https://doi.org/10.31891/2219-9365-2025-83-19
UDC 004.9; 004.8

HAVRYLIUK Myroslav
Lviv Polytechnic National University
https://orcid.org/0000-0001-5259-7564
e-mail: myroslav.a.havryliuk@lpnu.ua

METHOD OF INTELLECTUAL ANALYSIS OF SHORT HIGH-DIMENSIONAL SAMPLES BASED ON BAGGING ENSEMBLE WITH DATA AUGMENTATION

One of the persistent and critical challenges in the application of machine learning and statistical analysis methods in the medical field remains the effective processing of small data – datasets containing a limited number of observations for practical, ethical or biological reasons. In contrast to large-scale population studies or broad epidemiological databases, many real-world clinical scenarios involve working with small samples: individual patient data, rare diseases, early stage studies or specialized diagnostic procedures. As a result, researchers and clinicians are often forced to work with incomplete, sparse, or highly unbalanced data in an effort to create accurate and robust models that can be used to inform important clinical decisions. Thus, the development of efficient, reliable, and interpretable methods for processing short data is not only a methodological necessity but also a practical requirement of modern medicine. One of the most common ways to partially solve the problem of small sample analysis is data augmentation. Increasing the number of instances in the training set often has a positive effect on the accuracy of models. However, in the case of augmented data, relying on a single modeling strategy is sometimes not enough. Often, combining augmentation and ensemble learning approaches can lead to significant improvements in model robustness and performance.

This article develops a new method for intellectual analysis of short high-dimensional data samples for solving regression modeling problems, based on the use of a bagging ensemble of artificial neural networks with an additional data augmentation procedure. Its training algorithm and results are described in detail. Using this method, two medical problems were solved: predicting the level of bone fragility in patients with osteoarthritis and the percentage of body fat. According to the results of comparing the main performance metrics of the developed approach and the baseline models, proposed method demonstrated the best results for both problems. The developed bagging ensemble can be used in cases where the amount of available data is limited and classical models do not provide the required accuracy.

Keywords: small data, high-dimensional data, generalized regression neural network, data augmentation, ensemble learning, bagging, regression.

ГАВРИЛЮК Мирослав
Національний університет «Львівська політехніка»

МЕТОД ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ КОРОТКИХ ВИСОКОВИМІРНИХ ВИБІРОК НА ОСНОВІ БЕГІНГОВОГО АНСАМБЛЮ З АУГМЕНТАЦІЄЮ ДАНИХ

Однією з постійних і критичних проблем у застосуванні методів машинного навчання та статистичного аналізу в медичній галузі залишається ефективна обробка малих даних – наборів даних, що містять обмежену кількість спостережень з практичних, етичних або біологічних причин. На відміну від масштабних популяційних досліджень або широких епідеміологічних баз даних, багато реальних клінічних сценаріїв передбачають роботу з невеликими вибірками: дані окремих пацієнтів, рідкісні захворювання, дослідження на ранніх стадіях або спеціалізовані діагностичні процедури. Як результат, дослідники та медики часто змушені працювати з неповними, розрідженими або дуже незбалансованими даними, щоб створити точні та надійні моделі, які можна використовувати для прийняття важливих клінічних рішень. Таким чином, розробка ефективних, надійних та інтерпретованих методів обробки коротких даних є не лише методологічною необхідністю, але й практичною вимогою сучасної медицини. Одним з найпоширеніших способів часткового вирішення проблеми аналізу невеликих вибірок є доповнення даних. Збільшення кількості екземплярів у навчальному наборі часто позитивно впливає на точність моделей. Однак у випадку доповнених даних покладатися на одну стратегію моделювання іноді недостатньо. Часто поєднання підходів аугментації та ансамблевого навчання може призвести до значного покращення стійкості та продуктивності моделі.

У цій статті розроблено новий метод інтелектуального аналізу коротких багатовимірних вибірок даних для розв'язання задач регресійного моделювання, заснований на використанні бегінгового ансамблю штучних нейронних мереж з додатковою процедурою аугментації даних. Детально описані його алгоритм навчання та результати. За допомогою цього методу було вирішено дві медичні задачі: прогнозування рівня крихкості кісток у пацієнтів з остеоартритом та відсотка жиру в організмі. За результатами порівняння основних метрик ефективності розробленого підходу та базових моделей, розроблений метод продемонстрував найкращі результати для обох задач. Розроблений бегінговий ансамбль може бути використаний у випадках, коли обсяг доступних даних обмежений, а класичні моделі не забезпечують необхідної точності.

