Ensuring accuracy of personal data processed in blockchain systems

Olexander Belej[1,*,†], Yulian Fedirko[1,†] and Oleksandr Markelov[1,†]

[1] *Lviv Polytechnic National University, 5 Mytropolyt Andrei str., Building 4, Room 324, Lviv 79013, Ukraine*

Abstract

Based on the analysis of the existing requirements and methods of ensuring data security, the relevance of developing a method of ensuring data security during processing in blockchain systems by using artificial neural networks has been confirmed. A method of ensuring the reliability of personal data processed in blockchain systems has been developed. As part of the development of this method, the category of methods for ensuring data reliability was expanded by using artificial neural networks to identify unreliable personal data when they are entered into the blockchain system. A method of analyzing the authorization behavior of information system users has been developed. As part of the development of this method, user behavior was formalized and the possibility of detecting anomalies in user behavior using artificial neural networks was demonstrated.

Keywords

data security, blockchain systems, personal data, reliability, identification, artificial neural networks 1

1. Introduction

Since 2009, information systems based on blockchain technology have been gaining more and more popularity. Blockchain is a data processing technology based on the following basic principles: a data storage structure is a blockchain containing information built according to certain rules; each block in the chain uses the hash value of the previous block. This information applies to the nearest block;

Blockchain consists of the following basic elements: "useful" service data from the previous block in the chain; "Useful" data can be any information that needs distributed storage. For example, additional information may include the time the block was created, its computational complexity, and the random number used to calculate the hash. The hash

MoDaST-2024: 6th International Workshop on Modern Data Science Technologies, May, 31 - June, 1, 2024, Lviv-Shatsk, Ukraine

* Corresponding author.

† These authors contributed equally.

Oleksandr.I.Belei@lpnu.ua (O. Belej); Yulian.A.Fedirko@lpnu.ua (Yu. Fedirko); Oleksandr.E.Markelov@lpnu.ua (O. Markelov)

0000-0003-4150-7425 (O. Belej); 0000-0001-9968-7313 (Yu. Fedirko); 0000-0002-2432-0768 (O. Markelov)

sum of the previous block is used to uniquely order the blocks. The exception is the hash sum of the previous block specified in the genesis block, which is usually randomly generated. The hash of the current block verifies the information contained in that block and relates it to the next block in the chain. In general, blockchain works like Figure 1. With the development of blockchain technology, its basic principles are also constantly evolving and changing.

To solve these problems, new blockchain privacy solutions are constantly emerging based on cryptographic privacy technology, which provides users with a mechanism of anonymity and control over their data when conducting any digital transaction on the ledger, following the principle of self-verification. Protected Sovereign Identity (PSI) template [1]. The author of [2] evaluated Acce-chain through experiments, and the results showed that the coverage performance is feasible under real VEC settings, and the query efficiency can be several orders of magnitude better than the baseline.

The proposed approach covers all aspects of the national health insurance scheme and therefore allows making changes to existing procedures without changing the rules of the health insurance system [3]. The author of [4] used Remix IDE to rigorously test the smart contract code in various scenarios and referenced the code on GitHub. Blockchain verifies the authenticity of IoT devices and cloud service providers added to the network and provides a mechanism to manage IoT data access policies. In addition, a prototype of the proposed framework was implemented using Hyperledger Fabric and Intel SGX, and an analysis of blockchain and SGX performance was also presented [5].

In [6] authors proposed an incentive mechanism to assess the value of publishers' efforts in managing and maintaining research data and creating new blocks. The results demonstrate the effectiveness of the proposed system in managing large datasets with low latency. This paper [7] proposes a new incentive mechanism that uses university degrees to save academic records and create new blocks. We conduct large-scale experiments to evaluate the performance of UniChain, and the results show the effectiveness of the proposal in processing large data sets with low latency.

The author [7] proposed a maritime transport communication system that supports the Internet of Things. The system is a decentralized system consisting of base stations and sea buoys. Agricultural insurance can help smallholder farmers in developing countries manage risks that they cannot manage on their own [8].

The study [10] examines the needs and prospects of using blockchain technology in the Internet of Things. Consider implementing encryption for open, decentralized IoT systems not to restrict viewing, but to limit unauthorized access. By integrating hash functions and digital signatures into the blockchain itself, it demonstrates the ability to protect data from unauthorized access. This is achieved using an encryption algorithm based on a pseudo-random number generator.

This study [11] is based on the above facts and aims to explore how to make blockchain

GDPR compliant. As such, it contains several proposals to make blockchain technology more GDPR compliant.

In the article [12], the authors describe the design of a system for the deployment and processing of survey data following the GDPR. It combines the Hyperledger Fabric blockchain for data immutability and the InterPlanetary File System (IPFS) for storage. Paper [13] developed a healthcare system that can securely manage personal medical data and create interactions between doctors, patients, insurance companies, and pharmacies or medical stores.

Today, blockchain technology can be conditionally divided into five generations.

The first generation of blockchain (Blockchain 1.0) is the basis of digital payment systems, the first and most popular representative of which is Bitcoin, launched in 2009.

One of the main disadvantages of Bitcoin is the hash calculation method. Since the task of calculating the hash value is solved in a decentralized manner (which is good because it increases the reliability of the chain), then only one calculation participant (miner) can become the winner. Therefore, most of the miner's work is wasted, because... the calculations performed are useless. As of December 2021, the total computational power of Bitcoin miners is approximately 174 petaflops per second. This leads to another disadvantage - the tendency to centralize calculations. In the past, individual miners may have been the winners, but today, as the total computing power of miners increases, so does the computational complexity, and the only way to calculate hashes (and get rewarded) is faster than other methods. Just unite the miners. According to a report by digital asset management company CoinShares, as of June 12, 2021, approximately 65% of the effective computing power of cryptocurrency mining equipment is concentrated in China.

Blockchain of the second generation (Blockchain 2.0) not only supports the functions of registration, confirmation, and transfer of currency but also supports other types of assets - various contracts and properties. Second-generation blockchain protocols can use Bitcoin's decentralized ledger or create their decentralized ledger (Figure 1).

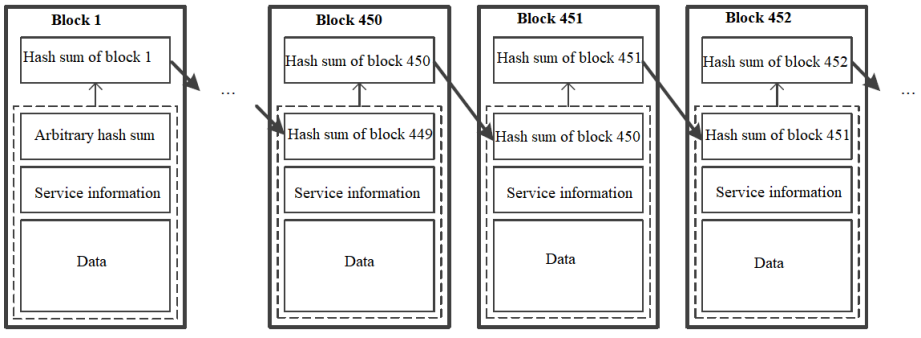

Figure 2: Blockchain diagram based on current block computing power: [10].

The areas of application of second-generation blockchain technology can be divided into intellectual assets; as part of this effort, it seems appropriate to clarify the basic information about smart assets and smart contracts. Intellectual property rights allow you to trade any property. After assets are registered in the decentralized ledger, control of the property is transferred to the key holder. Transfer of private keys means transfer of ownership.

The general meaning of smart contracts comes from the idea of smart assets. Smart contracts are a method of conducting transactions in a decentralized ledger based on the use of cryptocurrencies and smart assets to sign agreements through blockchain technology. An example of a smart contract is a transaction that remains inactive in the decentralized ledger until a certain date or event: the transfer of inheritance rights on the day of the death of the owner of the asset, the purchase or sale of an asset on the day of the death of the owner of the asset. date of death of ownership. date. If the owner of the asset dies, the exchange takes place after notification and ownership automatically pass from the finance company to the individual after all loans are paid off. Procedure from the point of view of judicial practice, high-quality contract drafting, and the introduction of automatic enforcement mechanisms can significantly reduce the number of disputes. The combined use of smart assets and smart contracts can create a lending system that uses the borrower's smart assets as collateral, thereby reducing the cost of insurance against fraud and abuse and making lending safer and more profitable. A distinctive feature of smart contracts is that there is no need for trust between participants — smart contracts are executed automatically using code running on blockchain technology, leaving no room for the human factor. However, the use of smart contracts today requires a strict regulatory framework to regulate the procedures for fulfilling contractual obligations.

