Network traffic analyzing algorithms on the basis of machine learning methods

R I Battalov[1], A V Nikonov[1], M M Gayanova[1], V V Berkholts[1] and R Ch Gayanov[2]

[1]Ufa State Aviation Technical University, K. Marks str., 12, Ufa, Russia, 4500082
[2]Higher School of Economics, Myasnitskaya str., 20, Moscow, Russia, 101000

e-mail: nikonovandrey1994@gmail.com, torina4@yandex.ru

Abstract. Traffic analysis systems are widely used in monitoring the network activity of users or a specific user and restricting client access to certain types of services (VPN, HTTPS) which makes content analysis impossible. Algorithms for classifying encrypted traffic and detecting VPN traffic are proposed. Three algorithms for constructing classifiers are considered - MLP, RFT and KNN. The proposed classifier demonstrates recognition accuracy on a test sample up to 80%. The MLP, RFT and KNN algorithms had almost identical performance in all experiments. It was also found that the proposed classifiers work better when the network traffic flows are generated using short values of the time parameter (timeout). The novelty lies in the development of network traffic analysis algorithms based on a neural network, differing in the method of selection, generation and selection of features, which allows to classify the existing traffic of protected connections of selected users according to a predetermined set of categories.

CEUR-WS.org/Vol-2416/paper56.pdf

1. Introduction

The term “deep package inspection” (DPI) [1] refers to the analysis of the network packet at the upper levels (application and presentation level) of the open systems interaction model (OSI) [2].

In addition to analyzing network packets [3] using standard patterns by certain standard patterns that can be used to unambiguously determine whether a package belongs to a specific application, for example, by the format of headers, port numbers, etc., the DPI system performs behavioural analysis of traffic. This allows to recognize applications that do not use known data headers and data structures for data exchange.

For identification, an analysis of the sequence of packets with the same characteristics is carried out. Analyzed characteristics are Source_IP: port - Destination_IP: port; packet size; frequency of opening new sessions per unit of time, etc. The analysis based on behavioral (heuristic) models corresponding to such applications.

The main component of the DPI solution [4] is the classification module. It is responsible for the classification of network flows. The classification can be performed with different accuracy depending on the purposes of the DPI application:

- the type of protocol or application (for example, Web, P2P, VoIP);
- a specific application-layer protocol (HTTP BitTorrent, SIP);

- applications using the protocol (Google Chrome, uTorrent, Skype).

Traffic analysis using traditional tools becomes impossible without selecting a key for streaming data with encryption (for example, TLS / SSL protocols). It takes a lot of resources to find the key. The relevance of hacking remains only at the governmental or military level [5].

In the [6] the classification of network encrypted traffic from Skype, Tor, PuTTY (SSHv2), CyberGhost (VPN) is discussed by application types for detecting security threats using such machine learning methods as the Naive Bayes, C4.5, AdaBoost and Random Forest algorithms. For the analysis, more than two million network packets from four applications that transmit encrypted traffic were collected: Skype, Tor, PuTTY (SSHv2), CyberGhost (VPN). Two different classification approaches were considered: the formation and analysis of flows for network packets whose IP addresses of the sender / recipient and the network protocol are the same, as well as the interception and analysis of each network packet [7]. When using each approach, the various attributes were identified and with the use of which the classification was made. Obtained results can be used to build traffic classifiers and intrusion detection systems, effectively processing the encrypted traffic used by various network applications.

In [8] it is shown that comparison of various algorithms for classifying network traffic is significantly difficult due to the lack of a generally accessible base of fully-fledged network routes on which it would be possible to make comparisons. One of the most actively developing areas at the moment is the use of various machine learning algorithms, graph and statistical analysis, because of their applicability to encrypted traffic (as opposed to the DPI approaches), whose share is growing rapidly. Another emerging focus is the development of combined approaches and classification systems. One of the reasons for development is an attempt to overcome the shortcomings of individual approaches (for example, low accuracy or processing speed) and use their advantages [9, 10].

Therefore, the development of algorithms that allow classifying the traffic of secure connections with the required level of detail by protocol is relevant [11].

Objective: improving the algorithms for analyzing the network traffic of secure connections.

The main tasks of the study:

1. Analysis of algorithms for network traffic classification;
2. Development of the network traffic analysis system structure;
3. Development of the algorithm for analyzing the network traffic of secure connections on the basis of algorithms of feature generation and selection for construction a neural network classifier;
4. Software implementation of algorithms for analyzing network traffic and evaluating the effectiveness of the proposed solution on the basis of full-scale data.

2. Analysis of algorithms and systems of network traffic treatment

Traffic classification [12] allows to identify various applications and protocols transmitted over the network. Also, the classification function is the management of this traffic, its optimization and prioritization. All packets become marked by belonging to a specific protocol or application after classification. This allows network devices to apply quality of service policies (QoS) based on these labels and flags.

