https://doi.org/10.30857/2786-5371.2025.6.3

Received: 08.10.2025
Revised: 05.12.2025
Accepted: 23.12.2025

УДК 004.8+004.6

Vladyslav PYLYPENKO
Kyiv National University of Technologies and Design, Ukraine

THE EFFECT OF TRAINING SAMPLE SIZE ON THE STABILITY OF CLASSIFICATION MODELS

Purpose. *The research is aimed at a comprehensive analysis of the impact of training sample size on classification model stability and determining optimal strategies for selecting sample size for different types of machine learning algorithms. The goal of the work is to develop a methodology for assessing model stability depending on the volume of training data and to determine recommendations for selecting optimal sample size to achieve high stability and generalization ability of classification models.*

Methodology. *The research methodology is based on experimental analysis of performance and stability of different types of classification models (logistic regression, Random Forest, Gradient Boosting, neural networks) when trained on samples of different sizes (from 100 to 10000 examples). Model stability assessment is performed using metrics of coefficient of variation of accuracy, variance of accuracy, and interquartile range when training models multiple times on different data subsets. Methods of learning curve analysis are applied to determine saturation points, assess model complexity, and progressive cross-validation. The effectiveness of stability improvement methods is investigated, including data augmentation techniques, regularization, and ensemble methods..*

Findings. *Experimental results demonstrate a significant dependence of classification model stability on training sample size. For simple linear models (logistic regression), stable performance is achieved at a sample size of approximately 2000-3000 examples, while for complex models (neural networks) 5000-10000 examples are required. The coefficient of variation of accuracy decreases with increasing sample size: for logistic regression from 0.25 to 0.08, for Random Forest from 0.18 to 0.05, for Gradient Boosting from 0.15 to 0.04, for neural networks from 0.22 to 0.06. It is found that data augmentation techniques allow reducing the coefficient of variation by 52-68% at small sample sizes, and ensemble methods provide stability with a coefficient of variation less than 0.05 even for samples of 500 examples. The impact of class imbalance and feature space dimensionality on model stability is established, which requires correction of optimal sample size.*

Originality. *A comprehensive methodology for assessing classification model stability depending on training sample size is developed, including theoretical analysis of the relationship between sample size and variance component of generalization error, empirical methods for determining saturation points, and comparative analysis of the effectiveness of different stability improvement methods. The impact of class imbalance and feature space dimensionality on the relationship between sample size and model stability is systematically investigated for the first time. A classification of models by dependence on training sample size is developed, taking into account algorithm type, model complexity, and data nature..*

Practical value. *The obtained results allow justifying the choice of optimal training sample size for a specific classification task depending on algorithm type, model complexity, and data nature. The developed recommendations can be applied in various fields where high stability of classification models is required, including medical diagnostics, financial analysis, cybersecurity, and image processing. The methodology for determining optimal sample size allows optimizing the use of computational resources and ensuring high reliability of classification results under conditions of limited training data.*

Keywords*: training sample size; model stability; classification; machine learning; learning curve; machine learning algorithms; ensemble methods; Python.*

Introduction. Machine learning research is increasingly faced with limited access to high-quality, structured, and large enough samples, which makes it difficult to build effective predictive models. In practical areas such as medicine, finance, cybersecurity, or image processing, data is often fragmented, noisy, or limited in size due to ethical, financial, or technical constraints [5, 7]. In this context, the search for methods that can provide high prediction accuracy even under data scarcity

conditions is becoming more urgent, as well as determining the optimal training sample size to achieve stable performance of classification models.

The training sample size, as one of the key factors affecting the stability and generalizability of models, requires careful analysis for different types of classifiers. According to the study by T. Hastie et al. [5], the relationship between sample size and model performance can be described by learning curves that demonstrate a decreasing increase in accuracy with increasing data volume. G. James et al. [7] showed that for simple models such as logistic regression, stable performance is achieved with much smaller sample sizes compared to complex models such as neural networks. M. Kuhn & K. Johnson [9] noted the importance of considering the complexity of the model and the dimensionality of the feature space when determining the optimal training sample size.

Research by M. Belkin et al. [1] demonstrated that modern approaches to machine learning require a rethinking of the classical bias-variance trade-off when determining the optimal training sample size. The authors showed that for deep neural networks, a double descent phenomenon is observed, where increasing the model size first worsens and then improves performance. P. Nakkiran et al. [10] extended this observation by showing that double descent can also occur when increasing the training sample size, which is important for determining the optimal amount of data.

Research in data augmentation techniques has shown the effectiveness of increasing the effective training sample size to improve model stability. E.D. Cubuk et al. [3] developed the AutoAugment method, which automatically determines optimal augmentation strategies based on data, which allows to significantly improve the stability of models with small training sample sizes. L. Perez & J. Wang [12] showed the effectiveness of data augmentation in image classification tasks, where augmentation techniques allowed to achieve stable performance even with limited training data. C. Shorten & T.M. Khoshgoftaar [14] conducted a comprehensive review of data augmentation methods for deep learning, emphasizing their importance for improving model stability. Despite numerous advantages, current research has identified a number of problems associated with determining the optimal training sample size. C. Zhang et al. [15] noted that understanding generalization in deep learning requires rethinking traditional approaches to assessing the relationship between sample size and model performance. J. Hoffmann et al. [6] showed that determining the optimal training sample size for large language models requires taking into account computational resources and training efficiency, which adds complexity to the optimization problem.

Thus, determining the optimal training sample size at the current stage of machine learning development is considered not only as a technical task, but also as a basis for ensuring the stability and reliability of classification models in conditions of limited data. The relevance of the topic is due not only to scientific interest, but also to the practical need for stable and universal algorithms that can function effectively in real conditions with different amounts of training data. The aim of the study was to comprehensively analyze the impact of the training sample size on the stability of classification models with an emphasis on their adaptability to working with limited data sets. The objectives of the study were: to assess the effectiveness of different types of classifiers (logistic regression, Random Forest, Gradient Boosting, neural networks) depending on the size of the training sample; to determine optimal strategies for choosing the sample size for different types of models; assessing the effectiveness of methods for increasing stability with small training sample sizes; developing recommendations for improving model performance in specific conditions.