One of the key directions in the development of third-generation blockchain technology is the use of methods based on directed acyclic graphs (DAG). A directed acyclic graph is a topology tree-based data processing structure. The arrangement of blocks in this structure does not have to be contiguous and provides direct communication between any transactions on the chain. The chain in this structure is not built by blocks but by transactions. The hash value is calculated from the parent transaction and passed to the next related transaction. The main advantages of using direct acyclic graphs are speed, ease of growth, and increased security of data processing systems. In the first generation of blockchains, it took about 10 minutes to create a new block. Creating a second-generation blockchain takes just 20 seconds. When using a direct acyclic graph, there is no need to collect transactions into blocks, and theoretically, hundreds of thousands of transactions per second can be guaranteed. The developers of blockchain-based systems refuse to avoid the high complexity of hash calculations, which leads to the need not to organize miners into mining pools, which in turn leads to a more decentralized network and therefore higher profits.

In general, third-generation blockchains, regardless of whether direct acyclic graphs are used or not, can solve the currency-independent problems of market transactions. Examples of such solutions include:

- email security system KeyID, a system that combines 32-bit alphanumeric identification codes with human-readable names called OneName and BitID, a system that identifies Bitcoin wallet addresses based on the Bihandle type;

- provision of services for authentication of full documentary confirmations (regarding confirmation of authenticity of wills, contracts, powers of attorney, medical certificates, promissory notes, etc.) without disclosing the information contained in them;

- a personalized government that provides instant cryptocurrency payments for active PR and commissions for event organizers;

- control of some traditional public services;

- an identification system that provides people with foreign passports that are not tied to a specific country;

- WikiLeaks and Twitter document solutions to combat online censorship.

One of the technical solutions based on fifth-generation blockchain technology is the Telegram Open Network (TON) project. TON is a platform for creating a blockchain ecosystem that provides storage of personal data in cloud storage and registration in services that require authentication. TON consists of the following main elements:

1. Blockchain is the main component of TON;

2. TON network - provides communication between all TON components;

3. Services, services, and applications Platform that provides services for TON applications;

4. TON Payments - provides payment services.

The TON blockchain includes several chains:

1. Master chain. Contains information such as system parameters, working chain state, hashes of all recent blocks, and the number of GRAM tokens issued.

2. Work chain. They connect chains of "shards". Each worker thread has a unique ID and logic and can have its own virtual machine and address format. TON supports a total of 232 work chains, and each work chain can contain up to 260 segment chains.

3. Broken chain. Ensure system expansion. You can share messages. Follow the chain of command rules.

4. Chain of accounts. virtual chain.

5. Part of a fragment chain. They are a kind of register of incoming and outgoing messages for a certain account.

The architecture used in TON provides a solution to two important problems - the large size of the blockchain and the high complexity of making changes to the blockchain architecture. The first problem is solved by special methods of data storage - the file can be stored off-chain and store only the hash value of the file, or the smart contract containing this data can be stored in the corresponding block. Information about conditions. data in the block. The document is stored in the chain. The second problem is addressed by the infinite sharding paradigm, which groups account chains into shard chains such that each shard chain block contains a shard chain block. At the beginning of 2018, $1.7 billion was raised for the development of the project as part of the ICO. Closed beta testing began in April 2019. As of December 2021, one GRAM token was worth approximately $0.004.

In general, blockchain systems can be divided into two categories based on the differences: public and private.

In a public system, access to the participating network is open, and anyone can create new entries and have read access to existing entries. Such solutions are recommended for cryptocurrencies, an example of such a system is Bitcoin,

In a private system, permissions are required to create new records or read existing records. Applications for such systems include enterprise systems as well as manufacturing and supply chains. Examples of such systems are Hyperledger, Hashgraph, R3 Corda, and Quorum.

When determining the feasibility of using blockchain technology and determining the type of blockchain system, the following basic data processing conditions must be taken into account:

- do you need data storage?
- multiple users are required to record data;
- lack of reliable data confirmed by third parties;
- is anonymity required to determine the type of blockchain system required – public or private;
- whether a public profile check is required (to determine whether to use a public exclusive system or a private exclusive system.

Some experts believe that the bills need serious changes.

In the United States, the IRS considers bitcoin a valuable asset and imposes a capital gains tax on bitcoin transactions. Meanwhile, some US government agencies are trying to regulate

Bitcoin as a currency.

Blockchain technology has the potential to become Occam's Razor, the most efficient, direct, and natural means of coordinating all human behavior in response to the natural desire for balance.

2. Problem statement

It is assumed that personal data (PD) should be processed in the blockchain system at least during the entire life cycle of the subject of personal data. During the entire process of personal data processing, their safety, confidentiality, availability, integrity, and reliability should be ensured.

Currently, the challenges of creating a blockchain system architecture that is free from the main threats associated with blockchain systems include ensuring the reliability of the processed data.

Therefore, the construction of a protected decentralized registry of personal data (DRPD) boils down to solving the following tasks:

- to determine the composition of a black hole, it must be processed in DRPD;
- define the overall DRPD architecture;
- determine the order of data storage;
- determine mechanisms for reaching consensus including procedures for providing rewards to users to ensure the operation of the DRPD and procedures for automatic assessment of the risks of processing unreliable personal data;
- select the hash function calculation method;
- determine the general sequence of development of the DRPD.

It seems appropriate to use machine learning techniques to implement automated risk assessment within consensus mechanisms. For example, it is recommended to identify four characteristics that can be used to conduct a risk analysis:

- the degree of formal connection between the consensus node and the confirmation object;
- mutual confirmation of participants in a PD network can indicate collusion and user behavior that is distributed among multiple people and is therefore particularly difficult to detect;
- the amount of potential compensation to PD subjects;

- reliability of consensus nodes and verification objects.

During the creation of the DRPD, the following issues must be resolved:

- determine the purpose of PD processing and the corresponding PD components, and their processing should be carried out in a decentralized system;
- define the overall DRPD architecture;
- determine the order of data storage;
- develop a consensus mechanism to encourage user participation in ensuring the functioning of the DRPD and conducting automated assessments of the risks of entering and processing unreliable personal data in the DRPD;
- determine the method of calculating the hash function;
- to determine the general sequence of development of DRPD;
- defines the method for calculating PD trade-offs when processing PD trade-offs in DRPD.

Figure 2 shows the recommended approach to protecting personal data when using DRPD.

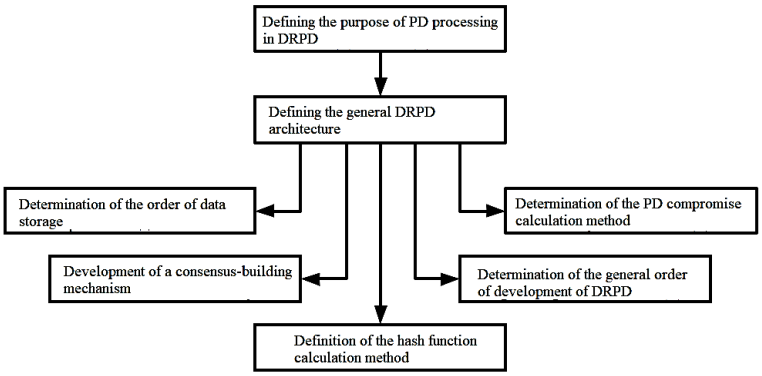

Figure 2: The holistic approach of a decentralized registry of personal data to protecting human resources.