There are two main methods of traffic classification (Figure 1):

- Classification based on data blocks (Payload-Based Classification). It is based on the analysis of data packet fields. This method is the most common but does not work with encrypted and tunneled traffic.
- Classification based on statistical analysis (time between packets, session time, etc.).

A universal approach to traffic classification based on information in the header of the IP packet. This is usually IP address (Layer 3), MAC address (Layer 2) and the protocol used. This approach has its limitations [13].

Deep package inspection (DPI) allows to implement more advanced classification . The main mechanism for identifying applications in DPI is signature analysis [14]. Each application has its own unique characteristics, which are entered in the database of signatures. Comparing the sample from the database with the analyzed traffic allows to determine the application or protocol. However, new

applications periodically appear, the signature database also needs to be updated to ensure high identification accuracy.

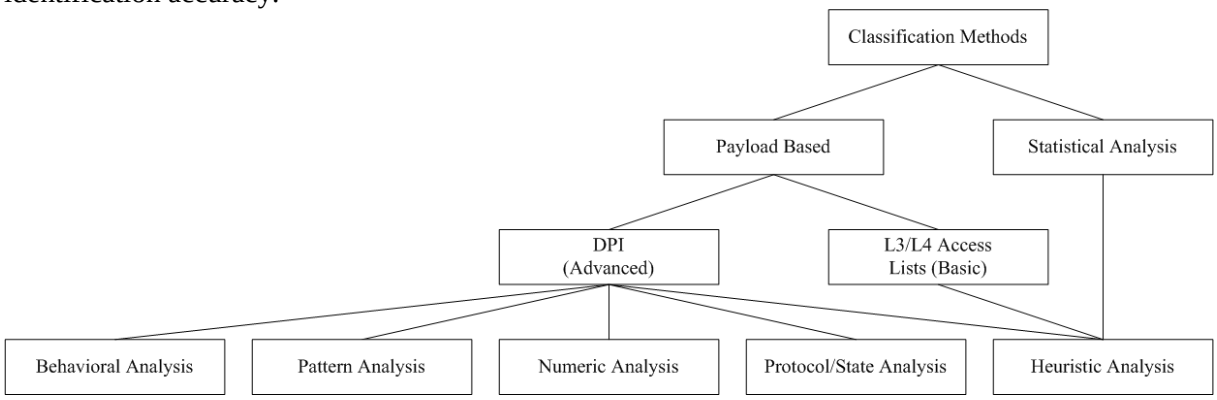

Figure 1. Methods for classifying network traffic.

There are several methods of signature analysis:

Pattern analysis

The applications contain certain sample sequences in the package data block. They can be used for identification and classification. Not every package contains a sample application data, so the method does not always work.

Numerical analysis

Numerical analysis uses the quantitative characteristics of the sequence of packets, such as: size of the data block, response time, interval between packets. Simultaneous analysis of multiple packets is time consuming, which reduces the effectiveness of this method.

Behavioral analysis, Heuristic analysis

Method is based on the analysis of traffic dynamics of the running application. While the application is running, it creates traffic that can also be identified and labeled [15].

Protocol/state analysis

Protocols of some applications are a sequence of certain actions. Analysis of such sequences allows to accurately identify the application. reduces the effectiveness of this method.

Behavioral and heuristic analysis are used when working with encrypted traffic. For more accurate identification, cluster analysis is used, which combines the methods of heuristic and behavioral analysis [16, 17].

3. Development of the network traffic system analysis based on machine-learning training

The development of analysis algorithm for classifying network traffic of secure connections of dedicated users according to a pre-defined set of categories is actual [18-20].

Two scenarios of network traffic analysis are considered:

- analysis of encrypted traffic;
- analysis of encrypted traffic passing through a virtual private network (VPN).

The figure 2 shows the organization's LAN structure with an analysis module for network traffic of encrypted connections.

Traffic arrives from the edge router. There is a seizure and preprocessing of traffic using the libpcap library. The primary features of the data flow are extracted from the received pcap files. Vector of primary features and sessions with a duration of 15, 30, 60 and 120 seconds is formed. The generation and selection of features for the neural network classifier training is performed. The prepared vector of features is fed to the neural network analysis module of user sessions. The settings for training and work are set by the administrator.

This process is shown in the figure 3.

