with Pinyin replacement, the noise such as similar shape words and word order change is less likely to be detected by readers, so it should be preferred. In conclusion, the above methods can be further improved.

3 Method

The mainstream Chinese social media text classification models are divided into character-level and word-level models. The character-level models directly encode and process the characters in the text. The advantage is that it is not limited by the scale of word embedding space and can analyze the special symbols, emojis and other irregular data in the social media text. But the model needs to aggregate semantic units from scratch to learn high-level semantic features. The word-level model uses the word segmentation algorithm to slice the text into words and trains the word-level embedding vector on this basis, making it easier for the model to learn high-level semantic features. Still, the overall effect of the model will be constrained by the performance of the word segmentation models. Considering that these two types of models are widely used in Chinese social media classification tasks, an adversarial text generation method CSMTP, which can attack both character-level and word-level classification models at the same time, is proposed. The method is mainly composed of a key tokens searcher and an adversarial noise selector.

3.1 Key Tokens Searcher

For social media texts, the contribution of different characters and words to the classification results of neural network models is quite different. Therefore, how to locate the key tokens that need to add noise is the basis for generating highly hidden and deceptive adversarial texts. When CSMTP algorithm generates adversarial examples, the key tokens searcher based on query scoring mechanism and visual perception weight is used to find the optimal position of noise addition. First, the importance of the Chinese characters or words in the sentence is estimated by calculating the targeted delete score for each Chinese character or phrase. Then the visual similarity of the text is constrained by calculating the visual perception weight of the position of the tokens. Then, the hyperparameter adaptively adjusted according to the text length is used to control the intensity of noise further and ensure the concealment of adversarial examples.

3.1.1 Query Scoring Mechanism

The text adversarial examples generation methods often use the temporal score mechanism [5] to locate keywords, as in Equation (2). The mechanism takes the target word as the center, divides the input sentence into two parts, and measures the importance of the word by calculating the change in the confidence of the model prediction results before and after deleting the target word.

$$\begin{aligned} TS(\boldsymbol{w}_i) &= THS(\boldsymbol{w}_i) + \lambda \times TTS(\boldsymbol{w}_i) \\ &= [f(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_i) - f(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{i-1})] + \\ &\lambda \times [f(\boldsymbol{w}_i, \cdots, \boldsymbol{w}_n) - f(\boldsymbol{w}_{i+1}, \cdots, \boldsymbol{w}_n)] \end{aligned} \quad (2)$$

However, this method of splitting sentences into two steps may lead to unreliable evaluation results. Inspired by the WordHandling algorithm [21], CSMTP uses TDS (Targeted Delete Score) to find the optimal position. As shown in Equation (3), the importance of the token is evaluated by calculating the change in the probability that the sentence is predicted to be the label y after deleting the target word.

$$\begin{aligned} TDS^{(y)}(w_i) \quad = \quad & f^{(y)}(w_1, \cdots, w_{i-1}, w_i, w_{i+1}, \cdots, w_n) \\ & -f^{(y)}(w_1, \cdots, w_{i-1}, w_{i+1}, \cdots, w_n) \end{aligned} \quad (3)$$

3.1.2 Visual Perception Weight

A survey shows that due to the influence of digital technology, the length of human attention per time has been shortened from 12 seconds in 2000 to 8 seconds today. This means that the perturbation in the middle of the text is less likely to be detected when reading the text. Therefore, the CSMTP algorithm measures the possibility of human detection by introducing VPW (Visual Perception Weight) when searching for key tokens. As shown in Equation (4), where i is the position of the target character, μ is the center of the sensory field of vision, that is, the point where the perturbation is most challenging to be found. The higher the value of VPM, the less noticeable the change to the character of this position. As shown in Figure 1, the distribution of VPM value can be dynamically adjusted according to the length of the input text.

$$\begin{aligned} VPW(i) = 1 + \beta \times \sqrt{n} \times Gaussian(i; \mu, \sigma), \\ where \quad \mu = \frac{n}{2}, \sigma = \frac{n}{4} \end{aligned} \quad (4)$$

3.2 Adversarial Noise Selector

The adversarial noise selector mainly provides a series of transformation algorithms that can generate various noisy texts. A policy network based on deep reinforcement learning is also implemented, which is used to match the appropriate transformation algorithm to generate adversarial examples according to the input key tokens and target sentences.

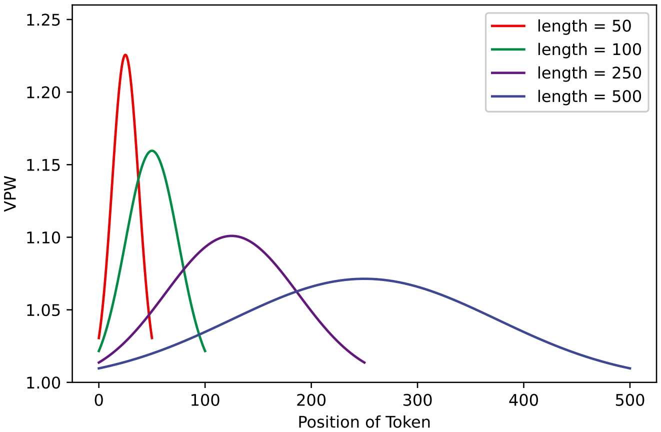


Figure 1: Distribution of VPM values with different text lengths

3.2.1 Text Transformation Algorithm

To ensure the similarity of the semantic features of the text before and after the disturbance is added, that is, the generated adversarial examples can deceive the target models and try not to affect human reading and understanding, five text transformation strategies are provided to convert key tokens into noise tokens, as shown in Table 1.

Among them, transformations 1, 2, and 5 can be used to generate adversarial examples for character-level and word-level models at the same time, while transformations 3 and 4 are only used for word-level models. These transformations can modify the tokens that contribute greatly to the classification. The target models can map them to the word vectors that are meaningless to the correct results or are not stored in the word embedding space to change the prediction results of the models.

3.2.2 Policy Network

To select the algorithm with a higher attack success rate from the five transformations, CSMTP introduces the task-related policy network to analyze the features of key tokens and sentences to be attacked and finally select the appropriate transformation. The structure of the policy network is shown in Figure 2. The network adopts a double input structure, which accepts the key tokens and target sentences to be transformed as inputs, respectively. It contains two outputs, one for outputting the index of the recommended transformation algorithm and the other for the pre-training process of the network.

Because the output of the policy network is the action, it isn't easy to use the traditional supervised learning method for end-to-end training. Therefore, a two-stage training method based on "pre-training - deep reinforcement learning" is adopted by CSMTP.

Pre-training stage. In order to make the policy network represent the feature space of the text, the datasets are

Table 1: Text transformation strategies

Index	Transformations	Noise Level	Original Tokens	Adversarial Tokens
1	shape-similar characters replacement	both	愉 快	榆 快
2	traditional characters replacement	both	忧 伤	憂 傷
3	words order perturbation	word-level	愉快	快愉
4	words splitting	word-level	开心	开-心
5	Pinyin rewriting	both	高 兴	Gao Xing

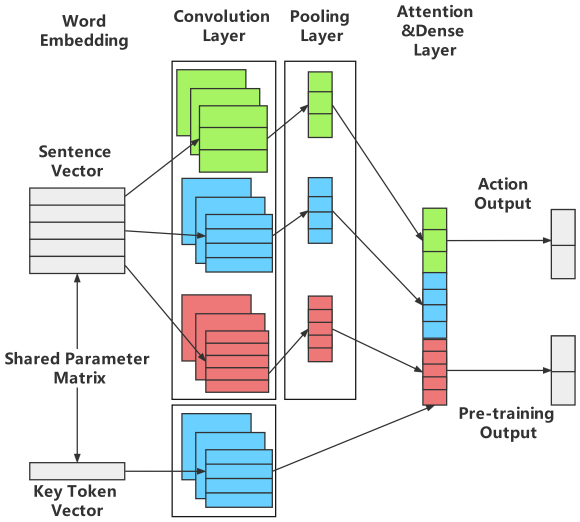


Figure 2: The structure of the policy network

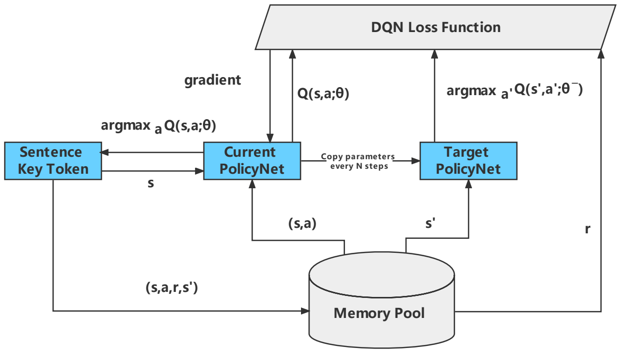


Figure 3: Training process of the policy network

preprocessed into the format of "label, key token, target sentence". Then the supervised learning method is used to pre-train the network, that is, to fit the mapping relationship between the inputs and the downstream tasks.

Deep reinforcement learning stage. Freeze the embedding and convolution layer near the inputs as the feature extractors, and use the deep Q-learning method [16] to retrain the policy network. The traditional Q-learning is a reinforcement learning algorithm based on iterative value, in which Q represents $Q(s, a)$, which is the expectation of income after taking action a in a certain state s. the environment will feedback corresponding rewards according to the action of Q-learning agent. Its optimization process is shown in Equation (5). Finally, Q-learning can construct s and a into a Q-table to store the Q value and then sample the actions that can obtain the maximum benefit according to the Q value.

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta \times (R_{t+1} + \gamma \times \max_a Q(s_{t+1}, a) - Q(s_t, a_t)). \tag{5}$$

However, in the task of selecting an appropriate transformation algorithm according to key tokens and target sentences, the input state s is a high-dimensional word vector, which is difficult to process by the traditional Q-learning method. Therefore, deep Q-learning is introduced to generate Q-values by fitting a neural network instead of Q-table, to map high-dimensional word vectors into actions. The main process is shown in Figure 3.

The details of the training policy network using deep Q-learning are shown in Algorithm 1. The algorithm accepts the policy network PolicyNet, target network TargetModel and state space S as inputs. S consists of all (key token vectors, sentence vectors) tuples, and the output of PolicyNet represents the type of transformation applied to the key tokens.

Algorithm 1 Training PolicyNet by DQN

```
Procedure DQN (PolicyNet, TargetModel, $S$):
    Initialize replay memory $D$ to capacity $n$
    $PolicyNet_{\theta^-} \leftarrow PolicyNet_\theta$
    for $e$ interations do:
        Reset sequence $S$
        for $t$ interations do:
            $action_t \leftarrow$ ArgMax$_{action}$(PolicyNet $(\phi(S_t), action; \theta))$
            $r_t \leftarrow R(action_t, \phi(S_t)$, TargetModel)
            $D \leftarrow (S_t, action_t, r_t, S_{t+1})$
            Sampling $(S_j, action_j, r_j, S_{j+1})$ from $D$
            if j = T-1: $y \leftarrow r_j$
            else: $y \leftarrow r_j + \gamma \times$
                $Max_{action}$(PolicyNet $(\phi(S), action; \theta^-))$
            Gradient descent step on $(y$
                -PolicyNet $(\phi(S_j), action_j; \theta))^2$
            Every $c$ step: PolicyNet$_{\theta^-} \leftarrow$ PolicyNet$_\theta$
    return PolicyNet
```

The reward function of deep Q-learning is also adjusted in the training process, as shown in Equation (6). *PertScore* is the perturbation score, which mainly rewards and punishes according to whether the output actions of the policy network can deceive the target model. *TypeScore* refers to the score of action type used to ensure the concealment and diversity of noise. It will give higher reward scores to the four transformations of traditional characters replacement, shape-similar characters replacement, words order disturbance and words splitting. In comparison, it will provide lower reward scores for the transformation of Pinyin rewriting, which is easier to attract readers' attention. *FreqScore* is the action frequency reward. This score prevents the policy network from learning shortcuts, that is, only selecting specific types of transformations. The *FreqScore* of the most frequently used actions will be reduced during training, while the *FreqScore* of the least frequently used actions will be increased.

$$\begin{aligned} R(\boldsymbol{s}_t, a_t) = &\ PertScore(\boldsymbol{s}_t, a_t, TargetModel) \\ &+ TypeSore(a_t) + FreqScore(a_t) \end{aligned} \tag{6}$$

3.3 Algorithm Description

The CSMTP algorithm changes the prediction results of the model by modifying key tokens, which includes two steps: Firstly, the TDS mechanism is used to search the key tokens in the sentence. Then, the policy network is used to match the appropriate algorithm from five text transformations to add noise to the text. For untargeted attacks, the algorithm modifies the words that contribute more to the original label of the result as much as possible, squeezes the confidence of the original label and improves the probability of other predicted labels. For targeted attacks, the algorithm attempts to attack tokens that hinder the prediction of the target labels and indirectly improves the probability of the model predicting the labels. This process of focusing only on the original labels or target labels simplifies the step of determining the attack direction of the WordHandling algorithm so that the algorithm can be extended to attack multi-classification models.

The process of generating adversarial examples using CSMTP is shown in Algorithm 2. The algorithm will receive the text W, the original label y^*, and the targeted attack switch *is_targeted*, targeted attack target label y' and the maximum number of attacks limit dynamically calculated according to the text length, and noise ratio are used as inputs.

4 Experiments & Analysis

4.1 Experimental Datasets

To verify the attack performance of CSMTP algorithm, Chinese social media text datasets with different targets and data characteristics were selected for training target models and testing adversarial examples. The details of the datasets are shown in Table 2 and Figure 4.

Algorithm 2 CSMTP

```
Procedure CSMTP($\boldsymbol{W}, limit, y^*, is\_targeted, y'$):
    $S \leftarrow$ dict()
    $y \leftarrow$ TargetModel($\boldsymbol{W}$)
    if ($y \neq y^*$ and not $is\_targeted$) or $y = y'$:
        return $\boldsymbol{W}$
    $\boldsymbol{A} \leftarrow \boldsymbol{W}$
    for $i$ in length of sentence $\boldsymbol{W}$:
        if $is\_targeted : t \leftarrow$ -TDS($\boldsymbol{W}_i, y'$)
        else: $t \leftarrow$ TDS($\boldsymbol{W}_i, y$)
        if $t > 0 : S_i \leftarrow$ VPM($i$) $\times t$
    SortbyKey($S$)
    for $limit$ interations do:
        for $key$ in $S$:
            $action \leftarrow$ PolicyNet($\boldsymbol{W}, \boldsymbol{W}_{key}$)
            $\boldsymbol{A}_{key} \leftarrow$ Transform($\boldsymbol{W}_{key}, action$)
    return $\boldsymbol{A}$
```

4.2 Comparative Experiments

In the experiment, RandomAttack and CWordAttacker [19] were selected as the baseline methods to compare with CSMTP algorithm, and the attack effects of their targeted attack and untargeted attack versions on three target models were compared.

For untargeted attacks, the decline in the accuracy of the models after the attack was used to evaluate the attack effect of the algorithms. The greater the reduction, the stronger the ability of the algorithm to mislead the models. The experimental results are shown in Table 3 and Table 4, and the optimal results are marked in black bold.

Targeted attacks are stricter in evaluating the effect. The algorithm is successful only when the misleading model predicts the example as the specified label. Therefore, the algorithm uses the attack success rate as a measure, as shown in Equation (7). Specifically, the attacker will select a label as the target label, and then the samples belonging to this category in the dataset will be deleted. Finally, the proportion change of the label in the prediction results before and after the models are attacked will be calculated as the results.

$$\begin{aligned} ASR_{targeted} &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{I}(f(A_{adv}^{i}) = y') \\ &\quad - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{I}(f(W_{org}^{i}) = y') \end{aligned} \tag{7}$$

In the equation, y' is the target label, and I is the indication function. The experimental results of word-level and character-level targeted attacks are shown in Table 5 and Table 6, respectively. In the Tables, ER is the abbreviation for error rate, which means the proportion of

Table 2: Information of experimental datasets

Datasets	Weibo Sentiment	Weibo Rumor	E-commerce Review	THUCNews
Tasks	Sentiment Analysis	Rumor Detection	Sentiment Analysis	Topic Classification
Categories	2	2	2	10
Label Distribution	Balance	Balance	Balance	Balance
Train Set	18000	2000	9600	50000
Validation Set	2000	200	2000	5000
Test Set	2000	1000	1200	10000

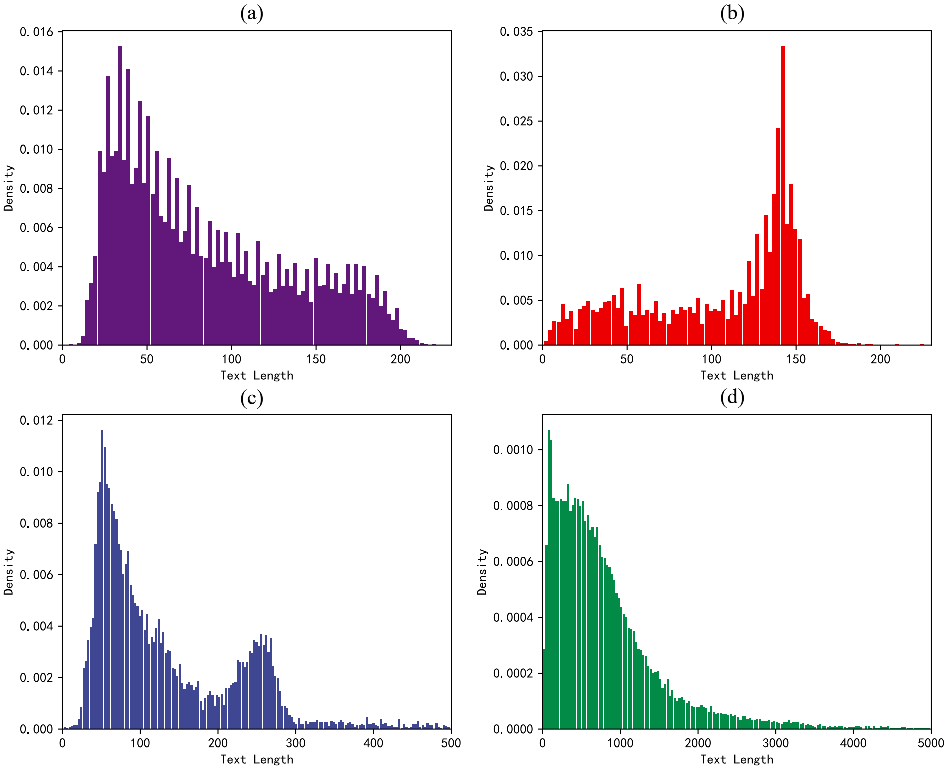


Figure 4: Text lengths distribution of the datasets. (a) Weibo sentiment. (b) Weibo Rumor. (c) E-commerce Review. (d) THUCNews

Table 3: Untargeted attack experiments against word-level models

(a) LSTM Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	96.40	92.15	4.25	58.10	38.30	54.00	**42.40**
Weibo rumor	85.70	84.20	1.50	55.20	30.50	33.20	**52.50**
E-com. review	88.25	85.67	2.58	41.33	46.92	30.58	**57.67**
THUCNews	92.39	91.59	0.80	49.10	43.29	45.52	**46.87**

(b) TextCNN Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	96.30	91.25	5.05	55.40	40.90	51.20	**45.10**
Weibo rumor	86.60	85.90	0.70	67.10	19.50	42.70	**43.90**
E-com. review	89.00	87.33	1.67	39.75	49.25	30.42	**58.58**
THUCNews	94.04	93.40	0.64	63.82	30.22	58.96	**35.08**

(c) Att-CNN Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	96.05	90.75	5.30	57.60	38.45	53.75	**42.30**
Weibo rumor	85.80	84.60	1.20	67.70	18.10	57.40	**28.40**
E-com. review	90.08	88.25	1.83	51.08	39.00	42.75	**47.33**
THUCNews	92.48	90.95	1.53	68.07	24.41	63.02	**29.46**

Table 4: Untargeted attack experiments against character-level models

(a) LSTM Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	95.75	90.25	5.50	55.75	40.00	51.40	**44.35**
Weibo rumor	84.10	80.20	3.90	41.90	42.20	24.70	**59.40**
E-com. review	87.42	80.50	6.92	39.92	47.50	25.83	**61.59**
THUCNews	87.30	66.83	20.47	13.95	73.35	11.09	**76.21**

(b) TextCNN Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	95.65	90.75	4.90	54.15	41.50	49.75	**45.90**
Weibo rumor	84.70	79.10	5.60	56.50	28.20	42.10	**42.60**
E-com. review	89.25	82.42	6.83	40.25	49.00	30.17	**59.08**
THUCNews	90.71	67.62	23.09	26.92	63.79	24.39	**66.32**

(c) Att-CNN Model

Datasets	**Ori_Acc**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		Accuracy	**Reduction**	**Accuracy**	**Reduction**	**Accuracy**	**Reduction**
Weibo sentiment	96.05	91.30	4.75	59.75	36.30	55.85	**40.20**
Weibo rumor	85.20	82.10	3.10	39.90	45.30	37.40	**47.80**
E-com. review	88.67	79.92	8.75	48.83	39.84	39.00	**49.67**
THUCNews	90.35	67.75	22.60	27.39	62.96	25.49	**64.86**

examples with wrong predictions in the total number of examples. To conform to the attacker's behavior in the real scene, in the targeted attack task of emotion analysis, E-commerce review and rumor detection tasks, negative emotion to positive emotion, negative reviews to positive reviews, and rumor to non-rumor were selected as the attack targets. The ”technology” category was randomly selected as the attack target for the news classification task.

By analyzing the above experimental results, it can be found that CSMTP algorithm can launch effective adversarial attacks against Chinese social media text classification systems based on LSTM, TextCNN and self-attention mechanism, and the attack performance was generally better than the CWordAttacker and RandomAttack.

Firstly, from the analysis of the perturbation level of adversarial examples, by comparing the attack effects of character-level attacks and word-level attacks, it can be found that there was an anti-intuitive phenomenon: In general, there were more noise transformation strategies for word-level attacks, and the search space for adversarial noise was larger. Therefore, after cooperating with the policy network, it should have a higher attack success rate, but the experimental results showed that the attack performance of the character-level was usually better. One possible reason was that in the word-level models, the scale of the word embedding matrix was limited and fewer tokens were recorded. As a result, in the model training stage, most of the words in the normal examples were also mapped to the UNK word vector (representing the words not recorded by the word embedding layer). Finally, these models were robust to the adversarial noise. The samples contained fewer UNK vectors in the training process for the character-level models, so the models were more vulnerable to antagonistic noise.

Secondly, from the perspective of datasets, for short text datasets, the CSMTP algorithm can achieve good experimental results in both targeted and untargeted attack tasks. Untargeted attacks can reduce the accuracy of the three target models by more than 40% in most scenarios, making the attacked model approximate random guessing when predicting adversarial examples. For the news classification dataset with long text lengths, CSMTP still had an advantage in the success rate of untargeted attacks. Still, the performance of character-level targeted attacks was slightly lower than that of the CWordAttacker, and the performance of word-level targeted attacks was also close to the baseline. In addition to the fact that the length of the text was too long, the contribution of the tokens was scattered, which limited the attack effect of perturbation noise. Another main reason was that the pinyin rewriting has a better attack effect on the target models trained on the news dataset. However, the policy network of CSMTP gave a lower reward for this noise transformation during the training process, so the probability of CSMTP choosing this noise attack was smaller, resulting in a lower attack effect than the CWordAttacker.

Finally, from the perspective of attack type, CSMTP was more stable for targeted and untargeted attacks on binary classification tasks, but the difference between the two attacks on the multi-class dataset was obvious. Taking the LSTM model as an example, the accuracy of the model before and after the character-level attack on the subset that did not contain the target category data dropped by about 70%, but the attack success rate for the target label was only 2%, which proved that the perturbation of the targeted attack was not enough to mislead the model to predict the example as the target label.

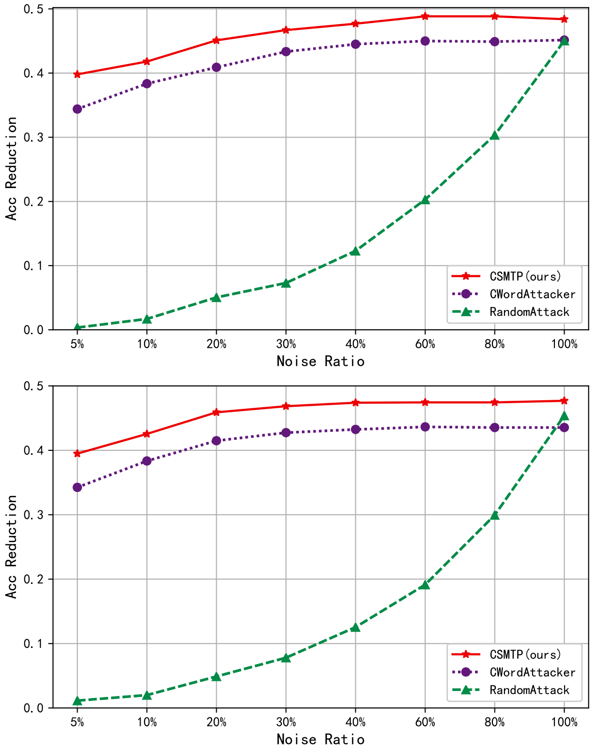


Figure 5: Influence of adversarial noise ratio on untargeted attack performance. Above: Word-level attacks. Below: Character-level Attacks

4.3 Hyperparameter Analysis

The attack performance of the CSMTP algorithm was not only affected by the features of the datasets, the target models and the attack modes but also by the noise intensity, the number of categories and other factors. Taking the TextCNN model with faster running efficiency as an example, with the weakening of the disturbance limit, the performance of the attack algorithm on the sentiment analysis task is shown in Figure 5. The performance of CSMTP, CWordAttacker, and random attacks consistently improved as the proportion of noise in the text increased, both in character-level and word-level attacks. However, CSMTP and CWordAttacker can locate key tokens more accurately thanks to the score query mechanism. After the noise ratio reached 30%, the attack performance gradually converged to the upper bound, which

Table 5: Targeted attack experiments against word-level models

(a) LSTM Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	0.10	12.20	12.10	73.00	**72.90**	73.00	**72.90**
Weibo rumor	17.20	17.80	0.60	30.00	12.80	39.60	**22.40**
E-com. review	6.59	7.09	0.50	21.79	15.20	31.25	**24.66**
THUCNews	1.06	1.13	0.07	9.46	**8.40**	8.83	7.77

(b) TextCNN Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	2.90	13.90	11.00	89.00	**86.10**	89.00	**86.10**
Weibo rumor	19.80	23.80	4.00	72.00	52.20	78.40	**58.60**
E-com. review	13.51	14.19	0.68	73.65	60.14	74.49	**60.98**
THUCNews	1.47	1.52	0.05	17.39	15.92	18.37	**16.90**

(c) Att-CNN Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	4.40	17.20	12.80	86.30	81.90	86.60	**82.20**
Weibo rumor	21.00	25.60	4.60	47.60	26.60	48.00	**27.00**
E-com. review	13.34	17.23	3.89	64.86	51.52	67.57	**54.23**
THUCNews	1.27	1.30	0.03	10.10	8.83	10.60	**9.33**

Table 6: Targeted attack experiments against character-level models

(a) LSTM Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	1.50	13.80	12.30	87.80	**86.30**	87.50	86.00
Weibo rumor	15.00	24.60	9.60	86.00	71.00	89.60	**74.60**
E-com. review	9.80	14.02	4.22	61.99	52.19	68.58	**58.78**
THUCNews	1.37	2.18	0.81	5.54	**4.17**	3.47	2.10

(b) TextCNN Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	2.50	13.20	10.70	88.90	86.40	89.40	**86.90**
Weibo rumor	21.20	22.20	1.00	56.40	35.20	58.20	**37.00**
E-com. review	10.81	25.68	14.87	76.18	65.37	77.53	**66.72**
THUCNews	0.70	2.62	1.92	7.04	**6.34**	6.87	6.17

(c) Att-CNN Model

Datasets	**Ori_ER**	**RandomAttack**		**CWordAttacker**		**CSMTP (ours)**	
		ER	**ASR**	**ER**	**ASR**	**ER**	**ASR**
Weibo sentiment	1.50	13.10	11.60	75.70	74.20	75.90	**74.40**
Weibo rumor	13.40	18.00	4.60	60.80	47.40	62.00	**48.60**
E-com. review	11.32	30.91	19.59	75.68	64.36	79.73	**68.41**
THUCNews	0.87	6.59	5.72	29.30	**28.43**	26.68	25.81

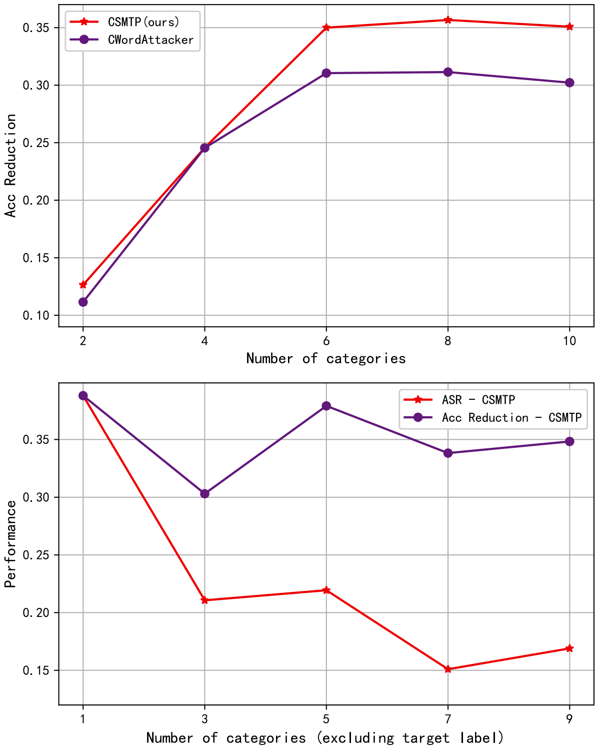


Figure 6: Influence of the number of categories on attack performance. Above: Untargeted Attacks. Below: Targeted Attacks

reduced the model classification accuracy to the benchmark level of random guessing. The noise added by random attacks was lack of pertinence, so even when the proportion of noise reached 80%, its attack effect was still far from that of CSMTP with a small perturbation proportion. Moreover, since CSMTP used a policy network that can match adversarial noise types according to the features of key tokens and target sentences, the attack performance was also better than the CWordAttacker algorithm based on a random transformation strategy.

We also additionally studied the effect of the number of categories on the attack success rate. In the experiment, we selected sub-datasets containing different categories from THUCNews to train TextCNN as the target models. The details of the target models and the experimental results are shown in Tables 7 and Figure 6, respectively.

Table 7: Original accuracy of TextCNN models

Categories	2	4	6	8	10
Accuracy	99.65	97.37	95.22	94.41	94.04

According to Figure 6, the performance of untargeted attacks was positively correlated with the number of categories when there were few categories in the datasets and gradually tended to a stable state after the categories increased to a certain number. At the same time, it also fluctuated due to the influence of data distribution of different categories. On the other hand, the success rate of targeted attacks decreased with the increase in the number of categories. Further comparison of the difference between the attack success rate and the decline of model accuracy showed that the reason for this phenomenon was that with the increase in the number of categories, the classification interval between the original label and the target label might increase due to the addition of other category intervals. As a result, more samples were classified into other categories between the original label and the target label. The principle is shown in Figure 7.

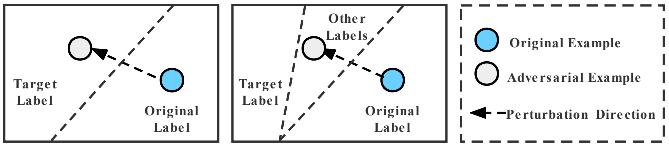


Figure 7: Adding categories may increase the classification gap between the original and target labels

5 Conclusions

Aiming at the vulnerability of the current Chinese social media text classification models, this paper proposes an adversarial examples generation method CSMTP based on reinforcement learning. Experiments show that CSMTP can launch effective targeted and untargeted attacks on Chinese social media text classification models such as LSTM and TextCNN trained on various datasets. Furthermore, the generated adversarial examples have the features of less noise and strong deception.

However, there are still gaps in our work. For example, CSMTP mainly relies on the scoring mechanism to locate key tokens, but as the sentence length increases, the TDS is gradually dispersed. This will lead to inaccurate positioning and affect the attack effect, resulting in the limited effectiveness of the algorithm when generating adversarial examples for long texts. At the same time, the social text classification API deployed in some practical scenarios often limits the access speed and times, and the key tokens location method based on query score may affect the efficiency of generating adversarial examples. Therefore, combining TDS with other positioning methods to further improve the efficiency of searching key tokens will be the next improvement direction of the CSMTP algorithm.

Acknowledgments

This research was funded by Henan Provincial Key R&D and Promotion Special Project (Science and Technology Tackling) (212102210165); National Social Science

Foundation Key Project (20AZD114); Henan Provincial Higher Education Key Research Project Program (20B520008); National Key Research and Development Program of China (2022YFC3300800). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," in *International Conference on Learning Representations*, 2018.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[3] X. Ding, T. Liu, J. Duan, and J.-Y. Nie, "Mining user consumption intention from social media using domain adaptive convolutional neural network," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2389–2395.

[4] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 31–36.

[5] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 50–56.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] S. Hochreiter and S. J, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2021–2031.

[9] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8018–8025.

[10] S. Karimi, A. Shakery, and R. M. Verma, "Enhancement of twitter event detection using news streams," *Natural Language Engineering*, pp. 1–20, 2022.

[11] J. D. M. W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[13] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," in *26th Annual Network and Distributed System Security Symposium*, 2019.

[14] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 4208–4215.

[15] M. Ling, Q. Chen, Q. Sun, and Y. Jia, "Hybrid neural network for sina weibo sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 983–990, 2020.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[17] N. Papernot, P. McDaniel, A. Swami, and R. Harang, "Crafting adversarial input sequences for recurrent neural networks," in *MILCOM 2016-2016 IEEE Military Communications Conference*. IEEE, 2016, pp. 49–54.

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[19] X. Tong, L. Wang, R. Wang, and J. Wang, "A generation method of word-level adversarial samples for chinese text classification (in chinese)," *Netinfo Secur*, vol. 20, no. 09, pp. 12–16, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] W. Wang, R. Wang, L. Wang, and B. Tang, "Adversarial examples generation approach for tendency classification on chinese texts (in chinese)," *Journal of Software*, vol. 30, no. 8, pp. 2415–2427, 2019.

[22] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, "Towards a robust deep neural network against adversarial texts: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[23] E. H. Xu, X.-L. Zhang, Y.-P. Wang, S. Zhang, L.-X. Liu, and L. Xu, "Adversarial examples generation method for chinese text classification," *International Journal of Network Security*, vol. 24, no. 4, pp. 587-596, 2022.

[24] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[25] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[26] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

Biography

Xin Tong is a Ph.D. student at the School of Information and Cyber Security, People’s Public Security University of China. He received his master’s degree at the People’s Public Security University of China in 2022. He is a committee member of Professional Committee of Computer Security of China Computer Federation. His current research interests include natural language processing and adversarial examples.

Jingya Wang received the master’s degree from Computer software and theory major, department of computer science, northwestern university, Xi’an China. She is currently a professor in School of Information Technology and Cyber Security, People’s Public Security University of China. Her research interests include natural language processing and deep learning.

Binjun Wang received the doctor’s degree from Computer software and theory major, department of computer science, northwestern university, Xi’an China. He is currently a professor in School of Information Technology and Cyber Security, People’s Public Security University of China. His research interests include natural language processing and AI security.

Hanming Zhai studied information security as an undergraduate and received her bachelor’s degree from Renmin University of China in 2021. She attended the School of Information and Cyber Security at the People’s Public Security University of China during her graduate studies. Her current research interests are mainly in natural language processing and knowledge graph.

Kaidi Zhang studied information security as an undergraduate and received his bachelor’s degree from People’s Public Security University of China in 2020. He attended the School of Information and Cyber Security at the People’s Public Security University of China during her graduate studies. His current research interests are mainly in distributed learning and multimodal sentiment.

Non-linear and Non-steady Time Series Forecasting Method Based on EMD and OSELM

Xuebin Xu, Meijuan An, Shuxin Cao, Longbin Lu, Liangxu Su, Tao Yang, and Jiaqi Luo
(Corresponding author: Xuebin Xu)

School of Computer Science, Xi'an University of Posts and Telecommunications
Chang'an District, Xi'an, Shaanxi Province, China
Email: xuxuebin@xupt.edu.cn

(Received June 7, 2022; Revised and Accepted Dec. 8, 2022; First Online Dec. 13, 2022)

Abstract

In reality, time series such as wind speed and stock price are always nonlinear and non-stationary and face significant forecasting challenges. To improve the prediction accuracy of time series, the EMD and OSELM models are updated. First, the time series is decomposed into components by EMD; then, an OSELM model is built on each component for forecasting, and the results are obtained by summing. Combining empirical mode decomposition (EMD) and online sequential extreme learning machine (OSELM), an improved method is proposed for forecasting nonlinear and unsteady time series. The wind speed dataset is used to test the forecasting model. EMD algorithm is used to decompose a time series into a finite number of intrinsic mode functions (IMFs) and a trend term to reduce the complexity of the time series. At the same time, OSELM is used to predict IMFs and a trend term, respectively, and the final results are added to each forecasting. Ultimately, the wind speed time series are predicted and analyzed, and the proposed model is compared with the traditional time series forecasting algorithms. The final results show that the improved model has a more vital generalization and accurate forecasting ability.

Keywords: Empirical Mode Decomposition; Intrinsic Mode Functions; Online Sequential Extreme Learning Machine; Time Series Forecasting

1 Introduction

As an important source of information, time series have attracted more attentions in many fields. Time series forecasting is widely used in water economics [11], medicine [1, 8], environmental protection [2], energy sources [12], and Internet [3], which are affected by many factors, such as large volatility, high complexity, strong nonlinearity and instability. How to effectively deal with complex time series and make the forecasting results more accurate has become an urgent problem. Currently, machine learning algorithm for time series forecasting has applied maturely ANN [20], SVM [19], KNN and Naive Bayes [16]. With the rapid development of neural network, they are widely used in time series forecasting. Chuang *et al.* [5] improved two moving average methods to see if we can find a better method that more accurately analyzes time series datasets to identify trends in past problems and possible future trends.

Shen *et al.* [15] used LSTM to predict trade trends and Tadjer *et al.* [18] proposed DeepAR to predict hydrocarbon production. However, machine learning algorithm and traditional neural networks still have some problems, such as long learning time, falling into local minima and so on [6, 9] . As the training time increases, the prediction results of the model become more and more inaccurate [13]. EMD has been widely used in the prediction of time series [4, 17]. According to data smoothing, the noise of the sequence can be effectively removed and the prediction accuracy can be improved. Extreme learning machine (ELM) was proposed by Huang [10] and widely used in various scenarios [21]. Ebermam *et al.* [7] combined EMD and ELM to predict time series. Based on the training results of ELM, OSELM algorithm updates the training data in real time and improves the accuracy of prediction. In order to improve the prediction accuracy of time series, the EMD and OSELM models are improved. Time series are decomposed into components by EMD, then OSELM model is built on each component for prediction and the final result is obtained by summing. The wind speed dataset is used to test the prediction model.

2 Algorithm

2.1 Empirical Mode Decomposition

The EMD algorithm can decompose a complex time series into several Intrinsic Mode Functions (IMFs) and a trend term, the expression is:

$$x(t) = \sum_{i=0}^{n-1} c_i(t) + r_n(t)$$

where $x(t)$ is the original series; $c_i(t)$ is the ith IMF; $r_n(t)$ is a trend term. The corresponding partial maximum and minimum points are obtained from the original data $x_0(t)$, and the upper and lower envelope lines $u_0(t)$ and $l_0(t)$ are calculated by the cubic spline interpolation method, as shown in Figure 1.

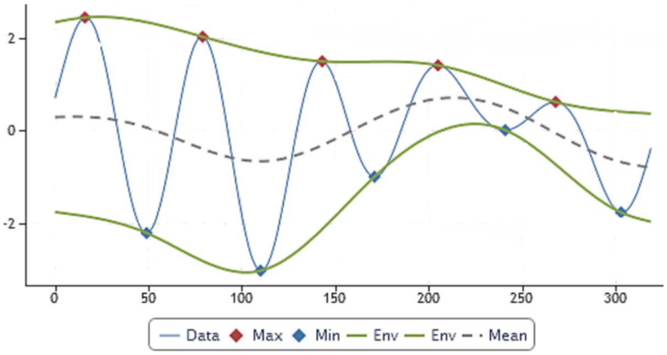


Figure 1: Envelope of the original signal

First, according to the upper and lower envelope lines $u_0(t)$ and $l_0(t)$, the average value is calculated to obtainthe average envelope line $m_0(t)$:

$$m_0(t) = \frac{u_0(t) + l_0(t)}{2}$$

After that, it is necessary to further calculate the difference between the average envelope line and the original series, i.e. the residual series $p_0(t)$:

$$p_0(t) = x(t) - m_0(t).$$

If the obtained $p_0(t)$ satisfies that the difference between zero points and extreme points of the entire series is at most 1, and $m_0(t) = 0$. It is considered that an intrinsic mode component $c_1(t)$ is obtained. If the above conditions are not satisfied, $p_0(t)$ would be as anew $x(t)$ to calculate the upper and lower envelope line, and the average envelope line and the residual series.This process is repeated until $c_1(t)$ is generated to satisfy the conditions. The residuals are found after obtaining the mode components:

$$r_1(t) = x(t) - c_1(t).$$

Residual $r_1(t)$ will be used as $x(t)$ in the next round. The previous process is repeated until $r_n(t)$ obtained is a monotonic function or a constant. The overall process is shown in Figure 2.

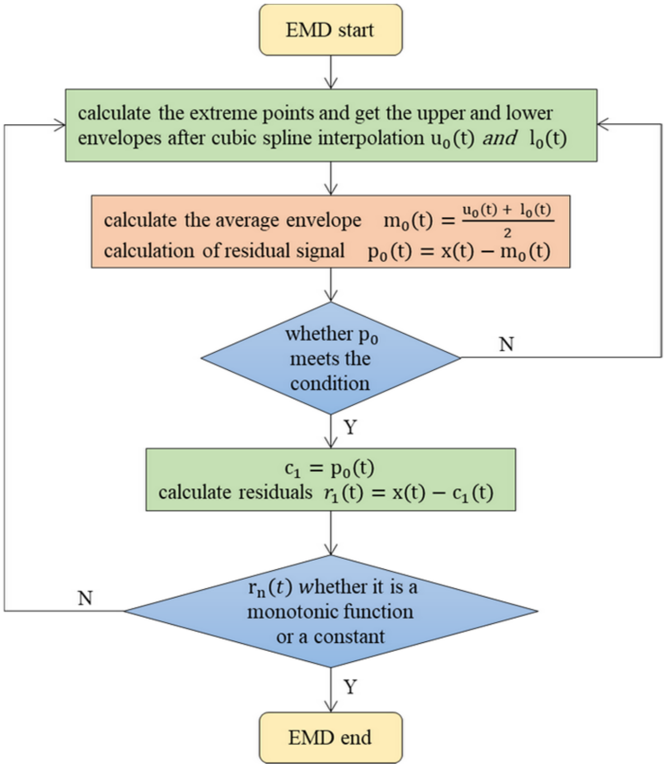

EMD start
calculate the extreme points and get the upper and lower envelopes after cubic spline interpolation $u_0(t)$ and $l_0(t)$
calculate the average envelope $m_0(t) = \frac{u_0(t) + l_0(t)}{2}$
calculation of residual signal $p_0(t) = x(t) - m_0(t)$
whether p_0 meets the condition
N
Y
$c_1 = p_0(t)$
calculate residuals $r_1(t) = x(t) - c_1(t)$
$r_n(t)$ whether it is a monotonic function or a constant
N
Y
EMD end

Figure 2: EMD decomposition flow chart

2.2 Extreme Learning Machine

ELM is proposed to improve the back propagation (BP) algorithm, which can improve learning efficiency and simplify parameter setting. The standard ELM is a single hidden layer feed forward neural networks (SLFNs) structure. The output function of SLFNs with L hidden nodes is given as:

$$f_L(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x), x \in R^n, a_i \in R^n$$

where a_i and b_i are the mapping parameters, and β_i is the output weights connecting the ith hidden layer and the output layer.

$G(a_i, b_i, x)$ is the sigmoid function of the ith hidden layer neuron of the input variable x.In supervised learning, there are N random samples $(x_i, t_i) \in R^n \times R^m$, where x_i is the input vector of $n \times 1$ and t_i is the target vector of $m \times 1$. There are β_i, a_i and b_i satisfying the following equation:

$$f_L(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = t_j, j = 1, \cdots, N$$

which can be abbreviated as:

$$H\beta = T.$$

H is the output matrix of the hidden layer, where the ith row is the output vector of the hidden node corresponding to x_i and the jth column is the output vector of the jth hidden node of the input $x_1, \cdots, x_n$.

The output weight matrix with the minimum error is solved by the following equation:

$$\min_{\beta \in R_{L \times m}} \|H\beta - T\|^2$$

Similar to an improved least square estimation method, it can be solved directly in the following way:

$$\hat{\beta} = H^{\psi} T$$

H^{ψ} is the Moore-Penrose generalized inverse. The following solution is obtained by orthogonal projection:

$$\hat{\beta}(c) = \left(H^T H + \frac{1}{c} \right)^{-1} H^T T$$

Generally, ELM sets random weights and thresholds between the input layer and the hidden layer to improve computational efficiency and increase generalization ability. Finally, the output layer performs regression using the regular or pseudo-inverse form.

2.3 Online Sequential Extreme Learning Machine

Sometimes the data is not always available at the beginning but arrives in blocks during training. In the previous algorithm, the output weights should be updated as the data increases, i.e. Online Sequential ELM (OSELM).

Given an initial training set containing $N_0 (\geq L)$ data, consider minimizing $\|H_0 \beta - T_0\|$ using a batch ELM, the solution tothe minimization problem is:

$$\begin{aligned} \beta^{(0)} &= K_0^{-1} H_0^T T_0 \\ K_0 &= H_0^T H_0 \end{aligned}$$

When another block containing N_1 data arrives, these two blocks are considered, and the output weights are:

$$\beta^{(1)} = K_1^{-1} \begin{bmatrix} H_0 \\ H_1 \end{bmatrix}^T \begin{bmatrix} T_0 \\ T_1 \end{bmatrix} \quad (1)$$

$$K_1 = \begin{bmatrix} H_0 \\ H_1 \end{bmatrix}^T \begin{bmatrix} H_0 \\ H_1 \end{bmatrix} \quad (2)$$

In order to update the weights, $\beta^{(1)}$ is defined as a function of $\beta^{(0)}, K_1, H_1, T_1$, written as:

$$K_1 = \begin{bmatrix} H_0^T & H_1^T \end{bmatrix} \begin{bmatrix} H_0 \\ H_1 \end{bmatrix} = K_0 + H_1^T H_1 \quad (3)$$

$$\begin{aligned} \begin{bmatrix} H_0 \\ H_1 \end{bmatrix}^T \begin{bmatrix} T_0 \\ T_1 \end{bmatrix} &= H_0^T T_0 + H_1^T T_1 \\ &= K_0 K_0^{-1} H_1^T T_0 + H_1^T T_1 \\ &= K_0 \beta^{(0)} + H_1^T T_1 \\ &= (K_1 - H_1^T H_1) \beta^{(0)} + H_1^T T_1 \\ &= K_1 \beta^{(0)} - H_1^T H_1 \beta^{(0)} + H_1^T T_1 \end{aligned} \quad (4)$$

Combined Equation (2) with Equation (5), we can obtain

$$\begin{aligned} \beta^{(1)} &= K_1^{-1} \begin{bmatrix} H_0 \\ H_1 \end{bmatrix}^T \begin{bmatrix} T_0 \\ T_1 \end{bmatrix} \\ &= K_1^{-1} (K_1 \beta^{(0)} - H_1^T H_1 \beta^{(0)} + H_1^T T_1) \\ &= \beta^{(0)} + K_1^{-1} H_1^T (T_1 - H_1 \beta^{(0)}) \end{aligned} \quad (5)$$

where $K_1 = K_0 + H_1^T H_1$. When new data arrives, the recursive algorithm and corresponding parameters of the least square solution are updated. When the $k+1$ data block is received, the parameter can be updated to

$$K_{K+1} = K_K + H_{K+1}^T H_{K+1}$$

$$\beta^{(k+1)} = \beta^{(k)} + K_{k+1}^{-1} H_{k+1}^T (T_{k+1} - H_{k+1} \beta^{(k)})$$

where

$$T_{k+1} = \begin{bmatrix} t^T \left(\sum_{j=0}^{k} N_j \right) + 1 \\ \vdots \\ t^T \sum_{j=0}^{k+1} N_j \end{bmatrix}$$

$$H_{k+1} = \begin{bmatrix} G(a_1, b_1, x_{(\sum_{j=0}^{k} N_j)+1}) & \cdots & G(a_L, b_L, x_{(\sum_{j=0}^{k} N_j)+1}) \\ \vdots & \cdots & \vdots \\ G(a_1, b_1, x_{\sum_{j=0}^{k} N_j}) & \cdots & G(a_L, b_L, x_{\sum_{j=0}^{k} N_j}) \end{bmatrix}_{N_{k+1} \times L}$$

The above formula derivation indicates that OSELM [14] can sample the training data in the form of piece of incremental learning. At any time, when the new data involves in learning, as long as the learning process is complete, the data can be discarded immediately and no longer used in order to reduce the waste of space resources. In addition, in online learning, the update calculation of the model is completed recursively based on the results of last iteration and the latest arrived data, and there is no need to save and relearn the previous training samples, which greatly reduces the storage and computation costs. Many experiments and applications also show that compared with other popular online learning algorithms,OSELM not only has better generalization ability, but also has obvious advantages in learning speed.

2.4 EMD-OSELM Forecasting Process

The simulation forecasting of the EMD-OSELM model is followed, and the process is shown in Figure 3.

Step 1. Decompose the time series using EMD to obtain multiple IMFs and trend term RES.

Step 2. Use the OSELM method to predict each IMF and trend term RES.

Step 3. Sum the prediction results of each component to obtain the final prediction results.

Step 4. Update the time series and the corresponding parameters.

3 Simulation Analysis

3.1 Data Preprocessing

We selected the wind speed monitoring data for a certain period of time in a certain place in China and recorded it at 10-second intervals. A total of 8784 data points were obtained as shown in Figure 4.

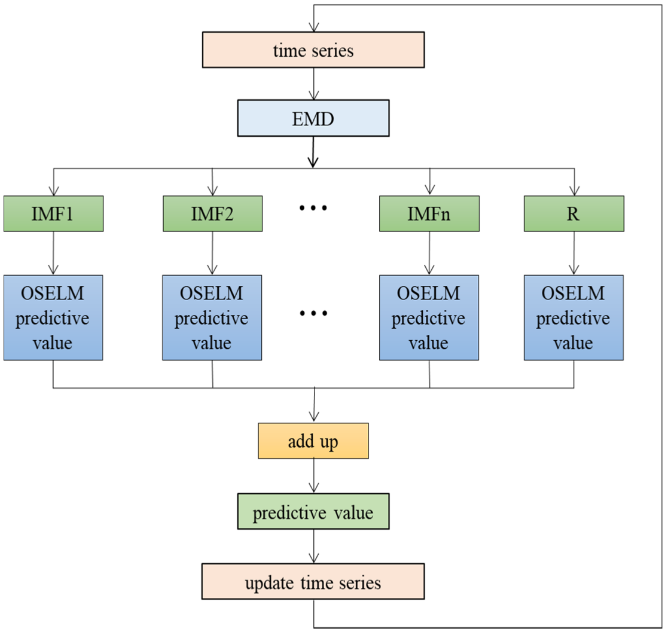


Figure 3: Flow chart of time series forecasting method based on EMD and OSELM

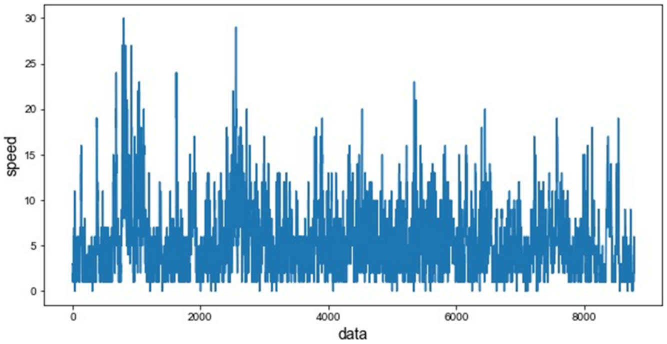


Figure 4: Time series of wind speed

It can be seen from Figure 4 that the original sequence of wind speed has strong nonlinearity and instability, with large fluctuations and high complexity. Using the EMD algorithm to process the original wind data sequence can remove noise and reduce the complexity of the sequence. The result after decomposition is shown in Figure 5.

After constructing the dataset, the first 20% of the data is selected as the training set to establish the ELM model. The rest data will be divided into five groups(each group contains 1300 data points), and then be added into the model in proper order to update the parameters.The last 500 data points in the dataset are selected as the test set to obtain the final results by predicting a trend term and each IMF component. The division of the dataset is shown in Figure 5.

The prediction results of each component are added to obtain the final forecasting results,as shown in Figure 6.

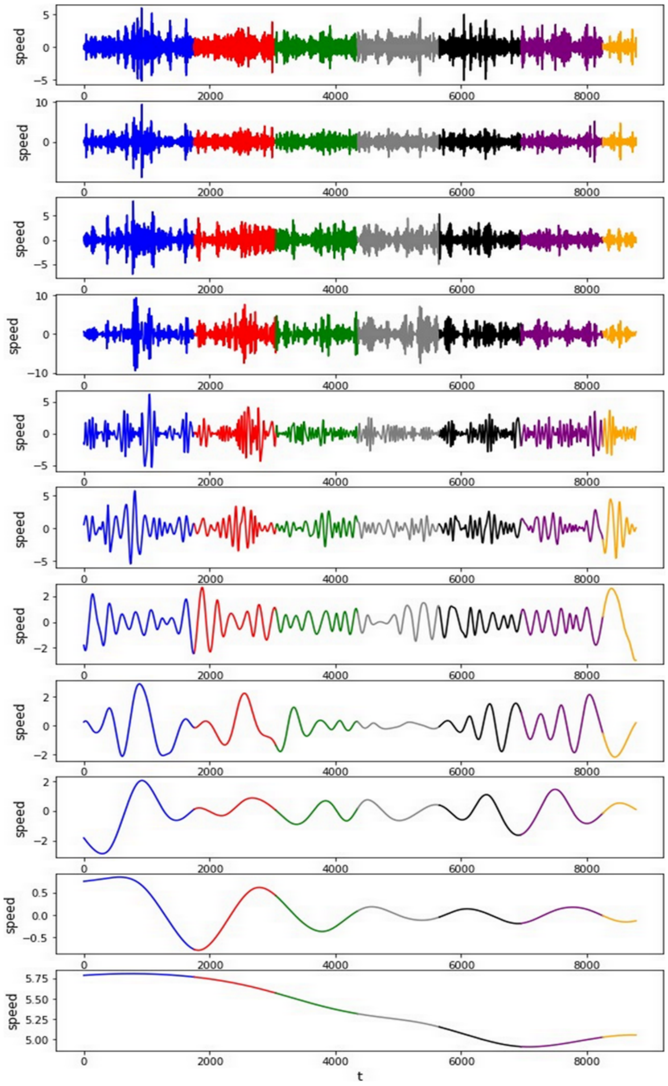


Figure 5: Update data for each time series after EMD decomposition

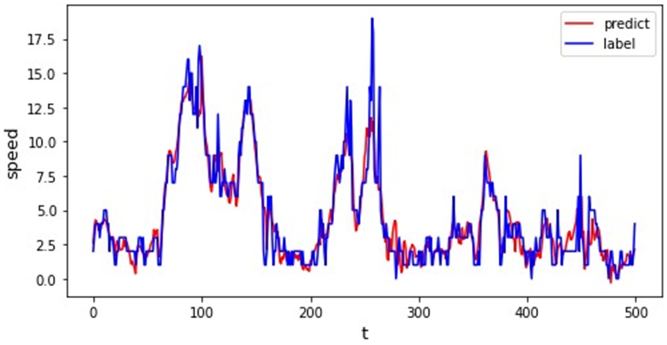


Figure 6: Final prediction result after addition

3.2 Algorithm Performance Analysis

Root mean squared error (RMSE), mean absolute error (MAE) and mean relative error (MRE) are used as indicators to measure the forecasting ability of the EMD-OSELM algorithm for time series.The complete time series sequence is divided into multiple sample blocks to simulate the scenario where data arrives in batches under real conditions, and the model is updated with new data at each time it arrives. After the empirical mode of the data is decomposed, the influence of updates on RMSE, MAE and MRE of the model is shown in Table 1.

Table 1: Forecasting index for different update sample times

Updates	RMSE	MAE	MRE
0	1.540946	1.141189	0.494791
1	1.440463	1.013902	0.390861
2	1.377619	0.978138	0.371081
3	1.342217	0.934270	0.340725
4	1.331585	0.924462	0.328778
5	1.322387	0.918793	0.329281

Table 1 shows the prediction indicators calculated from the prediction results according to the updates. The more data is updated, the more accurate the prediction results are, thus OSELM is better than ELM.

In order to verify the accuracy of the EMD-OSELM algorithm, ELM, OSELM, LSTM, BP, EMD-ELM, EMD-LSTM, EMD-BP are used to establish a prediction model for the original data sequence, and RMSE, MAE and MRE are shown in Table 2.

Table 2: Predictive indicators obtained by seven methods

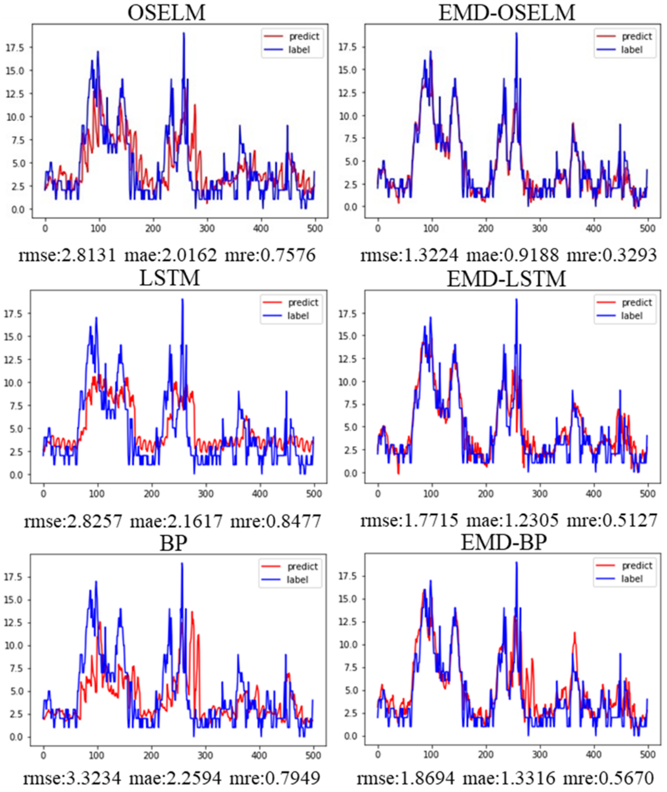
Model	**RMSE**	**MAE**	**MRE**
ELM	2.954801	2.187337	0.843599
OSELM	2.813081	2.016215	0.757628
EMD-OSELM	**1.322387**	**0.918794**	**0.329281**
LSTM	2.825711	2.161665	0.847699
EMD-LSTM	**1.771473**	**1.230466**	**0.512658**
BP	3.323371	2.259400	0.794925

According to the indicators in Table 2 and Figure 7, OSELM performs better than BP and LSTM, and EMD can effectively reduce the nonlinear and unsteady degree of time series to improve the performance of various models.

Considering the arrival of time series samples in blocks, the OSELM method, which can be learned online, is better than the end-to-end LSTM and BP methods. The experimental results also confirm this.

The EMD algorithm is used to decompose the sequence to reduce the complexity of the original sequence and facilitate model learning. Thereby the EMD algorithm can significantly improve the prediction accuracy.

Compared with the EMD-LSTM model and the EMD-BP model, the EMD-OSELM model has more accurate prediction ability and lower complexity. ELM has a great advantage in training speed (See Table 3). After new data is obtained, the OSELM model can be quickly updated online, while LSTM and BP need to integrate the new added data and retrain the model with a lot of time.



rmse:2.8131 mae:2.0162 mre:0.7576 rmse:1.3224 mae:0.9188 mre:0.3293

rmse:2.8257 mae:2.1617 mre:0.8477 rmse:1.7715 mae:1.2305 mre:0.5127

rmse:3.3234 mae:2.2594 mre:0.7949 rmse:1.8694 mae:1.3316 mre:0.5670

Figure 7: The prediction results of each model

Table 3: Training time of three models

Training Time (s)	Time 1	Time 2	Time 3	Average
EMD-OSELM	0.168940	0.230158	0.180210	0.193103
EMD-LSTM	146.327691	151.012072	154.613179	150.650981
EMD-BP	54.716220	55.605869	54.357104	54.893064

4 Conclusion

The wind speed data set is used to test the prediction model. Combined EMD with OSELM, an improved method for nonlinear and unsteady time series prediction is proposed. EMD algorithm is used to decompose the time series into finite IMFs and a trend term to reduce the complexity of time series. OSELM is used to predict IMFs and the trend term respectively, and the final results are added to each prediction result. Then the wind speed time series are predicted and analyzed to compare with traditional time series forecasting algorithms. The final results show that the model has strong generalization ability and more accurate forecasting results.

Using the data of wind speed time series and the measure indicators containing RMSE, MAE and MRE, the EMD-OSELM model and algorithms such as LSTM and BP are commonly used to analyze time series. The results of the analysis are as follows:

1) For a single prediction model, the EMD algorithm can effectively improve the accuracy of the model;

2) For the decomposed sequence, the OSELM algorithm

has a higher accuracy rate than the commonly used neural network-based LSTM and BP algorithm;

3) EMD-OSELM model has a faster training speed than the LSTM model and consumes fewer resources.

The above results show that the EMD-OSELM model has a good performance in the forecasting of time series, and also has potential applications in engineering practice.

Acknowledgments

This paper work was supported by the National Natural Science Foundation of China (No. 61673316); by the department of Education Shaanxi Province, China, under Grant16JK1697; by the Key Research and Development Program of Shaanxi province, under Grant 2017GY-071; by the Technical Innovation Guidance Special Project of Shaanxi province, under Grant 2017XT-005; by the research program of Xianyang city under Grant 2017K01-25-3. This paper is also supported by the key discipline special fund construction project of Shaanxi Province.

References

[1] M. Abdullah, K. Kolo, P. Aspoukeh, R. Hamad, and J. R. Bailey, "Time series modelling and simulating the lockdown scenarios of covid-19 in kurdistan region of iraq.," *Journal of infection in developing countries*, vol. 15, no. 3, pp. 370–381, 2021.

[2] M. Akyol and E. Ucar, "Carbon footprint forecasting using time series data mining methods: the case of turkey," *Environmental Science and Pollution Research*, vol. 28, pp. 1–11, 08 2021.

[3] A. Badr, T. Makarovskikh, P. Mishra, M. Abotaleb, A. M. G. A. Khatib, K. Karakaya, S. Redjala, A. Dubey, and E. Attal, "Modelling and forecasting of web traffic using holt's linear, bats and tbats models," *Journal of Mathematical and Computational Science*, 2021.

[4] T. Cha and H. Xue, "A study on ship collision conflict prediction in the taiwan strait using the emd-based lssvm method," *PLoS ONE*, vol. 16, 2021.

[5] C. H. Chuang, W. F. Lu, Y. C. Lin, and J. C. Chen, "Visual exploration for time series data using multivariate analysis method," *Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013*, pp. 1189–1193, 04 2013.

[6] C. W. Deng, G. B. Huang, J. Xu, and J. X. Tang, "Extreme learning machines: new trends and applications," *Science China Information Sciences*, vol. 58, pp. 1–16, 2014.

[7] E. Ebermam, G. G. De Angelo, H. Knidel, and R. A. Krohling, "Empirical mode decomposition, extreme learning machine and long short-term memory for time series prediction: A comparative study," *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 492–497, 2018.

[8] S. Govindarajan and R. Swaminathan, "Extreme learning machine based differentiation of pulmonary tuberculosis in chest radiographs using integrated local feature descriptors," *Computer Methods and Programs in Biomedicine*, vol. 204, p. 106058, 2021.

[9] G. B. Huang, N. Y. Liang, H. J. Rong, P. Saratchandran, and N. Sundararajan, "Online sequential extreme learning machine with the increased classes," *Computers & Electrical Engineering*, vol. 90, p. 107008, 2021.

[10] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006.

[11] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1, pp. 307–319, 2003. Support Vector Machines.

[12] W. Liu, W.D. Liu, and J. W. Gu, "Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network," *Journal of Petroleum Science and Engineering*, vol. 189, p. 107013, 2020.

[13] H. H. Nguyen and C. W. Chan, "Multiple neural networks for a long term time series forecast," *Neural Computing & Applications*, vol. 13, pp. 90–98, 2003.

[14] Y. Qiao, Z. He, and X. Zhao, "Research on improved oselm algorithm for dynamic data flow," *Journal of Beijing Electronic Science and Technology Institute*, vol. 28, no. 3, pp. 1–12, 2020.

[15] M. L. Shen, C. F. Lee, H. H. Liu, P. Y. Chang, and C. H. Yang, "Effective multinational trade forecasting using lstm recurrent neural network," *Expert Systems with Applications*, vol. 182, p. 115199, 2021.

[16] B. Soepriyanto, "Comparative analysis of k-nn and naïve bayes methods to predict stock prices," *International Journal of Computer and Information System*, vol. 2, no. 2, pp. 49–53, 2021.

[17] W. Sun and C. M. Ren, "Short-term prediction of carbon emissions based on the eemd-psobp model," *Environmental Science and Pollution Research*, vol. 28, pp. 56580 – 56594, 2021.

[18] A. Tadjer, A. J. Hong, and R. B. Bratvold, "Machine learning based decline curve analysis for short-term oil production forecast," *Energy Exploration & Exploitation*, vol. 39, pp. 1747 – 1769, 2021.

[19] F. E. Tay and L. J. Cao, "Application of support vector machines in financial time series forecasting," *Omega-international Journal of Management Science*, vol. 29, pp. 309–317, 2001.

[20] C.P. Tsai and T. L. Lee, "Back-propagation neural network in tidal-level forecasting," *Journal of Waterway Port Coastal and Ocean Engineering-asce*, vol. 125, pp. 195–202, 2001.

[21] Y. F. Xu, S. Zhang, Z. T. Cao, Q. Q. Chen, and W. D. Xiao, "Extreme learning machine for heartbeat classification with hybrid time-domain and wavelet time-frequency features," *Journal of Healthcare Engineering*, vol. 2021, 2021.

Biography

Xuebin Xu is mainly engaged in the research of artificial intelligence and biometric recognition.

Meijuan An is mainly engaged in the research of artificial intelligence and medical image processing.

Shuxin Cao is mainly engaged in the research of artificial intelligence and human ear recognition.

Longbin Lu is mainly engaged in the research of artificial intelligence and biometric recognition.

Liangxu Su is mainly engaged in the research of artificial intelligence and biometric recognition.

Tao Yang is mainly engaged in the research of artificial intelligence and biometric recognition.

Jiaqi Luo is mainly engaged in the research of artificial intelligence and biometric recognition.

Privacy-preserving Electronic Medical Records Sharing Solution Based on Blockchain

Mingqiang Shao, Momeng Liu, and Zhenzhen Wang

(Corresponding author: Momeng Liu)

Shaanxi Key Laboratory of Clothing Intelligence, State and Local Joint Engineering Research Center for Advanced Networking and Intelligent Information Services, School of Computer Science, Xi'an Polytechnic University
No. 58 Shan-gu Road, Lintong, Xi 'an 710600, China
Email:liumomeng@163.com

(Received July 27, 2022; Revised and Accepted Dec. 9, 2022; First Online Dec. 13, 2022)

Abstract

In the era of big data, the extensive applications of electronic medical records (EMRs) improve the efficiency of consultations and patient data management. However, it may cause privacy leakage and data damage during inappropriate uses. Fortunately, the advantages of blockchain, such as anonymity, decentralization, and immutability, can provide a favorable platform for health data sharing. This paper proposes a privacy-preserving EMRs sharing scheme based on blockchain. In our scheme, we offer the fine-grained access control property using an identity-based signcryption scheme and ensure no single point of failure through the interplanetary file system (IPFS). Furthermore, for traceability, smart contracts and blockchains are jointly used to record the processes of data storage and access. In the end, we make a security analysis of our proposed scheme to illustrate that it can meet data privacy and integrity requirements. Moreover, we conducted several experiments by timing the consumption of data encryption and decryption to indicate that our work has a lower computation cost than other related works.

Keywords: Blockchain; EMRs; Identity-based Signcryption; IPFS; Smart Contracts

1 Introduction

The development of information technology promotes the applications of electronic medical records (EMRs) in health system. In a nutshell, EMRs are digital records that providers use to record patient health data. Compared with paper-based records, they are more efficient at managing payments, scheduling patient visits, sharing information and recording patient health data [7]. EMRs can accurately document a patient's medical history so as to reduce misdiagnoses and provide appropriate care. If used properly, it can not only improve the cure rate, but also stimulate the development of disease research. However, once the data center is attacked, EMRs may be used illegally, resulting in privacy leakage. Therefore, it is crucial to realize the secure storage and fine-grained access control of EMRs.

At present, most healthcare institutions still manage EMRs with centralized systems, which are prone to single points of failure [13]. In addition, data between various institutions is independent of each other, resulting in serious data fragmentation and difficulty in sharing [10]. Furthermore, treatment records and inspection data are not immutable, unforgeable and traceable, causing it hard to gain the trust of workers in different institutions [8]. Therefore, it directly leads to repeated examinations when patients visit different medical institutions, which not only wastes medical resources, but also brings additional burdens to patients.