Ключові слова: малі дані, високорозмірні дані, нейронна мережа узагальненої регресії, аугментація даних, ансамблеве навчання, беггінг, регресія.

Стаття надійшла до редакції / Received 19.07.2025
Прийнята до друку / Accepted 14.08.2025

INTRODUCTION

The dependence of modern medicine on data-driven technologies opens up new possibilities for diagnosis, prognosis, treatment planning and clinical decision-making. However, one of the persistent and critical challenges in the application of machine learning and statistical analysis methods in the medical field remains the effective processing of small data – datasets containing a limited number of observations for practical, ethical or biological reasons [1]. In contrast to large-scale population studies or broad epidemiological databases, many real-world clinical scenarios involve working with small samples: individual patient data, rare diseases, early stage studies or specialized diagnostic procedures. Despite their limited volume, such data often have high information value and can be the key to the development of personalized medicine, early detection of diseases or adaptation of treatment to individual patient characteristics.

The problem of small data is particularly acute in the fields of rare disease research, pediatrics, and critical care, where large-scale data collection is impossible or extremely difficult due to low prevalence, high collection costs, or ethical constraints. In addition, privacy concerns, strict data sharing regulations, and fragmented storage of medical information across institutions further limit the availability of large, unified databases [2]. As a result, researchers and clinicians are often forced to work with incomplete, sparse, or highly unbalanced data in an effort to create accurate and robust models that can be used to inform important clinical decisions.

From a methodological perspective, processing small datasets poses a number of unique challenges. The application of classical machine learning models often requires a large number of samples to avoid overfitting, so data limitations make their effective use impossible [3]. In addition, medical data often have high dimensionality (e.g., genetic markers, laboratory tests) relative to the number of available observations, which complicates analysis due to the so-called "curse of dimensionality". Such conditions require the development of specialized approaches that can compensate for the lack of data while maintaining interpretability and reliability of results – critically important properties in the medical environment.

The importance of addressing this problem goes far beyond academic interest – it has a direct impact on patient health, the efficiency of healthcare, and the ability to make timely clinical decisions. For example, in emergency or intensive care settings, decisions often have to be made on the basis of very limited and dynamic information [4]. Similarly, in the early stages of drug development, preliminary conclusions based on small patient groups require particularly careful analysis. In such situations, robust methods for processing short data can significantly affect the timeliness of diagnosis, treatment adjustment, and the overall quality of medical decisions.

Thus, the development of efficient, reliable, and interpretable methods for processing short data is not only a methodological necessity but also a practical requirement of modern medicine. In this work, we aim to contribute to this direction by proposing a new approach that takes into account the unique challenges of short data in the medical environment.

RELATED WORKS

One of the most common ways to partially solve the problem of small sample analysis is data augmentation [5]. Increasing the number of instances in the training set often has a positive effect on the accuracy of models. However, in the case of augmented data, relying on a single modeling strategy is sometimes not enough. Therefore, augmentation and ensemble learning approaches often complement each other and, when combined, can lead to significant improvements in model robustness and performance.

In [6], the authors review existing ensemble methods for solving regression problems. They note that ensemble techniques for classification problems cannot always be successfully used for regression problems. The authors divide the ensemble approach into three key stages: generation (creation of a set of basic regression models), prunning (selection of the most effective models from all candidates), and integration (combining the results). [6] also provides a classification of ensemble techniques and demonstrates the key differences between them.

The paper [7] considers the problem of unbalanced regression, when certain ranges of values of the target variable occur much less frequently than others (similar to the problem of imbalanced classes). To improve the efficiency of the analysis of such samples, the authors propose a modified version of bagging, where classical bootstrapping is supplemented with resampling taking into account imbalance. According to the results of experiments on 20 regression datasets from various industries (including medical), this method demonstrated an advantage in accuracy compared to standard bagging.

[8] presents another approach to solving the problem of imbalanced regression, which is based on a combination of weighted resampling and data augmentation. For data augmentation, the researchers use several existing techniques (for example, from [9] and [10]). The authors demonstrate that their augmentation approach significantly improves the quality of predictions in regression problems with imbalanced target variables compared to baseline methods. Thus, the works [7] and [8] are examples of a successful combination of augmentation and ensembles for problems where the amount of data is insufficient.

Another approach of this type is the work [11], which proposes a method of doubling inputs for processing short data samples. The authors take traditional machine learning models, such as support vector regression, as a basis and, using specific augmentation of the dataset, achieve better analysis efficiency compared to basic models. The essence of this approach is to use the Cartesian square of the support sample to significantly increase its size.