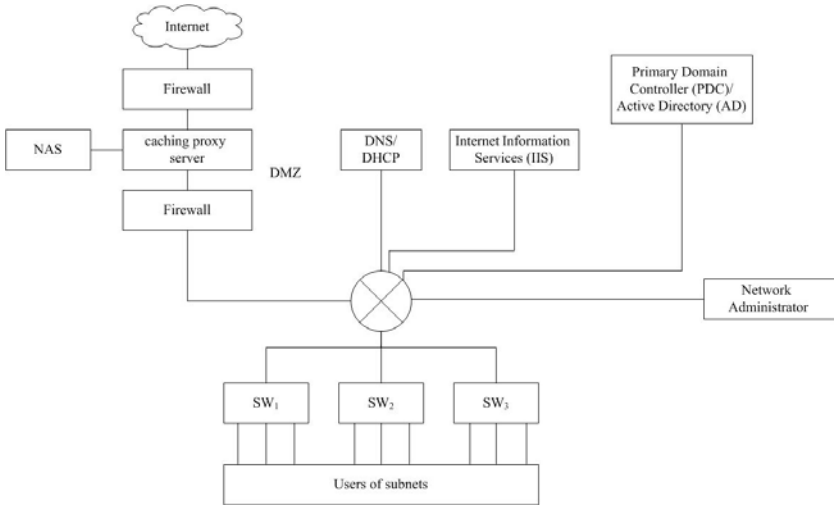

Figure 2. The structure of the LAN with an expanded system for analyzing network traffic.

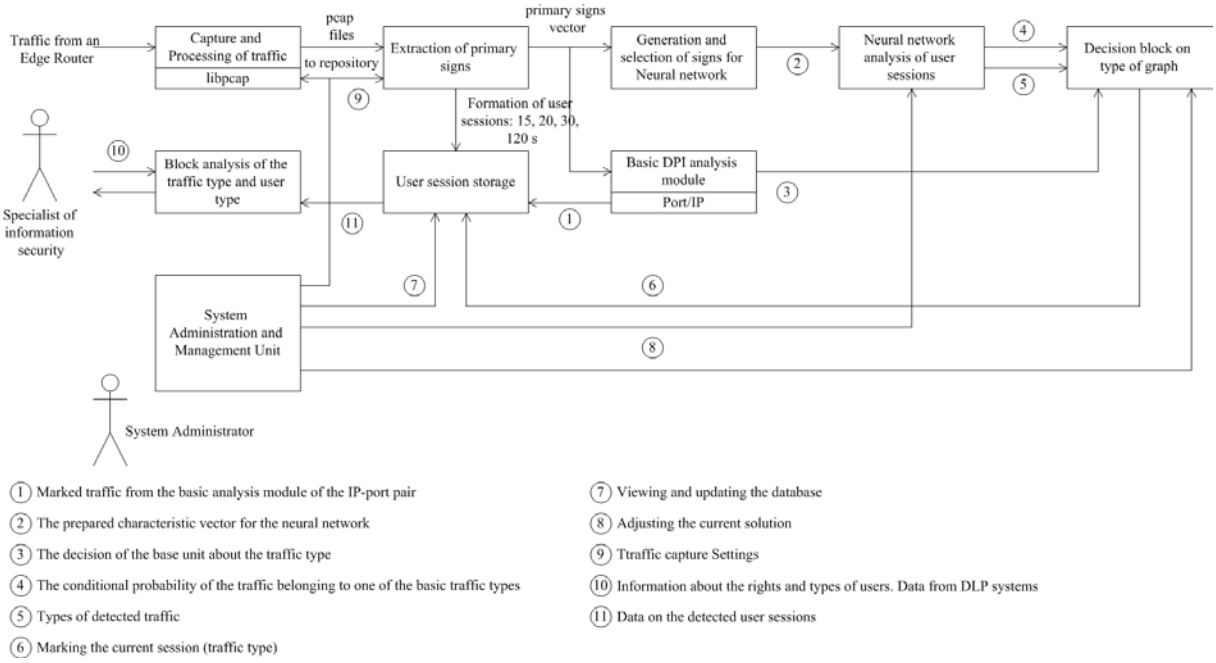

Figure 3. Structure of the network traffic analysis system.

Further, the following information comes to the decision block: the decision of the base block on the type of traffic, the probability of the traffic belonging to one of the basic types and the types of recognized traffic from the NA block analyzing user sessions. Administrator can perform the adjustment of the current solution on the decision block on the type of traffic.

Then, the current traffic from the decision block and marked traffic from the basic traffic analysis module (sender's IP, recipient's IP, port, sender, port of the recipient) are sent to the user sessions store.

Further, data on recognized user sessions is sent to the traffic type analysis module and the user type. An information security specialist receives information about the types of users and their rights. The administrator interacts with the repository to view and replenish the database and sets the parameters for capturing traffic.

Development of an algorithm for analyzing network traffic

At the first step, the fragment of the intercepted traffic is downloaded, then the classifier scenario is selected. Based on the features indicated in the scenario, a training sample is formed to build the initial

knowledge base. After analyzing the given features on the test sample, the accuracy of the classifier is determined. If the accuracy satisfies the requirements, the current state is saved, otherwise the cycle returns to the definition of the type of the script. The block diagram of the network traffic classifier is shown in Figure 4.