The merits of blockchain, e.g. anonymity, decentralization and immutability, boost the rapid development of the study on EMRs [21, 22, 29]. The works of [4, 15] propose to use blockchain for data storage to ensure traceability. However, when the amount of data is huge, this method cannot satisfy both efficient storage and data retrieval. Targeting to the problem, the works of [6, 24, 26] propose to outsource encrypted EMRs to a third-party cloud server for management, and use blockchain for retrieval. Although it enhances storage scalability and provides efficient data retrieval, the entire system will be paralyzed once cloud servers fail. To solve this problem, the works of [12, 19] propose to combine interplanetary file system (IPFS) with blockchain. Considering IPFS is a decentralized storage protocol, the single points of failure can be avoided. However, it cannot ensure fine-grained access control. Aiming at the problem, the works of [5, 23] propose to use attribute-based encryption to achieve. The medical data is encrypted according to attributes of users, thereby determining who can decrypt it. However, the time for key generation, data encryption and decryption grows rapidly as the increase of users' attributes, resulting in low retrieval efficiency.

Focusing on the above problems, we propose a privacy-preserving EMRs sharing solution based on blockchain. To summarize, our contributions are as follows:

1) In order to avoid single points of failure, we use IPFS as a storage platform to store encrypted data. Further more, smart contracts are used to upload retrieval index returned by IPFS to the blockchain.

2) We provide fine-grained access control by using an identity-based signcryption scheme.

3) All storage and access requests initiated by patients or users will be recorded in the blockchain through smart contracts, enabling traceability.

2 Preliminaries

In this section, we briefly introduce some technical backgrounds involved in our scheme.

2.1 Blockchain

Blockchain [16, 28, 30] is a technology that emerged with the increasing popularity of cryptocurrencies such as bitcoin [18]. It records all past transactions and historical data by establishing a database jointly maintained by nodes on the network. The data will be stored in a distributed manner and cannot be tampered with. Any user on the network can reach a credit consensus through contracts or digital signatures without any central authority. A blockchain is essentially a series of linked blocks (e.g., $Block_{j-1}$, $Block_j$, $Block_{j+1}$), as shown in Figure 1. Each block consists of a block header and a block body. The block header includes a block version indicating which set of block verification rules to follow, a hash value pointing to the previous block and a hash value of all transactions (e.g., $TX_1, ..., TX_i$) in the block called Merkle Root. The block body consists of all transactions. Block size and the size of each transaction determine the maximum transactions that a block can contain.

2.2 Smart Contract

A smart contract [2, 3] is an automatically-executing contract for trusted transactions and agreements among disparate, anonymous parties without the need for a central authority. The terms of the agreement between buyer and seller are written directly into lines of code. The code and agreements contained therein exist across a distributed, decentralized blockchain network. Codes control the execution of transactions, which are traceable and irreversible. Ethereum [25] is currently the most widely used smart contract platform with the following features:

Distribution. Smart contracts are replicated and distributed across all nodes in the Ethereum network.

Consistency. Only pre-set operations will be executed, and the results of any node execution are consistent.

Automaticity. Smart contracts will automatically execute when pre-set conditions are met.

Immutability. It cannot be changed once a transaction occurs, which will be permanently recorded on the chain.

2.3 IPFS

Interplanetary file system (IPFS) [1] is a protocol for storing and sharing in a distributed system. Any resources stored in the system will generate a retrieval index, through which the stored resources can be quickly found. Due to the protection of the cryptographic algorithm, the index is immutable and undeleteable. Once stored in IPFS, it is permanently. In addition, IPFS has a deduplication mechanism, which can effectively avoid duplication of data storage and save storage space. Considering blockchain is not suitable for storing large amounts of data, in this paper, we utilize IPFS to store encrypted EMRs of patients.

2.4 Intractable Problems

We briefly review two intractable problems in this part.

Elliptic Curve Discrete Logarithm (ECDL) Problem [11]: Given a prime number p, consider the equation $Q = kP$ on the Abel group $E_p(a, b)$ formed by elliptic curves, where $P, Q \in E_p(a, b)$, $k < p$. It can easily get Q from k and P, but difficult to compute k from Q and P. If there is an algorithm $\mathcal{A}$ that can solve the *ECDL problem* in polynomial time, the corresponding probability $Adv^{ECDL}(\mathcal{A}) = \Pr[\mathcal{A}(P, Q) = k]$ is negligible.

Computational Diffie-Hellman (CDH) Problem [9]: Let q be a large prime number, and G be a cyclic group on an elliptic curve with order q. P is a generator in G. Given a tuple (P, aP, bP) for a randomly chosen generator P and random $a, b \in Z_q^*$, the *CDH problem* is to compute abP. If there is an algorithm $\mathcal{A}$ that can solve the *CDH problem* in polynomial time, the corresponding probability $Adv^{CDH}(\mathcal{A}) = \Pr[\mathcal{A}(P, aP, bP) = abP]$ is negligible.

2.5 Identity-based Signcryption

The notion of identity-based signcryption is first proposed by Malone-Lee in [14], which consists of four polynomial-time algorithms (*Setup*, *KeyGen*, *Signcrypt*, *Unsigncrypt*) such that

$params \leftarrow Setup(\lambda)$. Take as input a security parameter λ and output the system parameters $params$.

$(sk_{ID}, pk_{ID}) \leftarrow KeyGen(ID, params)$. Take as input a string ID represents identity and system parameters $params$, and output a key pair (sk_{ID}, pk_{ID}).

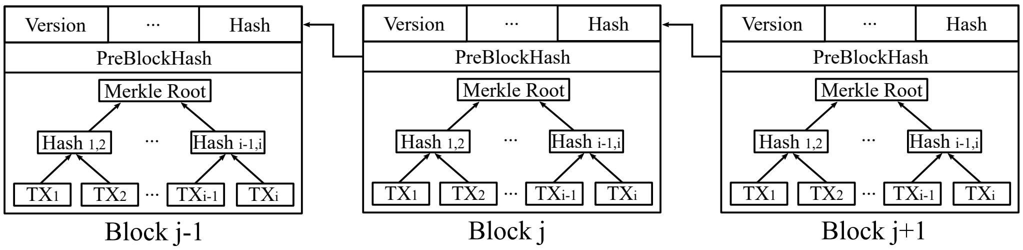


Figure 1: Example of blockchain

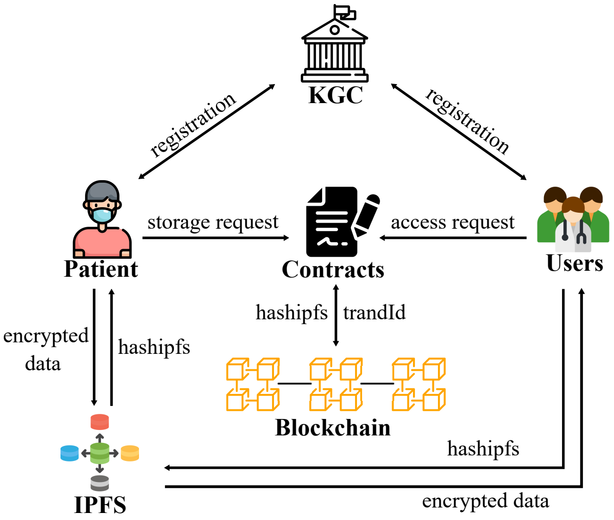


Figure 2: System model of our scheme

$\sigma \leftarrow Signcrypt\left(sk_{ID_a}, \{pk_{ID_i}\}, \{ID_i\}, m\right)$. Take as input private key sk_{ID_a}, public key groups $\{pk_{ID_i}\}$, identity groups $\{ID_i\}$ and message m to generate the ciphertext σ as output.

m/$\perp \leftarrow Unsigncrypt\left(ID_a, pk_{ID_a}, sk_{ID_i}, \sigma\right)$. Take as input ID_a, public key pk_{ID_a}, secret key sk_{ID_i} and the ciphertext σ, and output message m or $\perp$. The symbol $\perp$ indicates that the ciphertext was invalid.

3 System Model and Security Requirements

In this section, we describe how does the participants involved in our scheme work and give the security requirements it needs to meet.

3.1 System Model

The system model of our scheme is shown in Figure 2, which involves the following entities.

KGC. As a fully trusted third party, responsible for generating system parameters and keys for registrants (patient or users).

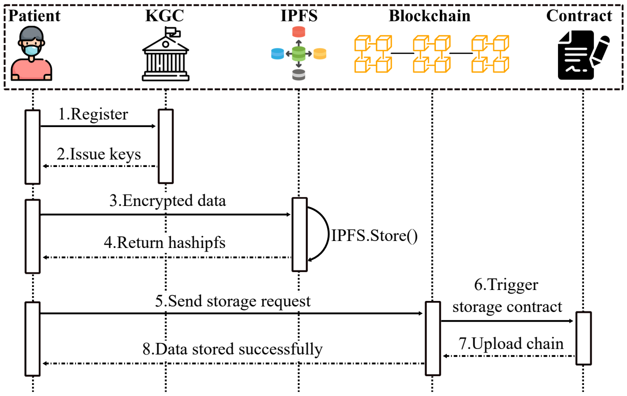


Figure 3: Workflow of patient storage

Patient. As owner of electronic medical records (EMRs), patient encrypts his/her EMRs to realize access control.

Users. As demanders of patient's EMRs, users (e.g. doctors, nurses, medical research institutions) request to obtain encrypted EMRs.

IPFS. As the storage platform for EMRs. A patient uploads encrypted EMRs to IPFS, which then returns a retrieval index.

Smart contract. Smart contracts are deployed on blockchain. If a patient initiates a storage request or access request, the storage contract or access contract will be triggered.

Blockchain. If a smart contract on the blockchain is triggered, the storage or access operations will be permanently recorded on the blockchain.

Under the system model in Figure 2, our solution is divided into two scenarios, patient storage and users access.

Patient storage. Patient first joins the system through the KGC. Before sharing medical data, the patient encrypts the data and predetermines who can access it. Then the patient uploads encrypted data to IPFS, which will return an index to retrieve the data.Nextly, the patient initiates a storage request

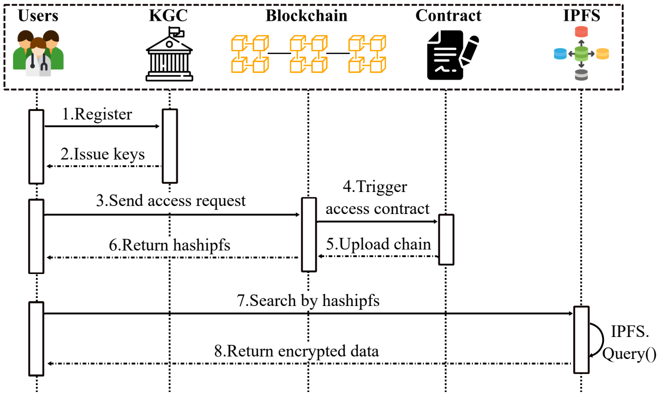


Figure 4: Workflow of users access

to the blockchain, which will trigger the storage contract and store the index on the blockchain. The whole process is shown in Figure 3.

Users access. Firstly, users join the system through KGC. If a user tries to access data, he/she needs to initiate an access request to the blockchain, which will trigger the access contract and return the index stored by patient. Through this index, encrypted data stored in IPFS can be obtained. Authorized users can decrypt the encrypted data to obtain the original data. The whole process is shown in Figure 4.

3.2 Security Requirements

Considering the sensitivity and privacy of EMRs, our scheme is supposed to meet the following four security requirements.

Confidentiality. EMRs should be encrypted before being uploaded to IPFS, and patient' private information cannot be disclosed to any third party.

Integrity. The index uploaded to the blockchain cannot be altered, and the authenticity of the data uploader should be verifiable.

Non-repudiation. The patient cannot deny the data they had uploaded, nor can users deny the data they had requested.

Anonymity. No one except the predetermined users can obtain the real identity of the patient from the intercepted ciphertext.

4 The Proposed EMRs Sharing Solution

In this section, we describe our scheme and security analysis in detail.

4.1 Detailed Execution of Our Scheme

The important notations used in our scheme are summarized in Table 1. The detailed execution consists of six phases.

Table 1: Notations and descriptions

Notations	Descriptions
λ	Security Parameter
msk	Master Secret Key
mpk	Master Public Key
ID_{pts}	Identity of patient
$\{ID_1, ID_2, ..., ID_n\}$	Set of n users,$i = 1, ..., n$
(X_i, Y_i)	User's public key pair
(x_i, y_i)	User's private key pair
MD	Patient's medical data
H_1, H_2, H_3, H_4	Cryptographic hash functions
$hashipfs$	Ciphertext retrieval index in IPFS
$\oplus$	Exclusive-OR (XOR) operation

System initialization phase. Given a security parameter λ, KGC generate two prime numbers p and q, where p, q satisfy the condition $q|p-1$. G is a cyclic group on an elliptic curve with order q, and P is a generator of G. The KGC randomly chooses msk as the master private key, where $msk \in Z_q{}^*$, and set mpk as the master public key, where $mpk = msk \cdot P$. KGC then selects four cryptographic hash functions, respectively $H_1 : \{0,1\}^l \times G \times G \to Z_q{}^*$, $H_2 : \{0,1\}^* \times G \to Z_q{}^*$, $H_3 : \{0,1\}^* \to \{0,1\}^u$, $H_4 : G \times G \times G \to Z_q{}^*$, where l represents the length of identity and u represents the length of the data. In addition, KGC defines a special function $F(A, ID)$ with inputs A and ID, where $A \neq 0$ and $ID \in \{0,1\}^l$. When $ID = \emptyset$, $F(A, ID)$ outputs 0, otherwise outputs A. KGC then publishes system parameters $params = (p, q, G, P, mpk, H_1, H_2, H_3, H_4, F)$.

Key generation phase. Users or patient interacts with KGC to generate keys. Through this phase, the registration in the system is completed. Taking a user as an example, details are as follows:

a. A user ID_i randomly selects the secret value $x_i \in Z_q{}^*$ as the first part of the private key, and computes $X_i = x_i \cdot P$ to get the first part of the public key.

b. The user sends the identity ID_i and the X_i obtained in the previous step to KGC.

c. When getting ID_i and X_i, KGC then randomly selects the secret value $z_i \in Z_q{}^*$ and computes $Y_i = z_i \cdot P$, which is another part of public key. Next, computing $y_i = z_i + msk \cdot H_1(ID_i, X_i, Y_i)$, y_i will be used as another part of private key.

d. KGC sends y_i to the user ID_i through a secure channel and stores the user's public key (X_i, Y_i) in a public table.

e. When the user ID_i gets y_i, he/she can check the validity of y_i by verifying the formula

$$y_i \cdot P = Y_i + H_1(ID_i, X_i, Y_i) \cdot mpk$$

If the formula holds, the private key is (x_i, y_i) and the public key is (X_i, Y_i).

Patient signcryption phase. Suppose a patient ID_{pts} be the data owner, $Reveivers = \{ID_1, ID_2, ..., ID_n\}$ be the n users predetermined by the patient. Assuming that the EMRs to be encrypted is $MD \in \{0,1\}^u$, patient ID_{pts} generate ciphertext through this phase. Details are as follows:

a. Through registering in the system, the patient gets his/her public key (X_{pts}, Y_{pts}) and private key (x_{pts}, y_{pts}).

b. The patient ID_{pts} randomly selects a secret value $d \in Z_q{}^*$, then computes $D = d \cdot P$.

c. For $i = 1, ..., n$, the patient ID_{pts} compute $J_i = x_{pts}(X_i + Y_i + h_i \cdot mpk)$ and $T_i = H_4(J_i)$, where $h_i = H_1(ID_i, Y_i, X_i)$.

d. Patient ID_{pts} randomly selects value $t \in Z_q{}^*$, and constructs a polynomial $f(x)$ with degree n as follows:

$$\begin{aligned} f(x) &= \prod_{i=1}^{n}(x - T_i) + t(mod\ q) \\ &= a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n \end{aligned}$$

where $a_i \in Z_q{}^*$.

e. The patient ID_{pts} then computes

$$V = F(\frac{x_{pts} + y_{pts}}{dR}, ID_{pts})$$

$$W = F(H_3(t), ID_i)$$

where $R = H_2(ID_{pts}, MD, D)$.

f. The patient ID_{pts} computes $C = W \oplus MD$ and gets the ciphertext $\sigma = (a_0, a_1, ..., a_{n-1}, V, C, D)$.

Patient storage phase. The patient uploads the ciphertext obtained in the previous phase to IPFS in advance, which then returns a retrieval index $hashipfs$ to the patient. Next, the patient initiates a storage request to the blockchain network, which triggers the storage contract to store $hashipfs$ on the blockchain. The storage contract is as follows (Algorithm 1).

Algorithm 1 Storage Contract

```
Input: $ID_{pts}$, $hashipfs$
Output: $transactionID_{pts}$
$timestamp \leftarrow system.time$
if $ID_{pts} == \emptyset$ then
    identity initialization fail
end if
if $hashipfs == \emptyset$ then
    storage initialization fail
end if
$transactionID_{pts} \leftarrow blockchain.upload(hashipfs)$
return $transactionID_{pts}$
```

Algorithm 2 Access Contract

```
Input: $ID_{pts}$, $ID_i$
Output: $transactionID_{pts}$
$timestamp \leftarrow system.time$
if $ID_i \neq \emptyset$ then
    if $ID_{pts} \neq \emptyset$ then
        $transactionID_{pts} \leftarrow blockchain.search(ID_{pts})$
        return $transactionID_{pts}$
    end if
    illegal access
end if
access initialization fail
```

Users access phase. A user initiates an access request to the blockchain network through the patient's identity ID_{pts}. After receiving the request, the access contract will be triggered, then the retrieval index $hashipfs$ on the blockchain will be sent to the user. Meanwhile, the access request will be recorded on the blockchain. The access contract is as follows (Algorithm 2). Through the retrieval index $hashipfs$, the user downloads encrypted medical data σ from IPFS.

Users unsigncryption phase. When the ciphertext $\sigma = (a_0, a_1, ..., a_{n-1}, V, C, D)$ is retrieved by receivers $\{ID_1, ID_2, ..., ID_n\}$, a user ID_i can decrypt and verify. Details are as follows:

a. The user ID_i reconstructs the n degree polynomial $f(x)$ according to $(a_0, a_1, ..., a_{n-1})$ to get

$$f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$$

b. The user ID_i computes $J_i{}' = (x_i + y_i)X_{pts}$ through his/her private key and the patient's public key. Next, computes $T_i{}' = H_4(J_i{}')$ and $t' = f(T_i{}')$.

c. The user ID_i computes $MD = C \oplus W'$, where $W' = F(H_3(t'), ID_i)$. Therefore, the user can get the patient's medical data MD.

d. The user ID_i checks the validity of the MD and the identity of the patient by verifying the formula

$$RVD = X_{pts} + Y_{pts} + h_{pts} \cdot msk$$

where $h_{pts} = H_1(ID_{pts}, Y_{pts}, X_{pts})$ and $R = H_2(ID_{pts}, MD, D)$. If the formula holds, the MD can be decrypted, and the identity of the patient ID_{pts} can be authenticated.

4.2 Security Analysis

We make a security analysis in this part. In our scheme, a user encrypts medical data through the identity-based signcryption scheme [20]. The attack model relies on certificateless cryptography, which defines two types of adversaries as follows.

a. Type I adversary (Adv_I) may request public keys and replace public keys arbitrarily. However, he/she is not allowed to achieve the master secret key.

b. Type II adversary (Adv_{II}) does have access to the master secret key but may not replace the public keys of entities.

The signcryption scheme satisfies *confidentiality* under adaptive chosen ciphertext attacks and *unforgeability* under chosen message attacks in games between an adversary (Adv_I or Adv_{II}) and a challenger $\mathcal{C}$ described in [27,31]. The security properties of our scheme are analyzed in terms of the security requirements listed in Section 3, details are as follows:

Confidentiality. The confidentiality of the data MD is guaranteed by the keys generated by KGC. If the adversary wants to get the secret random value t from T_i, he has to compute $T_i{}' = H_4((x_i + y_i)X_{pts})$ from $X_{pts} = x_{pts} \cdot P$ and $X_i + Y_i + h_i \cdot mpk = (x_i + y_i) \cdot P$. Due to the intractability of *CDH*, the adversary cannot compute T_i to get the keys.

Integrity. The integrity of the data MD is guaranteed by the signcryption of ID_{pts}. The patient ID_{pts} outputs $\sigma = (a_0, a_1, ..., a_{n-1}, V, C, D)$. Each user ID_i can decrypt C to get the pseudo-identity ID_{pts} and medical data MD. Next, ID_i check whether the equation $RVD = X_{pts} + Y_{pts} + H_1(ID_{pts}, Y_{pts}, X_{pts}) \cdot msk$ holds. If the equation holds, MD is integrated and has not been altered. Due to the intractability of *ECDL*, the σ is unforgeable.

Non-Repudiation. Only the legitimate patient can generate σ to authenticate with a user ID_i. It cannot be forged by anyone under the intractability of *ECDL*. Only predetermined users can decrypt the σ and authenticate. Each user can check the validity of MD. If it is valid, the patient cannot deny that he had provided users with medical data, and the users cannot deny that he had requested data from the patient.

Anonymity. During the whole processes of our scheme, the patient and users use pseudo-identities to register and exchange data. In order to obtain someone's true identity, the adversary needs to solve the *CDH* instance since the adversary has to compute $T_i{}' = H_4((x_i + y_i)X_{pts})$ and $X_i + Y_i + h_i \cdot mpk = (x_i + y_i) \cdot P$. Due to the intractability of *CDH*, our scheme provides the anonymity for patient and users.

5 Performance Evaluation

In this section, we conduct two experiments on patient encryption and users decryption. The experiments are run on a laptop with a 64-bit Windows 10 operating system equipped with an Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz and 16.0 GB of memory. They are implemented through C programming language, leveraging the Pairing-based Cryptography (PBC) library [17] to get results. Specially, only three cryptographic operations are considered including multiplication, exponentiation, and pairing operations since the cost of hash operations and polynomial operations is negligible.We make a comparison with other medical data sharing schemes proposed by [4,15,26].

In our scheme, a patient needs to compute J_i to encrypt data, where $J_i = x_{pts}(X_i + Y_i + h_i \cdot mpk)$, $i = 1, 2, ..., n$, n represents the number of users. Therefore, the patient needs $2n$ point multiplication to encrypt the data MD, and only one point multiplication is needed to compute the signature V, D. Any user recovers the source data with just one point multiplication and verifies the signature with two point multiplications. We compare the computation cost of encryption for patient and decryption data for users in our scheme with other schemes in Table 2. Among them, T_m denotes the execution time of one point multiplication operation, T_e denotes the execution time of one exponentiation operation, and T_p denotes the execution time of one pairing operation. It is obviously that our scheme is implemented without exponentiation and pairing operations, while other schemes operate on pairing, point multiplication or exponentiation.

Table 2: Computational overhead comparisons of patient and users

Schemes	Patient	Users
[4]	$nT_p + 5T_m$	$3T_p + T_m$
[15]	$4(n+1)T_e$	$11T_e + 5T_p$
[26]	$nT_p + 4T_m$	$3T_p + T_m$
this work	$(2n+1)T_m$	$3T_m$

We compute the time consumption of a patient to encrypt medical data, and compare it with the experiments in [4,15,26]. The comparison results are shown in Figure 5. It can be seen from the figure that with the growth of the number of users, the time for patient to encrypt data in our scheme increases at a lower rate, which is lower than that of other schemes. In addition, we compare the time consumption of each user to decrypt medical data in our scheme with [4,15,26]. Note that the number of specialist physicians in practical applications will not exceed ten generally, so we set the number of users to ten. The comparison results are shown in Figure 6. It can be seen from the figure that the time for a user to decrypt data in our scheme is obviously lower than that of other schemes.

6 Conclusions

Targeting to the shortcomings of the current electronic medical records (EMRs) sharing scheme, we propose a privacy-preserving EMRs sharing solution based on blockchain. The storage and access of EMRs are realized through the mode of on-chain storage and off-chain encryption. Among them, an identity-based signcryption is used to implement fine-grained access control of EMRs, so as to achieve privacy-preserving. It can be seen from the experiments that our scheme is more efficient.

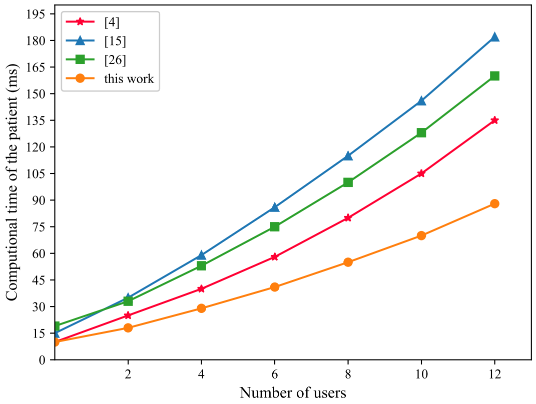


Figure 5: Comparison of encryption time consumption with regard to the number of users

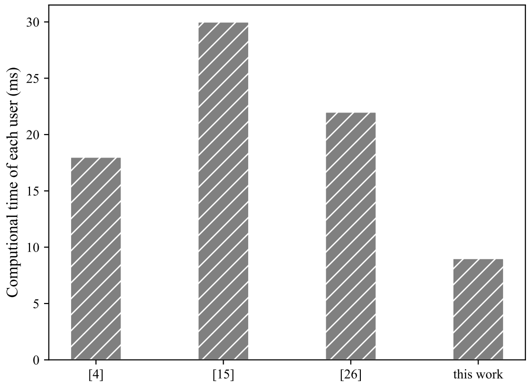


Figure 6: Comparison of decryption time consumption for each user

In addition, our scheme can still make some improvements in terms of identity secondary authentication. In future work, we consider placing the secondary authentication of patients and users on the blockchain to further improve the reliability.

Acknowledgments

This work is supported by the National Natural Science Foundations of China (Grant No.61902303),the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No.21JK0651), the Young Talent fund of University Association for Science and Technology in Shaanxi, China (Program No.20210116), the Shaanxi Key Laboratory of Blockchain and Secure Computing (No.N-KY-XZ-1101-202110-7349).

References

[1] J. Benet, "Ipfs-content addressed, versioned, P2P file system," *arXiv preprint arXiv:1407.3561*, 2014.

[2] V. Buterin, "A next-generation smart contract and decentralized application platform," *white paper*, vol. 3, no. 37, pp. 2–1, 2014.

[3] W. Chai, M. Liu, Z. Zhang, and L. Lv, "Blockchain-based privacy-preserving electronic voting protocol," *International Journal of Network Security*, vol. 24, no. 2, pp. 230–237, 2022.

[4] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "Medblock: Efficient and secure medical data sharing via blockchain," *Journal of medical systems*, vol. 42, no. 8, pp. 1–11, 2018.

[5] H. Guo, W. Li, E. Meamari, C.-C. Shen, and M. Nejad, "Attribute-based multi-signature and encryption for ehr management: A blockchain-based solution," in *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC'20)*, pp. 1–5, 2020.

[6] H. Huang, P. Zhu, F. Xiao, X. Sun, and Q. Huang, "A blockchain-based scheme for privacy-preserving and secure sharing of medical data," *Computers & Security*, vol. 99, p. 102010, 2020.

[7] T. Hulsen, "Sharing is caring data sharing initiatives in healthcare," *International Journal of Environmental Research and Public Health*, vol. 17, no. 9, p. 3046, 2020.

[8] H. Jin, Y. Luo, P. Li, and J. Mathew, "A review of secure and privacy-preserving medical data sharing," *IEEE Access*, vol. 7, pp. 61 656–61 669, 2019.

[9] A. Joux and K. Nguyen, "Separating decision diffie–hellman from computational diffie–hellman in cryptographic groups," *Journal of cryptology*, vol. 16, no. 4, pp. 239–247, 2003.

[10] D. G. Katehakis, "Electronic medical record implementation challenges for the national health system in greece," *International Journal of Reliable and Quality E-Healthcare*, vol. 7, no. 1, pp. 16–30, 2018.

[11] N. Koblitz, A. Menezes, and S. Vanstone, "The state of elliptic curve cryptography," *Designs, codes and cryptography*, vol. 19, no. 2, pp. 173–193, 2000.

[12] R. Kumar, N. Marchang, and R. Tripathi, "Distributed off-chain storage of patient diagnostic reports in healthcare system using ipfs and blockchain," in *International Conference on*

COMmunication Systems & NETworkS (COMSNETS'20), IEEE, pp. 1–5, 2020.

[13] S. G. Langer, "Challenges for data storage in medical imaging research," *Journal of digital imaging*, vol. 24, no. 2, pp. 203–207, 2011.

[14] J. M. Lee, "Identity-based signcryption." *IACR Cryptol. ePrint Arch.*, vol. 2002, p. 98, 2002.

[15] X. Liu, Z. Wang, C. Jin, F. Li, and G. Li, "A blockchain-based medical data sharing and protection scheme," *IEEE Access*, vol. 7, pp. 118 943–118 953, 2019.

[16] Y. Liu and G. Xu, "Fixed degree of decentralization dpos consensus mechanism in blockchain based on adjacency vote and the average fuzziness of vague value," *Computer Networks*, vol. 199, p. 108432, 2021.

[17] B. Lynn, *The Pairing-Based Cryptography (PBC) library*, June 4, 2013. (https://crypto.stanford.edu/pbc/)

[18] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[19] A. Shahnaz, U. Qamar, and A. Khalid, "Using blockchain for electronic health records," *IEEE Access*, vol. 7, pp. 147 782–147 795, 2019.

[20] J. Shen, Z. Gui, X. Chen, J. Zhang, and Y. Xiang, "Lightweight and certificateless multi-receiver secure data transmission protocol for wireless body area networks," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[21] S. Shi, D. He, L. Li, N. Kumar, M. K. Khan, and K.-K. R. Choo, "Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey," *Computers & security*, vol. 97, p. 101966, 2020.

[22] M. Sookhak, M. R. Jabbarpour, N. S. Safa, and F. R. Yu, "Blockchain and smart contract for access control in healthcare: a survey, issues and challenges, and open issues," *Journal of Network and Computer Applications*, vol. 178, p. 102950, 2021.

[23] J. Sun, X. Yao, S. Wang, and Y. Wu, "Blockchain-based secure storage and access scheme for electronic medical records in ipfs," *IEEE Access*, vol. 8, pp. 59 389–59 401, 2020.

[24] C.-J. Wang, X.-L. Xu, D.-Y. Shi, and W.-L. Lin, "An efficient cloud-based personal health records system using attribute-based encryption and anonymous multi-receiver identity-based encryption," in *Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, IEEE, pp. 74–81, 2014.

[25] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

[26] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, "Medshare: Trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, vol. 5, pp. 14 757–14 767, 2017.

[27] G. Xu, J. Dong, C. Ma, J. Liu, and U. G. O. Cliff, "A certificateless signcryption mechanism based on blockchain for edge computing," *IEEE Internet of Things Journal*, 2022.

[28] G. Xu, Y. Liu, and P. W. Khan, "Improvement of the dpos consensus mechanism in blockchain based on vague sets," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4252–4259, 2020.

[29] I. Yaqoob, K. Salah, R. Jayaraman, and Y. Al-Hammadi, "Blockchain for healthcare data management: opportunities, challenges, and future recommendations," *Neural Computing and Applications*, pp. 1–16, 2021.

[30] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends," in *IEEE international congress on big data*, pp. 557–564, 2017.

[31] C. Zhou, W. Zhou, and X. Dong, "Provable certificateless generalized signcryption scheme," *Designs, codes and cryptography*, vol. 71, no. 2, pp. 331–346, 2014.

Biography

Mingqiang Shao is a master student of the School of Computer Science, Xi'an Polytechnic University, China. His current research interests is designing privacy-preserving protocols based on blockchain.

Momeng Liu is an associate professor in the School of Computer Science, Xi'an Polytechnic University, China, and a member of Shanxi key Laboratory of Clothing Intelligence. In 2018, she earned her Ph.D. degree in cryptography from Xidian University, China. Her research interests mainly focus on designing protocols built upon lattice-based cryptography and providing privacy-preserving solutions in blockchain-based scenarios.

Zhenzhen Wang is a master student of the School of Computer Science, Xi'an Polytechnic University, China. His current research interests is designing privacy-preserving protocols based on blockchain.

Common Knowledge Based Secure Generation and Exchange of Symmetric Keys

Hexiong Chen[1], Jiaping Wu[2], Wei Guo[1], Feilu Hang[1], Zhenyu Luo[1], and Yilin Wang[2]
(Corresponding author: Jiaping Wu)

Information Center of Yunnan Power Grid Co. Ltd[1]
Kunming, Yunnan 650000, China
Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China[2]
Quzhou, Zhejiang 324000, China
Email: 13981671425@163.com

(Received Apr. 27, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 13, 2022)

Abstract

Man-in-the-middle attacks bring severe security threats to the Diffie-Hellman (DH) based symmetric key generation and exchange. This paper proposed a common knowledge-based technique for secure key generation and exchange to ensure the security of the key generation between the communication peers. A message matching mechanism with low cost is designed without interfering with or changing the standard communication mechanism. Therefore, the common knowledge about the communication peers' historical communication messages is obtained using the shared knowledge generation and verification algorithm. Based on the common knowledge and the proposed sizeable prime number generation algorithm, DH symmetric key agreement with low information exchange is implemented to improve the security in generating the symmetric key. Theoretical analysis and simulation results show that the proposed schemes can considerably improve the security in the process of the symmetric key generation with low resource cost.

Keywords: Common Knowledge; Consistency Verification; Key Exchange; Message Matching Scheme; Symmetric Key

1 Introduction

In recent years, with the booming development of information technology, the scale of information networks has continued to expand, and the applications of information networks have become increasingly abundant. These applications generally have the characteristics of a large size of transmitted data and strong real-time communication requirements. The use of a symmetric encryption mechanism to encrypt data can better meet application requirements.And the key generation and exchange are crucial technologies of symmetric encryption. In 1976, Diffie and Hellman jointly proposed the Diffie-Hellman key exchange(DHKE) protocol to securely generate symmetric keys for the the communication peers [11], which has been widely used in network security protocols such as SSL (Security Socket Layer) and IPsec (Internet Protocol Security). However, the DHKE protocol has always had the problem of MITM attacks in practice [2], which leads to the disclosure of session content and cannot guarantee the security of the communication process.

To solve the problem of MITM attack of DHKE protocol, many studies have been done in academia and industry. In 2010, Yoon *et al.* [12] proposed a secure DHKE protocol based on Chebyshev polynomials and chaotic mapping. Since then, some researchers have extended the application of this protocol to achieve identity authentication in different scenarios [1,8,17,19]. But these studies increased interactions and communication cost. In 2014, Shen *et al.* [24] proposed a technique for communication key transmission between device-to-device (D2D) based on DHKE protocol. In 2015, Khader *et al.* [16] analyzed the MITM attack problem of the DHKE protocol and proposed a scheme using the Geffe binary number generator, to improve the key generation mechanism and resist the MITM attacks. In 2019, Zhang *et al.* [26] proposed that both parties used shared secret matrix eigenvalues for key agreement. In 2021, Shen *et al.* [23] presented a novel in-band solution for defending the MITM attack during the key establishment process for wireless devices. Though these studies have enhanced the security of the DHKE mechanism, they have led to high communication costs or long key generation time.

The research of protecting key exchange information is as follows. In [7], Bui et.al proposed a key exchange protocol using blockchains and other public ledger structures. In [21], Naher *et al.* proposed a DHKE protocol based on a shared CRC (Cyclic Redundancy Check) polynomial, which can detect MITM attacks in the process of key information exchange, but it cannot prevent MITM attacks. In [25], Thwe *et al.* proposed to protect the key exchange information of the DHKE protocol by hash function, to resist the MITM attacks encountered during the key generation process. In 2020, Chunka [9] improved the mechanism of the DHKE protocol and prevented the key exchange information from being tampered with by using digital signature technology. Ali *et al.* [4] proposed an approach to defend against attacks by generating a hash of each value transmitted over the network. However, these studies have high computational complexity in protecting key exchange information.

Meanwhile, the key generation mechanism based on communication context has emerged. In 2020, Dar *et al.* [10] proposed to calculate the level of confidentiality of each message based on context-aware computing, and select the optimal encryption algorithm according to the confidentiality level. This study has reduced the resource cost and time delay of the encryption and decryption process. However, context-aware computing has greater computational complexity. In [15, 20], the authors proposed to perceive communication data using context-aware computing in real time and combine it with attribute encryption technology to achieve access control encryption for communication data. However, these studies only use the key attributes in the message for contextual computing, such as time, location, and device status. This study is vulnerable to attacks third-party. Consequently, if the computational complexity of the context-awareness computing is great, it will increase the encryption and decryption time. When the message is altered, diverted, or leaked, it will bring threats to the security of future communication by using contextual computing in historical messages to encrypt new messages.

In order to reduce the risk of key leakage and improve the security of communication, this paper proposes a common knowledge-based symmetric key generation and exchange technique (CK-SKGET). Based on an elaborate message matching and consistency verification mechanism, CK-SKGET uses historical communication messages to generate consistent common knowledge among communication peers. Then, CK-SKGET realizes secure key generation and exchange with low costs based on common knowledge. This paper designs a low-cost message matching mechanism without interfering or changing the normal communication mechanism. And we propose an algorithm for generating common knowledge through the historical communication messages of the communication peers. Furthermore, we provide a new idea for key exchange using communication messages. Thus, CK-SKGET has better security in key exchange and can be applied in broader network scenarios. Our contributions are summarized as follows.

1) A common knowledge generation and verification algorithm is designed. The communication peers use mutual communication messages to quickly establish common knowledge based on minimizing the number of additional information interactions and the risk of key leakage.

2) An algorithm for generating large prime numbers is designed which is based on the common knowledge established by the communication peers. Then, the communication peers use the DHKE protocol to realize the generation and exchange of symmetric keys, which improves the security in the process of symmetric key generation and exchange.

3) Through theoretical analysis and simulation comparisons, it was proved that CK-SKGET significantly improves the security of the symmetric key generation process under reasonable resource cost, also the feasibility and effectiveness of the scheme.

The rest of the paper is organized as follows. Section 2 introduces the system architecture, including the definition of common knowledge and system model. Section 3 introduces the generative process of common knowledge and the verification algorithm of common knowledge. Then, we use the common knowledge to compute large prime numbers and obtain symmetric keys with the DHKE protocol in Section 4. The security analysis of the proposed scheme is given in Section 5. Our proposed scheme was compared with others in Section 6, with the conclusion of our proposed scheme outlined in Section 7.

2 System Architecture

This section will define the common knowledge in this paper, and specifically explain the basic ideas and system architecture for the communication peers to build common knowledge and generate symmetric keys under the existing communication mode.

2.1 Definition of Common Knowledge

Common knowledge: To meet the user's application requirements, the communication peers will continuously exchange information and send encapsulated data frames to each other. During the exchange of data frames, we will gradually establish a shared and exclusive shared historical message database for them. The communication peers can calculate and abstract these historical messages to obtain a consistent understanding of the communication process and content-common knowledge. Since the common knowledge has the characteristics of mutuality, consistency, and exclusivity of the communication peers, we can apply it to the security protection process such as key generation and exchange, to make the communication process more secure.

Specifically, according to the role of each field in the message, these fields can be divided into two categories: control fields M^c that guarantee the communication process, and content fields M^d that are exchanged in communication. The format and length of the control field are deterministic and consistent, which is convenient for unified processing. The format and length of the content field are affected by communication requirements, and have variability and differences, and need to be truncated or zero-filled, etc. After the communication peers calculate the data in the control field and the content field, they will establish common knowledge of communication process and content.

Among the historical messages exchanged during the communication process between the communication peers, those historical messages M_i $(i \in \mathbb{N}^+)$ shared by the communication peers after negotiation and confirmation are called effective messages. Each effective message is represented as M_i^e $(i \in \mathbb{N}^+)$, and each filed in M_i^e is represented as $m_{i,j}$ $(i, j \in \mathbb{N}^+)$. These fields have two types: control field and content field, depending on the specific communication protocol used. For example, in the Ethernet frame protocol, a message M_i consists of seven fields: lead code $m_{i,1}$, frame start $m_{i,2}$, destination address $m_{i,3}$, source address $m_{i,4}$, protocol type $m_{i,5}$, data packet $m_{i,6}$, and parity code $m_{i,7}$. $m_{i,1}$, $m_{i,2}$, $m_{i,3}$, $m_{i,4}$, $m_{i,5}$, and $m_{i,2}$ belong to control fields, while $m_{i,7}$ belongs to content fields.

Therefore, the computation of common knowledge is primarily dependent on long-term communication between communication peers (depending on the security level, it can be in minutes, hours, days, months, or even years), continuous communication history, and a small amount of negotiation. It is theoretically possible for an attacker to eavesdrop on the entire process and steal the common knowledge based on the complete information overheard. However, due to the implementation cost and difficulty, it is almost impossible to carry out long-term continuous eavesdropping on each group of communication peers in the entire network. Besides, its eavesdropping is also less covert and easier to detect. Therefore, it can be considered that common knowledge has the characteristics of mutuality, consistency, and exclusivity. We can apply it to the security protection process such as key generation and exchange to make the communication process between the communication peers more secure.

2.2 System Model

The system structure of CK-SKGET is shown in Figure 1. After the communication frequency between communication peers reaches a certain threshold F_0 $(F_0 > 0)$, H_1 can decide to negotiate with H_2 and start building common knowledge of the communication process. The negotiation content is a 7-tuple $< T_0, N_m, N_p, c, d, R_C, R_D >$, where T_0 is the start time of collecting messages, $N_m (N_m \in \mathbb{N}^+)$ is the number of messages collected, c $(c > 3, c \in \mathbb{N}^+)$ is the number of columns of the control matrix, d $(d \in \mathbb{N}^+)$ is the number of columns of the content matrix, R_C $(R_C \in \mathbb{N}^+)$ and R_D $(R_D \in \mathbb{N}^+)$ are random number. If the message matching fails for the first time, it means that the communication message sets collected by H_1 and H_2, $S_1 = \{M_{1,1}, M_{1,2}, \cdots, M_{1,N_m}\}$ and $S_2 = \{M_{2,1}, M_{2,2}, \cdots, M_{2,N_m}\}$, are inconsistent.

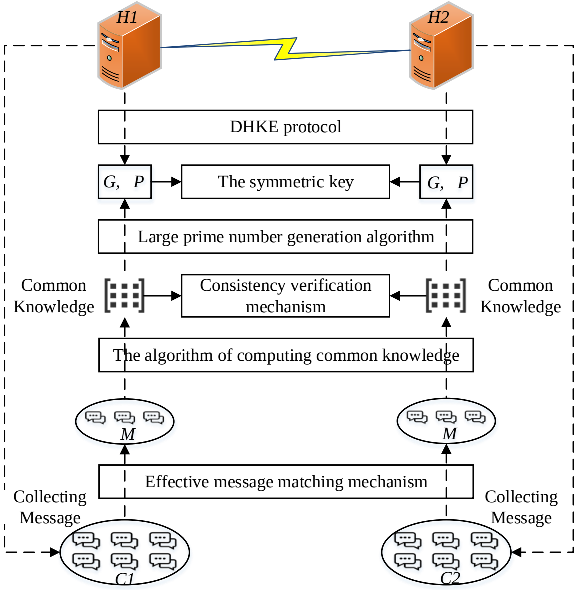

H1
H2
DHKE protocol
G, P
The symmetric key
G, P
Large prime number generation algorithm
Common Knowledge
Consistency verification mechanism
Common Knowledge
The algorithm of computing common knowledge
M
M
Effective message matching mechanism
Collecting Message
Collecting Message
C1
C2

Figure 1: The Structure of CK-SKGET

Then, based on the agreed rules, the communication peers take subsets s_i^1 and s_i^2 $(i \in \mathbb{N}^+)$ of sets S_1 and S_2, respectively. And they match s_i^1 and s_i^2, which is partial message matching. N_p is the maximum number of partial message matching.

The communication frequency F between H_1 and H_2 changes in real time. Suppose the time required to collect messages is T_m, the time for generating common knowledge and symmetric keys through historical messages is T_c, and the value of N_m is decided by $\psi(F, T_c)$. The design principle of the function $\psi(F, T_c)$ is as follows. 1) When F is small, N_m should be reduced to shorten the message collecting time, which aims to ensure that the key can be updated in time. 2) When F is large, N_m should be increased to make $T_m > T_c$ guaranteeing there is sufficient time to update the key. At the same time, the security of the key generated by CK-SKGET is related to the number of effective messages. When F is smaller, the minimum number of N_m is $N_0 = \psi(F, T_c)$ to ensure the security of the generated key. In order to update the key in time and improve the security of the key, H_1 and H_2 should negotiate the 7-tuple in each round of generating common knowledge. After H_1 and H_2 negotiate to generate symmetric key through CK-SKGET, they store the communication messages in their respective caches C_1 and C_2 starting from T_0.

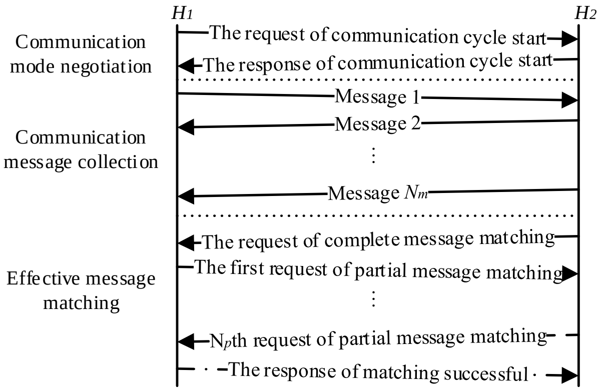
When the message cached by H_1 reaches the predetermined number N_m, it sends an effective message matching request to H_2. They obtain an effective message set for generating common knowledge through an effective message matching mechanism M^v. Then, the communication peers use the common knowledge generation algorithm to obtain the common knowledge matrix K, and verify the consistency of K. Then, they use the eigenvalues of the matrix and the large prime number generation algorithm to get the same large prime number G and primitive root P. Finally, they execute the DHKE protocol to reach agreement of the symmetric key. The symmetric key generated by common knowledge will be used in the subsequent encryption and decryption of communication data to ensure the security of the data in the communication process.

3 Common Knowledge Generation and Verification Algorithm

The process of generating common knowledge must comply with the following basic principles: 1) H_1 and H_2 cannot interfere with the normal communication process and cannot change the mechanisms of the existing communication protocol. For example, operations such as retransmission and reassembly of lost and out-of-sequence data frames must be determined by the communication protocol, and no additional additions or deletions are permitted. 2) The messages in S_1 and S_2 are inconsistent, and it is necessary to make sure that the messages used to compute common knowledge are the same. 3) The key generation is based on common knowledge, and the common knowledge generated by the communication peers should be exactly the same. Based on the above principles, the establishment of common knowledge between H_1 and H_2 requires three steps: effective messages matching, computing common knowledge based on effective messages, and consistency verification of common knowledge.

3.1 Effective Messages Matching Mechanism

The effective messages matching mechanism is shown in Figure 2. It includes three stages: communication mode negotiation, communication messages collection and effective messages matching. In the stage of communication mode negotiation, H_1 and H_2 negotiate to start a new round of communication cycle and related information. H_1 sends the request of communication cycle start to H_2. H_2 responds to the request after receiving it. In the stage of communication messages collection, they carry out the normal communication interaction process. When the number of message buffers is N_1 , it will send effective messages matching requests. H_1 sends a complete message matching request to H_2, and after H_2 receives the request, they perform a complete message matching. If the match is successful, H_2 sends a response message to H_1, or they start partial message matching to obtain effective messages that can generate common knowledge. If they still fail to match after N_p times of partial message matching, they abandon the messages in this round of communication cycle and start a new round of communication cycle. After communication peers are successfully matched with effective messages, they obtain the set $M^e = \{M^e_1, M^e_2, \cdots, M^e_N\}$ of N effective messages, and send a matching success response to the other.



H_1
H_2
Communication mode negotiation
The request of communication cycle start
The response of communication cycle start
Communication message collection
Message 1
Message 2
⋮
Message N_m
Effective message matching
The request of complete message matching
The first request of partial message matching
⋮
N_pth request of partial message matching
The response of matching successful

Figure 2: Effective message matching mechanism

When matching the complete message with the partial message, H_1 and H_2 will hash the message arranged in chronological order. When the hash results of the communication peers are consistent, the effective message is matched successfully. When performing partial message matching, H_1 and H_2 hash the set $s^1_i = \{M_{1,1}, M_{1,2}, \cdots, M_{1,k}\}$ and $s^2_i = \{M_{2,1}, M_{2,2}, \cdots, M_{2,k}\}$ ($k \in [1, N_m]$), respectively. The message $M_{1,j}$ and $M_{2,j}$ ($j \in [1, k]$) in the set s^1_i and s^2_i, respectively, need to sample in S_1 and S_2 by function $S(x)$. The design principle of the function $S(x)$ is to ensure that in the expected N_p hashing result comparisons, there is as little overlap between the sampled messages as possible and the probability of a single message appearing in the two hashing comparison is as small as possible. In this paper, $S(x) = a \times i + 1$ is Step function, where $a = 2, 3, 4$ *et al.* The value of a is related to N_m and N_p. In actual applications, $S(x)$ is adjusted or redesigned according to the requirements of the communication scenario.

3.2 Computing Common Knowledge

To compute common knowledge, H_1 and H_2 firstly sample data from M^c_i and M^d_i in M^e_i to get the control matrix U and the content matrix V, respectively. Then the common knowledge matrix can be computed based on U and V.

In the construction matrix U and V, H_1 and H_2 obtain a row vector $\vec{u} = [u_1, u_2, \cdots, u_c]$ of U and $\vec{v} = [v_1, v_2, \cdots, v_d]$ of V by sampling M^c_i and M^d_i, respectively. When they sample the data in M^c_i, the bytes $b_{i,j}$ of M^c_i containing the fields are first sequentially concatenated into a field having l^c_i bytes. Then, they sample data from c, R_C, and l^c_i. The rules of sampling data are as follows.

1) If $l^c_i < c$,

$$u_y = \begin{cases} b^c_{i,j}, & j \leq l^c_i; \\ 0, & j > l^c_i. \end{cases} \tag{1}$$

2) If $l_i^c \geq c$, H_1 and H_2 divide the content of $b_{i,j}$ to c groups, and sample data from each group. The data size of each group is $\xi_i = \lfloor \frac{l_i^c}{c} \rfloor$, where "$\lfloor \bullet \rfloor$" is round-down operation. The initial value of R_i^s is R_C, and H_1 and H_2 adjust R_i^s according to ξ_i.

$$R_i^s = \begin{cases} 1 & , \xi_i = 1; \\ \lfloor \frac{R_i^s}{\xi_i} \rfloor & , R_i^s \geq \xi_i. \end{cases} \quad (2)$$

So that H_1 and H_2 obtain $u_y = b_{i,j}^c$ $(1 \leq y \leq c)$, j mod $\xi_i = R_i^s$ and $j \leq c \times \xi_i$. When H_1 and H_2 sample data in M_i^d, they replace d, R_D, l_i^d with c, R_C, l_i^c, respectively. Then, they obtain each data $v_y = b_{i,j}^d$ $(1 \leq y \leq d)$ of $\vec{v}$.

For the above sampling rules, the bytes in the message field are all regarded as sampled byte data when the number of bytes in the message field is less than the number of samples, and the rest is filled with zeros. When the number of bytes in the message field can support random sampling, the value of the sampling random number will be readjusted according to the number of groups to enable uniform data sampling from the message field. So that H_1 and H_2 construct a matrix U and a matrix V.

$$U = \begin{bmatrix} u_{1,1} & \cdots & u_{1,c} \\ \vdots & \ddots & \vdots \\ u_{N,1} & \cdots & u_{N,c} \end{bmatrix}, V = \begin{bmatrix} v_{1,1} & \cdots & v_{1,d} \\ \vdots & \ddots & \vdots \\ v_{N,1} & \cdots & v_{N,d} \end{bmatrix}$$

Then, on the basis of U and V, matrix W is

$$W = U^T \times V = \begin{bmatrix} w_{1,1} & \cdots & w_{1,d} \\ \vdots & \ddots & \vdots \\ w_{c,1} & \cdots & w_{c,d} \end{bmatrix} \quad (3)$$

The meaning of the matrix W is to project the content information into the space corresponding to the control information, so that H_1 and H_2 can understand the content information field in the same way as the control information field. But W is not a square matrix, so we can't find the eigenvalues directly. Multiply W with its transposed matrix W^T to get the common knowledge matrix of square matrix.

$$K = W \times W^T = \begin{bmatrix} k_{1,1} & \cdots & k_{1,c} \\ \vdots & \ddots & \vdots \\ k_{c,1} & \cdots & k_{c,c} \end{bmatrix} \quad (4)$$

After solving the eigenvalues and eigenvectors of the matrix K, a set $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_c\}$ of eigenvalues and a matrix $\gamma = \{\vec{\gamma_1}, \vec{\gamma_2}, \cdots, \vec{\gamma_c}\}$ of corresponding eigenvectors can be obtained. The computing process of H_1 and H_2's common knowledge is shown in Algorithm 1.

Algorithm 1 Common Knowledge Generation

```
Input: $T_0, N_m, N_p, c, d, R_C, R_D$
Output: $K, \lambda, \gamma$
Effective messages $M^e \leftarrow$ Effective message matching mechanism$(T_0, N_m, N_p)$.
for Message $M_i^e$ in $M_e$ do
    $\vec{u}_i$, $\vec{v}_i \leftarrow$ Sampling $M_i^e$ by $c, d, R_C, R_D$.
end for
$W = U^T \times V$.
$K = W \times W^T$.
$\lambda, \gamma \leftarrow$ Decompose the eigenvalues of the matrix $K$.
End
```

3.3 Consistency Verification of Common Knowledge

The matrix K is a real symmetric matrix according to (4), and the eigenvectors corresponding to different eigenvalues are orthogonal to each other. This paper uses the eigenvectors of K to verify the consistency of the matrix calculated by H_1 and H_2. However, the eigenvector does not participate in the calculation process of large prime number, and the eigenvalue cannot be deduced from the eigenvector. So that transferring the characteristic vector in the communication network will not affect the security of the key agreement process.

The consistency verification process of common knowledge includes eigenvector multiplication verification and eigenvector hash verification. Before H_1 sends the consistency verification information of common knowledge, it selects n eigenvectors from γ to make up verification matrix $\eta_1 = [\vec{\gamma_1'}, \vec{\gamma_2'}, \cdots, \vec{\gamma_n'}]^T$, and sorts the remaining $c - n$ eigenvectors according to the size of the corresponding eigenvalues to form a matrix γ_1. η_1 and $hash_{\gamma_1}$ constitute the consistency verification information of common knowledge. After receiving the consistency verification information, H_2 uses γ to multiply η_1 to verify it. And the matrix ρ can be obtained by multiplying with η_1 and γ.

$$\begin{aligned} \rho &= \gamma \times \eta_1 \quad (5) \\ &= \begin{bmatrix} \gamma_{1,1} & \cdots & \gamma_{1,c} \\ \vdots & \ddots & \vdots \\ \gamma_{c,1} & \cdots & \gamma_{c,c} \end{bmatrix} \times \begin{bmatrix} \gamma'_{1,1} & \cdots & \gamma'_{n,1} \\ \vdots & \ddots & \vdots \\ \gamma'_{1,c} & \cdots & \gamma'_{n,c} \end{bmatrix} \end{aligned}$$

Since the eigenvectors corresponding to different eigenvalues are orthogonal to each other, there are only n non-zero values in ρ and each column has only one non-zero value. When $\rho_{i,j} \neq 0$ $(i, j \in [1, c], \mathbb{N}^+)$, $\rho_{i,j} = \vec{\gamma_i} \times \vec{\gamma_j'} = 1$, and $\vec{\gamma_i} = \vec{\gamma_j'}^T$, H_2 verifies η_1 by non-zero value in $\rho_{i,j}$ and obtain the n feature vectors making up η_1. Then H_2 gets a matrix γ_1' composed of $c - n$ eigenvectors, and hash γ_1' to obtain $hash_{\gamma_1'}$, after which it compares $hash_{\gamma_1}$ and $hash_{\gamma_1'}$ to verify whether the remaining $c - n$ feature vectors are the same. If the above conditions are met, H_2 can determine that the common knowledge generated with H_1 is the same, and at the same time prevent a third party from tampering with the verification information of the common knowledge. Next, H_2 samples n' $(n' \in [1, c - n])$ eigenvectors from γ_1 to get verification matrix η_2. And H_2 obtains matrix γ_2 based on $c - n'$ eigenvectors in γ to obtain the consistency verification information of common knowledge of H_2 to H_1. H_1 verifies this verification information as H_2. The consistency verification process of the common knowledge using the verification message (including γ, η and $hash_{\gamma_x}(x \in 1, 2)$) is shown in Algorithm 2.

Algorithm 2 Consistency Verification of Common Knowledge

```
Input: $\gamma, \eta, hash_{\gamma_x}$
Output: $True$ or $False$
$\rho = \gamma \times \eta$.
for $i = 1 \rightarrow c, j = 1 \rightarrow n$ do
    if $\rho_(i, j) \neq 0$ and $\vec{\gamma_i} \neq \vec{\gamma_j}^{'T}$ then
        return $False$.
    end if
end for
$\gamma^{'} = \gamma - \eta$.
if $Hash(\gamma^{'}) \neq hash_{\gamma_x}$ then
    return $False$.
end if
return $True$.
End
```

4 Key Agreement

This section mainly introduces the algorithm of using common knowledge to compute large prime numbers and the combination of DHKE protocol to obtain symmetric keys. The large prime number generation algorithm includes two parts: factor base update and large prime number computation.

4.1 Factor Base Update

The Miller-Rabin prime number detection algorithm is a widely used plasticity detection algorithm, which is based on the Fermat theorem. The Fermat theorem is as follows: there are prime number n and integer a, which satisfy $\gcd(a, n) = 1$, so $a^{n-1} \equiv 1(\mod n)$ [22]. The Miller-Rabin prime number detection algorithm is derived from Fermat theorem. Since n is a prime number, $n = 2^s \times r + 1$, where $r = 2 \times k + 1$ ($k \in \mathbb{N}^+$). For integer a, $\gcd(a, n) = 1$, so $a^r \equiv 1(\mod n)$ or $a^{2 \times j \times r} \equiv -1(\mod n)$ with $0 \leq j \leq s - 1$. The Miller-Rabin prime number detection algorithm is a probability detection algorithm, which means some strong pseudo prime numbers may be detected incorrectly [18]. After k times of testing, the probability of being wrongly judged as a composite number is $(\frac{1}{4})^k$. The speed of this prime number detection algorithm is much higher than that of other detection algorithms (such as Solovay-Strassen detection algorithm) [14].

This paper combines the rapid generation algorithm of large prime numbers from small prime numbers in [27] and the incremental prime number generation algorithm in [18] to generate odd numbers. And we use the Miller-Rabin prime number detection algorithm for primality detection to realize symmetric key generation based on common knowledge.

Not all elements in the set of eigenvalues λ are prime number. When constructing the small prime number set based on the set λ, we replace the odd number closest to each λ_i ($i \in [1, c]$, $\mathbb{N}^+$). Then we use the Miller-Rabin algorithm to detect the primality of λ_i. If λ_i is not the prime number, do $\lambda_i = \lambda_i + 2$ until λ_i is a prime number. Finally, after arranging λ_i in the set of small prime numbers by value, the factor base is obtained $B = \{b_1, b_2, \cdots, b_c\}(b_i = \lambda_i)$.

Suppose the factor base of the i-th round update is B_i, B_0 is the initial factor base. The prime number of B_i is $b_{i,j}$ ($j \in [1, c]$, $\mathbb{N}^+$). The formula of updating $b_{i,j}$ is as follows.

$$b_{i+1,j} = \prod_{k=1, k \neq c+1-j}^{c} b_{i,j}^{\alpha} + 2 \times \delta. \tag{6}$$

α is the power factor, through which we can adjust the speed of updating B, and $\alpha = 1$ in this paper. δ is the distance between $b_{i+1,j}$ and the odd number $\prod_{k=1, k \neq c+1-j}^{c} b_{i,j}^{\alpha}$ obtained by multiplying multiple prime factors, the initial value of which is 0. If the Miller-Rabin algorithm detects that $b_{i+1,j}$ computed from Formula (6) is not prime, do $\delta = \delta + 1$ to obtain new $b_{i+1,j}$ until $b_{i+1,j}$ is a prime number. This process is called incremental prime number generation algorithm. After computing the prime number $b_{i+1,c}$, we can obtain B_{i+1}.

Before updating B_i, it will first compare the length $f(b_{i+1,j})$ of the largest prime $b_{i+1,j}$ in B_{i+1} with the length $\tilde{P}$ of the prime P. So that we can judge whether the factor base B_i is the last round of factor base. The function $f(x)$ for estimating the number of prime numbers $b_{i,j}$ is as follows.

$$f(x) = \lceil \log_2 b_{i,j} \rceil \tag{7}$$

"$\lceil \bullet \rceil$" is round-up operation, such as $f(50) = \lceil \bullet \rceil = 6$. Suppose the product of two integers a_0 and a_1 is a_2, so $\log_2 a_2 = \log_2 a_1 + \log_2 a_0$, and the length of a_2 satisfies

$$f(a_0) + f(a_1) - 1 \leq f(a_2) \leq f(a_0) + f(a_1). \tag{8}$$

Put the above formula into Formula (6),

$$\sum_{k=1, k \neq c+1-j}^{c} f(b_{i,k}) - c + 1 \leq f(b_{i+1,j}) \leq \sum_{k=1, k \neq c+1-j}^{c} f(b_{i,k}). \tag{9}$$

The number of prime numbers calculated in the next round is less than $\tilde{P}$. When $\tilde{P} > f(b_{i+1,c})$, shown B_i is not the last round factor base B_l. So we need compute B_{i+1} and retain $b_{i,c}$ to add it to the set ϕ. If not, $\tilde{P} < f(b_{i,c})$, shown $B_l = B_i$, we stop updating B_{i+1}.

The number of length of $b_{i+1,j}$ depends on the prime factors involved in the process of calculating it according Formula (9), and the length difference between $b_{i+1,j}$ and $b_{i,j}$ is huge. The differences in the lengths of prime numbers in B_0 will affect the prime length differences after the factor base is updated. When the difference in lengths between $b_{0,1}$ and $b_{0,c}$ are little, so the difference in lengths between $b_{1,1}$ and $b_{1,c}$ are also little after the factor base is updated. But the length difference between $b_{1,j}$ and $b_{0,j}$ is huge. Thus $f(b_{i,j}) \approx f(b_{i,j+1})$ and $f(b_{i+1,j}) \approx (c-1) \times f(b_{i,j})$. So that the update end condition of updating B_i is to determine whether $f(b_{i+1,c})$ is greater than $\tilde{P}$. Meanwhile, $b_{l,j}$ in B_l is not fully in involved the compute process of P, so the update speed of the factor base can be improved by reducing the number of calculations of $b_{l,j}$. Because of $\theta_l = \lfloor \frac{\tilde{P}}{f(b_{l,c})} \rfloor$ and $\theta \in \mathbb{N}^+$, θ_l is the number of participating in the compute process of $\tilde{P}$, we can obtain the primes numbers $\{b_{l,c-\theta_l-1}, \cdots, b_{l,c}\}$ in B_l.

Due to the uneven distribution of prime numbers, in the process of computing the prime number $b_{i+1,j}$, the number of prime number determinations with Miller-Rabin algorithm is random [18]. But the time of computing $b_{i+1,j}$ is increasing with the length of $b_{i+1,j}$. Therefore, it takes a certain amount of time to update B_i, and the subsequent prime numbers cannot be calculated directly from the initial factor base B_0, preventing a third party from quickly cracking the prime numbers in the key agreement process.

4.2 Large Prime Number Computation

After updated the factor base, the set $\{b_{l,c-\theta_l}, \cdots, b_{l,c}\}$ in B_l and the set $\phi = \{b_{0,c}, b_{1,c}, \cdots, b_{l-1,c}\}$ make up the direct factor base B_ϕ to compute P. Because of

$$L_1 = \sum_{j=\theta_l}^{c} f(b_{l,j}) \leq \tilde{P} < \sum_{j=\theta_l-1}^{c} f(b_{l,j}) = L_2, \tag{10}$$

when b_{l,θ_l-1} is involved in the compute of prime numbers, the length of computed odd number will be greater than P. We only need to compute θ_l prime numbers.

The number of $b_{i,c}$ in B_ϕ is θ_i ($\theta_i \in \mathbb{N}$).Firstly, due to $\Delta L = \tilde{P} - L_1$, $\Delta L - \theta_i \times f(b_{i,c}) \leq f(b_{i,c})$, the number of $b_{l-1,c}$ is θ_i. Then, $\Delta L = \Delta L - \theta_i \times f(b_{i,c})$, we repeat this process to get the set

$$\begin{aligned}\Phi = &\{\{b_{l,\theta_l}, \cdots, b_{l,c}\}^{\theta_l}, \{b_{0,c}, \cdots, b_{0,c}\}^{\theta_0}, \cdots, \{b_{l-1,c}, \cdots, \\ &b_{l-1,c}\}^{\theta_{l-1}}\}.\end{aligned} \tag{11}$$

Through

$$P = 2^\beta \times \prod_{i=0}^{l-1} b_{i,c}^{\theta_i} \times \prod_{j=\theta_l}^{c} b_{l,j} + 2 \times \delta, \tag{12}$$

we can compute P. However, since the length of $P_d = \prod_{i=0}^{l-1} b_{i,c}^{\theta_i} \times \prod_{j=\theta_l}^{c} b_{l,j}$ and $\tilde{P}$ are not necessarily equal, we multiply it with 2^β ($\beta \in \mathbb{N}^+$) to get an odd number whose lengths is $\tilde{P}$. Through $\beta = \tilde{P} - f(P_d)$, we obtain the prime number P by the incremental prime number generation algorithm.

4.3 Key Information Exchange

The basic idea of the DHKE protocol is to take advantage of the difficulty in computing the discrete logarithm. The communication parties send only part of the data generated by the key and retain the other part of the data, respectively, to achieve key security agreement [11]. After H_1 and H_2 generate the large prime numbers P and G using their common knowledge, they negotiate the key using the DH key exchange protocol as follows:

1) H_1 generates a random number R_1 ($R_1 \in \mathbb{N}^+$), then according to the following exchange information it can be calculated the key exchange information $X_1 = G^{R_1} \mod P$ which can be transmitted in the public channel, and transmit X_1 to the H_2.

2) H_2 generates a random number R_2 ($R_2 \in \mathbb{N}^+$), after receiving the key exchange message from H_1, and transmits X_2 and to H_1, then the symmetric key is calculated, where $Key = X_1^{X_2} \mod P$.

3) H_2 calculates the symmetric key, where $Key = X_2^{X_1} \mod P$.

From $Key_1 = Key_2 = G^{R_1 \times R_2} \mod P$, H_1 and H_2 can get the same key. Decisional Diff-Hellman (DDH) hypothesis states that it is difficult to distinguish the tuples (g, g^x, g^y, g^{xy}) and (g, g^x, g^y, g^z), among which g is a generator and $\{x, y, z\}$ is a set of random numbers; when G and P are very large, and R_1 and R_2 cannot be obtained at the same time, it is difficult to calculate the shared key, though the DHKE protocol cannot resist MITM attacks. In the process of symmetric key negotiation, the common knowledge matrix K cannot only generate shared large primes against the MITM attacks, but also provide identity authentication for key information exchange which guarantees the security of the whole key generation process.

5 Security Analysis

This section analyzes the security of the proposed scheme in theory and compares the security performance of the proposed scheme with the existing mechanism in combination with typical attack modes.

5.1 Security Model

Based on the Random Oracle model proposed in [6], a security analysis model for CK-SKGET scenario is designed. Suppose the legal participants are H_1 and H_2, and an probabilistic polynomial time (PPT) enemy $\mathcal{A}$. Through inquiring the session instance between H_1 and H_2 ($\prod_{H_1}^n, \prod_{H_2}^n$), the enemy attempts to obtain the key. The specific model is as follows:

1) $Execute(\prod_{H_1}^n, \prod_{H_2}^n)$ query: simulating the passive attack launched by the adversary, $\mathcal{A}$ make communication between H_1 and H_2 through the query, and obtain all the contents in the communication process.

2) $Hash(M_i)$ query: H_1 or H_2 inquires the hash value, $Hash(M_i)$, which are corresponding to the message, M_i ($i \in \mathbb{N}^+$) from $\mathcal{A}$. If the hash value exists in the list, L_H, it is returned to A; otherwise, a random number $R_i^H \in \mathbb{Z}_q^+$ is as the hash value of M_i returning to $\mathcal{A}$, and $M_i, R_i^H \in \mathbb{Z}_q^+$ are stored in L_H.

3) $Send(\prod_{H_1}^n / \prod_{H_2}^n, M_i)$ query: simulating an active enemy attack, $\mathcal{A}$ sends a random message M_i to H_1 or H_2, after receiving the message, according to the numerical procedure H_1 or H_2 will calculate and return the result to $\mathcal{A}$.

4) $Reveal(\prod_{H_1}^n / \prod_{H_2}^n)$ query: simulating an attack from the enemy knowing the session key, after H_1 or H_2 receiving the query, the key negotiated by the scheme is returned to $\mathcal{A}$.

5) $Knowledge(CK)$ query: simulating an attack from the enemy knowing the common knowledge, after H_1 or H_2 receiving the query, the large prime calculated by the scheme is returned to $\mathcal{A}$.

6) $Test(\prod_{H_1}^n / \prod_{H_2}^n)$ query: $\mathcal{A}$ selects H_1 or H_2 for the session test, if $\mathcal{A}$ calculates the key, then return $\perp$ and stop the query; otherwise, through a coin toss to determine the value returned to $\mathcal{A}$, if it is head, the real key is returned to $\mathcal{A}$; otherwise, return a random number as long as the real key. But $\mathcal{A}$ is only allowed for one coin toss.

