Problems of identification in information systems

UDC 004.732.056 doi: https://doi.org/10.20998/2522-9052.2023.4.01

Oleksii Hornostal, Svitlana Gavrylenko

National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, Ukraine

APPLICATION OF HETEROGENEOUS ENSEMBLES IN PROBLEMS OF COMPUTER SYSTEM STATE IDENTIFICATION

Abstract. **The object of the study is** the process of identifying anomalies in the operation of a computer system (CS). **The subject of the study is** ensemble methods for identifying the state of the CS. **The goal of the** study is to improve the performance of ensemble classifiers based on heterogeneous models**. Methods used:** machine learning methods, homogeneous and heterogeneous ensemble classifiers, Pasting and Bootstrapping technologies. **Results obtained:** a comparative analysis of the use of homogeneous and heterogeneous bagging ensembles in data classification problems was carried out. The effectiveness of various approaches to the selection of base ensemble classifiers has been studied. A method for identifying the state of a computer system, based on the heterogeneous bagging ensemble was proposed. Experimental studies made it possible to confirm the main theoretical assumptions, as well as evaluate the efficiency of the constructed heterogeneous ensembles. **Conclusions.** Based on the results of the study, the method for constructing a heterogeneous bagging ensemble classifier, which differs from known methods in the procedure for selecting base models was proposed. It made possible to increase the classification accuracy. Further development of this research could include the creating and integration of dissimilarity metrics as well as other quantitative metrics for a more accurate and balanced base model selection procedure, which would further improve the performance of the computer system state classifier.

Keywords: computer system; anomaly detection; machine learning; bagging; homogeneous ensembles; heterogeneous ensembles; decision trees; k-nearest neighbors method; multilayer perceptron neural network.

Introduction

The development of information technologies has led to the fact that information systems have become an integral part of most spheres of modern life. From the banking system that processes financial transactions to the medical databases that store the medical history of patients, information systems have permeated every aspect of our society. In this digital era, the efficiency and stability of such systems have become not just important, but also vital factors affecting safety, quality of service and even public health.

Failures and malfunctions in information systems can have catastrophic consequences. For example, in the financial sphere, even a small error in the processing of transactions can lead to serious losses and loss of customer trust. In the medical field, improper storage or transmission of medical data can endanger the health of patients. In the energy industry, automated control systems can affect the stability of the energy supply and, consequently, the vital functions of society [1].

Taking into account the above, ensuring the stability and security of information systems has become one of the priority directions in modern information technology. And it is here that ensembles of models, such as bagging, can play a decisive role, helping to identify anomalies, detect threats, and ensure more reliable functioning of information systems in various areas of human life.

In this context, the need for continuous improvement of methods and technologies used to monitor and identify the state of computer systems becomes obvious. At the same time, constantly changing security threats and the complexity of modern attacks require more advanced and adaptive methods for detecting and preventing all kinds of incidents.

Object, subject and methods of research. The main idea of the work is to study the possibility of using heterogeneous bagging ensembles to improve the accuracy of identifying the state of computer systems and the efficiency of anomaly detection. The object of the research is the process of identifying anomalies in the computer system operation. The subject of the research is ensemble methods for identifying the state of the computer system. The main objectives of the study are as follows:

1. Improving the accuracy of classification and anomaly detection.
2. Analysis of the impact of a variety of machine learning methods on the ensemble performance.
3. Comparison of the effectiveness of homogeneous and heterogeneous bagging ensembles.
4. Study of the effectiveness of strategies for selecting base models for the formation of a heterogeneous ensemble.
5. Development of a method for increasing the accuracy of identifying the state of computer systems through the use of heterogeneous ensembles.
6. Experimental testing of theoretical assumptions about the use of various machine learning methods as basic classifiers.
7. Formation of recommendations for the practical application of the obtained results.

The study is aimed at assessing the impact of the diversity of basic models and model selection strategies on the quality of ensemble performance in problems of classification of CS states in order to identify possible anomalies in their operation.

Statement of the research problem

We will assume that the functioning of the CS is characterized by a set of its indicators:

$$X = \{x_{i1}, x_{i2}, ..., x_{im}\} .$$

Marked pairs of objects are used as initial data for the given task:

$$\{(x_i, y_i\}_{i=1}^{N} ,$$

where x_i is an indicator of the CS state or training sample, y_i is a class label (normal or abnormal state).

There is an unknown "target dependence" - mapping $f : X \rightarrow Y$ whose value is known only on the objects of the final training sample $(X, Y) = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

It is necessary to form the structure of the ensemble classifier ***F,*** which is able to classify an arbitrary object $x \in X$ and adjust its parameter values w to bring the predicted value $\hat{y}$ closer to the actual value of y:

$$F(f(w, x), \hat{y}) = y.$$

This model should be able to predict $\hat{y}$ not only for objects from the training sample, but also for new objects.

Related works analysis

In classification problems, which include the task of identifying the computer system state, ensemble methods have proven themselves well. The ability to combine diverse models, evaluate their dissimilarity, and integrate predictions allows us to more accurately and reliably identify anomalies and respond to potential threats in a timely manner. This becomes critical because even small disruptions to information systems can have far-reaching consequences for business, health, safety and society as a whole.. In addition, we can highlight two main advantages of using ensembles in problems of identifying the computer system state:

- Increased reliability. Ensembles combine several base classifiers, which helps to increase the reliability of the system. Instead of using a single algorithm, which may make mistakes or, for example, may skew its predictions due to high noise levels, the ensemble uses multiple algorithms, reducing the likelihood of false positives and increasing overall accuracy.
- Improved generalization. Ensembles enable more accurate generalization of data. They reduce the tendency to overfit, which is especially important when working with large and complex data sets, which are often found in computer systems.