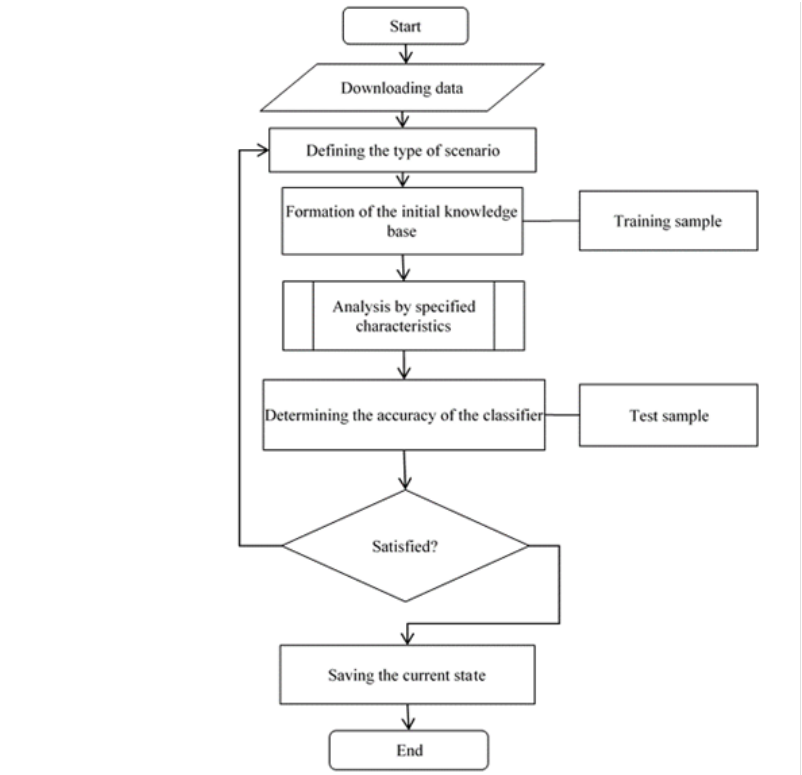

Figure 4. Block diagram of the network traffic classifier.

4. Realization of network traffic analysis algorithms and experiment on natural data

Traffic classification is based on the analysis of the temporal characteristics of the intercepted network packets stream for the formation of encrypted and VPN-traffic features (time-related features). The temporal characteristics of the flow make it possible to reduce the computational cost of building a set of features extracted from the encrypted network traffic by reducing the set of fixed parameters.

The experiment uses a dump of network traffic [21-29] with 14 different traffic type tags generated by different applications (7 for conventional encrypted traffic and 7 for VPN traffic).

Table 1. Fixed time-dependent network traffic parameters.

Function	Description
duration	Flow duration
fiat	The interval between 2 packets sent from the network client (average, min, Max).
biat	The interval between the two packets sent to the network client (average, min, Max).
flowiat	The interval between two packets sent in any direction (average, min, Max).
active	The interval during which the network exchange was active, before switching to the standby mode (average, min, Max).
idle	The interval during which the network exchange was in the standby mode (average, minimum, maximum).
fb psec	Intensity of the flow, bytes per second
fp psec	Intensity of flow, packets per second

The quality criterion for traffic classification is the accuracy of classifying samples. Assessment of the classification accuracy can be carried out by cross-validation [30]. The separation into the training and test sets is performed by dividing the sample in a certain pro-portion – the training set is two-thirds of the data and the test set is one-third of the data.

To solve the classification problem, the following algorithms are considered:

- Random Forest algorithm (RFT);
- K-Nearest Neighbor method (KNN);
- Multilayer Perceptron (MLP) [31].

As the source data, the real traffic generated by such applications and services as Skype, Facebook, etc. is used. Table 1 provides a complete list of the different types of traffic and applications included in the source dataset.

For each type of traffic (VoIP, P2P, etc.), open sessions and sessions are used in the created VPN tunnel, therefore there are in total 14 categories of traffic: VoIP, VPN-VoIP, P2P, VPN-P2P, etc.

Table 2. List of captured protocols and applications.

Traffic	Content	Method of generation
Web-browsing	Firefox, Chrome	HTTPS traffic generated by users while viewing or executing any task that involves using the browser.
Email	SMPTS, POP3S, IMAPS	Samples of traffic generated by the Thunderbird client and Gmal accounts. Clients are configured to deliver mail through SMTP / S, receive and use it using POP3 / SSL in one client and IMAP / SSL in another
Chat	ICQ, AIM, Skype, Face-book, Hangouts	The chat label defines applications for instant messaging (Face-book and Hangouts via web browser, Skype, and IAM and ICQ).
Streaming video	Vimeo, YouTube	The streaming label defines multimedia applications that require a continuous and stable data flow. For example, the services of YouTube (HTML5 and Flash version) and Vimeo, using Chrome and Firefox.
File Transfer	Skype, FTPS and SFTP using FileZilla and external service	The label identifies the application traffic for sending or receiving files and documents. The files were transferred to Skype, FTP through SSH (SFTP) and FTP through SSL (FTPS).
VoIP	Voice calls on Facebook, Skype	The IP Telephony Label groups all traffic generated in voice ap-plications (Facebook, Hangouts and Skype)
P2P	uTorrent and transfer (BitTorrent)	A label is used to identify file sharing protocols, such as Bit-Torrent