3. Defining a distributed ledger architecture

Taking into account the purpose and composition of the PD, as well as the large volume of data that must be processed in the DRPD, it is recommended. DRPD has several independent private blockchains, one for each subject area:

- the main chain blockchain is for data identification information (BDI);
- blockchain of Work (BDE) for recording educational data;
- the Jobchain (BDS) blockchain is for skills data;
- blockchain Smart Asset Information (BDSA) workflow for asset data;
- smart Contract Information Dedicated Blockchain Work Chain (BDSC) for Smart Contract Information (PD Category - Other);
- use second-generation blockchain technology as the basis of the main chain, replacing transactions with blocks to reduce traffic and load on network node computing resources;
- using third-generation blockchain technology, a direct acyclic graph is used as the basis for the BDE, BDS, BDSA, and BDSC work chains, as this approach will allow references to specific files and other blocks to be included in the blockchain.

It is recommended that each blockchain contains a unique user ID and the following PD:

- BDI - Information about identity cards and passport data: series and number, issue and time of issue, sample personal signature, surname, first name, patronymic, gender, date of birth, place of birth, place of residence, military information, family information about status, information about children, information about a previously issued passport;
- BDE - information about education: name of the educational institution, teachers and specialties, years of study, composition of subjects, and success rate;
- BDS - information about professional skills: name of the place of work, unit and position, work experience, job duties and components, key skills;
- BDSA – Asset Intelligence: non-cash funds, stocks, mutual funds, bonds, cryptocurrencies, real estate, vehicles;
- BDSC - smart contract data: employment contracts, contracts for the provision of various services, and contracts for the purchase and sale of goods.

Figure 3 shows a general block diagram.

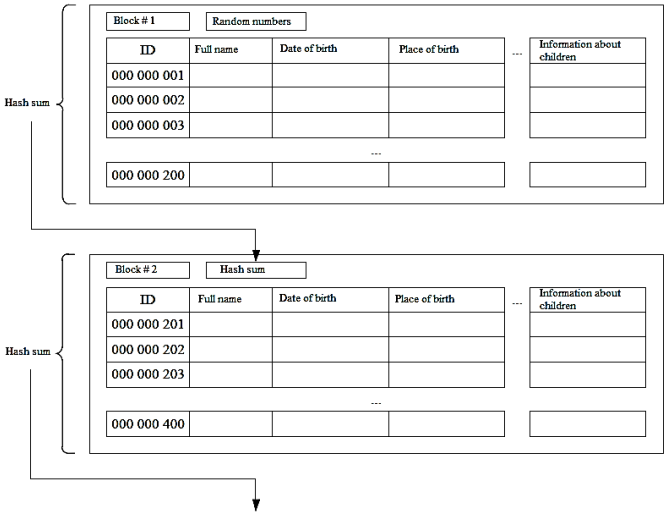

Figure 3: Block structure of various chains.

The generalized DRPD hierarchy proposed by the authors is shown in Figure 4.

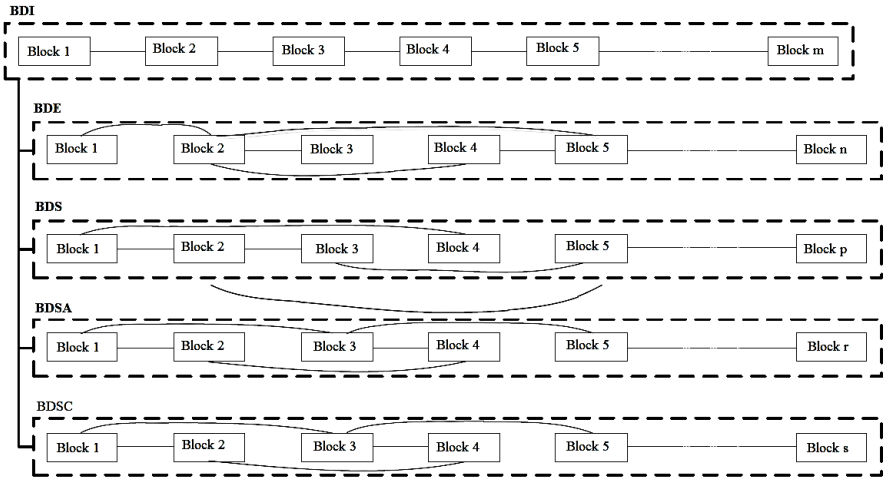

Figure 4: General hierarchy of decentralized registry of personal data.

It is recommended that DRPD nodes be divided into three types: consensus nodes, audit nodes, and thin clients. Consensus nodes must participate in the formation of new blocks by contributing PD to the block and distributing it throughout the network. The audit node must contain a copy of the blockchain and ensure load distribution across the network, jointly acting as a content delivery network (CDN), i.e. providing: data transfer between light clients and consensus nodes; reducing the volume of transit used to prevent delays, breakdowns and loss of communication in congested corridors and their intersections. Thin clients are designed to be installed on platforms with lower performance characteristics, including mobile platforms, and may contain only the necessary host information.

Therefore, people who use DRPD can be divided into two categories: users and

operators:

1. User:

- Submit your PD to DRPD;
- If necessary, obtain and provide third parties with access to your data;
- If necessary, provide a personal device for storing data in an encrypted form.

2. Operator:

- carry out technical control over the activities of DRPD;
- Make sure the PD entered in the DRPD is correct, create a new block, and enter the DRPD.

As a basis for DRPD, you can use ready-made solutions or develop new ones. The core of the DRPD platform is proposed to be implemented in the Java 8 programming language using a NoSQL database. Provides interaction between the RESTful API architecture and the core of the platform. The Erachain platform has taken a similar approach to creating a decentralized login code architecture specifically designed to handle personal data. Safe software development practices are recommended when developing DRPD. In addition, the methods of calculating the reliability of complex systems can be used in the design of DRPD.

This block should contain approximately 1.5 KB of identification information per person, 12 KB of education and skills information per person, and 1 KB of information per smart asset and smart contract.

Based on the above, Table 1 guides the appropriate number of blocks that should be generated initially during system creation and how often new blocks should be generated.

Table 1

A proposal for the number and frequency of creation of new blocks

Name of the blockchain	The volume of one block	Number of blocks initially created	Approximate volume of the blockchain		Average approximate frequency of creating new blocks
			at the start	Increase per year	
BDI	300 KB	6 250	1.9 GB	0.1 GB	1 per day
BDE	2,400 KB	1 001	2.4 GB	2.6 GB	3 times a day
BDS	2,400 KB	2 815	6.8 GB	7.0 GB	8 times a day
BDSC	200 KB	50 000	10.0 GB	7.6 GB	1 time every 10 minutes
BDSA	200 KB	50 000	10.0 GB	7.6 GB	1 time every 10 minutes

Therefore, the current DRPD data volume will reach approximately 31.1 GB at system launch and will grow to 24.9 GB each year.

Due to the specificity of the thematic fields, BDE does not publish new blocks every day but mainly contains information about additional professional education received. But the fall and spring will be the season when new quarters with information about secondary and higher education will appear.

It is recommended to store large amounts of data on DRPD user personal data storage media as audit nodes or consensus nodes. If biometric data needs to be processed, masking compression methods based on weighted image structure models can be used. To ensure the confidentiality of personal data, it is recommended to ensure that it is encrypted. Current tasks also include the development and certification of decentralized systems using blockchain technology to meet the requirements of the Cryptographical Information Protection Facility (CIPF).

PD-distributed storage must be able to store large files. In addition, decentralized PD storage requires a version control file system with persistent access capabilities that can uniquely map unique files to their hash values to verify file integrity and the absence of undeclared functions. An example of a system that can provide such functionality is the InterPlanetary File System (IPFS) project. IPFS combines BitTorrent's peer-to-peer file-sharing technology with the capabilities of Git, a decentralized version control system created to manage software development, but can be used for any digital resource. Transactions listed in a blockchain block may contain references to files stored outside the network and methods of accessing them. In addition, IPFS is designed using direct acyclic graph technology, is compatible with the technical architecture of cryptocurrency, and rewards file-sharing nodes in the form of Filecoin coins. Therefore, IPFS can serve as a technical solution for processing large volumes of data.