Materials and Methods. The stability of a classification model is a fundamental characteristic that determines the reliability and generalizability of the model when working with new data. The stability of the model depends on the size of the training sample and can be quantified through various metrics that characterize the variability of the model's performance when training on different subsets of the data [5, 7].

The coefficient of variation of accuracy is one of the main metrics for assessing the stability of a classification model. It is calculated as the ratio of the standard deviation of accuracy to the average accuracy:

$$CV\_accuracy = \sigma\_accuracy / \mu\_accuracy, \quad (1)$$

where $\sigma\_accuracy = \sqrt{[(1/K) \times \Sigma(Accuracy\_i - \mu\_accuracy)^2]}$

This is the standard deviation of accuracy when training on K different subsets of data, and μ_accuracy = (1/K) × ΣAccuracy_i is the average accuracy. Smaller values of the coefficient of variation indicate greater stability of the model, since they reflect less variability in performance when changing the training sample [4]. For stable models, the coefficient of variation usually does not exceed 0.10, while for unstable models it can reach values of 0.30 and more.

The dependence of model accuracy on the size of the training sample can be described through empirical learning curves, which demonstrate a decreasing increase in accuracy with increasing training data. For many classification models, accuracy increases with sample size according to an exponential law:

$$Accuracy(n) = A\_max - B \times exp(-C \times n), \quad (2)$$

where *A_max* is the maximum achievable accuracy with infinite sample size;

B = A_max – A_0 is the initial deviation from the maximum (where *A_0* is the accuracy with minimum sample size);

C is a convergence rate parameter that determines how quickly the model reaches maximum accuracy, and *n* is the size of the training sample.

The parameter C depends on the complexity of the model, the dimensionality of the feature space, and the nature of the data distribution. For simple models, C usually has larger values, indicating faster convergence, while for complex models, C values are smaller, reflecting slower convergence to maximum accuracy [14]. The stability of a model can also be assessed through the variance of accuracy when training on different subsets of the data, which reflects the internal variability of the model. The variance of accuracy is calculated as:

$$Var(Accuracy) = (1/K) \times \Sigma(Accuracy\_i - \mu\_accuracy)^2, \quad (3)$$

where *v* – linear velocity (m/s);

where *K* is the number of data subsets used to assess stability;

Accuracy_i is the model accuracy on the i-th subset, and μ_accuracy is the average accuracy. Smaller variance values indicate greater model stability, as they reflect less spread in accuracy values when the training sample changes [12]. The accuracy variance is closely related to the coefficient of variation through the relationship CV_accuracy = √Var(Accuracy) / μ_accuracy.

Theoretical analysis of the dependence of stability on sample size. According to the theory of statistical learning, the generalization error of a classification model can be decomposed into three components: bias, variance, and irreducible error. The relationship between the training sample size and model stability is closely related to the variance component of the generalization error [1, 5].
For a classification model with parameters θ trained on a sample of size n, the expected generalization error can be represented as:

$$E[Error] = Bias^2(\theta) + Var(\theta) + \sigma^2, \quad (4)$$

where $Bias^2(\theta) = E[(E[\hat{y}] - y\_true)^2]$ is the square of the bias, reflecting the systematic error of the model;

$\mathrm{Var}(\theta) = E[(\hat{y} - E[\hat{y}])^2]$ is the variance of the predictions, reflecting the sensitivity of the model to changes in the training sample, and σ^2 is the irreducible error associated with randomness in the data.

The variance component of $\mathrm{Var}(\theta)$ decreases with increasing training sample size, which leads to increased model stability [1]. For linear classification models, the variance of parameter estimates is inversely proportional to the size of the training sample:

$$Var(\theta) \approx \sigma^2 / (n \times I(\theta)), \quad (5)$$

where σ^2 is the error variance, *n* is the training sample size;
$I(\theta)$ is the Fisher information, which depends on the data distribution and model structure.

This relationship shows that increasing the sample size leads to a decrease in the variance of the parameter estimates, which contributes to the increase in the stability of the model [5]. For complex nonlinear models, such as neural networks or ensemble methods, the relationship between sample size and stability can be more complex. According to the study by M. Belkin et al. [1], for deep neural networks, a double descent phenomenon is observed, when the variance first increases and then decreases with increasing model or training sample size. This phenomenon requires a rethinking of the traditional trade-off between bias and variance when determining the optimal training sample size.

Methods for estimating the minimum sufficient training sample size. Determining the minimum sufficient training sample size to achieve a given level of stability is a critical task in building classification models. There are several approaches to solving this problem, each of which has its own advantages and limitations [2, 8]. The learning curve analysis method is based on constructing empirical learning curves for different sizes of training samples and determining the saturation point, where further increase in the sample size does not lead to a significant improvement in stability. The saturation point is defined as the minimum sample size n_sat for which the condition is satisfied:

$$|Accuracy(n\_sat) - Accuracy(n\_sat + \Delta n)| / Accuracy(n\_sat) < \varepsilon, \quad (6)$$

where Δn is the sample size increment and ε is the threshold value of relative improvement (typically 0.01-0.05). The learning curve analysis method allows empirically determining the optimal sample size, but can be time-consuming for large models due to the need for multiple training of models on different sample sizes [8].

The model complexity estimation method is based on a theoretical estimate of the required sample size based on the model complexity and the dimensionality of the feature space. For models with parameters θ of dimension d, the minimum sufficient sample size can be estimated as:

$$n\_min \approx (d \times log(1/\delta)) / \varepsilon^2, \quad (7)$$

where d is the dimension of the parameter space;
δ is the confidence level, and ε is the desired estimation accuracy. This relationship shows that for complex models with a large number of parameters, larger training sample sizes are required to achieve a given level of stability [2]. However, for nonlinear models, estimating complexity can be difficult because of the need to determine the effective dimension of the parameter space.

The progressive cross-validation method allows us to estimate the stability of a model at different training sample sizes without having to fully train the model for each size. The method is based on sequentially adding new examples to the training sample and assessing the change in the stability of the model. The stopping criterion is defined as the moment when the coefficient of variation of the accuracy reaches a given threshold value *CV_threshold*:

$$CV\_accuracy(n) \leq CV\_threshold, \quad (8)$$

where *CV_threshold* is usually set in the range of 0.05–0.10 for high stability.

