Research of Methods of Identifying the Computer Systems State based on Bagging Classifiers

Svitlana Gavrylenko, Oleksii Hornostal and Viktor Chelak

October 3, 2022

Research of Methods of Identifying the Computer Systems State based on Bagging Classifiers

Svitlana Gavrylenko
Department of Computer Engineering and Programming
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
ORCID: 0000-0002-6919-0055

Oleksii Hornostal
Department of Computer Engineering and Programming
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
ORCID: 0000-0001-5820-9999

Viktor Chelak
Department of Computer Engineering and Programming
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
ORCID: 0000-0001-8810-3394

***Abstract*—Peculiarities of constructing ensemble bagging classifiers for identifying the state of a computer system under conditions of noisy data are studied. Decision trees and multilayer perceptron were used as basic classifiers. It was found that the accuracy of the bagging algorithm with decision trees as basic classifiers with standard settings ranges from 84.4% to 88.7%. The use of Bootstrap algorithms for the formation of data samples: Pasting, Bootstrapping, Random Subspace, Random Patches Ensemble and the selection of the number of basic classifiers in the ensemble made it possible to increase the classification accuracy to 90.2%. The following parameters were added to improve the accuracy of bagging classifiers based on the multilayer perceptron: the algorithm for forming data samples, the number of basic classifiers in the ensemble, the function of optimizing the neural network, the function of activating hidden layer, size of hidden layers. The recommendation was made to choose the value of the analyzed parameters for the creation of bagging ensembles with multilayer perceptrons, which made it possible to increase the accuracy of computer system identification up to 92.2%. The obtained results have further practical significance and can be used in improving the methods of identifying the state of computer systems.**

***Keywords—computer system, state identification, machine learning, bagging, decision trees, multilayer perceptron*.**

I. Introduction

When solving problem related to the diagnosis and protection of computer information resources, the central task is the prompt detection of anomalous behavior of the computer system under conditions of external influences.

Cyber threat statistics record a significant increase in the annual number of attacks, which leads to significant economic, moral and reputational losses. Thus, in 2021, the average number of cyber attacks and data breaches increased by 15.1% [1]. As experts predict, the number of attacks will also increase. This requires improving the infrastructure, reviewing the information security strategy, computer system architecture and methods of implementation and means of identifying their state, especially in the conditions of constant transformation of the global security system.

The object of the research is the process of identifying the state of the computer system.

The subject of the research is the methods of identifying the state of computer system.

The purpose of the research is to develop a method for identifying the state of a computer system based on the use of bagging classifiers.

II. Literature Review

The functioning of the computer system (CS) is characterized by a large number of processes. Complex mathematical algorithms based on machine learning methods are used to analyze this data and classify it. The most popular machine learning algorithms are given in [2,3].

One of the best classification methods is the support vector machine [4]. Disadvantages of the support vector method are the ability to perform only binary classification. Also, model parameters are difficult to interpret.

Bayes based classifier [5] characterized by a simple implementation of the algorithm in the form of a program, high speed of operation, easy interpretation of the results of the algorithm. Despite the above advantages, the Bayes method has insufficient classification accuracy and is unable to take into account the dependence of the classification result on the combination of features.

The advantages of the k-nearest neighbors method [6] are: simple implementation, the presence of a good theoretical base, adaptation to the required task by choosing a metric or a kernel. The disadvantages of this method include: insufficient performance in real tasks, since the number of neighbors used for classification will be quite large; difficulties in setting appropriate weights and determining which features are necessary for classification; dependence on the chosen metric of the distance between objects.

Neural networks [7, 8] are quite effective because they generate a large number of regression models (which are used in solving classification problems by statistical methods). However, any method based on neural networks will never provide a classifier of the desired quality if the set of training samples is not complete enough for the task that the system will have to work with.

The method of decision trees (DT) is characterized by high training and forecasting performance. DT can be easily visualized and interpreted. The disadvantage of the method is the relatively low accuracy of forecasts, since the construction of the classifier significantly depends on the input parameters, the data structure, and the nature of their occurrence [9]. To overcome the above disadvantages, methods based on the use of ensembles of several classifiers have been developed. Ensembles improve the quality, reduce the dependence of models on the studied data and input parameters, increasing the stability of the results.

