Modeling of Pseudo-Random Sequences Generated by Data Encryption and Compression Algorithms

Alexander V. Kozachok [1], Vasiliy I. Kozachok, Andrey A. Spirin[1], Andrey K. Trofimenkov[2]

[1] *Academy of the Federal Guard Service of the Russian Federation, 35 Priborostroitelnaya ul., Orel, 302015, Russia*
[2] *Orel Branch of the Federal Research Center "Informatics and Management" of the Russian Academy of Sciences (OF FIC IU RAS), 137 Moskovskoe shosse., Orel, 302025, Russia*

Abstract

Reports of information and analytical agencies indicate a high proportion of internal violators as sources of confidential data leaks in Russia. One of the possible channels of data leaks can be their transmission in encrypted form. Modern data analysis tools are not able to reliably detect the transfer of information after its processing by cryptographic algorithms. In addition, an attacker can embed digital signatures specific to compressed data in encrypted data, thereby disguising them as legitimate file types. The paper presents an approach to the classification of encrypted and compressed data based on the developed model of pseudorandom sequences and the algorithm for their classification. The accuracy of the proposed method was 0.97.

Keywords

Statistical data analysis, machine learning, classification of encrypted and compressed data, binary sequence classification

1. Introduction

According to the report of the expert and analytical center of the Infowatch group of companies, the share of internal violators as sources of confidential data leaks in Russia increased in 2020 [1], in more than 79% of cases, confidential data leaks were caused by an internal violator.

To protect information from leaks, software tools for detecting and preventing leaks of confidential data and various systems for deep traffic analysis are used.

Methods of traffic analysis, depending on the feature spaces used in its classification, can be divided into several groups: calculation of the entropy of all or part of the data [2-6]; service information of data transmission protocols [7–10], statistical characteristics and byte distribution [11–14]. However, there are ways to circumvent these security methods, such as using encryption, data compression, or encapsulation in other protocols [15,16]. In papers [17, 18] the authors conclude that the entropy approach reduces the information about the distribution to a single number, thereby reducing the features available for analysis. To overcome this problem, deep neural networks were used to classify data of two classes: encrypted by the aes algorithm and compressed by the rar, zip/gzip, jpeg, png, mp3, and pdf algorithms. This mechanism, according to the authors, will automatically determine the most significant features inherent in the analyzed sequences and improve the accuracy of their classification.

In [19], a combined approach based on recurrent networks and extremely randomized trees is used to detect potentially malicious actions and intrusions into information systems. Service network information was used as attributes.

In the paper [20] notes that the byte distributions in encrypted and compressed sequences tend to be evenly distributed. This fact is explained by the fact that encryption algorithms disperse the original statistics in messages and compression algorithms tend to reduce the redundancy of the original

BIT-2021: XI International Scientific and Technical Conference on Secure Information Technologies, April 6-7, 2021, Moscow, Russia
EMAIL: a.kozachok@academ.msk.rsnet.ru (A. 1); kosachok@list.ru(A. 2); spirin_aa@bk.ru (A. 3), andjeytrf@rambler.ru (A. 4)
ORCID: 0000-0002-6501-2008 (A. 1); 0000-0001-5384-2269 (A. 2); 0000-0002-7231-5728 (A. 3); 0000-0002-5576-875X (A. 4)

messages. The article [21] describes the use of a feature space based on byte allocation for detecting potentially malicious software.

In the reviewed studies, the following machine learning algorithms were used: decision tree (DT), support vector machine (SVM), random forest (RF), Markov chain (MC), hidden Markov chain (HMC), boosting (BG), convolutional neural networks (CNN), recurrent neural networks (RNN), determination of autocorrelation of distributions (AC). The considered methods use container and file headers for data analysis, which contain "magic" bytes – digital signatures that uniquely identify the transmission protocol or the compressed data container. At the same time, software and hardware protection against information leaks do not have mechanisms for analyzing encrypted or compressed data, in the absence of information about the compression algorithm [3]. An overview of the methods considered is presented in Table 1.