It is also recommended to include provisions for archiving unused blocks in the blockchain. You can archive using the Internet Archive, the Wayback Machine, or similar systems.

For DRPD, it is necessary to ensure the first level of security:

- when used with DRPD, your computer may have software with undeclared features;
- DRPD includes biometrics and other PD categories.

Creating new blocks doesn't have to be a time-consuming task. Considering the specifics of the considered blockchain system, the Proof-of-Authority algorithm appears to be the most appropriate, designed to ensure the operation of a private network and allow the identification of privileged validators. Its functionality is proposed to be expanded with the help of a program that automatically evaluates the reliability of data entered into the blockchain system.

Figure 5 shows an overview of the PD record verification procedure proposed by the authors in their notebooks.

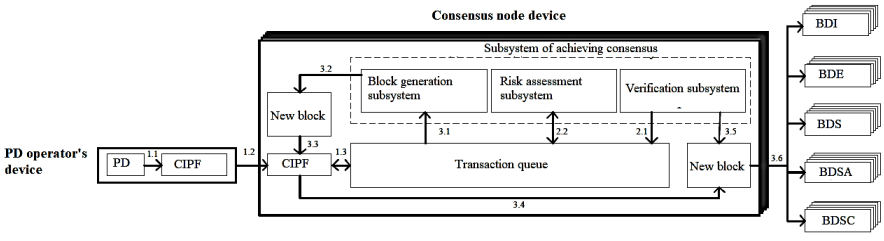

Figure 5: Check the general register data input scheme.

Since DRPD requires user computing resources, which are used to store data and perform calculations when new blocks are created, and users' confirmation of new blocks, a user reward system must be established. Developed

One possible option is to implement a performance-based (PR) system.

The value of PR can increase based on success in education, career, contract fulfillment, etc.

The reward for the creation of new blocks can be realized with the help of individual accounting units of reward (AUR). At the same time, the opportunity to participate in the creation of a new block and the probability of success should depend on the user's PR.

Reward methods and their definition require careful study and are beyond the scope of this study. In Table 2, the authors provide general examples of possible reward values to demonstrate the general principles of rewarding users. Fuzzy set theory can be used to demonstrate the value of rewards. In the proposed example, users with a PR value of 6 can become consensus nodes.

Thus, the unit of remuneration in accounting for mining activity is:

This applies to the BDSA network only;

After all, we are talking about a decentralized registry that randomly selects a PR value of at least a given value and expresses its willingness to become a user of the consensus node.

The probability of a consensus node winning depends on the user's PR value.

Table 2

Generalized example of possible reward values

Chain	Basis for remuneration	Reward amount by which the PR value increases
All	Registration	0.5
All	Data storage from 10% to 100% chain	0.1–0.5
BDE	Getting an education	0.5/1/2
BDE	Advanced training/professional retraining	0.5/1
BDE	Obtaining an academic degree	3/6

Registration of such information can satisfy the subject's desire for economic benefit, high social status, etc. To ensure the reliability of entering registration information, it is recommended to implement a mechanism for automatic assessment based on machine learning methods of the risks of entering and processing unreliable personal data. As initial

data, it is recommended to determine the factors that create prerequisites for entering and processing unreliable PDs in the DRPD. These factors can be:

- increased probability of collusion between identified subjects and targets;
- the possibility of substantial compensation;
- the reliability of the confirmed object is low low PR value, reflecting low material happiness and, a lack of necessary knowledge and skills.

For example, it is recommended to identify the four characteristics described in subsection 2.2 and take into account that a risk analysis can be carried out:

- the degree of formal connection between the consensus node and the confirmation object;
- degree of participation in networks of mutual recognition of personal data;
- the amount of potential compensation to the PD subject;
- reliability of consensus nodes and validation objects.

Therefore, in the considered paradigm, the risk of encountering an unreliable PD will be represented by the risk of collusion between consensus nodes and validation objects. Each presented feature can be represented by the coefficient $x_n,\ n \in \{1;\ 4\}$, where n is the number of the feature:

- x1 - The degree of correlation between the consensus node and the confirmation object;
- x2 - Participation in the mutual confirmation of the PD network;
- x3 - the potential amount of remuneration of the subject of personal data;
- x4 is a consensus node and confirms the reliability of objects.

When confirming PD, it is recommended to use ANN as a mathematical tool for risk assessment. Therefore, the input end of the neural network must have four input signals x1-x4, and the construction of the ANN is reduced to solving the following problems:

- determine the required type of ANN;
- develop a method of assigning numerical values (x1-x4) to the input signals of ANNs expressing analytical features;
- determine the number of necessary ANN layers and the number of neurons in the ANN layer;
- selection of ANN training methods;
- selection of the activation function;
- select the NET output range to indicate the level of risk confirmed by the PD.

Figure 6 presents a summary diagram of the necessary ANN, proposed by the authors.

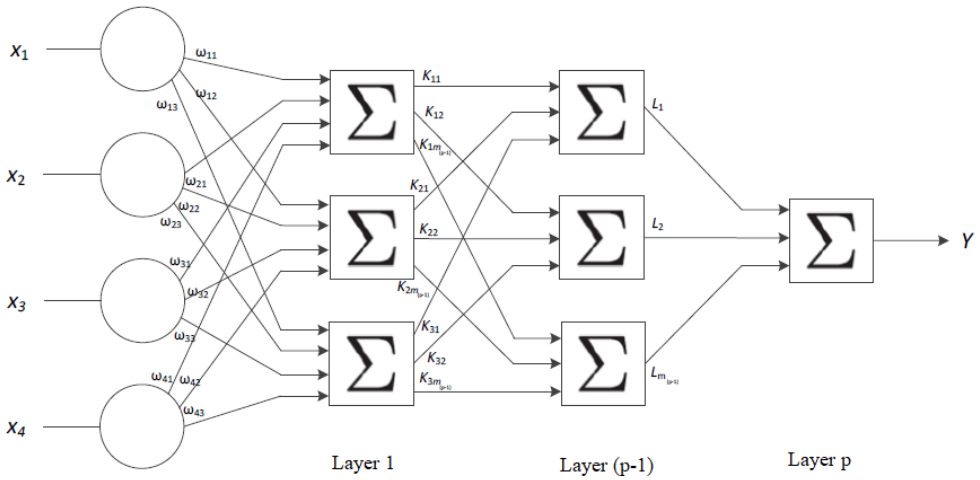

Figure 6: Generalized graph of ANN for determining the reliability of PD.

When learning ANNs, the importance of input values is determined by changing the weight coefficients of neural connections. when building a neural network, the following is recommended:

- use the mathematical method of fuzzy set theory to assign values to input signals;
- use the well-studied multi-layer fully connected perceptron as a feedback-free neural network;
- a neural network consists of three layers;
- use the backpropagation algorithm as a learning method;
- to minimize the RMS error of the neural network when training the neural network, use the hyperbolic tangent as the activation function;
- the range of initial values [-1;1] should be interpreted as follows: -1 - the minimum risk of PD unreliability, and 1 - the maximum risk of PD unreliability.

Within the framework of the problem under consideration, the symbol xn is proposed to be considered as:

The membership of $\mu_A(u)$ to the eigenfunction of the set of values, A represents the increased probability of reaching a given unreliable PD on the universal set U,

The value of the elements of the set U that belong to the set A is equal to 1, and the value of the elements that do not belong to the set A is equal to 0:

$$\mu_A(u) = \begin{cases} 1, if\ u \in A \\ 0, if\ u \in A \end{cases}. \quad (1)$$

In this case, it is necessary to consider its own set for each membership function. Examples of the four functions of object ownership are:

- a function belonging to the set of values of the degree of connection between consensus nodes and verified objects, under which the most favorable conditions of collusion are created;

- a membership function for a set of intermediate confirmation values that demonstrate an increased probability of participation in a conspiracy;
- determination of the membership function of the set of values of the reward object, which provides the greatest incentive to participate in the conspiracy;
- the overall value of the functionality and reliability of the verified objects included in the group of consensus nodes creates minimal prerequisites for participation in the conspiracy.