One of the most popular classification methods is ensemble models based on bagging [10,11].

III. Approaches and Methods

Bagging is a simple technique based on the idea of combining independent weak classifiers. At the same time, classifiers are trained in parallel, using the same learning algorithm.

Bagging is based on Condorcet's theorem:

$$\mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i} , \qquad (1)$$

where μ – the probability of making a correct decision by the classifier, N – number of decision trees, m – the minimum value of the required majority of correct classifier solutions, p – the probability of making a correct decision of the classifier, C_N^i – the number of permutations of N objects taken i at a time.

Bagging will mainly focus on obtaining an ensemble model with less spread than its components and will aim to reduce variance by averaging the results. At the same time, a weighted average is used to solve the regression problem

$$S_L(.) = \frac{1}{L}\sum_{i=1}^{L} w_l(\cdot) , \qquad (2)$$

where L – number of classifiers, $w_l(\cdot)$ – work result of *l* – base classifier.

Majority voting is usually used to solve the classification task

$$S_L(.) = \arg_k \max[card(l \mid w(\cdot) = k] \qquad (3)$$

In addition, classification problems determine the probabilities of each class predicted by all models. In the future, the values should be averaged and the class with the highest average probability should be saved (soft voting). Averages or votes may be simple or weighted if any appropriate weights are used.

When forming an ensemble of models, a DT is most often used as the basic classifier. DT is easy to interpret, requires little prior data preparation, can work with both numerical and categorical data, uses a white-box model, and is easily explained in Boolean logic. However, DT is very sensitive to small changes in training data and classifier settings, has high sensitivity to noise, and low accuracy.

To overcome the above disadvantages in the work as basic classifiers, classifiers based on an elementary perceptron [12] with simple A- and R-elements and transfer functions of relations of the form were studied:

$$C_{ij}(t) = w_{ij}(t) U\text{out}.i\,(t - \tau_{ij}), \qquad (4)$$

where $w_{ij}(t)$ – the weight of the connection between the *i*-th and *j*-th neurons at the moment of time *t*; $U\text{out}.i\,(t - \tau_{ij})$ – the output signal of the *i*-th neuron at the moment of time ($t - \tau_{ij}$); τ_{ij} – signal transmission time $U\text{out.i}\,(t - \tau\text{ij})$ rom the output of the *i*-th neuron to the input of the *j*-th element.

In addition, when building a multilayer perceptron, various functions are used to optimize the weights of neural networks, namely:

- ‘lbfgs’ – an optimizer from the family of quasi-Newton methods [13];
- ‘sgd’ – stochastic gradient descent [14];
- ‘adam’ – the optimizer that is based on principles of stochastic gradient descent and proposed by researchers Kingma, Diederik and Jimmy Ba [15].

In addition, when building a multilayer perceptron, it is also necessary to choose the activation function of the hidden layer. The following functions are the most popular:

- ‘identity’ – no-op activation, useful for implementing a linear bottleneck, returns f(x) = x;
- ‘logistic’ – logistic sigmoid function that returns f(x) = 1 / (1 + exp(-x));
- ‘tanh’– the hyperbolic tan function that returns f(x) = (x);
- ‘relu’ – rectified linear unit function that returns f(x) = max(0, x).

The efficiency of bagging is achieved due to the fact that the errors of the basic algorithms trained on different subsamples are mutually compensated during voting, as well as due to the fact that outlier objects may not fall into some training samples.

To build classifiers, various algorithms for forming data samples are used: Pasting, Bootstrapping, Random Patches, Random Subspaces [16,17]. At the same time, about 60% of the raw data $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, are used as samples for training, the rest – only for testing.

According to the meta-algorithm Pasting subsamples contain all the original features $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, are formed randomly, are unique and unrepeatable. The main disadvantage of this process is that each subsample cannot be repeated and this creates a problem when the data set is not large enough [16].

According to the meta-algorithm Bootstrapping the subsamples containing all the original features $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, are randomly generated and can be repeated.