Table 1
Analysis of the subject area of research

The authors	Features	Algorithms	Accuracy
Mamun M.S.I. [2]	Entropy	DT	0,981
Tang Z. [3]	Entropy	SVM, RF	0,979
Shen M. [7]	service traffic information	MC	0,912
Chen Y. [8]	service traffic information	BG	0,987
Obasi T.C. [9]	service traffic information	CNN, RF, DT	0,96
Yao Z. [10]	service traffic information	HMC	0,99
Choudhury P. [11]	Bytes distribution	AC	0,99
Baldini G. [13]	Bytes distribution	SVM	0,8
Shen M. [14]	Bytes distribution	RF	0,882
De Gaspari F. [17]	Auto extraction	DNN	0,85-0,99
Kasongo, S. M. [19]	service traffic information	Extra trees, RNN	0,99

Machine learning algorithms are also used in related areas of information security. So in the study, the authors used the XGBoost algorithm to search for abnormal behavior of players on the stock exchange who own insider information and use it improperly [22]. The signs were various specific signs and financial indicators inherent in conducting transactions on the exchange. Despite some differences in the subject area, the authors solve a similar problem of optimizing machine learning algorithms to improve the accuracy of classifying illegitimate actions of agents.

In a number of studies, an algorithm of extremely randomized trees is used to detect intrusions into corporate networks. This approach also applies to the ensemble method as well as the random forest, but differs in the way of selecting the partition threshold. Instead of searching for the most optimal thresholds, as happens in the random forest algorithm, the thresholds are selected randomly for each possible feature, and the best one is selected as a rule for dividing the node. The use of extremely randomized trees slightly reduces the variance of the model due to a slightly larger increase in the bias [23-25].

Gradient boosting refers to ensemble machine learning algorithms that can be used for classification or regression tasks. Ensembles are built on the basis of decision tree models. Trees are added one by one to the ensemble and trained to correct prediction/classification errors made by previous models. The models are trained using any arbitrary differentiable loss function and a gradient descent optimization algorithm. This explains the name of the algorithm "gradient boosting", since the loss gradient is minimized as the model is trained, like a neural network, and the addition of trees causes the term boosting [23].

Special attention in Work 21 is paid to the definition of hyperparameters of the classifier. The maximum depth of the trees limits the growth of the trees used in depth. A higher value increases the complexity of the model, but at the same time can lead to overfitting. The learning rate controls the weight of an individual model in the ensemble prediction. Smaller speed values may require more decision trees in the ensemble.

In work [24], the authors search for hyperparameters of the classifier based on the lightgbm algorithm to increase its accuracy, the main ones are: maximum depth of trees – this parameter determines the complexity of the model, the higher it is, the more accurately classifies the data, but can lead to overfitting and a decrease in accuracy on test data; learning rate – the parameter assigns the weight of each tree in the ensemble, lower values indicate a low weight of each tree in the ensemble. The special role of the feature selection and normalization process is also noted, since erroneous data and omissions in the feature values can significantly reduce the accuracy of the classification algorithms used. The statistical characteristics of the obtained distributions of the values of the feature space were used as features: the minimum and maximum values, the arithmetic mean, and the variance.

The analysis of the works considered in the study allows us to conclude that machine learning algorithms are widely used in the field of information security and their high accuracy in solving problems of data classification of various classes. In many works, the entropy approach, the distribution of bytes and subsequences of different lengths, is used to form the feature space. Based on the analysis, it was suggested that it is possible to form a feature space based on byte distributions and subsequences of limited bit length.

2. Pseudo-random sequences model

The analysis of sources in the subject area of research revealed the frequent use of the byte distribution as a feature space, its statistical characteristics, and the results of counting the frequencies of subsequences of different lengths with different steps.

It was suggested that the feature space based on combining the frequency of occurrence of independent bit subsequences of different length N (in bits) without taking into account the complete overlap of each subsequence and the byte distribution can be used to solve the problem of improving the accuracy of the classification of PRS and improving existing solutions in the field of information protection from leaks. For example, for sequence = 100101111010, the frequency of occurrence of subsequences with length N = 3 bits is shown in Table 2.