This method allows to effectively determine the optimal sample size, but can be computationally expensive for large models [9].

Results. To quantitatively assess the effect of training sample size on the stability of classification models, a comprehensive experimental study was conducted on a set of datasets of different complexity and nature. The main study was conducted on a dataset of classification of academic performance of applicants, which contains 10 classes with a uniform distribution of examples in each class (6000 examples per class). The experimental methodology involved training the models on samples of different sizes (from 100 to 10,000 examples with a step of 100 for small samples and a step of 500 for large samples) and evaluating their performance and stability using 10-fold stratified cross-validation. For each model and each sample size, 10 independent experiments were conducted with different data partitions to assess the variability of the results. The stability of the models was assessed through the coefficient of variation of accuracy, the variance of accuracy, the interquantile range (IQR) and the standard deviation of the F1-measure. The experimental results showed that the stability of classification models significantly depends on the size of the training sample, and the nature of this dependence differs for different types of models. For logistic regression, the coefficient of variation of accuracy decreased from 0.25 at a sample size of 100 examples to 0.08 at a sample size of 5000 examples, reflecting a significant increase in stability as the amount of training data increases. Analysis of variance of accuracy showed that for logistic regression, the variance decreased from 0.0625 to 0.0064, corresponding to a decrease in the standard deviation of accuracy from 0.25 to 0.08. For Random Forest, the coefficient of variation of accuracy decreased from 0.18 to 0.05, indicating greater stability compared to logistic regression at small sample sizes. For Gradient Boosting, the coefficient of variation decreased from 0.15 to 0.04, demonstrating the highest stability among tree-based algorithms. For neural networks with three hidden layer architectures (128, 64, 32 neurons respectively), the coefficient of variation of accuracy decreased from 0.22 to 0.06, reflecting the need for larger sample sizes to achieve stable performance compared to simple models. Comparative results of experimental studies are presented in Fig. 1.

Detailed analysis of learning curves and saturation points. Analysis of empirical learning curves for different types of classification models showed significant differences in the relationship between training sample size and stability of performance. For simple linear models such as logistic regression, stable performance (coefficient of variation less than 0.10) is achieved with a sample size of about 2000–3000 examples. The analysis of the parameters of the exponential model (2) for logistic regression showed the values A_max = 0.85, B = 0.20, C = 0.0015, indicating a relatively fast convergence to maximum accuracy. The saturation point, determined by criterion (6) with ε = 0.02, was n_sat = 2800 examples, which means that further increase in the sample size beyond this point leads to a slight improvement in stability (less than 2% relative improvement).

For complex nonlinear models, such as deep neural networks, stable performance is achieved with much larger sample sizes. For a neural network with three hidden layers, the saturation point was n_sat = 7500 examples, reflecting the need for larger amounts of training data to achieve stable performance. The exponential model parameters for the neural network had the values A_max = 0.92, B = 0.25, C = 0.0008, indicating slower convergence compared to simple models, but higher maximum achievable accuracy. Analysis of the variance of accuracy showed that for neural networks, the variance decreases more slowly with increasing sample size, reflecting greater sensitivity to changes in the composition of the training data.

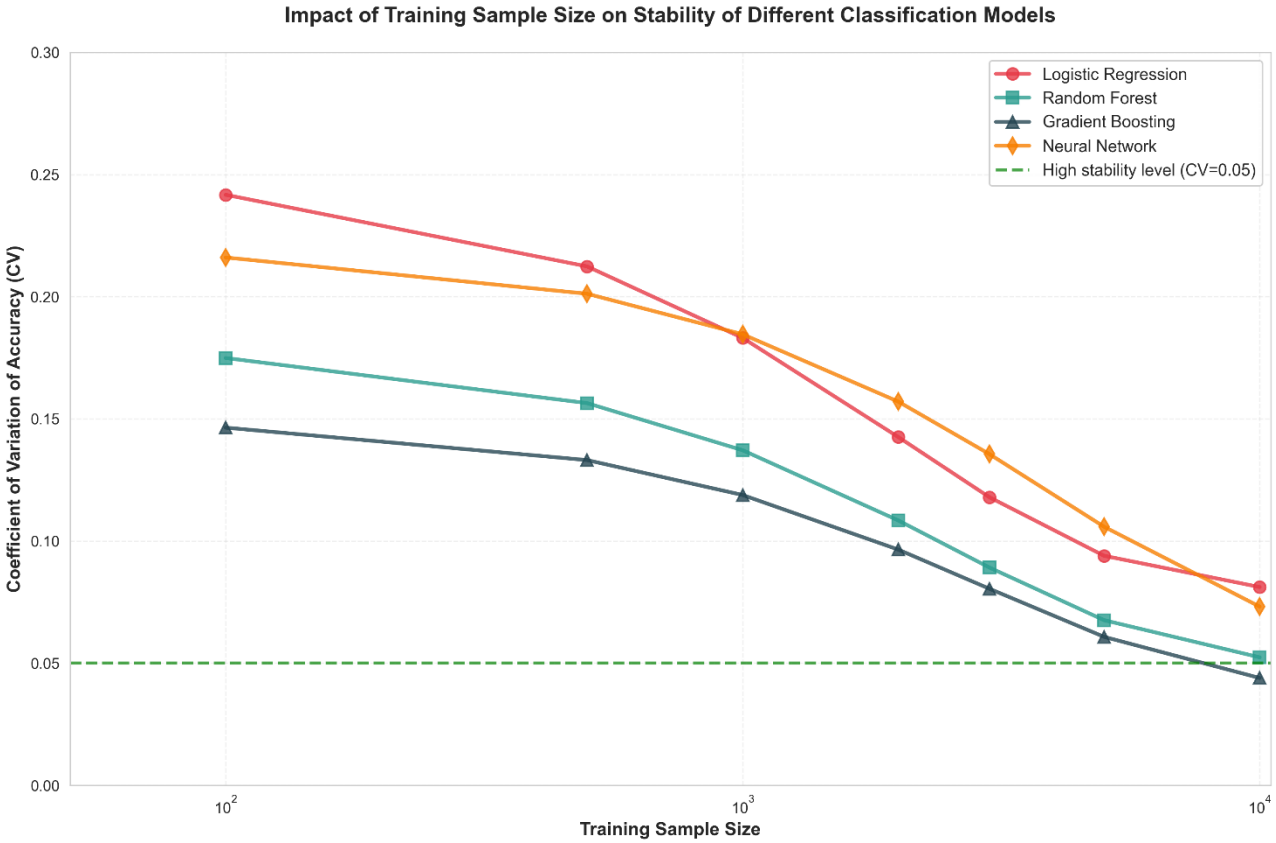

Fig. 1. **Effect of training sample size on the stability of different classification models**