According to the meta-algorithm Random Patches subsamples are created by randomly selecting part of the features $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, and may be repeated.

According to the meta-algorithm Random Subspaces subsamples are created by randomly selecting part of the features $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$, are unique and unrepeatable [18].

IV. Experimental Research and Efficiency Evaluation

The main task of the conducted experiment is the research and development of methods for identifying the state of the computer system based on ensemble classifiers, their adjustment taking into account the peculiarities of the input data to improve the quality of work.

The use of the above algorithms for the formation of data samples: Pasting, Bootstrapping, Random Subspace and Random Patches Ensemble was analyzed to generate samples of the initial data of the basic classifiers.

The parameters of the operation of the CS (loading of the central processor, memory, volume of traffic, number of read/write operations to the disk, signatures of intrusions; statistical data of the analysis of system events (the number of operations of working with the system registry, the file system, the number of processes and etc.) At the same time, in order to increase the accuracy of the classification at the boundary of the delimitation, the input data were additionally made noisy.

At the beginning, the performance quality of the standard version of the bagging algorithm with decision trees as basic classifiers was analyzed. At the same time, such setting parameters as the above algorithms for forming data samples and the number of basic classifiers in the ensemble were investigated. It was found that the minimum accuracy is equal to 84.4%, and the accuracy under standard settings (using the Bootstrapping algorithm for forming data samples and with the number of classifiers equal to 10) is equal to 88.7%. Thanks to the selection of optimal parameters, it was possible to increase the accuracy to 90.2%. The results of the research of bagging classifiers with decision-making trees as basic classifiers are shown in Fig. 1.

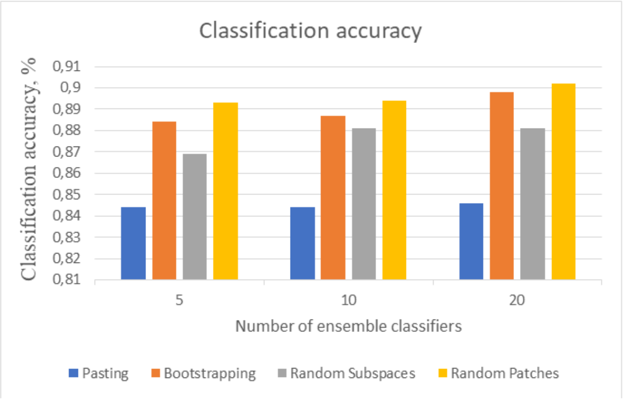

Fig. 1. Comparative analysis of the dependence of the classification accuracy of the bagging classifier based on decision trees on the number of base classifiers.

To improve the accuracy of the ensemble work, it was decided to use multilayer perceptrons as the basic classifiers instead of the usual decision trees .

Combinations of selected parameters-settings that affect the performance quality of the ensemble classifier based on a multilayer perceptron were studied, namely:

1. The algorithm for forming data samples.
2. The number of basic classifiers in the ensemble.
3. The set of parameters of the basic ensemble classifier (function for optimizing neural network weights and hidden layer activation function, sizes of the first and second hidden layers).

The values of the parameters-settings of the classifier considered in the research are given in the Table 1.

TABLE I. THE VALUES OF THE PARAMETERS-SETTINGS

Parameter number	Parameter name	Possible values
1	The algorithm for forming data samples	Pasting, Bootstrapping, Random Patches, Random Subspaces
2	The number of basic classifiers in the ensemble	5, 10, 20
3	The function used to optimize the neural network weights	lbfgs, sgd, adam
4	The hidden layer activation function	identity, logistic, tanh або relu
5	The size of the first hidden layer of the perceptron	10, 100, 200
6	The size of the second hidden layer of the perceptron	0 (is not used), 5, 10, 20

The essence of the experiment is to create classification models using the brute force method of the specified settings parameters and build a trend line for each of the studied settings. The index of the settings vector acts as the abscissa axis. At the same time, the index of the settings vector is ordered by increasing accuracy for the studied setting parameter.

For each set of parameters-settings, 3 classifiers (5184 classifiers in total) were built to perform the cross-validation procedure with the same set of parameters-settings, and the model with the highest accuracy was selected (resulting in 1728 classifiers).