Definition 1 (Security define of CK-SKGET). *For any adversary $\mathcal{A}$, event Succ means that the key can be obtained by $\mathcal{A}$ through the above query processes, i.e., the attack is successful; in CK-SKGET, the winning edge of $\mathcal{A}$ is $Adv_{\mathcal{A}} = |Pr(S) - 1/2|$; if the value of $Adv_{\mathcal{A}}$ is negligible, the key generated from common knowledge is secure.*

5.2 Security Proof

Theorem 1 (Security of CK-SKGET). *If H: $\{0,1\}^+ \to \{0,1\}^l$ is a random predictor, when the DHH hypothesis is true, the margin of $\mathcal{A}$ attacking DHH hypothesis is negligible, and the biggest margin of A to attack CK-SKGET is*

$$Adv_{CK-SKGET} \leq \frac{(q_e+q_s)^2}{2^{l_c}} + \frac{q_h{}^2}{2^{l_c}} + \frac{q_s{}^2}{2^{l_g+l_p}} + q_h Adv_{DDH}, \tag{13}$$

where the maximum number of times that $\mathcal{A}$ can initiate Hash query, Execute and Send query are q_h, q_e and q_s respectively, and the bit lengths of consistency verification information, large prime G and P outputted by hash function are l_c, l_g and l_p respectively.

Proof. the security of CK-SKGET is verified by a series of games(G) among participants and enemy. G includes five games from $Game_0$ to $Game_4$, define $Succ_i$ represents $\mathcal{A}$ wins $Game_i$, and the ability of $\mathcal{A}$ increases gradually according to the query process in 5.1.

$Game_0$: The game is a real scene, and known from **Definition 1**.

$$Adv_{CK-SKGET} = |\Pr(S) - 1/2| \tag{14}$$

$Game_1$: based on $Game_0$, $\mathcal{A}$ can steal the communication between H_1 and H_2. H_1 and H_2 run Plan $\mathcal{P}$, and maintain the list $L_H = \{M, h\}$, which is used to store the query and answer to $\mathcal{A}$. It is known that when $\mathcal{A}$ can get the content of communication history, its margin is still equal to the one of attacking the DHH security hypothes which is negligible, i.e., $Game_0$ and $Game_1$ are indistinguishable in the predictor,

$$\Pr(Succ_1) - \Pr(Succ_0) = 0 \tag{15}$$

$Game_2$: based on $Game_1$, $\mathcal{A}$ can send message to H_1 or H_2 actively to form the content of communication history. But $\mathcal{A}$ does not know the generation mechanism of common knowledge, then consistency verification information can only be formed through hash collision, which terminates when hash collision occurs. From the birthday paradox, the probability of collision in the predictor is $\max(q_h/2^{l_c})$; if $\mathcal{A}$ cannot obtain the content of communication history, the collision probability is $\max[(q_e+q_s)^2/2^{l_c}]$, then

$$\Pr(Succ_2) - \Pr(Succ_1) \leq \frac{(q_e+q_s)^2}{2^{l_c}} + \frac{q_h{}^2}{2^{l_c}} \tag{16}$$

$Game_3$: based on $Game_2$, $\mathcal{A}$ conjectures the large prime number G and P through the common knowledge, then the margin of A winning the game is

$$\Pr(Succ_3) - \Pr(Succ_2) \leq \frac{q_s{}^2}{2^{l_c+l_p}} \tag{17}$$

$Game_4$: during the game, $\mathcal{A}$ can only guess the key by the predictor attempting to solve the problem of DHH, where the margin is

$$\Pr(Succ_4) - \Pr(Succ_3) = q_h Adv_{DHH} \tag{18}$$

From the five games above, after casting all attacks, the margin of $\mathcal{A}$ under the security model described in Section 5.1 is

$$Adv_{CK-SKGET} \leq \frac{(q_e+q_s)^2}{2^{l_c}} + \frac{q_h{}^2}{2^{l_c}} + \frac{q_s{}^2}{2^{l_g+l_p}} + q_h Adv_{DDH} \tag{19}$$

□

5.3 Typical Attack Analysis

5.3.1 Typical Attack Modes

On the basis of security proof, the security attributes of CK-SKGET, such as bidirectional authentication, anti-replay attack, anti-password guessing attack and forward security, are further analyzed.

1) Bidirectional authentication. Based on the common knowledge established, only when the generation method of common knowledge matrix K is mastered, can H_1 and H_2 pass feature vector multiplication verification and feature vector hash verification which is used to verify the identities of the two in the consistency verification phase of shared knowledge.

2) Anti-replay attack. If $\mathcal{A}$ sent message to H_1 and H_2 and cannot pass the consistency verification of shared knowledge since the lack of calculation method of matrix K, $H_1$and H_2 are able to detect the presence of $\mathcal{A}$ and stop key negotiation.

3) Anti-password guessing attack. If enemy $\mathcal{A}$ send randomly guessed key exchange messages to H_1 and H_2, as the digits and the number of large prime numbers increase, the probability of $\mathcal{A}$ using the guessed large prime number for key negotiation is extremely low.

4) Forward security. As the communication interaction between H_1 and H_2 continues, the key can be continuously generated for communication encryption and decryption. Therefore, even if $\mathcal{A}$ can obtain the key of the current communication process through other methods, it is still difficult to obtain the correct large prime number and the negotiated key without CK-SKGET.

5.3.2 Security Comparison

This section compares several schemes related to the proposed scheme from five security attributes, such as bidirectional authentication and anti-replay attack, and it shows that CK-SKGET generates symmetric keys with better security. The results are shown in Table 1.

In [3], two-level private keys are designed to improve the DH key exchange protocol and hash the final negotiated key at the same time, but the identity of H_1 and H_2 cannot be bidirectional authentication. In [16], the server sends large prime numbers to H_1 and H_2, uses asymmetric encryption to ensure the security of data transmission, bidirectional identity authentication of H_1 and H_2 and anti-MITM attack, and Geffe binary generator can resist password guessing attack. However, H_1 and H_2 cannot resist the replay attack by the enemy by sharing large prime information, nor can they guarantee the forward security of the communication process. In [26], the authors can carry out bidirectional authentication on H_1 and H_2 by using the shared secret matrix to resist password guessing attacks and MITM attacks, but cannot resist replay attacks by using previously used eigenvalues. CK-SKGET implements bidirectional identity authentication for H_1 and H_2 and resolves the MITM attacks, anti-key exchange information replay attack, anti-password guessing attack, and forward security in the DH key exchange protocol.

6 Simulation Results and Performance Analysis

This section conducts simulation verification for CK-SKGET to evaluate its performance. The simulated physical device is H_1 and H_2 connected through an Ethernet, and the data transmission rate is set to 100 Mb/s. The communication behavior per second between H_1 and H_2 obeys the Poisson distribution. The application layer data length in the Ethernet frame sent obeys the uniform distribution over the interval $[46, 1500]$. Each byte sent in communication data obeys the uniform distribution over the interval $[1, 255]$. The language used in the simulation in this paper is Python. Firstly, analyze the time to crack CK-SKGET, the traditional DHKE protocol and the RSA key exchange protocol. And to effectively evaluate the CK-SKGET, the storage resource cost and the calculation time cost caused by the key generation process of the three schemes are compared through the average analysis of 100 experiments on the PyCharm platform.

6.1 Creaking Time Analysis

In CK-SKGET, the guarantee of key security is provided by the DH key agreement mechanism and the process of generating shared large prime numbers. The security of the DHKE protocol is a discrete logarithm problem, which is determined by the number of large prime numbers in the key exchange information [11]. The security of the shared large prime number generation process is guaranteed by common knowledge. If an adversary tries to obtain the communication key of H_1 and H_2 by creaking the shared large prime number, it is necessary to obtain the shared knowledge matrix from the effective message when the adversary want to pass the consensus verification process of the common knowledge and the message verification code process during key information exchange. Then the key agreement process between H_1 and H_2 can be successfully attacked. Therefore, the adversary mainly cracks the common knowledge from the effective message and then cracks the CK-SKGET.

The security of the traditional DHKE protocol and the RSA key exchange protocol is related to the key transmission mechanism and the key. The key transmission mechanism of the two protocol guarantee the security based on the mathematical problem principle. For the security of the key, the DHKE protocol is determined by the shared large prime number. However, the generation mechanism of shared large prime numbers is the same as the key generation mechanism of the RSA key exchange protocol. The random number generator gives a random number that meets the requirements of the number of key lengths, and then obtains the prime number through the prime number generation algorithm and the prime number judgment algorithm. Therefore, the security of the key in the traditional key exchange protocol is mainly determined by the number of prime numbers [5].

Suppose that the adversary can obtain the effective

Table 1: Comparison of security attributes

Lterature	**[3]**	**[16]**	**[26]**	**CK-SKGET**
Two-way authentication	N	Y	Y	Y
Anti-replay attack	Y	N	N	Y
Anti-password guessing attack	Y	Y	Y	Y
Forward safety	Y	N	Y	Y
Anti-MITM attacks	Y	Y	Y	Y

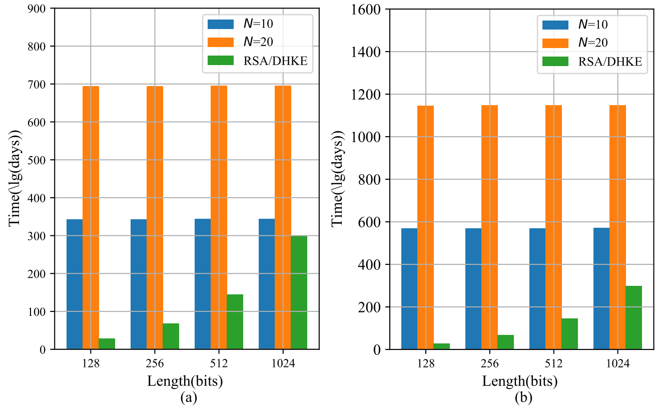


Figure 3: Comparison of time cost for creaking large prime number. (**a**) The time of cracking large prime number when $d = 15$. (**b**) The time of cracking large prime number when $d = 30$.

message between the sink node and the cloud, and it is clear that the process of obtaining matrix K from matrix U and V and computing shared large prime numbers. The time cost required for the adversary to crack the scheme is analyzed. When the adversary cracks the CK-SKGET, it constantly tries to obtain the vector $\vec{u}$ of length c and the vector $\vec{v}$ of length d from each valid message, and then generate matrix U and V. The number of counts the adversary crack CK-SKGET is mainly related to the number of effective messages N and the columns d of the content matrix. The lengths of prime number P to be found is the abscissa and the logarithm of the cracking time is the ordinate. When $c = 5$, we respectively get CK-SKGET in the case of $d = 15$ and $d = 30$, the number of effective messages and the key cracking time cost of the traditional key exchange method, as shown in Figure 3.

When cracking CK-SKGET, the adversary is mainly trying to obtain different matrix U and V, and the number of lengths of large prime numbers only affects its final calculation of large prime numbers. For traditional methods, the number of prime numbers directly affect the time taking for the adversary to crack the key. Therefore, in Figure 3, when N is unchanged, the cracking time corresponding to different $\tilde{P}$ is basically unchanged. But different N and d will seriously affect cracking time. The number of primes is larger, the time of cracking key is longer for traditional methods. Therefore, when H_1 and H_2 use CK-SKGET for key agreement, increasing the number of valid messages and the number of content matrix columns can improve the security.

6.2 Storage Resource Cost Analysis

In the process of generating common knowledge, the matrix U and V will take up a lot of space, it's about $m \times (c + d)$ bytes. In the process of prime number generation, the factor base updated in the last-round will occupy a large space and is related to the number of digits of each prime factor. Then the updated average number of prime factors can be known by Formula (9), which is

$$(L_0 - z) \times (c - 1)^z \leq f(\bar{b}_i) \leq L_0 \times (c - 1)^z. \qquad (20)$$

L_0 is the average length of the initial factor base, z is the number of update rounds of the factor base. Analyzing the calculation process of matrix K, we can get $L_0 = f(\bar{\lambda}_i) \leq \lceil \log_2 r \times N^2 \times \tau^4 \rceil$, where τ is the average value of each byte in the message. The number of messages has little effect on the storage resource consumption during prime number generation. We use the length of the prime number (bits) as the abscissa and the storage resource overhead as the ordinate. When $c = 15$ and $d = 15$, we respectively draw the matrix U and V, the storage resource cost of prime number generation and the traditional method, as shown in Figure 4.

In Figure 4(a), when the number of messages and $\tilde{P}$ are small, the storage resources occupied by the attribute matrix are greater than the prime number generation process. When $\tilde{P} = 512$ or $\tilde{P} = 1024$, the storage resource of the prime number generation process is larger than the attribute matrix. In Figure 4(b), when the number of messages is larger, the space occupied by the attribute matrix will always be greater than the prime number generation process. However, the storage resource cost of CK-SKGET will always be greater than the storage resources occupied by the prime number generation process used in the traditional DHKE protocol and the RSA key exchange protocol. The storage resource occupied by CK-SKGET is maintained at 1-100 Kb, and the increased storage overhead is still acceptable compared to the storage capacity of current smart devices.

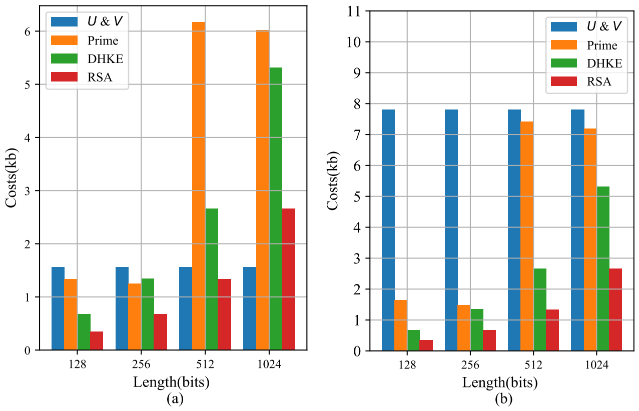


Figure 4: Comparison of storage resource cost. (**a**) The cost of storage resource when the number of messages is $m = 10$. (**b**) The cost of storage resource when the number of messages is $m = 50$.

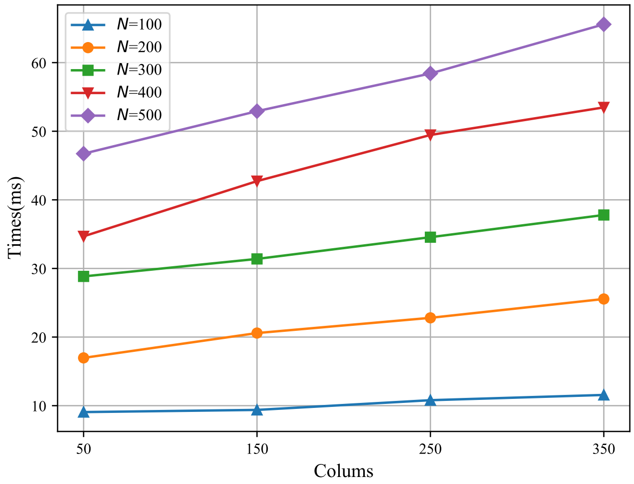


Figure 5: Comparison of computing cost for generating common knowledge

6.3 Computing Cost Analysis

The computing cost of CK-SKGET is mainly reflected in the process of common knowledge generation and the key generation based on common knowledge. The main computing cost of the latter comes from the large prime number generation process and the key agreement process. And the computing cost of the key agreement process is less than that of the large prime number generation process. Therefore, the computing cost of CK-SKGET mainly consumes time in the process of calculating common knowledge and computing large prime numbers.

6.3.1 Cost Analysis of Computing Common Knowledge

In the common knowledge generation stage, the main computing cost includes matrix operations and the eigenvalue decomposition of the common knowledge matrix. The computing cost of the former is related to effective messages N and columns of the content matrix d, but the computing cost of the latter is related to columns of the control matrix c. We take d as the abscissa and the computing cost of common knowledge as the ordinate. In the case of $c = 5$, we draw the computing cost curves under different situations, respectively, in Figure 5.

In Figure 5, when N is smaller, increasing d has a small impact on the time of the common knowledge generation process, and the computing cost of curves corresponding to different N is larger. But the value of N is larger, the impact on computing cost is higher. Therefore, the computing cost of common knowledge generation can be kept at a stable level by adjusting the number of messages and the number of columns in the content matrix.

6.3.2 Cost Analysis of Generating the Prime Number

In the prime number computing stage, the time of searching large prime numbers is a random value. And the cost of computing prime numbers increases with the length of prime numbers. The prime numbers in the factor base are better to compute than the large prime numbers sought. When changing the value of c, we can obtain a new number of rounds updating B since the count of eigenvalues is decided by the control matrix columns. After updated B, we compute the direct prime factor of the large prime number from the method in Section 4.2. Then we will compute P_d and find the prime number P closest to P_d. The time consumed in this process is related to the number of prime numbers, which is described in detail in the [18]. The two traditional key exchange methods to generate prime numbers are the same. So we compared the computing cost of CK-SKGET and traditional key exchange methods to generate prime numbers. And We use the number of prime numbers to be sought as the abscissa, and the time spent in computing large prime numbers under different prime factors is the ordinate, as shown in Figure 6.

In Figure 6, the cost of the prime number generation process is mainly related to the number of prime numbers to be sought and the number of prime factors (the number of columns c in the control matrix). From formula 20, we know that the rounds of updating B_i is related to the lengths $f(b_{i,j})$ of the prime number in B_i when seeking prime numbers with different lengths. The $f(b_{i,j})$ is larger, the rounds is more and the cost of time is longer. For example, when $\tilde{P} = 512(bits)$, $f(b_{i,j}) \leq 199(bits)$in the curve of $c = 6$, which is smaller than the curves of $c = 3, 4, 5$. But when $\tilde{P} = 1024(bits)$, $f(b_{i,j}) \leq 996(bits)$in the curve of $c = 6$, which is larger than the curves of $c = 3, 4, 5$, its cost of time is the largest. When the number of prime numbers is smaller, the com-

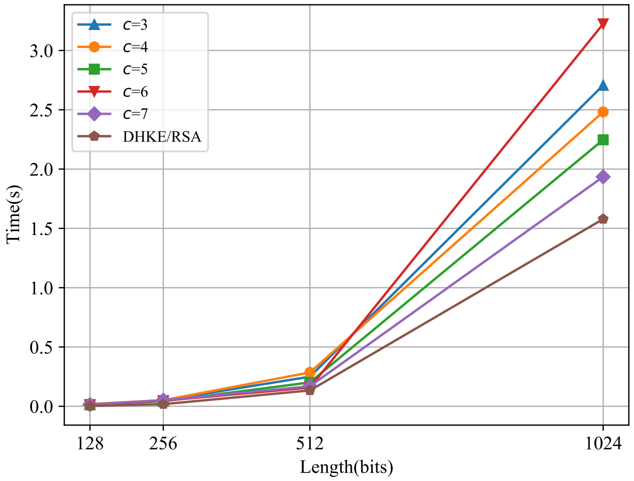


Figure 6: Comparison of computing cost for large prime numbers

puting cost of prime numbers is smaller. Therefore, the count of prime numbers in B_i base has a direct impact on the computing cost of prime number. Meanwhile, according to the prime number theorem [13], due to the larger interval between 512 bits and 1024 bits, the increase in computation time is larger than that between 256 bits and 512 bits [13]. Furthermore, the lowest computing cost of CK-SKGET is also close to that of the traditional key exchange method.

Comparing Figure 5 and Figure 6, when the number of prime numbers is larger, the computing cost of the common knowledge generation process is much smaller than the cost of the prime number generation process. So the cost of computing CK-SKGET mainly comes from the prime number generation process, considering that the prime number generation process also exists in the traditional key exchange method. The computing cost added by CK-SKGET only accounts for a very small proportion of the overall cost, which does not significantly increase the system computational burden, compared with the traditional key exchange method.

7 Conclusions

To resist the MITM attack in the key agreement process, this paper proposes a symmetric key generation and exchange technology based on common knowledge. Firstly, the communication peers obtain the same common knowledge through the common knowledge generation and verification algorithm. Then, they conduct symmetric key generation and exchange based on the common knowledge, and resist the MITM attacks in the key generation process. Theoretical analysis and simulation results show that CK-SKGET has theoretically provable security and can resist some typical attacks, thereby ensuring the security of the communication. Besides, compared with the traditional key agreement scheme, the storage cost and computing cost of CK-SKGET are not obvious. As for future work, the construction efficiency of common knowledge can be further improved, and the common knowledge can be extended to a wider range of application scenarios such as enhancing privacy protection, improving communication efficiency, and enabling tacit communication.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (U1705263), the Science Project of Yunnan Power Grid (YNKJXM20200172) and the Special Science Foundation of Quzhou (2021D013). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] M. D. Abbasinezhad and M. Nikooghadam, "Efficient anonymous password-authenticated key exchange protocol to read isolated smart meters by utilization of extended chebyshev chaotic maps," *IEEE Transactions on Industrial Informatics.*, vol. 14, no. 11, pp. 4815–4828, 2018.

[2] D. Adrian, K. Bhargavan, Z. Durumeric, P. Gaudry, M. Green, J. A. Halderman, N. Heninger, D. Springall, E. Thomé, L. Valenta, B. VanderSloot, E. Wustrow, S. Zanella-Béguelin, and P. Zimmermann, "Imperfect forward secrecy: How diffie-hellman fails in practice," in *Proceeding of 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 5–17, New York, NY, USA, 2015.

[3] H. M. Ahmed and R. W. Jassim, "Distributed transform encoder to improve diffie-hellman protocol for big message security," in *Proceedings of 3rd International Conference on Engineering Technology and its Applications*, pp. 84–88, IEEE, Iraq, 2020.

[4] S. Ali, A. Humaria, M. S. Ramzan, I. Khan, S. M. Saqlain, A. Ghani, J. Zakia, and B. A. Alzahrani, "An efficient cryptographic technique using modified diffie–hellman in wireless sensor networks," *International journal of distributed sensor networks*, vol. 16, no. 6, p. 1550147720925772, 2020.

[5] R. Alvarez, G. C. Caballero, J. Santonja, and A. Zamora, "Algorithms for lightweight key exchange," *Sensors.*, vol. 17, no. 7, pp. 15–17, 2017.

[6] M. Bellare and D. R. Pointcheval, "Authenticated key exchange secure against dictionary attacks," in *International conference on the theory and applications of cryptographic techniques*, pp. 139–155, Springer, 2000.

[7] T. Bui and T. Aura, "Key exchange with the help of a public ledger," in *Cambridge International Workshop on Security Protocols*, pp. 123–136, 2017.

[8] Y. Cao, N. Li, and J. Pan, "Key agreement protocol for dynamic identity authentication based on

chaotic mapping," *Mathematics in Practice and Theory*, vol. 51, no. 14, p. 9, 2021.

[9] C. Chunka, S. Banerjee, S. Nag, and R. S. Goswami. "A secure key agreement protocol for data communication in public network based on the diffie-hellman key agreement protocol,". in *Micro-Electronics and Telecommunication Engineering*, vol. 106, pp. 531–543, Singapore, 2020.

[10] Z. Dar, A. Ahmad, F. A. Khan, F. Zeshan, R. Iqbal, H. H. Sherazi, and A. K. Bashir, "A context-aware encryption protocol suite for edge computing-based iot devices," *The Journal of Supercomputing.*, vol. 76, no. 4, pp. 2548–2567, 2020.

[11] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE transactions on Information Theory.*, vol. 22, no. 6, pp. 644–654, 1976.

[12] J. Y. Eun and S. J. Il, "An efficient and secure diffie–hellman key agreement protocol based on chebyshev chaotic map," *Communications in Nonlinear Science and Numerical Simulation*, vol. 16, no. 6, pp. 2383–2389, 2011.

[13] P. Hoffman, *The Man Who Loved Only Numbers*, vol. 45. New York: Hyperion Books, 1998.

[14] J. Hurd, "Verification of the miller–rabin probabilistic primality test," *The Journal of Logic and Algebraic Programming*, vol. 56, no. 1-2, pp. 3–21, 2003.

[15] S. Inshi, R. Chowdhury, M. Elarbi, s. H. Ould, and C. Talhi, "LCA-ABE: Lightweight context-aware encryption for android applications," in *International Symposium on Networks, Computers and Communications*, pp. 1–6, IEEE, 2020.

[16] A. S. Khader and D. Lai, "Preventing man-in-the-middle attack in diffie-hellman key exchange protocol," in *22nd International Conference on Telecommunications*, pp. 204–208, IEEE, 2015.

[17] H. Lai, M. A. Orgun, J. Xiao, J. Pieprzyk, L. Xue, and Y. Yang, "Provably secure three-party key agreement protocol using chebyshev chaotic maps in the standard model," *Nonlinear Dynamics.*, vol. 77, no. 4, pp. 1427–1439, 2014.

[18] F. Li, Z. Gong, F. Lei, S. Gu, and P. Gao, "Overview of fast prime numbers generation," *Journal of Cryptologic Research.*, vol. 6, no. 4, pp. 463–476, 2019.

[19] L. Liu and J. Cao, "Analysis of one key agreement scheme for bans based on physiological features," *International Journal of Electronics and Information Engineering*, vol. 13, no. 4, pp. 142–148, 2021.

[20] B. Majid and R. A. Mohammad, "An attribute based key agreement protocol resilient to kci attack," *International Journal of Electronics and Information Engineering*, vol. 2, no. 1, pp. 10–20, 2015.

[21] N. Naher and M. M. Haque, "Authentication of diffie-hellman protocol against man-in-the-middle attack using cryptographically secure crc," in *International Ethical Hacking Conference 2018*, vol. 811, pp. 139–150, Singapore, 2019.

[22] R. P. Sah, U. N. Roy, A. K. Sah, and S. K. Sourabh, "Early proofs of fermat's little theorem and applications," *International Journal of Mathematics Trends and Technology.*, vol. 64, no. 2, pp. 74–79, 2018.

[23] W. Shen, Y. Cheng, B. Yin, J. Du, and X. Cao, "Diffie-hellman in the air: A link layer approach for in-band wireless pairing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 11894–11907, 2021.

[24] W. Shen, W. Hong, X. Cao, B. Yin, D. Shila, and Y. Cheng, "Secure key establishment for device-to-device communications," in *IEEE Global Communications Conference*, pp. 336–340, Austin, USA, 2014.

[25] P. P. Thwe and M. Htet, "Prevention of man-in-the-middle attack in diffie-hellman key exchange algorithm using proposed hash function," *International Journal of Advances in Scientific Research and Engineering (IJASRE)*, vol. 5, p. 10, 2019.

[26] Y. Zhang, Z. Wang, Z. Wang, and H. Chen, "Verifiable three-party secure key exchange protocol based on eigenvalue," *Journal of communication.*, vol. 40, no. 12, pp. 149–154, 2019.

[27] L. Zhou and T. Hu, "Safe primes construction algorithm based on laime primes judgment theorem," *Computer Engineering and Applications.*, vol. 52, no. 13, pp. 152–156, 2016.

Biography

Hexiong Chen, born in 1984, received the master degree from Kunming University of Science and Technology in June 2011, and now working at Yunnan Power Grid Company Limited. His research interests include Network Technology, Network Security, Network Operation and Maintenance.

Jiaping Wu, born in 1998. He is currently a M.S. candidate in University of Electronic Science and Technology of China. His research interests include Internet of Things Security and Blockchain Technology.

Wei Guo, born in 1986, received the bachelor degree and now working at Yunnan Power Grid Company Limited. His research interests include Network and Network Security Maintenance and Management.

Feilu Hang, born in 1984, received the master degree from Yunnan University in June 2014, and now working at Yunnan Power Grid Company Limited. His research interests include Network Security Attack and Defense Technology.

Zhenyu Luo, born in 1985, received the bachelor's degree and now working at Yunnan Power Grid Company Limited. His research interests include Network Technology, Network Security, Network Operation and Maintenance.

Yilin Wang, born in 1998. She is currently a M.S. candidate in University of Electronic Science and Technology of China. Her research interests include Internet of Things Security and Blockchain Technology.

Construction and Deployment of a Distributed Firewall-based Computer Security Defense Network

Chunjuan Wang
(Corresponding author: Chunjuan Wang)

Shaanxi Xueqian Normal University
No. 69, Xingshan Temple East Street, Yanta District, Xi'an, Shaanxi 710061, China
Email: miaotuichun788@126.com

(Received Sept. 22, 2019; Revised and Accepted Dec. 5, 2022; First Online Dec. 13, 2022)

Abstract

The development of the Internet has greatly facilitated the exchange of information between people, but it has also provided a convenient channel for transmitting malicious information. Malicious data transmitted on the Internet will negatively impact normal users; thus, prevention measures are needed, and a firewall is an effective protection measure. This paper introduced the traditional firewall and the distributed firewall based on a software-defined network (SDN) structure. First, a local area network (LAN) was built with several servers in a laboratory. Then, simulation experiments were carried out on the traditional firewall and the distributed firewall to test the network's throughput under both firewalls and the protection against attacks from the internal and external networks. The results suggested that the existence of a firewall affected the throughput of the network; the impact of the distributed firewall on the network throughput was smaller than that of the traditional firewall; both the traditional firewall and the distributed firewall could effectively intercept abnormal data from the external network, but only the distributed firewall could effectively intercept the abnormal data from the internal network.

Keywords: Distributed; Firewall; Network Defense; Software-Defined Network

1 Introduction

The advent of computers has greatly satisfied people's computing needs. As computers perform increasingly better, the tasks they can perform have become more and more diverse [3]. The Internet is developed together with computers. The Internet is a network in which a plurality of computers exchanges data under agreed communication rules. Data from computers can be transmitted either wired or wirelessly. As computers are difficult to face each other, so only when the data are received can the computer judge whether the data is malicious or not. If there is no proper protection, then when the user finds the malicious data, it is already too late to stop the damage to the computer [15].

Firewall technology is a network security defense means that divides the Internet into intranet and extranet like a wall. The protected computer is on the intranet. Unknown data will first pass through the firewall when transmitted from the extranet to the intranet, and the firewall will intercept the data that are in line with the interception rules to protect intranet data. The traditional firewall simply divides the network into intranet and extranet. Although it can intercept risky data from the Internet, the data from the intranet is not guarded. Once the attack is launched from the intranet by bypassing the firewall, even though the firewall can intercept the continuous attack from the extranet, the data that causes damage on the intranet cannot be intercepted [1].

In order to solve the shortcomings of traditional firewalls, distributed firewalls were brought up. Distributed firewalls distribute the defense function in the intranet, which simply means that the intranet is no longer considered a trusted area. Filipek *et al.* [6] defined a security model for a mobile ad-hoc network using public key infrastructure (PKI), firewalls, and intrusion prevention system (IPS) and found through experiments that the model helped mitigate and prevent the most common attacks. Tran *et al.* [14] faced the problem of time-setting delays and controller overhead in software-defined network (SDN) firewalls and proposed a topology-aware selective firewall distribution scheme. They conducted simulations and found that the scheme significantly reduced the firewall setup traffic and the firewall-violated traffic travel route and was suitable for large-scale SDN. Chang *et al.* [4] proposed a robust algorithm for distributing security policies from firewalls to distributed SDN devices in a cloud cluster environment, validated the performance of the algorithm through experiments, and solved the memory limitation of Ternary Content Addressable Memory (TCAM) in SDN devices. After introducing traditional firewalls, this paper proposed using the SDN structure to build distributed firewalls and conducted simulation experiments on traditional and distributed firewalls in the laboratory local area network (LAN).

2 Distributed Firewall

2.1 Traditional Firewall

The implementation principle of the firewall technology used for computer network security protection is similar to separating the fire area from the safe area using the firewall body in real firefighting measures, and that is why it is called "firewall" [9]. Figure 1 shows the basic structure of a traditional firewall. The internal network is the protected area, and the Internet is the untrusted area with risks. They are interconnected by a series of components (e.g., routers, bastion hosts, etc.), which are firewalls. The exchange of data between the Internet and the internal network can only take place through the ports provided by the firewall. In terms of the functions, firewalls can be broadly classified as packet filter firewalls, application layer gateway firewalls, and content filter firewalls [8]; in terms of topology, they can be divided into dual-host host structures, screened host structures, and screened subnet structures [5].

Internet
Firewall
Intranet

Figure 1: The basic structure of a traditional firewall

2.2 SDN-Based Distributed Firewall

In order to improve the shortcomings of traditional firewalls that prevent the outside but not the inside, distributed firewalls are brought up. While the traditional firewall divides the network into intranet and extranet in terms of physical topology, the distributed firewall changes the intranet from the physical sense to the logical sense. The overall structure of distributed firewalls is divided into a network firewall, a host firewall, and a policy management center [13]. The network firewall is the traditional firewall, which is used to isolate the intranet and the extranet, but it is no longer responsible for all the protection work. The host firewall is responsible for the main defense work in every server or personal computer in the intranet, which simply means that the information transmitted between different servers in the intranet also has to pass through the firewall. The policy management center is responsible for formulating the security rules of the firewall. The administrator sets the security rules in the policy management center, and the management center distributes the security rules to every firewall. Every firewall performs defense work according to the assigned rules.

In practice, if a distributed firewall works in a small network, it is possible to set up a firewall for every server because the structure is relatively simple and the number of servers in the intranet is small [10], but as the size of the network increases, the structure becomes complex, and the number of servers also increases, so the cost of configuring a firewall for every server will also increase. This paper proposes a distributed firewall under the SDN architecture. Its structure diagram is shown in Figure 2. The actual application process will require more equipment; limited by space, Figure 2 has been simplified to some extent. The basic structure of an SDN can be divided into the application layer, control layer, and data layer, where the application and control layers are usually in the same position in the physical structure, i.e., the SDN controller in Figure 2. The function of the application layer is to visualize the information from the data layer and the control layer, which is convenient for the administrator to change the network rules. The control layer is the core of the whole network, which parses the firewall rules from the application layer, converts them into OpenFlow rules, and sends them to the switches in the data layer. The plural SDN switches constitute the data layer of SDN, and the switches execute the work of data forwarding under the rules issued by the controller [2].

In the distributed firewall with an SDN architecture, the SDN controller, SDN switch, and the client are all in the intranet, the intranet is connected to the Internet through the filtering router, and the filtering router and intranet are also connected to the bastion host. It is overall similar to the screened host structure of the traditional firewall, but the difference is that the intranet is an SDN structure and the application of the SDN controller provides a firewall for the intranet. The SDN controller is only responsible for formulating and issuing data forwarding rules in the SDN architecture, and the SDN switch is only responsible for executing the rules to forward data [7].

When the SDN-based distributed firewall is working, the administrator first designs the firewall rules in the management interface provided by the application layer of the SDN controller. The firewall rules can be formulated for the firewall of a single switch or a firewall group composed of multiple switches. The designed rules are first saved to the database so that the rules do not need to be set again after the firewall is restarted. Then, the rules are transmitted to the control layer of the controller through the Application Programming Interface (API) [11]. The firewall module in the control layer parses the rules, converts them into OpenFlow rules, and sends them to the SDN switch. When the Internet wants to access the client server on the intranet, it first passes through a filtering router with a firewall. Since this router is the channel connecting the intranet and the extranet, the firewall rules

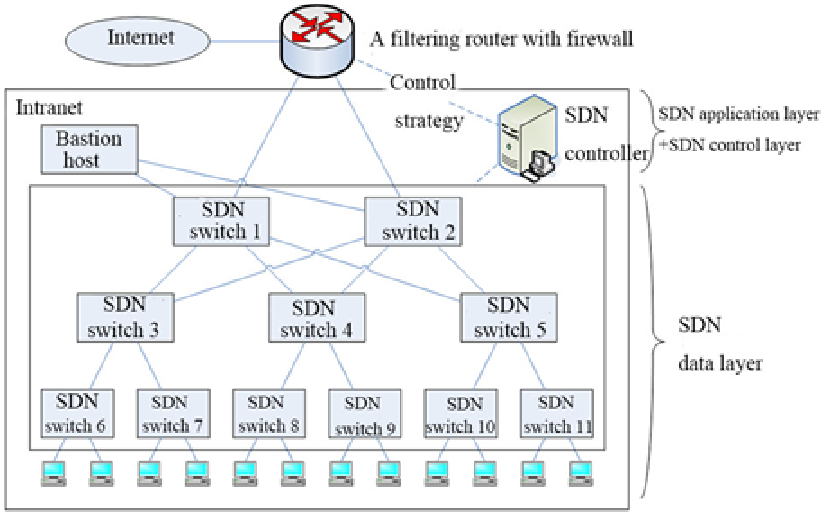


Figure 2: Schematic diagram of SDN-based distributed firewall architecture

are generally simple. After passing through the filtering router, the access data enter the intranet and reaches the target server through the forwarding of the switch. Once the access data meet the firewall rules of some switch in this process, they will be intercepted. Data access between servers on the intranet is also like this. The access data reaches the target port through the switch and will be directly intercepted once they match the firewall rules [12].

3 Simulation Experiments

3.1 Experimental Environment

The simulation experiments were conducted in a LAN built in a lab. The server configurations were 16 G memory and a Windows 10 operating system.

3.2 Experimental Setup

Fourteen servers from 0 to 13 were used to build the SDN structure. Server 0 served as the SDN controller, server 12 served as the bastion host, and server 13 acted as the extranet. Servers 1 to 11 served as the SDN switches. Although only one SDN controller was used to manage and control the network in the SDN architecture, there were plural switches used for data forwarding. Every switch could act as a firewall, which posed some difficulties for management; therefore, the switches were grouped, where switches 1 and 2 constituted the core group, switches 3, 4, and 5 constituted the aggregation group, and switches 6 ∼ 11 constituted the access group. Interconnecting links between the core group, aggregation group, and access group are shown in Figure 2. The access group was mainly responsible for the access of the user side in the intranet, and every switch in the access group accessed two user sides. The aggregation group grouped the switches of the access group together, and every switch of the aggregation group accessed two switches of the access group. The core group was the transit core of the edges of the intranet and extranet, and every switch in the core group was connected to every switch in the aggregation group. The controller sent firewall rules to the filtering routers connected to the intranet and extranet using the Secure Sockets Layer (SSL) protocol and to the switches in the intranet using the OpenFlow protocol. Table 1 shows the firewall rules issued by the controller to the switches.

In order to verify the performance of the SDN-based distributed firewall, a traditional firewall with a screened host structure was also built to compare with the distributed firewall. In order to ensure that only the firewall settings were different between the two firewalls for comparison experiments, the SDN topology in the distributed firewall was still used when setting up the traditional firewall with the screened host structure, and the difference was that the SDN controller only sent firewall rules to the filtering routers and sent regular OpenFlow rules to the switches in the intranet.

3.3 Experimental Projects

1) Throughput comparison between distributed and traditional firewalls
Firstly, server 13 sent data to the user side in the intranet at a rate of 30 mb/s. At this moment, the firewall has not been turned on in both networks. After the data transmission lasted for 20 s, both networks turned on the firewall by issuing the corresponding firewall rules through the SDN controller, and the rules of the distributed firewall are shown in Table 1. The rule issued by the traditional firewall to the filtering router was that packets passing through port 22 were discarded. The throughput of the filtering router was tested before and after the firewall was

Table 1: Firewall rules issued by the SDN controller to different groups of switches

Group	Number in the group	Firewall rules
Core group	1, 2	None
Aggregation group	3, 4, 5	Packets passing through port 22 are discarded
Access group	6 ∼ 11	Packets passing through port 445 are discarded

turned on in both networks.

2) Defense of two firewalls against attacks on the intranet and extranet
Firstly, the defense against attacks on the extranet was tested. Server 13 sent 500 packets to the user side of the intranet, of which 300 packets were normal data and were transmitted through port 80, and the remaining 200 packets were abnormal data and were transmitted through port 22. Both networks sent 500 packets from server 13 before and after opening the firewall. The packets received by the user side before and after opening the firewall were detected.

Then, the defense against intranet attacks was tested. The user side connected to switch 6 sent 500 packets to other user sides in the intranet, 300 packets of which were normal data and were transmitted through port 80, and the remaining 200 packets were abnormal data and were transmitted through port 445. Both networks sent 500 packets from the user side accessing switch 6 before and after opening the firewall. The packets received by other user sides before and after opening the firewall were detected.

3.4 Experimental Results

Figure 3 shows the throughput changes of the filtering routers of the two networks before and after the firewall is turned on. It was noticed from Figure 3 that the throughput of the filtering routers in the first 20 s before opening the firewall in both networks was almost the same, slightly less than 30 mb/s, and remained stable; while at 20 s, due to the opening of the firewall, the throughput of the filtering routers changed significantly: the throughput of the filtering routers in the network with the traditional firewall significantly reduced, while the throughput of the filtering routers in the network with the distributed firewall reduced less significantly. The reason is as follows. When the firewall was not turned on, the topology of the two networks was the same, and the transmission of data in the extranet was not blocked, so the difference in throughput between the two networks was not significant and close to the transmission rate; however, when the firewall was turned on, the firewall needed to judge the data first and decide whether the data could pass or not, resulting in the router not being able to transmit data at the original rate. In particular, the traditional firewall network set the firewall in the filtering router, i.e., the only interface of the intranet and extranet, which directly impacted the throughput of the entire intranet after opening the firewall; the network with the distributed firewall only added simple firewall rules to the filtering router and did not set the firewall in the core group switch with the largest traffic, which made the throughput of the intranet reduce less.

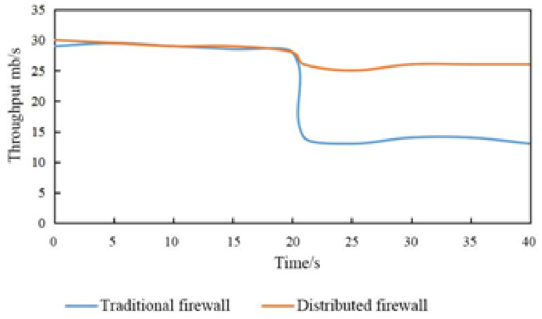


Figure 3: Changes in filtering router throughput in both networks before and after turning on the firewall

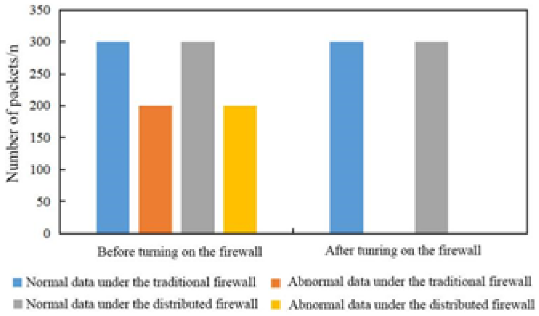


Figure 4: Packet throughput of the two networks before and after opening the firewall under extranet attacks

Figures 4 and 5 show the packet throughput within the two networks before and after turning on the firewall under the attack from the extranet and intranet, respectively. It was seen from Figure 4 that the attack of abnormal data came from the extranet, there were normal data and abnormal data in the data transmitted from the extranet to the user side of the intranet, and both networks let the normal data and abnormal data pass before turning on the firewall; after turning on the firewall, both firewall networks still let the normal data pass, while

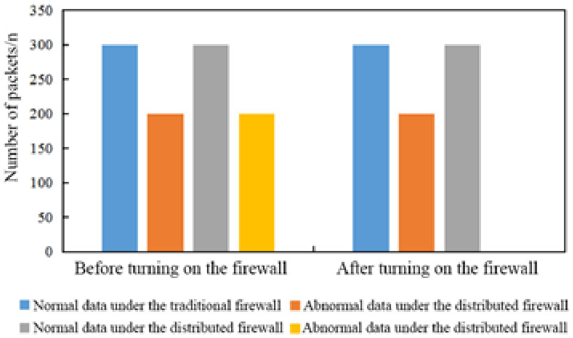


Figure 5: Packet throughput of the two networks before and after opening the firewall under intranet attacks

the abnormal data could not pass because the port was closed, which meant that both the traditional firewall and the distributed firewall could effectively block the attacks from the extranet.

Then, it was seen from Figure 5 that the attack of abnormal data came from the intranet, and there were both normal and abnormal data in the data transmitted from the user side of the intranet to the other user side; before turning on the firewall, both firewalls allowed normal and abnormal data to be transmitted in the intranet; after turning on the firewall, normal and abnormal data were still transmitted in the intranet under the traditional firewall, while only normal data were transmitted in the intranet under the distributed firewall. These results indicated that the traditional firewall could not play the role of normal interception in the face of an attack from the intranet, but the distributed firewall could effectively intercept the abnormal data.

4 Conclusion

This paper introduced the traditional firewall and the SDN structure-based distributed firewall. A LAN was built in the lab using several servers, and simulation experiments were conducted on the traditional firewall and the distributed firewall to test the throughput of the network and the defense against attacks from the intranet and extranet under the two firewalls. The results are shown below. (1) The network throughput reduced when either the traditional firewall or the distributed firewall was turned on, and the network' throughput decreased more significantly when the traditional firewall was turned on. (2) After turning on the firewall, both the traditional firewall and the distributed firewall could effectively intercept the abnormal data from the extranet; the traditional firewall could not intercept the abnormal data from the intranet, but the distributed firewall could.

References

[1] F. K. Abid, A. M. Makhlouf, F. Zarai, M. Guizani, "DVF-fog: distributed virtual firewall in fog computing based on risk analysis," *International Journal of Sensor Networks*, vol. 30, no. 4, pp. 242, 2019.

[2] H. Bedi, S. Shiva, S. Roy, "A game inspired defense mechanism against distributed denial of service attacks," *Security and Communication Networks*, vol. 7, no. 12, pp. 2389-2404, 2015.

[3] M. Caprolu, S. Raponi, R. Di Pietro, "FORTRESS: An efficient and distributed firewall for stateful data plane SDN," *Security and Communication Networks*, vol. 2019, pp. 1-16, 2019.

[4] Y. Chang, T. Lin, "Cloud-clustered firewall with distributed SDN devices," in *IEEE Wireless Communications & Networking Conference*, pp. 1-5, 2018.

[5] A. K. Chawan, S. S. Mahajan, "Solving firewall policy anomalies using generic algorithm," *International Journal of Engineering Trends and Technology*, vol. 20, no. 4, pp. 200-203, 2015.

[6] J. Filipek, L. Hudec, "Securing Mobile Ad Hoc Networks using distributed firewall with PKI," *IEEE International Symposium on Applied Machine Intelligence & Informatics*, pp. 321-325, 2016.

[7] J. Hwang, T. XIe, F. Chen, A. X. Liu, "Fault localization for firewall policies," in *Proceedings of the IEEE Symposium on Reliable Distributed Systems*, pp. 100-106, 2015.

[8] Y. Jarraya, A. Eghtesadi, S. Sadri, M. Debbabi, M. Pourzandi, "Verification of firewall reconfiguration for virtual machines migrations in the cloud," *Computer Networks*, vol. 93, no. DEC.24, pp. 480-491, 2015.

[9] F. Kamoun-Abid, A. Meddeb-Makhlouf, F. Zarai, Guizani, "Distributed and cooperative firewall/controller in cloud environments," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pp. 1-10, 2018.

[10] S. Kumari, P. Singh, R. K. Upadhyay, "Virus dynamics of a distributed attack on a targeted network: Effect of firewall and optimal control," *Communications in Nonlinear Science and Numerical Simulation*, vol. 73, no. JUL., pp. 74-91, 2019.

[11] H. Lee, S. Lee, K. Kim, H. K. Kim, "HSViz: Hierarchy simplified visualizations for firewall policy analysis," *IEEE Access*, vol. 9, pp. 71737-71753, 2021.

[12] M. Lyu, H. H. Gharakheili, C. Russell, V. Sivaraman, "Hierarchical anomaly-based detection of distributed DNS attacks on enterprise networks," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1031-1048, 2021.

[13] V. Sobeslav, L. Balik, O. Hornig, J. Horalek, O. Krejcar, "Endpoint firewall for local security hardening in academic research environment," *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 2, pp. 1475-1484, 2017.

[14] T. V. Tran, H. Ahn, "A network topology-aware selectively distributed firewall control in SDN," in *International Conference on Information & Communication Technology Convergence*, pp. 89-94, 2015.

[15] B. Wahyuaji, "CNDS-SYN flood prevention using distributed firewall in software-defined WAN architecture," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1, pp. 182-186, 2019.

Biography

Chunjuan Wang, born in 1981, has received the master's degree. She is a lecturer in Shaanxi Xueqian Normal University. She is interested in the distributed System application.

The Validation Code System of Websites Based on Multi-dimensional Chaotic Logistic Map

Dahua Song[1], Chang Liu[1], and Jiahui Liu[2]
(Corresponding author: Chang Liu)

Center of Educational Technology and Information, Mudanjiang Medical University[1]
Mudanjiang 157011, China
Email: bsuljh@163.com
College of Computer Science and Technology, Harbin University of Science and Technology[2]
Harbin 150080, China

(Received July 4, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 13, 2022)

Abstract

Validation code systems have been widely used in websites and forums as CAPTCHA. The attack on the validation code system poses a significant threat to the website. Many validation code systems use random number generators to generate validation code sequences. The logistic map can produce a chaotic, random sequence. However, chaotic sequences have the problem of short periods. In this paper, the multi-dimensional chaotic Logistic map is used to generate the chaotic, random sequence, and the chaotic sequence is arranged by random allocation in the validation code graph for websites. Experiments show that the proposed method can produce good random distribution sequences and effectively resist the attack on the validation code system.

Keywords: Chaotic Sequence; Logistic Map; Multi-dimensional Chaos; Short Period; Validation Code

1 Introduction

The validation code system originated from CAPTCHA system, which is the abbreviation of ”Completely automated public Turing test to tell computers and humans apart” [9,14]. It is an automatic program to distinguish whether the user is a computer or a human [3].

The validation code system can prevent malicious password cracking, ticket swiping and irrigation for forums. It effectively prevents a hacker from constantly trying to log in to a specific registered user by brute force cracking with the specific program. In fact, the use of the validation code is a common protection method for many websites [20]. Using the validation code system, the questions can be generated and judged by the computer, but only human beings can answer them. Because the computer cannot answer CAPTCHA questions, the user who answers the questions can be regarded as human beings [16,18].

Most validation code systems generate random sequences through random number generators, and then embed them into graphs with the validation code for users to recognize. Chaotic random number is a widely used random number generator [8,12,15]. Logistic map is a typical chaotic dynamic equation. In chaotic dynamics theory, when the trajectory of Logistic map changes with time in the continuous real number domain, Logistic map shows its own complex changes, especially its aperiodic, unpredictable and good randomness. However, in order to avoid short period [21], improvements for Logistic map are needed.

At present, the validation code system has not been clearly defined. Generally speaking, the validation code is a string of the validation information sent from the server computer to the client end, when the internet users or the mobile terminal need to access the online server computer system. If the server computer system confirms that it is manually input by users and the input the string of the validation information is correct, this user can log in to the online server computer system for getting service of operation, communication, and so on.

The validation code system has been widely used [4]. Some examples are as follows. When online banking customers log in to their own account for query or other operations, they need to input a string of verified characters or numbers and other information on the login page of websites. After entering correctly, the user can enter their personal account. In addition, in some websites for the secure email service, when the user enters the wrong login password for the first time, the system will prompt the user to enter a string of validation code information. If it is correct, the user can enter the account smoothly. Chinese railway order-ticketing system of 12306 website needs to input the validation code online during query, order and other operations. If the validation code is entered correctly, the user can carry out next operations in

12306 website.

The main function of validation code is to protect the security of online server computer system and services [19, 24]. For example, the validation code can prevent hackers and illegal users from brutally cracking passwords by using computer programs. Hackers can use the computer program to automatically access to the user login page of the online server computer system and to input the user password with the computer program [17]. If there is no protection of the validation code system, hackers will use the computer program to conduct continuous password input test, so as to achieve the purpose of violent cracking. The validation code system can prevent malicious password cracking and protect the security of user password. In addition, the validation code system can prevent hackers and unauthorized users from malicious damage to the website. The validation code system can prevent hackers from using computer programs to register maliciously.

Due to the protection of the validation code system, the validation code system is often attacked by hackers [7]. With the development of computer technology and the popularization of multi-core technology, the computing and data processing ability of computer has been greatly improved. Therefore, many validation code systems have many problems [2, 13]. Here are some typical examples.

Many validation code systems use random numbers for verification. Random numbers are generated by computer programs. Can the limited random numbers meet the validation code requirements of the online server computer system? For example in one day, if the computer generates a random number per second, the computer will generate 60(second)*60(minute)*24(hour)=86400 every day. Theoretically, the computer needs 86400 validation codes per day. In fact, if the random number is not generated by the parallel program or many server computers, the validation code generated by the random number generator of the single core computer program is outputted continuously. In a certain period of time, validation codes that can be intercepted are continuous. Therefore, if the time period is very short, the number of validation codes generated is very limited. If hackers choose a specific time period, they use computer programs to attack the validation code system, which is easy to crack.

Besides, the random number generated by the random number generator of some validation code systems has a short period and weak randomness, which is easy to be cracked by computer programs, resulting in a threat to the online server computer system and the destruction of the website. Therefore, the selection of random source is very important for the validation code system.

Moreover, some CAPTCHA systems focus on the imperceptible system with cursor trajectories [22], the emerging-image motion system [5] and the graphical password system [23]. The validation code system is added with uncertainty to the process [10]. Besides, CAPTCHA systems are in different languages, such as interactive text-based handwritten Arabic system for mobile devices [1].

When the traffic of some websites increases sharply in some special periods, some validation code systems are difficult to respond to the request response of the client quickly, resulting in problems such as long waiting time and service delay.

The validation code system is widely used in current social networks and login systems. The main purpose is to prevent hacker cracking and program automatic attacks. The structure of this paper is as follows: the first part describes the validation code system and the main problems. The second part includes Logistic map and periodic point problem. The third part introduces the implementation of the validation code system with the multi-dimensional Logistic map. In the fourth part, the performance and experimental analysis are carried out. Finally, the conclusion and prospect of this paper are given.

2 Chaotic Logistic Map

Random number generator is a method to generate random numbers with long period and good statistics through the independent function of random number. Many random number generators have the problem of short period and periodic point. Logistic map is widely used in random number generator in computers, such as generating random number, generating hash value and so on.

2.1 Logistic map

Logistic map is a typical chaotic dynamic equation. In chaotic dynamics theory, when the trajectory of Logistic map changes with time in the continuous real number field, Logistic map shows its own complex changes, especially its aperiodic, unpredictable and good randomness. Characteristics of Logistic map in the real orbit of continuous real number field are simulated by computer. Because the discrete numerical value is used to represent the change of orbit in the computer, the degradation characteristics in the real orbit are caused.

The mathematical definition of Logistic map is different from that in computers. In the continuous real number field, the time series of real orbit changes continuously, and its value will not have round-off errors. In the discrete computing orbit of the computer, due to the discrete numerical representation, points of the computing orbit already have quantization errors, and such round-off errors are difficult to control and predict. Therefore, there may be that the discrete computing orbit has actually deviated from the real orbit and jumped between different orbits. In order to formalize Logistic map in computer, which is described in Equation (1) as follows:

$$x_{n+1} = c * x_n * (1 - x_n) \tag{1}$$

where x_n is the time sequence of Logistic map, n represents time. x_0 represents the initial value of x_n, and c

represents the control parameter of Logistic map. x_n is set in the variation range of a positive real number between 0 and 1. The variation range of c is a positive real number between 3.56 and 4, including the boundary value of 4.

The trajectory of Logistic map numerically simulated in the computer has two parts. The first part is the transition orbit. However, with the limitation of the calculation accuracy in the computer the orbit of Logistic map finally enters the cycle orbit under the discrete calculation in computers. No computer can provide infinite accuracy to simulate and calculate Logistic map. Therefore, the characteristics of orbit calculated by numerical simulation are difficult to really approach the real orbit. Some typical examples have been studied in Ref. [11]. The ideal chaotic orbit is greatly affected by the accuracy in the computer of discrete numerical calculation.

In theoretical research and experimental tests, when the control parameter of Logistic map is close to 4, the change of Logistic map is the most complex, and the chaotic characteristics are close to the real orbit.

The control parameter of Logistic map enters the chaotic state after 3.56. When the control parameter of Logistic map is close to 4, the bifurcation points of Logistic map are all over the whole space, and the complex characteristics of chaos are close to the real orbit. When the calculation accuracy of Logistic map is higher, the characteristics of its orbit are closer to the ideal chaos. Therefore, the accuracy has a great impact on the chaotic orbit.

2.2 Fixed Point of Logistic Map

There are also fixed points and the short period problem in chaotic sequences [6]. Since the fixed point and short period have important influence on random number generator, the research on the fixed point and short period is not only an important research problem in the theoretical research of chaos, but also the key problem in the application of the chaotic random sequence. Effectively avoiding the fixed point and short period of the orbit in the discrete field is very important to produce a good random chaotic sequence. Analysis from the perspective of theory and experiment is also an effective means to judge the advantages and disadvantages of random number generator.

This section mainly studies the fixed point in the orbit of Logistic map and the phenomenon of the short period near the fixed point. The analysis of the short period phenomenon near the fixed point is not only an effective way to reveal the relationship between the short period and the fixed point, but also an effective experimental method to study the characteristics of chaotic equations in the computational orbit of discrete numerical simulation.

Fixed point refers to the phenomenon that the orbital sequence of chaos remains unchanged at one point or some points with the change of time in the iterative process. Such points are called fixed points. From the definition of Logistic map, it can be deduced that when $c = 4$, the chaotic orbit of Logistic map is close to real orbit. However, Logistic map has fixed points under certain conditions.

The control parameter setting value of Logistic map is $c = 4$, the initial value of x_n is set to 0.75 in Equation (1). In addition, the given accuracy in computing Logistic map is 2. The iterative process of Logistic map is as follows:

At x_1, the calculation process is $x_1 = c * x_0 * (1 - x_0) = 4 * 0.75 * (1 - 0.75) = 4 * 0.75 * 0.25 = 0.75$. At x_2, it is still 0.75, that is $x_2 = 0.75$. With the change of time, x_n is still 0.75, that is $x_n = 0.75$.

No matter how large the value of x_n is, the time series will not change. It can be seen that as long as the given accuracy is greater than or equal to 2, Logistic map has a fixed point $x_n = 0.75$ in the orbit with the condition that the control parameter is $c = 4$ and the initial value of x_0 is 0.75. The trajectory of Logistic map does not change at this point.

Even in the continuous real number field, the fixed point with $x_n = 0.75$ exists. In addition, when the control parameters of Logistic map are $c = 4$ and $x_0 = 0.25$ in Equation (1), Logistic map will also enter the fixed point $x_n = 0.75$. Obviously, at the fixed point $x_n = 0.75$, $1 - x_0 = 0.75$ and $x_0 = 0.25$, it finally enters the fixed point. $x_n * (1 - x_n)$ has the same value in the initial value of $x_0 = 0.75$ and $x_0 = 0.25$ in Equation (1). Therefore, the following section uses $x_0 = 0.75$ as the initial setting for exploring the short period of the orbit in Logistic map.

2.3 Short Period

The short period in the initial value of $x_0 = 0.75$ is studied for Logistic map in this section. The chaotic orbit of Logistic map can be divided into transition period part and cycle period part. Assume that the point of the chaotic orbit of Logistic map is as follows:

$$X_1, X_2, X_3, \cdots, X_{i-1}, X_i, X_{i+1}, X_{i+2}, \cdots, X_m, \cdots, \\ X_{i+k-1}, \cdots, X_i, X_{i+1}, \cdots, X_m, \cdots$$

Among points, the sequence $X_1, X_2, X_3, \cdots, X_{i-1}$ with a length of $(i - 1)$ is the transition period part. The point of sequences $X_i, X_{i+1}, X_{i+2}, \cdots, X_m$ has a cycle in the chaotic sequence. The length is $(m + 1)$, which is defined as the cycle period part. In the finite precision with the change of time Logistic map finally enters the cycle.

Logistic map are set to $c = 3.8$ and $x_0 = 0.75$ in Equation (1). The computing accuracy is set to 2. With changes of time the sequence of Logistic map obtains as follows:

$x_1 = 0.71, x_2 = 0.78, x_3 = 0.65, x_4 = 0.86,$
$x_5 = 0.45, x_6 = 0.94, x_7 = 0.21, x_8 = 0.63,$
$x_9 = 0.88, x_{10} = 0.4, x_{11} = 0.91, x_{12} = 0.31,$
$x_{13} = 0.81, x_{14} = 0.58, x_{15} = 0.92, x_{16} = 0.27,$
$x_{17} = 0.74, x_{18} = 0.73,$
$x_{19} = 0.74, x_{20} = 0.73,$
$x_{21} = 0.74, x_{22} = 0.73,$
$x_{23} = 0.74, x_{24} = 0.73.$

It can be seen that from x_1 to x_{16} the sequence is the transition period part, and the period length is 16. From x_{17} to x_{24} the sequence has cycles, and the value of 0.74 and 0.73 has cycles. Therefore, this part is the cycle period part with the length of 2.

Furthermore, the parameter of Logistic map is set to $c = 3.7$ and $x_0 = 0.75$, and the computing accuracy is set to 2. The sequence obtains as follows:

$x_1 = 0.69, x_2 = 0.79, x_3 = 0.61,$
$x_4 = 0.88, x_5 = 0.39,$
$x_6 = 0.88, x_7 = 0.39.$

The sequence from x_1 to x_3 is the transition period part, and the period length is 3. There are cycles from x_4 to x_7, and the cycle length is 2 in the value of 0.88 and 0.39.

Although the transition period length of Logistic map is longer when the control parameter is 3.8 than $c = 3.7$, it is obvious that the cycle period is particularly short, and the length is only 2. It poses a great threat to the chaotic sequence. As mentioned above, at $x_0 = 0.75$ Logistic map has the problem of the short period.

3 Proposed Method

Generally, the function of the random number can be independent or combined. For example, the random number sequence generated by one function is used as the basis, and then another generator is used to rearrange the random number sequence.

Pseudo random number is a sequence of random numbers calculated by deterministic algorithm, which seems to come from the uniform distribution. It is not really random, but it has statistical characteristics similar to random numbers, such as uniformity, independence and so on. When calculating pseudo-random number, if the initial value as seed used is unchanged, the sequence of pseudo-random number is also unchanged. Generally, the pseudo-random number with extremely long cycle period and passing the random number test is used in the simulation to ensure the randomness of the sequence. Figure 1 shows the structure of chaotic sequence with multi-dimensional Logistic map.

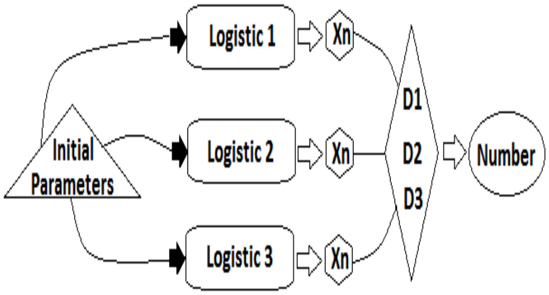


Figure 1: The structure of chaotic sequence with multi-dimensional Logistic map

The set of initial parameters includes as follows: three Logistic maps, the initial iteration number of Logistic map, the given computing precision, and so on.

$$IPSet = \left\{ \begin{array}{l} x_0^1,\ x_0^2,\ x_0^3,\ c_1,\ c_2,\ c_3, \\ iniIter_1,\ iniIter_2,\ iniIter_3, \\ OutNum, ModNum, Opt_{Position}, \end{array} \right\}$$

where x_0^1, c_1 and $iniIter_1$ are set to the first Logistic map which is named as Logistic 1. x_0^1 means the initial value of x_0 for Logistic 1. c_1 means the control parameter c of Logistic 1. The $iniIter_1$ is the initial iteration number of Logistic 1. $x_0^2, x_0^3, c_2, c_3, iniIter_2, iniIter_3$ are given to Logistic 2 and Logistic 3, respectively. $Opt_{Position}$ means the given computing precision in computers.

Each Logistic map is iterated with the initial value of x_0 and the control parameter of c in Equation (1). After Logistic map iterates with the number of the initial iteration under the computing precision of $Opt_{Position}$, every iteration step of Logistic map will output x_n which is belong to $(0, 1)$. x_n is converted to an integer number. $D1$ means the converting number of x_n for Logistic 1. $D2$ and $D3$ are referred to as the converting number of x_n for Logistic 2 and Logistic 3, respectively. Each Logistic map is computed with different parameters which include the initial value of x_0, the control parameter c and the number of the initial iteration. Notice that

$$\left\{ \begin{array}{l} x_0^1 \neq x_0^2 \neq x_0^3 \\ c_1 \neq c_2 \neq c_3 \\ iniIter_1 \neq iniIter_2 \neq iniIter_3 \end{array} \right\}.$$

Different parameters of three Logistic maps make the orbit change more complex, so the influence of stable point orbit can be avoided.

A new integer number D consists of $D1, D2$ and $D3$. The $ModNum$ means the number of modular operation. An output number is named as $OutNum$ which is converted by the number of D in Equation (2) as follows:

$$OutNum = ASCIICode(\ D\ mod\ ModNum) \qquad (2)$$

where $ASCIICode$ means the function which is converted from the number of D to the digit, uppercase or lowercase letter in ASCII. Figure 2 plots the picture with validation code and interference image.

In Figure 2, the function of random assignment performs randomly to assign a number in the graph. The random assignment of $\{N1, N2, N3, N4\}$ makes the combination of numbers in the graph with more and larger period, and avoids the short period problem of Logistic map. The position of $\{N1, N2, N3, N4\}$ can be arranged as following:

$$\left\{ \begin{array}{llllll} 1: & N1 & N2 & N3 & N4 \\ 2: & N2 & N1 & N3 & N4 \\ 3: & N3 & N1 & N2 & N4 \\ & \cdots & \cdots & \cdots & \cdots \\ i: & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots \\ n: & N4 & N1 & N2 & N3 \end{array} \right\}$$

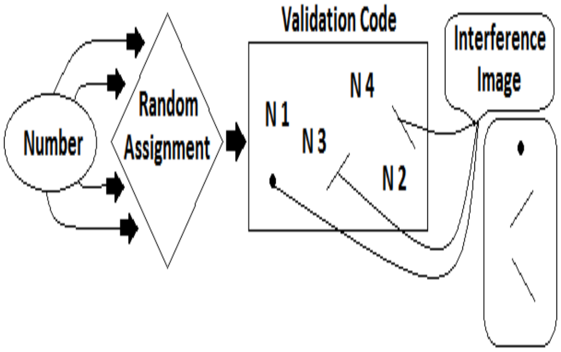


Figure 2: The graph with the validation code and interference image

Obviously, the method of the multi-dimensional Logistic map and the random assignment can be further optimized in comparison with one-dimension Logistic map. The arrangement and combination of the output number also can enhance sequence period.

After the number of $N1, N2, N3$ and $N4$ is mapped to the corresponding positions of the graph, the interference image will be added to the graph with the validation code.

The interference pattern is mainly used to realize interference in the validation code picture. If the machine is used to identify or attack, the interference image can increase the difficulty of identifying numbers or letters. Figure 3 shows the real validation code picture.

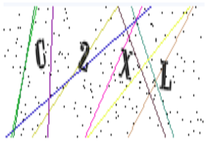


Figure 3: The real validation code graph in the real website

It is a rendering of the output with the validation code system used in the real website. It is clear that those points do not affect the user's recognition of letters and numbers in the validation code graphics.

4 Analysis

In this section we analyze the performance of the proposed method.

4.1 Key Space Analysis

The key space of Logistic map is determined by the accuracy of computers. In C language, single precision floating-point numbers occupy 4 bytes and 32 bits of the memory space. Double precision takes up 8 bytes of 64 bit of the memory space. The floating-point type of Double and Float can be converted to other types of integers.

Figure 4 plots the chaotic sequence with the floating-point type of Double and Float.

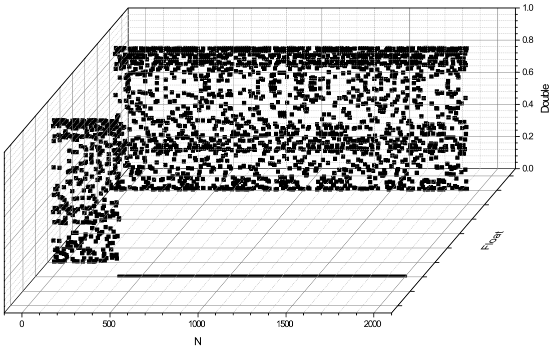


Figure 4: The chaotic sequence with the floating-point type values of Double and Float

Under the same initial value of parameters, we compare the two types of sequence key spaces of Double and Float. In Figure 4, the initial value of x_0 and the control parameter are set to 0.117 and 3.9, respectively. When the floating-point type of Float is used, the sequence of x_n happens cycle after 500 iteration steps. Clearly, in case of the floating-point type of Double the sequence does not happen cycle in spite of 2000 iteration steps. As mentioned above, when the random number space is large, the key space of Logistic map is large.

4.2 Sensitivity Test

The initial value of x_0 is set to 0.1 and 0.100001 for Logistic map. The control parameter is set to 3.9. Figure 5 plots the orbit of Logistic map with $\{x_0 = 0.1, c = 3.9\}$ and $\{x_0 = 0.100001, c = 3.9\}$.

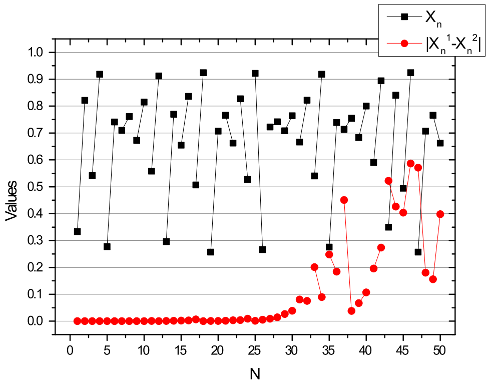


Figure 5: The orbit of Logistic map with $\{x_0 = 0.1, c = 3.9\}$ and $\{x_0 = 0.100001, c = 3.9\}$

The sequence of x_n means the sequence of Logistic map with $x_0 = 0.1$ and $c = 3.9$. The sequence of $|x_n^1 - x_n^2|$

means the absolute value of the difference for the sequence of $\{x_0 = 0.1, c = 3.9\}$ and $\{x_0 = 0.100001, c = 3.9\}$.