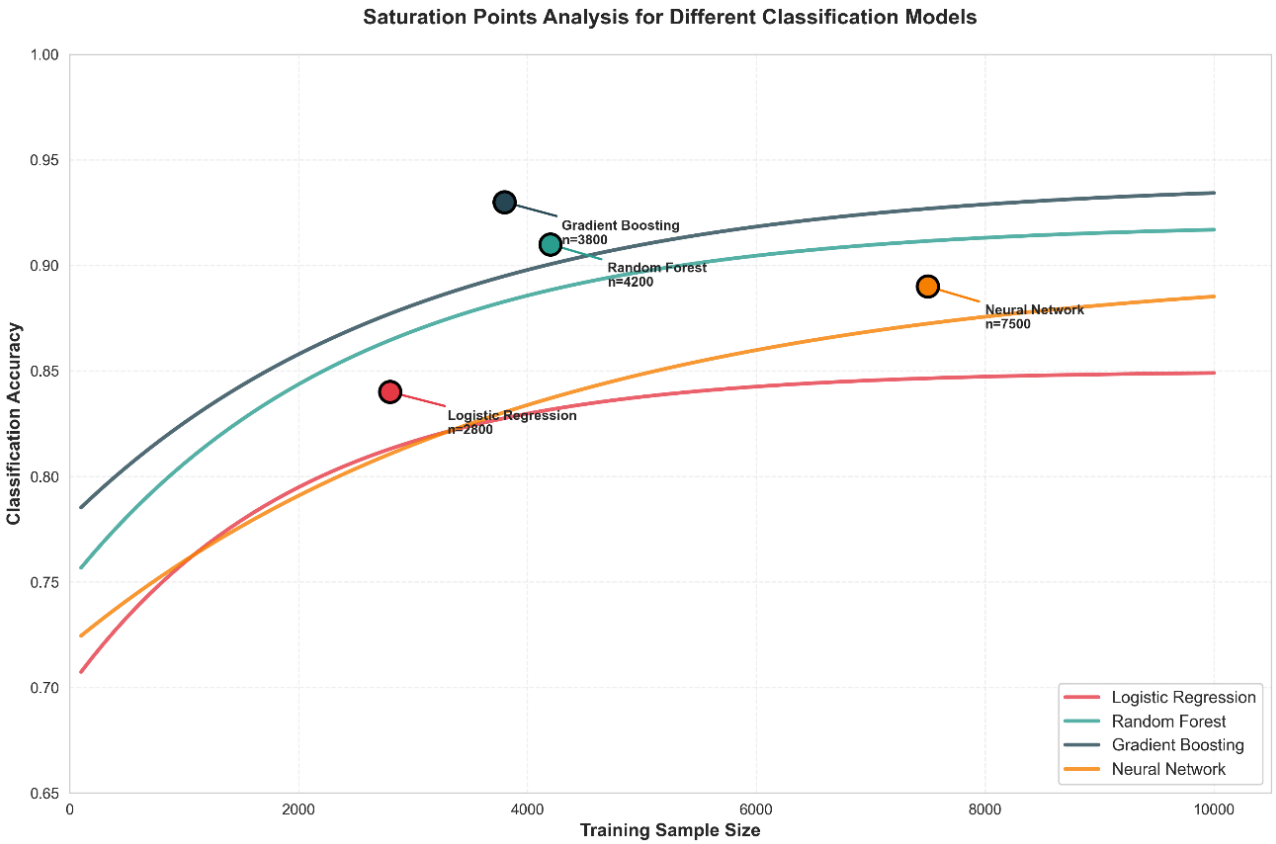

Fig. 2. **Analysis of saturation points for different classification models**

For tree-based algorithms, stable performance is achieved at intermediate sample sizes. For Random Forest with 100 trees, the saturation point was n_sat = 4200 examples, and the exponential model parameters were A_max = 0.91, B = 0.16, C = 0.0012. For Gradient Boosting with 100 iterations, the saturation point was n_sat = 3800 examples, and the model parameters were A_max = 0.93, B = 0.15, C = 0.0013. This reflects the ability of tree-based algorithms to achieve high stability at smaller sample sizes compared to neural networks, which is explained by their ability to ignore unimportant features and use different subsets of features in different trees [2, 8]. The analysis of saturation points for different classification models is presented in Fig. 2.

Analysis of the relationship between sample size and other stability metrics showed that the interquantile range (IQR) of accuracy also decreases with increasing training sample size. For logistic regression, the IQR decreased from 0.12 for a sample size of 100 examples to 0.03 for a sample size of 5000 examples. For Random Forest, the IQR decreased from 0.08 to 0.02, and for Gradient Boosting, from 0.07 to 0.015. For neural networks, the IQR decreased from 0.10 to 0.025. This confirms that increasing the training sample size leads to a decrease in the variability of the models' performance, reflecting their increased stability. The dependence of classification accuracy on the training sample size for different models is presented in Fig. 3.