There are several well-proven subtypes of ensembles, but in our research we focused on bagging ensembles [2], as they have a number of undeniable advantages:

- Reducing Dispersion. Bagging (Bootstrap Aggregating) is based on the bootstrap principle, which creates several random subsamples from the original data set. This allows us to reduce the spread of the algorithm, since each base classifier is trained on different data. In the context of anomaly detection, where data can be noisy and variable, scatter reduction is especially useful.
- Reducing Correlation: Bagging also helps reduce the correlation between base classifiers. This is important because correlated algorithms may produce almost identical results and will not provide much benefit in an ensemble. Bagging helps to diversify base models by generating input sequences using a special bootstrap procedure when training each base model.
- Simple implementation. Bagging is a relatively simple method that does not require complex setup and is suitable for various types of base classifiers.

In general, the use of bagging in problems of identifying the computer system state makes it possible to increase the reliability and efficiency of an anomaly detection system, which is critically important in the face of constantly emerging digital threats.

Previous research has found that homogeneous bagging ensembles, such as those based on decision trees, are successful in detecting anomalies in computer systems [3]. However, these ensembles are composed of similar models, so they are often limited in their ability to improve their performance. This is because structurally similar basic models may make similar errors and may not provide enough diversity to effectively reduce dispersion.

In this regard, there is an assumption that the use of various base classifiers of the ensemble, as well as their combinations, can significantly improve the performance of the ensemble [4, 5]. This approach complements the basic idea of bagging, which is to use base models with high variance to create a more powerful ensemble.

In addition, research confirms that diversity in underlying models can significantly improve an ensemble's anomaly detection ability, as different methods can identify different characteristics of anomalous behavior [6]. This improves the stability and accuracy of the ensemble [7], making it more adaptive to changing conditions and new threats in computer systems.

Overview of approaches and methods

In the study of homogeneous and heterogeneous bagging ensembles, various combinations of the following machine learning methods were used: decision trees, k-nearest neighbors, support vector machines, naive Bayes classifier, logistic regression and multilayer perceptron. The choice of basic models is due to the variety of their advantages and disadvantages.

Decision trees are a graphical model designed for decision making. At each node of the tree, the data is divided into two or more subgroups based on the value of one of the features. Predictions are made based on leaves. The main advantage of decision trees is their easy interpretability and the ability to handle both categorical and numerical features. The limitation is their tendency to overfit [8].

The k-nearest neighbors (k-NN) method determines the class of a new object based on the classes of its nearest neighbors using a distance measure. The main advantage of k-NN is its simplicity in implementation and ability to work with different types of data. The limitation is the computational complexity with a large amount of data and the dependence of performance on the value of k [9].

The support vector machines (SVM) method constructs a hyperplane that best separates data classes by maximizing the distance to the nearest points of each class. The main advantage of SVM is its ability to process linearly separable and linearly inseparable data, as well as to generalize the results with new data. The limitations are computational complexity for large data volumes and the need to select parameters such as the kernel [10, 11].

A naive Bayesian classifier uses probabilistic methods to classify objects, assuming independence of features. The main advantage of the naive Bayesian classifier is its simplicity in implementation and the ability to process multidimensional data. The limitations include the assumption of independence of features and the potential unsuitability of the method for working with data with complex relationships [12, 13].

Logistic regression is a method widely used in binary and multi-class classification problems. The principle of its operation is to model the probability of an object belonging to a certain class based on a linear combination of its characteristics. This probability is then transformed using a logistic function (sigmoid), which constrains its values to be between 0 and 1. When training logistic regression, model parameters are tuned to maximize the likelihood of the data, allowing the model to accurately separate classes based on feature values. The main advantage of logistic regression is its simplicity and interpretability, which allows us to understand the influence of features on classification. However, the main limitation of this method is the assumption of a linear relationship between the features and the target variable, which may limit the model's ability to correctly describe complex nonlinear relationships [14, 15].

Multilayer Perceptron (MLP) is a multilayer neural network consisting of input, hidden and output layers that transmit signals taking into account weights and activation functions. The main advantage of MLP is its ability to model complex nonlinear relationships in data. The limitation is the need for a large amount of data and the risk of overfitting, as well as a long training time compared to other methods [16, 17].

The following characteristics, widely used in classification problems, were used as metrics for assessing the quality of work of ensemble classifiers: *Accuracy* (1), *Precision* (2), *Recall* (3) and *F1-Score* (4):

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN}; \quad (1)$$

$$Precision = \frac{TP}{TP+FP}; \quad (2)$$

$$Recall = \frac{TP}{TP + FN}; \quad (3)$$

$$F1\text{-}score = \frac{2}{\frac{1}{Pr\,e\,cision} + \frac{1}{Re\,c\,all}} = \frac{TP}{TP + 0.5(FP + FN)}. \quad (4)$$

These metrics allow you to evaluate both the overall performance of the classifier and its ability to find and classify positive examples (recall) and avoid false positives (precision). The F1-measure is the harmonic mean between precision and recall and is used when it is necessary to balance between these two characteristics.

Experimental part

To test the theoretical assumptions, software was developed to conduct an experiment consisting of four stages.

At the first stage of the study, a standard bagging ensemble was created using decision trees as basic models. Decision trees are well established for their ability to process different types of data and identify important features and complex patterns. They can be called a classic choice when building a bagging ensemble. After training this ensemble, its effectiveness was assessed.