Table 2
Example of bit subsequences frequency count

Subsequences	Number	Frequency
000	0	0
001	1	0,1
010	2	0,2
011	1	0,1
100	1	0,1
101	2	0,2
110	1	0,1
111	1	0,1

In order to prevent the use of digital signatures and correctly classify encrypted and compressed sequences that have a uniform byte distribution and are called pseudorandom (PSP), a PSP model based on statistical features was developed. When it is formed, the header part of the file with a length of 10 KB is discarded, the rest is subject to statistical analysis. In its formal form, the PSP model is defined by the expression 1.

$$V = \{F(B_{i\in[0,\dots,255]}), B_{mean}, B_{sko}, B_{min}, B_{max}, F(S_{j\in[0,\dots,511]})\}, \tag{1}$$

where $B_{i\in[0,\dots,255]}$ – byte distribution in the analyzed PRS; $S_{j\in[0,\dots,511]}$ – distribution of frequencies of occurrence of subsequences of length 9 bits in the analyzed PRS; B_{mean} – mean of the byte frequency in PRS; B_{sko} – standard deviation of the byte distribution; B_{min}, B_{max} – minimum and maximum byte frequency values.

Mean of the byte frequency in PRS and standard deviation of the byte distribution is calculated according to expression 2.

$$B_{mean} = \frac{\sum_{i=0}^{255} n(b_i)}{256}, B_{sko} = \sqrt{\frac{\sum_{i=0}^{255}(n(b_i) - B_{mean})^2}{256}} \quad (2)$$

where $n(b_i)$ – the number of occurrences of byte i in the analyzed PRS.

The frequency of occurrence of subsequences in the PRS is calculated according to expression 3.

$$F(S_{j\in[0,\dots,511]}) = f_j = \frac{n_j}{M - N + 1}, \quad (3)$$

where n_j – the number of occurrences of the subsequence j in PRS; M – PRS length in bits; N – length of the subsequence j in bits.

To test the adequacy of the model, experiments were conducted to classify the generated set of encrypted and compressed data with a size of 600 Kbytes. 2000 files containing meaningful text in Russian were generated, then they were processed by encryption algorithms (AES, DES, RC4, Camellia, GOST34.12 "Kuznechik") and compression algorithms (RAR, ZIP, 7Z, GZ, BZ2, XZ). Thus a data sample consisting of two classes and containing 22,000 files was obtained.

3. Practical application of the developed PRS model

In order to select the most suitable machine learning algorithm, experiments were conducted to evaluate the accuracy of the PRS classification. In view of the dependence of the classification accuracy on the number of features used in the PRS model, experiments were conducted using a different number of features ranked by discriminating ability [26].

The results of evaluating the accuracy of the classification of PRS by machine learning algorithms depending on the number of features are shown in the figure 1.

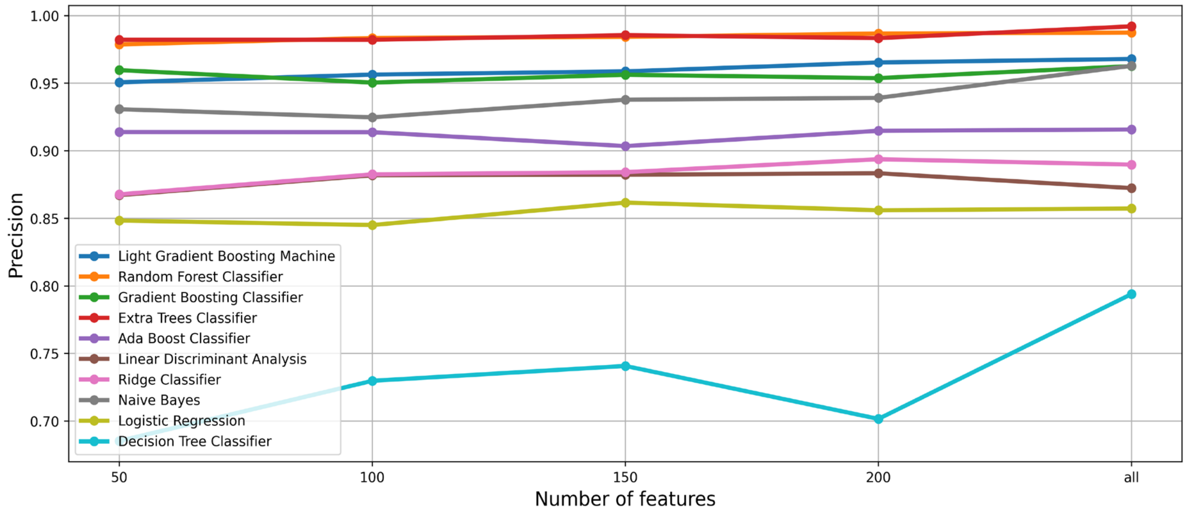

Figure 1: Evaluation of the accuracy of the classification of PRS depending on the number of used features of the modeled PRS.