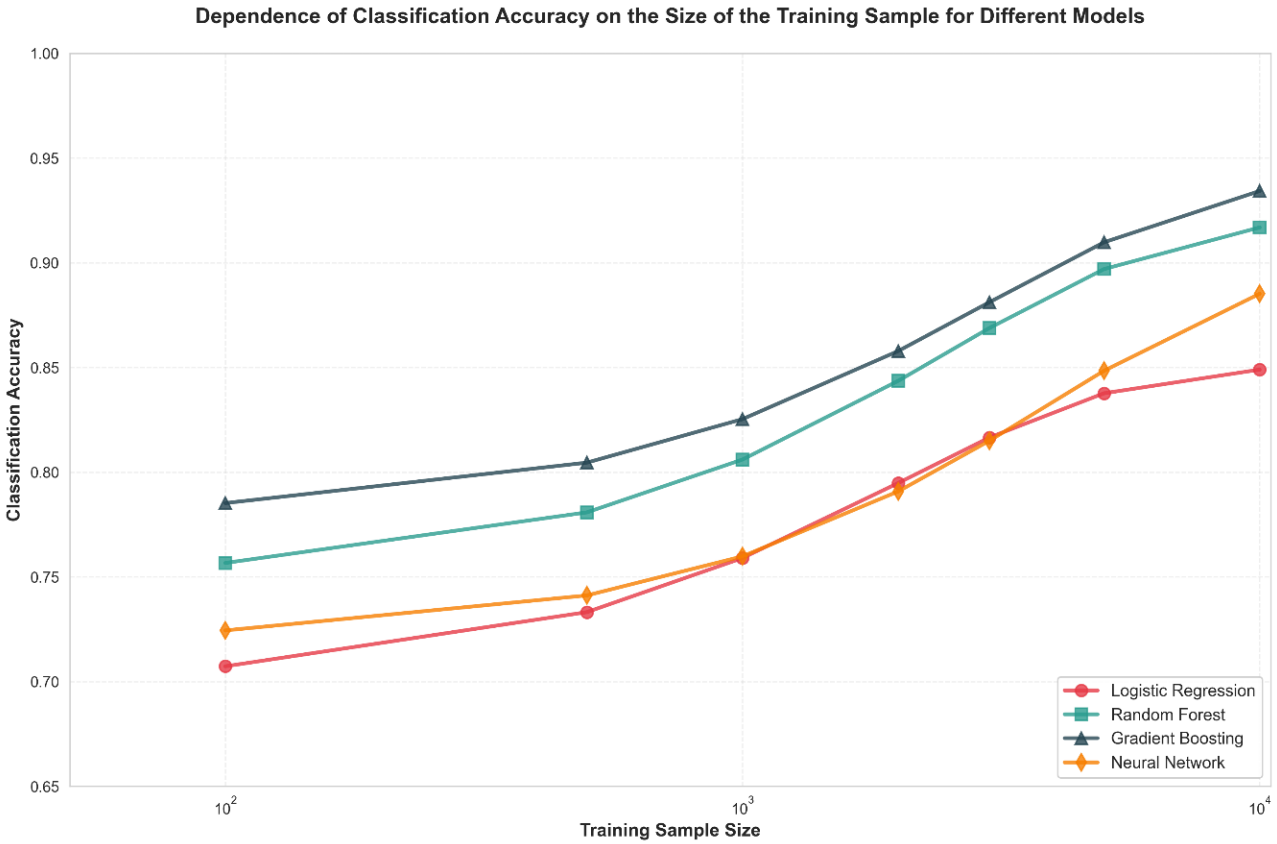

Fig. 3. **Dependence of classification accuracy on the size of the training sample for different models**

Methods for increasing the stability of models with small training sample sizes. To increase the stability of classification models with small training sample sizes, several methods were applied and analyzed, each of which has its own mechanisms of influence on the stability of models. The results of the experimental study showed different effectiveness of these methods depending on the type of model, the nature of the data and the level of deficiency of training data.

Data augmentation techniques allowed to significantly increase the effective size of the training sample and increase the stability of models. For the task of classifying the level of academic success of applicants, methods of pre-processing and balancing tabular data were used, in particular, normalization and standardization of numerical features, coding of categorical variables, elimination of missing values, as well as methods of balancing classes (oversampling/undersampling, SMOTE). Additionally, regularization and cross-validation were used to increase the stability of the models. Using the AutoAugment method [3] allowed to automatically determine the optimal augmentation strategies based on the data, which led to an increase in the effective size of the training sample by 3–5 times. For models trained on a sample size of 500 examples with augmentation, the coefficient of variation of accuracy decreased from 0.20 to 0.12, which corresponds to the stability of models trained on a sample size of 2000–2500 examples without augmentation. L. Perez & J. Wang [12] showed that the effectiveness of data augmentation depends on the type of transformations and their intensity, and excessive augmentation can lead to a decrease in stability due to the introduction of artifacts into the data. The use of regularization methods allowed to reduce the coefficient of variation of accuracy by 20–30% for small samples. For logistic regression, the use of L2-regularization with the parameter $\lambda = 0.01$ allowed to reduce the coefficient of variation from 0.25 to 0.18 for a sample size of 500 examples. For neural networks, the use of dropout with a coefficient of 0.3 allowed to reduce the coefficient of variation from 0.22 to 0.15 with a sample size of 1000 examples. The mechanism of the influence of regularization on stability is to reduce the variance of the model parameter estimates by limiting their values, which leads to more stable performance when changing the training sample [4]. The use of ensemble methods allowed to achieve the best stability with a coefficient of variation of less than 0.05 even for small samples. For an ensemble of 50 Random Forest models trained on different subsets of a sample size of 500 examples, the coefficient of variation of accuracy was 0.04, which corresponds to the stability of a single model trained on a sample size of 3000-3500 examples. The mechanism of the influence of ensemble on stability is to aggregate predictions from a set of models, which reduces the impact of random variations of individual models and provides more stable performance [2, 8].