The results of the study of dependence of classification accuracy on the used forming data samples algorithm are presented in Fig. 2. Analysis of the results showed that the use of the Random Patches algorithm is the most qualitative. The worst results were obtained when using the Bootstrapping algorithm. The other two algorithms showed nearly the same results. At the same time, under certain settings, the use of the Pasting algorithm made it possible to obtain the highest accuracy classifier.

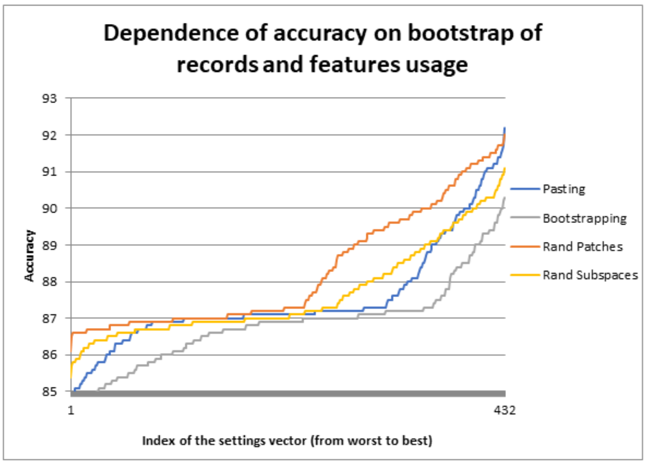

Fig. 2. Study of the influence of using the Bootstrap procedure for input data and features on the classification accuracy of the Bagging ensemble.

The study of the dependence of classification accuracy on the number of basic classifiers in the ensemble showed that the difference in accuracy is not significant (Fig. 3). Thus, it is recommended to take five classifiers if the speed of model building is a priority, or to increase their number if the priority is accuracy.

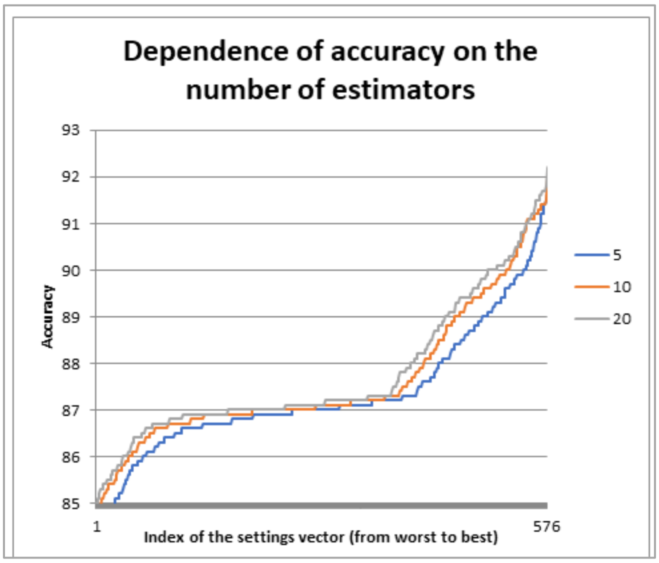

Fig. 3. Study of the influence of the number of independent classifiers in the ensemble on the classification accuracy of the Bagging ensemble.

A study of the dependence of neural network weight optimization functions on the classification accuracy of the Bagging ensemble is shown in Fig. 4. The performed analysis of the algorithms showed that the use of the *sgd*-function is not appropriate. The choice between the other two algorithms requires research and depends on the specific input data set.