Traffic was captured using the Wireshark sniffer. For VPN-traffic, the external service of the VPN provider was used, the connection was made using OpenVPN. To generate SFTP and FTPS traffic, use an external service provider and FileZilla as the client.

Let's define two different scenarios A and B (Figure 5). Four different time duration values were used to generate the data sets.

Scenario A: The purpose is to select the features of encrypted traffic with VPN identification, for example, distinguishing between voice calls (VoIP) and voice calls passing through VPN (VPN-VoIP). As a result, there are 14 different types of traffic: 7 regular types of encrypted traffic and 7 types of traffic passing through the VPN. The first classifier uses VPN and non-VPN traffic separation, and then each traffic type classified separately (VPN and non-VPN).

Scenario B: in this case we use a mixed data set. The classifier's input is regular encrypted traffic and VPN traffic, and the output is allocated the same 14 different categories.

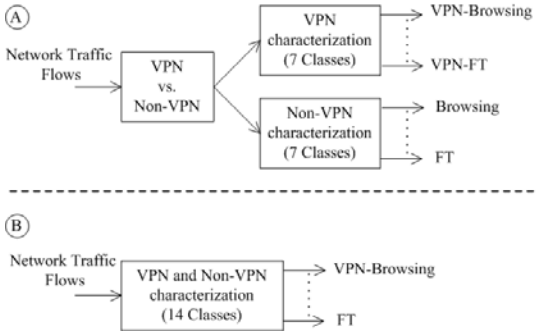

Figure 5. Scenarios for handling captured traffic.

Using general definition of traffic, it is determined by the sequence of packets with the same values for: destination IP address, sender's IP address, sender's port, sender's port, and protocols (TCP or UDP).

Streams are considered bidirectional. Together with the generation of traffic, the features associated with each type of traffic are determined. TCP stream are usually terminated when the connection is broken (by the FIN packet), and the UDP stream are terminated by a thread interruption. The value of the thread interrupt can be assigned arbitrarily. In particular, we set the duration of the streams to 15, 30, 60 and 120 seconds [32-34].

Scenario A analysis

MLP is a neural network of direct propagation. It consists of two hidden layers. In the first hidden layer there are 30 neurons, and in the second there are 15 neurons. Size of input vector of features is 23. The activation function of neurons is a hyperbolic tangent [35]. As a numerical metric, the root-mean-square error is used to estimate the network error. As a learning algorithm, the method of conjugate gradients is used. The number of learning epochs is 5000.

When testing the KNN algorithm, the neighbor number parameter is 50.

When testing the random tree method, the classifier type “Bag” is used. The number of ensemble training cycles is 150.

Table 3. Obtained data for scenario A1.

Session Length and Type	Classifier	Training sample			Test sample		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
120s-VPN	MLP	0.799	0.851	0.826	0.748	0.822	0.787
120s-VPN	KNN	0.786	0.715	0.749	0.763	0.712	0.736
120s-VPN	RFT	0.957	0.998	0.978	0.805	0.848	0.828
15s-VPN	MLP	0.820	0.831	0.826	0.796	0.783	0.789
15s-VPN	KNN	0.798	0.776	0.786	0.784	0.764	0.773
15s-VPN	RFT	0.975	0.999	0.988	0.848	0.866	0.857
30s-VPN	MLP	0.826	0.847	0.837	0.774	0.819	0.798
30s-VPN	KNN	0.753	0.773	0.764	0.764	0.774	0.770
30s-VPN	RFT	0.969	0.998	0.984	0.831	0.880	0.857
60s-VPN	MLP	0.831	0.791	0.813	0.787	0.729	0.761
60s-VPN	KNN	0.826	0.687	0.764	0.826	0.659	0.750
60s-VPN	RFT	0.964	0.996	0.978	0.838	0.808	0.824

The cross validation (Tables 4, 5) was performed in order to assess how classifiers are able to work with real data, while producing a result whose accuracy is correlated to accuracy in the test sample [36].

Table 4. Cross-validation results for scenario A1.