Figure 7 presents the Zade diagram 33, which shows the possible dependence of the value of the characteristic membership function on the set of values of the degree of connection between consensus nodes and verified objects, which creates the most favorable conditions for a consensus conspiracy. The degree of contact between the node and the confirmed object, at this time:

Ua - a set of values indicating the degree of connection between the consensus node and the object being checked $u_a = [u_a, u_a \in R: 0 \leq u_a \leq 10]$;

Aa is a set of values of the degree of connection between the consensus node and the object under test. With this value, the most favorable conditions for participation in the conspiracy are created;

$\mu_{A_a}(u_a)$ is a characteristic function of this group value, which belongs to the degree of connection between the consensus node and the verification object? With this characteristic function, the most favorable conditions for participation in the conspiracy are created;

x(u_a) – The degree to which the eigenvalue of the function belongs to the set of relation values $\mu_{A_a}(u_a)$ between consensus nodes and verified objects, among which the most favorable entry condition belongs to the collision.

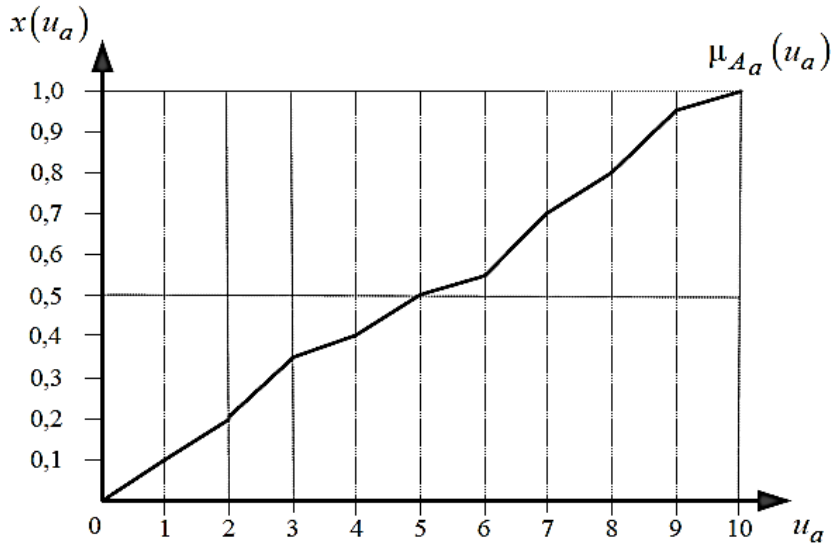

Figure 7: Dependence on the degree of connection between the verification object and the consensus node and favorable conditions for collusion.

In the given example, it is assumed that 0 corresponds to different degrees of

connectivity, 6 along the abscissa axis.

After PR, as the degree of contact between the consensus node and the verification object increases and the most favorable collusion conditions appear, the membership possibilities continue to increase.

For a more convenient interpretation of data when forming training samples and, if necessary, their normalization, it seems recommended to reduce the obtained results to a general representation of fuzzy subsets:

$$A_\alpha = \sum_{u_a=0}^{10} \mu_{A_a}(u_a)/u_a = \sum_{u_a=0}^{0} 0{,}00/u_a + \cdots + \sum_{u_a=10}^{10} 1{,}00/u_a. \quad (2)$$

When forming training samples, it seems recommended to determine values that can negatively affect the learning process of neural networks - these values do not allow us to draw clear conclusions about anomalies in user behavior. In the theory of fuzzy sets, this value is determined by the transition point. For the membership function μ A, such a point is $_a = 5$.

Figure 8 shows the Zade diagram, which shows the possible dependence of the value of the attribution function in the set of probability of collusion of the PD subject on the degree of mutual confirmation of participation:

U_b - a set of intermediate confirmation values;

A_b - a set of confirmed intermediate values showing an increased probability of participation in a conspiracy;

$\mu_{A_b}(u_b)$ is a characteristic function of the degree of belonging to a set of intermediate confirmation values, which demonstrates an increased probability of participation in a conspiracy;

$x(u_b)$ is the value of the characteristic membership function μ Ab (ub) for the set of intermediate values of the confirmation, which represents the increased probability of participation in the conspiracy.

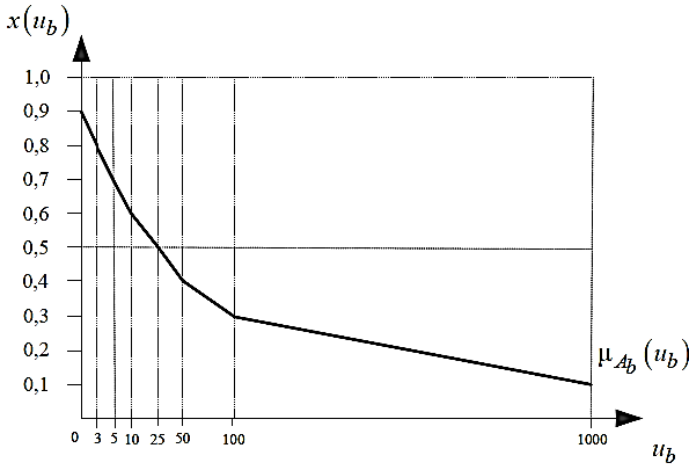

Figure 8: The relationship between the level of participation and the probability of mutual confirmation of the conspiracy IoT.

In our research, the horizontal axis shows the u_b values corresponding to the intermediate confirmation numbers: 0, 3, 5, 10, 25, 100, 1000.

Since PR, as the number of mutual confirmations increases (as the value of u_b increases), the probability of collusion decreases (the value of $x(u_b)$ decreases) and the membership function also decreases.

The general form of recording fuzzy subsets will have the following form:

$$A_b = \sum_{u_a=0}^{1000} \mu_{bA_b}(u_b)/u_b = \sum_{u_b=0}^{0} 0{,}9/u_b + \cdots + \sum_{u_b=1000}^{1000} 0{,}1/u_b. \quad (3)$$

The transition point of the membership function *x(ub) = 0.5* is equal *to ub = 25*.

Figure 9 shows a Zade plot illustrating the possible dependence of the eigenvalues of the membership functions in the set of conspiracy motivation values on the potential reward size of the confirmed object, where:

Uc – identifies a set of potential object reward values expressed as PR values;

Ac is the set of reward values that provide the maximum incentive to collude for confirmed objects;

$\mu_{A_c}(u_c)$ – Determine the characteristics of the membership function that provides the maximum incentive for collusion within a set of target reward values;

$x(u_c)$ is the value of the characteristic membership function, which determines the size of the object's reward and provides the greatest incentive for collusion.

The examples below assume the following:

- the higher the potential reward, the higher the incentive to collude;
- the following key salary values can be distinguished, reflecting certain achievements: 0.1;
- the goal is to demonstrate with concrete numerical examples the general principles of membership functions that shape the characteristics of neural networks and future input values.

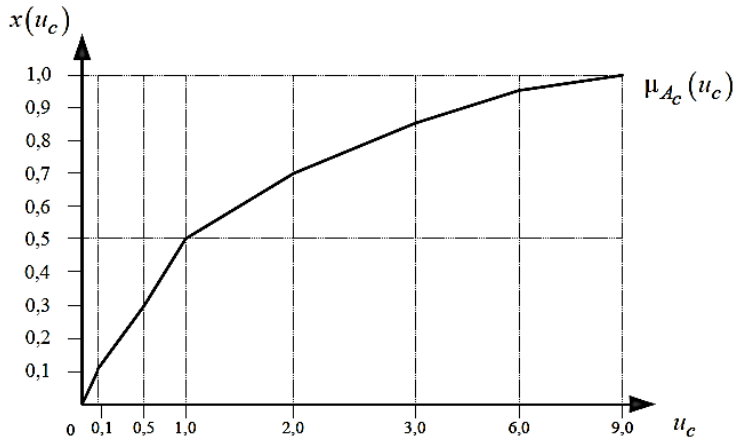

Figure 9: The relationship between the size of the reward and the motivation to participate in the conspiracy.

Due to PR, as the reward increases (as u_c increases), the incentive to collude increases as *x(uc)* increases, and the membership function also increases.