The results of the first stage of the study confirmed previous studies and showed that standard bagging with decision trees demonstrates good performance in the tasks of identifying the computer system state and detecting anomalies, especially when using the Bootstrapping procedure when generating input data and with the optimal choice of the main parameters of the base models and bagging meta-algorithm. However, quality indicators demonstrated the need to improve classification efficiency.

The second stage of the study includes the construction of homogeneous bagging ensembles using various machine learning methods as base classifiers. The following basic classifiers were used: the k-nearest neighbors method, the support vector machine, several subtypes of the naive Bayes classifier, logistic regression as well as a multilayer perceptron. Each ensemble consisted of similar models.

The process of constructing a homogeneous bagging ensemble using various machine learning methods as base models is as follows:

1. Selecting the type of base ensemble classifier, such as decision trees, logistic regression, etc.
2. Selecting an algorithm and generating initial data samples for each basic classifier. This study used the Bootstrapping algorithm, in which the samples contain all the original features, are generated randomly and can be repeated.
3. Train base ensemble classifiers in parallel using different data samples obtained in step 2.
4. Aggregation of results obtained from base models. In the case of a classification problem, majority rule voting is used to determine the most popular class.
5. Evaluate the performance of the model using the quality metrics, as well as the time it takes to train and test the model.

It is important to note that the resulting settings for the parameters of the bagging meta-algorithm (the number of base models, methods of aggregating results) remain unchanged when constructing all models, which further allows for a more accurate assessment of their quality.

The results of the second stage of the study are presented in fig. 1–6. The results obtained show that the use of different models has different effects on the classification quality. Thus, models showing low accuracy when working in an ensemble were further excluded from further research.

Based on the results of the second stage of the study, it was decided to use 5 methods for further research. For example, homogeneous ensembles based on a multilayer perceptron (MLP) and the k-nearest neighbors method (KNN) demonstrated the best quality of work. Support vector machines and logistic regression provide less but good accuracy. In addition, previous studies have shown that by selecting optimal tuning parameters, decision trees also have the potential to improve the accuracy of ensemble performance. Using different variations of the Naive Bayes classifier does not lead to a significant increase in accuracy during ensemble. Thus, the results obtained emphasize the importance of choosing basic classifier models when constructing ensembles and adjusting their parameters when using specific initial data.