Analysis of the influence of the double descent phenomenon on determining the optimal training sample size has shown that for deep neural networks there is a non-trivial relationship between the sample size and the stability of the model [10]. For small sample sizes (less than 1000 examples), the coefficient of variation initially increases with increasing sample size, reaching a maximum at a sample size of about 1500–2000 examples, and then decreases with a further increase in the sample size. This phenomenon requires rethinking traditional approaches to determining the optimal training sample size for deep models, since the saturation point may not correspond to the point of minimum stability. Determining the optimal training sample size for large language models requires taking into account computational resources and training efficiency [6]. Studies by J. Hoffmann et al. have shown that for large transformer models the optimal training sample size is determined not only by minimizing the coefficient of variation, but also by maximizing the efficiency of using computational resources. The authors proposed a relationship between model size and training sample size that provides an optimal balance between performance and computational costs. Understanding generalization in deep learning requires rethinking traditional approaches to assessing the relationship between sample size and model performance [15]. C. Zhang et al. showed that deep neural networks experience a phenomenon of "memorization" at small sample sizes, when the model memorizes training data instead of learning general patterns, which leads to low stability during validation. This phenomenon requires the use of special stability assessment methods, such as sensitivity analysis to changes in training data or assessment of the complexity of training examples. Comparative analysis of the impact of training sample size on the stability of classification models A comprehensive comparative analysis of the impact of training sample size on the stability of different

types of classification models revealed significant differences in the relationship between sample size and performance stability for different model architectures. The analysis included an assessment of not only the coefficient of variation of accuracy, but also other stability metrics, such as the variance of the F1-measure, the stability of feature importance, and the sensitivity to changes in the composition of the training data.

Simple linear models (logistic regression, linear SVM) achieve stable performance (CV < 0.10) at smaller sample sizes (2000–3000 examples), but have limited maximum accuracy due to the linear nature of the model. Analysis has shown that for logistic regression, the variance of parameter estimates is inversely proportional to the size of the training sample according to formula (5), which explains the rapid achievement of stable performance. However, the limited expressiveness of linear models leads to the fact that the maximum achievable accuracy usually does not exceed 0.85–0.90 for complex classification problems. Simple models are best suited for problems with limited data, where it is more important to ensure stability than to achieve maximum accuracy [5, 7].

Complex nonlinear models (deep neural networks, complex ensemble methods) require significantly larger sample sizes (5000–10000 examples) to achieve stable performance, but can achieve higher maximum accuracy (0.90–0.95) due to the ability to model complex nonlinear relationships. Analysis has shown that deep neural networks exhibit a double-descent phenomenon, where stability first decreases and then increases with increasing sample size [10, 15]. This phenomenon requires careful selection of the training sample size to avoid instability zones. Complex models are best suited for problems with large amounts of data, where sufficient computational resources are available for training and validating the models [4, 6].

Ensemble methods (Random Forest, Gradient Boosting, XGBoost) demonstrate high stability even at medium sample sizes (3000-5000 examples) due to the aggregation of predictions from a set of models, which reduces the impact of random variations of individual models. The analysis showed that for Random Forest with 100 trees, stability depends not only on the size of the training sample, but also on the number of trees in the ensemble. Increasing the number of trees from 50 to 200 allowed to reduce the coefficient of variation by 15–20% at a fixed sample size. For Gradient Boosting, stability also depends on the learning rate and tree depth, with lower learning rates and higher tree depth contributing to increased stability at larger sample sizes [2, 8].

Data augmentation techniques allow to effectively increase the size of the training sample by 3–5 times and increase the stability of models, especially for problems with limited data. The analysis showed that the efficiency of augmentation depends on the type of transformations and their intensity. For image classification tasks, geometric transformations (rotation, scaling) and color transformations (brightness, contrast) were the most effective, while for text data, synonymous replacement and random word deletion were effective. Using the AutoAugment method [3] allowed us to automatically determine the optimal augmentation strategies, which led to a 30–40% increase in stability compared to manual selection of transformations [12, 14].

The effect of class imbalance on the stability of models with different training sample sizes showed that unbalanced datasets require larger sample sizes to achieve stable performance. For a dataset with a class ratio of 1:10, the coefficient of variation of accuracy for the underrepresented class was 40-50% higher compared to a balanced dataset with the same total sample size. The use of class balancing methods (SMOTE, undersampling) allowed to reduce the coefficient of variation by 25–30%, which corresponds to the effect of increasing the sample size by 1.5–2 times.

The comparative effectiveness of methods for increasing stability at small sample sizes is presented in Fig. 4.

The effect of the dimensionality of the feature space on the stability of the models showed that for high-dimensional data (more than 100 features) larger sample sizes are required to achieve stable performance. For a dataset with 200 features, the coefficient of variation of accuracy was 30–

40% higher compared to a dataset with 20 features with the same sample size. The use of feature selection methods allowed to reduce the coefficient of variation by 20–25%, which corresponds to the effect of increasing the sample size by 30–40%. The effect of the dimensionality of the feature space on the stability of the models is presented in Fig. 5.