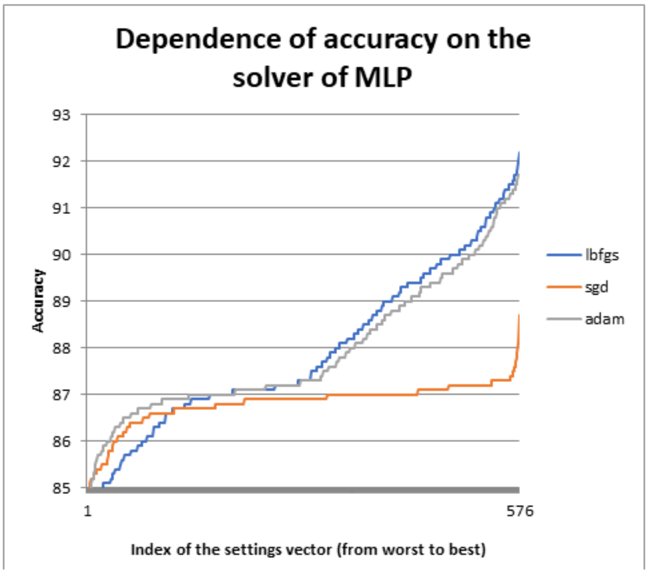

Fig. 4. Study of the influence of neural network weight optimization functions on the classification accuracy of the Bagging ensemble.

The analysis of the effect of the available activation functions of the hidden layer of the multilayer perceptron on the classification accuracy of the Bagging ensemble showed that ensembles using multilayer perceptrons with the identity activation function (Fig. 5) have the lowest accuracy. The use of other activation functions is appropriate and also depends on the set of specific input data.

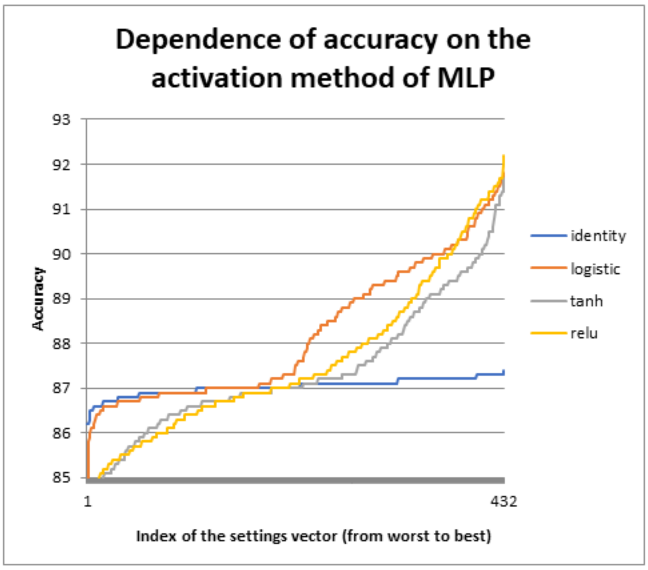

Fig. 5. Study of the influence of the activation functions of the hidden layer of the multilayer perceptron on the accuracy of the Bagging ensemble classification.

The study of the optimal sizes of two hidden layers turned out to be the most interesting and most promising for further research (Fig. 6-7). The value 0 in Fig. 7 corresponds to the situation when the layer is not used. According to the results of the study, it was found that in the first hidden layer it is possible to use a much smaller number of neurons than with standard settings. It is enough to take 10 neurons in the first layer, but the size of the second layer can be increased depending on the task and the features of the input data set.

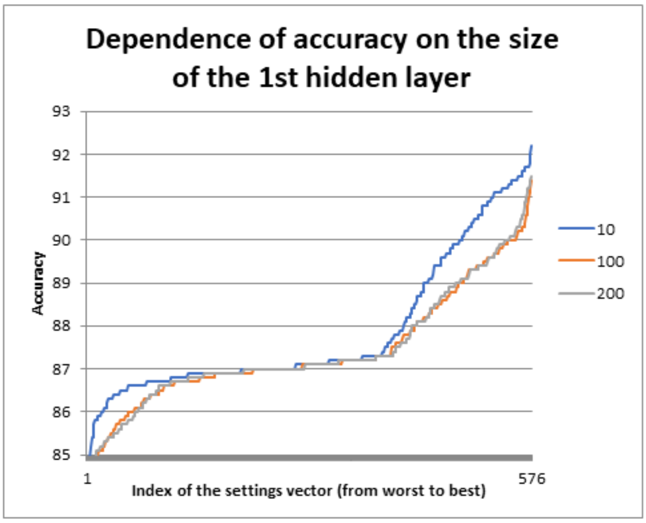

Fig. 6. Study of the influence of the size of the first hidden layer of the multilayer perceptron on the accuracy of the Bagging ensemble classification.