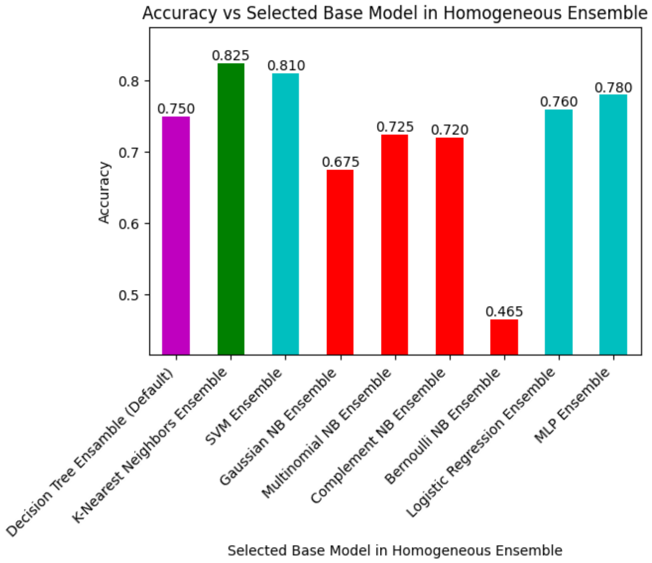

Fig. 1. Comparison of accuracy of homogeneous bagging ensembles

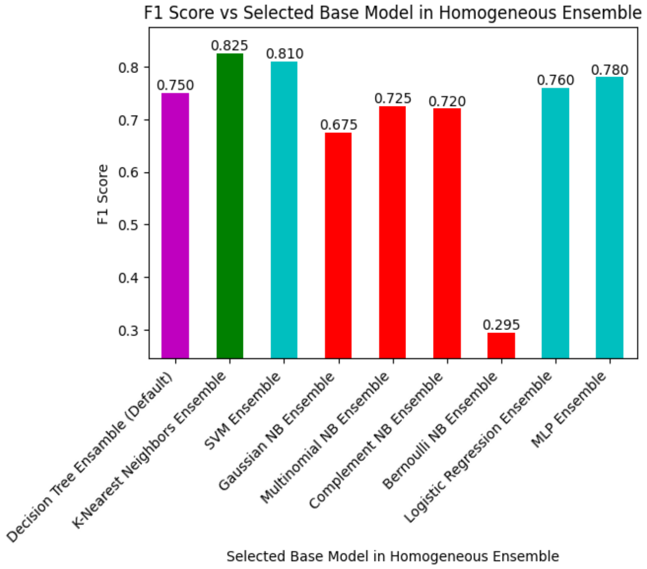

Fig. 2. Comparison of the F1 Score metric of homogeneous bagging ensembles

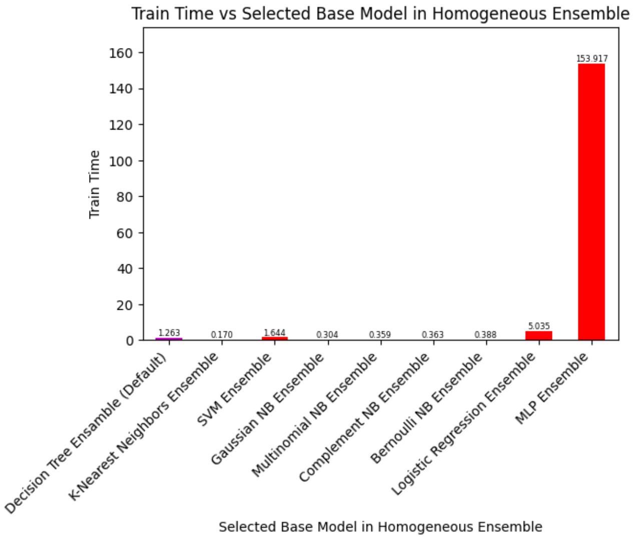

Fig. 3. Comparison of training time of homogeneous bagging ensembles

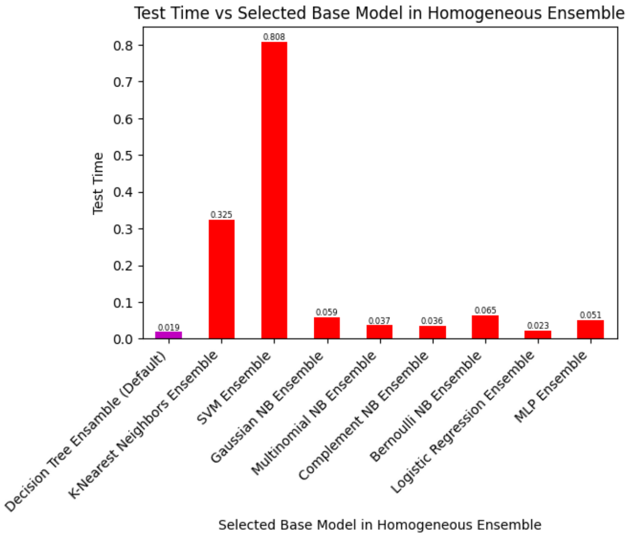

Fig. 4. Comparison of identification time on a test sample of homogeneous bagging ensembles

The third stage of research includes the selection of the most effective basic models' types for combining them into a heterogeneous ensemble. At the same time, the procedure of two-stage selection of basic models and Pasting technology was used when choosing the basic classifiers of the ensemble. This approach allowed us to assess how a variety of base models can improve ensemble performance.

At the first stage, the previously selected methods were taken and 5 different homogeneous ensembles were trained on their basis. Each ensemble included 100 models of the same type. At the second stage, pairs were created that included combinations of all homogeneous ensembles. From each pair, its trained classifiers were taken and placed in the classifier pool. Each pool contained 200 models. Using Pasting technology (random selection without repetitions), 100 classifiers were selected from each pool and combined into a new ensemble. For each ensemble with a pair of methods, quality metrics were calculated, and training and testing times were estimated. Next, similar actions were performed with the construction of heterogeneous ensembles based on three types of basic classifiers for all possible combinations.

The results of studying heterogeneous ensembles using two and three types of different basic models are presented in Fig. 5–12.

The best results have been achieved using k-nearest neighbors (KNN), multilayer perceptrons (MLP), and decision trees (DT) as base models for ensembles. It was found that heterogeneous ensembles based on their combination can improve the classification accuracy to 9.5%

The fourth stage of the study includes the construction of a heterogeneous ensemble using all considered machine learning methods. As a result, it was found that such an ensemble does not allow obtaining the desired increase in accuracy, but it works faster than ensembles based on a smaller number of machine learning methods, since some of them significantly increase the classification time on the test sample.

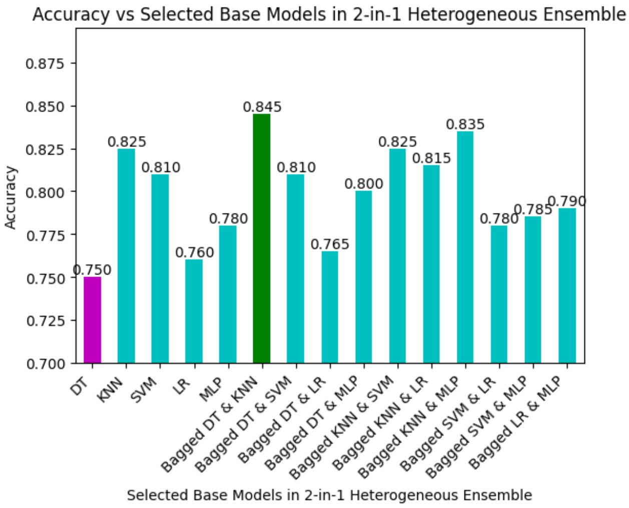

Fig. 5. Comparison of the accuracy of homogeneous and heterogeneous bagging ensembles (with 2 methods)

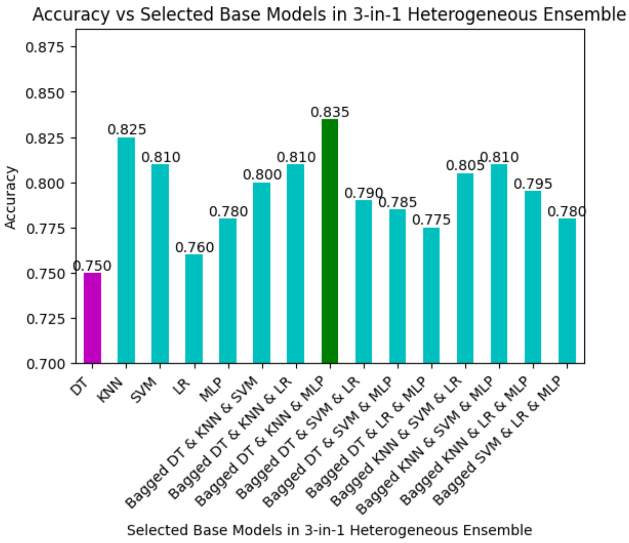

Fig. 6. Comparison of the accuracy of homogeneous and heterogeneous bagging ensembles (with 3 methods)