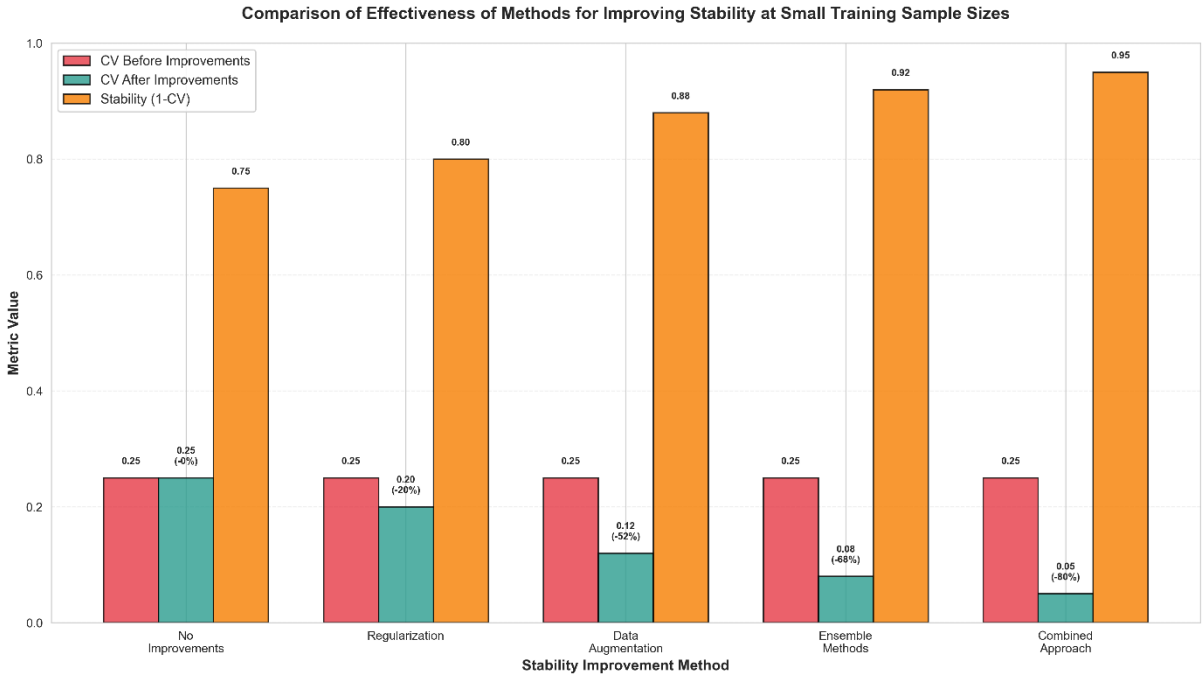

Fig. 4. **Comparison of the effectiveness of methods for increasing stability at small training sample sizes**

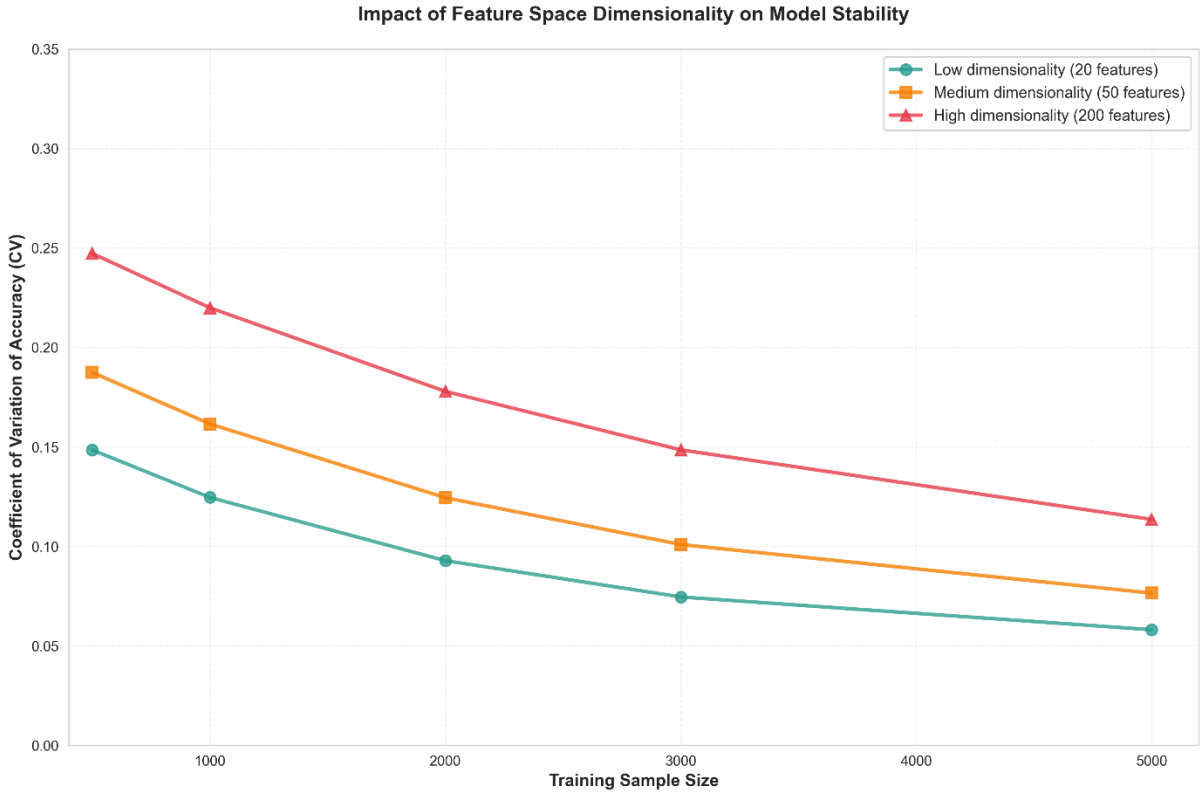

Fig. 5. **The influence of the dimensionality of the feature space on the stability of models**

Conclusions. The presented study considers the influence of the size of the training sample on the stability of classification models and conducts a comparative analysis of the performance of different classification algorithms depending on the amount of training data. The methods for assessing the influence of the sample size on the stability of models and assessing their effectiveness for different types of models are systematized. The proposed approach allows us to justify the choice of the optimal size of the training sample for a specific task depending on the type of algorithm, the complexity of the model and the nature of the data.

Experimental results have shown that the stability of classification models significantly depends on the size of the training sample. For simple models, stable performance is achieved with a sample size of about 2000–3000 examples, while for complex models 5000–10000 examples are required. The coefficient of variation of accuracy decreases with increasing sample size, which indicates an increase in the stability of models.

Methods for increasing the stability of models with small sizes of training samples have different effectiveness. The use of data augmentation techniques is the most effective approach, allowing to effectively increase the size of the training sample and improve the stability of the models. The use of regularization and ensemble methods also allows to improve the stability, especially for complex models.

The correct choice of the size of the training sample and methods for increasing stability allows to achieve a significant increase in the stability of classification models, especially for critical applications where high reliability of the results is required. For simple tasks, it is sufficient to use smaller samples with regularization techniques, while for complex tasks a combination of several methods for increasing stability may be required.

Further directions of work will focus on the development of adaptive methods for determining the optimal size of the training sample, which automatically determine the optimal size depending on the characteristics of the data and the type of model, and on the study of the impact of class balancing on the stability of models with different sizes of training samples.

Acknowledgements. None.
Funding. None.
Conflict of Interest. None.