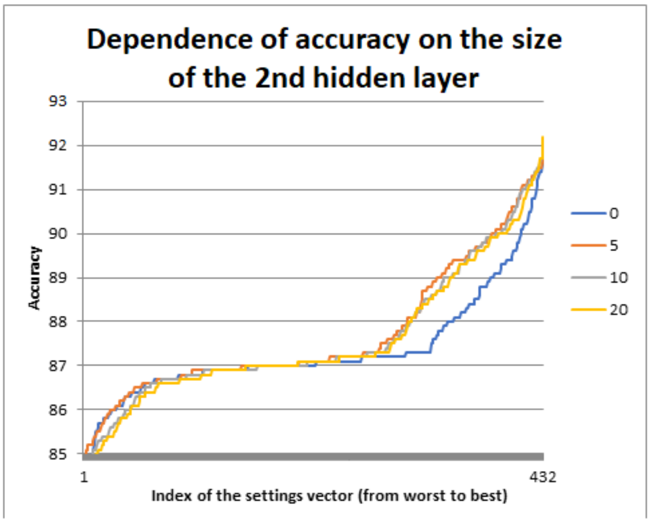

Fig. 7. Study of the influence of the size of the second hidden layer of the multilayer perceptron on the accuracy of the Bagging ensemble classification.

V. Conclusions

In this work, the peculiarities of the construction of ensemble bagging classifiers based on decision trees and multilayer perceptrons as basic classifiers for identifying the state of the computer system have been studied. The results of the research are implemented programmatically using the Google Colab cloud service based on Jupiter Notebook.

The performance indicators of the CS were used as the initial data. At the same time, in order to increase the accuracy of the classification at the boundary of the delimitation, the input data were additionally noisy.

It was found that the accuracy of the bagging algorithm with decision trees as basic classifiers with standard settings ranges from 84.4% to 88.7%. The use of Bootstrap algorithms for forming data samples: Pasting, Bootstrapping, Random Subspace and Random Patches Ensemble and the number of basic classifiers in the ensemble were analyzed as setting parameters. The selection of optimal setting parameters made it possible to increase the classification accuracy to 90.2%.

To increase the accuracy of classification, the work of bagging ensembles with multilayer perceptrons as basic classifiers is considered. The influence of the following parameters-settings on the accuracy of the model was studied: the algorithm for forming data samples, the number of basic classifiers in the ensemble, the optimization function of the neural network weights, the activation function of the hidden layer, the sizes of the first and second hidden layers).

Based on the results of the research, the following recommendations have been developed regarding the selection of the values of the considered parameters-settings of the bagging ensembles with multilayer perceptrons:

1. When choosing an algorithm for forming data samples, it is necessary to take into account that the Random Patches algorithm is more qualitative, and the use of the Bootstrapping algorithm is impractical.
2. Classification accuracy does not significantly increase with an increase in the number of basic classifiers. Thus, it is recommended to take five classifiers if the speed of model building is a priority, or to increase their number if the priority is accuracy.
3. When choosing a function to optimize the weights of neural networks, the use of the sgd-function should be avoided. Other functions should be selected taking into account other parameters-settings.
4. In further studies, it is impractical to use the identity activation function of the hidden layer, and other functions must be used depending on the characteristics of the input data.
5. The study of the optimal sizes of two hidden layers showed that in the first hidden layer it is possible to use a much smaller number of neurons than with standard settings. This makes it possible to significantly reduce the time of building classifiers at the training stage. It is enough to take 10 neurons in the first layer, but the size of the second layer can be increased depending on the task and the features of the input data set.

The developed recommendations for the classifier settings made it possible to increase the accuracy of the bagging ensemble with multilayer perceptrons as basic classifiers to 92.2%.

In further research, to improve the accuracy of the classifiers, it is necessary to focus on the selected values of the setting parameters and focus on the pre-processing of the data taking into account the selected approaches and functions when building the classifier.

Acknowledgment

The work was supported by LLC “Advanced Digital Solutions” as part of R&D project “Models and methods of information processing and protection in computer systems” (0122U200526).