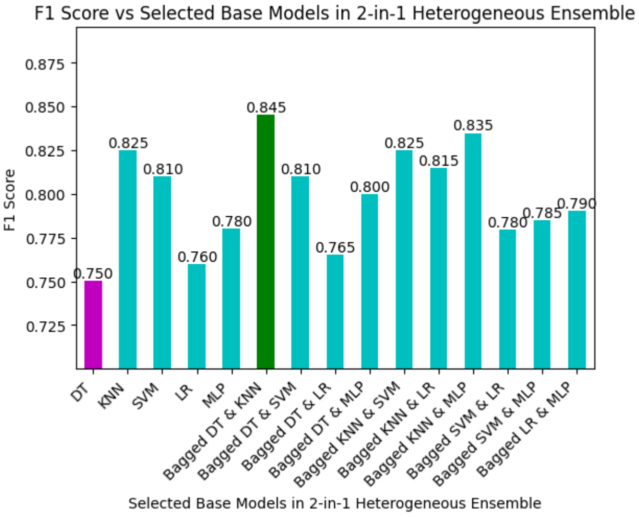

Fig. 7. Comparison of the F1 Score metric of heterogeneous bagging ensembles (with 2 methods)

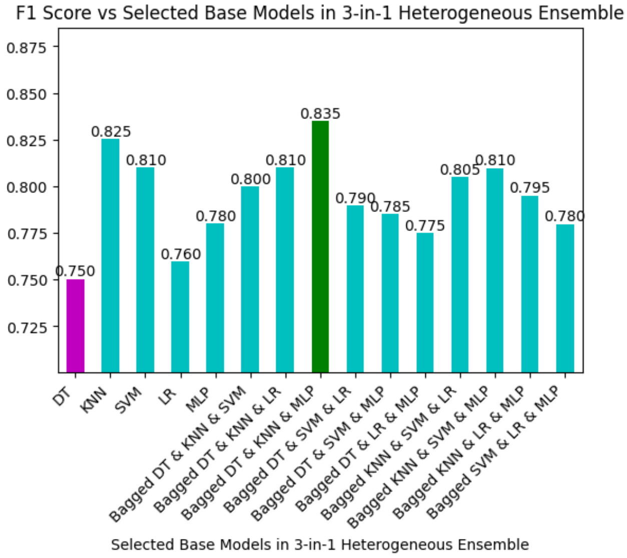

Fig. 8. Comparison of the F1 Score metric of heterogeneous bagging ensembles (with 3 methods)

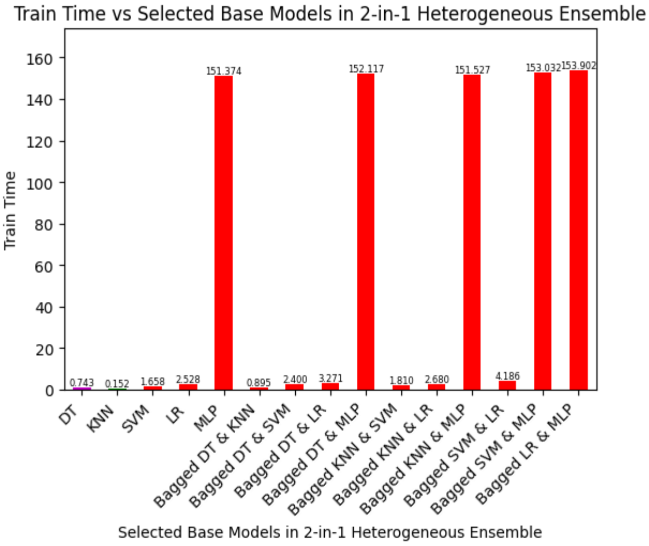

Fig. 9. Comparison of training time of heterogeneous bagging ensembles (with 2 methods)

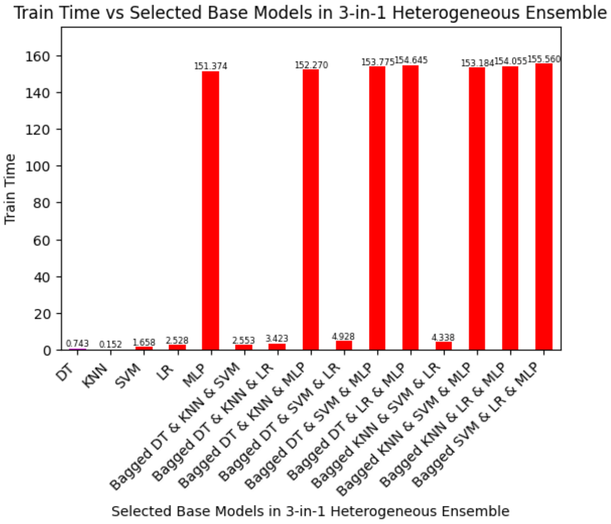

Fig. 10. Comparison of training time of heterogeneous bagging ensembles (with 3 methods)

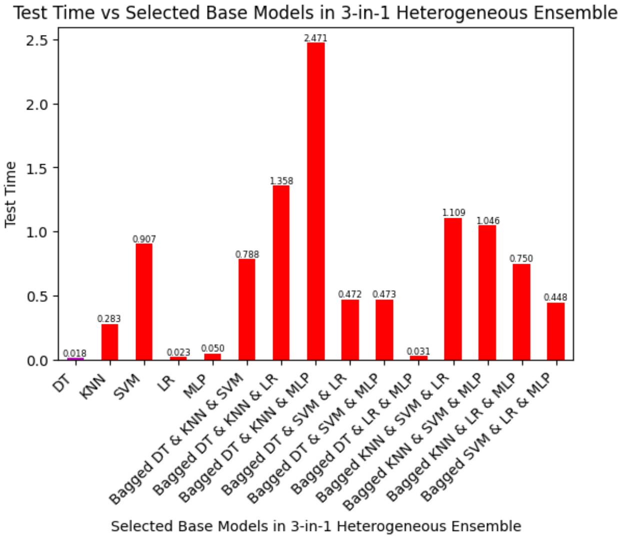

Fig. 11. Comparison of identification time on a test sample of heterogeneous bagging ensembles (with 2 methods)