Session Length and Type	Classifier	Average accuracy	Deviation
120s-VPN	KNN-crossval	0.730	0.015
120s-VPN	RFT-crossval	0.826	0.007
15s-VPN	KNN-crossval	0.770	0.010
15s-VPN	RFT-crossval	0.858	0.009
30s-VPN	KNN-crossval	0.747	0.013
30s-VPN	RFT-crossval	0.840	0.008
60s-VPN	KNN-crossval	0.751	0.013
60s-VPN	RFT-crossval	0.815	0.007

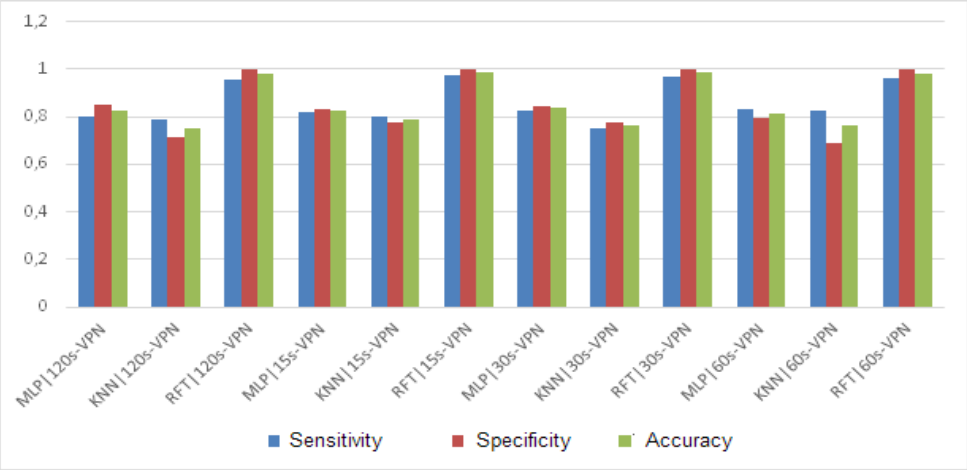

Figure 5. Classification results on the training sample for scenario A.

There is a direct relationship between the length of the captured session of the thread and the performance of the classifiers. When using the RFT classifier, the accuracy on the test sample decreases from 0.857 with a flow time of 15 seconds to 0.828 using a 120 second flow. Similar behavior is observed for the KNN and MLP algorithms. The best results are achieved using the RFT algorithm, with the time required to create the stream equal to 15. These results show that, using shorter time-out values for the traffic classifier, you can in-crease the accuracy value.

The second part of scenario A focuses on the separation of VPN and non-VPN traffic. The input is classified according to traffic categories. The results for shorter duration values are better than the results for larger values, albeit with some exceptions. In the case of VPN classifier, as VPN-mail, where the best result is obtained with the value of ftm equal to 30 seconds. In the case of a non-VPN classifier, the same thing happens.

Analysis of scenario B

All encrypted streams and VPN traffic are mixed in one set of data. The goal is to classify traffic without prior VPN separation from non-VPN traffic. There are 14 types of traffic: 7 encrypted and 7 VPN traffic types.

The short duration of the session of the captured stream does not provide the greatest accuracy. For example, for the MLP algorithm, the test accuracy is 0.795 and 0.51 for 15 seconds, and for a session time of 30 seconds, the accuracy on the test sample for the same algorithm is 0.798 and 0.637. The highest accuracy on the test sample for different interrupt values is 0.847 (RFT algorithm with a flow time of 120 seconds).

Table 5. Cross-validation results for scenario A2.

Session Length and Type	Classifier	Average accuracy	Deviation
120s-NO-VPN	KNN-crossval	0.776	0.012
	RFT-crossval	0.882	0.007
120s-VPN	KNN-crossval	0.679	0.020
	RFT-crossval	0.838	0.009
15s-NO-VPN	KNN-crossval	0.807	0.012
	RFT-crossval	0.889	0.006
15s-VPN	KNN-crossval	0.721	0.019
	RFT-crossval	0.830	0.009
30s-NO-VPN	KNN-crossval	0.798	0.008
	RFT-crossval	0.875	0.006
30s-VPN	KNN-crossval	0.710	0.017
	RFT-crossval	0.846	0.008
60s-NO-VPN	KNN-crossval	0.744	0.013
	RFT-crossval	0.854	0.006
60s-VPN	KNN-crossval	0.670	0.016
	RFT-crossval	0.817	0.009

Table 6. Obtained data for scenario B.