In [12] the approach from [11] was further developed. The authors modified the input doubling method for the case of analyzing high-dimensional short-volume data. The developed method, unlike the original one, does not multiply the number of attributes in the augmented sample. This characteristic allows avoiding potential problems associated with a large number of features during training and application of the model. This approach demonstrated a lower prediction error for tasks in the field of medicine, compared to the original one. However, the accuracy of this model still has the potential to be improved when using ensemble principles. Thus, further development of this direction of short-volume data analysis is relevant.

The aim of this work is to develop a new method based on ensemble principles and an augmentation approach to improve the efficiency of intellectual analysis of short high-dimensional datasets in the medical field.

Materials and methods

The approach proposed in [12] is based on the augmentation of short datasets with subsequent prediction using a generalized regression neural network (GRNN), developed by D. Specht [13]. This model does not require training in the traditional sense, instead it uses a support sample to calculate the value of the output variable.

The generalized regression neural network consists of 4 layers. Its structural scheme is presented in Fig. 1.

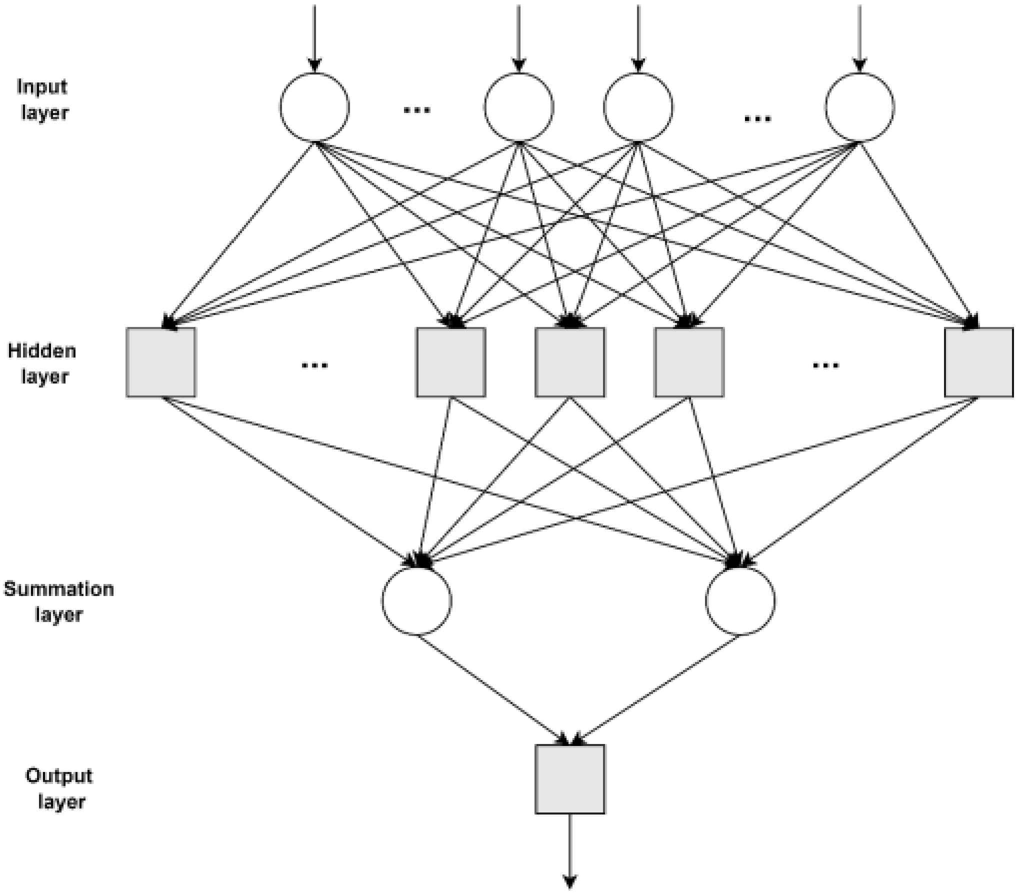

Fig. 1. Structural scheme of GRNN

The output value (prediction) *y* of a generalized regression neural network can be calculated as:

$$y(x)=\frac{\sum_{i=1}^{N} y_i * \exp(-\frac{(R_i)^2}{2\sigma^2})}{\sum_{i=1}^{N} \exp(-\frac{(R_i)^2}{2\sigma^2})} \tag{1}$$

where *x* is the input vector; y_i is the value of the target variable for the *i*-th vector of the support sample of size *N*; R_i is the Euclidean distance from the input vector to the *i*-th vector of the support sample; σ is the smoothing factor.