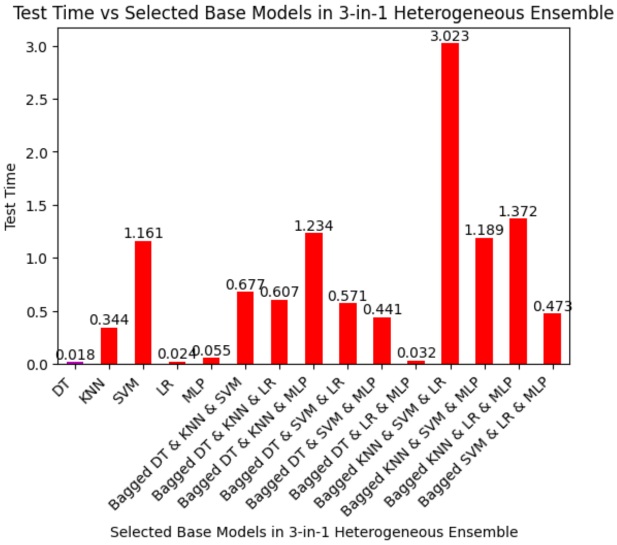

Fig. 12. Comparison of identification time on a test sample of heterogeneous bagging ensembles (with 3 methods)

The results of studying a heterogeneous ensemble using all the considered methods are presented in Fig. 13–16.

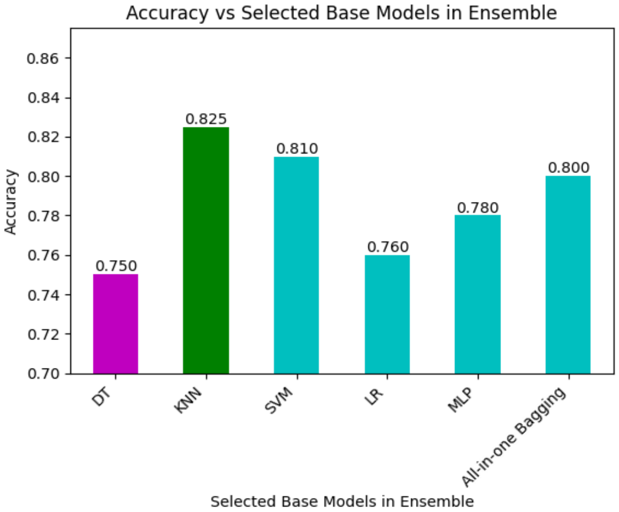

Fig. 13. Comparison of the accuracy of homogeneous and heterogeneous bagging ensembles

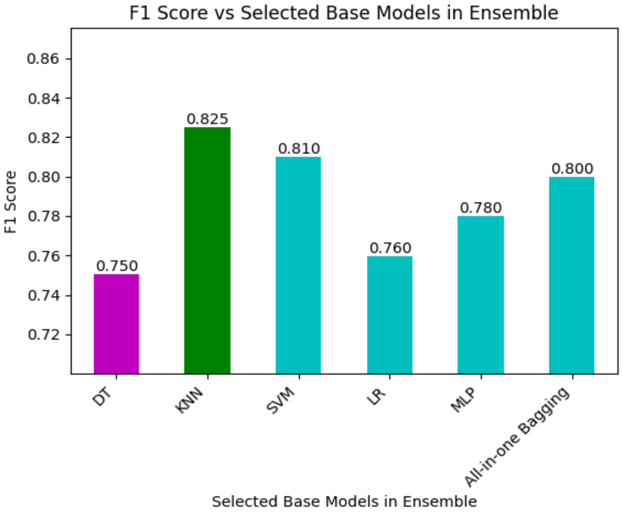

Fig. 14. Comparison of the F1 Score metric of homogeneous and heterogeneous bagging ensembles

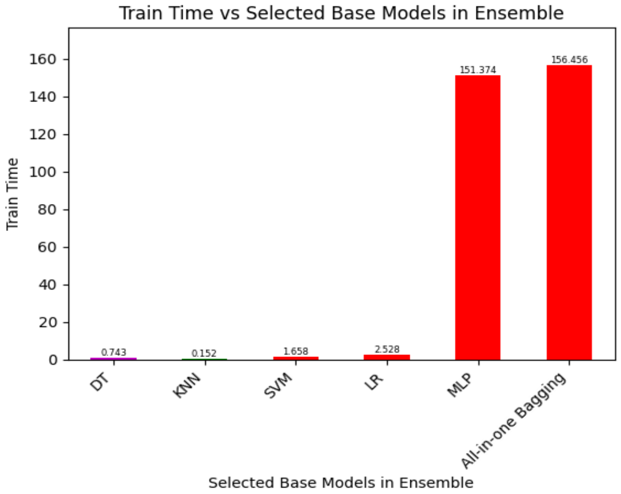

Fig. 15. Comparison of training time of homogeneous and heterogeneous bagging ensembles