Session Length and Type	Classifier	Training sample			Test sample		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
120s-AllinOne	MLP	0.922	0.872	0.802	0.909	0.859	0.787
120s-AllinOne	KNN	0.911	0.772	0.739	0.915	0.762	0.735
120s-AllinOne	RFT	0.913	0.999	0.965	0.507	0.984	0.847
120s	MLP	0.853	0.895	0.629	0.829	0.886	0.607
120s	KNN	0.816	0.832	0.604	0.818	0.835	0.618
120s	RFT	0.972	1	0.964	0.457	0.993	0.767
15s-AllinOne	MLP	0.896	0.945	0.834	0.860	0.929	0.795
15s-AllinOne	KNN	0.8782	0.8802	0.769	0.878	0.873	0.752
15s-AllinOne	RFT	0.979	0.999	0.984	0.622	0.968	0.846
15s	MLP	0.870	0.839	0.521	0.879	0.841	0.510
15s	KNN	0.656	0.904	0.662	0.615	0.905	0.653
15s	RFT	0.996	1	0.982	0.565	0.984	0.776
30s-AllinOne	MLP	0.915	0.941	0.830	0.882	0.930	0.798
30s-AllinOne	KNN	0.892	0.858	0.744	0.889	0.844	0.737
30s-AllinOne	RFT	0.967	0.999	0.979	0.605	0.970	0.844
30s	MLP	0.811	0.922	0.661	0.780	0.916	0.637
30s	KNN	0.727	0.892	0.634	0.722	0.892	0.629
30s	RFT	0.986	1	0.980	0.537	0.983	0.767
60s-AllinOne	MLP	0.911	0.930	0.795	0.889	0.913	0.754

60s-AllinOne	KNN	0.902	0.842	0.734	0.882	0.840	0.717
60s-AllinOne	RFT	0.952	0.999	0.973	0.576	0.970	0.810
60s	MLP	0.799	0.922	0.622	0.750	0.911	0.602
60s	KNN	0.757	0.891	0.615	0.763	0.883	0.606
60s	RFT	0.973	0.999	0.971	0.544	0.985	0.737

Table 7. Inaccuracy matrix of all traffic with a flow duration of 60 seconds.

Inaccuracy matrix		**Actual data of the session**						
		Web-browsing	Chat	File Transfer	Email	P2P	Streaming video	VoIP
The result of the classifier's work	Web-browsing	189	18	10	1	7	15	53
	Chat	19	373	27	7	13	21	24
	File Transfer	15	30	234	3	0	4	14
	Email	0	3	6	74	1	3	1
	P2P	7	9	1	3	803	41	9
	Streaming video	29	54	16	18	28	340	8
	VoIP	69	66	19	16	19	28	1131

The cross-validation (Table 8) was performed to assess the effectiveness of the KNN method and the RFT method.

Table 8. Cross-validation results for scenario B.

Session Length and Type	**Classifier**	**Average accuracy**	**Deviation**
120s-AllinOne	KNN-crossval	0.722	0.014
	RFT-crossval	0.829	0.009
120s	KNN-crossval	0.582	0.013
	RFT-crossval	0.733	0.011
15s-AllinOne	KNN-crossval	0.757	0.008
	RFT-crossval	0.845	0.009
15s	KNN-crossval	0.643	0.014
	RFT-crossval	0.768	0.012
30s-AllinOne	KNN-crossval	0.733	0.018
	RFT-crossval	0.835	0.008
30s	KNN-crossval	0.616	0.017
	RFT-crossval	0.757	0.013
60s-AllinOne	KNN-crossval	0.721	0.011
	RFT-crossval	0.819	0.009
60s	KNN-crossval	0.594	0.014
	RFT-crossval	0.724	0.011

5. Conclusion

Traffic analysis systems are widely used in monitoring the network activity of users or a specific user and restrict the client's access to certain types of services (VPN, HTTPS). This makes analysis of content impossible. Algorithms for classification of encrypted traffic and detection of VPN traffic are proposed. Three algorithms for constructing classifiers: MLP, RFT and KNN, are considered.

The effect of the session length of the captured data stream on the accuracy of the classification is established. The developed classifier demonstrates the accuracy of recognition on the test sample to 80%. Algorithms MLP, RFT and KNN had almost identical indicators in all experiments.

It is also established that the proposed classifiers work better when network traffic flows are generated using short time-out values.

The novelty lies in the development of algorithms for analyzing network traffic on the basis of a neural network. This method differs in the way of features generation and selection, which allows classifying the existing traffic of protected connections of selected users according to a predefined set of categories.

The developed algorithms can improve the security of the data transmission network by improving the algorithms for analyzing network traffic as part of a data leak prevention system.