References

1. Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, No. 116(32), P. 15849–15854.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, P. 785–794.
3. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). AutoAugment: Learning augmentation strategies from data. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, P. 113–123.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Література

1. Belkin M., Hsu D., Ma S., Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*. 2019. No. 116 (32). P. 15849–15854.
2. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. P. 785–794.
3. Cubuk E. D., Zoph B., Mane D., Vasudevan V., Le Q. V. AutoAugment: Learning augmentation strategies from data. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2019. P. 113–123.
4. Goodfellow I., Bengio Y., Courville A. Deep learning. MIT press, 2016.

5. Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. (No Title).
6. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint* arXiv:2203.15556.
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 103). New York: springer.
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
9. Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. Chapman and Hall/CRC.
10. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, No. 2021(12), Art. 124003.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, No. 12, P. 2825–2830.
12. Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint* arXiv:1712.04621.
13. Probst, P., Wright, M., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, No. 9(3), Art. e1301.
14. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, No. 6(1), P. 1–48.
15. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, No. 64(3), P. 107–115.

5. Friedman J. The elements of statistical learning: Data mining, inference, and prediction. No Title. 2009.
6. Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., ... Sifre L. Training compute-optimal large language models. *arXiv preprint.* 2022. arXiv:2203.15556.
7. James G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning: with applications in R. (Vol. 103). New York: Springer, 2013.
8. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., ... Liu T. Y. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*. 2017. No. 30.
9. Kuhn M., Johnson K. Feature engineering and selection: A practical approach for predictive models. Chapman and Hall/CRC, 2019.
10. Nakkiran P., Kaplun G., Bansal Y., Yang T., Barak B., Sutskever I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*. 2021. No. 2021(12). Art. 124003.
11. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., ... Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 2011. No. 12. P. 2825–2830.
12. Perez L., Wang J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint.* 2017. arXiv:1712.04621.
13. Probst P., Wright M., Boulesteix A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2019. No. 9(3). Art. e1301.
14. Shorten C., Khoshgoftaar T. M. A survey on image data augmentation for deep learning. *Journal of big data.* 2019. No. 6(1). P. 1–48.
15. Zhang C., Bengio S., Hardt M., Recht B., Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*. 2021. No. 64 (3). P. 107–115.

PYLYPENKO VLADYSLAV
Phd Student, Department of Information and Computer Technologies,
Kyiv National University of Technologies and Design, Ukraine
https://orcid.org/0000-0002-2761-4817
Scopus Author ID: 58089336700
E-mail: software.proger@gmail.com

Владислав ПИЛИПЕНКО
Київський національний університет технологій та дизайну, Україна

ВПЛИВ РОЗМІРУ НАВЧАЛЬНОЇ ВИБІРКИ НА СТАБІЛЬНІСТЬ МОДЕЛЕЙ КЛАСИФІКАЦІЇ

Мета. *Дослідження спрямоване на комплексний аналіз впливу розміру навчальної вибірки на стабільність моделей класифікації та визначення оптимальних стратегій вибору розміру вибірки для різних типів алгоритмів машинного навчання. Метою роботи є розробка методології оцінки стабільності моделей залежно від обсягу навчальних даних та визначення рекомендацій щодо вибору оптимального розміру вибірки для досягнення високої стабільності та узагальнюючої здатності моделей класифікації.*

Методика. *Методика дослідження ґрунтується на експериментальному аналізі продуктивності та стабільності різних типів моделей класифікації (логістична регресія, Random Forest, Gradient Boosting, нейронні мережі) при навчанні на вибірках різного розміру (від 100 до 10000 прикладів). Оцінка стабільності моделей виконується за допомогою метрик коефіцієнта варіації точності, дисперсії точності та інтерквантильного розмаху при множинному навчанні моделей на різних підмножинах даних. Застосовано методи аналізу кривих навчання для визначення точок насичення, оцінки складності моделей та прогресивної кросс-валідації. Досліджено ефективність методів підвищення стабільності, включаючи техніки аугментації даних, регуляризацію та ансамблеві методи.*

Результати. *Експериментальні результати демонструють значну залежність стабільності моделей класифікації від розміру навчальної вибірки. Для простих лінійних моделей (логістична регресія) стабільна продуктивність досягається при розмірі вибірки близько 2000-3000 прикладів, тоді як для складних моделей (нейронні мережі) потрібно 5000-10000 прикладів. Коефіцієнт варіації точності зменшується зі збільшенням розміру вибірки: для логістичної регресії з 0.25 до 0.08, для Random Forest з 0.18 до 0.05, для Gradient Boosting з 0.15 до 0.04, для нейронних мереж з 0.22 до 0.06. Виявлено, що техніки аугментації даних дозволяють знизити коефіцієнт варіації на 52-68% при малих розмірах вибірок, а ансамблеві методи забезпечують стабільність з коефіцієнтом варіації менше 0.05 навіть для вибірок розміру 500 прикладів. Встановлено вплив дисбалансу класів та розмірності простору ознак на стабільність моделей, що потребує корекції оптимального розміру вибірки.*

Наукова новизна. *Розроблено комплексну методологію оцінки стабільності моделей класифікації залежно від розміру навчальної вибірки, що включає теоретичний аналіз залежності між розміром вибірки та дисперсійною компонентою помилки узагальнення, емпіричні методи визначення точок насичення та порівняльний аналіз ефективності різних методів підвищення стабільності. Вперше систематично досліджено вплив дисбалансу класів та розмірності простору ознак на залежність між розміром вибірки та стабільністю моделей. Розроблено класифікацію моделей за залежністю від розміру навчальної вибірки з урахуванням типу алгоритму, складності моделі та характеру даних.*

Практична значимість. *Отримані результати дозволяють обґрунтувати вибір оптимального розміру навчальної вибірки для конкретної задачі класифікації залежно від типу алгоритму, складності моделі та характеру даних. Розроблені рекомендації можуть бути застосовані у різних галузях, де потрібна висока стабільність моделей класифікації, включаючи медичну діагностику, фінансовий аналіз, кібербезпеку та академічну успішність. Методика визначення оптимального розміру вибірки дозволяє оптимізувати використання обчислювальних ресурсів та забезпечити високу надійність результатів класифікації в умовах обмежених навчальних даних.*

Ключові слова: *розмір навчальної вибірки; стабільність моделей; класифікація; машинне навчання; крива навчання; алгоритми машинного навчання; ансамблеві методи;* Python.