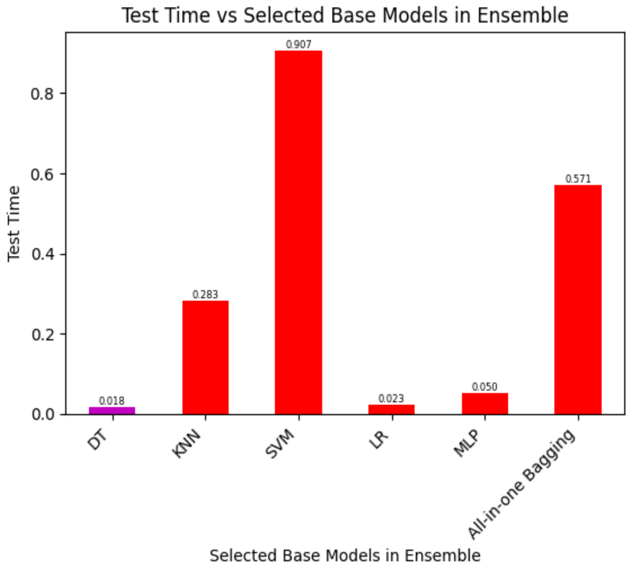

Fig. 16. Comparison of identification time on a test sample of homogeneous and heterogeneous bagging ensembles

Based on the study, we can conclude that the use of heterogeneous bagging ensembles can improve the accuracy of identifying the computer system state and detecting anomalies.

Heterogeneous ensembles combining diverse base models have shown high potential for performance improvements over homogeneous ensembles. These results confirm the value of using ensemble methods in the task of identifying the state of computer systems and highlight the importance of choosing a variety of models to create more effective monitoring and anomaly detection systems.

Conclusions

This paper examines the effectiveness of using homogeneous and heterogeneous bagging classifiers in the context of identifying the computer system state and detecting anomalies.

The results of the study showed that the use of heterogeneous ensembles can improve classification accuracy in these tasks.

Such ensembles combine different types of models or algorithms. This helps to increase the diversity of forecasts. Different models may have different generalization abilities, and in certain situations one model may make more accurate predictions than another. By combining these models into an ensemble, the risk of overfitting can be reduced and generalization ability can be improved..

During the experiment, it was revealed that the greatest accuracy was obtained when constructing a bagging ensemble, which included models based on decision trees and the k-nearest neighbors' method as base classifiers.

The use of a bagging ensemble based on these methods makes it possible to increase the accuracy of the model on a test sample by up to 9.5% in comparison with a standard homogeneous bagging ensemble based on decision trees.

Bagging ensembles based on a combination with multilayer perceptrons also have relatively high accuracy, however, their use leads to an increase in classification time on the test set.

The influence of this negative factor can be neutralized in the future by using the ensemble pruning technique. Combining other methods either leads to a significant increase in testing time or does not provide the desired increase in classification accuracy.

Thus, based on the results of the study, a method for identifying the computer system state has been proposed, which differs from known methods by using a heterogeneous bagging meta-algorithm and includes a two-stage selection process for base classifier models based on Pasting technology. The use of this method made it possible to increase the classification accuracy.

A promising direction for further research is the creation and integration of various metrics that assess the diversity of models and other quantitative indicators characterizing the basic models for the purpose of their more accurate and balanced selection. These steps will further increase the efficiency of computer system state classification. In addition, it is important to pay attention to speeding up the classification process on the test set, for example, using ensemble pruning technology.

References

1. Gavrylenko, S. and Chelak, V. (2023), "Development of method base on fuzzy decision trees for identification of the computer systems state", *Control, Navigation and Communication Systems,* Is. 1(71), pp. 78–83, doi: https://doi.org/10.26906/SUNZ.2023.1.078
2. Hornostal, O. and Gavrylenko, S. (2021), "Development of a method for identification of the state of computer systems based on bagging classifiers", *Advanced Information Systems*, Vol. 5, no. 4, pp. 5–9, doi: https://doi.org/10.20998/2522-9052.2021.4.01
3. Hornostal, O. and Gavrylenko, S. (2023), "Method of identifying the state of a computer system based on ensemble classifiers with an improved voting procedure", *Control, Navigation and Communication Systems*, Is. 3(73), pp. 79–85, doi: https://doi.org/10.26906/SUNZ.2023.3.079
4. Kuo-Wei, Hsu (2017), "A Theoretical Analysis of Why Hybrid Ensembles Work", *Computational Intelligence and Neuroscience*, Vol. 2017(1), pp. 1–12, doi: https://doi.org/10.1155/2017/1930702
5. Nascimento, D., Canuto, A., Silva, L. and Coelho A. (2011), "Combining different ways to generate diversity in bagging models: An evolutionary approach", *Proceedings of the International Joint Conference on Neural Networks*, pp. 2235–2242, doi: https://doi.org/10.1109/IJCNN.2011.6033507
6. Feng, Y., Wang, X. and Zhang, J. (2021), "A Heterogeneous Ensemble Learning Method For Neuroblastoma Survival Prediction", *IEEE Journal of Biomedical and Health Informatics*, Vol. 26, pp. 1472–1483, doi: https://doi.org/10.1109/JBHI.2021.3073056