Almost all machine learning algorithms demonstrate an increase in classification accuracy with an increase in the dimension of the feature space, which meets the expected results, since the features are ranked according to their classifying ability. The most accurate classification according to the precision metric was shown by the random forest algorithm and the extra trees algorithm.

Machine learning algorithms have different time complexity, and to assess their applicability in real systems, experiments were conducted to determine the training time of a model based on the corresponding machine learning algorithm. The results of the experiments are shown in Figure 2.

The proposed PRS model based on the developed PRS classification algorithm allows classifying encrypted and compressed data with an accuracy of 0.97 in 0.5 seconds for a sequence of 600 Kbytes in length. The classification does not take into account digital signatures, file extensions, and other service information, but only statistical features of byte distributions and 9-bit subsequences [27].

The developed solution can be implemented in existing DLP systems and perform statistical analysis of data transmitted beyond the controlled perimeter of the corporate data network, the flow diagram is shown in Figure 2.

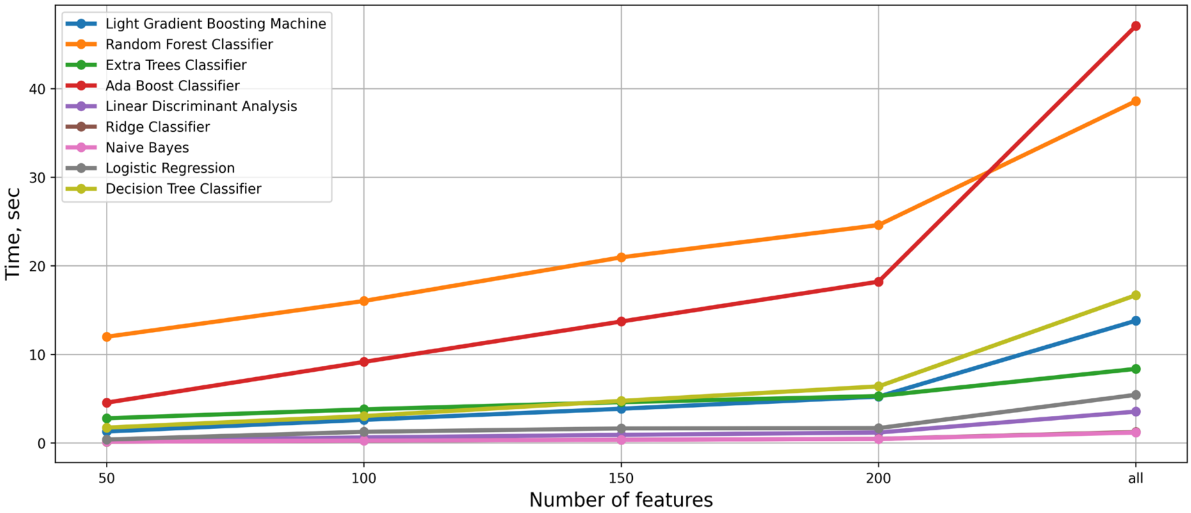

Figure 2: Estimating the time complexity of building machine learning models.

The gradient boosting algorithm was excluded from the experiment because it has the longest training time of 250 seconds on a given sample size. The classifier based on adaptive boosting has the maximum time to build the model. A classifier based on a random forest and extremely randomized trees have similar time complexity on this task.

Since the classifier based on a random forest is superior to the classifier based on extremely randomized trees with the most optimal values of the number of parameters equal to 200 features, it was chosen to build a prototype for the classification of encrypted and compressed data in information leak prevention systems. The first 10 features with the maximum discriminating ability are shown in Figure 3.