The general form of recording fuzzy subsets will have the following form:

$$A_c = \sum_{u_c=0}^{9} \mu_{A_c}(u_c)/u_c = \sum_{u_c=0}^{0} 0{,}9/u_c + \cdots + \sum_{u_c=9{,}0}^{9{,}0} 1{,}00/u_c. \quad (4)$$

The transition point of the membership function $x(u_c) = 0.5$ is $u_c = 1$.

Figure 10 shows a Zade plot showing the possible dependence of the values of the characteristic membership functions in the set of conspiracy motivation values on the reliability of consensus nodes and confirmation objects, where:

- Ud is a consensus node, which is a set of values confirming the overall reliability of the object represented by the PR value;
- an announcement is a set of trust points that creates the minimum prerequisites for participating in a conspiracy;
- $\mu_{A_d}(u_d)$ is a characteristic function belonging to a set of values that creates minimum prerequisites for participation in a conspiracy;
- x(ud) –A functionally important characteristic of a set of reliability values, which creates minimal prerequisites for participation in a conspiracy.

This research assumes the following:

- the higher the overall authority and verification goal of the consensus node, the lower the incentive for collusion;
- the PR value of the consensus node is 8, which is 100% reserved after registering with DRPD;
- fields of scientific degree and availability of candidates of sciences.

As an example, below are the values of control points PR and ud of matched nodes:

- confirmation that the goal has been achieved: 13 have successfully repaid a loan of 10 million rubles;
- consensus points reached: 46.5, 52.5, 70.5, 73.1, 74.

The goal is to demonstrate with concrete numerical examples the general principles of membership functions that shape the characteristics of neural networks and future input values.

As PR increases along with the trustworthiness of consensus nodes and validators as *ud* increases, the probability of collusion decreases as *x(ud)* increases and the membership function increases.

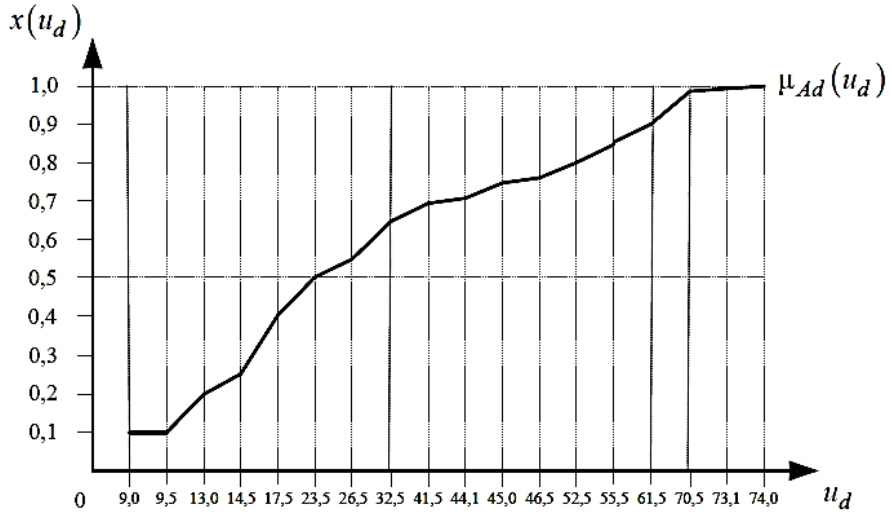

Figure 10: Dependence on the reliability of the verification object and the consensus node and the probability of collusion.

Therefore, the general form of recording fuzzy subsets will have the following form:

$$A_d = \sum_{u_d=0}^{74} \mu_{A_d}(u_d)/u_d = \sum_{u_d=0}^{9,0} 0{,}10/u_c + \cdots + \sum_{u_d=74}^{74} 1{,}00/u_d. \quad (5)$$

The transition point of the membership function *x(ud) = 0.5*.

Any continuous function of m variables on the unit interval [0 1] can be expressed as the sum of a finite number of one-dimensional functions:

$$f(x_1, x_2, \ldots, x_n) = \sum_{p=1}^{2m+1} g\left(\sum_{i=1}^{m} \lambda_i \varphi_p(x_i)\right), \quad (6)$$

where the functions g and φ_p are one-dimensional and continuous, *i = const* for all *i*. It follows that any continuous function can be approximated by a three-layer neural network with m input neurons, *2m + 1* hidden neuron, and 1 output neuron. This result is extended to multilayer networks using the backpropagation algorithm.

Thus, the final neural network contains three layers. As the number of neurons in the hidden layer increases, on the one hand, the accuracy of the artificial neural network increases, but on the other hand, if the scale of the hidden layer is too large, it will cause the network to be overloaded and result in the network being too large. Accuracy is also degraded. The generalizing ability of ANNs. Therefore, the number of neurons in the network should be minimized.

According to the proposal to determine the number of neurons in a neural network based on the number of training pairs, it is recommended to use the following formula:

$$2(m_1 + m_2 + m_3) < L < 10(m_1 + m_2 + m_3), \quad (7)$$

where m1 is the number of input layer neurons, m2 is the number of hidden layer neurons,

m3 is the number of output layer neurons, and L is the number of training pairs.

Taking into account the degree of use of neurons in the layer can be expressed as:

$$6m_1 + 4 < L < 30m_1 + 20, \quad (8)$$

when training a neural network, use samples drawn from a distribution close to the true one. The distribution used has a ratio of invalid to valid data of approximately 1:99.

So, the neural network contains 1001 neurons, 333 of which are in the input layer, 667 in the hidden layer, and 1 in the output layer.

The initial configuration is a three-layer fully connected homogeneous feedback-free perceptron with four inputs, a thousand neurons, and a hyperbolic tangent as the activation function. The first layer contains 333 neurons, 667 hidden layers, and 1 output layer.

When forming the training set, validation set, and test samples:

1. Use the following principles:

- principles of sequential experiments;
- standardization of factors;
- validation and testing samples should be drawn from the same data distribution - approximately 1% unreliable PDs and 99% reliable PD`s.

2. Make the following assumptions:

- the frequency of errors during the classification of training, validation, and test samples (errors in marked examples before neural network training) is 1%;
- due to the large size of the training set, it seems more appropriate to split the test set into eye sample PR and black box selection PR;
- the execution time of the algorithm will never exceed the maximum allowable value.

According to the recommendations, the number of training pairs L is determined by the following formula:

$$2(m_1 + m_2 + m_3) < L < 10(m_1 + m_2 + m_3), \quad (9)$$

where m1, m2, m3 are the number of neurons in the layer. Therefore, the number of training pairs L should take values in the range [2002;10010].

Use the 10,000 training pairs as training samples to express your own set of features and create images that will be fed to the neural network input.

Since the range of values of the hyperbolic tangent is limited, the training set is rescaled to the appropriate range of values.

The neural network is trained on the training set until a given mean squared error is reached.

Validation and test samples include 3000 pairs. To enable rapid manual estimation of classification error and to avoid overfitting, the samples were split into eyeball samples and black box samples consisting of 500 and 2500 pairs, respectively. Assuming a mislabeled sample rate of 1%, a sample of 500 pairs of eyeballs will contain approximately 5 mislabeled

samples. It turns out that the number of unclassified examples is insufficient for error analysis.

However, to avoid overtraining the network, it was decided not to include all 3000 test samples in the eye samples, but to prioritize the selected samples into a black box, which will generate up to 5 new eyes 500 per sample.

If there are large errors, it is concluded that the neural network is underequipped and additional training is performed. If the deviation is small, but the variation is large, it can be concluded that the neural network is overloaded. The neural network is trained until the bias and distribution reach a certain target value.

To determine the quality of neural networks, it is recommended to use multi-parameter PR metrics, including:

Satisfaction indicators:

- the average correspondence between accuracy and completeness is not less than 0.6;
- dispersion - no more than 0.5%;
- the value of the PR shift of the optimization indicator should not exceed 1%.

During the training process, the possible output characteristics of training and testing include two categories: valid data input and invalid data input. At the initial stage, the class consists of the following pairs:

Training set - 4960 pairs of reliable data and 40 pairs of unreliable data;

There are 3970 pairs of valid data and 30 pairs of invalid data in the control sample and the test sample.