According to the augmentation method presented in [12], a new support sample is formed according to the following formulas:

$$F_{Train}=X_{Train}\otimes j_N + j_N \otimes(-X_{Train}) \tag{2}$$

$$l_{Train}=p_{Train}\otimes j_N \tag{3}$$

$$r_{Train} = j_N \otimes p_{Train} \tag{4}$$

$$A_{Train} = \left[F_{Train} \quad l_{Train} \quad r_{Train} \right] \tag{5}$$

where X_{Train} is the initial support sample of data in the form of a matrix; p_{Train} is the column vector of predictions made using GRNN for each of the vectors of the initial support sample; j_N is a column vector of ones of size N; $\otimes$ is the Kronecker product operation.

Among the main advantages of this method:

- for each pair of vectors of the initial sample, the differences of the corresponding attributes are calculated, which are used as new features in the augmented sample, allowing to avoid the use of artificial instances and noise;
- predictions obtained using the classical GRNN model complement the instances of the augmented sample with new informative features;
- the dimensionality of the augmented sample increases by only 2 features, which avoids a multiple increase in the number of attributes.

The described features of the algorithm from [12] allow us to expect a potential improvement in its performance when used in combination with ensemble approaches, in particular bagging. Bagging involves building a set of independent models on different random subsets of the training data, , after which the results of these models are averaged. This approach reduces the variance of the predictions, makes them less sensitive to noise and random features of the sample, and also increases the overall efficiency of the model. The use of augmented data subsamples in the process of training each model enhances the diversity of the ensemble, which is especially important for small sets, and provides a synergistic effect in increasing the accuracy and reliability of regression predictions.

In this work, we have developed an ensemble method that improves the efficiency of analyzing short high-dimensional samples of medical data by taking advantage of the combination of the augmentation approach from [12] and bagging. The main steps of the proposed algorithm are:

1. Formation of three random subsets of support sample vectors of a certain size (90% of the initial sample).
2. Augmentation of each of the subsets using the method from [12].
3. Creating generalized regression neural network models, using each of the augmented samples obtained in step 2.
4. Obtaining predictions for test examples using each of the models created in step 3.
5. Calculating final predictions by averaging the results of all models.

The scheme of the developed ensemble approach is visualized in Fig. 2.

We applied the proposed ensemble method to solve two practical problems in the medical field, each of which has significant practical importance for the diagnosis and prevention of diseases:

- predicting the level of bone fragility for patients with osteoarthritis (dataset 1 – [14]);
- body fat percentage prediction (dataset 2 – [15]).

Both datasets using which we evaluated the performance of our method are short and high-dimensional:

- dataset 1 contains 34 instances (6 attributes);
- dataset 2 contains 24 observations (7 attributes).

MODELING, RESULTS, AND COMPARISON

Four key metrics were selected to assess the model’s performance, each of which provides a comprehensive description of accuracy and computational cost. These include: mean absolute error (MAE), which measures the average deviation of predicted values from actual values; mean square error (MSE), which reflects the average value of the squares of deviations and is more sensitive to large errors; root mean square error (RMSE); and median error (MedE), which characterizes the central tendency of errors [16].

The simulation was performed using the Python programming language and a wide range of its scientific libraries, including numpy for array and numerical operations, scipy for mathematical calculations and optimization, and sklearn for implementing machine learning tools. Before performing the main procedure, the augmented data sample was scaled using the MaxAbsScaler transformation.

The search for optimal values of the two key parameters σ1 and σ2 was carried out in the range from 0.001 to 10 using the global optimization algorithm Dual Annealing, which combines elements of simulated annealing and local search to increase the accuracy and reliability of the result [17]. To calculate the values of the metrics, a 5-fold cross-validation procedure was used, which allows for a more objective assessment of the quality of the predictions. The final results of the experiments for dataset 1 are given in Table 1, and for dataset 2 - in Table 2.