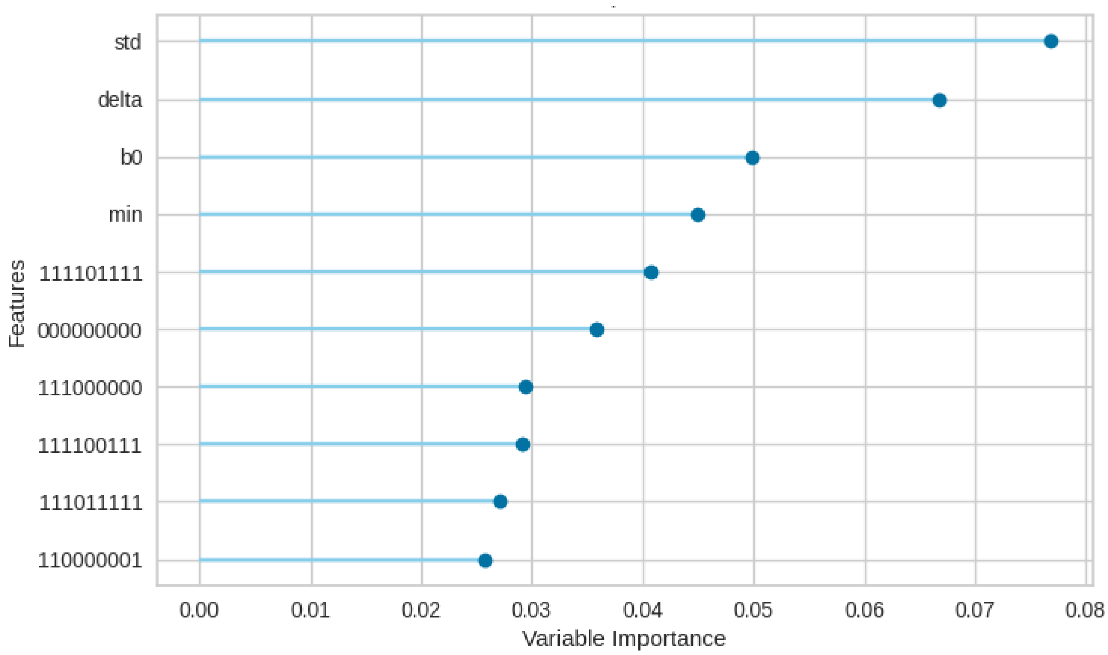

Figure 3: First 10 features with maximal discriminative power.

Thus, the feature space is a vector of statistical properties with a length of 110 elements that have the maximum discriminating ability. Figure 4 shows the result of a multidimensional data visualization algorithm for the 5 most discriminating features. Statistical features are located around the circle, the points are the values of the features for each class of PSP, located independently of each other. The points inside the circle are arranged so that its values are normalized along the axes from the center to each arc.