The results displayed by the neural network at the initial stage of training are shown in Table 3:

1. The reliability of the neural network is equal to:

$$Accuracy = \frac{12 + 1818}{12 + 1152 + 18 + 1818} \sim 0{,}61, \tag{10}$$

2. The accuracy coefficient of the neural network is equal to:

$$Precision = \frac{12}{12 + 1152} \sim 0{,}01, \tag{11}$$

3. The integrity of the neural network is equal to:

$$Recall = \frac{12}{12 + 18} = 0{,}4, \tag{12}$$

4. F1-measure of the neural network:

$$F1score = 2\frac{0{,}01 \times 0{,}4}{0{,}01 + 0{,}4} \sim 0{,}02, \quad (13)$$

5. Deviation of the training sample - 34%, deviation of the test sample - 39%.
6. The spread is 5%.

Table 3
Results of the initial stage of neural network training

Parameter	Meaning
Learning results on the training set	
True Positive (TP)	16
False Positive (FP)	1336
False Negative (FN)	24
True Negative (TN)	2624
Results of testing on the validation set	
True Positive (TP)	12
False Positive (FP)	1152
False Negative (FN)	18
True Negative (TN)	1818

Since a 34% bias in the neural network results was found in the early stages of PR training, it was decided to increase the size of the neural network by adding neurons to the input layer, according to the proposal. that's why:

- the size of the neural network increased from 1001 neurons to 1004 neurons: 334 neurons in the input layer, 669 neurons in the hidden layer, and 1 neuron in the output layer;
- the number of training, validation, and test sample pairs does not change, as the number of training pairs falls into a new range [2008;10040].

The results after completing the training of the extended neural network are shown in Table 4. So, after completing the training on the test sample:

1. The reliability of the neural network is equal to:

$$Accuracy = \frac{25 + 2953}{25 + 17 + 5 + 2953} \sim 0{,}99, \quad (14)$$

2. The accuracy coefficient of the neural network is equal to:

$$Precision = \frac{25}{25 + 17} \sim 0{,}60, \quad (15)$$

3. The integrity of the neural network is equal to:

$$Recall = \frac{25}{25+5} \sim 0{,}83, \tag{16}$$

4. F1-measure of the neural network:

$$F1score = 2\frac{0{,}60 \times 0{,}83}{0{,}60 + 0{,}83} \sim 0{,}69, \tag{17}$$

Table 4
Results of the neural network after training

Parameter	Meaning
Learning results on the training set	
True Positive (TP)	35
False Positive (FP)	19
False Negative (FN)	5
True Negative (TN)	3941
Results of testing on the validation set	
True Positive (TP)	25
False Positive (FP)	17
False Negative (FN)	5
True Negative (TN)	2953
Results of testing on a test sample	
True Positive (TP)	23
False Positive (FP)	21
False Negative (FN)	7
True Negative (TN)	2949

The deviation of the training set is 0.6%, and the deviation of the test set is about 0.7%. The difference (deviation between training and test samples) is about 0.13%. On the test sample:

1. The reliability of the neural network is equal to:

$$Accuracy = \frac{23+2949}{23+21+7+2949} \sim 0{,}99, \tag{18}$$

2. The accuracy coefficient of the neural network is equal to:

$$Precision = \frac{23}{23+21} \sim 0{,}52, \tag{19}$$

3. The integrity of the neural network is equal to:

$$Recall = \frac{23}{23+7} \sim 0{,}77, \tag{20}$$

4. F1-measure of the neural network:

$$F1score = 2\frac{0{,}52 \times 0{,}77}{0{,}52 + 0{,}77} \sim 0{,}62. \quad (21)$$

5. The deviation is about 0.93%.
6. The deviation between the control sample and the test sample is 0.2%.

The final configuration is a three-layer homogeneous open-loop perceptron with 4 inputs, 1004 neurons, and a hyperbolic tangent as the activation function. The first layer contains 334 neurons, the hidden layer - 669, the output layer - 1.

4. Discussing

The topic of encryption is beyond the scope of this study. But since PR blockchain systems are by definition based on the use of cryptographic methods, it seems appropriate to provide general advice on the choice of methods for calculating hash functions.

Since a private blockchain is chosen as the basis for the blockchain system, methods based on symmetric encryption rather than asymmetric encryption as in public blockchain systems should be used as a method of cryptographic protection for blocks and transactions.

Given the urgency of the task of creating a post-quantum cryptosystem, it is necessary to foresee the possibility of making changes to the process of calculating hash functions in the developed scratchpad architecture.

DRPD access and PD storage are expected to use encryption to protect information. Even if the key is broken, the confidentiality of the protected data can be ensured by ensuring the confidentiality of the carrier signal energy and the confidentiality of these structures. Signals reliability of detection, complexity of the signal structure. However, this trade-off seems reasonable given the need to consider data transmitted over the Encrypted PD protocol.

Data paths in DRPD can be divided into two categories: non-overlapping paths and overlapping paths. DRPD is an on-call system.

The proposed method of calculating the probability of data leakage assumes the following:

- the sending node and the receiving party are protected, that is, the possibility of being attacked by hackers is zero;
- if one segment of a path (a node, a data link, or a combination thereof) is compromised, all data traveling along that segment will also be compromised.

Suppose that the following initial information is known:

- *pji* is the probability of damage to the j-th segment of the i-th path;
- *Qi* is the number of damaged segments on the ith path.

Then the probability that the i-th trajectory consisting of airborne debris will be destroyed can be calculated using the following formula:

$$p_i = (1 - p_{i1})(1 - p_{i2}) \dots \left(1 - p_{iM_i}\right) = 1 - \prod_{j=1}^{Q_i}(1 - p_{ij}), \tag{22}$$

when the data is divided into *N* parts *(N, N)* according to the Shamir scheme and transmitted through Q paths, the probability of data leakage is determined by the following expression:

$$P_{msg} = \prod_{i=1}^{Q} p_i, \tag{23}$$

where Q is the number of non-intersecting paths used to route data elements – the probability that the *i*-th path segment is compromised.

In the case of using intersecting routes with a series of connecting line segments and parallel structures, the probability of data theft is calculated according to the following formula:

$$P_{msg} = 1 - \prod_{j=1}^{\widetilde{N}}(1 - \tilde{p}_j), \tag{24}$$

where $\widetilde{N}$ is the total number of sequence segments in the series-parallel structure of the considered intersecting paths; $\tilde{p}_j$ is the probability of destruction of the *j*-th segment.

To demonstrate the general principle of the proposed method for calculating the probability of data leakage, Figure 11 shows a rather simplified structure of the DRPD, which consists of two connected paragraphs:

- The first segment includes a parallel connection of the communication channel 1→3 and the sequence of channels 1→2 and 2→3;
- The second segment is represented by the communication channel 3→4.

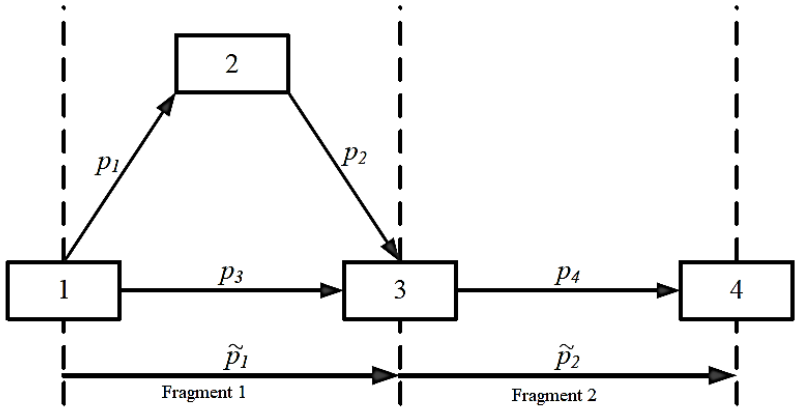

Figure 11: Examples of serial and parallel connection of components

The probability of leakage of the first and second points is determined by the probability of leakage of the communication channel that they form:

$$p_i = 1 - (1 - \widetilde{p_1})(1 - \widetilde{p_2}), \quad (25)$$

The probability of data leakage is calculated as follows:

$$\widetilde{p_1} = (1 - (1 - p_1)(1 - p_2))p_3, \widetilde{p_2} = p_4 \quad (26)$$

Among them, ə is the total number of parallel segments; ə is the probability of destruction of the *j*-th segment.