From Figure 5, we can see that in the first 30 iteration steps the absolute difference values are very close. In order to overcome this weakness of Logistic map the initial iteration number is given. Figure 6 plots the orbit of Logistic map with the initial iteration step of 100.

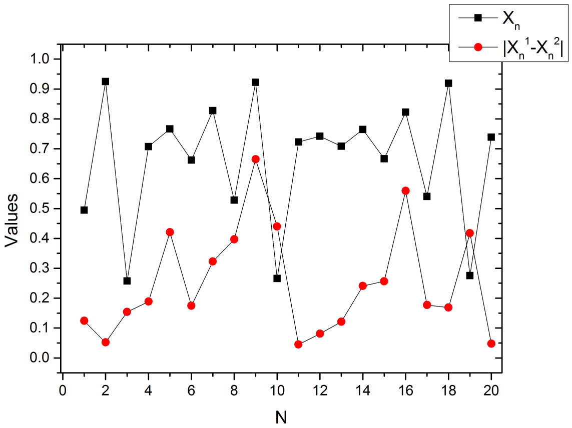


Figure 6: The orbit of Logistic map with initial iteration steps of 100

From Figure 6, after the first 100 iteration steps the absolute difference values are very different.

4.3 Statistics

We count the validation code in the form of numbers, uppercase letters and lowercase letters in three sample spaces with the value of 500, 2000 and 5000. We give a test as follows.

The output number of $OutNum$ in Equation (2) ranges from 0 to 84, which is converted to a digit. $OutNum$ in Equation (2) ranges from 85 to 170, which is converted to uppercase letter. $OutNum$ in Equation (2) ranges from 171 to 255, which is converted to lowercase letter. Figure 7, 8, and 9 shows the distribution of the digit, uppercase and lowercase letter with the proposed method. Figure 7 shows the distribution for the sample space of 500.

When the sample space is 500, the proportion of lowercase letters is the largest, about 41%. Figure 8 and 9 show the distribution for the sample space of 2000 and 5000, respectively.

With the increasing of sample space the proportion of lowercase letters is gradually decreasing and close to the average probability.

The proportion of uppercase letters is the smallest for the sample space of 500, about 28%. However, when the sample space enhances, the proportion of uppercase letters is gradually increasing and close to the average probability. In addition, the proportion of numbers is gradually

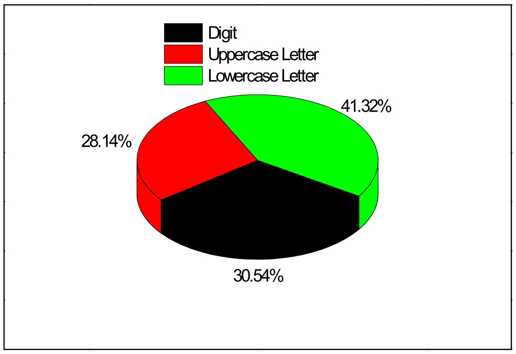


Figure 7: The distribution for the sample space of 500

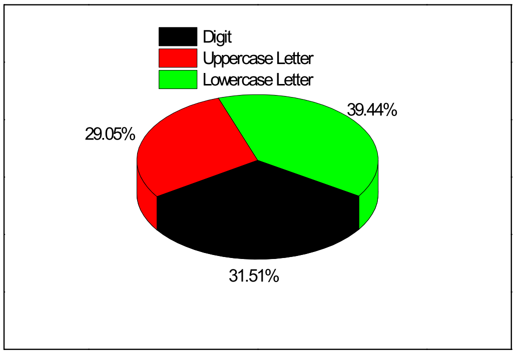


Figure 8: The distribution for the sample space of 2000

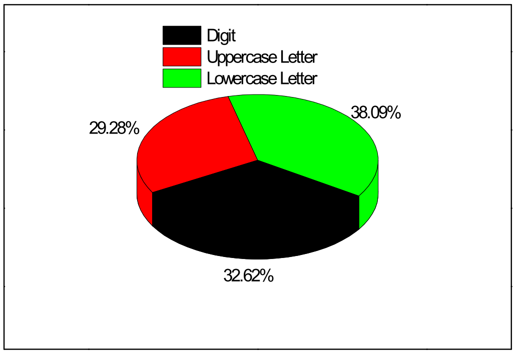


Figure 9: The distribution for the sample space of 5000

increasing and close to the average probability with the increasing of sample space.

5 Conclusions

In this paper, we generate the validation code through multi-dimensional Logistic map, effectively avoid the problem of the short period and the fixed point of the chaotic map. In addition, the period of the validation code obtains optimal when using the initial iteration number and the random mapping position of the graph. Random assignment can effectively increase the period of the output chaotic sequence and avoid the stable point, making the random sequence distribution more uniform. It can effectively resist the attacks of hackers on the validation code system and avoid the threat of websites.

Acknowledgments

This study was supported by the Basic Scientific Research Foundation and Scientific Research Project of Provincial Colleges and Universities in Heilongjiang (2018-KYYWFMY- 0093). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] S. A. Alsuhibany and A. A. Alnoshan, "Interactive handwritten and text-based handwritten arabic captcha schemes for mobile devices: A comparative study," *IEEE Access*, vol. 9, pp. 140991–141001, 2021.

[2] J. Chen, X. Luo, and Y. Liu *et al.*, "Selective learning confusion class for text-based captcha recognition," *IEEE Access*, vol. 7, pp. 22246–22259, 2019.

[3] R. Datta, J. Li, and J. Z. Wang, "Exploiting the human-machine gap in image recognition for designing captchas," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 504–518, 2009.

[4] H. Gao, M. Tang, and Y. Liu *et al.*, "Research on the security of microsoft's two-layer captcha," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1671–1685, 2017.

[5] S. Gao, M. Mohamed, and N. Saxena *et al.*, "Emerging-image motion captchas: Vulnerabilities of existing designs, and countermeasures," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 6, pp. 1040–1053, 2019.

[6] M. Garcia-Bosque, A. Pérez-Resa, and C. Sánchez-Azqueta *et al.*, "Chaos-based bitwise dynamical pseudorandom number generator on fpga," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 1, pp. 291–293, 2019.

[7] D. Hitaj, B. Hitaj, and S. Jajodia *et al.*, "Capture the bot: Using adversarial examples to improve captcha robustness to bot attacks," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 104–112, 2021.

[8] L. C. Huang, M. S. Hwang, and L. Y. Tseng, "Reversible data hiding for medical images in cloud computing environments based on chaotic Henon map," *Journal of Electronic Science and Technology*, vol. 11, no. 2, pp. 230–236, 2013.

[9] A. Kolupaev and J. Ogijenko, "Captchas: Humans vs. bots," *IEEE Security and Privacy*, vol. 6, no. 1, pp. 68–70, 2008.

[10] S. Kwon and S. Cha, "A paradigm shift for the captcha race: Adding uncertainty to the process," *IEEE Software*, vol. 33, no. 6, pp. 80–85, 2016.

[11] C. Li, Y. Chen, and T. Chang *et al.*, "Period extension and randomness enhancement using high-throughput reseeding-mixing prng," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 20, no. 2, pp. 385–389, 2012.

[12] C. Li, K. Tan, and B. Feng *et al.*, "The graph structure of the generalized discrete arnold's cat map," *IEEE Transactions on Computers*, vol. 71, no. 2, pp. 364–377, 2022.

[13] T. V. Nguyen, Z. Huang, and S. Bethini *et al.*, "Secure captchas via object segment collages," *IEEE Access*, vol. 8, pp. 84230–84238, 2020.

[14] C. Pope and K. Kaur, "Is it human or computer? defending e-commerce with captchas," *IT Professional*, vol. 7, no. 2, pp. 43–49, 2005.

[15] C. E. C. Souza, D. P. B. Chaves, and C. Pimentel, "Digital communication systems based on three-dimensional chaotic attractors," *IEEE Access*, vol. 7, pp. 10523–10532, 2019.

[16] M. Tang, H. Gao, and Y. Zhang *et al.*, "Research on deep learning techniques in breaking text-based captchas and designing image-based captcha," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2522–2537, 2018.

[17] P. Wang, H. Gao, and Q. Rao *et al.*, "A security analysis of captchas with large character sets," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 2953–2968, 2021.

[18] H. Weng, B. Zhao, and S. Ji *et al.*, "Towards understanding the security of modern image captchas and underground captcha-solving services," *Big Data Mining and Analytics*, vol. 2, no. 2, pp. 118–144, 2019.

[19] Y. Xu, G. Reynaga, and S. Chiasson *et al.*, "Security analysis and related usability of motion-based captchas: Decoding codewords in motion," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 5, pp. 480–493, 2014.

[20] J. Yan and A. S. ElAhmad, "Captcha security: A case study," *IEEE Security and Privacy*, vol. 7, no. 4, pp. 22–28, 2009.

[21] A. Ynnerman, S. C. Chapman, and P. Ljung *et al.*, "Bifurcation to chaos in charged particle orbits in a magnetic reversal with shear field," *IEEE Transactions on Plasma Science*, vol. 30, no. 1, pp. 18–19, 2002.

[22] H. Yu, S. Xiao, and Z. Yu *et al.*, "Imcaptcha: Imperceptible captcha based on cursor trajectories," *IEEE Consumer Electronics Magazine*, vol. 9, no. 1, pp. 74–82, 2020.

[23] B. B. Zhu, J. Yan, and G. Bao *et al.*, "Captcha as graphical passwords - a new security primitive based on hard ai problems," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 891–904, 2014.

[24] Y. Zi, H. Gao, and Z. Cheng *et al.*, "An end-to-end attack on text captchas," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 753–766, 2020.

Biography

Dahua Song is an associate professor at the computer science from Mudanjiang Medical University, China. Her main research interests include information security and chaos theory.

Chang Liu is an associate professor at the computer science from Mudanjiang Medical University, China. Her main research interests include information security.

Jiahui Liu is a professor at the computer science from Harbin University of Science and Technology, China. His main research interests include information security, computer software and network security.

Secure Search over Encrypted Enterprise Data in the Cloud

Kaishi Wang[1,2] and Jiaqi Guo[3]

(Corresponding author: Jiaqi Guo)

School of Economics and Management, Northwest University, Xi'an 710069, China[1]

Helios Power Corporation, Xi'an 710075, China[2]

School of Computer Science and Technology, Xidian University, Xi'an 710071, China[3]

Email: jqguo_echo@stu.xidian.edu.com

(Received July 10, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 13, 2022)

Abstract

With the rise of cloud computing, more and more enterprise data are stored in cloud servers to save local resources and facilitate access. To protect sensitive data, enterprises can encrypt the data before outsourcing. However, encrypted data is complex for others to query and access. This paper proposes a Secure enterprise Data Search (SEDS) scheme to search encrypted data. The SEDS scheme considers a use case that the data belonging to multiple sub-enterprises in a group enterprise. The scheme is based on symmetric searchable encryption and can efficiently search the data from multiple sub-enterprises by generating a trapdoor through a combined key. Meanwhile, the cloud server can search and return several most relevant data to save transmission bandwidth. The SEDS scheme can be securely and effectively applied to search enterprise data through security analysis and performance tests on real data sets.

Keywords: Cloud Storage; Enterprise Data; Multi-Keyword; Ranked Search; Secure Search

1 Introduction

The past decade witnessed the rapid development of cloud computing [7], accelerating the business data continue to migrate to the cloud. Due to the high performance, vast storage, low maintenance costs, and other advantages, it has been widely concerned by the industry. Increasing numbers of enterprises choose to store the data in the cloud server, which can reduce the cost of local storage and maintenance, access the data anytime and anywhere, or share data for teamwork. However, data stored in the cloud server is not directly controlled by the enterprise, so it faces the risk of leaking to more adversaries on the Internet [4, 15, 22]. If some sensitive enterprise data is leaked, such as equipment quotations, transaction orders, or technology patents, it will cause huge losses to the enterprise. Although the cloud service providers build many security measures to protect the data stored on them, they are also curious about the data may be for greater commercial interests or other reasons [12,13]. Therefore, when storing sensitive data, enterprises encrypt it locally before uploading it [8]. However, ciphertext makes it difficult to use data, such as data search. The intuitive approach is downloading all the data locally and decrypting it before searching, but this is obviously unreasonable. Therefore, the secure search of encrypted data is the main obstacle for enterprises to employ cloud computing, which also has become the focus of academia and industry [9].

To solve the problem of secure search on encrypted data, a new cryptography primitive, Searchable Encryption (SE), is proposed. It can perform searches on encrypted data without revealing the search content. The content of the file is divided into blocks and encrypted directly, and the full text should be scanned and matched during the search, which caused a large computational overhead and is not suitable for searching massive data. To improve the search efficiency, Goh *et al.* [5] propose an index-based scheme, which employs an index called Z-IDX using the Bloom Filter. It simply searches the index to retrieve the desired data directly. However, the probability of false positives of Bloom Filter may return the wrong search results. Besides the above symmetric searchable encryption schemes, the public key encryption with keyword search (PEKS) [**?**, 2] schemes are proposed, but they contain a large number of bilinear pairing or modular exponentiation operations, which brings huge overhead so that it is not in reality. The above schemes are all searched by a single keyword, but in actual application, it is often searched by multiple keywords or even a text paragraph. In addition, returning all the search results usually contains a large amount of data, which consumes the transmission bandwidth. A useful feature of plaintext search, multi-keyword ranked search, is also concerned in ciphertext search. The first secure search scheme to support both multi-keyword and results ranking is proposed by Cao *et al.* In this scheme, the files and queries are both represented as binary vectors, and calculates the similarity by inner product similarity, and protect the privacy of vectors by the technique of secure k-Nearest Neighbor (kNN). However, it does not consider the weight of keywords. For higher search precision, Sun *et al.* [18] propose a scheme based on the TF×IDF model, which introduces the vector space model in plaintext search into ciphertext search. To improve the search efficiency, a KBB-tree-based index is employed in [21]. In recent years, a large number of searchable encryption schemes have been proposed to achieve rich features, such as access control [16,23], result verification [11,14], dynamic update [1,6,17] and so on.

However, it is difficult to directly apply the above schemes to search actual enterprise data. For example, in a enterprise group composed of multiple sub-enterprises, each sub-enterprise encrypts the data with its own key and generates a secure index. They store the data and the secure indexes in the cloud server but would like to share some data with other authorized sub-enterprises. It becomes a problem that searches the data from multiple sub-enterprises through a trapdoor and ranks the search results encrypted with different keys by the cloud server. Therefore, aiming at the problem, this paper proposes a Secure Enterprise Data Search (SEDS) scheme. In the SEDS scheme, each sub-enterprise chooses a secret key to encrypt the index, and the index is encrypted by the secure kNN technique. When searching, the user generates a trapdoor with a combined key to search the data from multiple sub-enterprises. In addition, the cloud server searches on indexes of each sub-enterprise through the indication matrices and sorts all the results. At the same time, considering the strategic deployment of the group enterprise, such as the establishment, bankruptcy, or merger of subsidiaries, the dynamic update approach is presented. Through security analysis and performance tests on real data, the SEDS scheme can securely and efficiently search the data from the enterprise group. The main contributions are summarized as follows.

- We study the problem of data security search in large group enterprises and propose a SEDS scheme, which can perform a multi-keyword ranked search for data encrypted with different keys.
- For the dynamic deployment of group enterprises, a dynamic update approach is presented.
- We analyze the security of SEDS and conduct experiments on a real-world data set to confirm the efficiency of the scheme.

The structure of this paper is arranged as follows. The system model, the threat model, the design goals, the notations, and preliminaries are all given in Section 2. The details of the SEDS scheme are presented in Section 3. Then, Sections 4 and 5 analyze the security and performance of the scheme, respectively. Finally, the conclusion and prospects are discussed in Section 6.

2 Problem Formulation

2.1 Notations

First, the basic notations used in this paper is given in Table 1.

Table 1: Notations

Notations	Descriptions
o	A sub-enterprise that acts as a data owner
u	A sub-enterprise that acts as a data user
n	The number of keywords in the dictionary n' represents the number of redundant keywords.
m	The number of sub-enterprises in the conglomerate. m' represents the number of redundant sub-enterprises
F	The plaintext file set
C	The ciphertext file set
I	An $n + n'$-dimension index vector
I_s	The encrypted form of I
T	The plaintext index tree
T_s	The index tree with the encrypted index vector
Q	The query vector build from keywords of interest
TD	The encrypted form of Q
M_{com}	The combined matrix
M_{ind}	The indicator matrix

2.2 System Model

There are three parties in the system, namely the data owner, data user, and cloud server. We assume that there are multiple sub-enterprises in a group, each sub-enterprise encrypts files with its own key and uploads them to a cloud server for storage, and other authorized sub-enterprises can search for files of interest in the system. The system model of our secure search scheme is illustrated in Figure 1.

- *Sub-enterprises(data owners):* The sub-enterprises acted as data owners outsource the data to the cloud server. They encrypt their sensitive files and generate secure indexes. All the sensitive files and indexes are outsourced to the cloud server in encrypted form.
- *Sub-enterprises(data users):* The sub-enterprises act as data users query data from the cloud server. A sub-enterprise can be either a data owner or a data user. It encrypts the query vector containing several keywords of interest to generate a trapdoor. The key for encrypting the query is generated by the keys from data owners who authorize the data user. The

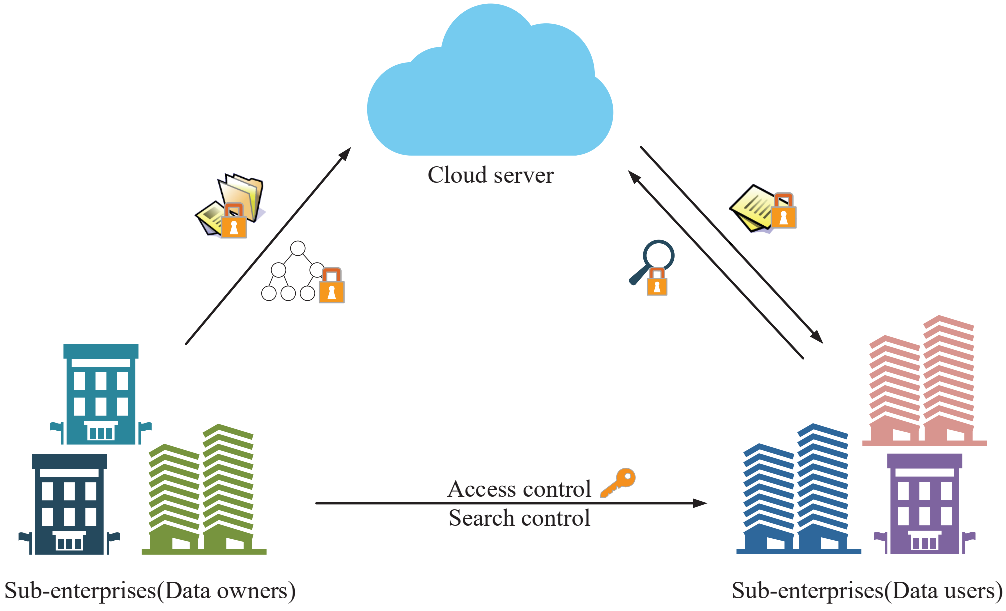


Figure 1: System model

Data user sends the trapdoor and a constant k to the cloud server for the top-k related files.

- *Cloud server:* When receiving the trapdoor from a data user, the cloud server searches over the file collections of each data owner, and returns the top-k related files to the data user.

2.3 Threat Model

Enterprise data contain much sensitive information. The cloud server has some security measures to protect against external attackers, but it may also gather and analyze the information for higher interests. Therefore, we assume that the cloud server in our scheme is ”honest but curious” [3, 21], that is, the cloud server honestly executes our proposed scheme, but it is curious about the potential information of all the data stored on it. We choose a known ciphertext model, the cloud server can learn the encrypted files and secure index tree of each data owner, and the trapdoor of each data user.

2.4 Design Goals

In order to securely and efficiently search the encrypted data of multiple sub-enterprises stored on the cloud server, the design goals of this solution are as follows:

- *Search efficiency:* The scheme can efficiently search the encrypted data of multiple sub-enterprises by one trapdoor, and neither the trapdoor generation time nor the search time will change with the number of queried keywords.
- *Ranked search:* Although the files are from multiple sub-enterprises and each sub-enterprise generates the secure index using a different secret key, the cloud server can search and return the top-k related documents.
- *Privacy goals:* The scheme protects the privacy of each sub-enterprise, so that the cloud server cannot obtain information about files, indexes, and queries. Specifically, this involves the follows:
  - *Data confidentiality*: The cloud server cannot obtain the plaintext information in the files.
  - *Index confidentiality*: The cloud server cannot obtain the plaintext keyword in the index and the plaintext TF value in the index vector.
  - *Query confidentiality and unlinkability*: The cloud server cannot obtain the plaintext keywords in the query and the plaintext IDF value in the query vector, nor can it determine whether the two trapdoors are generated by the same query keyword.

2.5 Preliminaries

2.5.1 Vector Space Model

Vector space model [19] is one of the significant mathematics tools in secure multi-keyword search schemes [3, 18, 21]. Specifically, each document can be represented as an index vector whose length equals to the size of the dictionary, each element in an index vector represents a normalized TF value of the keyword in this document. Also, the query can be expressed as a vector called trapdoor whose dimension is equal to the index vector's, where the elements in the location of query keywords represent their normalized IDF values. When calculating the inner product of an index vector and a trapdoor, we can learn the relevance score of the document and the query, and obtain the query results by sorting all the relevance score finally.

we can calculate the relevance score as:

$$Score(I,Q) = I \cdot Q = \sum_{w \in W_q \cap W} TF_w \times IDF_w$$

where w is a keyword, TF_w is the normalized TF value of w in I, IDF_w is the normalized IDF value of w in the document collection. When v is a index vector of the document, $TF_w = (1 + lnN_{f,w})/\sqrt{\sum_{w\in W}(1+lnN_{f,w})^2}$, where $N_{f,w}$ is the number of occurrences of keyword w in document f. The IDF_w is calculated as $ln(1+N/N_w)/\sqrt{\sum_{w\in W_q}(ln(1+N/N_w))^2}$, where N is the total number of documents and N_w is the number of documents containing w.

2.5.2 Keyword Balanced Binary (KBB) Tree

KBB-tree [21] is a dynamic data structure for multi-keyword ranked search. Each node v of the KBB-tree stores a tuple shown as follows:

$$v = \langle ID, I, PT_l, PT_r, FID\rangle$$

ID is the node identity, and FID is the file identity. PT_l and PT_r are two pointers to the left child and right child of v. When constructing a KBB-tree, I in a leaf node is a vector with normalized TF values, while I in internal nodes are generated by its left and right child nodes as follows:

$$I[i] = max\{i.PT_l \to I[i], j.PT_r \to I[i]\}, i = [1, 2, ..., n].$$

For example, in Figure 2, there are 4 files presented as $\{f_1, f_2, f_3, f_4\}$ with different TF values and $n = 3$ keywords in the dictionary. The query vector is set to $Q = (0.7, 0, 0.9)$, and the process for searching $top - 2$ files is shown in Figure 2. In our scheme, sub-enterprises build up and encrypt KBB-trees separately.

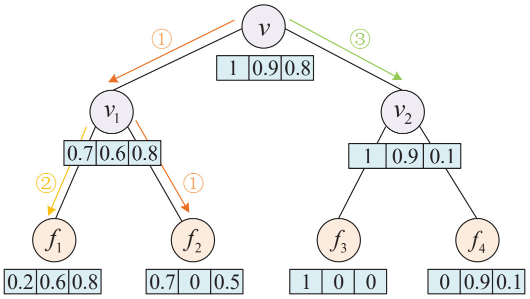


Figure 2: An example of searching by the KBB-tree based index

3 Secure Enterprise Data Search Scheme

In this scheme, a enterprise group model composed of several sub-enterprises is considered, in which secure multi-keyword ranked search can be realized. In this section, we first present the framework of the scheme and then describe each algorithm in detail.

3.1 The Framework of SEDS

- $\texttt{Setup}(1^\lambda) \to V$: It takes in the security number λ to generate a random binary vector.
- $\texttt{KeyGen}(1^\lambda) \to \{SK_o, sk_o\}$: Each sub-enterprise generates its secret key SK_o for encrypting indexes and sk_o for encrypting files.
- $\texttt{IndexGen}(I, SK_o) \to I_s$: The sub-enterprise o takes in the plaintext index vector I and the secret key SK_o to generate the secure index vector I_s.
- $\texttt{FileEnc}(F_o, sk_o) \to C_o$: The ciphertext set of the sub-enterprise o is encrypted by the secret key sk_o.
- $\texttt{TrapGen}(Q, SK_u) \to TD$: The sub-enterprise u encrypts its query vector Q by the SK to generate a trapdoor TD for search.
- $\texttt{Search}(T_s, k, TD) \to SList$: The cloud server takes in each T_s and the trapdoor to search the top-k related files.
- $\texttt{FileDec}(C_r, sk_o) \to F_r$: The results in encrypted form can be decrypted by the secret key sk_o of each sub-enterprises.

The specific file encryption algorithm is not considered, because the existing symmetric key cryptography such as AES can be easily employed. However, to describe the scheme completely, FileEnc and FileDec is still presented in the framework.

3.2 SEDS Scheme

The algorithm is described in detail as follows.

- Setup.

The system vector is generated as $\texttt{Setup}(1^\lambda)\ rightrow V$. V is a random binary vector of $n+n'$-bit, where n is equal to the size of the dictionary and n' is some redundant bits. The system vector is available to each sub-enterprise in the system.

- KeyGen.

Any sub-enterprise o that needs to outsource data needs to generate a query key SK_o and file key sk_o.

Each sub-enterprise generates two $(n + n') \times (n + n')$ invertible matrices $\{M_1, M_2\}$ as its query key. Then, each sub-enterprise chooses a random number r and computes

$SK'_o = \{rM_1, rM_2\}$ and sends SK'_o to other authorized sub-enterprises through a secure channel. In addition, each child enterprise also generates an indicator matrix $M_{ind} = [O|\cdots|(1/r)E|\cdots|O]$, where E is a unit matrix and O is a zero matrix both with size $(n+n') \times (n+n')$. There are only one unit matrix and $m-1$ zero matrices, and m is the number of enterprises in the group.

When receiving the SK_o of other sub-enterprises, as a data user, the sub-enterprise u combines the query keys to generate $SK_u = \{M_{com1} = [rM_1^{-1}|rM_1^{-1}|\cdots|rM_1^{-1}], M_{com2}2 = [rM_2^{-1}|rM_2^{-1}|\cdots|rM_2^{-1}]\}$. M_{com1} and M_{com2} are two combined matrices of m matrices of each sub-enterprise, each rM_1 or rM_2 is different and chosen by different sub-enterprise. If u is not authorized by a sub-enterprise, then u disposes of the matrix that of that sub-enterprise as a zero matrix. For example, there are 5 sub-enterprises in the group, u with number 3 obtains three keys from the sub-enterprises with number $\{1,2,5\}$, SK_u is shown as follows:

$$SK_u = \left\{ \begin{array}{l} M_{com1} = [rM_1^{-1}|rM_1^{-1}|rM_1^{-1}|O|rM_1^{-1}], \\ M_{com2} = [rM_2^{-1}|rM_2^{-1}|rM_2^{-1}|O|rM_2^{-1}] \end{array} \right\}$$

Besides the query key, the sub-enterprise generates a file key sk_o for encrypting its files. For details, refer to the symmetric encryption algorithm used, such as AES. For other sub-enterprises, o authenticates its identity and sends the file key to the authenticated sub-enterprises through a secure channel.

Finally, each o keeps $\{SK_o, sk_o\}$ as secret to encrypt the index and files, while u keeps SK_u for trapdoor generation.

- `IndexEnc.`

Each sub-enterprises constructs an index tree and encrypts it by the query key SK_o. First, each sub-enterprise builds an index tree from its file set. Each leaf node is a $n+n'$-dimensional vector, the first n elements of which are the normalized TF value of the keyword, and the last n' are set to 0. According to the leaf nodes, the index tree is constructed from bottom to top. Then, the sub-enterprise o encrypts the index vector of each node, including leaf nodes, intermediate nodes, and root nodes in the index tree. Each node stores a vector I, which is split into two random vectors I', I'' according to the system vector S, the splitting process is as follows.

$$\left\{ \begin{array}{ll} I'[i] = I''[i] = I[i] & \text{if } V[i] = 0, \\ I'[i] + I''[i] = I[i] & \text{if } V[i] = 1. \end{array} \right.$$

Note that $i \in \{1, \cdots, n+n'\}$ and $V[i]$ represents the i-th elements in V. Specifically, if $V[i]$ is 0, $I'[i]$ and $I''[i]$ are set as the same value as $I[i]$; if $V[i]$ is 1, $I'[i]$ and $I''[i]$ can be set as random values while the sum of them should be equal to I.

Then the sub-enterprise o encodes the two split vectors of each node by SK_o as $I_s = \{M_1^T I', M_2^T I''\}$. The secure index tree T_s is constructed.

- `FileEnc.`

Before outsourcing, each sub-enterprise encrypts its file set by a file key sk_o. Then, each sub-enterprise outsources the encrypted file set with the secure index tree T_s to the cloud server.

- `TrapGen.`

When a sub-enterprise u being interested to some content, u generates a trapdoor with several interested keywords.

First, u generates a query vector Q, which is an $n+n'$-dimensional vector. If the keyword is of interest, the corresponding element in Q is set to the normalized IDF value; otherwise, the element is set to 0. Then, similar to index splitting, Q is split into two random vectors $\{Q', Q''\}$. The difference is that if $V[i]$ is 0, $Q'[i]$ and $Q''[i]$ is set to two random numbers whose sum equals $Q[i]$; if $V[i]$ is 1, $Q'[i]$ and $Q'[i]$ are set to $Q[i]$.

$$\left\{ \begin{array}{ll} Q'[i] + Q''[i] = Q[i] & \text{if } V[i] = 0, \\ Q'[i] = Q''[i] = Q[i] & \text{if } V[i] = 1. \end{array} \right.$$

Then, the trapdoor is generated by multiplying the combined matrices in SK_u, i.e. $TD = \{M_{com1}Q', X_{com2}Q''\}$.

Finally, the sub-enterprise u sends the trapdoor TD_u to the cloud with a parameter k to obtain the top-k related files.

- `Search.`

When receiving a trapdoor from a sub-enterprises u, the cloud server searches each secure index tree for the results.

In the search process, greedy depth-first search (GDFS) is used to search iterated on the secure index tree T_s of each sub-enterprise. We present the search process as Algorithm 1. The cloud server does not need to consider whether the sub-enterprise is authorized or unauthorized, and only searches the secure index tree of each sub-enterprise in turn. When searching, a k-element list is always maintained. Each element in the list is a tuple containing a file ID and the relevance score of the file and TD, represented as $\langle FID, Score(I_{sv}, TD)\rangle$. v represents the node in the secure index tree. In the beginning, the score is set to a number slightly greater than 0 to prevent unauthorized data from being included in the results. The elements in the $SList$ are in descending order of score, so the score in the $k-th$ element is the smallest. $v.hchild$ and $v.lchild$ are the two children of v, where $v.hchild$ has a higher score and $v.lchild$ has a lower score. The relevance score of a node in T_s and a trapdoor TD is

Algorithm 1 Search

```
Input: the secure index trees $T_s$(the number of $T_s$ is $m$), $k$, the indicator matrices $M_{ind}$, the trapdoor $TD$
Output: the list of search results $SList$
for ($i = 1 \rightarrow m$) do
    Search($T_s$, $TD$, $M_{ind}$, $k$)
end for
Search(IndexTreeNode $v$, Trapdoor $TD$, int $k$)
if the node $v$ is not a leaf node then
    if Score ($I_v$, $TD$) >k-thScore then
        Search($v.hchild$, $TD$, $k$);
        Search($v.lchild$, $TD$, $k$);
    else
        return
    end if
else
    if Score ($I_v$, $TD$) >k-thScore then
        Delete the last element in $SList$ and insert the new one;
        Sort all the elements of $SList$;
    else
        return
    end if
end if
```

calculated as follows.

$$
\begin{aligned}
&\mathrm{Score}(I_{sv}, TD) = I_{sv}\cdot(M_{ind}TD)\\
&= (M_1^T I'_v)\cdot(M_{ind}M_{com1}^T Q') + (M_2^T I''_v)\cdot(M_{ind}M_{com2}^T Q'')\\
&= (M_1^T I'_v)\cdot(M_1^{-1}Q') + (M_2^T I''_v)\cdot(M_2^{-1}Q'')\\
&= {I'_v}^T M_1 M_1^{-1} Q' + {I''_v}^T M_2 M_2^{-1} Q''\\
&= {I'_v}^T Q' + {I''_v}^T Q'\\
&= I_v \cdot Q\\
&= \mathrm{Score}(I_v, Q)
\end{aligned}
$$

From the above equation, we can see that the score of the node in the secure index tree and the trapdoor of u is equal to the score of the plaintext index vector and the plaintext query vector, so the query correctness is guaranteed.

In a secure index tree, each element of the vector stored by the root node is the largest in the tree. This is very efficient when searching index trees of multiple sub-enterprises because if the relevance score of the root node and trapdoor in an index tree is less than the k-th score of the current result list, the other nodes of the tree are not traversed. If u is not authorized by the sub-enterprise, when searching the index tree of the sub-enterprise, $\{M_{ind}TD = M_{ind}M_{com1}^T Q', M_{ind}M_{com2}^T Q''\} = \vec{0}$, which means that the query vector is zero vector. Therefore, the relevance score of the files belonging to the sub-enterprise and the trapdoor are all equal to 0. The search results returned by the cloud server do not contain any data of the sub-enterprise. After searching the index tree of all the sub-enterprises, the cloud server gets a final list of no more than K files. Eventually, the cloud server returns the files on the list to the sub-enterprise u in encrypted form.

- `FileEnc.`

When receiving the search results C_r in encrypted form, the sub-enterprise u decrypts them by the corresponding secret key sk_o.

3.3 Dynamic Update

A conglomerate may dynamically adjust enterprise-scale or architecture, such as the merger of sub-enterprises or the establishment of new sub-enterprises. Due to dynamic adjustment, the secret keys of some sub-enterprises will change. If the sub-enterprise goes bankrupt, its files stored on the cloud are deleted, and its serial number is no longer valid which can be given to a new sub-enterprise. If two sub-enterprises merge, one performs the same operation as bankruptcy and re-encrypts its files with the secret key of the other sub-enterprise and uploads them. If a new sub-enterprise is established, it can get a serial number and generate its secret key (both query key and file key) based on the system parameters. In the initial stage of system setup, the conglomerate size can be estimated, and the number of sub-enterprises can be set to $m + m'$ which is larger than the conglomerate size m. When a new sub-enterprise is established in the future, the idle serial number can be chosen by it. Since the number of sub-enterprises remains unchanged, the indicator matrices of other sub-enterprises and the size of SK_u also remain unchanged.

As presented in Section 3.2, several redundant bits are set $n + n'$ for dictionary updating. When a new file is added, if it contains a keyword that is not in the dictionary, the keyword can be added to the redundant bits. Files are deleted in lazy mode, which means the dictionary remains unchanged when a file deleting. After a long time of update, the whole system needs to be updated, including updating system parameters, system dictionary, and TF and IDF values of each keyword. The sub-enterprises generate the secret keys based on the new system parameters and re-encrypt the files to ensure security.

4 Security Analysis

In Section 2.4, we define the privacy goals in the SEDS scheme. Now we demonstrate the security of the SEDS scheme by analyzing it according to these goals. Note that SEDS is presented in the known ciphertext model, which is not secure enough in the known background model. However, some approaches [3, 18, 21] can be directly applied to the SEDS to meet the known model under the background of security.

- Data confidentiality

The symmetric encryption algorithm such as AES is used in SEDS to encrypt files. As long as the secret key sk_o is not leaked to the cloud server, the ciphertext will not be decrypted, and the data confidentiality is effectively protected in SEDS.

- Index confidentiality

In SEDS, the index vector of each file is split and obfuscated into two vectors for generating the secure indexes. Each sub-enterprise holds a different secret key SK_o, and only the authorized users can obtain the keys. As long as the secret key SK_o is kept confidentially, the cloud server cannot learn the original information of the index. In the known ciphertext model in [20], it is proved that the attacker cannot learn the keywords or the TF and IDF values in the indexes or queries from the search results. Therefore, index confidentiality is well guaranteed in SEDS.

- Query unlinkability and query confidentiality

Like generating secure indexes, generating a trapdoor is also a split and obfuscating of the query vector. In SEDS, the sub-enterprise that needs to search has a secret key SK_u, as long as the SK_u is not revealed, the cloud server cannot know the specific keywords in the query. The query confidentiality is well protected.

When generating a trapdoor, the query vector is first randomly divided into two vectors, so different trapdoors will be generated even if the same keyword is queried. Therefore, query unlinkability can be realized. However, it is also possible for the cloud server to link to the same query with the same search path and relevance score. For accuracy in actual searches, SEDS does not consider defending against such attacks. But there are approaches [3,18,21] to resist, and SEDS can be enhanced by introducing these approaches if needed.

5 Performance Analysis

The performance of the scheme is theoretically analyzed, and experiments and analyses are carried out on real data sets.

5.1 Theoretical Analysis

- `IndexEnc`

Each sub-enterprise constructs an index tree according to the files to be outsourced and encrypts it to generate a secure index tree. We assume that each sub-enterprise has f files, so there are $O(f)$ nodes to be encrypted. Each node of the tree stores an index vector of length $n+n'$. So, for each node, it takes $O(n+n')$ time to split the vector and $O((n+n')^2)$ time to multiply the $(n+n') \times (n+n')$ matrix. As a whole, the time complexity of encrypting the index tree of a sub-enterprise is $O(f(n+n')^2)$. Since n' is much less than n, we can also denote it as $O(fn^2)$.

- `TrapGen`

When generating the trapdoor, similar to index encryption, it takes $O(n+n')$ time to split the query vector. The size of the combined matrices in secret key SK_u used to encrypting the query is $((m+m')(n+n') \times (n+n'))$, so the time complexity of trapdoor generation is $O((m+m')(n+n')^2)$, which can be denoted as $O(mn^2)$.

- `Search`

When searching, we assume that there are m sub-enterprise in the system, and the number of index trees to search is also m. The number of leaf nodes that need to be traversed of a sub-enterprise is represented as n_v. When calculating the relevance score, the time complexity is $O(m^2(n+n'))$. According to the structure of the KBB-tree, the height of the tree is $logf$. Therefore, the whole time complexity is $O(m^2(n+n')n_v logf)$, and we also denote it as $O(m^2nn_v logf)$.

5.2 Experimental Analysis

To test the efficiency of SEDS, we implement it using the Python language on a real data set: the abstracts of articles in arXiv. arXiv is a free distribution service and an open-access archive for numerous scholarly articles. We choose 10000 articles from physics, mathematics, computer science, and other fields. First, we consider a special case that each sub-enterprise holds the same files. Without losing generality, we also consider that each sub-enterprise holds files from different fields and compare the time cost of the two cases. The experimental results of `TndexEnc` and `TrapGen` are obtained with an Intel Core i5-4200M CPU (2.50GHz) and 4GB memory, while the results of `Search` are obtained with an Intel Core i7-7700 CPU (3.6GHz) with 64GB memory to simulate a cloud server.

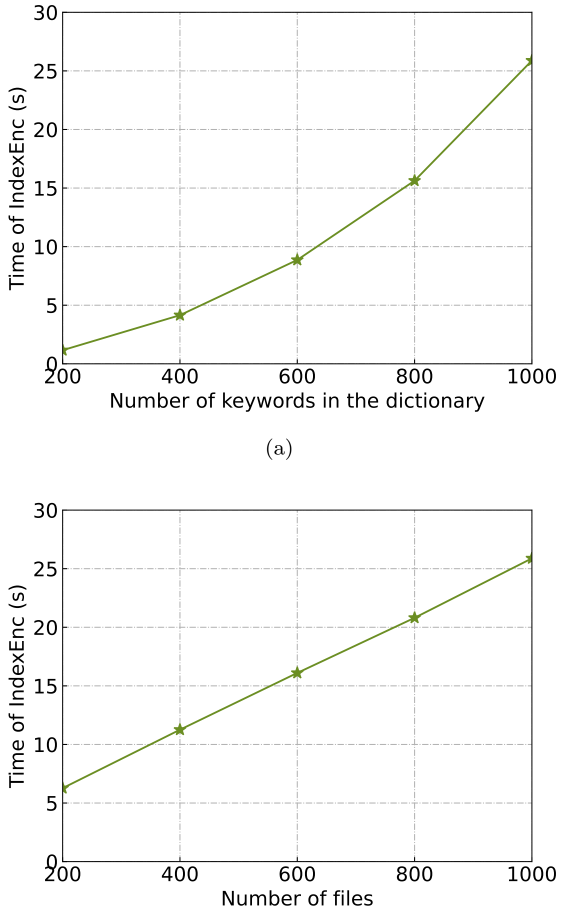
- `IndexGen`

Figure 3 shows the time cost of a sub-enterprise encrypting the index tree. In Figure 3(a), the number of files is fixed to 600, and the number of keywords in the dictionary is set from 200 to 1000. The redundant bits n' is set to $10\%n$, that is, if the dictionary size n=600, the index length is $n+n' = 660$. From Figure 3(a), it can be observed that the time cost for a sub-enterprise to generate the secure index tree is approximately proportional to the number of keywords in the dictionary.

In Figure 3(b), the number of keywords in the dictionary is fixed to 600, while the number of files varies from 200 to 1000. The time cost of encrypting the index tree is nearly linear with the number of files. Therefore, the time to run `IndexEnc` depends on the number of keywords in the dictionary and the number of files.

- `TrapGen`

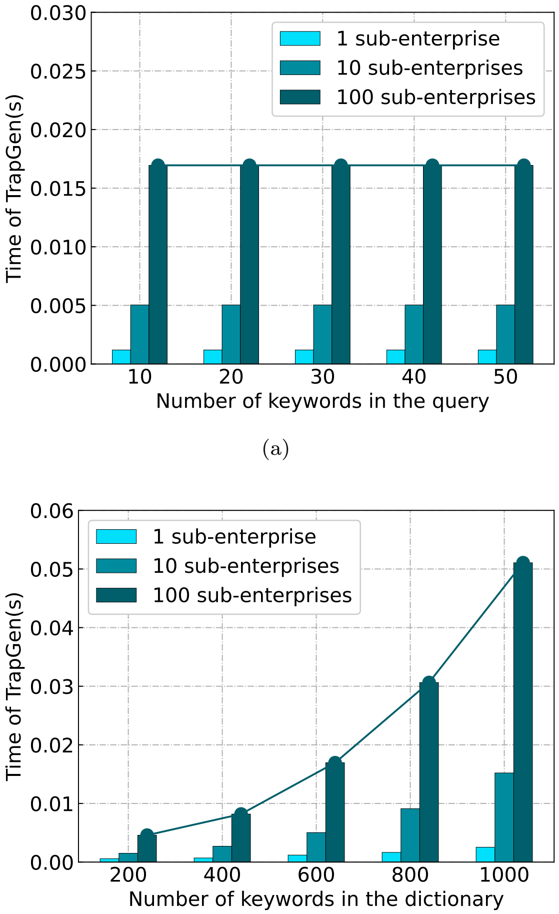
The time cost of generating a trapdoor by a sub-enterprise in the system is presented in Figure 4. The number of sub-enterprises in the group is set to 1, 10, 100, which can be



(a)

(b)

Figure 3: Time cost for `IndexEnc`



(a)

(b)

Figure 4: Time cost for `TrapGen`

applied to small enterprises without sub-enterprises, small group enterprises, and large-scale group enterprises. In Figure 4(a), the number of keywords in the dictionary is fixed to 600, and it can be observed that the number of keywords in the query has little influence on the time cost of generating a trapdoor, so that plenty of keywords can be searched.

In Figure 4(b), the number of keywords in the query is fixed to 30, the dictionary size is set to 200 to 1000. It indicates that the time of `TrapGen` grows with the number of keywords in the dictionary.

Both Figure 4(a) and 4(b) demonstrate that the more sub-enterprises in the group, the more time taken to generate the trapdoor. Because the size of the secret key SK_u used to generate the trapdoor is proportional to the number of sub-enterprises.

- `Search`

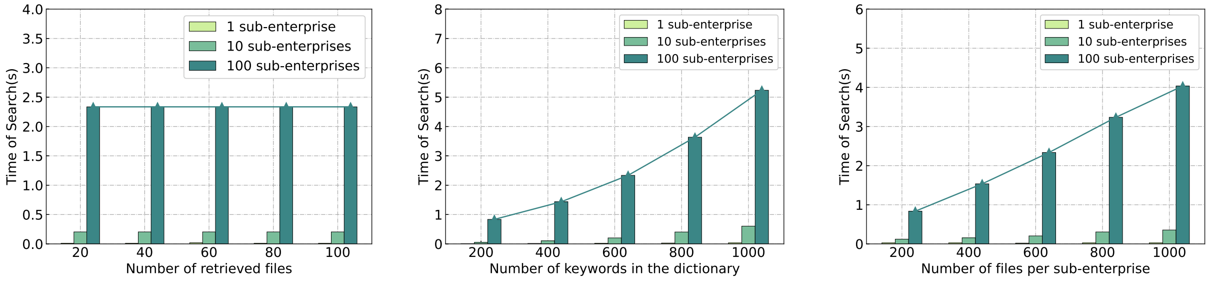
Figure 5 illustrates that the time to run `Search` varies with the number of keywords in the query, the number of keywords in the dictionary, and the number of files per sub-enterprise.

In Figure 5(a), the number of keywords in the dictionary is fixed to 600 and the number of files per sub-enterprise is fixed to 600. When the number of keywords varies from 10 to 50, the time of searching remains constant.

In Figure 5(b), the number k of files retrieved is fixed to 60 and the number of files per sub-enterprise is fixed to 600. It can be indicated that the time cost of `Search` grows with the dictionary size. The reason is that the length of the secure index vector and trapdoor both are proportional to the number of keywords in the dictionary.

In Figure 5(c), the number k of files retrieved is fixed to 60 and the dictionary size is fixed to 600. It can be observed that the time cost of searching is nearly linearly with the number of files per sub-enterprise, because more files mean more operations of calculating the inner product will be performed. The files of each sub-enterprise are set to be the same, and the sub-enterprise acts as the data user has the secret key of all the sub-enterprises, so the index tree of each sub-enterprise is traversed in the search process. In practice, the main business of each sub-enterprise is different, and the sub-enterprise as data user may has secret keys of several sub-enterprises, so the search time is much lower.

From both sub-figures of Figure 5, it can be observed that the time for searching grows with the number of sub-



(a) (b) (c)

Figure 5: Time cost for `search`

enterprises in the group. Because more sub-enterprises own more files, which means more operations to compute the inner product, and the length of the trapdoor is also related to the number of sub-enterprises. However, `Search` is implemented by the cloud server with huge computing power. So, it will take less time in practice and not be a burden on each sub-enterprise.

- `Comparison on different data sets`

We test the time cost of the three algorithms in this part. The number of sub-enterprises in the system is set to 10, and the articles are classified into 10 fields and each field belongs to a sub-enterprise. The number of keywords in the dictionary is set to 600, the number of files per sub-enterprise is set to 600, and the comparison is shown in Table 2. It can be observed that the time cost of `Search` is much lower if each sub-enterprise holds articles from different fields, which is closer to the practical. The time cost of `IndexEnc` is relatively long, but it is a one-time operation, so it can be accepted by the sub-enterprises. The time cost of `TrapGen` and `Search` is in milliseconds, which is very efficient.

Table 2: Time cost comparison

DataSet	`IndexGen`	`TrapGen`	`Search`
Unclassified	4.19s	5.04ms	20.43ms
classified	4.19s	5.04ms	6.01ms

6 Conclusions

This paper proposes a secure enterprise data search scheme named SEDS to solve the problem of secure data storage and search in the cloud for group enterprise. In this scheme, each sub-enterprise generates a secure index and stores it on the cloud server. The authorized sub-enterprise can generate a trapdoor to search the top-k related files. Through security and performance analysis, SEDS scheme can be applied to search the enterprise group data on the cloud securely and effectively. In this paper, text search is mainly considered, but enterprise data also contains a large number of images such as design drawings. Therefore, we will focus on the secure search on other forms of data such as images in the future.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] N. Aaraj, C. Marcolla, and X. Zhu, "Exipnos: An efficient verifiable dynamic symmetric searchable encryption scheme with forward and backward privacy," in *International Conference on Cryptology in India*, Springer, pp. 487–509, 2021.

[2] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, pp. 506–522, 2004.

[3] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, 2013.

[4] P. S. Chung, C. W. Liu, and M. S. Hwang, "A study of attribute-based proxy re-encryption scheme in cloud environments", *International Journal of Network Security*, vol. 16, no. 1, pp. 1-13, 2014.

[5] E.-J. Goh, "Secure indexes," *Cryptology ePrint Archive*, 2003.

[6] F. Hahn and F. Kerschbaum, "Searchable encryption with secure and efficient updates," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 310–320, 2014.

[7] B. Hayes, "Cloud computing," 2008.

[8] M. S. Hwang, C. C. Lee, and S. T. Hsu, "An elgamal-like secure channel free public key encryption with keyword search scheme," *International Journal of Foundations of Computer Science*, vol. 30, no. 02, pp. 255–273, 2019.

[9] S. T. Hsu, C. C. Yang, and M. S. Hwang, "A study of public key encryption with keyword search", *International Journal of Network Security*, vol. 15, no. 2, pp. 71–79, Mar. 2013.

[10] M. S. Hwang, T. H. Sun, C. C. Lee, "Achieving dynamic data guarantee and data confidentiality of public auditing in cloud storage service," *Journal of Circuits, Systems, and Computers*, vol. 26, no. 5, 2017.

[11] H. Li, T. Wang, Z. Qiao, B. Yang, Y. Gong, J. Wang, and G. Qiu, "Blockchain-based searchable encryption with efficient result verification and fair payment," *Journal of Information Security and Applications*, vol. 58, p. 102791, 2021.

[12] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of public auditing for shared data storage with user revocation in cloud computing", *International Journal of Network Security*, vol. 18, no. 4, pp. 650–666, 2016.

[13] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of attribute-based access control with user revocation in cloud data storage," *International Journal of Network Security*, vol. 18, no. 5, pp. 900–916, 2016.

[14] Z. Liu, T. Li, P. Li, C. Jia, and J. Li, "Verifiable searchable encryption with aggregate keys for data sharing system," *Future Generation Computer Systems*, vol. 78, pp. 778–788, 2018.

[15] E. U. Opara and O. J. Dieli, "Enterprise cyber security challenges to medium and large firms: An analysis," *International Journal of Electronics and Information Engineering*, vol. 13, no. 2, pp. 77–85, 2021.

[16] J. Ren, L. Zhang, and B. Wang, "Decentralized multi-authority attribute-based searchable encryption scheme," *International Journal of Network Security*, vol. 23, no. 2, pp. 332–342, 2021.

[17] E. Stefanov, C. Papamanthou, and E. Shi, "Practical dynamic searchable encryption with small leakage," *Cryptology ePrint Archive*, 2013.

[18] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," in *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, pp. 71–82, 2013.

[19] I. H. Witten, I. H. Witten, A. Moffat, T. C. Bell, T. C. Bell, E. Fox, and T. C. Bell, *Managing gigabytes: compressing and indexing documents and images*, Morgan Kaufmann, 1999.

[20] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 139–152, 2009.

[21] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 340–352, 2015.

[22] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation Computer Systems*, vol. 28, no. 3, pp. 583–592, 2012.

[23] L. Zou, X. Wang, and S. Yin, "A data sorting and searching scheme based on distributed asymmetric searchable encryption." *International Journal of Network Security*, vol. 20, no. 3, pp. 502–508, 2018.

Biography

Kaishi Wang received the B.S. degree in information security from Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012. He currently works for Shaanxi Non-ferrous Photovoltaic Technology Co., Ltd, and is a MBA student in Northwest University, Xi'an. His research interests include data security, digital models and decision-making.

Jiaqi Guo received the B.S. degree in information security from Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012. Her M.S. degree was received in communication and information system from Lanzhou University of Technology in 2017. She is currently working toward her PhD degree in the Institute of Computing Theory and Technology, Xidian University, Xi'an, P.R. China. Her research interests include applied cryptography, cloud computing security and privacy.

Element Extraction Method of Industrial Internet Security Situation Analysis Based on SDAE-IRF

Peng-Shou Xie, Ying-Wen Zhao, Shuai Wang, Wei Li, Wan-Jun Shao, and Tao Feng
(Corresponding author: Ying-Wen Zhao)

School of Computer and Communications, Lanzhou University of Technology
No. 36 Peng Jia-ping Road, Lanzhou, Gansu 730050, China
Email: 1376144882 @qq.com

(Received July 24, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 13, 2022)

Abstract

Security situational analysis element extraction is the bottom layer of the Industrial Internet security situational visualization and early warning model. It is also the basis of Industrial Internet security situational awareness. However, the limitations of traditional situational analysis element extraction methods for multi-featured, high-dimensional nonlinear data processing capabilities lead to imprecise and inefficient situational analysis element extraction. In response to the above problems, this paper establishes an Industrial Internet security posture element extraction model based on the traditional network security posture element extraction model. It proposes an Industrial Internet security posture analysis element extraction method based on SDAE-IRF. SDAE is used to extract features from the original input data by a series of nonlinear transformations, IRF filters the base classifier, the original random forest algorithm is improved by using weighted majority voting, and the final classification results are obtained by training the reduced-dimensional data for classification. On this basis, experimental validation of the effectiveness of the proposed method is carried out in this paper using a natural gas pipeline dataset. The experimental results show that compared with the traditional security situational analysis element extraction method, the method achieves dimensionality reduction processing and feature extraction of Industrial Internet security data, which effectively improves the efficiency and accuracy of the industrial Internet security situational analysis element extraction method.

Keywords: Denoising Autoencoder; Element Extraction Method; Improving Random Forests; Industrial Internet; Security Situation Analysis

1 Introduction

The Industrial Internet is a concept that combines industrial systems and Internet technologies, a new product that highly integrates global industrial technologies with advanced computer technologies, analytics, sensor technologies, and networking technologies [18].

The Industrial Internet breaks through the relatively occludable environment of the traditional industrial system and connects all systems and production units to the external network, with the consequent exposure of defects in the production system to the Internet [9]. Since production equipment systems are usually continuous and used for a long time, the resulting defects cannot be regularly repaired and updated, making them vulnerable to network attacks and thus posing an increasing threat to the security of the enterprise. If a safety accident occurs, it will not only cause huge economic losses but also endanger public security.

The core computer network of the Guri hydroelectric power plant in Venezuela was maliciously attacked in March 2019, and approximately 30 million people were affected by the power outage [22]. In January 2020, Iran created ransomware called Snake, which is a malicious attack that specifically targets network systems in the industrial sector and not only deletes the volume of shadow copies of computers but also kills industrial control processes such as SCADA, causing serious damage to industrial systems [19].

With security incidents of different scales occurring from time to time, it is evident that industrial Internet security is as urgent as a star, and there is a growing demand for new technologies to achieve real-time monitoring and early warning of network security conditions. Industrial Internet security situational awareness technology is a new security technology that can quantitatively and qualitatively assess the current and future network security situation and monitor and alert it in real time [16].

The accuracy of the situational analysis element extraction will have an important impact on the overall performance of the entire industrial network security assurance system. The main purpose of situational analysis element extraction is to remove redundant data and extract potentially valuable situational elements [20], This provides a data basis for the next step of situational analysis and situational visualization and warning. Zhang Xin [23] addressed the sample imbalance problem by using a convolutional neural network as the base classifier, extracting deep features in the network, and using GAN to generate adversarial network extensions to overcome the sample imbalance problem and effectively improve the classification accuracy of cyber security posture elements. A new way of thinking is proposed for the problems of incomplete data features and implied relationships between data packets in the current network security posture element extraction.

Lu-zhe Cao [2] proposed a method combining CNN and BiLSTM for situational element extraction, which extracts the temporal and spatial features of the data in both time and space while mining the hidden relationships between the data, which extracts the features of the data more comprehensively and improves the effect of situational element extraction.

Gyoung S. Na [15] proposed a feature extraction method based on an unsupervised subspace extractor, which does not require any rigorous assumptions on the data from any auxiliary information and, in addition, uses subspaces that can be found generated from nonlinear combinations of input features and automatically determines the optimal dimensionality of the subspace given the nonlinear combinations. Although researchers at home and abroad have conducted a lot of researches on network security situational element extraction methods, when faced with multi-featured, high-dimensional nonlinear data their limitations lead to certain limitations in their processing capabilities, resulting in imprecise and inefficient extraction of situational analysis elements [7].

To address the above issues, this paper analyzes the status of situational extraction in the industrial Internet environment by studying and analyzing the feature extraction and classification methods in situational analysis element extraction methods [24].

Using the effectiveness of feature extraction and the strengths and weaknesses of classification algorithms as evaluation metrics for situational analysis element extraction methods. Using a noise-reducing self-encoder algorithm to select redundant data in the dataset for rejection. Secondly, by improving both the selection of classifiers and the voting method of the traditional random forest classification algorithm, the extracted features can be accurately classified, thus providing a more accurate method for the extraction of situational analysis elements to a certain extent.

2 Industrial Internet Security Posture Analysis Element Extraction Model

In this section, we introduce an element extraction model and specific extraction method design for industrial Internet security situation analysis.

2.1 Extraction Model

According to the working principle of industrial Internet security situational analysis element extraction, the extraction model of industrial Internet security situational analysis elements established in this paper is shown in Figure 1. The model starts by obtaining the raw data from the industrial Internet and performing preprocessing operations on it [10], where the data are digitally transformed and normalized. Because this dataset contains three non-numeric types of data, and there is usually a large amount of extreme difference between the individual feature quantities in the dataset, direct experiments on the original data would lead to a decrease in the speed of solving the optimal solution of the algorithm and a decrease in the classification accuracy [3]. Finally, the SDAE-IRF factor extraction method proposed in this paper is used to extract and classify the features of the pre-processed data to obtain the set of analysis factors required for situational analysis.

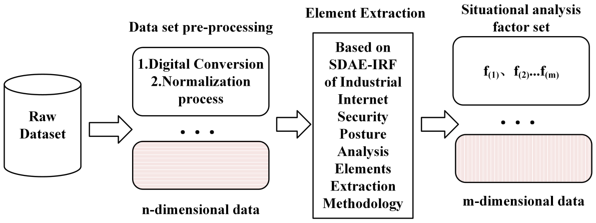

Data set pre-processing
Element Extraction
Situational analysis factor set
Raw Dataset
1.Digital Conversion 2.Normalization process
Based on SDAE-IRF of Industrial Internet Security Posture Analysis Elements Extraction Methodology
f(1)、 f(2)...f(m)
n-dimensional data
m-dimensional data

Figure 1: Industrial Internet security posture analysis element extraction model

2.2 Extraction Method Design

Since the stacked noise reduction self-encoder algorithm (SDAE) has good complex function approximation capability, it has good feature learning capability for a large amount of data, and can effectively select important attributes and remove redundant attributes. Moreover, the Improved Random Forest algorithm (IRF) can improve the overall classification accuracy and generalization ability by filtering the base classifiers from both the classification accuracy and diversity of the base classifiers for the random forest against the drawback that the traditional random forest has direct majority voting for the base classifiers and ignores the impact of redundant classifiers on the overall classification accuracy. And because its model training time is short and the matching ability is extremely strong, especially for big data, this paper

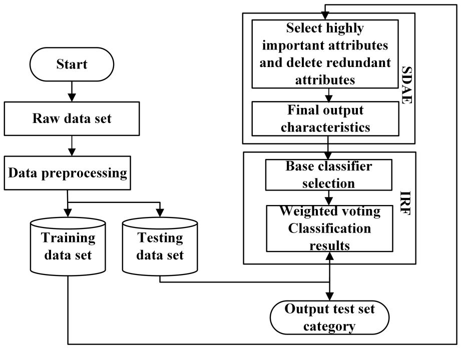

Start
Raw data set
Data preprocessing
Training data set
Testing data set
Select highly important attributes and delete redundant attributes
Final output characteristics
SDAE
Base classifier selection
Weighted voting Classification results
IRF
Output test set category

Figure 2: Block diagram of SDAE-IRF extraction method

combines the two algorithms for industrial Internet security situational analysis element extraction, and finally obtains the extraction results of situational analysis elements. The overall framework of the extraction method is shown in Figure 2.

As can be seen from Figure 2, the processed data are fed into the noise-reducing self-encoder algorithm for dimensionality reduction and feature extraction to obtain attributes with higher importance and remove redundant attributes with poor complementarity. These features are then fed into the improved random forest algorithm to generate a certain number of base classifiers, and then the base classifiers with higher classification accuracy are filtered out, and then the final classification results are obtained by doing weighted voting on the base classifiers.

3 SDAE-IRF Based Element Extraction Method for Situational Analysis

3.1 Noise-cancelling Self-encoder SDAE

Autoencoder (AE) is an unsupervised neural network structure, which has a three-layer structure, bounded by the hidden layer, the lower input layer is the encoder (encoder), and the upper output layer is the decoder (decoder). Its structure is shown in Figure 3. In the hidden layer, the input information is first encoded and then the input data is reconstructed [21]. The reconstruction error between the output information and the input information is made close to the minimum value, to obtain an abstract representation of the data.

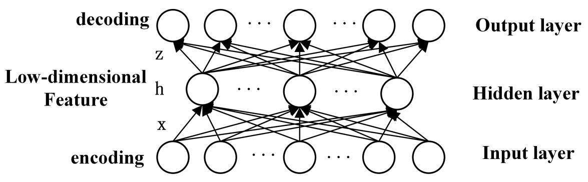


Figure 3: The architecture of AE for feature extraction

Suppose the number of neurons in the input layer and hidden layer of the encoder are D_i and D_h respectively. Given a set of unlabeled datasets $X = \{x_1, x_2, ..., x_n\}, x_i \in R^n$.The training process consists of two parts: an encoder and a decoder. During the encoding process, the encoder uses the mapping function f_θ to transform the input vector x_i into the hidden layer vector through a series of nonlinear mapping changes As shown in Formula (1).

$$h(x_i) = f_\theta(x_i) = f(W_1 x_i + b_1). \tag{1}$$

Among them, the parameter set $\theta = \{W, b\}$, $W_1 \in R^{D_i \times D_h}$ which represents the encoding weight matrix, $b_1 \in R^{D_h}$ is the offset vector.

In the decoding process, the mapping function g_θ is also used to linearly map the hidden layer vector h to reconstruct the input vector x_i, and obtain the output vector z as shown in Formula (2).

$$Z_i = g_\theta(h(x_i)) = W_2 h(x_i) + b_2 \tag{2}$$

Among them, the parameter set $\theta = \{W_2, b_2\}$ The decoding weight matrix is represented by $W_2 \in R^{D_h \times D_i}$, $b_2 \in R^{D_i}$ represents the offset vector.

From the above training model, the main task of the autoencoder is to minimize the reconstruction error of the input vector x and the output vector z by adjusting the relevant parameters of the encoder and decoder. That is, the cost function of input and output is minimized [17]. The cost function is defined as Formula (3).

$$J(W,b) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\|z_i - x_i\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{m_i-1}\sum_{i=1}^{m_i}\sum_{j=1}^{m_i+1}\left(W_{ij}^{(l)}\right)^2 \tag{3}$$

The first part of the above formula is the mean square error term, and the second part is the regularization term, which aims to reduce the size of the weights to prevent overfitting. λ the regularization coefficient [11].By minimizing the cost function $J(W, b)$, the relevant weight matrix W and the offset vector b can be obtained.

The essence of traditional autoencoders is to make the original input data and reconstructed output data equal by learning a nonlinear mapping. The biggest problem with this mapping directly obtained by encoding and decoding is that when the features of the test samples and training samples are not completely in the same feature space, especially the main features are not in the same feature space, the training effect is relatively poor, and the generalization ability is not good [8]. Compared with the traditional auto-encoder, since the sample space of the noise reduction auto-encoder (SDAE) is destroyed, it obtains a better feature expression during the training process of learning to remove the noise data and reconstruct the original sample space, and the generalization ability is also better outstanding. In terms of parameter adjustment, this paper studies the selection of the number of hidden layer nodes in the SDAE module determines the optimal parameters, and improves the efficiency of feature extraction.

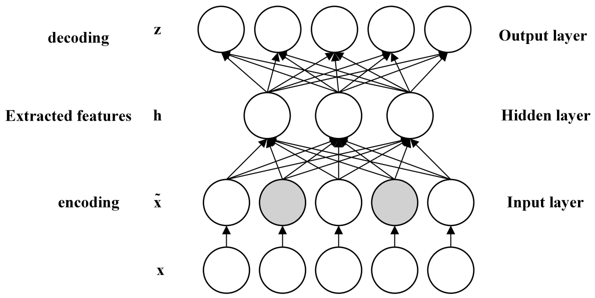


Figure 4: Structure diagram of denoising autoencoder

As shown in Figure 4, SDAE realizes the abstract feature representation of the original data by stacking multiple denoising autoencoders and reconstructing the input. This means that it trains the first hidden layer with input data and then takes the input of the first hidden layer as the first input of the second hidden layer until the desired data representation is completed [12].

Add noise to the original input data via random mapping:$\tilde{x} \to q(\tilde{x} \mid x)$, and map it to a hidden layer representation as shown in Formula (4).

$$h = f_\theta(\tilde{x}) = f\left(W_1 \tilde{x}_i + b_1\right). \tag{4}$$

The input data is reconstructed using the same reconstruction method as the autoencoder as shown in Formula (5).

$$Z_i = g_\theta(h) = W_2 h + b_2 \tag{5}$$

Likewise, the denoising autoencoder parameters $\{\theta, \theta'\}$ are obtained by minimizing the cost function as shown in Formula (6). Considering that SDAE is a stack of multilayer denoising encoders, W_{all} represents the weight matrix of the stacked denoising autoencoder network, and b_{all} represents the offset vector matrix [1].

$$W_{all}, b_{all} = \underset{\theta_i, \theta'}{\arg\min} J(W, b) \tag{6}$$

$l \in \{1, \ldots, L\}$ represents the number of hidden layers of SDAE, h_l represents the output vector of layer l, W_l is the weight of the layer l, and b_l Indicates the offset of the l layer, Using SDAE pre-training, the final output features can be obtained as shown in Formula (7).

$$h_{l+1} = f\left(W_{l+1} + b_{l+1}\right). \tag{7}$$

where $f(x)$ is the activation function, h_{l+1} representing the final output feature.

The above is the training process of a single-layer denoising autoencoder. The feature extraction module in this paper is trained by stacking multiple denoising autoencoders, and the training process is similar to a single-layer denoising autoencoder: After reconstructing the features of the first layer, they are input to the next layer as a hidden layer for information training, and this process is repeated until the end of the training. After the multi-layer training is completed, the model obtains features that are good enough to represent the original input data, but the denoising autoencoder cannot yet deal with classification problems. Generally speaking, a specific classifier needs to be added to the bottom layer for classification.

3.2 Improved Random Forest Algorithm IRF

The SDAE algorithm completes the first step of factor extraction for industrial Internet security situation analysis and feature extraction. However, feature extraction ultimately deals with classification problems, so the model needs to design a classifier at the bottom of SDAE. In this paper, IRF is used as the underlying classifier, mainly based on the following points: the traditional classification algorithm with high accuracy is not high in processing ability of large-scale data; the algorithm with better processing time often has higher requirements for data preprocessing. The IRF algorithm decomposes the classification problem of large-scale data into small-scale problems processed by each sub-classified, thereby achieving fast and accurate classification. And the algorithm improves the original random forest algorithm from the selection of the base classifier and the voting method. Compared with the original random forest algorithm, the selected base classifier has higher classification accuracy, the voting method for the base classifier is more scientific, the implementation is relatively simple, and the application field is wide. The basic steps of the original random forest to achieve classification are shown in Figure 5.

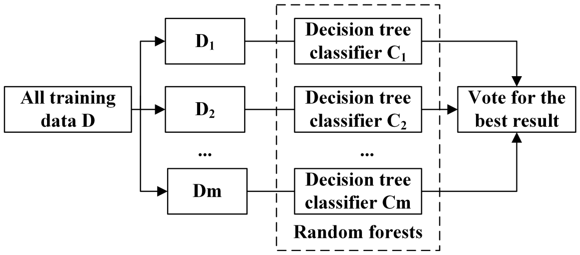

All training data D
D₁
D₂
...
Dm
Decision tree classifier C₁
Decision tree classifier C₂
...
Decision tree classifier Cm
Random forests
Vote for the best result

Figure 5: Original random forest classification diagram

Summarize the overall classification steps of the random forest algorithm from Figure 5:

1) The randomness of the sample: The original dataset is divided into two parts according to a certain principle [4], one part is used for training and the other part is used for testing. $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ is the training dataset, The bootstrap algorithm is used to randomly select the same number of samples as the original data from the sample set, and implement random sampling with replacement to obtain a training subset;

2) Randomness of features: from all the attribute features of each set of training subsets, randomly select K features, and select the best segmentation tree as a node to build M decision trees. There is no correlation between ;

3) Repeat the operation of Step (2) for each generated decision tree so that a single decision tree with a better effect can be generated;

4) The M decision trees form a random forest, and the classification of data is determined in the form of voting. The voting mechanism includes a one-vote veto system, a majority voting method, a weighted majority voting method, etc. The majority voting method is shown in Formula (8):

$$H(X) = \arg\max \sum_{i=1}^{k} I\left(h_i(X) = Y\right) \tag{8}$$

where $H(X)$ represents the final output element in the random forest algorithm, h_i represents the voting result of a single decision tree, and $I(.)$ represents an indicative function.