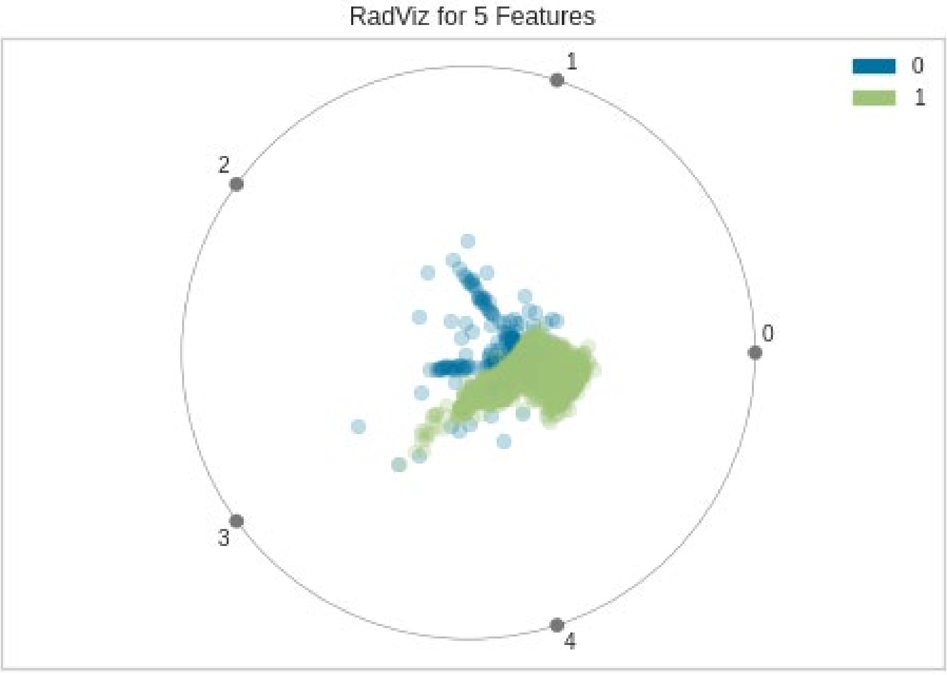

Figure 4. The top 10 most discriminating features.

The total values of the hyperparameters of the generated PRS classifier are presented in the table 3.

Table 3

Hyperparameters of the generated PRS classifier

Hyperparameter	Value

Number of estimators	100
Number of features	200
Max depth of the trees	38
Length of the subsequences	9 bits
100	1
101	2
110	1
111	1

The developed PRS model together with the PRS classification algorithm can be implemented in the statistical data analysis module of existing DLP systems. The greatest threat to modern enterprise systems is cloud storage. If necessary, the internal intruder is able to use encryption and data masking tools, for example, by introducing digital signatures of known file formats: for ZIP format - "50 4B"; RAR - "52 61 72 21 1A"; pdf - "25 50 44 46 2D 31 2E". The developed solution can be implemented in existing DLP systems and perform statistical analysis of data transmitted outside the controlled perimeter of the corporate data network, the block diagram of which is shown in figure 5.

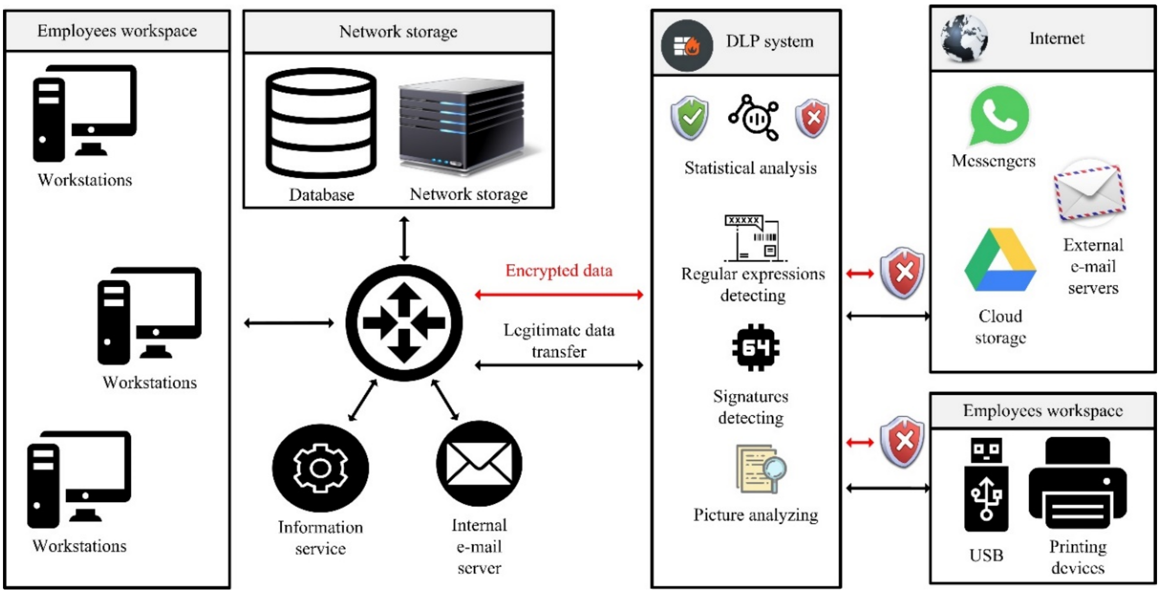

Figure 5: The pipeline of implementation of the developed approach of classification of PRS in the existing DLP systems.