The most common scenario for decentralized systems based on blockchain technology seems to be the use of intersecting paths with complex structures that allow network segments to be connected in series and parallel. For clarity, Figure 12 shows a general example of such a structure consisting of seven paragraphs:

- Fragments 1, 2, and 3 are connected in series, forming fragment 4;
- Segment 5 and Segment 6 are connected in series to form Segment 7;
- Fragments 4 and 7 are connected in parallel.

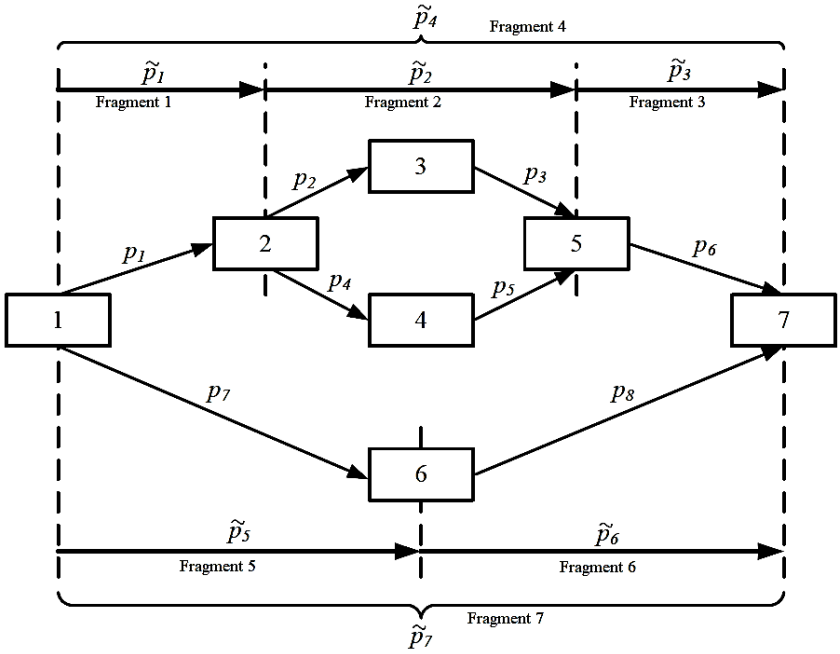

Figure 12: Example of a combination of elements

The probability of data leakage will be determined by the following formula:

$$P_{msg} = \widetilde{p_4}\widetilde{p_7}, \quad (26)$$

The probability of damage to segments 4 and 7 is represented by the probability of damage to the corresponding communication line:

$$\widetilde{p_4} = (1 - (1 - \widetilde{p_1})\big(1 - \widetilde{p_2})(1 - \widetilde{p_3})\big); \quad (27)$$
$$\widetilde{p_7} = 1 - (1 - \widetilde{p_5})(1 - \widetilde{p_6}) = 1 - (1 - p_7)(1 - p_8).$$

A method of ensuring the reliability of personal data processed in blockchain systems is

proposed. The approach includes recommendations for creating a common architecture for decentralized ledgers, an agreed PD for data storage, methods for reaching consensus, a common agreed PD for system implementation and development, and calculating the probability of data theft.

5. Conclusion

The category of data reliability methods is expanding due to the use of artificial neural networks to identify unreliable personal data when entered into blockchain systems.

The reliability of personal data processing in blockchain systems can be ensured using the proposed method:

- WIPO single-level cloud platform or at the country level;
- as part of ensuring compliance with the requirements for monitoring incorrect user actions when entering personal data.

This method differs from known methods by the unique architecture of the information system of personal data. This method differs from known methods in that it uses a conceptually new consensus approach that involves an automated assessment of the risks of implementing unreliable material handling. The theory of artificial neural networks and the theory of fuzzy sets.

Thus, the task proposed in the article has been solved to develop a method for ensuring the reliability of personal data processed in the blockchain system, and when the data enters the blockchain system, its reliability will be automatically evaluated.

References

[1] Bernal Bernabe, Jorge & Canovas Sanchez, Jose Luis & Hernández-Ramos, José & Torres Moreno, Rafael & Skarmeta, Antonio. (2019). Privacy-Preserving Solutions for Blockchain: Review and Challenges. IEEE Access. PP. 10.1109/ACCESS.2019.2950872.

[2] Y. Lu et al., "Accelerating at the Edge: A Storage-Elastic Blockchain for Latency-Sensitive Vehicular Edge Computing," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 11862-11876, Aug. 2022, doi: 10.1109/TITS.2021.3108052.

[3] I. Sharma, K. Kaushik and G. Chhabra, "Augmenting Transparency and Reliability for National Health Insurance Scheme with Distributed Ledger," 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1399-1405, doi: 10.1109/ICESC57686.2023.10193127.

[4] I. A. Omar, R. Jayaraman, K. Salah, H. R. Hasan, J. Antony, and M. Omar, "Blockchain-Based Approach for Crop Index Insurance in Agricultural Supply Chain," in IEEE Access, vol. 11, pp. 118660-118675, 2023, doi: 10.1109/ACCESS.2023.3327286.

[5] Y. Gao, H. Lin, Y. Chen and Y. Liu, "Blockchain and SGX-Enabled Edge-Computing-Empowered Secure IoMT Data Analysis," in IEEE Internet of Things Journal, vol. 8, no. 21, pp. 15785-15795, 1 Nov.1, 2021, doi: 10.1109/JIOT.2021.3052604.

[6] Daraghmi, Eman & Helou, Mamoun & Daraghmi, Yousef-Awwad. (2021). A Blockchain-Based Editorial Management System. Security and Communication Networks. 2021. 17.

10.1155/2021/9927640.
[7] Daraghmi, Eman & Daraghmi, Yousef-Awwad & Yuan, Shyan-Ming. (2019). UniChain: A Design of Blockchain-Based System for Electronic Academic Records Access and Permissions Management. Applied Sciences. 9. 10.3390/app9224966.
[8] T. Yang, Z. Cui, A. H. Alshehri, M. Wang, K. Gao, and K. Yu, "Distributed Maritime Transport Communication System With Reliability and Safety Based on Blockchain and Edge Computing," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 2, pp. 2296-2306, Feb. 2023, doi: 10.1109/TITS.2022.3157858.
[9] N. Kshetri, "Blockchain-Based Smart Contracts to Provide Crop Insurance for Smallholder Farmers in Developing Countries," in IT Professional, vol. 23, no. 6, pp. 58-61, 1 Nov.-Dec. 2021, doi 10.1109/MITP.2021.3123416.
[10] Belej O., Więckowski*** T., Staniec*** K. The need to use a hash function to build a crypto algorithm for blockchain // Advances in Intelligent Systems and Computing (AISC). – 2020. – Vol. 1173 : Theory and applications of dependable computer systems. Proceedings of the Fifteenth international conference on dependability of computer systems DepCoS-RELCOMEX, June 29 – July 3, 2020, Brunów, Poland. – P. 51–60.
[11] Salem, Yaman & Daraghmi, Eman. (2021). GDPR-BLOCKCHAIN COMPLIANCE FOR PERSONAL DATA: REVIEW PAPER. Journal of Theoretical and Applied Information Technology. 99.
[12] Martins Gonçalves, R.; Mira da Silva, M.; Rupino da Cunha, P. Implementing GDPR-Compliant Surveys Using Blockchain. Future Internet 2023, 15, 143. https://doi.org/10.3390/fi15040143
[13] G. Lodha, M. Pillai, A. Solanki, S. Sahasrabudhe and A. Jarali, "Healthcare System Using Blockchain," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 274-281, doi: 10.1109/ICICCS51141.2021.9432157.