For the improved random forest algorithm, the complementarity and independence between each group of base classifiers affect its classification effect and efficiency. Therefore, the first problem to be solved in the selection of base classifiers is how to remove classifiers with similar classification results [5]. This paper adopts a pairwise diversity measurement method that only focuses on the proportion of samples with different classifications between the two groups of classifiers. First, introduce the following symbols: Assuming that k base classifiers are established, where C_i and $C_j (i, j = 1, 2, 3 \ldots K, i \neq j)$ are any two different base classifiers in the base classifier, As shown in Table 1. $N_{11}\,(N_{00})$ is the number of elements of industrial Internet security situation analysis that both classifiers classify right (wrong), $N_{10}\,(N_{01})$ is the number of situation analysis elements with one right and one wrong in classification.

Table 1: Classification results of the two classifiers

C_i	C_j	
	1	0
1	N_{11}	N_{10}
0	N_{01}	N_{00}

From Table 2, the total amount of the situation analysis elements of the two base classifiers can be calculated as shown in Formula (9).

$$K = N_{11} + N_{00} + N_{10} + N_{01} \tag{9}$$

The inconsistency measure of classifier classification results is concerned with the samples of different classification results produced by the two classifiers C_i and C_j for the same feature classification, The inconsistent formulation between these two classifiers is shown in the Formula (10).

$$C_{ij} = \left(N_{10} + N_{01}\right)/K \tag{10}$$

It can be seen from the above formula that the more data points with different classification results between the two classifiers, the greater the mutual dissimilarity between the two classifiers, and the value range is between [0, 1].where C_{ij} is a classifier with similar classification results. After selecting the random forest classifier inconsistency index, the base classifier with the best performance can be obtained, thereby improving the classification accuracy of the random forest algorithm as a whole.

The voting method adopted by the original random forest is to perform a majority vote on all base classifiers with equal weights to obtain the extraction results of the final situation analysis elements [14]. In this way, the differences between different base classifiers are ignored, and there will always be a wrong voting phenomenon of the base classifiers with poor performance, which will finally affect the final classification result of the random forest algorithm. The weighted majority voting method is the most intuitive and commonly used comprehensive method. Through the weighted voting method, a larger voice can be assigned to a classifier with better performance. The relationship between the weight of each base classifier and the classification accuracy of the corresponding base classifier is shown in the Formula (11).

$$W = \ln \frac{p_i}{1 - p_i}, (i = 1 \ldots K) \tag{11}$$

where p_i is the classification accuracy of the ith base classifier, W is the weight corresponding to the p_i accuracy, In this paper, the above formula is used to calculate the weight of each base classifier after selection, and the corresponding formula between the final classification result and the weight is shown in Formula (12).

$$f_{\text{inal}}\ (x) = \arg\max \left\{ \sum_{m \in K, f_{\text{tree}}\ m(x) = i, i = 1, 2, \ldots, n} W_m \right\} \tag{12}$$

where f_{inal} is the final classification result, i is the number of categories, and m is the base classifier, Improve the overall classification effect by changing the voting method of the base classifier.

4 Experiments and Analysis of Results

4.1 Experimental Environment and Dataset

The experimental programming environment in this paper is Python 3.6 under Windows 10, and PyCharm Community 2017 is selected as the programming platform. Intel Core i7-6500U CPU 2.50GHz, 16GB RAM, The related algorithms are implemented with the help of the

Table 2: The corresponding relationship between data types and labels

Types of attacks	A detailed description	The label
Normal	Normal data	0
NMRI	Simple malicious response injection attack	1
CMRI	Complex malicious response injection attacks	2
MSCI	Malicious status command injection attack	3
MPCI	Malicious parameter command injection attack	4
MFCI	Malicious function command injection attacks	5
DoS	Denial of service attack	6
Recon	Reconnaissance attacks	7

machine learning library sci-kit-learn. The data set is selected from the public natural gas pipeline SCADA system data set. The data set contains a total of 274,628 instances, 26 dimension features, and a label column, and the category label contains seven different types of attacks.

Because the number of data samples in the data set is huge, it is difficult to conduct experimental research, so this paper randomly selects 5000 data from the original data set as experimental data for experiments. Select 4/5 of the experimental data as the training set and 1/5 of the data set as the test set [6]. Among them, 3137 normal data are extracted, 158 simple malicious response injection attack data, 783 complex malicious response injection attack data, 42 malicious status command attack data, 406 malicious parameter command injection attack data, and malicious function command injection attack data 19, 87 denial of service attack data, and 368 reconnaissance attack data.

4.2 Analysis of Experimental Results

To verify the effectiveness of this paper's industrial Internet security situation analysis factor extraction algorithm, this paper selects the accuracy rate and the false alarm rate as the performance indicators. The classification accuracy refers to the ratio between the number of correctly divided samples and the number of all samples. The higher the classification accuracy, the better the performance of the method; The false alarm rate refers to the ratio of the number of wrong samples in the number of detected positive samples to the total number of samples. The lower the value of the false alarm rate, the better the performance of the method. The calculation of the correct rate and false alarm rate is shown in Formula (13) and (14).

$$\text{Accuracy } = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$\text{Error } = \frac{FN}{TP+FN} \tag{14}$$

Among them, TP is positive, indicating that the classification result is a positive sample, which is a positive sample; FP is a false positive, indicating that the classification result is a positive sample, but it is a negative sample; TN is a true negative, indicating that the classification result is a negative sample, which is a negative sample; FN is a false negative, indicating that the classification result is a negative sample, which is a positive sample. In this paper, anomalous attacks in the dataset are set as positive samples.

Reference [13] discussed the training strategy of the deep neural network and pointed out that the increase in the number of network layers can enhance the representation learning ability of the denoising autoencoder, but too many layers will also lead to the reduction of the generalization of the network. Therefore, it is very important to choose the appropriate number of layers to construct the SDAE network structure.

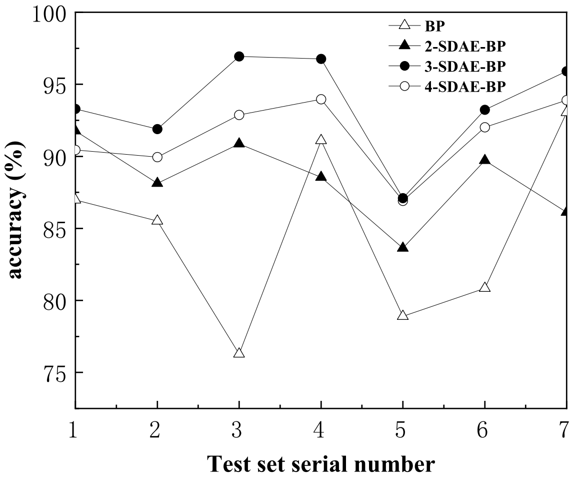


Figure 6: Comparison of SDAE feature extraction algorithms at different levels

In this paper, the SDAE method with coding layers 2, 3, and 4 layers is used to extract the situation analysis elements from the original data set. The performance comparison of the SDAE structure algorithm and BP feature extraction algorithm of different layers is shown in Figure 6. 2-SDAE-BP represents the BP algorithm with two hidden layers, 3-SDAE-BP represents the BP algorithm with three hidden layers, and 4-SDAE-BP represents the BP algorithm with four hidden layers.

As shown in Figure 6, when the number of hidden layers is 2, the advantage of the algorithm is not obvious compared with the BP algorithm. This is because when the number of hidden layers is small, the learning data reconstruction ability of the algorithm is not high, resulting in the feature information of the situation analysis elements cannot being sufficiently extracted. Due to the increase in the number of hidden layers, it is obvious that the feature learning ability of the algorithm is enhanced, and the extraction accuracy is also significantly improved. However, the extraction accuracy of the 4-SDAE-BP algorithm in the figure is slightly lower than that of the 3-SDAE-BP algorithm. This is because when the number of hidden layers increases, there are a large number of nonlinear transformations between layers, resulting in the existence of feature information of situation analysis elements. A certain loss is inevitable, and this loss is unavoidable, thereby reducing the accuracy of feature extraction. Therefore, this paper uses the 3-SDAE-BP algorithm with three hidden layers to extract the situation analysis elements from the original data set.

To verify the effectiveness of the IRF classification algorithm in this paper, Table 3 and Table 4 compare the two evaluation indicators of classification accuracy and false alarm rate obtained by different classification methods.

Table 3: Classification accuracy data of different methods

Attribute	**Classification accuracy(%)**		
	Literature [22]	**Literature [5]**	**This paper**
5	91.39	95.83	98.80
11	93.08	92.94	98.64
13	92.62	94.46	96.02
17	91.87	93.20	97.51

It can be seen from the data in Table 3 that when the number of attributes is 17, the average classification accuracy of the literature [22] method is 82.46%, The average classification accuracy of the literature [5] method is 85.78%, while the average classification accuracy of the IRF classification method in this paper is as high as 94.52%. The main reason is that the generated base classifiers are screened based on the original random forest, and the base classifiers with similar classification results are removed, thereby improving the classification accuracy of the classification algorithm as a whole.

From the data in Table 4, it can be seen that the false alarm rate of the classification algorithm IRF in this paper for different categories of attacks in the data set is lower than that of the literature [22] method and the literature [5] method. The voting of the controller has also been improved from the original majority vote to the weighted vote, thereby effectively reducing the false positive rate.

Table 4: False positive rate data for different methods

Category	**False alarm rate(%)**		
	Literature [22]	**Literature [5]**	**This paper**
Normal	12.6	13.2	2.1
NMRI	0.9	0.7	0.4
CMRI	0.8	0.6	0.5
MSCI	1.4	1.2	1.1
MPCI	2.8	4.7	1.6
MFCI	4.7	3.5	2.3
DoS	16.0	18.0	17.8
Recon	7.6	5.3	4.1
Weight	9.5	8.2	6.7

Figure 7 below represents the comparison graph of the correct rate of attack extraction for seven different types of attacks, including no-attack traffic, NMRI, CMRI, MSCI, MPCI, MFCI, DOS, Recon, for the three posture analysis element extraction methods of literature [22], literature [5] and the method in this paper, The different shaped bars in the figure represent different methods, the horizontal coordinates indicate the different types of attacks, and the vertical coordinates indicate the correct rate of situational analysis element extraction.

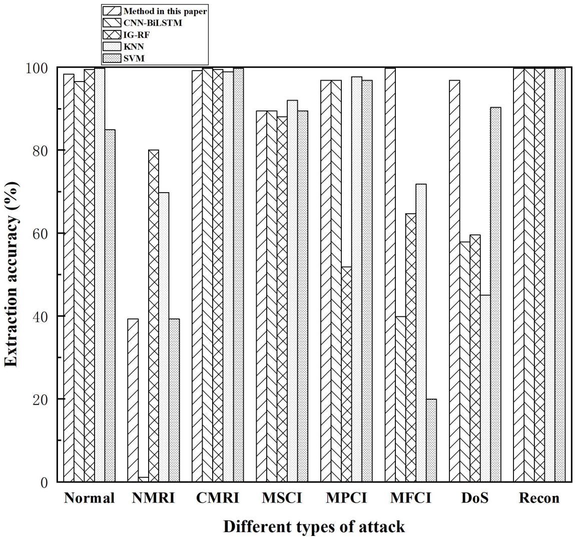


Figure 7: Extraction accuracy of different attack types

As shown in Figure 7, five different bar graphs represent the situation analysis element extraction method SDAE-IRF, convolutional neural network CNN and LSTM original situation element extraction method, information gains Random forest (IG-RF) situation element extraction method, KNN situation element extraction method, and SVM situation element extraction method respectively. It can be seen from the figure that the situation analysis element extraction algorithm SDAE-IRF in this paper is effective against normal data (Normal), complex malicious response injection attack (CMRI), malicious state command injection attack (MSCI), malicious parameter command injection attack (MPCI), The extraction of malicious function command injection attack

(MFCI), denial of service attack (DoS), and reconnaissance attack (Recon) is good, but the extraction of simple malicious response injection attack (NMRI) is not ideal, because This type of attack is sensitive to the sample data values used, resulting in poor extraction results. From the overall trend, the SDAE-IRF proposed in this paper has a better extraction effect among the above 5 algorithms, and its performance is relatively stable. Compared with other traditional machine learning algorithms, the accuracy has been significantly improved.

5 Conclusions

This paper first conducts systematic modeling for the extraction of industrial Internet security situation analysis elements. From the extraction model diagram in this paper, it can be seen that the performance of the situation analysis element extraction method will directly affect the dimension of the situation analysis element set, and it will also affect the security of the industrial Internet. Situational analysis results play a key role. Therefore, this paper designs a situation analysis element extraction method based on SDAE-IRF. Finally, through experiments, the performance of different levels of the SDAE structure feature extraction algorithm, the classification performance of different situation analysis element extraction methods, and different extraction algorithms are used to evaluate the attack type extraction efficiency. Comparative analysis. The experimental results show that the method used in this paper is an effective method for extracting the elements of situation analysis, which overcomes the limitations of the traditional method of extracting situation elements in the processing of multi-feature, high-dimensional and nonlinear data and the elements of situation analysis. It is difficult to extract inaccurate and low efficiency. It is more conducive to the development of the extraction of industrial Internet security situation analysis elements and provides data support for the subsequent industrial Internet security situation analysis and situational early warning work, but there are still shortcomings. Improving the efficiency of the algorithm and reducing the demand for storage space is the focus of the next phase of research work.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under Grants No.61862040 and No.61762059. The authors gratefully acknowledge the anonymous reviewers for their helpful comments and suggestions.

References

[1] A. Ashfahani, M. Pratama, E. Lughofer, and Y.-S. Ong, "Devdan: Deep evolving denoising autoencoder," *Neurocomputing*, vol. 390, pp. 297–314, 2020.

[2] L. Z. Cao, "Research on extraction and evaluation method of campus network security situation elements based on deep learning," Master's Thesis, The Chinese People's Public Security University, 2021.

[3] R.-H. Dong, C. Shu, and Q.-Y. Zhang, "Security situation assessment algorithm for industrial control network nodes based on improved text simhash," *International Journal of Network Security*, vol. 23, no. 6, pp. 973–984, 2021.

[4] R.-H. Dong, Y.-L. Shui, and Q.-Y. Zhang, "Intrusion detection model based on feature selection and random forest," *International Journal of Network Security*, vol. 23, no. 6, pp. 985–996, 2021.

[5] R.-H. Dong, Y.-L. Shui, and Q.-Y. Zhang, "Intrusion detection model based on feature selection and random forest," *International Journal of Network Security*, vol. 23, no. 6, pp. 985–996, 2021.

[6] Y. Duan, X. Li, X. Yang, and L. Yang, "Network security situation factor extraction based on random forest of information gain," in *Proceedings of the 2019 4th International Conference on Big Data and Computing*, pp. 194–197, 2019.

[7] R. F. Fouladi, O. Ermiş, and E. Anarim, "A ddos attack detection and defense scheme using time-series analysis for sdn," *Journal of Information Security and Applications*, vol. 54, p. 102587, 2020.

[8] X. Han, Y. Liu, Z. Zhang, X. Lü, and Y. Li, "Sparse auto-encoder combined with kernel for network attack detection," *Computer Communications*, vol. 173, pp. 14–20, 2021.

[9] J. He, J. Yang, K. Ren, W. Zhang, and G. Li, "Network security threat detection under big data by using machine learning." *International Journal of Network Security*, vol. 21, no. 5, pp. 768–773, 2019.

[10] W. Kehe, Z. Ying, and G. Minghao, "Situation awareness technology of lenet-5 attack detection model based on optimized feature set," in *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 269–272, 2020.

[11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, pp. 5530–5540, 2021.

[12] W.-H. Lee, M. Ozger, U. Challita, and K. W. Sung, "Noise learning-based denoising autoencoder," *IEEE Communications Letters*, vol. 25, no. 9, pp. 2983–2987, 2021.

[13] M. Ma, W. Hao, and P. Li, "Combining position-aware cnn and rnn for relation extraction," in *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pp. 427–432, 2019.

[14] S. Ma, J. Li, Y. Wu, C. Xin, Y. Li, and J. Wu, "A novel multi-information decision fusion based on improved random forests in hvcb fault detection application," *Measurement Science and Technology*, vol. 33, no. 5, p. 055115, 2022.

[15] G. S. Na and H. Chang, "Unsupervised subspace extraction via deep kernelized clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 1, pp. 1–15, 2021.

[16] W. Xi, W. Wu, and C.-Y. Yang, "A technical review on network security situation awareness," *International Journal of Network Security*, vol. 24, no. 4, pp. 671–680, 2022.

[17] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. Sabrina, "Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset," *IEEE Access*, vol. 9, pp. 140 136–140 146, 2021.

[18] W. Wu and C.-Y. Yang, "An overview on network security situation awareness in internet," *International Journal of Network Security*, vol. 24, no. 3, pp. 450–456, 2022.

[19] H. Yang, L. Cheng, and M. C. Chuah, "Deep-learning-based network intrusion detection for scada systems," in *2019 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–7, 2019.

[20] Y. Yang, C. Yao, J. Yang, and K. Yin, "A network security situation element extraction method based on conditional generative adversarial network and transformer," *IEEE Access*, vol. 10, pp. 107 416–107 430, 2022.

[21] M. Yu, T. Quan, Q. Peng, X. Yu, and L. Liu, "A model-based collaborate filtering algorithm based on stacked autoencoder," *Neural Computing and Applications*, vol. 34, no. 4, pp. 2503–2511, 2022.

[22] P. Yu, Y. Long, H. Yan, H. Chen, and X. Geng, "Design of security protection based on industrial internet of things technology," in *2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, IEEE, pp. 515–518, 2022.

[23] Z. J. Zhang Xin, "Sampling unbalanced network security situation factor acquisition," *Computer Engineering and Applications*, vol. 58, no. 1, p. 9, 2022.

[24] B. Zhu, Y. Chen, and Y. Cai, "Three kinds of network security situation awareness model based on big data", *International Journal of Network Security*, vol. 21, no. 1, pp. 115–121, 2019.

Biography

Peng-shou Xie was born in Jan.1972. He is a professor and a supervisor of master student at Lanzhou University of Technology. His major research field is Security on Internet of Things. E-mail: xiepsh lut@163. com

Ying-wen Zhao was born in Feb.1996. He is a master student at Lanzhou University of Technology. He major research field is network and information security. E-mail: 1376144882@qq. com

Shuai Wang was born in Aug.1995. He is a master student at Lanzhou University of Technology. His major research field is network and information security. E-mail: 627858493@qq. com

Wei Li was born in Mar.1998. He is a master student at Lanzhou University of Technology. His major research field is network and information security. E-mail: 2304118232@qq. com

Wan-jun Shao was born in Mar.1998. She is a master student at Lanzhou University of Technology. His major research field is network and information security. E-mail: 2443404684@qq. com

Tao Feng was born in Dec.1970. He is a professor and a supervisor of Doctoral student at Lanzhou University of Technology. His major research field is modern cryptography theory, network and information security technology. E-mail: fengt@lut. cn

Enhancing the Robustness of Deep Neural Networks by Meta-Adversarial Training

You-Kang Chang, Hong Zhao, and Wei-Jie Wang

(Corresponding author: You-Kang Chang)

School of Computer and Communications, Lanzhou University of Technology
36 Peng-jia-ping Road, Lanzhou, Gansu 730050, China
Email: 2507576651@qq.com

(Received July 29, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 13, 2022)

Abstract

Adversarial training can effectively defend against the impact of adversarial attacks on deep neural networks but suffers from poor generalization ability and low defense efficiency. To address this problem, this paper proposes a method combining meta-learning with adversarial training to enhance the robustness of deep neural networks. Firstly, a training dataset containing adversarial examples and clean examples is constructed, and conduct adversarial training on the deep neural network. Secondly, the features extracted from the adversarial training are learned using the meta-learning method, and the problem of the need to continuously input a large number of adversarial examples for training in adversarial training is solved by using the feature that meta-learning has strong adaptability in the face of new tasks. Experimental results show that this method can improve the robustness of deep neural networks and effectively resist standard classes of adversarial attacks.

Keywords: Adversarial Training; Adversarial Attack Defense; Meta-learning; Neural Networks; Robustness Studies

1 Introduction

Deep neural networks play an increasingly important role as deep learning is applied to an increasingly wide range of scenarios. For example, it has shown good performance in autonomous driving [22,26], medical image analysis [1,27] and image recognition [31]. However, research and practical applications have shown that deep neural networks are vulnerable to adversarial attacks [6,10], and are deceived by adversarial examples to produce wrong results. During the training phase of a deep neural networks, the attacker attacks by modifying the training dataset, changing the characteristics of the input data or the data labels. In the testing phase of deep neural networks, white-box attacks and black-box attacks can be used, where white-box attacks are performed by obtaining the structure of deep neural networks to generate adversarial examples, and black-box attacks are performed by querying the structure of network models and exploiting the transferability between adversarial examples.

In response to the vulnerability of deep neural networks to adversarial example attacks, researchers have successively proposed a variety of defense methods, which are mainly divided into three categories. The first category is data preprocessing, such as adversarial example denoising [29] and data compression [14], which are computationally fast and do not require modification of the network structure of the model. The disadvantage is that when modifying the input examples, the high-frequency information of the examples will be lost, making the network model unable to extract the correct feature regions and leading to the wrong classification of the neural network. The second category is to enhance the robustness of deep neural networks, such as adversarial training [8] defensive distillation methods [17] and deep compression network [7]. Such methods improve the stochasticity of the network model and the cognitive performance of the network to a certain extent. but their defensive efficiency decreases significantly if specific attacks are performed on a particular network. The third category of methods is the detection of adversarial examples before they are fed into the deep neural network, such as based on Generative Adversarial Network (GAN) [23], based on MagNet [16] and Defense Perturbation [21]. These methods have good generalization ability and good defense against black and gray box attacks in particular, however, their performance decreases substantially in the case of white box attacks.

For the second category of defense methods in the adversarial training defense mechanism, which serves as one of the most promising defense methods to improve the robustness of deep neural networks [13], it is necessary to add newly emerged adversarial examples to the training set for adversarial training in the face of never-appeared adversarial examples, and this method to improve the robustness of deep neural networks through violent training has the problems of long training time and poor generalization ability. To tackle this problem, this paper proposes an adversarial training defense method that introduces meta-learning technology, combining adversarial training with meta-learning method, and using the characteristics of meta-learning with strong generalization and high recognition accuracy to solve the problem of poor generalization of adversarial training.

A brief overview of our contributions is as follows:

1) Application of meta-learning methods to the adversarial training process of deep neural networks to enhance the robustness of deep neural networks;

2) The method is not only effective in defending against adversarial samples, but also has no impact on the accuracy of clean samples, which are the original data set, not generated by the adversarial attack algorithm;

3) The method can still maintain high accuracy with strong generalization in the face of unprecedented adversarial examples;

4) The method is not only applicable to white-box attacks, but also has strong defense capability in the face of black-box attacks.

This paper is organized as follows: The seco section reviews related work. The third section describes the Meta-adversarial training (Meta-adv training) defense method. The fourth section conducts the experimental design as well as the analysis of the results. Finally, the full paper is summarized in section fifth.

2 Related Work

2.1 Adversarial Attacks

Kurakin *et al.* [10] proposed the Basic Iterative Method (BIM) method, where the generated perturbations are added to the input image multiple times incrementally through multiple iterations along the direction of the gradient and the gradient direction is recalculated after each iteration. Carlini and Wagner [2] proposed the Carlini and Wagner (C&W) method to generate adversarial examples using the Adam-Optimizer optimizer. Moosavi-Dezfooli *et al.* [19] proposed the DeepFool method, which is based on a binary classification problem where the minimum perturbation vector added is the vertical distance vector between x0 and the straight line. Xie *et al.* [30] proposed Diverse-Input-Iterative FGSM(DI^2FGSM) and Momentum-Diverse-Input-Iterative FGSM(MDI^2FGSM), where DI^2FGSM performs a random transformation of the image with probability p during the generation of the adversarial example, MDI^2FGSM method improves the efficiency of the attack by adding momentum to the DI^2FGSM method to avoid local maximums.

Figure 1 illustrates the different adversarial examples and observes the differences between them from a visual perspective and finds that the effect of the added adversarial perturbations is not significant. However, their pixel values are displayed in three-dimensional coordinates for comparison, as shown in Figure 2, where yellow indicates the pixel value is higher and blue indicates the pixel value is lower, and it can be seen that the adversarial examples after adding the perturbation differ from the normal examples in terms of pixel intensity, and the pixel values of the clean examples are smoother compared to the adversarial examples. For example, the pixel values of the adversarial examples generated by the BIM method are continuous and constant in some areas, while the pixel values of the adversarial examples generated by the MDI^2FGSM attack method fluctuate more. The variation of pixel values causes the deep neural network to extract the wrong feature regions and eventually output the wrong results.

2.2 Adversarial Training and Meta-Learning

In response to the influence brought by adversarial attacks, adversarial training and its optimization methods have been successively proposed as the most effective defense methods at present. Zhang *et al.* [32] proposed a feature scattering-based adversarial training defense method to generate adversarial examples for training by feature scattering in the potential space. Zhang *et al.* [33] proposed Friendly Adversarial Training (FAT) defense method, they believed that in using PGD attack method to generate adversarial examples for adversarial training, it will affect the accuracy of clean examples and even cause the neural network not to converge, so the proposed method will stop in time during the process of generating adversarial examples using PGD iterations and return to the adversarial examples vicinity the decision boundary for training, gradually enhancing the robustness of the deep neural network and ensuring the accuracy of clean examples. The existing adversarial training uses the adversarial examples generated by one attack method to train the deep neural networks, which cannot effectively cover other types of adversarial examples, Kwon *et al.* [11] proposed a diverse adversarial training method using a combined training set of FGSM, I-FGSM, DeepFool and C&W for adversarial training to enhance the robustness against unknown adversarial attacks.

The defense method based on adversarial training enhances the robustness of the network model by improving the randomness and cognitive properties of the deep neural network, but this method needs to retrain the network model when facing unknown types of adversarial examples, which has the problems of poor generalization ability and large computational resource consumption. For this reason, this paper introduces the meta-learning method in the process of adversarial training.

Meta-learning is a learning approach that imitates biological use of prior knowledge to quickly learn new and unseen things, and it can be a good solution to the prob-

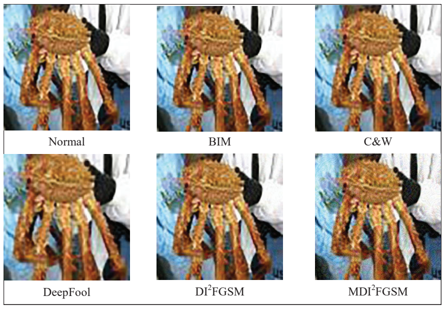


Figure 1: Comparison between different adversarial examples

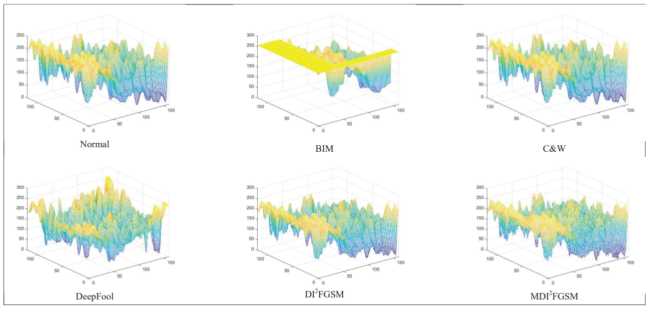


Figure 2: Three-dimensional visualization of different adversarial example pixels

lems of deep neural networks such as low robustness, poor generalization, difficulty in learning and adapting to unobserved tasks, and dependence on large-scale data. Research on meta-learning can improve the robustness and generalization of network models.

Meta-learning has now been applied to many fields, and good experimental results have been achieved by integrating meta-learning with other network models. Finn *et al.* [5] proposed the Model-Agnostic Meta-Learning (MAML) for Fast Adaptation of Deep Networks method, which is a model-independent meta-learning method for different learning problems. Firstly, The parameters of the network model are optimized using gradient descent, and then in a new task, the parameters are fine-tuned by training a small amount of data to make the network model with good generalization performance. Later, Zhang *et al.* [34] proposed the MetaGAN meta-learning method, which combined meta-learning with the Generative Adversarial Network(GAN) model to help classifiers learn clearer decision boundaries in small example data and improve the generalization performance of the network model by introducing an adversarial generative model. Li *et al.* [12] proposed Adversarial Feature Hallucination Networks (AFHN) to ensure the distinguishability and diversity of few shot data. Mandal *et al.* [15] combined Graph Neural Networks (GNNs) with meta-learning to improve the generalization performance of GNNs in the face of few shot data.

3 Meta-Adversarial Training Methods

This paper proposes to enhance the robustness of deep neural networks using meta-adversarial training, which is divided into three main phases: firstly, clean examples are combined with generated adversarial examples into a training set for feature extraction using adversarial training; secondly, the extracted features are quickly adapted to a few shot learning task in the meta-learning training phase; and finally, the defense method is tested against the adversarial examples in the meta-learning testing phase, and the robustness of the deep neural network is evaluated at the same time.

Adversarial training: adversarial training is trained by fusing adversarial examples with clean examples, which can regularize the deep neural network to certain extent and adapt the network model to this change and enhance the generalization ability. Huang *et al.* [9] defined the Min-Max problem for the first time, where: Min refers to minimizing the classification error of the network model during training process. Max refers to finding the adversarial perturbation of the input example that maximizes the classification error of the network model and states that the key to solving the Min-Max problem is to find the adversarial example with stronger attack performance. Later, Shaham *et al.* [24] considered the Min-Max problem from the perspective of robust optimization and proposed a framework for adversarial training, as shown in Equation (1):

$$\min_{\theta} \mathbb{E}_{(Z,y)\sim\mathcal{D}}\left[\max_{\|\delta\|\leqslant\varepsilon} L\left(f_\theta(X+\delta),y\right)\right] \tag{1}$$

where the inner layer denotes maximization, X denotes the input example, δ denotes the perturbation added to the input example,$f_\theta\,()$ denotes the deep neural network, and y denotes the true label of the clean example. $L\left(f_\theta(X+\delta),y\right)$ denotes the loss between the output label of the adversarial example X+δ passing through the deep neural network and the true label. max (L) denotes the optimization objective, which aims to find the perturbation that maximizes the loss function so that the added perturbation should disturb the deep neural network as much as possible.

The outer layer represents the minimization formulation of the optimized deep neural network, which trains the deep neural network to minimize its loss on the training data when the adversarial perturbation has been determined, adversarial perturbation has been determined, allowing the network model to have some robustness to adapt to the perturbation. Equation (1) describes the idea of adversarial training, but it does not describe how to design a perturbation δ with strong attack performance. therefore, the researchers proposed a variety of attack methods to find the perturbation δ. In fact, during adversarial training, the stronger the attack performance of the perturbation δ can make the deep neural network more robust.

In the meta-adversarial training defense method, feature extraction is first performed on the data in the training set using the convolution operation, then in the meta-learning phase, the parameters of the feature extractor are learned by scaling and shifting transformations to make the deep neural network quickly adapt to few shot tasks; finally, the accuracy of the test data is output in the meta-testing phase. The specific process is as follows:

Feature extraction: The parameters of the feature extractor Θ and classifier θ are first initialized, and then some of the clean examples in the mini-ImageNet training set are replaced with the generated adversarial examples, and the parameters of feature extractor Θ and classifier θ are learned by gradient descent method using the ResNet network model, as shown in Equation (2):

$$[\Theta;\theta] = [\Theta;\theta] - \alpha\nabla \lim_{x\to\infty} \mathcal{L}_D([\Theta;\theta]) \tag{2}$$

where α denotes the learning rate and $\mathcal{L}_D$ denotes the cross-entropy loss function, as shown in Equation (3):

$$\mathcal{L}_D([\Theta;\quad\theta]) = \frac{1}{\mathcal{D}} \sum_{(x,y)\in\mathcal{D}} l\left(f_{[\Theta;\theta]}(x), y\right). \tag{3}$$

Meta-learning stage: The feature extractor parameters Θ learned in the feature extraction phase remain fixed during the few shot learning process, and they are scaled and shifted transformed in the meta-learning phase to quickly adapt to unseen data examples; however, the classifier parameters θ need to be reinitialized and updated due to the inconsistency in the number of categories between the feature extraction phase and the meta-learning phase. as shown in Equation (4):

$$\theta' \leftarrow \theta - \beta\nabla_\theta \mathcal{L}_{\mathcal{T}^{(tr)}}\left([\Theta;\quad\theta], \Phi_{S_{\{1,2\}}}\right) \tag{4}$$

where Φ_{S_1} denotes the scaling transformation, which is initialized to 1, Φ_{S_2} denotes the shifting transformation, initialized to 0, $\Phi_{S_{\{1,2\}}}$ denotes the scaling and shifting transformation, $\mathcal{T}^{(tr)}$ denotes the training data, and β denotes the learning rate. Different from θ in Equation (2), θ in Equation (4) focuses on a small number of classes in the meta-learning training task to classify in a few shot of data, $\theta^{'}$ denoting the parameters of the current classification task.

During the test process, the parameters of the scaling and shifting are optimized by calculating the loss values using the test data $\mathcal{T}^{(te)}$, while updating the parameters θ, as shown in Equations (5) and (6):

$$\Phi_{S_i} = \Phi_{S_i} - \gamma\nabla_{\Phi_{S_i}} \mathcal{L}_{\mathcal{T}^{te}}\left([\Theta;\theta'], \Phi_{S\{1,2\}}\right) \tag{5}$$

$$\theta = \theta - \gamma\nabla_\theta \mathcal{L}_{\mathcal{T}^{te}}\left([\Theta;\theta'], \Phi_{S\{1,2\}}\right) \tag{6}$$

For a given Θ, the i-th layer of the feature extractor Θ contains K neurons, that is, it contains K parameter pairs and $\{(\mathrm{W_{i,k}}, \mathrm{b_{i,k}})\}$ denotes the weights and bias respectively, and if the input is X, the formula for applying $\Phi_{S_{\{1,2\}}}$ to (W, b) is shown in Equation (7):

$$SS\left(X; W, b; \Phi_{S_{\{1,2\}}}\right) = \left(W \odot \Phi_{S_1}\right) X + \left(b + \Phi_{S_2}\right) \tag{7}$$

The weights trained on large-scale datasets are migrated to the meta-learning task using the already optimized scaling and shifting. which ensure fast convergence of the deep neural network in the face of few shot data and effectively reduce overfitting.

4 Experimental Design and Analysis of Results

4.1 Experimental Platform

The experimental platform for this study is based on ubuntu 18.04, with 128G of experimental running memory. Hardware equipment using a graphics card NVIDIA Tesla V100 GPU with 32G of video memory. The experimental environment uses the PyTorch deep learning framework that supports GPU accelerated computing, and the cuda environment is configured with NVIDIA CUDA 11.3 and cuDNN V8.2.1 deep learning acceleration library.

4.2 Dataset Setup

This experiment uses the miniImageNet [28] dataset to verify the effectiveness of the model. mini-ImageNet contains a total of 100 categories, with 64 categories in the training set, 16 categories in the validation set, and 20 categories in the test set, each containing 600 images, for a total of 60,000 data samples of size 84 × 84. During the experiments, the data are first preprocessed and the samples are upsampling to 299 × 299 pixel size, and then the adversarial examples are generated using the white-box adversarial attack methods BIM, C&W, DeepFool, DI^2FGSM, MDI^2FGSM, and the black-box adversarial attack methods P-RGF, RGF [4] and Parsimonious [18].

4.3 Parameter Setting

In the pre-training phase, the parameters of the model were optimized using the SGD optimizer, setting the learning rate α=0.1, the momentum set to 0.9, and the weight decay value to 0.0005; the parameters were updated using the cross-entropy loss function, specifying that the loss value decays 0.2 in the learning rate when the model does not decline in 30 rounds. The parameter settings are shown in Table 1.

Meta-learning training phase using the Adam optimizer for optimization of parameters, setting the learning rate $\beta = 0.01$. The cross-entropy loss function is also used for parameter updating, and the learning rate is specified to decay by 0.5 when the loss value does not decline in

Table 1: Pre-training phase parameter setting

Parameters	Setting
Learning rate	0.1
Epoch	100
Weight decay	0.0005
Batch_size	128
Learning rate decay	0.2
Momentum	0.9
Step_size	30

Table 2: Meta-learning training phase parameter setting

Parameters	Setting
Learning rate	0.01
Epoch	100
Train_query	15
Val_query	15
Learning rate decay	0.5
Step_size	10
Num_batch	100

10 consecutive rounds. The training process uses 100 different tasks, with 15 examples per category in each task selected for training and 15 examples selected for validation. The parameters are set as shown in Table 2.

4.4 Analysis of Experimental Results

4.4.1 Effect of the Proportion of Adversarial Examples

During the experiments, the effects of different proportions of adversarial examples on the robustness of the deep neural network ResNet-12 are compared. Firstly, clean examples in the training set are replaced with adversarial examples in different proportions of 10%, 30%, 50%, 70% and 90% for adversarial training. The test set uses the generated adversarial examples. The experimental results of meta-adversarial training are shown in Table 3 and Table 4.

Table 3 and Table 4 show the experimental results for 1shot-5way and 5shot-5way during the meta-adversarial training, respectively. It can be seen that with the increasing proportion of adversarial examples, the accuracy of the experimental results shows an overall increasing trend, the reason being that the deep neural network treats the adversarial examples as clean examples, fitting the distribution of the data, and the loss generated by the adversarial examples as part of the loss of the deep neural network, increasing the loss of the model without modifying the structure of the network model, producing a regularization effect.

Table 3: Accuracy of 1shot-5way on the adversarial example (%)

	10%	30%	50%	70%	90%
BIM	0.6588	0.6743	0.6694	0.6729	0.6720
C&W	0.6096	0.6090	0.6112	0.6030	0.6001
DeepFool	0.6068	0.5845	0.5577	0.5905	0.5661
DI^2FGSM	0.4528	0.4691	0.5285	0.5342	0.5932
MDI^2FGSM	0.4375	0.4579	0.4983	0.5264	0.5811

Table 4: Accuracy of 5shot-5way on the adversarial example (%)

	10%	30%	50%	70%	90%
BIM	0.8243	0.8086	0.8099	0.8281	0.8269
C&W	0.7695	0.7697	0.7494	0.7644	0.7604
DeepFool	0.7673	0.7496	0.7223	0.7536	0.7276
DI^2FGSM	0.6221	0.6282	0.7024	0.7047	0.7525
MDI^2FGSM	0.5888	0.6224	0.6742	0.6992	0.7407

Table 5: Accuracy of 1shot-5way on clean examples (%)

	Clean	10%	30%	50%	70%	90%
BIM	0.6045	0.6109	0.6043	0.6030	0.5956	0.5800
C&W		0.6029	0.5997	0.6122	0.6092	0.6082
DeepFool		0.6096	0.5969	0.5905	0.5922	0.5826
DI^2FGSM		0.6060	0.5832	0.5505	0.4297	0.4180
MDI^2FGSM		0.5970	0.5656	0.5359	0.3932	0.3819

The adversarial training has achieved good results against adversarial examples, and the next step will test the effect of adversarial training in the face of clean examples. and the experimental results are shown in Table 5 and Table 6.

Clean in Table 5 and Table 6 indicates that both the training and test sets are clean examples, and can achieve 60.45% and 76.25% accuracy in the 1shot-5way and 5shot-5way cases, respectively. However, when the adversarial examples are continuously added to the training set for adversarial training, the accuracy of clean examples shows an overall decreasing trend, such as when 90% of MDI^2FGSM adversarial examples are added for adversarial training, the accuracy of clean examples in the 1shot-5way and 5shot-5way cases is only 38.19% and 55.94%, respectively, which is different from the accuracy of clean examples. Therefore, after trade-off between the accuracy of the deep neural network against adversarial examples and clean examples, we add 20% of the adversarial examples randomly to the training set for adversarial training to ensure both the accuracy of clean examples and the robustness of the deep neural network against adversarial examples. Next, we will use 20% of the adversarial examples for adversarial training to validate the defensive

Table 6: Accuracy of 5shot-5way on clean examples (%)

	Clean	10%	30%	50%	70%	90%
BIM	0.7625	0.7674	0.7628	0.7656	0.7602	0.7368
C&W		0.7612	0.7604	0.7696	0.7693	0.7673
DeepFool		0.7697	0.7603	0.7514	0.7546	0.7474
DI^2FGSM		0.7673	0.7502	0.7271	0.6280	0.6126
MDI^2FGSM		0.7589	0.7305	0.7098	0.5707	0.5594

capability against migration between attack methods.

4.4.2 Migratory Defense Against Attacks

The proposed meta-adversarial training defense method allows the deep neural network to still show good robustness against unprecedented adversarial examples, and for this reason, this section verifies the migration between the meta-adversarial training method defense against different adversarial attacks. The experimental results are shown in Table 7 and Table 8. Training indicates that 20% of the adversarial examples are added to the clean examples for training, and Test verifies the adversarial training defense against different adversarial attack algorithms.

It can be seen from Table 7 and Table 8 that the meta-adversarial training method can effectively defend against different adversarial examples and enhance the robustness of the deep neural network. For example, meta-adversarial training using 20% of BIM adversarial examples can achieve recognition accuracy of 59.59%, 60.68%, 60.30%, and 57.28% for 1shot-5way in the face of C&W, DeepFool, DI^2FGSM, and MDI^2FGSM attack methods, and this result is slightly lower than that of using C&W adversarial examples for meta-adversarial training of 61.74%, but all are higher than the results of meta-adversarial training using DeepFool, DI^2FGSM and MDI^2FGSM adversarial examples.

In the 5shot5way case, the recognition accuracy has been improved substantially in all cases. When using C&W adversarial examples for meta-adversarial training, the recognition accuracy of BIM, DeepFool, DI^2FGSM, and MDI^2FGSM remains stable, and it is able to reach 74.22% even when facing the MDI^2FGSM attack algorithm, which has a strong attack capability. The reason for the good results is that the meta-learning method can fine-tune the parameters so that the deep neural network can quickly adapt to new tasks and has good generalization performance when facing unseen adversarial examples.

After the above-mentioned comparison experiments, the following experiments will verify the effectiveness of the proposed method in defending against white-box attacks and black-box attacks.

4.4.3 Defending Against White-box Attacks and Black-box Attacks

(1) Defending Against White-Box Attacks

For BIM, C&W, DeepFool, DI^2FGSM and MDI^2FGSM white-box attack algorithms, the proposed meta-adversarial training defense method is compared with CompareNets [25], Self-Supervised Learning (SSL) [3], and Neural Representation Purifier (NRP) [20] defense methods for performance comparison, where CompareNets and SSL are meta-learning methods that use the dataset consistent with the proposed meta-adversarial training method. The experimental results are shown in Table 9.

From Table 9, it can be concluded that meta-adversarial training achieves better accuracy on BIM, C&W, DeepFool, and DI^2FGSM white-box attack algorithms compared to CompareNets, but slightly lower accuracy in the face of the stronger MDI^2FGSM attack algorithm. Meta-adversarial training achieves better accuracy on BIM, C&W, DeepFool white-box attack algorithms compared to SSL, and slightly lower performance than SSL defense methods in the face of DI^2FGSM and MDI^2FGSM attack algorithms. Compared with NRP, meta-adversarial training was lower in accuracy than NRP in the 1shot5way case, but higher in accuracy than the NRP defense method in the 5shot5way case.

(2) Defense Against Black-Box Attacks

For the defense against RGF, P-RGF and Parsimonious black box attacks, the same three defense methods of CompareNets, SSL and NRP are used for comparison with meta-adversarial training. The experimental results are shown in Table 10.

It can be seen from Table 10 that meta-adversarial training outperforms CompareNets across the board in terms of defense effectiveness. Compared to SSL, meta-adversarial training is 0.36% and 0.6% less accurate in the 5shot-5way case when facing RGF and P-RGF adversarial attacks, however, all other metrics are higher than SSL. Compared with NPR, it can be seen that NRP cannot effectively defend against black box attacks and its accuracy is lower than meta-adversarial training in both 1shot-5way and 5shot-5way cases.

Compared with CompareNets, SSL, and NRP defense methods, the proposed meta-adversarial training shows better overall defense performance for both white-box and black-box attacks. The reason is that CompareNets simply superimposes the feature maps directly during feature extraction, and the extracted feature information is destroyed, resulting in its poor adaptability and low accuracy. SSL adopts a self-supervised learning method for few shot learning. Self-supervision can effectively prevent the overfitting phenomenon and enhance the generalization ability and robustness of deep neural networks, so it can maintain stable robustness in the face of white-box attacks and black-box attacks. NRP effectively uses the

Table 7: Migrability of 1shot-5way defense against attacks (%)

Training	Test				
	BIM	C&W	DeepFool	DI^2FGSM	MDI^2FGSM
BIM	**0.6601**	0.5959	0.6068	0.6030	0.5728
C&W	0.6079	**0.6174**	0.6006	0.6048	0.5800
DeepFool	0.5844	0.5887	**0.5825**	0.5881	0.5512
DI^2FGSM	0.5937	0.5921	0.5918	**0.4761**	0.3893
MDI^2FGSM	0.5856	0.5841	0.5820	0.5556	**0.4604**

Table 8: Migrability of 5shot-5way defense against attacks (%)

Training	Test				
	BIM	C&W	DeepFool	DI^2FGSM	MDI^2FGSM
BIM	**0.8241**	0.7541	0.7661	0.7617	0.7351
C&W	0.7748	**0.7735**	0.7649	0.7660	0.7422
DeepFool	0.7554	0.7510	**0.7453**	0.7511	0.7166
DI^2FGSM	0.7604	0.7585	0.7577	**0.6507**	0.5141
MDI^2FGSM	0.7560	0.7512	0.7487	0.7293	**0.6213**

Table 9: Comparison of performance against white-box attacks (%)

	Meta adv-training		CompareNets		SSL		NRP
	1shot5way	5shot5way	1shot5way	5shot5way	1shot5way	5shot5way	
BIM	0.6601	0.8241	0.5894	0.7327	0.6416	0.8080	0.6631
C&W	0.6174	0.7735	0.4935	0.6510	0.5736	0.7677	0.6802
DeepFool	0.5825	0.7453	0.4887	0.6548	0.5553	0.7522	0.6813
DI^2FGSM	0.4761	0.6507	0.4953	0.6397	0.5722	0.6634	0.6152
MDI^2FGSM	0.4604	0.6213	0.4757	0.6361	0.4633	0.6529	0.5936

Table 10: Comparison of performance against white-box attacks (%)

	Meta adv-training		CompareNets		SSL		NRP
	1shot5way	5shot5way	1shot5way	5shot5way	1shot5way	5shot5way	
RGF	0.6031	0.7632	0.4772	0.6243	0.5818	0.7668	0.5359
P-RGF	0.5960	0.7583	0.4780	0.6420	0.5772	0.7643	0.3648
Parsimonious	0.5672	0.7259	0.4925	0.6471	0.5534	0.7039	0.4089

information contained in the feature space of deep neural networks for self-supervised learning, and the method can effectively defend against white box attacks, but is less effective in defending against black box attacks.

5 Conclusions

In this paper, we propose a deep neural network defense method based on meta-adversarial training to defend against the ever emerging adversarial attack methods, which combines meta-learning with adversarial training. First of all, adversarial training as an effective defense method against attacks, it can effectively defend against most of the adversarial attack methods, but its generalization ability in the face of emerging adversarial examples is poor and its robustness is low. To tackle this problem, the meta-learning method is used in the process of adversarial training to improve the robustness of deep neural networks using its better adaptability and generalization ability in few shot tasks. The experimental results show that the overall defense performance of the proposed defense method is stronger compared with other defense methods.

Acknowledgments

This research is supported by the National Natural Science Foundations of China (62166025); Science and technology project of Gansu Province(21YF5GA073); Gansu Provincial Department of Education: Outstanding Graduate Student "Innovation Star" Project(2021CXZX-511, 2021CXZX-512)

References

[1] S. Budd, E. C. Robinson, and B. Kainz, “A survey on active learning and human-in-the-loop deep learning for medical image analysis,” *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[2] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

[3] D. Chen, Y. Chen, Y. Li, F. Mao, Y. He, and H. Xue, “Self-supervised learning for few-shot image classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1745–1749. IEEE, 2021.

[4] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, “Improving black-box adversarial attacks with a transfer-based prior,” *Advances in neural information processing systems*, vol. 32, 2019.

[5] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[7] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” *arXiv preprint arXiv:1412.5068*, 2014.

[8] J. Ho, B.-G. Lee, and D.-K. Kang, “Attack-less adversarial training for a robust adversarial defense,” *Applied Intelligence*, vol. 52, no. 4, pp. 4364–4381, 2022.

[9] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, “Learning with a strong adversary,” *arXiv preprint arXiv:1511.03034*, 2015.

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.

[11] H. Kwon and J. Lee, “Diversity adversarial training against adversarial attack on deep neural networks,” *Symmetry*, vol. 13, no. 3, p. 428, 2021.

[12] K. Li, Y. Zhang, K. Li, and Y. Fu, “Adversarial feature hallucination networks for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13479, 2020.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[14] R. Mahfuz, R. Sahay, and A. ElGamal, “Mitigating gradient-based adversarial attacks via denoising and compression,” *arXiv preprint arXiv:2104.01494*, 2021.

[15] D. Mandal, S. Medya, B. Uzzi, and C. Aggarwal, “Metalearning with graph neural networks: Methods and applications,” *ACM SIGKDD Explorations Newsletter*, vol. 23, no. 2, pp. 13–22, 2022.

[16] D. Meng and H. Chen, “Magnet: a two-pronged defense against adversarial examples,” in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 135–147, 2017.

[17] A. Mirzaeian, J. Kosecka, H. Homayoun, T. Mohsenin, and A. Sasan, “Diverse knowledge distillation (dkd): A solution for improving the robustness of ensemble models against adversarial attacks,” in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, pp. 319–324. IEEE, 2021.

[18] S. Moon, G. An, and H. O. Song, “Parsimonious black-box adversarial attacks via efficient combinatorial optimization,” in *International Conference on Machine Learning*, pp. 4636–4645. PMLR, 2019.

[19] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: A simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[20] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, “A self-supervised approach for adversarial robustness,” in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 262–271, 2020.

[21] F. Nesti, A. Biondi, and G. Buttazzo, "Detecting adversarial examples by input transformations, defense perturbations, and voting," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[22] F. Nezhadalinaei, L. Zhang, M. Mahdizadeh, and F. Jamshidi, "Motion object detection and tracking optimization in autonomous vehicles in specific range with optimized deep neural network," in *2021 7th International Conference on Web Research (ICWR)*, pp. 53–63, IEEE, 2021.

[23] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[24] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, 2018.

[25] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

[26] C. Tian, L. Wang, E. Zhou, K. Liu, S. Du, and J. Liu, "Integration and experimental study of automatic driving system for bus," in *2021 7th International Symposium on Mechatronics and Industrial Informatics (ISMII)*, pp. 96–103. IEEE, 2021.

[27] H. Veeraraghavan and J. Jiang, "Deep learning from small labeled datasets applied to medical image analysis,". in *State of the Art in Neural Networks and their Applications*, pp. 279–291, Elsevier, 2021.

[28] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[29] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 501–509, 2019.

[30] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

[31] J. Xiong, D. Yu, S. Liu, L. Shu, X. Wang, and Z. Liu, "A review of plant phenotypic image recognition technology based on deep learning," *Electronics*, vol. 10, no. 1, p. 81, 2021.

[32] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[33] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *International conference on machine learning*, pp. 11278–11287. PMLR, 2020.

[34] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018.

Biography

You-kang Chang was born in 1994. He is a doctor student at Lanzhou University of Technology. His major research field is adversarial attacks and defense adversarial attacks. E-mail: 2507576651@qq.com.

Hong Zhao was born in 1971. He is a professor and a supervisor of doctor student at Lanzhou University of Technology. His major research field is System modeling and simulation, deep learning, natural language processing. E-mail: zhaoh@lut.edu.cn.

Wei-jie Wang was born in 1994. She is a doctor student at Lanzhou University of Technology.Her major research field is speaker recognition. E-mail: 1132744259@qq.com

A Hybrid-based Feature Selection Method for Intrusion Detection System

Xibin Sun[1], Heping Ye[1], and Xiaolin Liu[1],
(Corresponding author: Xibin Sun)

Guangdong Polytechnic of Science and Technology, Zhuhai, China[1]
Zhuhai 519090,China
Email: jacky5555@qq.com

(Received Aug. 11, 2022; Revised and Accepted Dec. 5, 2022; First Online Dec. 13, 2022)

Abstract

As we know, feature selection can improve the performance of machine learning algorithms for intrusion detection. This paper proposes a hybrid feature selection method, which ranks features according to two factors: relevancy and redundancy, and then adopts the forward search strategy to select the optimal feature subset from the ranked features. Experiments on the KDDCup'99 dataset showed that our proposed feature selection method could get better performance on the accuracy rate and false positive rate in intrusion detection compared with other feature selection approaches.

Keywords: Feature Selection; Hybrid; Intrusion Detection; Performance

1 Introduction

The intrusion detection system (IDS) as a part of network security infrastructure, can detect network attacks or abnormal behaviors. Traditional IDS can be categorized into two types: Signature-based IDS and Anomaly-based IDS. Signature-based IDS is good at detecting known network attacks and suffers from unknown or novel attacks. Anomaly-based IDS can detect novel attacks, yet it usually owns a high false positive rate. Therefore, there exist challenges in the traditional IDS when they are deployed into the real-world network environment. At the same time, as machine learning (ML) methods are applied to different fields successfully, many researchers introduce them into the intrusion detection domain for detecting network attacks. The authors [19] introduced ML algorithms to build classification models from the network datasets to predict the network attacks, such as the support vector machine (SVM) algorithms, the decision tree (DT) algorithms, the random forest (RF) algorithms, deep learning algorithms, and so on.

Though these built IDS by using ML methods can obtain better results in detecting network attacks, they often suffer from the high dimensional and massive network traffic data. Furthermore, the features of large-scale network traffic often contain redundancy, incompleteness, and irrelevance, which not only declines the performance of ML algorithms but also adds the time and complexity of building classification models. Therefore, it is significant to select the optimal feature subset from the initial network datasets before using ML algorithms to build classification models. In this paper, we have designed a hybrid-based feature selection method that consists of two phases. In the first phase, the filter method is used to sort the features from the original feature space of the network datasets according to their relevancy and redundancy. In the second phase, the wrapper feature selection approach is used to select the final feature subset from the sorted features of the first phase. To be specific, we start with the first feature of the sorted features and incrementally add a feature to the wrapper method one by one, and the feature subset that can make the classification model obtain the highest accuracy rate in detecting network attacks will retain the final selected features. The contributions of this paper are listed as follows:

1) We proposed a hybrid feature method that integrates the high efficiency of the filter feature selection method and the ability to extract optimal features of the wrapper feature selection method. The experimental results on the KDDCup'99 dataset showed our proposed method could get better performance than other relevant feature selection methods.

2) For other feature selection methods, such as MIFS [8], MIFS-U [16], MMIFS [7], and LCC-RF-RFEX [24], when they are used to select a feature subset from the initial feature space, the specific threshold or the number of final selected features need to be provided in advance. However, these values are often set based on the human experience, which is hard to set the optimal values when facing different network datasets. Our approach differs from the above feature selection methods, which don't provide a specific threshold or the selected fea-

ture number.

3) As we know, the traditional wrapper feature selection methods usually adopt a greedy search approach by evaluating all the possible combinations of features against the specific machine learning algorithm. Therefore, it is a time-consuming process when directly applying the wrapper methods to the initial features for feature selection. In the final phase of our approach, we use the wrapper method to select the final features from the sorted features rather than the initial features, which avoids the computational disaster of feature selection in the wrapper method and improves the efficiency of wrapper methods.

The rest of this paper is structured as follows. Section 2 presents the related works on feature selection approaches. Section 3 introduces our proposed feature selection method in detail. The analysis of experimental results shows in Section 4. Finally, Section 5 summarizes the works in this paper and points out the future works.

2 Related Works

The feature selection methods can reduce the time to build classification models or improve the accuracy of classification models. So far, we can divide the feature selection approaches into three categories: filter, wrapper, and hybrid. Filter methods mainly select feature subsets using the specific heuristic evaluation function to measure the relevance of features. Wrapper methods often adopt a classification algorithm to train a model to estimate the optimal feature subset. Therefore, filter methods are much faster than wrapper methods as they do not involve training models. However, the final selected features using the wrapper methods can often make the classification models obtain better performance than those using the filter methods. Hybrid methods often have the best performance by integrating the advantages of filter methods and wrapper methods.

The authors [5, 7, 8, 11, 14, 16, 20, 23, 24, 29] introduced the filter feature selection methods, such as [5, 7, 8, 16, 20, 24, 29] mainly presented the correlation-based feature selection (CFS) method for feature selection. At present, Linear Correlation Coefficient (LCC) and Mutual Information are the two main heuristic evaluation functions to evaluate the correlation between two random variables. For example, The authors [11] used the LCC function to measure the relevance between two features, then, ranked the features according to the calculated correlation values, and finally, selected the final feature subset by removing the features of which correlation values are below the specified thresholds. Similarly, The authors [5, 7, 8, 16, 20, 24, 29] introduced how to use Mutual Information (MI) methods to select features. The authors [8] first provided the mutual information method feature selection (MIFS) which maximizes the relevance between feature and class label and minimizes the redundancy of the selected features. MIFS-U [16], MMIFS [7], FMIFS [5], mRMR [20], and RPFMI [29] were all based on the improvement of MIFS. In MIFS-U [16], MMIFS [7], and mRMR [20], their feature selection algorithms need to provide the specific threshold as the input parameter before using them to select features. Though FMIFS [5] and RPFMI [29] overcome the above limitation, they belong to the filter method which mainly uses statistical techniques to evaluate the intrinsic relationship of features (i.e., the relevance and redundancy), and the final selected features are independent of the learning algorithm, which leads to the built IDS owning a lower detection accuracy.

The authors [1–4,9,10,13,21,22,25] introduced wrapper methods to select features. Such as, the authors [25] presented the wrapper method to select features, and SVM algorithm is used to build IDS based on the final selected features. Experimental results showed that the build IDS achieved 82.34% accuracy rate. The authors [22] used C4.5 tree and BN algorithms to select features, and got the higher accuracy rate and the lower false positive rate for the four types of attacks (Dos, Probe, R2L, and U2R) respectively by comparing the full 41 features. The authors [2] provided a new feature selection algorithm based on pigeon inspired optimizer(PIO) for IDS, and the experimental results showed that the PIO feature selection algorithm not only reduced the number of features of KDDCup'99, NSL-KDD, and UNSW-NB15 datasets respectively, but also maintained a high accuracy rate and reduced the required time for training the classification models significantly.

The authors [6,17,18,27] demonstrated the hybrid feature selection methods to select features. Such as, the authors [6] used the filter method to eliminate the irrelevant and redundant features from the initial feature space and then, the remained features were fed to the wrapper method LS-SVM to select the final feature subset. Experiments showed that the proposed method could get the 98.9% accuracy rate classification accuracy. The authors [17] designed the hybrid method: FGLCC-CFA, which combined the filter FGLCC method and the wrapper method CFA. It first used the FGLCC to rank the initial features and select the opimal feature subset, and then, the feature subset was input to the CFA method to select the final features. Experimental results showed the FGLCC-CFA method got a higher accuracy rate and detection rate equal to 95.03% and 95.23%, respectively, and a lower false positive rate of 1.65% compared with the filter FGLCC method and the wrapper CFA method.

3 Proposed Feature Selection Method

Through the analysis of the above-related literate, we find the current feature methods may exist the following defects:

1) Many feature selection methods exist a limit that

needs to specify the threshold or the number of final selected features before using them to select the final feature subset. Such as, the authors [7,8,16,24] needs to set a specific value for the redundancy parameter in their feature selection algorithms. Yet, there is no empirical value for the parameter, and how to set an appropriate value for the parameter is still a vexing question to answer, especially when facing tasks in different domains and different datasets.

2) For the wrapper methods, evaluating all the possible combinations of features by using the machine learning algorithms from the initial feature space is often a time-consuming process, which is called an NP-complete problem [15], especially for high-dimension feature space.

To overcome the aforementioned problems, we proposed a hybrid feature selection method that contains two phases. In the first phase, we use mutual information to rank the features by comprehensively considering their relevance and redundancy. In the second phase, we adopt the forward search strategy (FSS) to incrementally select features from the sorted feature set and then feed them to the specific classification algorithm to count the accuracy rate (AR). The feature subset which gets the maximum AR will be retained as the final selected features. Different from other filter approaches, our approach only ranks the initial features rather than selects features, so there is no need for setting the specific threshold beforehand. Furthermore, in the second phase of our approach, we use a forward search strategy (FSS) to select a feature subset from the ranked feature space rather than the initial feature space, which effectively avoids the NP-complete problem of the feature combination in the wrapper methods. The workflow of our approach is shown in Figure 1.

3.1 Mutual Information

We use mutual information [8] as a heuristic evaluation function to rank the features in our proposed approach. As we know, mutual information is widely used to measure the relevance between random variables. If two random variables U={u_1,u_2,...,u_n } and V={v_1,v_2,...,v_n } belong to the discrete variables, where n is the total number of samples, the mutual information (MI) of the two variables is defined as shown in Equation (1) [8]:

$$I(U;V) = \sum_{u\in U}\sum_{v\in V} p(u,v)\log\frac{p(u,v)}{p(u)p(v)}. \tag{1}$$

Where p(u) and p(v) are the probability distribution of U and V separately. p(u,v) is a joint probability distribution. For continuous variables, the MI is defined as shown in Equation (2) [8]:

$$I(U;V) = \int_u\int_v log\frac{p(u,v)}{p(u)p(v)}dudv. \tag{2}$$

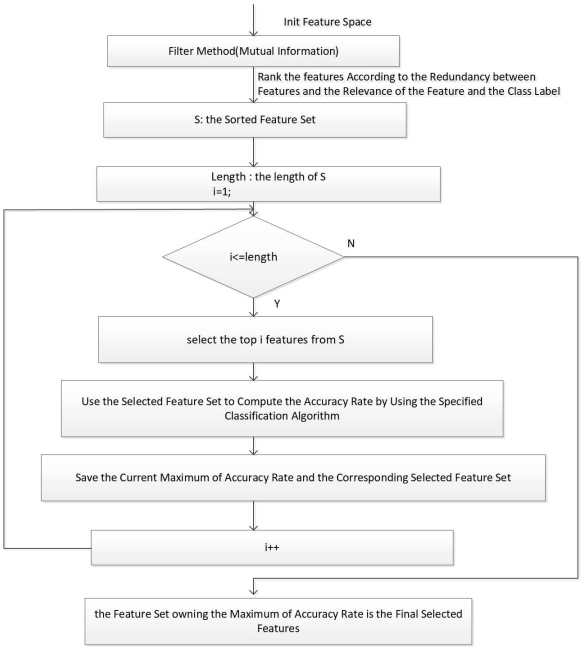

Init Feature Space
Filter Method(Mutual Information)
Rank the features According to the Redundancy between Features and the Relevance of the Feature and the Class Label
S: the Sorted Feature Set
Length : the length of S
i=1;
i<=length
N
Y
select the top i features from S
Use the Selected Feature Set to Compute the Accuracy Rate by Using the Specified Classification Algorithm
Save the Current Maximum of Accuracy Rate and the Corresponding Selected Feature Set
i++
the Feature Set owning the Maximum of Accuracy Rate is the Final Selected Features

Figure 1: The workflow of our feature selection approach

The MI value is larger, which presents that the two variables are closely related, and A zero value of MI indicates that the two variables are independent.

3.2 Proposed Feature Selection Algorithm

Our proposed feature selection method mainly contains two main phases: the first phase in which the filter method is used for feature ranking. Different from other filter methods aiming at feature selection, our approach mainly uses mutual information to rank the features according to the redundancy of features and the relevance of the feature and the class label. Inspired by [5], we use Equation (3) to decide the position of a feature in the final sorted feature subset.

$$G_{MI} = \text{argmax}_{f_i\in F}\left(I\left(C;f_i\right) - \frac{1}{|S|}\sum_{f_s\in S} MR\right) \tag{3}$$

Where F is the original feature set of datasets, f_i is the candidate feature, S is the sorted feature set from the original feature set F, f_s is the sorted feature in S, $|S|$ is the number of the final sorted features in S and C is the class label, MR in Equation (3) is the relative redundancy of feature f_i against feature f_s. MR is defined by Equation (4) [5]:

$$MR = \frac{I\left(f_i;f_s\right)}{I\left(C;f_i\right)} \tag{4}$$

Where I(C; f_i) is the mutual information value between the candidate feature f_i and the class C, and I(f_i; f_s) is the mutual information value between the candidate feature f_i and the sorted feature f_s. Equation (3) is intended to select a feature f_i from the F that maximizes I(C; f_i) and minimizes the average of redundancy MR simultaneously.

In the second phase, we use the wrapper method to select the feature subset from the ordered feature set S which is coming from the first phase of our proposed approach. In the second phase, we adopt the forward search strategy (FSS) to incrementally select features from the sorted feature set S, and then feed them to the specific classification algorithm to count the accuracy rate. The final feature subset which can get the maximum accuracy rate will be retained. The pseudo-code of our proposed feature selection method is shown in Algorithm 1.

Algorithm 1 The Proposed Feature Selection Algorithm

```
Input:
    $F$: Feature set $F = \{f_i | i = 1, .., n\}$
    $A$: The Specific Classification Algorithm(e.g. Decision Tree, Naive Bayes, Support Vector Machine, etc)
Output:
    $maxS$: The Final Selected Feature Subset
Begin
$S \leftarrow \varnothing$
Calculate I(C; $f_i$), for each feature $f_i$, i=1,..,n, C notes the class label.
if I(C;$f_i$)==0 then
    $F \leftarrow F \setminus \{f_i\}$
end if
Select the feature $f_i$: $f_i \in F$ that maximizes I(C; $f_i$).

$S \leftarrow S \cup \{f_i\}$
$F \leftarrow F \setminus \{f_i\}$
while $F \neq \varnothing$ do
    select the feature $f_i$ using Equation (3)
    $S \leftarrow S \cup \{f_i\}$
    $F \leftarrow F \setminus \{f_i\}$
end while
$maxAR \leftarrow 0$
$maxS \leftarrow \varnothing$
$length \leftarrow$ the length of S
for $i = 1; i \leq length; i++$ do
    $S_{sub} \leftarrow$ select the top i features from S
    Count Accurate Rate (AR) of the classification algorithm A by using $S_{sub}$
    if $AR > maxAR$ then
        $maxAR \leftarrow AR$
        $maxS \leftarrow S_{sub}$
    end if
end for
return maxS
End
```

4 Experiments and Results

4.1 Datasets for Evaluation

KDDCup'99 dataset [26] is one of the datasets for evaluating intrusion detection. It contains 39 attack types divided into four categories: Dos, Probe, U2R, and R2L. Furthermore, it also provides the training and test datasets for evaluating the machine learning algorithms. The training dataset contains about five million connection records, and the test dataset includes around two million records. Each connection record that contains 41 features is labeled as either normal or an attack. Considering that there are a large number of redundant records and the imbalance of the distribution of attack records in the KDDCup'99 dataset, We selected partial data from the KDDCup'99 dataset to generate the corresponding training dataset and test dataset for each of the four attack categories. Details of the generated datasets are shown in Table 1.

4.2 Performance Metrics

In this paper, we mainly use two metrics to evaluate the performance of our proposed feature selection method, and the performance metrics are accuracy rate (AR) and false positive rate (FPR) separately. AR can be formally defined as:

$$AR = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{5}$$

FPR is defined as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{6}$$

4.3 Experimental Results and Analysis

Python language is used to realize our proposed feature selection approach, and all the experiments were performed on a Windows platform having configuration i5 core 4 CPU 2.3 GHz, 8GB RAM. Table 2 shows the final selected features of four types of attacks by using our approach based on the decision tree algorithm. Figure 2 and Figure 3 show the AR and FPR of the built IDS based on the selected features by using our proposed approach and full features(41) separately. The results demonstrate that the IDS based on the selected features can achieve better performances in DR and FPR metrics by comparing IDS constructed with the full features(41). Furthermore, we also test the total consuming time(training and test times) of the built IDS by using selected features and the full features(41) separately. Table 3 shows that the IDS based on selected features consumes less time than IDS based on the full features. This is principally because our proposed approach deletes the redundant and irrelevant features from the full features, which causes not only to reduce the total consuming time of classification models but also to improve their performance.