4. Conclusion

The study examines existing approaches to the formation of feature spaces and machine learning algorithms used to build classifiers of encrypted and compressed data. Existing approaches show high classification accuracy, but all of them use the headers of the analyzed files or data packets, which contain digital signatures that uniquely determine the type of information transmitted. An attacker can take advantage of this flaw and transmit the information in encrypted form by changing the header part of the file or using encapsulation methods for traffic.

To improve the accuracy of the classification of PRS generated by data encryption and compression algorithms, a model of PRS was developed that uses the byte distribution and the frequency of occurrence of 9-bit subsequences in the analyzed PRS. The choice of the mathematical apparatus based on the random forest construction algorithm was justified. During the experiments, the adequacy of the model and the possibility of its application for the classification of the data types specified in the work with a precision of 0.98 were confirmed. The place of implementation of the developed PRS model in the means of detecting and preventing information leaks was proposed.

5. Acknowledgements

The reported study was funded by Russian Ministry of Science (information security), project number 18/2020.

6. References

[1] InfoWatch analytics report, 2021. URL: https://www.infowatch.ru/analytics/reports/30708.

[2] Mamun M.S.I., Ghorbani A.A., Stakhanova N. An Entropy Based Encrypted Traffic Classifier. // In: Qing S., Okamoto E., Kim K., Liu D. (eds) Information and Communications Security. ICICS 2015. Lecture Notes in Computer Science, vol. 9543. Springer, Cham. DOI: 10.1007/978-3-319-29814-6_23.

[3] Tang Z., Zeng X., Sheng Y. Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification. International Journal of ICIC, 2019, vol. 15, no. 3, pp. 845–860. DOI: 10.24507/ijicic.15.03.845.

[4] Belyaev S. and etc. Development of a Pseudo-Random Sequence Generation Function Based on the "Kuznechik" Cryptographic Algorithm. Voprosy kiberbezopasnosti [Cybersecurity issues], 2021, No 4 (44), pp. 25-34. DOI: 10.21681/2311-3456-2021-4-25-34. DOI: 10.21681/2311-3456-2017-5-30-41. (In Russ.)

[5] Livshitz I., Neklydov A. Assessment of Entropy of Information Security Systems. Voprosy kiberbezopasnosti [Cybersecurity issues], 2017. No 5(24), pp. 30-41. DOI: 10.21681/2311-3456-2017-5-30-41. (In Russ.)

[6] Zhou, K., Wang, W., Wu, C., & Hu, T. (2020). Practical evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks. *ETRI Journal*, *42*(3), 311-323. DOI: 10.4218/etrij.2019-0190.

[7] Shen M., Wei M., Zhu L. & Wang M. Classification of encrypted traffic with second-order markov chains and application attribute bigrams // IEEE Transactions on Information Forensics and Security, 2017, vol. 12, no. 8, pp. 1830-1843. DOI: 10.1109/TIFS.2017.2692682.

[8] Chen Y., Zang T., Zhang Y., Zhouz Y., & Wang Y. Rethinking Encrypted Traffic Classification: A Multi-Attribute Associated Fingerprint Approach. // 2019 IEEE 27th International Conference on Network Protocols (ICNP), Chicago, IL, USA, 2019, pp. 1–11. DOI: 10.1109/ICNP.2019.8888043.

[9] Obasi T.C. Encrypted Network Traffic Classification using Ensemble Learning Techniques // Doctoral dissertation, Carleton University, 2020. DOI: 10.22215/etd/2020-14171.

[10] Yao Z., Ge J., Wu Y., Lin X., He R., Ma Y. Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models. // Journal of Network and Computer Applications, 2020, vol. 166, p. 102711. DOI: 10.1016/j.jnca.2020.102711.

[11] Wang F., Quach T.T., Wheeler J., Aimone J.B., James, C.D. Sparse coding for n-gram feature extraction and training for file fragment classification. // IEEE Transactions on Information Forensics and Security, 2018, vol. 13, no.10, pp. 2553–2562. DOI: 10.1109/TIFS.2018.2823697.