7. Khreich, W., Murtaza, S. Sh., Hamou-Lhadj, A. and Talhi, C. (2018), “Combining Heterogeneous Anomaly Detectors for Improved Software Security”, *Journal of Systems and Software*, vol. 137, pp. 415–429, doi: https://doi.org/10.1016/j.jss.2017.02.050
8. Gavrylenko, S., Chelak, V. and Hornostal, O. (2022), “Construction Method of Fuzzy Decision Trees for Identification the Computer System State”, *Proceedings of the 32th International Scientific Symposium Metrology and Metrology Assurance*, pp. 1–5, doi: https://doi.org/10.1109/MMA55579.2022.9992878
9. Bicego, M., Rossetto, A., Olivieri, M., Londoño-Bonilla, J. and Orozco-Alzate, M. (2022), “Advanced KNN Approaches for Explainable Seismic-Volcanic Signal Classification”, *Mathematical Geosciences*, Vol. 55, pp. 59–80. doi: https://doi.org/10.1007/s11004-022-10026-w
10. Hlavcheva, D., Yaloveha, V., Podorozhniak, A. and Kuchuk, H. (2021), “Comparison of CNNs for Lung Biopsy Images Classification”, *2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering, UKRCON 2021 – Proceedings*, pp. 1–5, doi: https://doi.org/10.1109/UKRCON53503.2021.9575305
11. Khreich, W., Khosravifar, B., Hamou-Lhadj, A. and Talhi, C. (2017), “An Anomaly Detection System based on Variable N-gram Features and One-Class SVM”, *Information and Software Technology*, Vol. 91, pp. 186–197. doi: https://doi.org/10.1016/j.infsof.2017.07.009
12. Kuchuk, N., Mozhaiev, O., Mozhaiev, M. Kuchuk, H. (2017), “Method for calculating of R-learning traffic peakedness”, *2017 4th International Scientific-Practical Conference Problems of Infocommunications Science and Technology*, PIC S and T 2017 – Proceedings, pp. 359–362, doi: https://doi.org/10.1109/INFOCOMMST.2017.8246416
13. Salau A. O., Assegie T. A., Akindadelo A. T. and Eneh, J. N. (2023), “Evaluation of Bernoulli Naive Bayes model for detection of distributed denial of service attacks”, *Bulletin of Electrical Engineering and Informatics*, Vol. 12, no. 2, pp. 1203–1208, doi: https://doi.org/10.11591/eei.v12i2.4020
14. Kovalenko, A., Kuchuk, H., Kuchuk, N. and Kostolny, J. (2021), “Horizontal scaling method for a hyperconverged network”, *2021 Int. Conf. on Inf. and Digital Technologies (IDT)*, Zilina, Slovakia, doi: https://doi.org/10.1109/IDT52577.2021.9497534
15. Kamarudin, M. H., Maple, C., Watson, T. and Sofian, H. (2015), “Packet Header Intrusion Detection with Binary Logistic Regression Approach in Detecting R2L and U2R Attacks”, *Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic (CyberSec)*, pp. 101–106, doi: http://dx.doi.org/10.1109/CyberSec.2015.28
16. Dun B., Zakovorotnyi, O. and Kuchuk, N. (2023), “Generating currency exchange rate data based on Quant-Gan model”, *Advanced Information Systems*, Vol. 7, no. 2, pp. 68–74, doi: http://dx.doi.org/10.20998/2522-9052.2023.2.10
17. Paul, S. and Kundu, R. K. (2022), “A Bagging MLP-based Autoencoder for Detection of False Data Injection Attack in Smart Grid”, *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, doi: https://doi.org/10.1109/ISGT50606.2022.9817480

Надійшла (received) 12.09.2023
Прийнята до друку (accepted for publication) 15.11.2023

Відомості про авторів / About the authors

Горносталь Олексій Андрійович – аспірант кафедри "Комп'ютерна інженерія та програмування", Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна;
Oleksii Hornostal – PhD Student of Department of "Computer Engineering and Programming", National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gornostalaa@gmail.com; ORCID ID: https://orcid.org/0000-0001-5820-9999.

Гавриленко Світлана Юріївна – доктор технічних наук, професорка, професорка кафедри "Комп'ютерна інженерія та програмування", Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна;
Svitlana Gavrylenko – Doctor of Technical Science, Professor, Professor of Department of "Computer Engineering and Programming", National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: gavrilenko08@gmail.com; ORCID ID: https://orcid.org/0000-0002-6919-0055.

Застосування гетерогенних ансамблів у задачах ідентифікації стану комп'ютерних систем

О. А. Горносталь С. Ю. Гавриленко

Анотація. Об'єктом дослідження є виявлення аномалій у роботі комп'ютерної системи. **Предметом дослідження** є ансамблеві методи ідентифікації стану КС. **Метою дослідження** є підвищення продуктивності ансамблевих класифікаторів на основі гетерогенних моделей. **Методи, що використовуються:** методи машинного навчання, гомогенні та гетерогенні ансамблеві класифікатори, технології Pasting та Bootstrapping. **Отримані результати:** проведено порівняльний аналіз використання гомогенних та гетерогенних беггінг ансамблів у задачах класифікації даних. Досліджено ефективність різних підходів щодо вибору базових класифікаторів ансамблю. Запропоновано метод ідентифікації стану комп'ютерної системи на основі гетерогенного беггінг ансамблю. Експериментальні дослідження дозволили підтвердити основні теоретичні припущення та оцінити ефективність роботи побудованих гетерогенних ансамблів. **Висновки.** За результатами дослідження запропоновано метод побудови гетерогенного ансамблевого класифікатора, який відрізняється від відомих методів процедурою вибору базових моделей. Це дозволило підвищити точність класифікації. Подальший розвиток цього дослідження може включати розробку та інтеграцію метрик несхожості, а також інших кількісних метрик для більш точної та збалансованої процедури відбору базових моделей, що сприятиме подальшому підвищенню ефективності роботи класифікатора стану комп'ютерної системи.

Ключові слова: комп'ютерна система; виявлення аномалії; машинне навчання; беггінг; однорідні ансамблі; різнорідні ансамблі; дерева рішень; метод k-найближчих сусідів; багатошарова нейронна мережа перцептрона.