Table 1: Sample Distributions of Instances for Four Attack Types in Datasets

Attack Type	**Attack Name**	**Training Data**	**Test Data**
Dos	*normal*	20000	20000
	smurf	10000	10000
	neptune	5000	5000
	mailbomb	1500	1500
	back	500	500
	land	15	15
	teardrop	400	400
	processtable	350	350
	pod	100	100
	aparche2	250	250
	SubTotal	*training dataset:38115*	*test dataset:38115*
Probe	*normal*	10000	10000
	ipsweep	1247	306
	mscan	600	400
	nmap	130	100
	portsweep	540	500
	saint	400	300
	satan	800	600
	SubTotal	*training dataset:13717*	*test dataset:12206*
U2R	*normal*	10000	10000
	buffer_overflow	30	22
	httptunnel	158	158
	loadmodule	9	2
	perl	3	2
	rootkit	10	13
	SubTotal	*training dataset:10210*	*test dataset:10197*
R2L	*normal*	20000	20000
	ftp_write	8	3
	guess_passwd	53	4367
	imap	12	1
	multihop	8	19
	phf	6	3
	warezclient	1021	1021
	warezmaster	21	1603
	SubTotal	*training dataset:21129*	*test dataset:27017*

Table 2: Selected Features by Using Proposed Approach based on Decision Tree Algorithm

Attack Type	**Selected Features**
Dos	5, 37, 23, 3, 31, 12, 25, 36, 2, 6, 26, 16, 32, 13, 24, 39, 8
Probe	5
R2L	5,22,11
U2R	5, 14, 17, 13, 40, 18, 10, 11, 27, 15, 9, 16, 41

In addition, we also compared the performance of our approach with the filter methods, such as the linear correlation-based feature selection (LCFS) algorithm [7], the mutual information-based feature selection (MIFS) algorithm [8], and the wrapper methods, such as the Random Forest-Recursive Feature Elimination (RF-RFE) [12]. As shown in Figure 4, compared to other filter and wrapper methods, our approach has a higher AR, which indicates that the IDS based on the hybrid feature selection method has better performance than IDS based on a single filter method or wrapper method. Moreover, as mentioned in section III, unlike other wrapper methods based on initial features, we use the wrapper method on the ranked features to select the final feature subset in

Table 3: The Total Consume Time(training time and test time) of Decision Tree Algorithm Based on Selected Features and Full Features(41)

Attack Type	**Total Consume Time(s)**	
	Selected Features	**Full Features(41)**
Dos	0.21875	0.29688
Probe	0.03125	0.12500
R2L	0.09375	0.28125
U2R	0.03125	0.06250

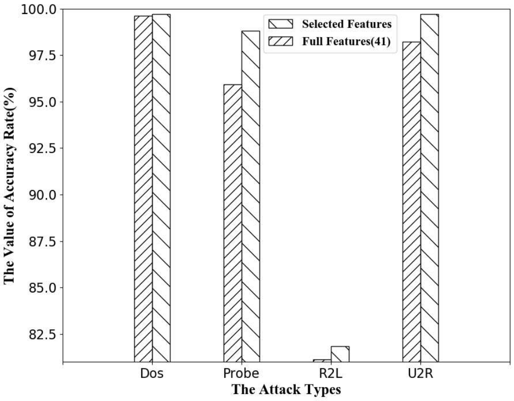


Figure 2: The Accuracy Rate(AR) of Decision Tree Algorithm With Selected Features and Full Features(41)

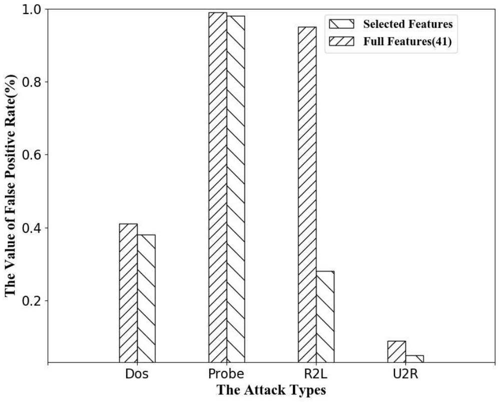


Figure 3: The False Positive Rate(FPR) of Decision Tree Algorithm With Selected Features and Full Features

the second phase of our proposed method. Table 4 shows that wrapper methods based on sorted features have more efficient time performance than wrapper methods based on the original features.

Furthermore, we evaluated the IDS based on our feature approach and the recent hybrid feature selection

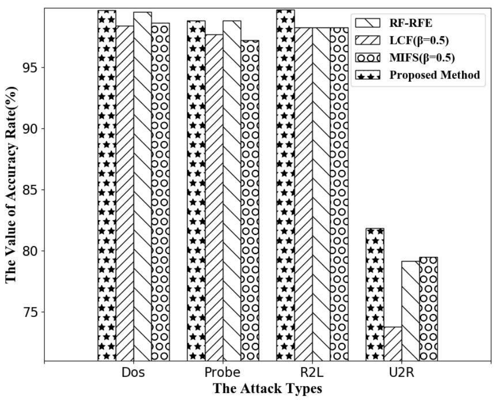


Figure 4: Accuracy Rate of Decision Tree Algorithm With non-hybrid Feature Selection Methods and Proposed Method

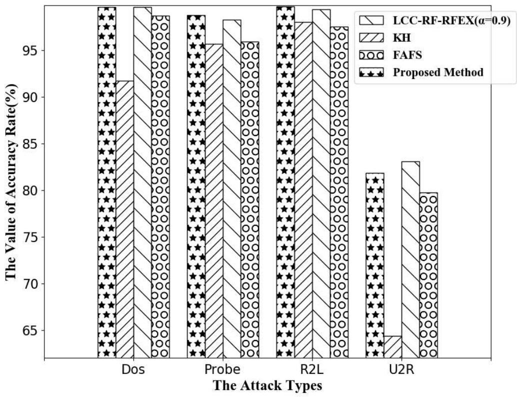


Figure 5: Accuracy Rate of the Classification Model based on Decision Tree Algorithm With Hybrid-based Feature Selection Methods

methods, such as the LCC-RF-RFEX [24], KH [28], and FAFS [22] methods. Figure 5 shows the accuracy rate of classification models based on the decision tree algorithm with hybrid feature selection methods. Experimental results show that our proposed approach outperforms

Table 4: Consume Time of Selecting Feature Subset

Attack Type	**Consume Time(s)**	
	Wrapper Method based on Sorted Features **(Proposed Method)**	Wrapper Method based on Original Features **(RF-RFE)**
Dos	7.26562	7.56250
Probe	2.64062	4.54688
U2R	2.23438	2.90625
R2L	4.75000	7.75000

Table 5: Accuracy Rate of Our Feature Selection Method based On Different Classification Algorithms

Classification Algorithm	**Dos**	**Probe**	**U2R**	**R2L**
Random Forest	99.89%	98.25%	99.37%	83.11%
Naive Baye	90.99%	95.69%	98.03%	64.35%
Multi perceptron	98.61%	95.91%	97.54%	79.75%
Support Vector Machine	97.69%	97.65%	98.37%	77.29%
Decision Tree	99.62%	98.80%	99.70%	81.84%
Logistic Regression	93.73%	98.30%	98.77%	78.65%

these hybrid-based feature selection methods (except for the R2L attack type). This is mainly because we adopt the forward search strategy (FSS) to incrementally select features from the sorted feature set. The feature subset which gets the maximum AR will be saved as the final selected feature subset. Finally, we evaluate the performance of our feature selection method based on different classification algorithms. Table 5 shows that the IDS based on the Random Forest (RF) and Decision Tree (DT) can obtain better AR by comparing with the IDS based on other classification algorithms.

5 Conclusions and Future Work

This paper proposed a hybrid feature selection method for intrusion detection, which absorbs the advantages of the filter feature selection methods and the wrapper methods. Different from other filter feature selection methods, we use mutual information to rank the original features rather than select features. Unlike other wrapper approaches, we use the wrapper method to select a feature subset from the sorted features rather than initial features. Furthermore, we adopt the forward search strategy (FSS) to incrementally select features from the sorted feature set and then feed them to the specific classification algorithm to count the AR. The feature subset which gets the maximum AR will be retained as the final selected subset. Therefore, there is no need to specify the threshold or the number of the final selected features in advance when using our approach to select the final feature subset. Experimental results on the KDDCup'99 dataset showed that our approach could achieve better performance compared with other related feature selection methods.

So far, we only finish selecting the optimal feature subset from the labeled datasets. However, for unlabeled network traffic, how to use unsupervised technology to select the features of attacks will be considered in our future studies.

Acknowledgments

This study is mainly supported by the Guangzhou Science and Technology Program Project “Research on Feature Engineering Technology and Machine Learning algorithm Applied in the Intrusion Detection System” (Project No.: 202102080586). Furthermore, It is also supported in part by the Guangdong Province Educational Science Planning Project “Research on the Application Mode and Practice of VR Technology in Vocational Education” (Grant No.: 2018GXJK318).

References

[1] W. L. AI-Yaseen, A. K. Idrees, and F. H. Almasoudy, “Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system”, *Pattern Recognition*, vol. 132, no. 12, pp. 1–10, 2022.

[2] H. Alazzam, A. Sharieh, and K. E. Sabri. “A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer”, *Expert Systems With Applications*, vol. 148, no. 15, pp. 1–13, 2020.

[3] F. H. Almasoudy, W. L. Al-Yaseen, and A. K. Idrees. “Differential Evolution Wrapper Feature Selection

for Intrusion Detection System", *Procedia Computer Science*, vol. 167, no. 2, pp. 1230–1239, 2020.

[4] O. Almomani. "A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms", *Symmetry*, vol. 12, no. 6, pp. 1–20, 2020.

[5] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan. "Building an intrusion detection system using a filter-based feature selection algorithm", *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 2986–2998, 2016.

[6] M. A. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, and U. T. Nagar. "A Novel Feature Selection Approach for Intrusion Detection Data Classification", in *IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 82–89, Beijing, China, Sep 2014.

[7] F. Amiri, M. Mahdi, R. Yousefi, and A. Shakery. "Mutual information-based feature selection for intrusion detection systems", *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011.

[8] R. Battiti. "Using mutual information for selecting features in supervised neural net learning", *IEEE Transactions on Neural Networks*, vol. 5, no. 5, pp. 537–550, 1994.

[9] M. S. Bonab, A. Ghaffari, F. S. Gharehchopogh, and P. Alemi. "A wrapper-based feature selection for improving performance of intrusion detection systems", *International Journal of Communication Systems*, vol. 33, no. 12, pp. 1–25, 2020.

[10] K. Bouzoubaa, B. Nsiri, and Y. Taher. "Predicting DOS-DDOS Attacks: Review and Evaluation Study of Feature Selection Methods based on Wrapper Process", *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 131–145, 2021.

[11] H. F. Eid, A. E. Hassanien, T. H. Kim, and S. Banerjee. "Linear correlation-based feature selection for network intrusion detection model", in *International Conference on Security of Information and Communication Networks*, pp. 240–248, Cairo, Egypt, Sep 2013.

[12] B. Gregorutti, B. Michel, and B.P. Saint-Pierre. "Correlation and variable importance in random forests", *Stat Comput*, vol. 27, no. 3, pp. 659–678, 2017.

[13] M. Hasan, M. Nasser, and K. Molla. "Feature Selection for Intrusion Detection Using Random Forest", *Journal of Information Security*, vol. 7, no. 3, pp. 129–140, 2016.

[14] S. M. Kasongo and Y. X. Sun. "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset", *Journal of Big Data*, vol. 7, no. 105, pp. 1–20, 2020.

[15] B. Kumari and T. Swarnkar. "Filter versus wrapper feature subset selection in large dimensionality micro array: a review", *International Journal of Computer Science and Information Technologies*, vol. 2, no. 2, pp. 1048–1053, 2011.

[16] N. Kwak and C-H. Choi. "Input feature selection for classification problems", *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[17] S. Mohammadi and H. Mirvaziri. "Cyber intrusion detection by combined feature selection algorithm", *Journal of Information Security and Applications*, vol. 44, no. 8, pp. 80–88, 2019.

[18] H. Mohammadzadeh and F. S. Gharehchopogh. "A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection", *Computational Intelligence*, vol. 37, no. 1, pp. 176–209, 2021.

[19] U.S.Musa, M.Chhabra, A.Ali, and M.Kaur. "Intrusion Detection System using Machine Learning Techniques: A Review", in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 149–155, Trichy, India, Sep 2020.

[20] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information criteria of maxdependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[21] F.Salo, A. B. Nassif, and A. Essex. "Dimensionality reduction with IGPCA and ensemble classifier for network intrusion detection", *Computer Networks*, vol. 148, no. 15, pp. 164–175, 2019.

[22] B. Selvakumar and K. Muneeswaran. "Firefly algorithm based Feature Selection for Network Intrusion Detection", *Computers Security*, vol. 81, no. 2, pp. 148–155, 2019.

[23] M. A. Siddiqi and W. Pak. "Optimizing filter-based feature selection method flow for intrusion detection system", *Electronics*, vol. 9, no. 12, pp. 1–18, 2020.

[24] X. B. Sun, D. Zhang, H. O. Qin, and J.H.Tang. "Bridging the Last-Mile Gap in Network Security via Generating Intrusion-Specific Detection Patterns through Machine Learning", *Security and Communication Networks*, vol. 2022, no. 1, pp. 1–20, 2022.

[25] K. A. Taher, B. M. Y. Jisan, and M. M. Rahman. "Network intrusion detection using supervised machine learning technique with feature selection", in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, pp. 643–646, Dhaka, Bangladesh, Jan 2019.

[26] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set", in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, Ottawa, ON, Canada, Jul 2009.

[27] N. A. Umar, Z. F. Chen, and Y. Liu. "Network Intrusion Detection Using Wrapper-based Decision Tree for Feature Selection", in *In Proceedings of the 2020 International Conference on Internet Computing for Science and Engineering*, pp. 5–13, Male, Maldives, Jan 2020.

[28] L. Xin, Y. Peng, J. Yiming, and L. Tian. "LNNLS-KH: A Feature Selection Method for Network Intrusion Detection", *Security and Communication Networks*, vol. 2021, no. 1, pp. 1–22, 2021.

[29] F. Zhao, J. Y. Zhao, X. X. Niu, and Y. Xin. "A Filter Feature Selection Algorithm Based on Mutual Information for Intrusion Detection", *Applied Sciences*, vol. 8, no. 9, pp. 15–35, 2018.

Biography

XiBin Sun received his Ph.D of Computer Technology and Application in 2022 from Faculty of Information Technology, Macau University of Science and Technology, Macau, China. He is a lecturer in the computer science department of Guangdong Polytechnic of Science and Technology. His current research involves the machine learning and network intrusion detection.

HePing Ye received his Ph.D. in Technology of Computer Application in 2012 from the South China University of Technology, China. He is a lecturer in the computer science department of Guangdong Polytechnic of Science and Technology. His current research involves data mining and network security.

XiaoLin Liu received his master's degree in Software Engineering in 2006 from the South China University of Technology, China. He is a senior engineer in the computer science department of Guangdong Polytechnic of Science and Technology. His current research involves information hiding and network security.

An Efficient Attribute Encryption Scheme with Privacy-Preserving Policy in Smart Grid

Jing Cheng and Mi Wen
(Corresponding author: Jing Cheng)

College of Computer Science and Technology, Shanghai University of Electric Power
Shanghai, 200090, China
Email: 18355617975@163.com

(Received Aug. 15, 2022; Revised and Accepted Dec. 5, 2022; First Online Dec. 13, 2022)

Abstract

The smart grid is a new generation of power systems. The power information is generally collected through smart meters installed in homes and power monitors deployed outdoors; this massive amount of information will be stored in a smart cloud. Unfortunately, cloud computing scenarios are prone to leaking private information, hurting security. Ciphertext Policy-Attribute Based Encryption (CP-ABE) is a vital encryption technology for access control in smart grid scenarios. However, the access policy in attribute encryption is generally in plain text, which risks leaking sensitive information and requires a large amount of computation. This paper proposes an efficient and more privacy-preserving weighted attribute encryption method based on a ciphertext strategy for the smart grid. We completely hide the access policy, adopt online/offline encryption, and outsource decryption to reduce the amount of computation. We use the analytic hierarchy process (AHP) to assign appropriate weights to each attribute, providing fine-grained and flexible access control by adopting the Linear Secret Sharing Scheme (LSSS) access control mechanism. The analysis shows that while ensuring security, our scheme will also reduce the computational and communication overhead and is closer to the actual situation, which is suitable for application in cloud computing.

Keywords: Attribute-based Encryption; Decryption Outsource; Large Universe; Online/Offline Encryption; Policy Hidden

1 Introduction

The smart grid is the product of the rapid development of cloud computing and the Internet of Things (IoT), which has laid a solid foundation for the development of high-precision and high-efficiency power information technology in the future. It is based on an integrated, high-speed two-way communication network, through the application of advanced sensing and measurement technology, advanced equipment technology, advanced control methods, and advanced decision support system technology [3]. Smart grids rely on smart meters deployed outside to collect data. A large amount of data is transmitted every day across different regions.

When outsourcing power information to the smart cloud, the data owner loses control over the data. Also, cloud service providers are not fully trusted by users, which makes access control more challenging. Users may worry that cloud servers may make incorrect access decisions, intentionally or unintentionally, and leak their data to some unauthorized users. To enable users to control access to their data, some attribute-based access control schemes [13, 19, 22], have been proposed by utilizing attribute-based encryption [2, 5, 11]. In attribute-based access control, end users first define access policies for their data, and data is encrypted under these access policies. Only users whose attributes satisfy the access policy are eligible to decrypt data.

Data security is very important in the smart grid environment, and different responsible departments should collect power information in the corresponding area. But in most conventional schemes, the access policy is attached to the ciphertext in plaintext. To prevent privacy leakage in the access policy, a straightforward approach is to completely hide the attribute in the access policy. However, when attributes are hidden, not only unauthorized users but also authorized users to have no way of knowing which attributes are involved in the access policy, making decryption a challenging issue. For this reason, existing schemes do not hide or anonymize properties. Instead, they only hide the value of each attribute by using wildcards [26, 28], hidden vector encryption [1], and inner product encryption [8]. But hidden property values are only partially hidden methods, and property names also leak information.

Another issue in the smart grid scenario is that each department can be divided into different levels in the real scenario, such as the power grid city bureau, urban bureau, county bureau, substation, etc. If we intend to attach the overall situation to the access structure, the access structure will become relatively complex. We can grant the weights 1, 2,..., and n, respectively, indicating that the county bureau is the power grid company (supposed weight is 1) and the municipal bureau is the power grid company (supposed weight is 3). The power grid company of the municipal bureau is higher than the county bureau company, so you can obtain data that the county bureau does not have permission to obtain. In [14], Liu applied it to the real scene but did not give specific steps. In a word, there are some common problems in the existing schemes:

1) The current smart grid CP-ABE scheme has the problem of privacy leakage. The access policy in plain text will leak information, and malicious users can easily obtain sensitive data from the access policy illegally;

2) The computational overhead cost is fairly high for data owners and data users to encrypt and decrypt data;

3) Attributes in smart grid scenarios should be weighted to reflect their relevance and significance.

In this paper, we introduce an attribute-based encryption scheme based on weighted attributes that not only supports complete hiding but also introduces online/offline encryption and outsourced decryption and, at the same time, realizes a large attribute domain, which greatly reduces the communication cost and computing cost of this scheme. The contributions of this paper are summarized as follows:

1) In order to fully protect the privacy of the access policy in the scheme, this paper proposes a security-improved attribute, Bloom Filter (si-ABF), to reduce its false positive probability, thereby realizing the complete hiding of the access policy.

2) Utilize the methods of online/offline encryption and outsource decryption to reduce the computational overhead in the scheme. The concept of the weighted large universe is introduced, which reduces the communication overhead of the system and improves the flexibility of attribute expression in the system.

3) Analyze the security requirements and privacy protection capabilities of the proposed scheme, indicating that our scheme is secure in private information. Through performance experimental evaluation, our scheme can protect the privacy of any LSSS access policy and greatly reduce the computational overhead.

The rest of the paper is organized as follows: Sections 2 and 3 demonstrate related work and relevant preliminary knowledge. Sections 4 and 5 present the complete system solution. Section 6 will show the overhead and efficiency of this scheme. Finally, we will give general conclusions.

2 Related Work

2.1 Basic Classification of ABE

Attribute-Based Encryption (ABE) [1] is one of the most classic encryption methods. Sahai and Waters first proposed the concept of ABE, which is a public key cryptosystem with a one-to-many algorithm to protect data in the cloud. Here, the encryption of data is based on attribute sets. Its roots can be traced back to Identity-Based Encryption (IBE) [20], which can be seen as a special case of IBE. In addition, ABEs are classified into Ciphertext-Policy Attribute-Based Encryption (CP-ABE) [7] and Key-Policy Attribute-Based Encryption (KP-ABE) [6] according to whether attributes and policies are attached to the ciphertext or the user's private key. Only when the attributes in the private key satisfy the access structure in the ciphertext can the plaintext be obtained. In the smart grid scenario, the latter CP-ABE is generally considered.

2.2 Access Policy Hidden of ABE

Generally speaking, the access policy in the original CP-ABE also contains user privacy information. Since it is in plain text, everyone can see it, which has the risk of privacy leakage. This issue was first considered by Katz [8], who used Inner Product Predicate Encryption (IPE). While Nishide et al. first proposed the idea of a partial hiding strategy in CP-ABE [16]. Their scheme calls for the access policy to be a one-piece gate structure. Since then, policy hiding has gradually developed into partial policy hiding and complete policy hiding. Phuong [17] further studied the method of implementing policy hiding using AND gate CP-ABE. Lai *et al.* [9] divided the attributes in the access policy into two parts (attribute name and attribute value). Yang [24] proposed a new scheme that utilizes Attribute Bloom Filter (ABF) to achieve complete policy hiding. However, false positives for ABF may lead to decryption failure. Recently, Ying et al. proposed a policy complete concealment scheme (ACF) based on the cuckoo filter [25], but their scheme does not address the issue of attribute recovery.

2.3 Overhead Size of ABE

The size of the ciphertext and public key PK also increases linearly with the size of the attribute universe. Lewko and Waters [12] proposed the first unbounded KP-ABE scheme, which can support a large universe of attributes in complex ordered groups. In addition, Rouselakis and Waters [18] proposed KP-ABE and CP-ABE schemes to support a large attribute universe in units of prime order. Peng [27] proposed a composite scheme for large attribute universe. Recently, Cui [4] proposed a partially hidden CP-ABE scheme supporting LSSS policy and large attribute sum based on [18], but this scheme was proved to be secure in random Oracle models and could not achieve full safety. In particular, this scheme does not support decryption testing before full decryption, so even schemes designed based on prime order groups are inefficient. In contrast, this paper proposes a more efficient large attribute universe CP-ABE with complete policy hiding, which supports high expressivity and a large universe.

2.4 Attribute Importance of ABE

Although there are many attribute-based encryption schemes, these schemes share a common feature: they seldom consider the importance of attributes, that is, the status of attributes is equal. Since each attribute plays a different role in the system, the corresponding states they have in the system are also different. Therefore, in real life, attributes with weights have practical significance. Compared with the previous attribute-based encryption scheme, the introduction of weight in the system will make the scheme closer to the actual situation and have practical significance for the actual scenarios [21]. Liu [15] proposed the concept of attribute weights, but it was introduced in one sentence in a general way, lacking actual steps and algorithms. This paper considers the use of AHP in operations research to assign weights to all attributes in smart grid scenarios, making our scheme more realistic, simplifying access policies, and reducing communication overhead.

3 Preliminaries and Defintions

3.1 Bilinear Pairing

A bilinear pairing has three essential characteristics: Bilinear, Non-degenerate, and Computable. Suppose $\mathbb{G}$ and $\mathbb{G}_T$ are two multiplicative groups, and g is the generator of $\mathbb{G}$. If $\mathbb{Z}_p$ is an integer group. Then, the attributes of bilinear mapping function $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ are

- Computability: For any $a, b \in \mathbb{G}$, $e(a, b)$ is computed efficiently.
- Bilinear: $\forall\ x_1,\ x_2 \in \mathbb{Z}_p$, $e(g^{x_1}, g^{x_2}) = e(g, g)^{x_1 x_2}$.
- Non-degeneracy: $\forall\ a, b \in \mathbb{G}$, the $e(a, b) \neq 1$ holds.

3.2 Analytic Hierarchy Process (AHP)

AHP is a systematic, hierarchical, multi-objective, and comprehensive evaluation method. AHP can also play a role in a situation where the attributes to be evaluated of the evaluation object are complex and diverse, with different structures and difficult to quantify. It decomposes complex issues into various components, and then groups these factors according to the dominant relationship to form a hierarchical structure. Determine the relative importance of the factors in the hierarchy by way of pairwise comparison. Then, the overall ranking of the relative importance of the alternatives is determined. The whole process embodies the ideological characteristics of entry decomposition and problem-judgment synthesis. There are generally four steps:

- Analyze the problem, clarify the needs, determine the evaluation indicators, and establish the evaluation level relationship.
- Construct the judgment matrix for each node in the previous layer and the next layer.
- Obtain the relative weights between layers from the judgment matrix (single-level ranking and consistency check).
- Calculate the total weight of each layer to reach the total evaluation target (the total ranking of the layers) and obtain the evaluation results of each alternative.

3.3 Linear Secret Sharing Scheme (LSSS)

A secret sharing scheme $\prod$ over a set of parties P is called linear (over $\mathbb{Z}_p$) if

- The shares for each party form a vector over $\mathbb{Z}_p$.
- There exists a matrix A with l rows and n columns called the share-generating matrix $\prod$. For all $i = 1, ..., l$, the i^{th} row of A is labeled by a party $\rho_{(i)}$ (ρ is a function from $1, ..., l$ to P). When considering the column vector $v = (s,\ r_2,...,r_n)$, where $s \in \mathbb{Z}_p$ is the secret to be shared, and $r_2,...,r_n \in \mathbb{Z}_p$ are randomly chosen, then Av is the vector of l shares of the secret s according to $\prod$. The share $(Av)_i$ belongs to party $\rho_{(i)}$.

Suppose that $\prod$ is an LSSS for access structure A. Let S $\in$ A be any authorized set, and let I $\subset \{1, ..., l\}$ be defined as I $= \{$i $\mid \rho_{(i)} \in$ S$\}$. Then there exist constants $\{\omega_i \in \mathbb{Z}_p\}$, $\{\lambda_i\}$ are valid shares of any secret s according to $\prod$, then $\sum_{i\in I}\omega_i\lambda_i = s$. Let A_i denotes the i^{th} row of A, then have $\sum_{i\in I}\omega_i A_i = (1, 0, \cdots, 0)$. These constants $\{\omega_i\}$ can be found in time polynomial in the size of the share-generation matrix A. Note that, for unauthorized sets, no such constants $\{\omega_i\}$ exist.

4 System Model and Security Requirements

4.1 System Model

As depicted in Figure 1, five generic entities: the authority center (AC), smart cloud (SC), fog nodes (FN), data owner (DO), and data users (DU) are involved in the privacy-aware smart grid access control system, which is described below.

AC: AC is responsible for attribute weight assignment, initialization and user authorization. It is trustworthy, the computer capability of the certification center is very strong, and it is a completely trusted third party;

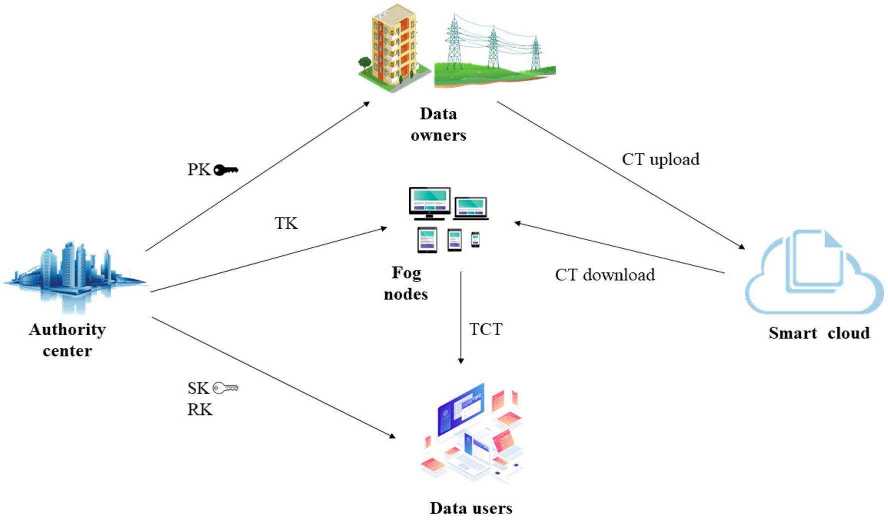


Figure 1: The system model

DO: DO manages the smart grid system and stores power data on the smart cloud in encrypted text with access policies completely hidden. Data from sensors or other smart devices is encrypted with the server's help and sent to the smart grid cloud to be shared with data users. Data owners are responsible for defining and enforcing access policies for encrypted information and completing online/offline encryption;

SC: SC has abundant storage capacity, stores encrypted ciphertext, has a completely hidden access policy, is honest and curious, and is untrustworthy;

FN : FN receives the encrypted complete ciphertext, has relatively large computing power, converts the original ciphertext into converted ciphertext, and sends it to the users;

DU: DU is a data consumer, usually a grid worker. Each department has a set of attributes and a key associated with the attribute group. They need access to the plaintext. So only when their attribute group satisfies the required access policy can they become legitimate users and finally decrypt the ciphertext.

4.2 Security Requirements

Security is very important for the smart grid. In our security model, it is generally believed that the AC is completely trusted and that the power information of the data owner will not be false. Suppose SC is honest and curious. According to the security model and system model of this scheme, the following security requirements should be met in secure smart grid communication.

The Data Confidentiality. To protect the information of the data owner (such as the power information received by the meter) from being attacked by attacker A, the page cannot obtain any relevant information from the access control policy during the process of incoming cloud communication. At the same time, during the decryption process, attacker A cannot match the identity information and cannot grant decryption permission. Even if A obtains either TK or RK, the ciphertext cannot be completely decrypted. In this way, the confidentiality of the power information can be guaranteed.

Collusion Resistance. Different users may work with the Smart Cloud (SC) to read power information that they do not have access to combine keys. For the data security of the smart grid, we should resist these collusion attacks.

Access Policy Privacy Protection. In the scheme of this paper, the content of the access policy should be sensitive and should be completely hidden to protect the privacy of the access policy.

5 The Proposed Scheme

The construction of our big data access control is based on the CP-ABE in [10]. Our scheme can also be applied to any CP-ABE schemes with LSSS structured access policies and consists of the following four algorithms: System Initialization, Key Generation, System Encryption, and System Decryption. The parameters involved at the same time are shown in Table 1.

5.1 System Initialization

The initialization algorithm first takes all the attributes U of the system as input, outputs the attribute segmentation set U^*, PK and MSK.

Table 1: Notations in system

Notations	Descriptions
U	The whole attributes
PK	The public key
MSK	The master secret key
S	An attribute set
SK	A set of secret keys associated with S
M	The plaintext
(A,ρ)	An access structure
CT	The ciphertext
CT’	Semi-encrypted ciphertext
TCT	Semi-decrypted ciphertext
U	The number of attributes in the set
S	Number of user attributes
L	The size of the access policy
E	The exponential action in $\mathbb{G}$, $\mathbb{G}_T$
Mt	A multiplication activity in group $\mathbb{G}_T$
M	A multiplication activity in group $\mathbb{G}$
I	The amount of attributes in secret key
n	The number of hash functions

The input of the attribute weight assignment algorithm is the entire attribute set U. Since each attribute in the system has a different weight, the trust center assigns a maximum weight allowed by the system to each attribute in the system, and then according to the weight, The whole attribute set U is transformed into the whole attribute weight segmentation set U^*. Then it outputs the segmentation set of all attribute weights.

5.1.1 Evalution Indicator

In AHP, three layers of hierarchical relationships are constructed: the target layer, criterion layer, and scheme layer.

A judgment matrix is constructed for each node in one layer and all its related nodes in the lower layer, and the judgment matrix describes the relative importance or superiority of the nodes in the next layer. To quantify the priority between nodes, the following judgment matrix scale definition Table 2 will be used.

According to the judgment matrix, for a node in the previous layer, the weight value of the important order of all the nodes in this layer that are related to it is calculated, to sort the importance order according to the weight.

The overall ranking of the hierarchy assigns weight to the importance of the scheme layer compared to the target layer. In this way, the decision result with the largest weight can be selected.

5.1.2 Setup

Then take a security parameter λ as input. It first runs $\mathbb{G}(\lambda)$ to obtain a bilinear group $(p, \mathbb{G}, \mathbb{G}_T, e)$, the bilinear map is $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$, where $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic groups of prime order p. It then chooses $g \in \mathbb{G}$, $\alpha, \beta \in \mathbb{Z}_p$, uniformly at random.Let L_{att} be the maximum bit length of attributes in the system. Let L_{row} be the maximum bit length of the row numbers of the access matrix. Let L_{si-ABF} be the size of the bit array of the si-ABF. Let n be the number of hash functions associated with the si-ABF.Then, computes $e(g,g)^{\alpha}$, g^{β} The public parameters are published as

$$\begin{aligned} PK &= (\mathbb{G}, \mathbb{G}_T, e, e(g,g)^{\alpha}, g, g^{\beta}, L_{att}, L_{row}, \\ &\quad L_{s_i-ABF}, HASH_1() \cdots HASH_n()) \\ MSK &= (\alpha, \beta). \end{aligned}$$

5.2 Key Generation

The key generation algorithm takes as input the public key, the master secret key, and a set S^* of attributes weights. The algorithm first randomly picks $t \in \mathbb{Z}_p$. Then, the secret key $SK = (S, K, K_0, K_x)$ is computed as $K = g^{\alpha} g^{\beta t}$, $K_0 = g^t$, $K_x = g_x^t \ \forall\ i \in S^*$.

5.3 System Encryption

Before the data is uploaded to the cloud, the data owner encrypts the plaintext data under the access policy in the form of LSSS by invoking the encryption algorithm. The encryption algorithm takes as input the public parameters PK, a message $M \in \mathbb{G}_T$ to encrypt, and an LSSS access structure $A = (A,\rho)$, where A is an $l \times n$ matrix which is a map from each row A_i of A to an attribute $\rho_{(i)}$. The algorithm first chooses a random vector $v = (s, v_2, . . . , v_n) \in \mathbb{Z}_p^n$. These values will be used to share the encryption exponent s. Then, for each row A_i of A, it chooses $r_i \in \mathbb{Z}_p$ uniformly at random. The ciphertext is $CT = (A, C, C', C_i, D_i)$, where $C = M\ \mathrm{e}(g,g)^{\alpha s}$, $C' = g^s$, $C_i = g^{\beta A_i v} g_{\rho_{(i)}}^{-r_i}$, $D_i = g^{r_i}$ $\forall$i $\in \{1, 2, ..., l\}$.

5.3.1 Offline Encryption

The input of the offline encryption algorithm is the access public key and the access policy of Linear Secret Sharing Scheme (LSSS). A is an $l \times n$ access matrix, and ρ maps its row vectors to attributes. Calculate and output the semi-encrypted ciphertext:

$$\begin{aligned} CT' &= (A, C', C_i, D_i), \quad \text{where } C' = g^s, \\ C_i &= g^{\beta A_i v} g_{\rho_{(i)}}^{-r_i}, \\ D_i &= g^{r_i}. \end{aligned}$$

In order to protect our power information, we need to completely hide the access policy, especially to remove the attribute mapping function ρ . However, at the same time, it will also have a great impact on the decryption of data users, and they do not know what attributes are involved in the scheme. To solve this problem, this paper proposes si-ABF, a security-improved attribute Bloom Filter algorithm.

Table 2: Judgment matrix scaling definition

Scaling	Meaning
1	The two elements are of equal importance
3	The former is slightly more important or advantageous than the latter
5	The former is more important or advantageous than the latter
7	The former is more important or advantageous than the latter
9	The former is absolutely important or has an advantage over the latter
2,4,6,8	The intermediate value between the above scales

The traditional BF can only provide the member query function, and the false positive rate is high. The ABF proposed by Yang [24] builds blocks with the help of an array of λ bits, λ is a security parameter, and ABF has a lower false-positive rate than BF. At the same time, in order to precisely locate the attribute corresponding to the row number in the access matrix, ABF uses a specific string as the element in the table.

si-ABF sets a unique hash value for each element. When there is a conflict in inserting a new element, we choose to create a new si-ABF table to store new elements instead of sharing them in ABF and specifying the table at the same time. The medium capacity must not exceed 70%, and once saturated, it will also actively create a new si-ABF table.

The data owner first takes the access policy (A, ρ) as input, and binds the attributes involved in the access policy with its corresponding row number in the access matrix A, where the ith row of the access matrix maps to the attribute att= ρ(i). In order to input the attribute into si-ABF, we bind the line number to the attribute,to find the attribute and the corresponding specific line number more accurately. The algorithm first shares elements with the secret sharing scheme by randomly generating $n-1$ λ bit strings $r_{1,att}$, $r_{2,att}$,... The atts are all expanded to the maximum bit length.

The algorithm hashes the attribute atte associated with element e using n independent and uniform hash functions $HASH_1(), \cdots, HASH_n()$ and get $HASH_1$(att), $HASH_2$(att), ..., $HASH_n$(att), where each $HASH$(att) ($i \in [1, n]$) represents the position index of ABF, it then stores the ith element share ri to the position indexed by $HASH$(att) in the si-ABF as $r_{1,el} \rightarrow HASH_1$(att) position in si-ABF;

$r_{2,el} \rightarrow HASH_2$(att) position in si-ABF;

$\vdots$

$r_{n,el} \rightarrow HASH_n$(att) position in si-ABF.

As elements continue to be added to the si-ABF, a position may already be occupied by previously added elements. If this happens, the system will open a new table and place it in the appropriate location of the new table. Instead of picking one at random or reusing the original location.

The specific procedure of access policy completely hidden can be computed by Algorithm 1. And the model of si-ABF refer to Figure 2.

Algorithm 1 Generation of the si-ABF

```
Input: Mapping function $\rho$ (row) = att, $L_{si-ABF}$
Input: n HASH functions $HASH_1()$,...$HASH_n()$
Output: The tables si-ABF
for i = 0 to $L_{si-ABF}$ -1 do
  Initialize si-ABF with NULL → si-ABF[i] = NULL

  for each attribute el=i|| att $\in S_e$ do
    pos=-1,finval=x
    for i=0 to n-1 do
      p=$HASH_{i+1}$(att) the index of the position
      if si-ABF[p]==NULL then
        if pos==-1 then
          pos=p
        else
          generate a random string $r_{p,el}$ with $\lambda$ bits
          si-ABF[p]=$r_{p,el}$,finval=finval $\oplus$ si-ABF[p]
        end if
      else
        finval=finval $\oplus$ si-ABF[p]
      end if
    end for
  end for
end for
The current table is saturated (70%), or the collision happens in the current position
return si-ABF
End
```

5.3.2 Online Encryption

The input of the online encryption algorithm stage is the public key, the plaintext M, and the semi-encrypted ciphertext CT', and the final ciphertext CT is calculated and output. $CT = (A, C, C', C_i, D_i)$, where $C = M$ e$(g, g)^{\alpha s}$, $C' = g^s$, $C_i = g^{\beta A_i v} g_{\rho(i)}^{-r_i}$, $D_i = g^{r_i}$.

5.4 System Decryption

The decryption algorithm takes as input the private key $SK = (S, K, K_0, K_i)$ for a set of attributes S and a ciphertext $CT = (A, C, C', C_i, D_i)$ where A is an $l \times n$ matrix. If S does not satisfy the access structure A, it outputs

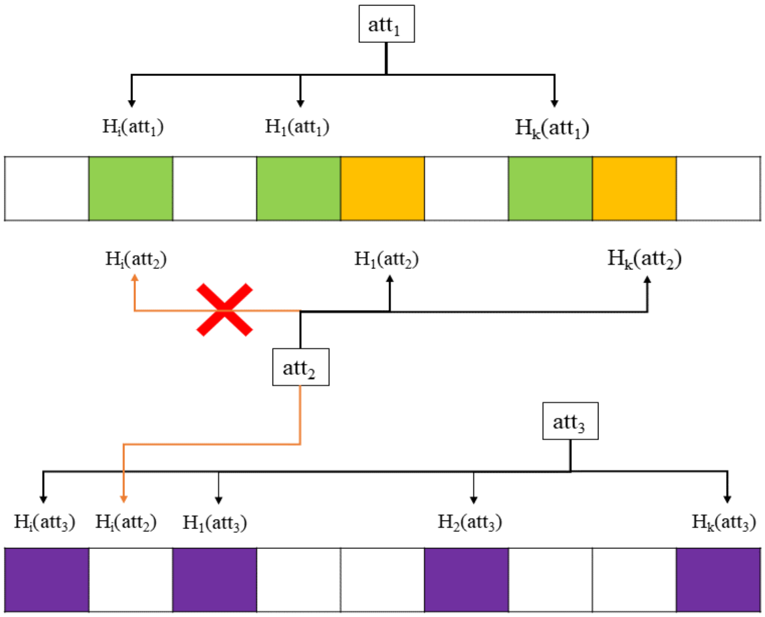


Figure 2: The model of si-ABF

$\perp$. Suppose that S satisfies the access structure A and let $I \subset \{1, 2, ..., l\}$ be defined as $I = \{i : \rho_{(i)} \in S\}$. It computes constant $\omega_i \in \mathbb{Z}_p$ such that $\sum_{i \in I} \omega_i A_i = (1, 0, ..., 0)$.

5.4.1 si-ABF Match

When accessing cloud data again, the data users must have their attributes to satisfy the access policy before decrypting the ciphertext. In traditional ABE systems, the access policy (A,ρ) is appended to the ciphertext in plaintext. Therefore, data users can easily check whether their attributes satisfy the access policy. We hide the attribute mapping function ρ, and should first check which attributes they have in the access matrix A by running Algorithm 2, si-ABF matching.

For each attribute att $\in$ S owned by the data users, the algorithm calculates each position index through the provided n $HASH$ functions and get

$HASH_1$(att), $HASH_2$(att),...,$HASH_n$(att).

Next, the corresponding string can be obtained through the position index.

$HASH_1$(att) position in si-ABF→$r_{1,el}$;

$HASH_2$(att) position in si-ABF→$r_{2,el}$;

$\vdots$

$HASH_n$(att) position in si-ABF →$r_{n,el}$.

We delete all zero bits on the left side of the corresponding string to get att. If the table is consistent with the user's att, the attribute is said to be in this access matrix. Next, get the line number of the attribute through the left bit of the string, if the above conditions are met, it means that the user attribute set satisfies this access control, and get the corresponding attribute set S_{si-ABF}.

Algorithm 2 si-ABF Match

```
Input: a security improved-Attribute Bloom Filter si-ABF, a set of attributes S, n hash functions $HASH_1(), \cdots, HASH_n()$
Output: The matching attribute set $S_{si-ABF}$.
for each att ∈ S do
  initialize the NewStr
  for i = 0 to n-1 do do
    p = $HASH_{i+1}$(att) get the index of the position
    NewStr=NewStr ⊕ si-ABF[p]
    Wait until the key bits of the attribute, and then remove the preceding zero bits.
    if att correspond to the same then
      According to access policy, the attributes corresponding to the line numbers in are expressed as $S_{si-ABF}$
    end if
  end for
end for
return $S_{si-ABF}$
End
```

5.4.2 Transkey Generation

The key transformation algorithm inputs the public key PK and the private key SK, randomly select θ.

$$\begin{aligned} K' &= K^{(1/\theta)} = g^{(\alpha/\theta)} g^{(\beta t/\theta)} \\ K_0' &= K_0{}^{(1/\theta)} = g^{(t/\theta)} \\ K_i{}' &= K_i{}^{(1/\theta)} = g_i{}^{(t/\theta)} \end{aligned} \qquad (1)$$

Accroding to Equation (1), we enter the conversion key

Table 3: Scheme feature comparison

Scheme	Policy Hidden	Online\Offline Encryption	Give Weight	Large Universe
Scheme [26]	Completely Hidden	✗	✗	✗
Scheme [28]	Partially Hidden	✗	✗	✓
Scheme [23]	✗	✗	✗	✗
Scheme [24]	Completely Hidden	✗	✗	✗
Our Scheme	Completely Hidden	✓	✓	✓

as TK=(S, K',K_0',K_i'), the local key RK= θ

5.4.3 Decryption

The outsourced decryption algorithm inputs the public key PK, the ciphertext CT, and the conversion key TK, then calculate TCT:

$$\begin{aligned}&\frac{e(C',K')}{\prod_{i\in I} e(C_i,K_0')e(K'_{\rho(i)},D_i)^{\omega_i}}\\ =&\frac{e(g,g)^{\alpha s/\theta}e(g,g)^{\beta ts/\theta}}{\prod_{i\in I}(g,g)^{\beta tA_i v w_i/\theta}}\\ =&e(g,g)^{\alpha s/\theta}\end{aligned} \qquad (2)$$

The decryption algorithm takes the semi-decrypted ciphertext TCT and the local key RK as input, according to Equation (2), we can get Equation (3), and through the calculation, the original ciphertext can be recovered:

$$\frac{C}{(TCT)^{\theta}} = \frac{Me(g,g)^{\alpha s}}{(TCT)^{\theta}} = M \qquad (3)$$

6 Preformance Evalution

By comparing the constructed scheme with other related schemes in terms of public key size, private key size, ciphertext size, whether to support weighted attributes, whether to complete the complete hiding of access policies, and whether to achieve a large attribute universe, the specific conclusions are shown in Table 3.

The concept of the weighted attribute is introduced in the ABE scheme, where the access policy is completely hidden, which does not affect the tracking performance. Since the weights of attributes in the system are determined by the authorization center AC before the initialization phase, there is no computational time cost to generate the weights of attributes.

To illustrate the issue of computational overhead, we program the scheme on an Ubuntu system by using the Java Pair-Based Cryptography (JPBC) library. Choose a class a pairing to complete the activity in the prime order group and construct an elliptic curve $y^2 = x^3 + x$, the system has a 2.4 GHz Intel Core i5 CPU and 3GB RAM with a base field size of 512 bits and an embedded degree of 2, making the security parameter equal to 1024 bits. The pairing( ) operation is implemented by calling the pairing function, and the powzn( ) function and the mul( ) function is called respectively to test the modular exponentiation and modular multiplication.

6.1 Communication Cost Comparison

As shown in Table 4, by comparing with other schemes. In our scheme, the PK size, SK size, and CT size are respectively constant, S+3, and 2L+3, which is better than other schemes. We also achieve that the size of PK remains stable no matter how many attributes are involved in our scheme, whereas in most other schemes, the size of PK and CT will increase linearly with the increase in the number of attributes. Although the scheme of [28] supports a large attribute domain, it only realizes the partial hiding of attributes and the assignment of attribute weights, and the security cannot be guaranteed. The length of the CT in [28] is 3L+4, which is longer than our proposed scheme, and the communication cost is higher than other schemes.

Table 4: Communication cost comparison

Scheme	PK size	SK size	CT size
Scheme [26]	U+8	S+2	2L+5
Scheme [28]	6	S+2	3L+4
Scheme [23]	7+U	2S+4	3L+3
Scheme [24]	U+n+6	S+2	L+2
Our Scheme	n+9	S+3	2L+2

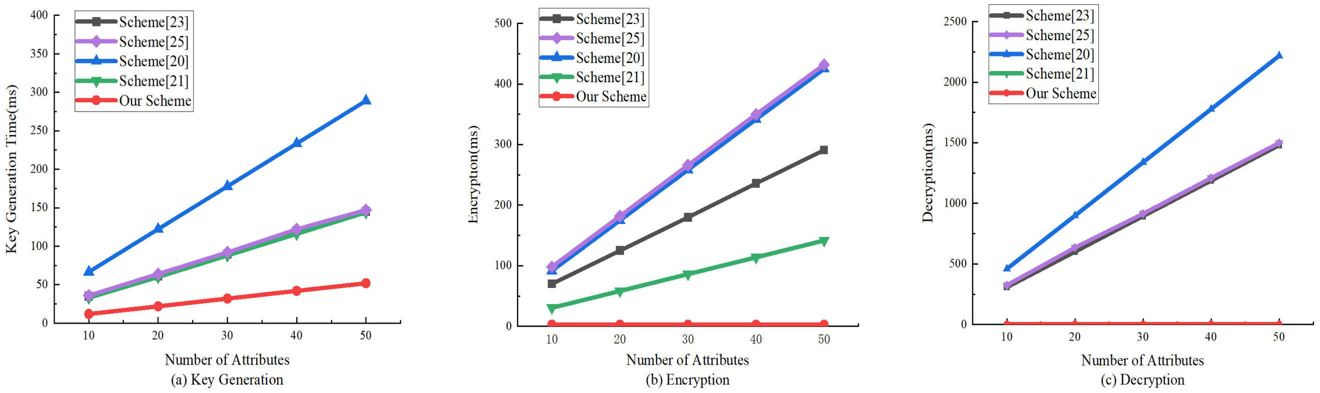
6.2 Computional Cost Comparison

The computational cost is shown in Table 5, where E represents the exponential action in $\mathbb{G}$, $\mathbb{G}_T$, P is the bilinear pair operation, M represents the multiplication activity in the $\mathbb{G}$ group, Mt represents the multiplication activity in the $\mathbb{G}_T$ group, and |I| represents the number of attributes in the decryption key that satisfy the conditions of the access policy.

Although the scheme of [24] is the most efficient and concise among other schemes, it implements fewer functions. Although other schemes achieve complete hiding of access policies and large attribute domains, their computational overhead becomes larger at the same time. While

Table 5: Computational cost comparison

Scheme	Key Generation	Encryption	Decryption
Scheme [26]	(S+2)E+M	(2L+5)E+Mt	(2I+1)P+M+2Mt
Scheme [28]	(S+2)E+3M	(3L+4)E+4Mt	(2I+1)P+3Mt+2E
Scheme [23]	(4+2S)E+M	(3+3L)E	(3I+1)P+Mt+2E+2M
Scheme [24]	(S+2)E+M	(L+1)E	(2I+1)P+3Mt+2E
Our Scheme	(2+S)	E	E



(a) Key Generation

(b) Encryption

(c) Decryption

Figure 3: Evalution of time cost

realizing these functions, we introduce online/offline encryption and outsource decryption. In the schemes of [23, 24, 26, 28], the computational overheads of the encryption stage are (25+2L)E+Mt, (4+3L)E+4Mt, and (3+3L)E, (L+1)E. However, in our scheme, not only the functions of the above schemes are realized, but also complex operations such as complex bilinear pairing are put into the offline stage, the remaining simple multiplication operations are completed in the encryption stage, and only one modular exponentiation operation is required. In the schemes of decryption stages [23, 24, 26, 28], the computational overheads are (2I+1)P+M+2Mt, (2I+1)P+3Mt+2E, (3I+1)P+Mt+2E+2M and (2I+1)P+2E+3Mt. Our solution outsources complex operations to edge fog nodes, and the plaintext can be obtained locally by a modular exponentiation operation. Compared with other schemes, our scheme not only realizes the function of hiding each attribute, ensuring the security of the scheme, but also realizes a large attribute domain, which reduces the communication overhead of the scheme to a certain extent, and the distributed operations of encryption and decryption also The calculation overhead is reduced. The time cost of the terminal device is further analyzed through experiments, including the time of the key generation process, encryption steps, and decryption phase, as shown in Figure 3.

7 Conclusion

This paper addresses data privacy and access control privacy issues by introducing an efficient weighted attribute encryption scheme with privacy-preserving policies in the smart grid. We use si-ABF to completely hide the access policies while preventing them from colliding and setting up a matching algorithm to avoid the risk of false positives in filters. For traditional CP-ABE, the time and computational costs required for the bilinear pairing of the calculation steps are too large, so this paper adopts an online and offline method for the encryption step and allocates a very large part of the calculation to the offline stage, while the remaining simple algorithms are completed online, which greatly reduces the burden of the scheme and improves the efficiency of the work. We also assign the maximum weight to all attributes in the system and use the AHP analytic hierarchy process in logic to ensure the rationality of the weight and, at the same time, provide convenience for the operation of the power grid system. The current solution is safer, more expressive, and more realistic.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.U1936213,61872230, Program of Shanghai Academic Research Leader No.21XD1421500, Shanghai Science and Technology Commission Project No.20020500600.

References

[1] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in *IEEE symposium on security and privacy (SP'07)*, pp. 321–334, 2007.

[2] P. S. Chung, C. W. Liu, and M. S. Hwang, "A study of attribute-based proxy re-encryption scheme in cloud environments", *International Journal of Network Security*, vol. 16, no. 1, pp. 1-13, 2014.

[3] I. Colak, R. Bayindir, And S. Sagiroglu, "The effects of the smart grid system on the national grids," in *8th International Conference on Smart Grid (icSmartGrid)*, IEEE, pp. 122–126, 2020.

[4] H. Cui, R. H. Deng, J. Lai, X. Yi, and S. Nepal, "An efficient and expressive ciphertext-policy attribute-based encryption scheme with partially hidden access structures, revisited," *Computer Networks*, vol. 133, pp. 157–165, 2018.

[5] S. Gao, G. Piao, J. Zhu, X. Ma, and J. Ma, "Trustaccess: A trustworthy secure ciphertext-policy and attribute hiding access control scheme based on blockchain," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 5784–5798, 2020.

[6] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM conference on Computer and communications security*, pp. 89–98, 2006.

[7] H. He, L.-h. Zheng, P. Li, L. Deng, L. Huang, and X. Chen, "An efficient attribute-based hierarchical data access control scheme in cloud computing," *Human-centric Computing and Information Sciences*, vol. 10, pp. 1–19, 2020.

[8] J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," in *annual international conference on the theory and applications of cryptographic techniques*, Springer, pp. 146–162, 2008.

[9] J. Lai, R. H. Deng, and Y. Li, "Expressive cp-abe with partially hidden access structures," in *Proceedings of the 7th ACM symposium on information, computer and communications security*, pp. 18–19, 2012.

[10] J. Lai, R. H. Deng, Y. Yang, and J. Weng, "Adaptable ciphertext-policy attribute-based encryption," in *International Conference on Pairing-Based Cryptography*, Springer, pp. 199–214, 2013.

[11] C. C. Lee, P. S. Chung, M. S. Hwang, "A survey on attribute-based encryption schemes of access control in cloud environments", *International Journal of Network Security*, vol. 15, no. 4, pp. 231-240, 2013.

[12] A. Lewko and B. Waters, "Decentralizing attribute-based encryption," in *Annual international conference on the theory and applications of cryptographic techniques*, Springer, pp. 568–588, 2011.

[13] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of attribute-based access control with user revocation in cloud data storage," *International Journal of Network Security*, vol. 18, no. 5, pp. 900–916, 2016.

[14] X. Liu, J. Ma, J. Xiong, Q. Li, and J. Ma, "Ciphertext-policy weighted attribute based encryption for fine-grained access control," in *2013 5th International Conference On Intelligent Networking And Collaborative Systems*, IEEE, pp. 51–57, 2013.

[15] Z. Liu, Z. Cao, and D. S. Wong, "Blackbox traceable cp-abe: how to catch people leaking their keys by selling decryption devices on ebay," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 475–486, 2013.

[16] T. Nishide, K. Yoneyama, and K. Ohta, "Attribute-based encryption with partially hidden encryptor-specified access structures," in *International conference on applied cryptography and network security*, Springer, pp. 111–129, 2008.

[17] T. V. X. Phuong, G. Yang, and W. Susilo, "Hidden ciphertext policy attribute-based encryption under standard assumptions," *IEEE transactions on information forensics and security*, vol. 11, no. 1, pp. 35–45, 2015.

[18] Y. Rouselakis and B. Waters, "Practical constructions and new proof methods for large universe attribute-based encryption," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 463–474, 2013.

[19] A. Sahai, "Non-malleable non-interactive zero knowledge and adaptive chosen-ciphertext security," in *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, IEEE, pp. 543–553, 1999.

[20] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Annual international conference on the theory and applications of cryptographic techniques*, Springer, pp. 457–473, 2005.

[21] J. Sun, Y. Yang, Z. Liu, and Y. Qiao, "Multi-authority criteria-based encryption scheme for iot," *Security and Communication Networks*, vol. 2021, 2021.

[22] A. Wu, Y. Zhang, X. Zheng, R. Guo, Q. Zhao, and D. Zheng, "Efficient and privacy-preserving traceable attribute-based encryption in blockchain," vol. 74, no. 7, Springer, pp. 401–411, 2019.

[23] X. Yan, X. Yuan, Q. Zhang, and Y. Tang, "Traceable and weighted attribute-based encryption scheme in the cloud environment," *IEEE Access*, vol. 8, pp. 38 285–38 295, 2020.

[24] K. Yang, Q. Han, H. Li, K. Zheng, Z. Su, and X. Shen, "An efficient and fine-grained big data access control scheme with privacy-preserving policy," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 563–571, 2016.

[25] Z. Ying, W. Jang, S. Cao, X. Liu, and J. Cui, "A lightweight cloud sharing phr system with access policy updating," *IEEE Access*, vol. 6, pp. 64 611–64 621, 2018.

[26] Z. Ying, W. Jiang, X. Liu, S. Xu, and R. Deng, "Reliable policy updating under efficient policy hidden fine-grained access control framework for cloud data sharing," *IEEE Transactions on Services Computing*, 2021.

[27] P. Zeng, Z. Zhang, R. Lu, and K.-K. R. Choo, "Efficient policy-hiding and large universe attribute-based encryption with public traceability for internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10963-10972, 2021.

[28] Y. Zhang, D. Zheng, and R. H. Deng, "Security and privacy in smart health: Efficient policy-hiding attribute-based access control," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2130–2145, 2018.

Biography

Jing Cheng Currently studying for a master's degree in the Department of Computer Science and Technology, Shanghai Electric Power University, China. Research interests include smart grid, attribute encryption and privacy protection.

Mi Wen(M'10) Received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2005, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008. She is currently a Professor of the College of Computer Science and Technology with Shanghai University of Electric Power, Shanghai, China. From May 2012 to May 2013, she was a Visiting Scholar at the University of Waterloo, Waterloo, ON, Canada. In 2016, she was awarded as Shanghai Municipal dawn scholar. Her research interests include privacy preserving in wireless networks, big data, smart grid, etc. She is an Associate Editor of Peer-to- Peer Networking and Applications (Springer). She is presently an director of Shanghai Computer Society. She acts as the track chairs of many conferences such as the IEEE VTC.

Bilinear Mapping and Blockchain-based Privacy-Preserving and Data Sharing Scheme for Smart Grid

Xiaoxu Zhang
School of Control and Computer Engineering, North China Electric Power University
Beijing 102206, China
Email: zhangxiaoxu@ncepu.edu.cn

(Received Aug. 24, 2022; Revised and Accepted Dec. 5, 2022; First Online Dec. 14, 2022)

Abstract

With the popularity of the smart grid, there are many problems, such as data security, high cost, and low efficiency in the query and access control of users' electricity data. To solve the above problems, this paper proposes a scheme of privacy-preserving and data sharing based on bilinear mapping and blockchain for an intelligent pricing system of a smart grid. Then, the secure encryption, signature, and authentication mechanism ensure the integrity and confidentiality of the message. An ASV-FLD algorithm is designed based on dichotomy to improve the speed of signature verification and fast recognition. First, the ID-based encryption scheme is used for data aggregation, which can significantly improve the calculation efficiency. Second, an ID-based proxy re-encryption scheme is used for data sharing. Finally, the simulation results show that the scheme has the characteristics of low storage cost, high efficiency of data calculation, flexible access control, and can meet the growing business needs of users.

Keywords: Blockchain; Data Sharing; Privacy-preserving; Smart Grid

1 Introduction

As a new generation power system, smart grid has the characteristics of high reliability, strong resistance to attacks, and real-time interactive friendly. The smart grid needs to frequently obtain user data and aggregate to analyze the user's electricity consumption characteristics, so as to realize the pricing of electricity consumption. However, there is a security risk of leaking user privacy data during the process of data aggregation or sharing. Therefore, the security of privacy-preserving has become a mainstream research trend in smart grid. Various protocols have been proposed for privacy-preserving in the power grid. Reference [20] proposed a scheme that uses Paillier homomorphic cryptography to encrypt structured data and directly aggregate it at the local gateway without decryption.

Reference [18] proposed a usage-based dynamic pricing system based on homomorphic encryption, which dynamically priced based on the user's electricity consumption, while protecting user privacy. However, the above scheme did not consider the feasibility, and it is difficult for the existing computing resources to be fully homomorphic. Reference [28] proposed the use of ring signatures to verify user identity and encrypt data to protect privacy.

Reference [21] proposed using proxy re-encryption anonymously aggregates multi-dimensional data, cutting off the association between entities and users in the system. Reference [25] proposed a data obfuscation mechanism that using elliptic curve signature technology authenticates the distributed obfuscation values [17,19]. Although the above scheme can better protect privacy but does not consider the calculation cost, as the number of users increases, the calculation cost of encryption and decryption and signature will also increase. A differential privacy model is proposed to protect privacy by increasing noise in power consumption [1]. However, the smart grid needs to frequently and accurately obtain user power consumption, so it directly affects the user's power consumption behavior, and this scheme will increase a lot of communication overhead.

Recently, a blockchain technology has been introduced into the research of distributed data security storage. Reference [2] combined traditional blockchain technology and digital signature to ensure the safe transaction of electric energy and the safe verification and storage of data. Because the consensus process of traditional blockchain technology needs the cooperation of all network nodes, resulting in huge network energy consumption. The above scheme is not suitable for smart grid sensor data storage. Reference [22] proposed to apply blockchain and smart contract to smart grid as an intermediary between power consumers and power producers to help reduce costs, improve transaction speed and enhance the security of transaction data generated. In their model, whenever a transaction occurs, there is a blockchain based instrument that updates the blockchain by creating a unique timestamp block in the distributed ledger. At the distribution level, system operators charge customers according to the data recorded on the blockchain.

Zyskind *et al.* [8] used blockchain for access control management and audit log security purposes, as an event tamper proof log is a distributed computing platform based on the optimized version of secure multiparty computing (SMPC). Different participants fully store and run data computing while maintaining complete privacy of data. Reference [27] proposed a data sharing framework based on blockchain, which fully solved the access control challenges related to sensitive data stored in the cloud by using the invariance and built-in autonomy of blockchain. They used the security cryptography technology to ensure the effective access control of the sensitive shared data pool using the allowed blockchain, and designed a data sharing scheme based on the blockchain.

Reference [11] proposed a distributed monitoring infrastructure, drams, based on blockchain for distributed access control system. The main motivation of DRAMs was to deploy a decentralized architecture that can detect policy violations in a distributed access control system under the assumption of a well-defined threat model. Access control management based on blockchain was described in more detail in Thomas and Alex's system. The chainanchor system provided anonymous but verifiable identity to entities attempting to execute transactions to the allowed blockchain. The enhanced privacy ID (EPID) zero knowledge proof scheme was used to realize and proved the anonymity and membership of participants [13].

The main innovations of this paper are as follows. In this paper, the ID-based aggregate signature algorithm is used for signature and authentication. The unidirectionality and non-collision of hash function ensure the reliability of signature scheme. When salesmen query user data, the control center will decrypt the data to the sharing chain after checking their permissions, and then salesmen can view the information. This mechanism can avoid salesmen directly contacting with the data on the private chain, and further improve the security by using the characteristics of blockchain.

The rest of this paper is organized as follows. The second part introduces the basic knowledge, mainly introduces the technology of data aggregation of encryption system, the algorithm of ID-based aggregate signature. The third section introduces the model and design objectives that will be used in this paper. In the fourth part, we propose a system model of data privacy-preserving for smart grid based on double blockchain and cloud storage. Next, we analyze the security of our model in Section five. In the sixth section, the security analysis and extensive evaluation of various performance indexes are carried out. Finally, giving our conclusion.

2 Preliminaries

2.1 Bilinear Pairing

Let $\mathbb{G}$ and $\mathbb{G}_T$ be two groups of prime order q, and g be a generator of $\mathbb{G}$. Consider a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ satisfies the following properties [14,16]:

1) Bilinear: For all $u, v \in \mathbb{G}$ and $a, b \in \mathbb{Z}_q^*$, we have $e(au, bv) = e(u, v)^{ab}$, and $e(u, (a+b)\cdot v) = e(u, a\cdot v)\cdot e(u, b\cdot v)$.

2) Nondegenerate: g should satisfy $e(g, g) \neq 1$.

3) Computable: $e(u, v)$should be computable.

2.2 ID-based Aggregate Signature Scheme

The ID-based aggregate signature [3,6,9] consists of 6 algorithms: $ParamGen$, $KeyGen$, $Sign$ and $Verify$ are the same as that in the ordinary ID-based signature, the signature aggregation algorithm $AggSign$ and the aggregate signature verification algorithm $AggVerify$ provide the aggregation capability. Let $\mathbb{G}$ and $\mathbb{G}_T$ be two groups of prime order p, and P be a generator of $\mathbb{G}$. $H_1 : \{0,1\}^* \longrightarrow \mathbb{G}$, $H_2 : \{0,1\}^* \longrightarrow \mathbb{Z}_q^*$. Consider a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$, described in detail as follows:

$ParamGen$**:** Given a security parameter k, let $s \in \mathbb{Z}_q^*$ denote the master private key and $P_{pub} = s \cdot P$. The outputs system parameters $Params = \{\mathbb{G}, \mathbb{G}_T, e, q, P, H_1, H_2\}$.

$KeyGen$**:** Let $U_1, U_2, \cdots, U_n$ denote all users to join the signing. The identity of U_i is denoted as id_i, the corresponding private-public key pair is $(PK_i = H_1(id_i), SK_i = s \cdot PK_i)$.

$Sign$**:** Given n different messages $m_1, m_2, \cdots, m_n$, without loss of generality, we assume that U_i signs message m_i. He randomly picks a number $r_i \in \mathbb{Z}_q^*$, computes and broadcasts $R_i = r_i \cdot P$, $h_i = H_2(m_i, R_i)$, $S_i = r_i \cdot P_{pub} + h_i \cdot SK_i$, and the signature $\sigma_i = (R_i, S_i)$.

$Verify$**:** Anyone can be designated to aggregate all these single signatures. The designated user (DU) first verifies the validity of each single signature. Having received all the single signatures, DU computes $T_i = R_i + h_i \cdot pk_i$. He accepts the signature if $e(P, S_i) = e(P_pub, T_i)$.

$AggSign$**:** DU comoutes $S = \sum_{i=1}^{n} S_i$. The aggregate signature on n different messages $m_1, m_2, \cdots, m_n$ is $\sigma = (R, S)$.

$AggVerify$**:** After receiving σ, the verifier computes $h_i = H_2(m_i, R = \sum_{i=1}^{n} R_i)$, $T = R + \sum_{i=1}^{n} h_i \cdot pk_i$. He accepts the aggregate signature if $e(P, S) = e(P_pub, T)$.

2.3 ID-based Encryption Scheme

Let $\mathbb{G}$ and $\mathbb{G}_T$ be two groups of prime order q, and P be a generator of $\mathbb{G}$, $H_1 : \{0,1\}^* \longrightarrow \mathbb{G}$. Consider a bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$, described in detail as follows [4,4,12,26]:

ParamGen: Given a security parameter k, it outputs system parameters $Params = \{\mathbb{G}, \mathbb{G}_T, e, q, P, H_1\}$.

KeyGen: Standard key generation algorithm. When the identity $ID \in \{0,1\}^*$ and the master key s are input, the public key $PK_{id} = H_1(id)$ and secret key $SK_{id} = s \cdot PK_{id}$ corresponding to the identity is output.

Encrypt: When M represents plaintext space, inputting a set of public parameters $params$, identity $id \in \{0,1\}^*$ and plain text $m \in M$, select a random$r \longleftarrow \mathbb{Z}_q^*$, and output $C_{id} = \left(r \cdot P, m \cdot e\left(s \cdot P, H_1(id)\right)^{r^2}, r \cdot SK_{id}\right)$, which is the cipher text under identity id.

Decrypt: The input $C_{id} = (C_1, C_2, C_3)$ ciphertext of id, and the output using C_3 to decrypt the ciphertext C_{id} is:

$$\frac{C_2}{e(C_1, C_3)} \quad (1)$$

Next, we consider the proxy re-encryption scheme (PRS) [7,12]. In the PRS, there are six phases: *ParamGen*, *KeyGen*, *Encrypt*, *PKGen*, *Reencrypt* and *Decrypt*. The first three phases are the same as the ID-based encryption scheme. We consider the last three phases.

PKGen: When entering the identity $\{id_1, id_2\} \in \{0,1\}^*$, select $x \in \mathbb{G}_T$, calculate $(R_1, R_2, R_3) = Encrypt(params, id_2, x)$, and output the re-encryption key $RK_{id_1 \longrightarrow id_2} = (R_1, R_2, -C_3 + H_1(x))$.

Reencrypt: In order to re-encrypt the cipher text of id_1 into the cipher text of id_2, the re-encryption key $RK_{id_1 \longrightarrow id_2} = (R_1^{'}, R_2^{'}, R_3^{'})$ are input, the re-encrypted cipher text C_id_2 is output as $c_{id_2} = (C_1, C_2 \cdot e(C_1, R_3^{'}), R_1^{'}, R_2^{'}, R_3^{'})$.

Decrypt: Before decryption, $x = R_2^{'}/e(R_1^{'}, R_3)$ needs to be calculated by $R_2^{'}$. The output using key sk_{id} to decrypt the ciphertext C_{id} is:

$$m = \frac{C_2 \cdot e\left(C_1, R_3^{'}\right)}{e\left(C_1, H_1(x)\right)} \quad (2)$$

In this paper, we use the user's public key to encrypt, and the control center decrypts according to the master key, in which the data is aggregated in the aggregation node to improve efficiency. However, the difference between the sharing phase and the encryption operation is whether the re-encryption key is generated after encryption.

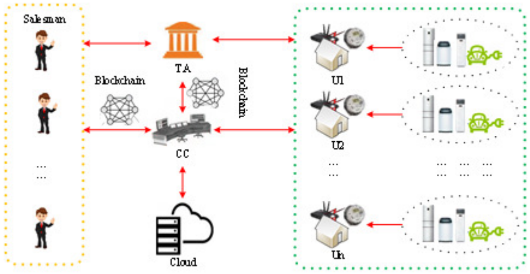


Figure 1: System model

3 System Model

3.1 Network Model

In our system model, the communication between various entities is shown in Figure 1. This article includes six entities: smart meter users (SM), aggregation nodes, control center (CC), cloud, trusted authority (TA), and foreign guest users (salesman). SM stands for a series of smart meters deployed on the user side and has sensing and communication modules. The CC is responsible for collecting data, maintaining shared blockchain, and interacting with outside visitors. The cloud is responsible for storing user data and returning the storage address. The TA is a completely trusted third party, responsible for system booting, key material distribution and agent re-encryption, and does not participate in the rest of the process. It is assumed here that there is a secure channel between the TA and other entities to transmit key material. Salesman is a user outside the CC system and has a need to access specific user data. One private chain stores a mapping of the user's real identity and pseudonym, and users have permission to view it; the other sharing chain is used to share information with outside visitors.

Transaction records: The transactions in our block are organized in a structure based on the Merkle tree, where leaf nodes represent data access transactions for mobile users. In the private chain, store smart meter IDs and hash values stored on the cloud; in the public chain, store encrypted data information that has been queried with access control rights.

Block header: The block header contains the following metadata to verify the data block.