[12] Choudhury P., Kumar K. P., Nandi S., Athithan G. An empirical approach towards characterization of encrypted and unencrypted VoIP traffic // Multimedia Tools and Applications, 2020, vol. 79, no. 1–2, pp. 603–631. DOI: 10.1007/s11042-019-08088-w.

[13] Baldini G., Hernandez-Ramos J. L., Nowak S., Neisse R., Nowak M. Mitigation of Privacy Threats due to Encrypted Traffic Analysis through a Policy-Based Framework and MUD Profiles. // Symmetry, 2020, vol. 12, no.9, p. 1576. DOI: 10.3390/sym12091576.

[14] Shen M., Liu Y., Zhu L., Xu K., Du X., Guizani N. Optimizing Feature Selection for Efficient Encrypted Traffic Classification: A Systematic Approach // IEEE Network, 2020, vol. 34, no. 4, pp. 20–27. DOI: 10.1109/MNET.011.1900366.

[15] Huang X. et al. A novel mechanism for fast detection of transformed data leakage //IEEE Access, 2018, vol. 6, pp. 35926–35936. DOI: 10.1109/ACCESS.2018.2851228.

[16] Cheng L., Liu F., Yao D. D. Enterprise data breach: causes, challenges, prevention, and future directions //Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, vol. 7, no. 5. DOI: 10.1002/widm.1211.

[17] De Gaspari F., Hitaj D., Pagnotta G., De Carli L., Mancini L. V. EnCoD: Distinguishing Compressed and Encrypted File Fragments. In *International Conference on Network and System Security* (pp. 42-62). 2020. Springer, Cham. DOI: 10.1007/978-3-030-65745-1_3.

[18] Lee K., Lee S. Y., Yim K. Machine learning based file entropy analysis for ransomware detection in backup systems. *IEEE Access*, *7*, pp. 110205-110215. 2019. DOI: 10.1109/ACCESS.2019.2931136.

[19] Kasongo S. M., Sun Y. A deep gated recurrent unit based model for wireless intrusion detection system. *ICT Express*, *7*(1), 81-87. 2021. DOI: 10.1016/j.icte.2020.03.002.

[20] Raff E., Zak R., Cox R., Sylvester J., Yacci P., Ward R., Nicholas C. (2018). An investigation of byte n-gram features for malware classification. *Journal of Computer Virology and Hacking Techniques*, *14*(1), pp. 1-20. DOI: 10.1007/s11416-016-0283-1.

[21] De Gaspari, F., Hitaj, D., Pagnotta, G., De Carli, L., & Mancini, L. V. (2021). Reliable Detection of Compressed and Encrypted Data. *arXiv preprint arXiv:2103.17059*.

[22] Deng S., Wang C., Li J., Yu H., Tian H., Zhang Y., Yang T. Identification of Insider Trading Using Extreme Gradient Boosting and Multi-Objective Optimization. *Information*, *10*(12), 367. 2019. DOI: 10.3390/info10120367.

[23] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Liu T. Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154. 2017. DOI: 10.5555/3294996.3295074.

[24] Minastireanu E. A., Mesnita G. Light gbm machine learning algorithm to online click fraud detection. *J. Inform. Assur. Cybersecur*, *2019*. DOI: 10.5171/2019.263928.

[25] Ge D., Gu J., Chang S., Cai J. Credit Card Fraud Detection Using Lightgbm Model. In *2020 International Conference on E-Commerce and Internet Technology (ECIT)* (pp. 232-236). IEEE. 2020. DOI: 10.1109/ECIT50008.2020.00060.

[26] Kozachok A.V., Spirin A.A., Golembiovskaya O.M. Algoritm classificazii psevdosluchainuh posledovatelnostei na osnove postroenija sluchainogo lesa. Dokladj Tomskogo gosudarstvennogo universiteta sistem upravlenija I radioelektroniki. – 2020. – V. 23. – no. 3. – P. 55–60. DOI: 10.21293/1818-0442-2020-23-3-55-60. (In Russ.)

[27] Kozachok A.V., Spirin A.A. Algoritm classificazii psevdosluchainuh posledovatelnostei. Sistemnij analyz i informazionnie tehnologii. – 2020. – no.1, pp. 87 – 98. DOI: 10.17308/sait.2020.1/2595.