6. References

[1] DPI Technology Overview – Deep Packet Inspection URL: https://habr.com/post/111054/
[2] Olifer V G, Olifer N A 2011 *Computer networks. Principles, technologies, protocols: Textbook for high schools* (SPb.: Peter) p 944
[3] Analyzers of network packets URL: https://compress.ru/article.aspx?id=16244
[4] Smith R 2008 Deflating the big bang: fast and scalable deep packet inspection with extended finite automata *ACM SIGCOMM Computer Communication Review* **38(4)** 207-218
[5] Traffic security URL: https://habr.com/post/46321/
[6] Kostin D V, Sheluhin O I 2016 Comparison of machine learning algorithms for encrypted traffic classification *T-Comm.* **10(9)** 43-52
[7] Get'man A I 2017 Overview of tasks and methods for solving them in the field of classification of network traffic *Proc. of the Institute for Syst. Prog. of the Russian Academy of Sciences* **29(3)**
[8] Lim Y 2010 Internet traffic classification demystified: on the sources of the discriminative power *Proc. of the 6th Int. Conf. (ACM)* p 9
[9] Moore A W, Zuev D 2005 Internet traffic classification using bayesian analysis techniques *ACM SIGMETRICS Performance Evaluation Review* **33(1)** 50-60
[10] Federal Law № 149-FZ "On Information, Information Technologies and Information Protection" URL: www.internet-law.en/law/inflaw/inf.htm
[11] Traffic classification and Deep Packet Inspection URL: https:// vasexperts.ru/blog/klassifikatsiya-trafika-i-deep-packet-inspection/
[12] Sukhov A M, Sagatov E S, Baskakov A V 2014 Analysis of Internet service user audiences for network security problems *2nd Int. Symp. on Telecommunication Technologies (ISTT)* 214-219
[13] The composition of the DPI system URL: https://vasexperts.ru/blog/sostav-sistemy-dpi/
[14] Moore A W, Zuev D 2005 Internet traffic classification using bayesian analysis techniques *Int. Conf. Measurement and Modeling of Comp. Syst.* URL: http://www.cl.cam.ac.uk/~awm22/ publications/moore2005internet.pdf
[15] Russian manufacturers of DPI and their platforms URL: https://vasexperts.ru/blog/rossijskie-proizvoditeli-dpi-i-ih-platfo/
[16] Foreign DPI manufacturers and their platforms URL: https://vasexperts.ru/blog/inostrannye-proizvoditeli-dpi-i-ih-platf/
[17] Sherry J, Lan C, Popa R A and Ratnasamy S *Blindbox: Deep packet inspection over encrypted traffic* URL: http://iot.stanford.edu/pubs/sherry-blindbox- sigcomm15.pdf
[18] Shen F, Pan C and Ren X 2007 Research of P2P traffic identification based on BP neural network *3rd Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing* **2** 75-78
[19] Raahemi B 2008 Classification of Peer-to-Peer traffic using incremental neural networks (Fuzzy ARTMAP) *Canadian Conf. on Electrical and Comp. Eng.* 719-724
[20] The UNSW-NB15 Dataset Description URL: https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA- NB15-Datasets/
[21] Tor-non Tor dataset (ISCXTor2016) URL: http://www.unb.ca/cic/datasets/tor.html

[22] Lotfollahi M et al 2017 Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning *arXiv preprint arXiv:1709.02656*
[23] Lopez-Martin M et al 2017 Network traffic classifier with convolutional and recurrent neural networks for Internet of Things *IEEE Access* **5** 18042-50
[24] Miller B 2014 I know why you went to the clinic: Risks and realization of https traffic analysis *Int Symp. on Privacy Enhancing Technologies* (Springer, Cham) 143-163
[25] Foremski P 2013 On different ways to classify Internet traffic: a short review of selected publications *Theoretical and Applied Informatics* **25(2)** 119-136
[26] Smit D 2017 Looking deeper: Using deep learning to identify internet communications traffic *Macquarie Matrix: Special edition, ACUR* **1** 1318-1323
[27] Michael A K J *Network traffic classification via neural networks* Technical Report (University of Cambridge, Computer Laboratory) p 25
[28] Belov S D 2008 Detection of patterns and recognition of abnormal events in the data stream of network traffic *Vestnik NGU: Information Technologies* **6(2)** 57-68
[29] Haykin S 2006 *Neural networks. Neural networks: a full course* (Moscow: Publishing house "Williams") p 1104
[30] Data science. What is cross-validation? URL: http://datascientist.one/cross-validation/
[31] Sukhov A M, Sagatov E S and Baskakov A V 2017 Rank distribution for determining the threshold values of network variables and the analysis of DDoS attacks *Procedia Engineering* **201** 417-27
[32] Galtsev A A, Sukhov A M 2011 Network attack detection at flow level *11th Int. Conf., NEW2AN: Lecture Notes in Computer Science* **6869** 326-334
[33] Salimov A S 2018 Application of SDN Technologies to Protect Against Network Intrusions *Int. Scien. and Techn. Conf. Modern Comp. Network Technologies (MoNeTeC)* 1-9
[34] Multilayer perceptron URL: http://www.aiportal.ru/articles/neural-networks/multi-perceptron.html
[35] Mastering fuzzy modeling methods and developing an algorithm to optimize the fuzzy classifier rules base based on observable data using a genetic algorithm URL: http:// refleader.ru/jgernayfsyfs.html

Acknowledgments
This work is partially supported by the Russian Science Foundation under grants № 17-07-00351.