1) Hash: The SHA256 hash of the block. The hash value can be expressed as $Hash = Hash(Hash1 + Hash2) = Hash[(Tx1.Hash) + (Tx2.Hash]$.

2) Last Hash: The hash of the previous block used for block verification.

3) Merkle Root: A structure that stores a set of transactions in each block.

4) Nonce: refers to a number generated by performing a proof of work operation on the Miner

node to generate a hash value lower than the target difficulty level.

5) Timestamp: refers to the time when the block was created. It is also the timestamp of the last transaction in the block.

3.2 Security Model

The security model in this paper considers the trusted state of each entity, assuming that the control center and the trusted center are completely trusted, and the aggregation node and the foreign guest are honest but curious. Honest but curious setting is on the one hand to honestly implement the protocol designed in this article, on the other hand to be curious and try to get the user's sensitive information. Suppose there is a malicious attacker A in the communication channel from the user to the aggregation node, trying to obtain the transmission data, it will also initiate some active attacks or inject useless data to destroy the data integrity.

Definition 1. *Computational Diffie-Hellman Problem (CDHP): If the discrete logarithm problem (DLP) in $\mathbb{G}$ and $\mathbb{G}_T$ is hard, given (P, aP, bP) in $\mathbb{G}$, compute abP. The hardness of $CDHP$ in $\mathbb{G}$ depends on the hardness assumption of DLP in $\mathbb{G}$ [5, 12].*

Definition 2. *Decisional Bilinear Diffie Hellman Assumption: Our encryption schemes are based on the assumed intractability of the Decisional Bilinear Diffie-Hellman problem (DBDH) in $\mathbb{G}$ and $\mathbb{G}_T$. This assumption is believed to hold in certain groups and used as the basis of several Identity-Based Encryption schemes, e.g. [4, 26].*

3.3 Design Goal

The design goal of this paper is to propose a privacy-preserving data aggregation and sharing scheme based on Bloom filter and dual blockchain under the condition of satisfying system model and security requirements. The problems to be solved in this paper are:

1) Unable to ensure trusted third-party aggregation;

2) No corresponding authority control strategy;

3) Fast identity authentication strategy.

The detailed objectives of this article are as follows:

Security and privacy: Our solution should meet all security requirements, such as confidentiality, authentication, integrity, and privacy-preserving requirements. This requires the control center and aggregation node to be able to recognize the illegal operation of attacker A. Collect reliable and complete data. The user's privacy needs should be met. No other users or aggregators can read the user's data. Only the control center and the guest who have access to the data can obtain the data.

Efficiency: Since the terminals in the smart grid, such as smart meters, are resource-restricted devices, this requires that our proposed scheme is lightweight, and the computational complexity of encryption and authentication operations is as small as possible, while the communication load should be as small as possible.

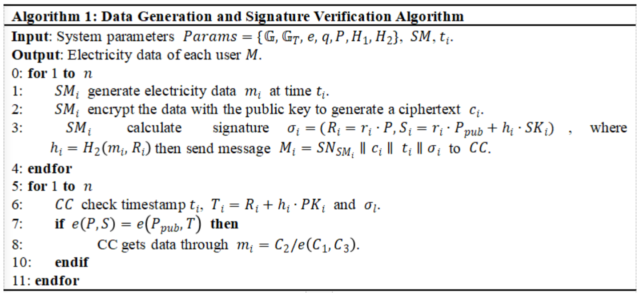
3.3.1 Bilinear Mapping and Blockchain-based Privacy-Preserving and Sharing Scheme

This paper considers two chains, one private chain is used as a side chain to store the mapping relationship between user identity and pseudonym; the other shared chain is used as the main chain to share information with foreign visitors. It is divided into four phases: system initialization, data generation, data aggregation, and sharing.

System initialization phase. Let $\mathbb{G}$ and $\mathbb{G}_T$ be two groups of prime order p, and P be a generator of $\mathbb{G}$. Consider a bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$. Given a security parameter, it outputs system parameters$Params = \{\mathbb{G}, \mathbb{G}_T, e, q, P, H_1, H_2\}$. Let $s \in \mathbb{Z}_q^*$ denote the master private key and $P_{pub} = s \cdot P$, $U_1, U_2, \cdots, U_n$ denote all users. The identity of U_i is denoted as id_i, the corresponding private-public key pair is $(PK_i = H_1(id_i), SK_i = s \cdot PK_i)$.

Data generation phase. To confuse the user's real identity, we use a pseudonym method to transmit data, that is, the serial number of the smart meter is used as a pseudonym, and the relationship between the user's real identity and the pseudonym is stored in the secondary chain. Each user selects a $r_i \in \mathbb{Z}_q^*$, assume that the electricity consumption data generated by smart meter SM_i at time t_i. Generate ciphertext $c_i = \left(r_i \cdot P, m_i \cdot e\left(s \cdot P, H_1(SN_{SM_i})\right)^{r^2}, r_i \cdot SK_i\right)$ and signature $\sigma_i = (R_i = r_i \cdot P, S_i = r_i \cdot P_{pub} + h_i \cdot SK_i)$, where $h_i = H_2(c_i, R_i)$. SM_i sends $M_i = SN_{SM_i} \parallel c_i \parallel t_i \parallel \sigma_i$to CC.

Signature verification phase. The CC receives n data $M_1, M_2, \cdots, M_n$ by $U_1, U_2, \cdots, U_n$ at time t_i, checks whether $e(P, S_i) = e(P_pub, T_i)$, where $S_i = r_i \cdot P_{pub} + h_i \cdot SK_i$ and $T_i = R_i + h_i \cdot PK_i$. If the aggregate signature is valid, the CC gets data through $m_i = C_2/e(C_1, C_3)$. Therefore, the billing center of the control center goes to the cloud storage center to query the electricity data to verify each user's pseudonym. If the user's identity is true and the data has not been tampered with, the billing standards for different periods of time calculate the dynamic electricity charges for each user and formulate more economical and suitable electricity consumption suggestions for the users based on these electricity consumption data. Assuming that the charging standard for different time periods is $p_1, p_2, \cdots, p_h$, the corresponding power consumption is $D_1, D_2, \cdots, D_n$ then the electricity billing rule is $S = p_1 \cdot D_1 + p_2 \cdot D_2 + \cdots + p_n \cdot D_n$. This process is shown in Figure 2.



```
Algorithm 1: Data Generation and Signature Verification Algorithm
Input: System parameters $Params = \{\mathbb{G}, \mathbb{G}_T, e, q, P, H_1, H_2\}$, $SM, t_i$.
Output: Electricity data of each user $M$.
for 1 to $n$
    $SM_i$ generate electricity data $m_i$ at time $t_i$.
    $SM_i$ encrypt the data with the public key to generate a ciphertext $c_i$.
    $SM_i$ calculate signature $\sigma_i = (R_i = r_i \cdot P, S_i = r_i \cdot P_{pub} + h_i \cdot SK_i)$, where
        $h_i = H_2(m_i, R_i)$ then send message $M_i = SN_{SM_i} \parallel c_i \parallel t_i \parallel \sigma_i$ to $CC$.
endfor
for 1 to $n$
    $CC$ check timestamp $t_i$, $T_i = R_i + h_i \cdot PK_i$ and $\sigma_l$.
    if $e(P, S) = e(P_{pub}, T)$ then
        CC gets data through $m_i = C_2/e(C_1, C_3)$.
    endif
endfor
```

Figure 2: Data Generation and Signature Verification Algorithm

3.3.2 ASV-FLD Algorithm

Aggregation signature verification and fast location algorithm based on Dichotomy (ASV-FLD) is an aggregate signature verification algorithm and a fast location algorithm based on dichotomy. The algorithm consists of two parts: aggregate signature verification and fast signature location. Generally, the signature verification method is one by one, as mentioned above. However, it would be more convenient for us to aggregate and verify the n signatures received.

The CC receives n signatures $\sigma_1, \sigma_2, \cdots, \sigma_i, \cdots, \sigma_n$ by $U_1, U_2, \cdots, U_i, \cdots, U_n$ at time t_i. For U_i, $\sigma_i = (R_i, S_i)$. Let $R = \sum_{i=1}^{n} R_i$, $S = \sum_{i=1}^{n} S_i$ and $T = \sum_{i=1}^{n} T_i$. Therefore, the aggregation signature verification is to judge whether $e(P, S) = e(P_{pub}, T)$. Its correctness is proved as follows.

By using aggregation method, all signatures can be verified by only one pairing operation, and multi-signature verification can be realized quickly without performing n-times pairing operation. In general, we think that pairing is more time-consuming than addition. However, this method has a problem, that is, if the signer is mixed with dishonest or malicious participants, the whole signature verification will fail. Therefore, we need to quickly locate the location of the failed signature when the aggregation verification fails, to eliminate the task so as not to affect the activities of other participants. To solve this problem, we design a fast location algorithm based on dichotomy. The main idea is shown in Figure 3.

$$
\begin{aligned}
e(P, S) &= e\left(P, \sum_{i=1}^{n} S_i\right) \\
&= e\left(P, S_1 + S_2 + \cdots + S_n\right) \\
&= e\left(P, S_1\right) \cdot e\left(P, S_2\right) \cdots e\left(P, S_n\right) \\
&= e\left(P, r_1 \cdot P_{pub} + h_1 \cdot SK_1\right) \cdot e(P, r_2 \cdot P_{pub} \\
&\qquad + h_2 \cdot SK_2) \cdots \left(P, r_n \cdot P_{pub} + h_n \cdot SK_n\right) \\
&= e\left(P, r_1 \cdot P_{pub}\right) \cdot e\left(P, h_1 \cdot SK_1\right) \\
&\qquad \cdot e\left(P, r_2 \cdot P_{pub}\right) \cdot e\left(P, h_2 \cdot SK_2\right) \\
&\qquad \cdots e\left(P, r_n P_{pub}\right) \cdot e\left(P, h_n \cdot SK_n\right) \\
&= e\left(P, r_1 s P\right) e\left(P, h_1 \cdot s \cdot PK_1\right) e\left(P, r_2 \cdot s \cdot P\right) \\
&\qquad \cdot e\left(P, h_2 \cdot s \cdot PK_2\right) \cdots e\left(P, r_n \cdot s \cdot P\right) \\
&\qquad \cdot e\left(P, h_n \cdot s \cdot PK_n\right) \\
&= e\left(P_{pub}, R_1\right) e\left(P_{pub}, h_1 PK_1\right) \cdot e\left(P_{pub}, R_2\right) \\
&\qquad \cdot e\left(P_{pub}, h_2 \cdot PK_2\right) \cdots e\left(P_{pub}, R_n\right) \\
&\qquad \cdot e\left(P_{pub}, h_n \cdot PK_n\right) \\
&= e\left(P_{pub}, R_1 + h_1 \cdot PK_1\right) \cdot e\left(P_{pub}, R_2 + h_2 \cdot PK_2\right) \\
&\qquad \cdots e\left(P_{pub}, R_n + h_n \cdot PK_n\right) \\
&= e\left(P_{pub}, T_1\right) \cdot e\left(P_{pub}, T_2\right) \cdots e\left(P_{pub}, T_n\right) \\
&= e\left(P_{pub}, T_1 + T_2 + \cdots + T_n\right) \\
&= e\left(P_{pub}, T\right).
\end{aligned}
$$

In Figure 3, we assume that user U_i provides a false signature, which results in the failure of aggregate signature verification. We want to quickly locate users who provide failed signatures. We can use dichotomy to achieve this. In the round 1, we divide n users into two groups equally, and then do aggregate signature verification respectively. For the left group, let $R_{left} = \sum_{j=1}^{i} R_j$, $S_{left} = \sum_{j=1}^{i} S_j$ and $T_{left} = \sum_{j=1}^{i} T_j$. Our goal is to determine whether $e(P, S_{left})$ and $e(P_{pub}, T_{left})$ are equal. In the same way, judge whether the aggregate signature verification of the group on the right meets. The result of the hypothesis test is that the group on the right-side passes, but the group on the left-side fails. Currently, we need to take the round 2 of verification and continue to use this method to determine the left and right aggregate signature results that fail to pass this group. And so on until the user with the failed signature is located. The time complexity of the algorithm is $O(logn)$.

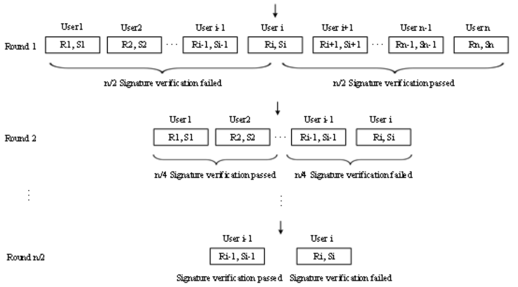
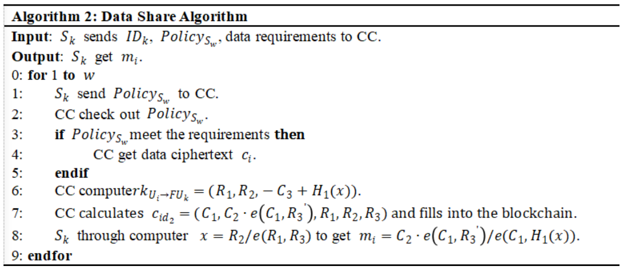


Figure 3: Fast location algorithm based on dichotomy

3.3.3 Data Share Algorithm

In the Figure 4, suppose an outside visitor S_k, such as a salesman, wants to visit m_i, he can get data through the CC and user authorization through the following operations. First, the salesman S_k sends the identity id_k, permission $Policy_{S_w}$ and data request to the CC.



```
Algorithm 2: Data Share Algorithm
Input: $S_k$ sends $ID_k$, $Policy_{S_w}$, data requirements to CC.
Output: $S_k$ get $m_i$.
for 1 to $w$
    $S_k$ send $Policy_{S_w}$ to CC.
    CC check out $Policy_{S_w}$.
    if $Policy_{S_w}$ meet the requirements then
        CC get data ciphertext $c_i$.
    endif
    CC computer$k_{U_i \rightarrow FU_k} = (R_1, R_2, -C_3 + H_1(x))$.
    CC calculates $c_{id_2} = (C_1, C_2 \cdot e(C_1, R_3^{'}), R_1, R_2, R_3)$ and fills into the blockchain.
    $S_k$ through computer $x = R_2/e(R_1, R_3)$ to get $m_i = C_2 \cdot e(C_1, R_3^{'})/e(C_1, H_1(x))$.
endfor
```

Figure 4: Data Share Algorithm

Then, after receiving its data request, the CC checks whether its permission is met. If satisfied, the CC extracts data and obtains ciphertext c_i, select $x \in \mathbb{G}_T$, calculate $(R_1, R_2, R_3) = Encrypt(params, id_k, x)$, computes the re-encryption key $RK_{U_i \longrightarrow S_k}$ by $RK_{U_i \longrightarrow S_k} = (R_1, R_2, -C_3 + H_1(x) \longrightarrow R_3^{'})$. The CC calculates the re-encrypted cipher text C_{id_2} and fills the information $P_i = SN_{SM_i} \parallel c_{S_k} \parallel T_i$ into the blockchain for S_k to access. S_k go to the main chain SB to get the data by computing $x = R_2/e(R_1, R_3)$ and decrypting the $m_i = C_2 \cdot e(C_1, R_3^{'})/e(C_1, H_1(x))$.

4 Security Analysis

We consider the security of the ID-based aggregate signature scheme and ID-based encryption scheme. We allow an adversary to corrupt all but one honest signer U_i while analyzing the security of our aggregate signature.

4.1 Security Proof

Theorem 1. *The ID-based aggregate signature scheme is secure against existential forgery on adaptively chosen message and ID attack in the random oracle model if CDHP in $\mathbb{G}$ is hard.*

Proof. Using the similar method given in [6], there exists an algorithm $\mathcal{A}_\infty$ for adaptively chosen message and given ID attack, if there has a polynomial time algorithm $\mathcal{A}_\prime$ with the same advantage for an adaptively chosen message and ID attack to our scheme. □

We assume that the identity of U_i is id_i, the corresponding public-private key pair is (pk_i, sk_i). According to the Forking Lemma in [24], if there exists an efficient algorithm $\mathcal{A}_\infty$ for an adaptively chosen message and given ID attack to the scheme, then there exists an efficient algorithm $\mathcal{B}_\prime$, which can produce two valid signatures $sigma_1 = (m, R, h_1, S_1)$ and $\sigma_1^* = (m, R, h_1^*, S_1^*)$, where $h_1 \neq h_1^*$. An algorithm $\mathcal{B}_\infty$, which is as efficient as $\mathcal{B}_\prime$. P, P_{pub} and pk_i are input to algorithm $\mathcal{B}_\infty$, which $P_{pub} = s \cdot P$ and $pk_i = t \cdot P$, for some $t \in \mathbb{Z}_q^{'}$. For the message m, runs $\mathcal{B}_\prime$ to obtain two forgeries σ_1 and $\sigma_1^{'}$, then $e(P, S_1) = e(P_{pub}, R + h_1 \cdot PK_i)$ and $e(P, S_1^*) = e(P_{pub}, R + h_1^* \cdot PK_i)$. Since e is non-degenerate, we have $(S_1 - S_1^*) - (h_1 - h_1^*) \cdot SK_i = 0$ and $SK_i = (h_1 - h_1^*)^{-1} \cdot (S_1 - S_1^*)$. So $e(P, (S_1 - S_1^*) - (h_1 - h_1^*) \cdot SK_i) = 1$. It means that algorithm $\mathcal{B}_\infty$ can solve an instance of CDHP in $\mathbb{G}$ since $sk_i = s \cdot PK_i$. Therefore, $\mathcal{A}_\infty$ will be able to succeed in forgery whenever $\mathcal{A}_\prime$ is successful. There is no efficient algorithm for an adaptively chosen message and given ID attack to our scheme since CDHP in $\mathbb{G}$ is hard. Therefore, the scheme is secure against existential forgery under adaptively chosen message and ID attack. Therefore, the theorem 1 is proved.

Security definitions for Identity-Based Encryption (see [9]) address the case where keyholders collude by combining secrets, as shown in Definition 2. We choose all these properties into a single game, by providing re-encryption keys to the adversary via an oracle. When the adversary possesses re-encryption keys, we must restrict it in some ways to avoid a trivial necessary condition. For example, to prevent it from obtaining a set of re-encryption keys leading from the challenge identity id_1 to some identity for which the adversary holds a decryption key. We consider chosen-plaintext security (CPA) and chosen-ciphertext secure (CCA).

Theorem 2. *(Security of Non-Interactive ID-based encryption scheme) Let $\mathcal{S}$ is an ID-based encryption scheme defined as a tuple of algorithms (KeyGen, Extract, Encrypt and Decrypt). Security is defined according to the following game $Exp^{\mathcal{A},IND-ID-ATK}$, where $ATK \in (CPA, CCA)$.*

We will describe the security proof for the proposed scheme with a game between a challenger C and a type attacker $\mathcal{A}$. The game is designed as follow:

Setup. C run $KeyGen(1^k)$ to get $(param, s)$, and give $params$ to $\mathcal{A}$.

Find phase 1. $\mathcal{A}$ makes the queries $(extract, encrypt, decrypt)$. If $ATK = CPA$, the queries $(extract, decrypt)$ are answered with $\perp$. On $(extract, id)$, Creturn $KeyGen(params, s, id)$. On $(decrypt, id, c)$, extract $sk = KeyGen(params, s, id)$and return $Decrypt(params, sk, c)$. At the end of this phase, $\mathcal{A}$ selects $id^* \in \{0,1\}^*$ and $(m_0, m_1) \in \mathcal{M}$. $\mathcal{A}$ is restricted to choices of id^* such that "trivial" decryption is not possible using keys extracted during this phase (e.g., by using re-encryption keys to translate from id to identity id^* for which $\mathcal{A}$ holds a decryption key).

Challenge phase. When $\mathcal{A}$ presents $(choice, id^*, m_0, m_1)$, C select $i \in \{0,1\}$ and compute $c^* = Encrypt(params, id^*, m_i)$, return c^* to $\mathcal{A}$. The id^* must be not queried in the Find phase 1.

Find phase 1. As in the Find phase 1, except that $\mathcal{A}$ cannot query the id^*'s private key and the decrypt query.

Guess phase. $\mathcal{A}$ makes queries with the following restrictions.

1) $\mathcal{A}$ is restricted from querying on $(extract, id)$ if there exists a challenge derivative (id, c).
2) $\mathcal{A}$ is restricted from querying on $(decrypt, id, c)$ if (id, c) is a challenge derivative.

The outcome of the game is determined as follows: if $i = i^{'}$, wins the game. $\mathcal{A}$'s advantage in the above game, $Adv_{\mathcal{A}}^{IND-ID-ATK}$ is defined as $\left|[i = i^{'}] - 1/2\right|$. For $ATK \in (CPA, CCA)$, we say that the ID-based encryption scheme $\mathcal{S}$ is $IND - ID - ATK$ -secure if for all probabilistic polynomial time algorithms $\mathcal{A}$, $Adv_{\mathcal{A}}^{IND-ID-ATK} \leq \varepsilon$, where ε is defined as a negligible function.

Theorem 3. *If a type adversary $\mathcal{A}$ that the probability of winning game in any polynomial time t is a non-negligible probability ε exists, there must be a challenger C which can solve the DBDH with no more than q_1 private key extraction queries, therefore this scheme is $(t, q_1, \varepsilon) - IND - ID - CPA$-secure.*

Proof. $\mathcal{A}$ is a type adversary winning the game with probability ε. C is a challenger in polynomial time that aims to solve DBDH and calls $\mathcal{A}$ as a subroutine. For user's private key sk_i, $\mathcal{A}$ must calculate the master key $msk = s$ to obtain or generate sk_i, because $sk_i = pk_i \cdot s$. To get s, consider $P_{pub} = s \cdot P$. Since P, P_{pub} is known, the problem is a discrete logarithm problem CBDH with computational complexity.

When $\mathcal{A}$ does not know the user private key SK_i and obtains ciphertext $c_i = (r_i \cdot P, m_i \cdot e(s \cdot P, H(SN_{SM_i}))^{r_i})$ through channel monitoring, $\mathcal{A}$ needs to calculate $e(s \cdot P, H(SN_{SM_i}))^{r_i} = e(P_{pub}, PK_i)^{r_i}$. Because P_{pub}, PK_i is known, $\mathcal{A}$ needs to know r_i to decrypt successfully. To get r_i, consider $R_i = r_i \cdot P$. Similarly, to obtain r_i, the problem is a discrete logarithm problem CBDH with computational complexity. The problem can be transformed into known P_{pub}, PK_i, P, R_i, solve $e(P_{pub}, PK_i)^{r_i}$. This is an DBDH complexity problem, which is equivalent to CBDH and has computational complexity. Using the similar method given in [15,29], Based on the Waters system, if DBDH holds, then $(t, q_1, \varepsilon) - IND - ID - CPA$ is safe. Therefore, Theorem 3 is proved. □

4.2 Security Analysis

Message integrity and non-repudiation. According to the theorem 1, we can know that when the CDHP is hard, no attacker can forge a valid signature in a polynomial time. Therefore, the receiver can check the message $(SN_{SM_i}, c_i, T_i, \sigma_i)$, by verifying if $e(P, S) = e(P_{pub}, T)$. Therefore, the proposed scheme satisfies the integrity and non-repudiation.

Privacy protection. According to the theorem 2 and theorem 3, we can know that when the EBDH is hard, no attacker can forge a forgery message r_i in polynomial time. Once it is hard to calculate r_i from $R_i = r_i \cdot P$, so as to m. Therefore, illegal participants cannot use to decrypt information $m = C_2 \cdot e(C_1, R_3)/e(C_1, H_1(x))$. Therefore, the proposed scheme satisfies the privacy protection.

Resistance against different types of attacks. Resist Forgery Attack. To forge a valid signature, the attacker must generate a message $(SN_i, c_i, T_i, \sigma_i^* = (R_i^*, S_i^*))$. According to the theorem 1, the receiver can calculate $T_i^* = R_i^* + h_i \cdot pk_i$, check $e(P, S_i^*) \neq e(P_{pub}, T_i^*)$. Consequently, our scheme can oppose forgery attack.

Resist man-in-the-middle and modification attack. According to the above theorem 1, theorem 2 and theorem 3, we can know that the communication among the user, salesman and the control center needs to be verified and valid message cannot be forged and modified. Therefore, this scheme can oppose man-in-the-middle attack.

Resist replay attack. Timestamp mechanism is an effective way to resist replay attacks. By embedding a timestamp in the message, the receiver can verify the time to prevent dishonest participants from sending past information. Therefore, this scheme can oppose replay attack.

5 Digital Experiment and Analysis

In order to assess our solution, we ran experiments on a Fedora 35 virtual machine platform with an Intel Core I5-10400 CPU@2.90 GHz processor and 8Gb of memory. The pYPBC 0.2 library was used to simulate our scheme. Furthermore, the curves that are employed for pairing are the basic curves in the PYPBC library, and the safety parameters of the curves are $qbits = 512$ and $Rbits = 160$. The algorithm, written in Python, estimates all the results by averaging 100 experiments.

5.1 Computational Complexity

The scheme can consider the computational complexity in stages. The time is: the exponential operation on each domain $\mathbb{Z}_{n^2}^*$ takes T_1, the exponential operation on each group $\mathbb{G}$ takes T_2, the multiplication operation on each group $\mathbb{G}$ takes T_3, and the multiplication operation on each field $\mathbb{Z}_{n^2}^*$ takes T_4, the logarithm operation on each domain $\mathbb{Z}_{n^2}^*$ during decryption takes T_5. In the initialization phase, registration, cloud upload, blockchain creation, and decryption read phase, all operations are performed offline, so their computational complexity is not considered. For calculation convenience, $K = 1$. In the data generation stage, generating one c_i requires performing two exponential operations on field $\mathbb{Z}_{n^2}^*$, and generat-

Table 1: Performance analyzes. Exp, Mul is the abbreviation of exponential and multiplication

Basic Operation	Exp on $\mathbb{Z}$	Exp on $\mathbb{G}$	Mul on $\mathbb{G}$	Mul on $\mathbb{Z}$	Logarithm
Time Complexity	$T_1 = 0.0772$	$T_2 = 0.7415$	$T_3 = 0.7315$	$T_4 = 0.0662$	$T_5 = 0.5333$
Computational	Our Data Generation	EPPA [20]	PEDA [10]	SEDA [23]	–
Time Complexity	$2nT_1 + 2nT_2 + nT_3$	$3nT_1 + 3nT_2 + nT_3)$	$8T_1 + 2nT_2 + nT_3 + 5T_5$	$3T_1 + 3nT_2 + 3nT_3 + 3T_5$	–
Computational	Our Signature	EPPA [20]	PEDA [10]	SEDA [23]	–
Time Complexity	$2(n+1)T_2 + (n+3)T_3 + 4T_5$	$(n+3)T_5 + (n+1)T_3$	$(n+1)T_5 + (2n+1)T_2 + (n+1)T_3$	$2T_5 + (6n+3)T_2 + nT_3$	–
Computational	Our Date Aggregation	EPPA [20]	PEDA [10]	SEDA [23]	–
Time Complexity	$3T_2 + 2T_3 + nT_4 + 2T_5$	$n(T_2 + T_3 + T_4 + T_5)$	$T_2 + T_3 + (n+5)T_5$	$2nT_4 + nT_3$	–
Computational	Our Data Reading	EPPA [20]	PEDA [10]	SEDA [23]	–
Time Complexity	$T_2 + T_3 + 4T_5$	$2T_2 + T_3 + 5T_5)$	$2T_2 + T_3 + nT_5$	$T_4 + (n+1)T_5$	–
Communication	Phase: User to CC	Phase: CC to Salesman	–	–	–
Load	$o_1 = n * (L_1 + L_2 + L_3 + L_4)$	$o_2 = L_1 + L_2 + L_3 + L_4$	–	–	–

ing one σ_i requires performing two exponential operation on group $\mathbb{G}$ and one multiplication operation on group $\mathbb{G}$. In the data aggregation stage, verifying the signature needs to perform one exponential operation on group $\mathbb{G}$, a multiplication operation on the group $\mathbb{G}$ and two logarithmic operations on the domain $\mathbb{Z}^*_{n^2}$. Aggregation node aggregates encrypted data needs to perform n multiplication operations on the domain $\mathbb{Z}^*_{n^2}$, and generate a signature σ_{A_j} needs to perform two exponential operation on group $\mathbb{G}$ and a multiplication operation on the group $\mathbb{G}$. In the data reading stage, the CC verifies the signature of the consensus node σ_{TC_t} needs to perform two logarithmic operations on the domain $\mathbb{Z}^*_{n^2}$, one exponential operation on group $\mathbb{G}$, a multiplication operation on the group $\mathbb{G}$ and the decrypted data needs to be in the domain $\mathbb{Z}^*_{n^2}$ perform two logarithm operations on it. In the sharing stage, the encrypted and decrypted data need to perform two exponential operations on the domain $\mathbb{Z}^*_{n^2}$.

The time complexity of the scheme is $T_{Total} = 2nT_1 + 2nT_2 + nT_3 + T_2 + T_3 + 2T_5 + nT_4 + 2T_2 + T_3 + 2T_5 + T_2 + T_3 + 2T_1 + 4T_1$. The time complexity of performing the multiplication operation on the domain $\mathbb{Z}^*_{n^2}$ is negligible compared with other operations, so the time complexity of the scheme is $T_{Total} \approx 2(n+3)T_1 + 2(n+2)T_2 + (n+3)T_3 + 4T_5$. And the time complexity of signature verification is $T_{sig} = 2(n+1)T_2 + (n+3)T_3 + 4T_5$. EPPA [20], PEDA [10], and SEDA [23] are all signature authentication algorithms designed in similar scenarios, and their time complexity is shown in Table 1. We first consider the comparison of our algorithm and scheme EPPA, PEDA, SEDA, as shown in the Figure 5. To sum up, the real-time performance of this scheme is better.

5.2 Communication Load

This paper considers the communication load of three processes, namely user to the control center, and the control center to a salesman. In the process from the user to the control center, SM_isend to aggregation node A_j is $M_i = SN_{SM_i} \| c_i \| T_i \parallel \sigma_i$, the communication load of SN_{SM_i} is L_1, the communication load of c_i is L_2, the communication load of T_i is L_3, and the communication load of σ_i is L_4, then the communication load of this process is $O_1 = n \cdot (L_1 + L_2 + L_3 + L_4)$. From the control center to the salesman, the data from the control center to fill the blockchain is $P_i = SN_{SM_i} \parallel c_{S_k} \parallel T_i$. Because the blockchain needs to fill the hash value of the previous block (SHA-1), it is equal to the length of the signature, then the communication load of this process is $O_2 = L_1 + L_2 + L_3 + L_4$. In summary, the communication load of the scheme is $O = O_1 + O_2 = (n+1)(L_1 + L_2 + L_3 + L_4)$, as shown in Table 1. Through analysis, it is found that in the SM to the control center and the control center to the external salesman phase, the communication cost can reach a constant level.

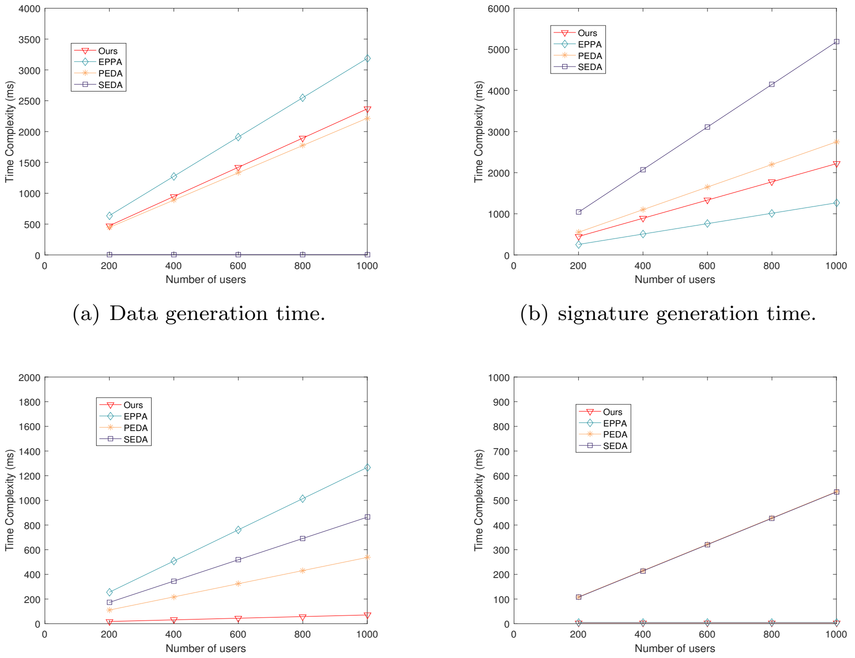
5.3 Simulation Experiment

To explore the time complexity relationship between each algorithm (signature, encryption and decryption, proxy re-encryption) and the number of users, we design a simulation experiment with the help of the PBC library. To fully observe the relationship between the number of users and the time complexity of the algorithm, we set the number of users to multiple levels. In this paper, set $n = \{1, 10, 50, 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$.It can be seen from Figure 5(a). In the data generation phase, Scheme SEDA [23] and our scheme take relatively short time. In the initialization phase of Scheme PEDA [10] and SEDA [23], RSA algorithm is used to generate the private key. It can be seen from Figure 5(b) that in the signature phase, the time of scheme EPPA [20] and our scheme is relatively short. It can be seen from Figure 5(c) that in the aggregate signature stage, our scheme adopts the ASV-FLD algorithm, which can greatly reduce the computational complexity while verifying the aggregate signature, and takes the shortest time. It can be seen from Figure 5(d) that in the Reading phase, our scheme still takes the shortest time.

After analyzing the time complexity of the three algorithms in this paper, we integrate each part of the scheme and simulate the whole scheme to explore the relationship between the time complexity and the number of users, as shown in Figure 4. From Figure 4, we can clearly see that each algorithm is approximately linear with the number of users.

6 Conclusion

This paper proposes a new privacy-preserving scheme. The concept of dual blockchain is proposed for the first time, which can improve query efficiency, reduce storage



(a) Data generation time.

(b) signature generation time.

(c) Aggregation time.

(d) Reading time.

Figure 5: Experiments on these steps

cost and further improve data security compared with single chain. Through security analysis, this scheme can effectively protect the integrity, traceability and confidentiality of user data. The experimental results show that compared with other schemes, it has less time complexity and communication overhead. However, this paper still has the trend of continuous optimization about the efficiency of real-time transmission.

References

[1] G. Ács and C. Castelluccia, "I have a dream!(differentially private smart metering)," in *International Workshop on Information Hiding*, Springer, pp. 118–132, 2011.

[2] N. Z. Aitzhan and D. Svetinovic, "Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 5, pp. 840–852, 2016.

[3] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in *Annual international cryptology conference*, Springer, pp. 213–229, 2001.

[4] D. Boneh and X. Boyen, "Efficient selective-id secure identity-based encryption without random oracles," in *International conference on the theory and applications of cryptographic techniques*, Springer, pp. 223–238, 2004.

[5] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the weil pairing," in *International conference on the theory and application of cryptology and information security*, Springer, pp. 514–532, 2001.

[6] X. Cheng, J. Liu, and X. Wang, "Identity-based aggregate and verifiably encrypted signatures from bilinear pairing," in *International Conference on Computational Science and Its Applications*, Springer, pp. 1046–1054, 2005.

[7] P. S. Chung, C. W. Liu, and M. S. Hwang, "A study of attribute-based proxy re-encryption scheme in cloud environments", *International Journal of Network Security*, vol. 16, no. 1, pp. 1-13, 2014.

[8] H. Deng, Z. Qin, Q. Wu, Z. Guan, R. H. Deng, Y. Wang, and Y. Zhou, "Identity-based encryption transformation for flexible sharing of encrypted data in public cloud," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3168–3180, 2020.

[9] H. Deng, Z. Qin, Q. Wu, Z. Guan, and Y. Zhou, "Flexible attribute-based proxy re-encryption for efficient data sharing," *Information Sciences*, vol. 511, pp. 94–113, 2020.

[10] C.-I. Fan, S.-Y. Huang, and Y.-L. Lai, "Privacy-enhanced data aggregation scheme against internal attackers in smart grid," *IEEE Transactions on Industrial informatics*, vol. 10, no. 1, pp. 666–675, 2013.

[11] M. S. Ferdous, A. Margheri, F. Paci, M. Yang, and V. Sassone, "Decentralised runtime monitoring for access control systems in cloud federations," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, pp. 2632–2633, 2017.

[12] M. Green and G. Ateniese, "Identity-based proxy re-encryption," in *International Conference on Ap-*

plied Cryptography and Network Security, Springer, pp. 288–306, 2007.

[13] T. Hardjono and N. Smith, "Cloud-based commissioning of constrained devices using permissioned blockchains," in *Proceedings of the 2nd ACM international workshop on IoT privacy, trust, and security*, pp. 29–36, 2016.

[14] M. S. Hwang, S. F. Tzeng, C. S. Tsai, "Generalization of proxy signature based on elliptic curves", *Computer Standards & Interfaces*, vol. 26, no. 2, pp. 73–84, 2004.

[15] M. Joye and G. Neven, *Identity-based cryptography*, IOS press, vol. 2, 2009.

[16] I. C. M. Kumar, K. Dutta, , "Impact of wormhole attack on data aggregation in hieracrchical WSN," *International Journal of Electronics and Information Engineering*, vol. 1, no. 2, pp. 70-77, 2014.

[17] L. H. Li, S. F. Tzeng, M. S. Hwang, "Generalization of proxy signature based on discrete logarithms", *Computers & Security*, vol. 22, no. 3, pp. 245–255, 2003.

[18] X. Liang, X. Li, R. Lu, X. Lin, and X. Shen, "Udp: Usage-based dynamic pricing with privacy preservation for smart grid," *IEEE Transactions on smart grid*, vol. 4, no. 1, pp. 141–150, 2013.

[19] E. J. L. Lu, M. S. Hwang, and C. J. Huang, "A new proxy signature scheme with revocation", *Applied Mathematics and Computation*, vol. 161, no. 3, PP. 799-806, Feb. 2005.

[20] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.

[21] S. Maiti and S. Misra, "P2B: Privacy preserving identity-based broadcast proxy re-encryption," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5610–5617, 2020.

[22] M. Mylrea and S. N. G. Gourisetti, "Blockchain for smart grid resilience: Exchanging distributed energy at speed, scale and security," in *2017 Resilience Week (RWS)*, IEEE, pp. 18–23, 2017.

[23] J. Ni, K. Alharbi, X. Lin, and X. Shen, "Security-enhanced data aggregation against malicious gateways in smart grid," in *2015 IEEE Global Communications Conference (GLOBECOM)*, IEEE, pp. 1–6, 2015.

[24] D. Pointcheval and J. Stern, "Security arguments for digital signatures and blind signatures," *Journal of cryptology*, vol. 13, no. 3, pp. 361–396, 2000.

[25] S. Tonyali, O. Cakmak, K. Akkaya, M. M. Mahmoud, and I. Guvenc, "Secure data obfuscation scheme to enable privacy-preserving state estimation in smart grid ami networks," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 709–719, 2015.

[26] B. Waters, "Efficient identity-based encryption without random oracles," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, pp. 114–127, 2005.

[27] Q. Xia, E. B. Sifah, A. Smahi, S. Amofa, and X. Zhang, "Bbds: Blockchain-based data sharing for electronic medical records in cloud environments," *Information*, vol. 8, no. 2, p. 44, 2017.

[28] C.-M. Yu, C.-Y. Chen, S.-Y. Kuo, and H.-C. Chao, "Privacy-preserving power request in smart grid networks," *IEEE Systems Journal*, vol. 8, no. 2, pp. 441–449, 2013.

[29] J. Zhang, Y. Zhao, J. Wu, and B. Chen, "Lpda-ec: A lightweight privacy-preserving data aggregation scheme for edge computing," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, IEEE, pp. 98–106, 2018.

Biography

Xiaoxu Zhang received the Bachelor degree from Tianjin University of Technology in 2020. She is currently studying for a Master degree at North China Electric Power University. Her current research interest focus on security and privacy issues in cloud computing, storage security, and applied cryptography. & Email: zhangxiaoxu@ncepu.edu.cn

Deep Security Detection Framework Based on ATT&CK

Rixuan Qiu[1], Yu Fu[1], Jian Le[2], Fuyong Zheng[1], Gan Qi[2], Chao Peng[1], Yi Zhang[3], Yuancheng Li[3], and Yan Liu[3]

(Corresponding author: Jian Le)

Dispatch Monitoring Center, State Grid Jiangxi Information & Telecommunication Company[1]
Nanchang, China
School of Electrical Engineering and Automation, Wuhan University[2]
Wuhan 430072, China
School of Control and Computer Engineering & North China Electric Power University[3]
Beijing 102206, China
Email: ncepua@163.com

(Received Aug. 28, 2022; Revised and Accepted Dec. 5, 2022; First Online Dec. 14, 2022)

Abstract

It is urgent to establish an efficient and comprehensive security system architecture facing the increasingly severe network threat. Aiming at the security problems in information systems, a deep security detection framework based on ATT&CK is proposed. There are three modules: information collection, analysis engine, and security response. The anomaly detection results of the technical layer based on ATT&CK knowledge base association analysis are mapped to the attack links of the tactical layer, thus realizing the bottom-up system security detection. An improved transformer anomaly detection method based on FLOATER positional coding is applied in the security detection framework to realize the early detection of network threats. Based on the continuous dynamic system, the positional coding function is learned to capture the temporal features of the sequence. Then multi-head attention is used to encode the input sequence. Moreover, the model training is optimized by the adversarial method. The experimental results demonstrated that the method applied in this work outperforms traditional deep learning models in terms of detection accuracy and F1 score.

Keywords: Anomaly Detection; ATT&CK; Security Architecture; System Security; Transformer

1 Introduction

With the advancement of network communication technology, an increasing number of businesses and organizations are conducting daily office and data storage through information systems, internal networks, or cloud services. This way simplifies people's lives and increases work efficiency, but it also introduces new risks. The impact will be enormous once the network and communication system is subjected to malicious attacks and significant security incidents. In recent years, network attackers have expanded their capabilities beyond short-term direct attacks and are now frequently launching APT attacks, posing a new threat to the traditional security detection architecture. APT attacks, also known as Advanced Persistent Threat, are launched by a group of attackers with the assistance of the government or major corporations. They are talented, well-organized, and do enormous damage. They also have defined goals [19]. Building a more effective and complete security detection framework is therefore crucial for security operation and maintenance.

The security detection architecture that exclusively depends on intrusion detection and firewall approaches is unable to deal with APT attacks because of their high concealment and protracted hidden time. The MITRE ATT&CK architecture illustrates a novel approach to APT attack detection and defence. The ATT&CK architecture records adversary attack behaviors and categorizes them into tactics, techniques, and procedures (TTPs) [1]. The knowledge base makes up for the shortcomings of the advanced attack model and pure technical analysis and can better build a security system and form a security closed loop by combining the specific threat behaviour at the bottom layer with the attack path at the top layer and conducting in-depth correlation analysis on the attacker's behaviour.

One popular approach for security detection is anomaly-based intrusion detection. Anomaly detection algorithms that incorporate deep learning have started to emerge with the development of high-speed large-scale networks, such as deep automatic encoder [23], convolutional neural network [12], recurrent neural network [21], and transformer [20]. Even though some network threats have been detected by the available techniques, more sophisticated and covert attacks can still go undetected due to a lack of broad detection techniques.

This research proposes a deep security detection framework based on ATT&CK, in which an enhanced transformer anomaly detection approach based on FLOATER positional coding is implemented, in order to enhance the threat detection effect of information systems. The framework incorporates ATT&CK's concepts, maps and analyses the attackers' fundamental tactics and techniques, and then realizes real-time system security protection by capturing the attackers' attack chains and final targets. In order to achieve more accurate and effective anomaly identification, the framework's anomaly detection technique models the position information based on a continuous dynamic system, inserts positional coding in each transformer encoder module, and performs adversarial training through two decoders. According to experimental findings, the suggested technique successfully detects two publicly available datasets, increasing system security and accuracy.

In this paper, our contributions are summarized as follows:

- For APT attacks, a deep security detection architecture is designed based on ATT&CK, which realizes the bottom-up security detection and protection of information systems.
- An improved transformer anomaly detection method based on FLOATER positional coding is implemented in the architecture, which improves the detection accuracy through temporal modelling and adversarial training.

The remainder of this paper is organized as follows. Section 2 presents a review of previous work related to anomaly detection, including classical methods and deep learning methods. The third chapter introduces the deep security detection framework designed in this paper based on ATT&CK. Section 4 introduces the anomaly detection methods applied in this framework, including data preprocessing, positional coding methods, and model details. Section 5 shows the detection performance results and comparative experiments, which are described in detail. Section 6 summarizes the full text.

2 Related Work

Intrusion detection is a long-term academic research topic that can be divided into two categories: rule-based intrusion detection methods and anomaly-based intrusion detection methods. Anomaly-based intrusion detection methods model the data information collected in the system or network to identify abnormal data, whereas rule-based intrusion detection methods establish a specific knowledge base based on known attack methods. Anomaly detection methods that are commonly used include traditional machine learning methods and deep learning methods.

Traditional machine learning methods. The unsupervised K-means clustering method was used by MacQueen *et al.* [10] to determine the threshold based on the distance from the entire data to the centroid. When the distance between a data point and the centroid exceeds a certain threshold, it is considered an abnormal point. Breunig *et al.* [3] proposed a LOF method based on density estimation that judges the abnormal degree of data by calculating the density of the local area as well as the density of its neighbouring points. Liu *et al.* [7] built an independent forest from independent trees and defined abnormal data as easily isolated outliers. The isolated outliers represented abnormal data in the case of recursive random segmentation of the dataset. Saha *et al.* [13]used the PCA principal component analysis method to reduce the dimensionality of the input data and calculate the eigenvectors, with abnormal data located far away from the eigenvectors for abnormality detection. Tian *et al.* [16] used a one-class support vector machine (SVM) for anomaly detection, separating the data and learning the ”normal” patterns using hyper planes in a multidimensional space.

Deep learning methods. DAGMM [23] proposed a deep autoencoder Gaussian mixture model, which is a density estimation-based anomaly detection method. The author employs the compression network to reduce the dimension of the feature space, and the obtained features, along with the reconstruction error, are fed into the estimation network to simulate the results of the Gaussian mixture model, after which the parameters of the Gaussian mixture model are obtained and the abnormal threshold is calculated. LSTM-NDT [4] employs LSTM as the neural network's backbone to model the input raw sequence nonlinearly and detect anomalies from prediction errors using a parameter-free dynamic thresholding method. At the same time, the author introduces the method of pruning and setting the minimum outlier to distinguish anomalies from noise in order to reduce the false positive and false positive rate. OmniAnomaly [15] proposed a multivariate time series anomaly detection method based on a stochastic recurrent neural network that generates reconstruction probabilities by combining gated recurrent units, variational autoencoders, and planar normalized flow. The authors take into account the time dependence of the input data as well as the randomness of multivariate time series, and use the Peak Over Threshold method to select an anomaly threshold automatically. The temporal hierarchical one-class network proposed by THOC [14] for time series anomaly detection is based on an extended recurrent neural network with skip connections to capture temporal dynamic features at different scales. During the hierarchical clustering process, the obtained multiple hyper spheres, that is, multi-scale vector data description, are used to define a one-class objective, the hyper spheres are encouraged to be orthogonal to each other during training, and a self-supervised task is added in the time domain. InterFusion [6] proposed an unsupervised method for modelling input data in multivariate time series data in terms of dependencies between different metrics and temporally sequential dependencies. The central idea is to use two random latent variables in conjunction with variational autoencoders to model normal patterns in multidimensional time series. In addition, the authors propose a Markov Chain Monte Carlo method for interpreting the detected anomalous results. Anomaly transformer [20] proposed an anomaly detection method based on reconstruction error, modelling time series by alternately stacking anomaly attention and feedforward layers. The abnormal attention is divided into two parts: the prior association, which is composed of the learnable Gaussian kernel, and the sequence association, which is calculated by the self-attention mechanism, and the difference is calculated by KL divergence. The author also proposes a minimax model optimization strategy that increases the reconstruction error between outliers and normal data.

3 Deep Security Detection Framework Based on ATT&CK

This paper proposes a deep security detection framework based on the ATT&CK framework (the structure is shown in the Figure 1), which is divided into three modules, namely information collection, analysis engine and security response.

Information collection module: It's split into two sections: data collection and knowledge base. The data collection section collects and monitors data in the information system network, such as system log files, internal and external network traffic data, audit records, system user behavior data, and other underlying data information. Network traffic data is captured using network traffic packet analysis tools such as sniffer, SNMP, NetFlow, Wireshark, and others. The setting of capture rules enables real-time monitoring of key assets and sensitive resources in the system, the filtering of invalid information, and the reduction of system threat false alarms. The data collection component obtains local system data information, and the knowledge base component acts as a supplementary knowledge base for local data, including the ATT&CK framework, the threat indicator IOC library, and the general vulnerability disclosure CVE. The ATT&CK framework supports the entire system architecture and serves as the prerequisite for realizing the correlation and mapping of data analysis results. The network threat information is then updated in real time with the help of shared threat intelligence in the IOC library and security flaws published in the CVE database to assist in detecting system security risks. System vulnerabilities and security threats can be discovered in time to improve security in today's increasingly severe world.

Analysis engine module: There are two sections: data processing and technical analysis. To avoid the problem of model accuracy decline caused by incomplete data and uneven distribution, the raw data obtained by the data collection part in the information collection module must be pre-processed through the data processing part, such as data cleaning, numericalization, regularization, normalization, and so on. Based on the ATT&CK framework, the technical analysis section covers operations such as deep learning models and mapping analysis. We performed the following actions: anomaly detection on collected network traffic, system calls, memory usage, and other monitoring values, using a natural language processing model for log files to obtain attack clues, and technical matching and correlation analysis based on the results of the above operations, combined with the ATT&CK framework. This section obtains abnormal discrimination criteria via deep neural network learning, performs a comprehensive detection and analysis of the system network, and then marks the sensitive operations and threat behaviours that occur in the system via system log analysis. The underlying technologies and attack strategies are associated, and potential attack paths in the system are mapped, based on the results of these two parts, according to the ATT&CK framework.

Security response module: Divided into two sections: security visualization and emergency response. The attack path, situational awareness, and system vulnerability analysis obtained by the analysis engine module are displayed through a user-friendly interface in the visualization section, making it easy for system security administrators to analyse the system security situation and review attack scenarios. In the emergency response section, the security detection system generates corresponding security alarms and emergency responses based on the analysis engine's output results, releases threat signals in a timely manner, blocks attack paths, and prevents attackers from further expanding and controlling the situation.

In general, the deep security detection framework designed in this paper is comprised of three modules: information collection, analysis engine, and security response. From asset monitoring to alarm response, the top-level attacker's behaviour is abstracted from the underlying abnormal or sensitive operations, allowing potential threats to the system to be detected in real time, intercepted in time, and responded to security via correlation mapping.

4 An Improved Transformer Anomaly Detection Method Based on FLOATER Positional Coding

The analysis engine module in the security detection framework of this paper applies an improved transformer anomaly detection method based on FLOATER positional coding, replaces the fixed coding method in the transformer with the coding method based on the contin-

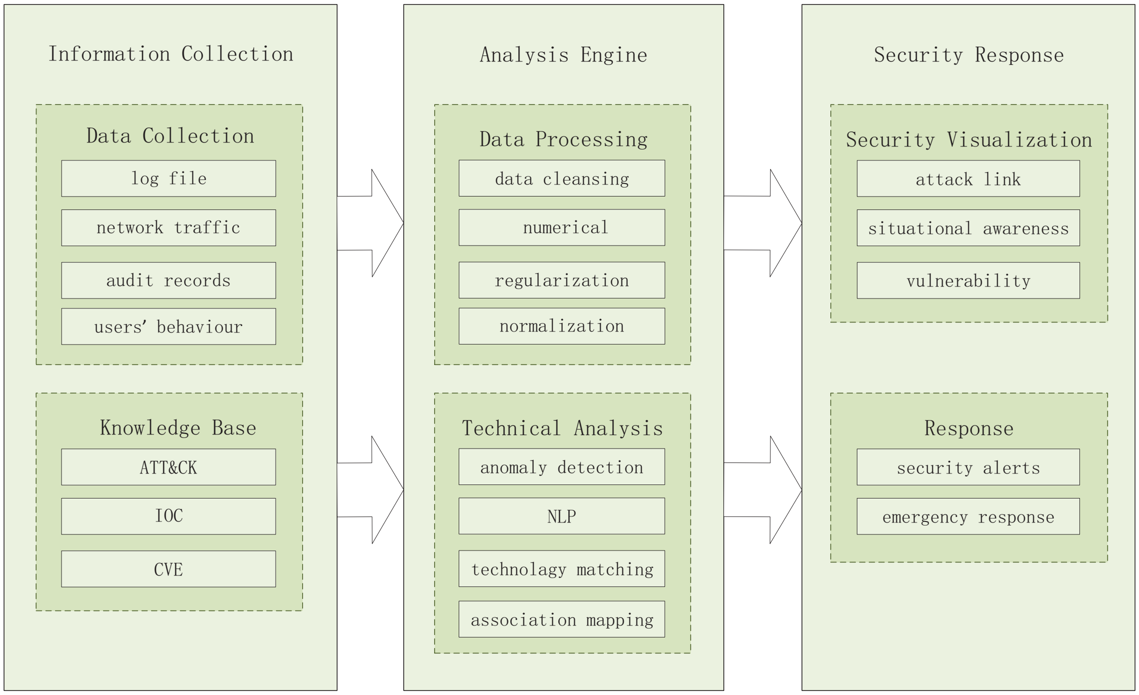


Figure 1: Deep security etection framework

uous dynamic system, and establishes a time model. Each encoder module is injected into the FLOATER block, so that it is trained at the same time to better grasp the temporal features. Convert the input sequence into a local context window sequence, input the two encoders Enc1 and Enc2 in the model, and conduct adversarial training through two identical decoders Dec1 and Dec2, and finally obtain the anomaly detection result. Model details are described in this section. The model structure as shown in Figure 2.

4.1 Data Pre-processing

For multivariate time series, suppose there are m observations at each time point, then the time series with length T is defined as:

$$\mathrm{T}=\{\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{T}\} \tag{1}$$

Among them, the time point data x_t has a timestamp t, and $\mathrm{x}_\mathrm{t}\in\mathbb{R}^m$. Normalize the time series as:

$$\mathrm{x}_\mathrm{t}=\frac{\mathrm{x}_\mathrm{t}-\min(\mathrm{T})}{\max(\mathrm{T})-\min(\mathrm{T})+\varepsilon'} \tag{2}$$

where $\min(\mathrm{T})$ and $\max(\mathrm{T})$ are the minimum and maximum variables over the entire pattern in the training set, and ε' is a small constant variable set to prevent division by zero.

In order to enhance the dependency of the time point data x_t, we turn the input time series into a local context window of length K:

$$\mathrm{W}_\mathrm{t}=\{\mathrm{x}_{\mathrm{t-K+1}},\ldots,\mathrm{x}_\mathrm{t}\} \tag{3}$$

When $\mathrm{t}<\mathrm{T}$, the window is copied and filled, and the constant variable x_t is added after the window. By working with window transformation, the input to the model will no longer be independent vectors, but data points with local contextual information [15].

4.2 FLOATER Positional Coding

The attention mechanism in the Transformer model [18] does not consider temporal information in the operation, and has position permutation invariance. However, time series information is an indispensable part in the analysis of time series, so additional positional coding processing is required. In this paper, a positional coding method named FLOATER [9] is adopted, which is a positional coding based on a continuous dynamic system, which is not limited by the input length. The encoding results are learned from the data and are suitable for use in transformer-based models with a small number of parameters introduced.

In the transformer model, the positional coding is a set of variables that are added to the input sequence, set to $\left\{p_i\in\mathbb{R}^d: i=1,\ldots,L\right\}$. Existing positional coding methods either use a fixed function (such as a sine function) to obtain $\{p_i\}$ or treat it as a non-learnable parameter, neither of which fails to capture the dependencies between positional coding $\{p_i\}$. This paper proposes to use the FLOATER method to replace the original position coding in the improved transformer, where the formula of the continuous dynamic system is as follows:

$$\begin{gathered}p(\mathrm{t})=p(\mathrm{s})+\textstyle\int_\mathrm{s}^\mathrm{t}\mathrm{h}(\tau,\mathrm{p}(\tau);\theta_\mathrm{h})\,d\tau,\\ 0\le s\le t<\infty\end{gathered} \tag{4}$$

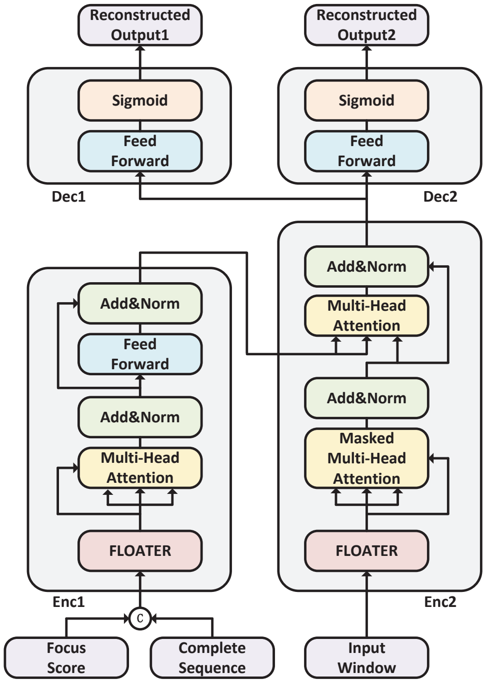


Figure 2: Improved transformer anomaly detection method based on FLOATER positional coding

Given an initial value $\mathrm{p}(0)$, where $\mathrm{h}(\tau, \mathrm{p}(\tau); \theta_{\mathrm{h}})$ is a neural network with parameters θ_{h}. The function value $\mathrm{p}(\mathrm{t})$ at time t is calculated from the function value $\mathrm{p}(\mathrm{s})$ at the previous time s and the increment (i.e. integral) of the function between time s and time t. Here $\mathrm{p}()$ is a function whose definition domain is positive real number domain, and the positional coding domain used in the model is the natural number domain. Therefore, the mapping relationship between the two can be established through the function $\mathrm{p}()$, thereby obtaining the transformer model in positional coding.

Injecting the positional coding into each encoder block of the transformer model can improve the performance of the model, for which we add a superscript (n) to describe the dynamics of the nth self-attention block:

$$\begin{aligned}&\mathrm{p}^{(\mathrm{n})}(\mathrm{t}) = \\ &\mathrm{p}^{(\mathrm{n})}(\mathrm{s}) + \textstyle\int_{\mathrm{s}}^{\mathrm{t}} \mathrm{h}^{(\mathrm{n})}\left(\tau, \mathrm{p}^{(\mathrm{n})}(\tau); \theta_{\mathrm{h}}^{(\mathrm{n})}\right) \mathrm{d}\tau\end{aligned} \tag{5}$$

Many dynamic models will introduce too many parameters and increase the training overhead, FLOATER solves this problem by sharing parameters on model blocks, as shown below:

$$\theta_{\mathrm{h}}^{(1)} = \theta_{\mathrm{h}}^{(2)} = \ldots = \theta_{\mathrm{h}}^{(\mathrm{N})} \tag{6}$$

In the standard transformer, the positional coding and the sequence to be processed are added together as the input of the model. FLOATER changes this addition method and can be initialized directly from the transformer model. Let the query matrix $\mathrm{Q}^{(\mathrm{n})}$ of the nth block in the standard transformer be:

$$\mathrm{q'}_{\mathrm{i}}^{(\mathrm{n})} = \mathrm{W}_{\mathrm{q}}^{(\mathrm{n})}\left(\mathrm{x}_{\mathrm{i}} + \mathrm{p'}_{\mathrm{i}}^{(\mathrm{n})}\right) + \mathrm{b}_{\mathrm{q}}^{(\mathrm{n})} \tag{7}$$

Where $\mathrm{W}_{\mathrm{q}}^{(\mathrm{n})}$ and $\mathrm{b}_{\mathrm{q}}^{(\mathrm{n})}$ are parameters in the transformer model, $\mathrm{p'}^{(\mathrm{n})}$ is the sinusoidal encoding, and $\mathrm{q'}_{\mathrm{i}}^{(\mathrm{n})}$ is the i-th row element of $\mathrm{W}^{(\mathrm{n})}$. The same, matrices $\mathrm{k'}_{\mathrm{i}}^{(\mathrm{n})}$ and $\mathrm{v'}_{\mathrm{i}}^{(\mathrm{n})}$ have similar structures.

In FLOATER, the new positional coding p_i is added as:

$$\begin{aligned}&\mathrm{q}_{\mathrm{i}}^{(\mathrm{n})} = \mathrm{W}_{\mathrm{q}}^{(\mathrm{n})}(\mathrm{x}_{\mathrm{i}} + \mathrm{p}_{\mathrm{i}}) + \mathrm{b}_{\mathrm{q}}^{(\mathrm{n})} \\ &= \mathrm{W}_{\mathrm{q}}^{(\mathrm{n})}\left(\mathrm{x}_{\mathrm{i}} + \mathrm{p'}_{\mathrm{i}}^{(\mathrm{n})}\right) + \mathrm{b}_{\mathrm{q}}^{(\mathrm{n})} + \mathrm{W}_{\mathrm{q}}^{(\mathrm{n})}\left(\mathrm{p}_{\mathrm{i}} - \mathrm{p'}_{\mathrm{i}}^{(\mathrm{n})}\right) \\ &= \mathrm{q'}_{\mathrm{i}}^{(\mathrm{n})} + \mathrm{b}_{\mathrm{q,i}}^{(\mathrm{n})}\end{aligned} \tag{8}$$

It can be seen that the change of the encoded embedding position is equivalent to adding the position-aware bias vector $\mathrm{b}_{\mathrm{q,i}}^{(\mathrm{n})}$ to each self-attention layer, so the dynamic system can finally be described as:

$$\begin{aligned}&\mathrm{b}_{\mathrm{q}}^{(\mathrm{n})}(\mathrm{t}) = \\ &\mathrm{b}_{\mathrm{q}}^{(\mathrm{n})}(0) + \textstyle\int_{0}^{\mathrm{t}} \mathrm{h}^{(\mathrm{n})}\left(\tau, \mathrm{b}_{\mathrm{q}}^{(\mathrm{n})}(\tau); \theta_{\mathrm{h}}^{(\mathrm{n})}\right) \mathrm{d}\tau\end{aligned} \tag{9}$$

4.3 Model Description

Transformer model is widely popular in natural language processing or computer vision field, and it also has good effect in time series anomaly detection problem [5]. Like other models of encoding-decoding structure, the input sequence of transformer will undergo transformation, which is based on attention mechanism. The improved transformer [17] anomaly detection model based on FLOATER positional coding proposed in this paper includes two encoders and two decoders. The inputs of the two encoders are the focus score, the local context window W and the sequence C with timestamps respectively. Converting the multivariate sequence window W and sequence C into matrices, the scaled self-attention is defined as:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{m}}\right)V \tag{10}$$

Among them, Q, K, V are matrices, m is the size of W and C matrices, and softmax generates convex combination weights for matrix V, which is convenient to compress and express to simplify the operation of downstream neural networks. Furthermore, multi-head attention is introduced, which enables the model to pay attention to the representation information in different subspaces at the same time. Multi-head attention is defined as:

$$\begin{aligned}&\mathrm{MultiHeadAtt}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{H}_1, \ldots, \mathrm{H}_h), \\ &\mathrm{where}\mathrm{H}_i = \mathrm{Attention}(\mathrm{Q}_i, \mathrm{K}_i, \mathrm{V}_i)\end{aligned} \tag{11}$$

The input I_1 of the encoder Enc1 consists of the positional coding, the focus score (initially a zero matrix) and the sequence C, which undergoes the following operations:

$$I_1^1 = LayerNorm\left(I_1 + \text{MultiHeadAtt}\left(I_1, I_1, I_1\right)\right) \quad (12)$$

$$I_1^2 = LayerNorm\left(I_1^1 + \text{FeedForward}\left(I_1^1\right)\right) \quad (13)$$

Here MultiHeadAtt () operates on the input I_1 to generate attention weights that capture temporal trends in the input sequence. The operation of the transformer at each time point does not depend on the output of the previous time point, which enables the model to perform calculations on the input sequence in parallel, which significantly improves the training efficiency.

The input of the encoder Enc2 is the window sequence W generated by pre-processing. The masked multi-head attention is used instead to mask the subsequent sequence data. The operation process is as follows:

$$I_2^1 = Mask\left(\text{MultiHeadAtt}\left(I_2, I_2, I_2\right)\right) \quad (14)$$

$$I_2^2 = LayerNorm\left(I_2 + I_2^1\right) \quad (15)$$

$$I_2^3 = LayerNorm\left(I_2^2 + \text{MultiHeadAtt}\left(I_1^2, I_1^2, I_2^2\right)\right) \quad (16)$$

The encoding I_2^1 of the complete sequence is input into the encoder Enc2 as the key and value, and the attention operation is performed together with the input window sequence matrix. The multi-head attention here is similar to that in Enc1, except that a masked attention operation is used to prevent the model from reading subsequent data in advance when sequences are input at the same time. Compared with the limited context relationship in the prior art, this mechanism enables the model to encapsulate and utilize more context information.

Finally, the structures of the decoders Dec1 and Dec2 are the same, and the outputs O_1 and O_2 are generated respectively. The calculation process is as follows:

$$O_i = Sigmoid\left(\text{FeedForward}\left(I_2^3\right)\right) \quad (17)$$

Where $i = \{1, 2\}$ refers to the two decoders respectively. The sigmoid activation function is used to generate outputs in the range $[0, 1]$ to match the normalized input window.

4.4 Model Training

GAN models perform well in detecting input anomalies, so we employ an efficient adversarial training method. We are able to reconstruct each input through the encoding-decoding process of the Transformer model. However, traditional encoder-decoder structures often fail to capture short-term temporal trends, and if the distinction between normal and abnormal is too small, the detection effect will be greatly reduced. Based on the idea of GAN, the prediction of the reconstruction window by our model is achieved in two stages.

In the first stage, the focus score is initialized to a zero matrix, and the model is used to generate an approximate reconstruction of the input window, which helps the attention network in the transformer model encoder to extract the temporal trend of the input sequence. During training, higher weights are learned for regions with large reconstruction errors. At this stage, the encoder converts the input window sequence W and focus scores into compressed latent representations via contextual attention. The above process is similar to that of the standard model, and is then transformed into reconstructed outputs O_1 and O_2 through operations.

In the second stage, the focus score is updated to the reconstruction error obtained in the first stage. After running the model again, we can get the output of the decoder Dec2, $O_2^{'}$. The focus score in the previous stage represents the error between the reconstruction and the given input sequence, and the second stage takes it as prior knowledge to modify the attention weights. Higher activation operations on specific input sub sequences enable the neural network to extract short-term temporal trends.

The benefits of the two-stage training approach we use are three-fold. First, this amplifies the disparity used for discrimination, and the reconstruction error acts as an activation in the encoder's attention block, making the anomalies easier to distinguish. Second, the encoder Enc2 in the model captures short-term temporal trends, reducing the false positive rate. Finally, adversarial training improves the overall generalization ability, further making the model robust to different input sequences [2,8].

Like other adversarial training frameworks, our model also faces training stability issues. To address this problem, we design an adversarial training procedure that uses two independent and identically structured decoders to separately complete the reconstruction task of the input sequence. In the above process, using the reconstructed outputs of the two decoders, the reconstruction errors of the decoders Dec1 and Dec2 are defined by the L2 norm as:

$$L_1 = \|O_1 - W\|_2 \quad (18)$$

$$L_2 = \|O_2 - W\|_2 \quad (19)$$

The goal of the subsequent loss function is to make the decoder Dec2 distinguish the input window from the reconstructed sequence generated by the decoder Dec1 by maximizing the difference, $\|\mathrm{O}_2^{'} - \mathrm{W}\|_2$, while the decoder Dec1 is trained to further optimize its reconstruction sequence. The result, that is, the training target can be expressed as:

$$\mathrm{min}_{\mathrm{Dec1}}\mathrm{max}_{\mathrm{Dec2}}\|\mathrm{O}_2^{'} - \mathrm{W}\|_2 \quad (20)$$

It can be seen that the goal of the decoder Dec1 is to minimize the reconstruction error of the output, and the decoder Dec2 is to maximize it, which is expressed by the loss function as:

$$L_1 = +\|\mathrm{O}_2^{'} - W\|_2 \quad (21)$$

$$L_2 = -\|O_2^{'} - W\|_2 \tag{22}$$

From this, we have a two-stage loss function. Combining the two of them, the loss function is defined to accumulate the reconstruction and adversarial losses in the training phase as:

$$L_1 = \varepsilon^{-n}\|O_1 - W\|_2 + (1-\varepsilon^{-n})\|O_2^{'} - W\|_2 \tag{23}$$

$$L_2 = \varepsilon^{-n}\|O_2 - W\|_2 + (1-\varepsilon^{-n})\|O_2^{'} - W\|_2 \tag{24}$$

Where n is the number of training epochs and ε is a parameter close to 1. The weight of the reconstruction error is defined high at initialization to facilitate stable training when the decoder reconstruction performance is not good enough. Poor reconstruction will affect the focus score of the second stage, making it less reliable. Therefore, in the initial stage of the process, the weight of the adversarial loss is low to avoid destabilizing the model training. As the reconstruction gets closer to the input window sequence, the focus score becomes more accurate and the weight of the adversarial loss increases accordingly. Since the training process does not assume that the data is available sequentially, the full time series is divided into (W, C) pairs, and the model is trained using batch-paired inputs. Masked multi-head attention allows us to run in multiple batches in parallel and speed up the training process.