Referenses

[1] Chuck Brooks. Alarming Cyber Statistics For Mid-Year 2022 That You Need To Know URL: https://www.forbes.com/sites/chuckbrooks/2022/06/03/alarming-cyber-statistics-for-mid-year-2022-that-you-need-to-know/?sh=542399877864

[2] Vipin Kumar. The Top Ten Algorithms in DataMining. Taylor & Francis Group, LLC, 2009, 2006 p.

[3] J. Kelleher, B. Namee and A. Archi. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies ,The MIT Pres, 2015, 642 p.

[4] V. Blanco, J. Puerto, and A.M. Rodr´ıguez-Ch´ıa. Support Vector Machines and Multidimensional Kernels, Journal of Machine Learning Research, 2021, vol.1, pp. 1-29. DOI: 10.48550/arXiv:1711.10332v1

[5] S. Wang. Adapting naive Bayes tree classification , Knowledge and Information system, 2015, vol. 44, No. 1, pp. 77–89. DOI: 10.1007/s10115-014-0746-y

[6] S. Sun and R. Huang.An adaptive k-nearest neighbor algorithm, 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010, pp. 91-94, DOI: 10.1109/FSKD.2010.5569740.

[7] M. L. Sidney. K. Heverton et al. Artificial intelligence-based antivirus in order to detect malware preventively. Progress in Artificial Intelligence, 2021, vol.10, pp.1–22. DOI: 10.1007/s13748-020-00220-4

[8] N. Demir, G. Dalkilic. Modified stacking ensemble approach to detect network intrusion. Turkish Journal of Electrical Engineering and Computer Sciences, 2018, vol. 26, No.1, pp. 418 – 433. DOI:10.3906/elk-1702-279

[9] S. Gavrylenko, V. Chelak and S. Semenov. Development of Method for Identification the Computer System State based on the Decision Tree with Multi-Dimensional, Radio Electronics, Computer Science, Control (RECSC), 2022, No. 2, pp.113-121. DOI: 10.15588/1607-3274-2022-2-11

[10] S. Gavrylenko,. V. Chelak and O. Hornostal. Ensemble approach based on bagging and boosting for Identification the Computer System State, Proceedings of the 30th International Scientific Symposium Metrology and Metrology Assurance, Sozopol, Bulgaria, 2021. DOI: 10.1109/MMA52675.2021.9610949

[11] S. Gavrylenko and V. Chelak. Method of computer system state identificationbased on boosting ensemble with special preprocessing procedure, Advanced Information Systems, 2022, vol. 6, No 1, pp. 12–18. doi: DOI:10.20998/2522-9052.2022.1.02

[12] Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal and Asif Ekbal. A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis, Conference: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, January 2017, DOI:10.18653/v1/D17-1057

[13] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, vol. 45, pp. 503–528. DOI: 10.1007/BF01589116

[14] Léon Bottou and Olivier Bousquet. "The Tradeoffs of Large Scale Learning". In Sra, Suvrit; Nowozin, Sebastian; Wright, Stephen J. (eds.). Optimization for Machine Learning. Cambridge: MIT Press., 2012, pp. 351–368. ISBN 978-0-262-01646-9.

[15] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, 2014, DOI: 10.48550/ARXIV.1412.6980

[16] Gilles Louppe and Pierre Geurts. Ensembles on Random Patches, Conference: Proceedings of the 2012 European conference on Machine Learning and Knowledge, September, 2012, vol.1, pp 346–361. DOI:10.1007/978-3-642-33460-3_28

[17] Ibrahim Kovan. Wisdom of the Crowd -Voting Classifier, Bagging-Pasting, Random Forest and Extra Trees, Aug 13, 2021, URL: https://towardsdatascience.com/wisdom-of-the-crowd-voting-classifier-bagging-pasting-random-forest-and-extra-trees-289ef991e723

[18] Ye Tian and Yang Feng. Random Subspace Ensemble Classification, Journal of Machine Learning Research, 2021, vol.22, pp.1-93. DOI: 10.48550/arXiv.2006.08855