5 Experiments and Results

In order to verify the effectiveness of the anomaly detection model in the system in this paper, this paper simulates APT attacks for anomaly detection experiments based on the public datasets SMD [15] and SWaT [11], and makes the proposed anomaly detection method based on the FLOATER encoder improved transformer. Comparative experiments with deep autoencoder Gaussian mixture model, DAGMM, OmniAnomaly model,and the CAE-M model [22] are carried out to verify the effect of the proposed method on anomaly detection in time series.

5.1 Experimental Setup

The method in this paper is written based on the PyTorch framework and uses the AdamW optimization method. The initial learning rate is 0.01, the number of encoder layers is 1, the number of feedforward unit layers is 2, the number of hidden units is 64, the dropout rate is 0.1, and the batch size is 128. During training, the data set is divided according to the same ratio for both our method in this paper and the comparison method. In order to prevent the over fitting of the model, we use an early-stop strategy for training, that is, once the training accuracy starts to decline, the training process is stopped immediately. This paper tests the model using two public datasets, both including anomalous (attacked) data and normal data, to simulate the detection of APT attacks:

Server Machine Dataset (SMD) : This is a five-week-long dataset collected at a large internet company. The data comes from 28 different servers, respectively recording the CPU load, network usage, memory usage and other information of the machine, including attacked data and normal data. The abnormal situation is given and marked by experts according to the incident report.

Secure Water Treatment Dataset (SWaT): This is a dataset with a time length of 11 days, collected in the operating water treatment plants, of which the normal operation time is 7 days and the rest is abnormal operation time. The attack scenario for the SCADA system is set as a mixed attack method of three types of attackers, including both network and physical attacks. This dataset records sensor data and operational operations in operation, such as the height of water levels, pipeline flow rates, and system valve conditions.

5.2 Experimental Results

5.2.1 Anomaly Detection Results

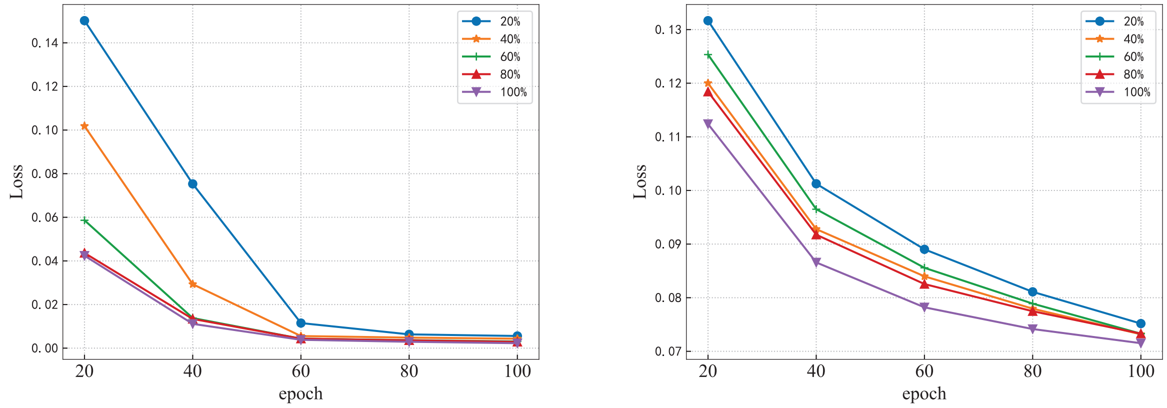
In order to verify the detection effect of the model in this paper under different datasets and data of different scales, we designed detection experiments. The following figures reflect the detection performance of the model in this paper on the SMD and SWaT datasets. In order to compare the influence of the data scale on the model detection effect, we adjusted the size of the training set and conducted experiments on the model. The results show that the model can also achieve better results in small-scale data sets. The specific method is as follows: We split the experimental training set and input it into the model, and the split ratios are 20%, 40%, 60%, 80% and 100% respectively. It can be seen that even in the case of using a small data set the model in this paper can achieve relatively ideal results (Figure 3).

5.2.2 Comparative Experiment

In order to further verify the effectiveness of the anomaly detection method of the improved transformer based on FLOATER positional coding and its advantages compared with other methods, we designed a comparative experiment and applied the model in this paper to the data set after data division. Based on the same settings, comparisons are made with the DAGMM model, the OmniAnomaly model and the CAE-M model.

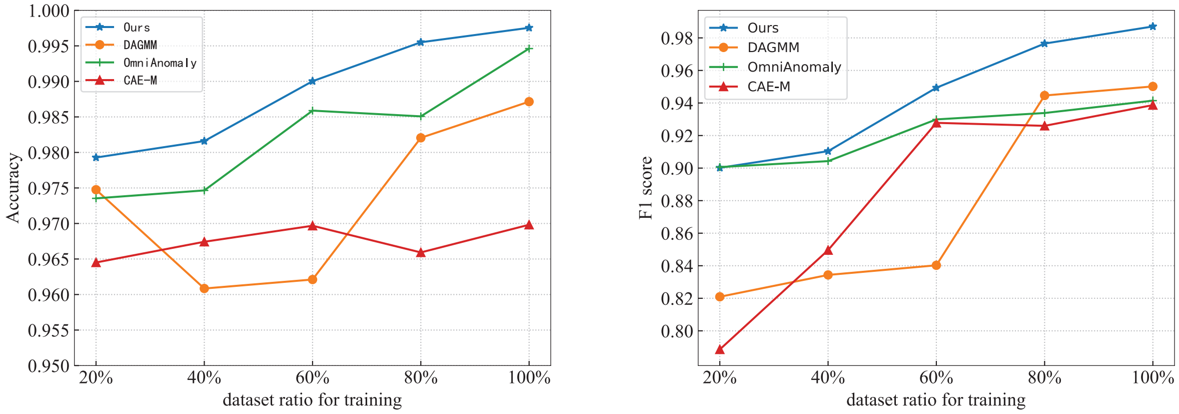
Table 1 shows the detection results of the model in this paper and the comparison model, where P represents precision, R represents recall, F1 represents F1 score, and AUC represents area under the receiver operating characteristic curve. It can be seen that when the input is a complete dataset, the anomaly detection effect of the model in this paper is generally better than the comparison model, a higher F1 score and AUC value are obtained.

Divide the training set into the same proportion and input it into the model and train it, and the obtained detection results are shown in the figure above. The deep



(a) SMD dataset (b) SWaT dataset

Figure 3: Training loss



(a) Accuracy (b) F1 score

Figure 4: Result with dataset ratio

Table 1: Performance comparison on the complete dataset

Method	**SMD**				**SWaT**			
	P	R	F1	AUC	P	R	F1	AUC
DAGMM	0.9104	0.9914	0.9491	0.9954	0.9933	0.6879	0.8128	0.8436
OmniAnomaly	0.8881	0.9985	0.9401	0.9946	0.9782	0.6957	0.8131	0.8467
CAE-M	0.9010	0.9783	0.9381	0.9723	0.9697	0.6957	0.8101	0.8464
Ours	0.9272	0.9974	0.9610	0.9975	0.9782	0.6997	0.8158	0.8491

learning method has strict requirements on the scale of the data set. The larger the scale of the data set obtained by the model and the more uniform the proportion of types, the better the learning effect will be. Since the data is not easy to obtain, the training effect of the model on the smaller dataset is very important. It can be seen that when the training set is input into the model at the proportion of 20%, 40%, 60%, 80% and 100%, the detection effect of the model in this paper is better than that of the comparison model, and the stability is higher (Figure 4).

6 Conclusion

In order to solve the increasingly serious information network security problem, a deep security detection framework based on ATT&CK is proposed, and an anomaly detection method of improved transformer based on FLOATER encoder is adopted. Security detection framework includes three modules: information collection, analysis engine and security response. By realizing bottom-up security detection and defense, the security of the system is improved. Among them, the anomaly detection method integrates the FLOATER positional coding into the training of transformer model, which can better learn the temporal characteristics of time series, strengthen the learning ability of the model through antagonistic training. Compared with the classical deep learning method, it has better detection effect.

Acknowledgments

This work was supported in part by the State Grid Jiangxi Information & Telecommunication Company Project "Research on de-boundary security protection technology based on zero trust framework" under Grant 52183520007V.

References

[1] O. Alexander, M. Belisle, and J. Steele, "Mitre att&ck for industrial control systems: Design and philosophy," *The MITRE Corporation: Bedford, MA, USA*, 2020.

[2] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3395–3404, 2020.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[4] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 387–395, 2018.

[5] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks," in *International conference on artificial neural networks*, Springer, pp. 703–716, 2019.

[6] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical intermetric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3220–3230. , 2021

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, IEEE, pp. 413–422, 2008.

[8] S. Liu, T. Wang, D. Bau, J.-Y. Zhu, and A. Torralba, "Diverse image generation via self-conditioned gans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14 286–14 295, 2020.

[9] X. Liu, H.-F. Yu, I. Dhillon, and C.-J. Hsieh, "Learning to encode position for transformer with continuous dynamical model," in *International conference on machine learning*, PMLR, pp. 6327–6335, 2020.

[10] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967.

[11] A. P. Mathur and N. O. Tippenhauer, "Swat: A water treatment testbed for research and training on ics security," in *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*, IEEE, pp. 31–36, 2016.

[12] S. Naseer and Y. Saleem, "Enhanced network intrusion detection using deep convolutional neural networks," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 10, pp. 5159–5178, 2018.

[13] B. N. Saha, N. Ray, and H. Zhang, "Snake validation: A pca-based outlier detection method," *IEEE signal processing letters*, vol. 16, no. 6, pp. 549–552, 2009.

[14] L. Shen, Z. Li, and J. Kwok, "Timeseries anomaly detection using temporal hierarchical one-class network," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 016–13 026, 2020.

[15] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

[16] J. Tian, H. Gu, C. Gao, and J. Lian, "Local density one-class support vector machines for anomaly detection," *Nonlinear Dynamics*, vol. 64, no. 1, pp. 127–130, 2011.

[17] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] C. Xiong, T. Zhu, W. Dong, L. Ruan, R. Yang, Y. Chen, Y. Cheng, S. Cheng, and X. Chen, "Conan: A practical real-time apt detection system with high accuracy and efficiency," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[20] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642*, 2021.

[21] B. Yan and G. Han, "La-gru: Building combined intrusion detection model based on imbalanced learning and gated recurrent unit neural network," *security and communication networks*, vol. 2018, 2018.

[22] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[23] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.

Biography

Ri-xuan Qiu is in State Grid Jiangxi Information & Telecommunication Company, Nanchang, China. Email: qiurixuanwork@163.com.

Jian Le was born in Huanggang, Hubei, China in 1975. He received his Ph.D. degree in electrical engineering from Tsinghua University (THU), Beijing, China in 2006. He is currently Associate Professor with the college of electrical engineering and automation at Wuhan University (WHU), where he has been working on smart grid operation and power quality control technology.

Fu-yong Zheng is in State Grid Jiangxi Information & Telecommunication Company, Nanchang, China. Email: 414833044@qq.com.

Gan Qi was born in Hengyang, Hunan, China in 1999. He received his B.S. degree from the School of Electrical Engineering and automation at Wuhan University (WHU), Wuhan, China, in 2021. He is now working towards a Master degree in electrical engineering at Wuhan University. He has been working on distribution generation control technology and power electronics.

Chao Peng is in State Grid Jiangxi Information & Telecommunication Company, Nanchang, China. Email: pengchao1988421@163.com.

Yi Zhang received the B.S. degree in information security from North China Electric Power University, Beijing, China, in 2019. She is currently pursuing the M.S. degree in School of control and Computer Engineering of North China Electric Power University. Her current research interests include information security and electricity trading.

Yuan-cheng Li was a postdoctoral research fellow in the Digital Media Lab, Beihang University. He has been with the North China Electric Power University, where he is a professor and the Dean of the Institute of Smart Grid and Information Security. He was a postdoctoral research fellow in the Cyber Security Lab, college of information science and technology of Pennsylvania State University.

Yan Liu received the B.S. degree in software engineering from Ludong University, Yantai, Shandong, in 2020. She is currently pursuing the M.S. degree in School of control and Computer Engineering of North China Electric Power University. Her current research interests include information security, blockchain, and electricity trading. Email: LySeeyouer@163.com

Emergent Cybersecurity Information Discovery in Support of Cyber Threat Prevention

Chia-Mei Chen, Jin-Jie Fang, Zheng-Xun Cai, Boyi Lee, and Dan-Wei Wen

(Corresponding author: Chia-Mei Chen)

Department of Information Management,
National Sun Yat-sen University, Kaohsiung, Taiwan
Email: cchen@mail.nsysu.edu.tw

(Received Aug. 2, 2022; Revised and Accepted Dec. 3, 2022; First Online Dec. 14, 2022)

Abstract

Businesses suffer from cyberattacks and demand a proactive defense strategy to adapt to the rapidly evolving landscape of cyberattacks. In order to acquire emergent attack trends, this research collects online cybersecurity news as the data source. It proposes an unsupervised learning approach that automatically extracts new cybersecurity information to find emergent cyber threat intelligence. The proposed discovery solution consists of two stages of clustering: the first stage identifies new topics, and the second phase further clusters the news articles of the new topics into groups so that those with similar contents are gathered in a group. The experimental results demonstrate that the proposed method can identify emergent threat intelligence effectively.

Keywords: Event Detection; Threat Intelligence; Topic Model

1 Introduction

Cyberattacks are becoming aggressive and widespread in the latest decades. Therefore, businesses should take proactive defense by receiving the latest cyber threat intelligence (CTI) so that the corresponding defense mechanism can be deployed in advance. Threat data can come from many sources, and external data sources such as cybersecurity news and reports contribute over 77% [6]. Gathering and analyzing such unstructured textual data is time-consuming. Therefore, this study proposes an approach of automatically collecting cybersecurity articles and retrieving emergent CTI.

Several research gaps were identified through the literature review. First, current CTI efforts rely on the use of auto-feeds from the websites to generate CTI, and most focus on extracting IoC (Indicator of Compromise) information with specific string patterns, which means that the current security measures are often handled reactively based on existing attack cases. Second, articles in the cybersecurity domain are different from common news as they contain security-related terms. Therefore, it requires an effective feature selection scheme to extract representative cybersecurity terms. On the other hand, discovering emergent CTI is different from hot topic discovery, as it might contain new terms or create a new topic different from the historical one. Extracting such new terms is thus a vital step in deciding what drives conversations, and doing this is critical in automatically retrieving new CTI topics. Finally, previous work mostly focused on identifying security information by classification with patterns and rarely explored unstructured news contents to discover emergent CTI by clustering. With these research gaps, the following research questions have been proposed to guide the study: (1) How to detect new CTI topics effectively from cybersecurity news articles from multiple sources? (2) How to discover new articles effectively?

The primary contribution of this study is proposing an unsupervised learning solution to extract new cybersecurity information effectively, which automatically collects and explores time-series news articles from multiple sources and adopts a two-stage discovery process.

2 Literature Review

Alves *et al.* [3] processed tweets to generate IoCs, where TF-IDF is applied to extract keywords and filtering and regular expression to retrieve IoCs. Li *et al.* [11] also applied filtering and regular expression to retrieve CTI terms and CVE. Behzadan *et al.* [4] developed two CNN models: a binary classifier to detect cyber-related tweets and a multi-classifier to categorize them into multiple types of cyber threats.

Abu *et al.* [1] applied an association rule mining method on network traffic to learn the relationship among the IoCs, mainly IP addresses. Najafi et al. [16] built up a knowledge graph that models the relationship among entities observed in DNS and proxy logs. Deliu *et al.* [7] adopted supervised and unsupervised machine learning techniques to analyze hacker forums and retrieve meaningful CTI. Huang and Chen [10] utilized hacker forums to identify the social networks of hackers.

3 The Proposed Method

Past studies proved that a two-stage clustering or a hybrid method yields promising results in topic detection and tracking. Therefore, this study proposed a top-down two-stage clustering approach. Inspired by past studies [2, 13], this study designs a top-down two-stage clustering algorithm that discovers emergent CTI. It leverages a topic model (LDA) to discover topics in a document set and then applies a centroid-based clustering algorithm (K-mean++) to detect emergent CTI events from a new topic cluster.

Figure 1 outlines the proposed solution consisting of the following steps: (1) collects news by a web crawler; (2) performs text preprocess; (3) selects features to represent articles; (4) clusters the historical and current news into topic clusters, respectively; (5) compares the similarity of a current topic cluster with the historical ones to identify new topic clusters; (6) clusters the new topic clusters into groups and extracts new CTI articles.

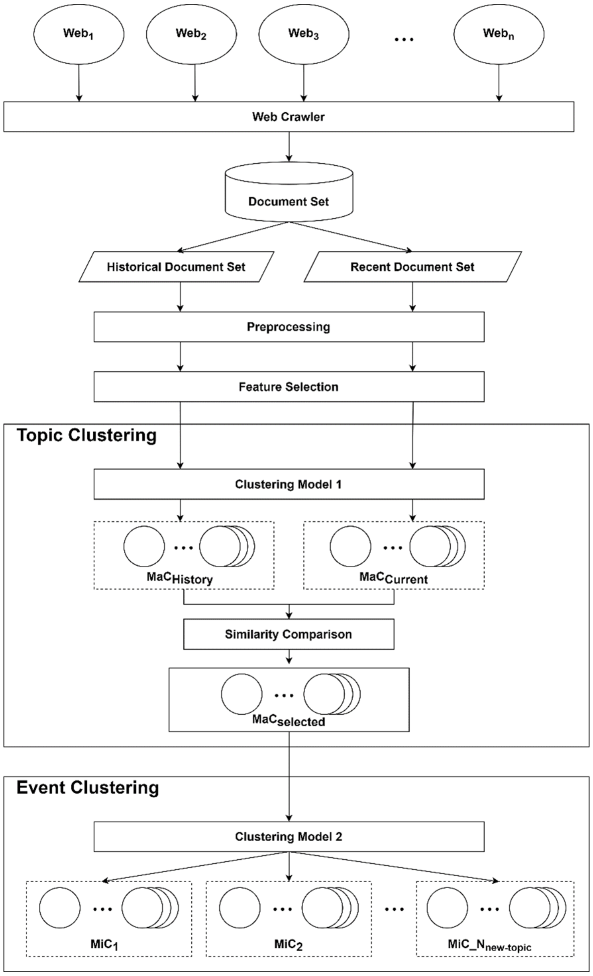


Figure 1: The proposed system architecture.

This study applies common preprocessing steps including stop word removal and tokenization. It is important to include emergent terms in the feature set, such as new malware or new hacker groups. Our preliminary study observed that such terms are capitalized. Therefore, this study includes all capital tokens, CVE, and IP addresses in the feature set.

A news article might be categorized into multiple topic clusters and each topic weights differently in the article, as topic modeling considers a document consisting of multiple topics. Let k be the number of topics covered by the document and θ_i be the weight of topic i in the document, and $\sum_{i=1}^{k} \theta_i = 1$. For example, an article focuses on reporting a phishing attack, but it mentions how to protect personal privacy at the end. In this article, "phishing" contributes 0.9 while "data leakage" 0.1. In the second clustering stage, each document is categorized into one single topic. To prevent the same news article appears multiple times in the second clustering, this study applies a modified LDA algorithm at the first clustering stage that categorizes each article into its most weighted topic cluster. Therefore, it categorizes the article in the "phishing" cluster as most terms are about phishing.

As for measuring cluster similarity, several past studies [8] [5] [9] applied JS divergence in topic modeling to quantify the difference between two topics and achieved effective results. Our preliminary study also concludes a similar outcome. Therefore, this study adopts Jensen-Shannon (JS) Divergence as the distance measurement for calculating the similarity of two topic clusters or two news articles.

4 System Evaluation

This study conducted the following experiments to address the proposed research questions. Exp 1 investigates the effectiveness of the topic clustering stage. Exp 2 investigates if the K-means++ can effectively divide documents with similar contents into the same groups.

4.1 Exp 1: The Effectiveness of Detecting New Topic Clusters

Exp 1 adopts topic coherence, UMass-coherence [12], to quantify the clustering quality and to determine the best number of topic clusters at the LDA clustering stage. After investigating different combinations of historical time frames and current time frames, $(T_{history}, T_{current})$, Table 1 lists the min JS divergences of different time frame combinations, and the results indicate that the best combination of (Thistory, Tcurrent)=(20,4). A large distance discrepancy implies it can efficiently distinguish historical and current topic clusters and can effectively identify new topic clusters.

Exp 1 also investigates the LDA parameter setting of α, β, and the number of iterations, L, where α represents document-topic density, β represents topic-

Table 1: The minimum JS divergences over $(T_{history}, T_{current})$

$T_{current}$ $T_{history}$	10	20	30
3	0.087055316	0.123762220	0.144177577
4	0.089493055	0.151379363	0.146341061

word density, and D specifies the maximum number of iterations allowed for convergence. After experimenting with different combinations of the parameters, Table 2 summarizes the determined parameter setting of the topic clustering stage.

4.2 Exp 2: The Effectiveness of Discovering New CTI Articles

Exp 2 investigates the effectiveness of the event clustering results to see if the applied K-means++ can identify new CTI effectively as well as aggregate similar contents into the same group. To determine the number of new events, $(N_{new-events})$, the evaluation adopts the two common metrics: Davies-Bouldin Index and Calinski-Harabasz Index, and investigates two clustering models: K-means++ and hierarchical cluster.

Figure 2 shows the K-means++ clustering results on different numbers of new event clusters. The two performance metrics collide between $N_{new-events} = 3$ and 4. After human verification, the number of new events in a new topic $(N_{new-events})$ is set to 3 as it achieves better between- and within-cluster dispersions.

Table 3 illustrates the clustering results of K-means++ on one emergent news topic dated between 2019/09/01 and 2019/09/04. By human validation, K-means++ can effectively group similar news articles from multiple data sources into a group. The results also demonstrate that the proposed solution is able to include new terms, such as Poshmark, into the keyword list.

5 Conclusion

This research collects and analyzes cybersecurity news from multiple sources and proposes a solution for emergent CTI discovery. The proposed unsupervised learning-based method applies two clustering stages, where the first clustering stage identifies emergent topics and the second clustering stage detects new CTI. The evaluation results prove that the proposed solution can detect emergent CTI effectively and demonstrate its applicability as well.

One future research direction can extend the data sources to acquire more emergent CTI. Another possible direction is to customize CTI retrieval based on the assets and systems in an organization so that the organization could obtain specific threats on their systems or assets.

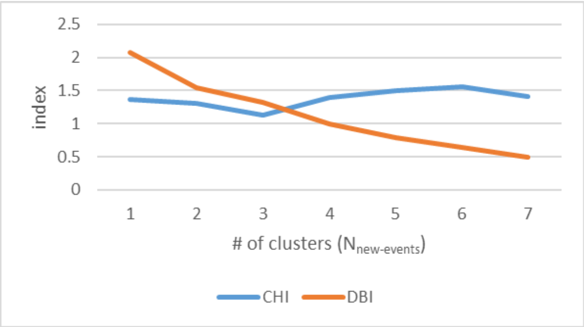


Figure 2: The evaluation of K-means++ clustering over different numbers of clusters.

Table 2: The parameter setting of the topic clustering stage.

	History	Current
Time frame $(T_{history}, T_{current})$	20	4
Average No. of articles $(\lvert H\rvert, \lvert C\rvert)$	450∼500	90∼105
No. of topic clusters $(NC_{history}, NC_{current})$	13	10
a	0.04	0.02
b	0.02	0.02
No. of iterations (L)	500	

References

[1] M. S. Abu, S. R. Selamat, R. Yusof, and A. Ariffin, "An attribution of cyberattack using association rule mining (ARM)," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.

[2] W. Ai, K. Li, and K. Li, "An effective hot topic detection method for microblog on spark," *Applied Soft Computing*, vol. 70, pp. 1010–1023, Sep. 2018.

[3] F. Alves, A. Bettini, P. Ferreira, and A. Bessani, "Processing tweets for cybersecurity threat awareness," *Information Systems*, vol. 95, p. 101586, 2020.

[4] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, *Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream*, 2018.

[5] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Applied Intelligence*, vol. 50, no. 1, pp. 138–156, 2020.

[6] R. Brown and R. M. Lee, "2021 SANS cyber threat intelligence (CTI) survey," https://www.cybersixgill.com/wp-content/uploads/2021/02/SANS_CTI_Survey_2021_Sixgill.pdf, 2021, accessed: 2022-06-04.

[7] I. Deliu, C. Leichter, K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," in *IEEE International Conference on Big Data*, 2017.

[8] Y. Fang, L. Si, N. Somasundaram, Z. Yu, "Mining contrastive opinions on political texts using cross-perspective topic model," in *Proceedings of the Fifth*

Table 3: An illustration of the K-means++ detection results.

Topic ID	Keywords in an event	News Title
New Topic ID 2	data, breach, customers, huge, expert, flights	Flight booking platform Option Way exposes customer and internal data
	Poshmark, cracked, password, data, details, bcrypt, login, million	Cracked Passwords for Poshmark Accounts Being Sold Online One million cracked Poshmark accounts being sold online
	XKCD, data, breach, forums, phpbb, password, offline, addresses	XKCD Forum Hacked Over 562,000 Users Account Details Leaked XKCD Forum Breach Exposes Emails, Passwords of 562,000 Users Hackers steal 560,000 user accounts in XKCD forum breach

ACM International Conference on Web Search and Data Mining, 2012.

[9] D. Hall, D. Jurafsky, C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[10] S. Y. Huang, H. Chen, "Exploring the online underground marketplaces through topic-based social network and clustering," in *IEEE Conference on Intelligence and Security Informatics*, 2016.

[11] K. Li, H. Wen, H. Li, H. Zhu, L. Sun, "Security OSIF: Toward automatic discovery and analysis of event based cyber threat intelligence," in *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, 2018.

[12] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[13] G. Petkos, S. Papadopoulos, Y. Kompatsiaris, "Two-level message clustering for topic detection in Twitter," in *SNOW-DC WWW*, 2014.

Biography

Chia-Mei Chen has joined in the Department of Information Management, National Sun Yat-sen University since 1996. She was the Section Chef of Network Division and Deputy Director, Office of Library and Information Services in 2009-2011. She had served as a coordinator of TWCERT/CC (Taiwan Computer Emergency Response Team/Coordination Center) during 1998 to 2013 and then as a consultant until 2018. Based on her CSIRT experience, she established TACERT (Taiwan Academic Network Computer Emergency Response Team) in 2009. She was a Deputy Chair of TWISC@NCKU, a branch of Taiwan Information Security Center during 2017 to 2020. She continues working for the network security society. Her current research interests include anomaly detection, network security, machine learning, text mining, and big data analysis.

Jin-Jie Fang has received his master's degree from the Department of Information Management, National Sun Yat-sen University. Currently he is a software engineer in Advanced Semiconductor Engineering, Inc.

Zheng-Xun Cai received his master's degree from the National Sun Yat-sen University in 2017 and continues pursuing the PhD degree at the same school. His research focuses on digital forensics, network analysis, and intrusion detection.

Boyi Lee received his PhD degree from the National Sun Yat-sen University. Currently he is a software engineer in an information security institute.

Dan-Wei Wen is an assistant professor at the Department of Business Administration, National Kaohsiung University of Science and Technology. She received her Ph.D. from the Department of Business Administration, National Cheng-Kung University. Her research interests include industry dynamics, catching-up strategy, and data mining.

An Efficient USM Weak Sharpening Detection Method for Small Size Image Forensics

Jie Zhao, Shuang Song, and Bin Wu

(Corresponding author: Bin Wu)

School of Computer and Information Engineering, Tianjin Chengjian University
Tianjin 300384, China
Email: wubin@tcu.edu.cn

(Received Aug. 7, 2022; Revised and Accepted Dec. 10, 2022; First Online Dec. 15, 2022)

Abstract

Unsharp masking (USM) sharpening is a typical image processing method to improve image quality. Therefore, revealing USM sharpening traces is essential to research in the field of forensics. In recent years, USM detection has attracted significant interest from image forensics professionals. As a result, several successful approaches have been developed for detecting USM sharpening in large-size images and detecting strong USM sharpening in small-size images. However, USM weak sharpening detection for small-size images is still tricky. To address this challenge, this paper proposes an efficient method called ViT-Small, a variant of the Vision Transformer, to detect the weak USM sharpening of small-scale images. Several experiments have been conducted on several public datasets to demonstrate that the proposed method outperforms the existing state-of-the-art USM detection methods.

Keywords: Image Forensics; Small-size Image Forensic; Vision Transformer Variants; Weak Sharpening Detection

1 Introduction

Images play a significant role in people's daily lives in the high information age and have an impact on all facets of those lives. Due to the emergence of advanced artificial intelligence (AI) technology and automatic image processing technology, as well as the simplicity that a large number of forged photos can be produced and transmitted online, people's confidence in the authenticity of images has been greatly reduced. At the same time, the challenge and current research hotpot is how to preserve the true and eliminate the false in various and complex images. As a result, more and more researchers are studying image forensics techniques [5, 13], which are used to determine whether a given image is an original image or a manipulated image. Since authenticity is a crucial component of both news media and judicial evidence, using advanced image forensics technology to determine the authenticity of images can increase the reliability of news reports and improve the effectiveness of forensic identification.

USM is an image sharpening method that increases the visual clarity of an image by enhancing the details of the image. It is also a common sharpening method in current image processing software. As described in [12], the operation of USM sharpening can also be used to conceal traces of illegal image forgery, so USM detection is an important and necessary research.

In 2009, Cao *et al.* [2] proposed an image sharpening detection for the first time, which was achieved by analyzing histogram aberration and measuring the strength of overshoot artifact to detect image sharpening. However, they later studied and found that the method of [2] can only detect images with wide histograms. To overcome this shortcoming, they proposed a more effective method [3]. This algorithm distinguishes the authenticity of the image by comparing a preset appropriate threshold and the overshoot artifact of the image. Compared with [2], the accuracy has also been improved to a certain extent. However, this method cannot cope with JPEG compression [19, 26] and is also sensitive to changes in sharpening parameters. Then, in 2013, Ding *et al.* [8] introduced the local binary pattern (LBP) [4,24] into sharpening detection, and used LBP to extract the local texture features of the image to separate the true and false pictures. Compared with the algorithm of Cao *et al.*, the accuracy of the accuracy of Ding *et al.* is significantly improved is significantly improved. Different from the algorithm ideas mentioned above, Lu *et al.* [18] proposed a new method based on the idea of anti-forensics. This method achieves the purpose of anti-forensics by removing overshoot artifacts, and successfully made the USM sharpening detection proposed by Cao *et al.* in [2] and [3] invalid. Since the overshoot artifact can be considered as a special texture feature, the LBP method in [9] has great performance in detecting sharpening, but the contradiction between the LBP window size and the LBP feature dimension makes this method still limited. Inspired by the LBP method, a novel method called Edge Perpendicular Binary Coding (EPBC) [10] was proposed by Ding *et al.* in 2015 to detect USM sharpening. It sets a window perpendicular to the edge points and detects USM sharpening by encoding the pixels corresponding to the window. Experiments show that this is a more advanced algorithm than the above algorithm. However, for weakly sharpened images, the EPBC method fails. For weakly sharpened detection, Ding *et al.* proposed a new algorithm called Edge Perpendicular Ternary Coding (EPTC) [7], this method works better than EPBC. In 2018, Ye *et al.* [27] first introduced CNN into USM sharpening detection, which achieved a leap from traditional manual feature extraction to deep learning feature extraction for USM sharpening detection. In 2020, Wang *et al.* [22] proposed a novel algorithm called DCT-CHDMY for the challenge of USM sharpening detection in small-size images.

The current USM sharpening detection algorithms can detect the USM sharpening of large-sized (more than 384*384 pixels) images, and small-sized images (smaller than 64*64 pixels) sharpened with strong sharpening strength, but these algorithms fail to detect the USM weak sharpening of small-sized images. To detect weak sharpening in images of small size, this paper proposes a method based on a variant of the Vision Transformer (ViT) called ViT-Small. A series of experiments on several public datasets show that our algorithm is effective enough for weak sharpening detection of small-size images and outperforms CNN and DCT-CHDMY, the current best algorithms for USM sharpening detection. The following are the main contributions of this work:

1) Compared with previous sharpening detection methods, the USM weak sharpening detection for small size images was first focused on and studied.

2) A novel VIT variant was proposed to perform an efficient USM weak sharpening detection method for small size image forensics.

3) A series of post-processing operations were done on the public datasets to evaluate the robustness of the algorithm.

The rest of this paper is organized as follows. Section 2 introduces the principle of USM sharpening; Section 3 illustrates the structural details of ViT and ViT-Small; Experimental results and discussions are presented in Section 4; It is summarized in Section 5.

2 USM Sharpening

USM is a popular sharpening method today. Its principle is to increase the high-frequency part of the image to improve image quality and detail clarity. The total form can be expressed as Equation (1):

$$I' = I + \lambda M. \tag{1}$$

I' represents the sharpened image, I represents the original image, M represents the unsharp mask, and λ is a parameter that controls the strength of sharpening.

In USM algorithm, unsharp mask can be generated by two ways. The first way is to high-pass filter the original image, as expressed by

$$M = I \otimes H. \tag{2}$$

Where H denote convolution operator and a high-pass filter, and $\otimes$ denotes a convolution operation.

However, this approach fails to reduce noise during USM sharpening. To compensate for this shortcoming, a different approach has been proposed. Instead of high-pass filtering directly, first low-pass filtering is performed on the original image, after that the low-frequency components are subtracted from the original image. The low-pass filter generally uses Gaussian filtering, that is

$$M = I - I \otimes G\sigma. \tag{3}$$

Where $G\sigma$ represent a Gaussian low-pass filter with a standard deviation of σ respectively.

3 Proposed Method

3.1 Color Space Model Choice

The three most popular image color space models are RGB, YCbCr, and Lab; RGB is the most fundamental of these. Red, green, and blue (R, G, and B) are superimposed in different ratios to create rich colors in RGB, and each of the three channel components has texture and color information in addition to being highly correlated [16]. In YCbCr, Y represents brightness, and Cb and Cr represent blue and red respectively, the Y component contains the texture information of the image, and the Cb and Cr components represent the color information of the image. Lab consists of three elements, one is the luminance channel L, and the other two channels a, b are color channels. Similar with YCbCr, L retains the detailed information of the image, and the a and b components contain all the color information of the image. The three image color space models and each channel are shown in Figure 1, Figure 2 and Figure 3.

Since USM sharpening mainly causes changes in image texture, inspired by [22], a set of experiments was done with the RGB, YCbCr, Lab, Y component of YCbCr, and the L component of Lab as the input of the model. Taking CASIA64 as an example, when the sharping permeant $\sigma = 0.5$, $\lambda = 0.5$, the results are summarized in Table 1.

From Table 1, it is clear that the USM sharpening detection accuracy is significantly higher when the input image is in the YCbCr color space than when the input image is in other color spaces. As a result, the experiment's datasets are transformed into a YCbCr model before being fed into the ViT-Small detection model in this paper.

Table 1: Detection accuracy of different color model on CASIA64

Color Space Model	RGB	YCbCr	Lab	Y channel	L channel
Accuracy	86.32%	**89.50%**	84.62%	78.32%	79.98%

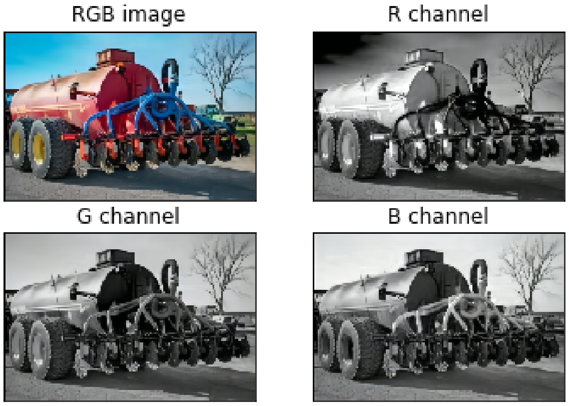


Figure 1: RGB image and R, G, B channel

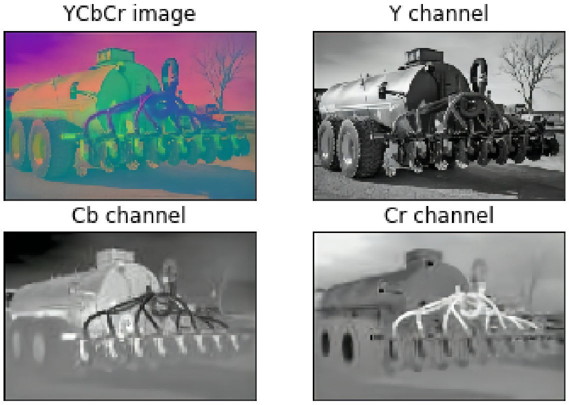


Figure 2: YCbCr image and Y, Cb, Cr channel

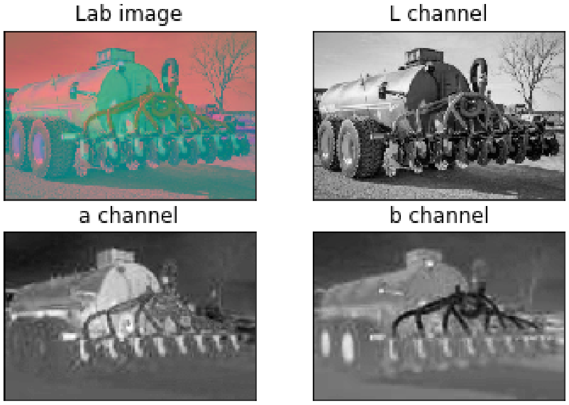


Figure 3: Lab image and L, a, b channel

3.2 Vision Transformer

Vision Transformer (ViT) [11] attempts to directly apply a standard Transformer to images and was inspired by a Transformer in the NLP field [6,23]. The encoder-decoder structure of the standard Transformer is maintained by the ViT, but multilayer perceptron (MLP) head is used in place of the complex decoder blocks. The entire image will first be split up into smaller image patches, and patch embedding will then be obtained through a linear mapping process, which is comparable to words and word embedding in NLP. The linear embedding sequence of these image patches will then be used as the input of the ViT. The training for image classification [14] will be done using the supervised learning approach. Finally, the MLP head layer is used to produce results for image classification. In this paper, a ViT variant structure called ViT-Small is introduced after image pre-processing, as shown in Figure 4.

3.2.1 Image Patch Embedding

A image of size $H \times W \times C$ (where H, W, C represent the height, width and number of channels of the image respectively) is divided into small patches, the size of each patch is $P \times P \times C$, and the number of patches is $(H \times W)/P^2$, denoted by N. This forms a sequence $\{x_1, x_2, x_3, \ldots, x_N\}$ of length N, where x_i $(i = 1, 2, \ldots, N)$ represents the i-th patch. After that, map each 2D patch to a 1D vector via a trainable linear projection, compressing the dimension to P^{2}C, which output is called as "patch embedding".

Before feeding the patches to the encoder, the image patches still need some processing, then an embedded matrix E is introduced. The sequence of image patches is multiplied by E and concats a learnable vector v_{class} prepared in advance for the classification task. If only this result sequence is fed into the transformer encoder, the encoder does not know the positional relationship between the sequences in the original image, so it is necessary to use a vector called Epos representing the position information and add the result sequence as the new encoder input z_0. This process is presented in Equation (4).

$$\begin{aligned} z_0 &= [v_{class}; x_1E; x_2E; \ldots; x_NE] \\ &\quad + E_{pos}, E \in R^{(P^2C) \times D}, Epos \in R^{(N+1) \times D} \end{aligned} \tag{4}$$

3.2.2 Vision Transformer Encoder

The transformer encoder consists of L identical blocks, and each block mainly consists of three parts, namely Layer Norm (LN), multihead self-attention (MSA) [17, 20, 21], and MLP. In each coding block, it always goes through the MSA layer first, followed by the MLP layer.

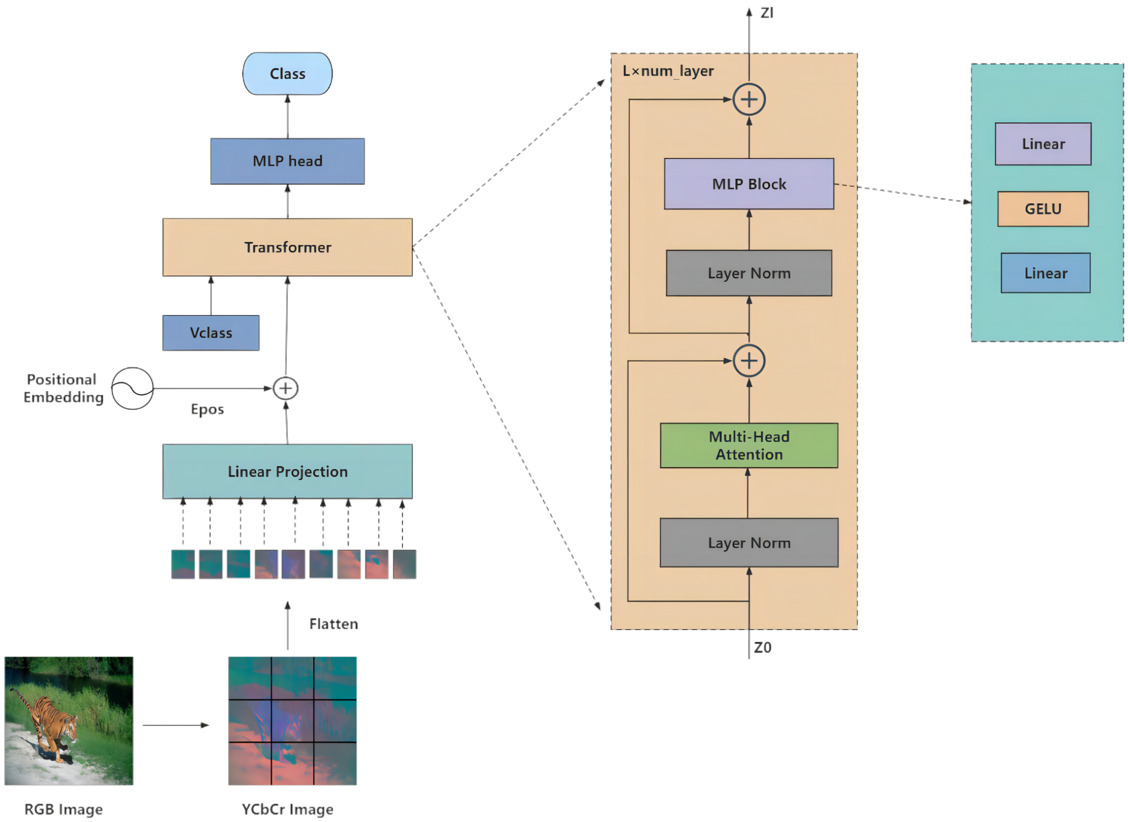


Figure 4: A Vision Transformer variant structure after image preprocessing is added

Both MSA and MLP layers are connected to an LN layer in front, and a residual is connected afterward, which can be described as Equation (5) and Equation (6). v_{class} in the output of the last coding block is represented by z_L^0, Finally, the z_L^0 is normalized and sent to the MLP head to obtain the final image prediction and classification result, as expressed in Equation (7).

$$z_l' = \text{MSA}\left(LN\left(z_{l-1}\right)\right) + z_{l-1}, 1 = 1 \ldots L \tag{5}$$

$$z_l = \text{MLP}\left(LN\left(z_l'\right)\right) + z_l', 1 = 1 \ldots L \tag{6}$$

$$y = LN\left(z_L^0\right) \tag{7}$$

The Layer Norm layer in the encoder, just like its name, is to normalize all neurons in an intermediate layer.

MLP block consists of an input layer, an output layer and at least one hidden layer. The neurons in each hidden layer in the network can receive the information transmitted from all neurons in the adjacent pre-sequential hidden layers, and after processing, the information is output to all the neurons in the adjacent subsequent hidden layers. In MLP, neurons in adjacent layers are usually connected in a "full connection" manner. MLP can simulate complex nonlinear functions, and the complexity of the simulated functions depends on the number of hidden layers in the network and the number of neurons in each layer.

For self-attention, its role is to encode each entity according to the global context information to capture the connections between all entities. The primary function of self-attention is to give the input sequence the proper weight. The weighted sum of all the values in the input sequence serves as a representation of its weight. Three learning matrices—W^Q, W^K, and W^V—are first defined. These three learning matrices are multiplied by the input sequence x to get the three values Q, K, and V. To determine the correlation between one element and other elements in the sequence, compute the dot product between the Q vector of that element and the K vector of the other elements. The results determine the relative importance of the patches in the sequence. Then the result of dot-product is fed into softmax. In order to prevent the gradient from becoming extremely small after softmax, the result of dot-product needs to be divided by $\sqrt{D_k}$, finally output of softmax is dot-multiplied with V to identify patches with high attention scores, witch result is self-attention B output as shown in Equation (8) and Equation (9). To sum up, the calculation process of self-attention is Equation (10) and self-attention detail map is shown in Figure 5.

$$\begin{cases} Q = w^Q X \\ K = w^K X \\ V = w^V X \end{cases} \tag{8}$$

$$\begin{cases} AK^T = Q/\sqrt{D_k} \\ A' = \text{softmax}(A) \\ B = A'V \end{cases} \tag{9}$$

$$\text{Attention}(\text{Q}, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \quad (10)$$

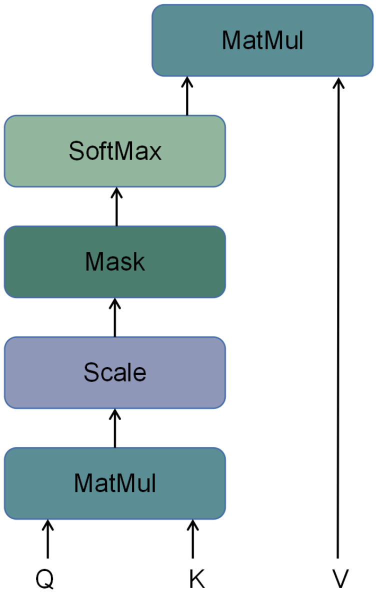


Figure 5: Self-attention detail map

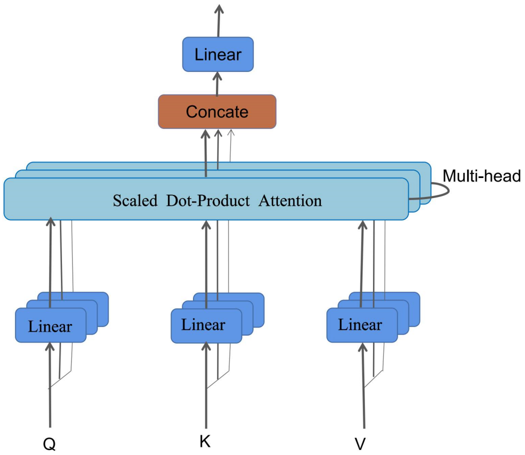


Figure 6: Multihead attention detail map

Multihead attention is an advanced version of self-attention which is displayed in Figure 6. Multihead attention allows the model to jointly attend to information from different representation subspaces at different positions [11]. Unlike self-attention, multiple K, Q, and V are used in multihead attention to obtain multiple similarities, which are then combined and multiplied by learnable parameter W^O matrix to provide the desired result, and the final result is obtained after fusion, which can be formalized as Equation (11).

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \quad (11)$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and $W^O \in R^{hdv \times D}$.

3.3 Vision Transformer Variants

Three different versions of the visual converter are proposed in [11]: ”ViT-Base”, ”ViT-Large” and ”ViT-Hugh”. The three versions differ in the number of layers of the encoder, the size of the hidden dimension, the number of attention heads used by the MSA layer, and the size of the MLP classifier. Each of these models is trained using patches of size 16 × 16 and 32 × 32, respectively. In this paper, due to the small experimental sample and the small image size, a smaller Vision Transformer is proposed, called ViT-Small. Various parameters are shown in Table 2.

4 Datasets and Experience

4.1 Dataset Description

In this paper, three public datasets are used for experiments: UCID, NRCS, CASIA v2.0 [1]. The images in the UCID and NRCS datasets are in TIF format, while those in the CASIA dataset are in JPEG format. UCID contains 1328 uncompressed color images with a size of 384×512 pixels, and NRCS contains 2259 gray images with a size of 1500×2100 pixels. 1000 images were randomly selected from these two databases, a total of 2000 images. At random positions in each image, we crop [25] out graphics with sizes of 64×64 pixels and 32×32 pixels as two real databases for the detection experiment of TIF format images and name them UN64, UN32. Besides, CASIA v2.0 contains 7491 true color images and 5123 tampered color images, which size ranging from 240×160 to 900×600 pixels. We randomly select 2000 real images, and crop out 64*64 and 32*32 images at random positions as two real databases for the detection experiment of JPEG format images, named CASIA64 and CASIA32 respectively.

Since the degree of USM sharpening is determined by the Gaussian kernel standard deviation σ and the sharpening intensity λ of the Gaussian filter, twelve sets of parameters about σ and λ were set to act on the four datasets respectively to form a tampered weakly USM sharpened image dataset. The experiment detected the real image and the tampered image with different weak sharpening parameters, and compared with current the most superior detection methods CNN network [27] and DCT-CHDMY [22].

The experiments in this paper were based on the PaddlePaddle framework. The dim is 128, the self-attention head is fixed as 16, the patch size and the number of transformer block are set to 4 and 5, respectively. Besides, the learning rate is initialized to 9e-5, and the decay strategy of the specified number of rounds is used. When the number of training rounds reached 20 and 30, the learning rate decays to 0.8 of the previous stage, a total of

Table 2: Parameter statistics for the base, large and huge variants of Vision Transformer

Model	Number of Layers	Hidden Size D	MLP Size	Heads	Number of Parameters
ViT-Small	5	48	198	128	1M
ViT-Base	12	768	3072	12	86M
ViT-Large	34	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 3: Detection accuracy of three methods on UN64

Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
$\sigma = 1$, $\lambda = 0.1$	**76.71%**	67.77%	56.82%
$\sigma = 1$, $\lambda = 0.2$	**80.22%**	70.16%	57.21%
$\sigma = 1$, $\lambda = 0.3$	**80.32%**	74.80%	58.64%
$\sigma = 1$, $\lambda = 0.4$	**84.43%**	78.35%	57.32%
$\sigma = 1$, $\lambda = 0.5$	**85.83%**	77.73%	58.64%
$\sigma = 1$, $\lambda = 0.6$	**87.93%**	80.76%	60.21%
$\sigma = 1$, $\lambda = 0.8$	**90.91%**	82.47%	62.64%
$\sigma = 0.8$, $\lambda = 0.1$	**72.07%**	65.33%	55.08%
$\sigma = 0.8$, $\lambda = 0.3$	**76.80%**	71.97%	66.04%
$\sigma = 0.7$, $\lambda = 0.1$	**70.65%**	64.35%	66.36%
$\sigma = 0.7$, $\lambda = 0.3$	**81.68%**	73.43%	67.06%
$\sigma = 0.5$, $\lambda = 0.5$	**79.44%**	73.35%	68.26%

50 epochs of training, and 64 pieces of data are fed into the training network each time. In the experiment, the method of double cross-validation [15] is used to evaluate the detection results. First, each data set is shuffled, half of the data is selected as the training set, and the other half is used as the test set, then the test data serves as training set, training data is regarded as the test set, and the average of the two test results is taken as the detection accuracy.

4.2 Experimental Setup

Three different sets of experiments were carried out. The UN64 and CASIA64 datasets are used in the first set of experiments to evaluate how well the ViT-Small performs for USM weak sharpening under various image formats. The accuracy of USM weak sharpening detection was tested using the UN32 and CASIA32 datasets in the second experiment, and the robustness of ViT-Small to image post-processing operations was examined in the third set of experiments. The detection effects of JPEG compression and additional noise after USM sharpening were combined to evaluate the robustness.

It is worth noting that in these experiments the performance of the algorithm was measured in terms of accuracy.

4.2.1 Experiment 1: Performance Preliminary Test

First, the UN64 data set was used for sharpening detection, which represented the detection effect of the three detection methods on TIF format images. Then use the CASIA64 dataset to conduct comparative experiments as the detection effect of JPEG format images. The detailed results were shown in Table 3 and Table 4, respectively.

From Table 3, it is obvious that our detection method surpasses CNN and DCT-CHDMY under these weak sharpening strengths, when λ increases, the accuracy rate will increase slightly in general. However, the relationship with σ is a little fuzzy, it is because the magnitude of σ is related to the intensity of overshoot artifacts, and the detection principle of the three algorithms being compared has little correlation with the intensity of overshoot artifacts.

Table 4 is a comparison of the detection accuracy of the three algorithms for 64*64 size JPEG images. Compared with the detection accuracy of TIF format images, the detection accuracy of ViT-Small and CNN all have a certain degree of improvement, this may be caused by different datasets. So it cannot be concluded that images in JPEG format are easier to detect USM sharpening than images in TIF format. Table 4 can only show that ViT-Small is superior to the other two algorithms in USM weak sharpening detection of small images in JPEG format.

Combining the analysis in Table 3 and Table 4, it can be concluded that whether the small image is in TIF or JPEG format, the accuracy rate of our proposed detection algorithm is better than that of the CNN and DCT-CHDMY algorithms.

4.2.2 Experiment 2: Image Size

In this section, we used the UN32 and CASIA32 datasets to perform sharpening detection experiments in turn, as the detection effect of 32*32 images. It should be noted that all pooling layers' pooling kernels were changed to 3*3 because the parameters in the CNN [27] structure were too large to be suitable for 32*32 images. Taking $\sigma = 1$, $\lambda = 0.1$; $\sigma = 0.5$, $\lambda = 0.5$; $\sigma = 1$, $\lambda = 0.8$ as an example, the experimental results are in Table 5 and Table 6.

From Table 5 and Table 6, we can observe when the image size is reduced to 32*32, for images in TIF format and JPEG format, the USM weak sharpening detection accuracy of each algorithm decreases. However, the performance of ViT-Small is still the best. It is quite obvious that the CNN method's accuracy rate drops to about 50%, which means that it is ineffective for USM weak sharpening detection of 32*32 pixels images.

Table 4: Detection accuracy of three methods on CASIA64

Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
$\sigma = 1$, $\lambda = 0.1$	**86.29%**	76.94%	55.69%
$\sigma = 1$, $\lambda = 0.2$	**88.70%**	82.27%	64.67%
$\sigma = 1$, $\lambda = 0.3$	**90.27%**	83.00%	68.26%
$\sigma = 1$, $\lambda = 0.4$	**90.72%**	85.01%	74.25%
$\sigma = 1$, $\lambda = 0.5$	**91.48%**	87.18%	77.84%
$\sigma = 1$, $\lambda = 0.6$	**91.52%**	88.18%	80.42%
$\sigma = 1$, $\lambda = 0.8$	**91.63%**	89.69%	84.80%
$\sigma = 0.8$, $\lambda = 0.1$	**85.83%**	77.58%	54.65%
$\sigma = 0.8$, $\lambda = 0.3$	**89.71%**	85.30%	72.37%
$\sigma = 0.7$, $\lambda = 0.1$	**85.63%**	74.26%	50.24%
$\sigma = 0.7$, $\lambda = 0.3$	**89.11%**	84.13%	69.06%
$\sigma = 0.5$, $\lambda = 0.5$	**89.50%**	82.56%	82.03%

Table 5: Detection accuracy of three methods on UN32

Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
$\sigma = 1$, $\lambda = 0.1$	**71.04%**	50.05%	65.27%
$\sigma = 1$, $\lambda = 0.8$	**82.08%**	50.07%	78.64%
$\sigma = 0.5$, $\lambda = 0.5$	**70.36%**	50.05%	68.46%

4.2.3 Experiment 3: Evaluating Robustness

In this set of experiments, we performed the robustness of ViT-Small to JPEG image compression [19] and Gaussian noise. Taking the CASIA64 dataset as an example, and the sharpening parameters were set to $\sigma = 0.5$, $\lambda = 0.5$. As the quality factor decreases, the image quality gets worse and the image texture and overshoot artifacts change more. The test results are displayed in Table 7 and Table 8.

As shown in Table 7 and Table 8, when the quality factor decreases, compared with Table 4 the USM weak sharpening detection accuracy of ViT-Small has not decreased significantly. Even if the quality factor $Q = 40$, the accuracy has only decreased by less than 5%. in addition, among these three algorithms, our accuracy drop is minimal. When Noise variance increase, the detection accuracy of ViT-Small almost did not decrease, while the accuracy of the other two algorithms decreased by more than 8%. Therefore, it can be said that ViT-Small is robust to JPEG compression and adding Gaussian noise.

Table 6: Detection accuracy of three methods on CASIA32

Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
$\sigma = 1$, $\lambda = 0.1$	**72.42%**	50.43%	60.06%
$\sigma = 1$, $\lambda = 0.8$	**84.97%**	49.56%	69.97%
$\sigma = 0.5$, $\lambda = 0.5$	**76.86%**	50.43%	58.25%

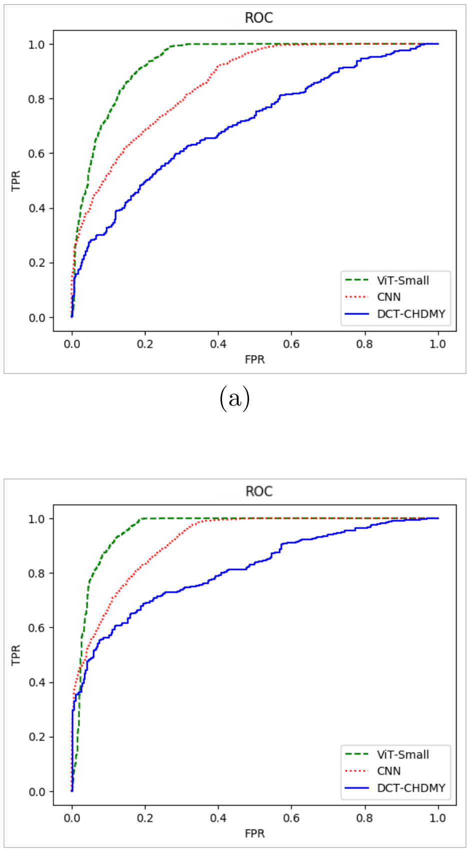
Table 7: Detection accuracy under JPEG compression

Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
Q=80	**88.50%**	78.22%	71.47%
Q=60	**85.36%**	76.46%	73.87%
Q=40	**83.83%**	74.56%	73.42%

Table 8: Detection accuracy under Gaussian noise

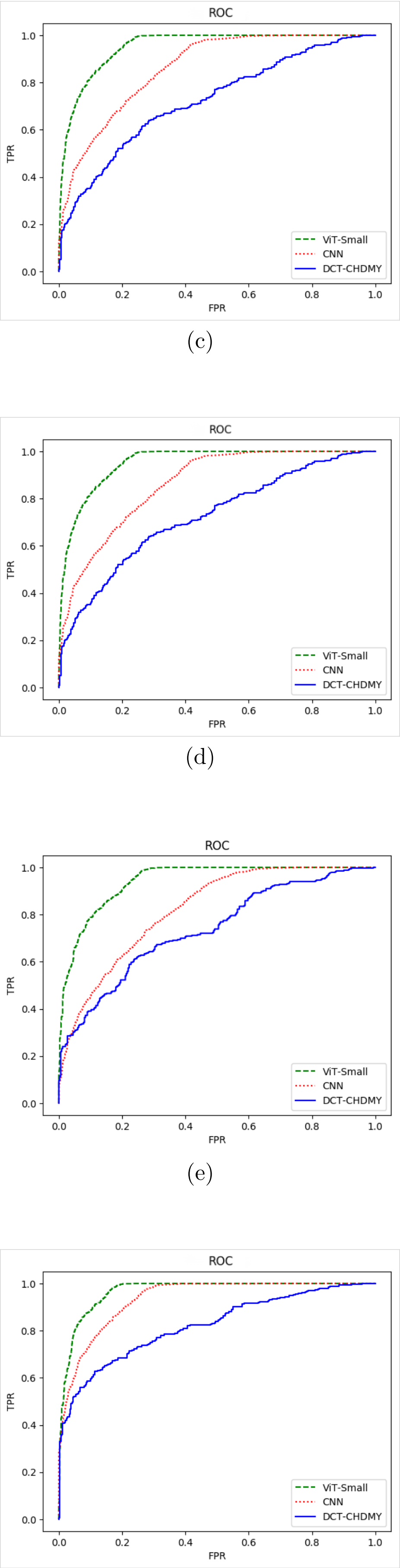
Parameters of USM	ViT-Small	CNN [27]	DCT-CH DMY [22]
1	**89.06%**	75.25%	68.46%
2	**88.40%**	73.55%	65.46%
3	**88.76%**	69.46%	61.41%

The receiver operating characteristic (ROC) curve was introduced in order to more easily compare the effectiveness of ViT-Small, CNN, and DCT-CHDMY. The relationship between true positive rate (TPR) and false positive rate (FPR) of the three algorithms under comparison can be seen in Figure 7 when the parameters were set to $\sigma = 0.7$, $\lambda = 0.3$; $\sigma = 0.7$, $\lambda = 0.3$; $\sigma = 0.8$, $\lambda = 0.1$; $\sigma = 0.8$, $\lambda = 0.3$; $\sigma = 1.0$, $\lambda = 0.1$; $\sigma = 1.0$, $\lambda = 0.3$; $\sigma = 0.5$, $\lambda = 0.5$. The outcome shows that ViT-Small outperforms the other two methods in terms of TPR for all specified FPR.



(a)

(b)

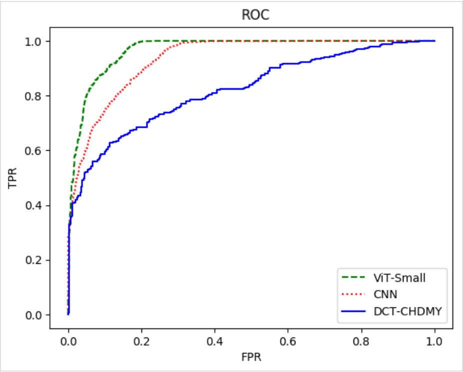


(c)

(d)

(e)

(f)



(g)

Figure 7: ROC curves of the three compared methods for USM sharpening detection as (a) $\sigma = 0.7$, $\lambda = 0.3$; (b) $\sigma = 0.7$, $\lambda = 0.3$; (c) $\sigma = 0.8$, $\lambda = 0.1$; (d) $\sigma = 0.8$, $\lambda = 0.3$; (e) $\sigma = 1.0$, $\lambda = 0.1$; (f) $\sigma = 1.0$, $\lambda = 0.3$; (g) $\sigma = 0.5$, $\lambda = 0.5$

5 Conclusion

In this paper, we for the first time focuse on small-size images USM weak sharpening detection, and propose a variant of the Vision Transformer model called ViT-Small, which overcomes the inefficiency of other USM sharpening detection in small-sized images. First, the image is converted into a YCbCr model, and the converted image is fed into ViT-Small. Extensive experiments prove that compared with current state-of-the-art method CNN [27] and DCT-DHCWY [22], our proposed structure has higher accuracy in detecting USM weak sharpening of small-size images in TIF format and JPEG format. Furthermore, it is robust to JPEG compression and Gaussian noise. All experiments were based on 2000 real images and 2000 tampered images. In the future, we will further investigate anti-forensic methods for USM sharpening.

References

[1] BIT, *Corel image database and the photographers*, Dec. 10, 2022. (http://forensics.idealtest.org/casiav2/)

[2] G. Cao, Y. Zhao, R. Ni, “Detection of image sharpening based on histogram aberration and ringing artifacts,” in Proc. *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2009.

[3] G. Cao, Y. Zhao, R. Ni, A.C. Kot, “Unsharp masking sharpening detection via overshoot artifacts analysis,” *IEEE Signal Process.:Lett.*,vol.18, no. 10, pp. 603–606, 2011.

[4] C. C. Chang, C. Chen, W. J. Kao, “A secure extended LBP data hiding scheme based on octagon-shaped shell,” *International Journal of Electronics and Information Engineering*, vol. 14, no. 5, pp. 497-508, 2021.

[5] M. S. Chang, C. P. Yen, "Forensic Analysis of Social Networks Based on Instagram," *International Journal of Network Security*, val. 21, no. 5, pp. 850-860, 2019.

[6] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding," *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, MI, USA, June 2019.

[7] F. Ding, G. Zhu, W Dong, "An efficient weak sharpening detection method for image forensics." *Journal of Visual Communication and Image Representation*, pp. 93–99, 2018.

[8] F. Ding, G. Zhu, Y. Q. Shi, "A novel method for detecting image sharpening based on local binary pattern," *in International Workshop on Digital-forensics and Watermaking (IWDW)*, 2013.

[9] F. Ding, G. Zhu, Y. Q. Shi, "A novel method for detecting image sharpening based on local binary pattern," in *International Workshop on Digital-forensics and Watermaking (IWDW)*, 2013.

[10] F. Ding, G. Zhu, J. Yang, J. Xie, Y. Q. Shi, "Edge perpendicular binary coding for usm sharpening detection," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 327–331, 2015.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al.*" An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint*: 2010.11929, 2020.

[12] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[13] Hany and Farid. "Image Forensics," *Annual review of vision science*, vol.5, pp.549-573, 2019.

[14] L. C. Huang, C. H. Chang, M. S. Hwang "Research on Malware Detection and Classification Based on Artificial Intelligence," *International Journal of Network Security*, vol. 22, no. 5, pp. 717-727, 2020.

[15] D. Krstajic, L. J. Buturovic, D. E. Leahy, *et al.*, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of cheminformatics*, vol. 6, no. 10, pp. 1-15, 2014.

[16] Z. Lai, E. Lu, W. Xie. "MAST: A memory-augmented self-supervised tracker," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR)*, 2020.

[17] D. Lepikhin, H. J. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006*.16668, 2020.

[18] L. Lu, G. Yang, M. Xia, "Anti-forensics for unsharp masking sharpening in digital images," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 5, no. 3, pp. 53–65, 2013.

[19] G. Singh, K. Singh, "Counter JPEG anti-forensic approach based on the second-order statistical analysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1194–1209, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, NY, USA, 2017.

[21] C. Wang, X. Bai, S. Wang, J. Zhou, P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 310-314, 2018.

[22] D. Wang, T. Gao. "An efficient USM sharpening detection method for small-size JPEG image," *Journal of Information Security and Applications*, val. 51, pp. 102451, 2020.

[23] X. L. Wang, R. Girshick, A. Gupta, K. He. "Nonlocal neural networks," *In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7794-7803, 2018.

[24] Y. Wang, X. He *et al.* "New image reconstruction algorithm for CCERT: LBP plus Gaussian mixture model (GMM) clustering." *Measurement Science & Technology* , vol. 32, no. 2, 2021.

[25] J. Yan, S. Lin, S. B. Kang, X. Tang, "Learning the change for automatic image cropping," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 971-978, 2013.

[26] J. Yang, J. Xie, G. Zhu, S. Kwong, Y. Q. Shi, "An Effective Method for Detecting Double JPEG Compression With the Same Quantization Matrix," *IEEE Transactions on Information Forensics and Security*, val. 9, no. 11, pp. 1933-1942, 2014.

[27] J. Ye, Z. Shen, P. Behrani, F. Ding, Y. Q. Shi, "Detecting USM image sharpening by using CNN," *Signal Processing: Image Communication*, vol. 68, pp. 258–264, 2018.

Biography

Jie Zhao received the B.S. degree in electronic information science & technology and the M.S. degree in communication & information systems from Ocean University of China, in 2006 and 2009, respectively. In 2015 he received the Ph.D. degree in information & communication engineering from Tianjin University, China. He is currently an Associate Professor at the department of electronic information engineering, Tianjin Chengjian University. His research interests include multimedia information security, image processing and computer vision.

Shuang Song is currently pursuing the M.S. degree in computer science & technology with Tianjin Chengjian University, China. Her current research interests include image forensics and deep learning.

Bin Wu received the M.S. degree in information & communication engineering from Beihang University, China, in 2002. He is currently a professor at the department of electronic information engineering, Tianjin Chengjian University. His research interests include machine vision, pattern recognition, FPGA logic front-end design, and embedded system development.

Guide for Authors
International Journal of Network Security

IJNS will be committed to the timely publication of very high-quality, peer-reviewed, original papers that advance the state-of-the art and applications of network security. Topics will include, but not be limited to, the following: Biometric Security, Communications and Networks Security, Cryptography, Database Security, Electronic Commerce Security, Multimedia Security, System Security, etc.

1. Submission Procedure

Authors are strongly encouraged to submit their papers electronically by using online manuscript submission at http://ijns.jalaxy.com.tw/.

2. General

Articles must be written in good English. Submission of an article implies that the work described has not been published previously, that it is not under consideration for publication elsewhere. It will not be published elsewhere in the same form, in English or in any other language, without the written consent of the Publisher.

2.1 Length Limitation:

All papers should be concisely written and be no longer than 30 double-spaced pages (12-point font, approximately 26 lines/page) including figures.

2.2 Title page

The title page should contain the article title, author(s) names and affiliations, address, an abstract not exceeding 100 words, and a list of three to five keywords.

2.3 Corresponding author

Clearly indicate who is willing to handle correspondence at all stages of refereeing and publication. Ensure that telephone and fax numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address.

2.4 References

References should be listed alphabetically, in the same way as follows:

For a paper in a journal: M. S. Hwang, C. C. Chang, and K. F. Hwang, ``An ElGamal-like cryptosystem for enciphering large messages," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 445--446, 2002.

For a book: Dorothy E. R. Denning, *Cryptography and Data Security*. Massachusetts: Addison-Wesley, 1982.

For a paper in a proceeding: M. S. Hwang, C. C. Lee, and Y. L. Tang, ``Two simple batch verifying multiple digital signatures," in *The Third International Conference on Information and Communication Security (ICICS2001)*, pp. 13--16, Xian, China, 2001.

In text, references should be indicated by [number].

Subscription Information

Individual subscriptions to IJNS are available at the annual rate of US$ 200.00 or NT 7,000 (Taiwan). The rate is US$1000.00 or NT 30,000 (Taiwan) for institutional subscriptions. Price includes surface postage, packing and handling charges worldwide. Please make your payment payable to "Jalaxy Technique Co., LTD." For detailed information, please refer to http://ijns.jalaxy.com.tw or Email to ijns.publishing@gmail.com.