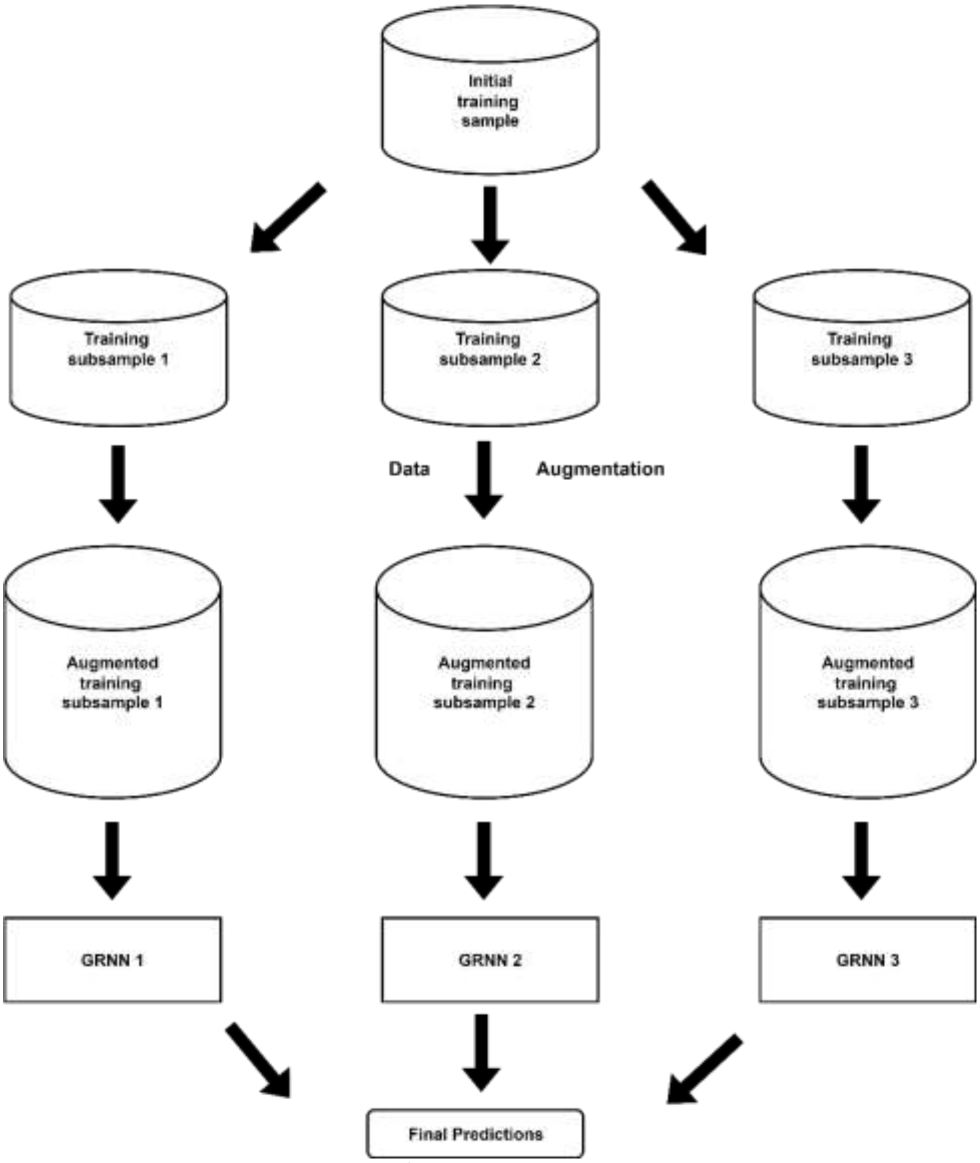

Fig. 2. Scheme of the developed ensemble approach

Table 1

Modeling results for dataset 1

Time, msec	σ1	σ2	MAPE	RMSE	MAE	MSE	MedAE
0,055 ±0.004	0,215	0,011	0,371±0.134	4,786±0.970	3,740±0.858	23,847 ±9.232	2,879±0.968
	0,009	0,150					
	0,105	0,162					

Table 2

Modeling results for dataset 2

Time, msec	σ1	σ2	MAPE	RMSE	MAE	MSE	MedAE
0,035±0.002	2,924	0,143	0,060±0.016	1,004±0.156	0,839±0.117	1,033±0.314	0,779±0.090
	6,263	0,117					
	2,357	0,136					

In addition, we compared the obtained performance metrics of the proposed comprehensive method and the approaches of which it consists:

- traditional generalized regression neural network;
- bagging (based on GRNN);
- augmentation algorithm (based on GRNN) from [12].

In general, the developed method demonstrated the best prediction accuracy. The comparison showed that the combination of different approaches is a promising direction for research.

A visualization of the comparison of RMSE and MAE for the first dataset can be seen in Fig. 3. As can be seen, the proposed ensemble significantly improved the values of the metrics compared to the baseline approaches. For example, the developed method reduced RMSE by 7.68% and MAE by 7.01% compared to the augmentation approach from [12].

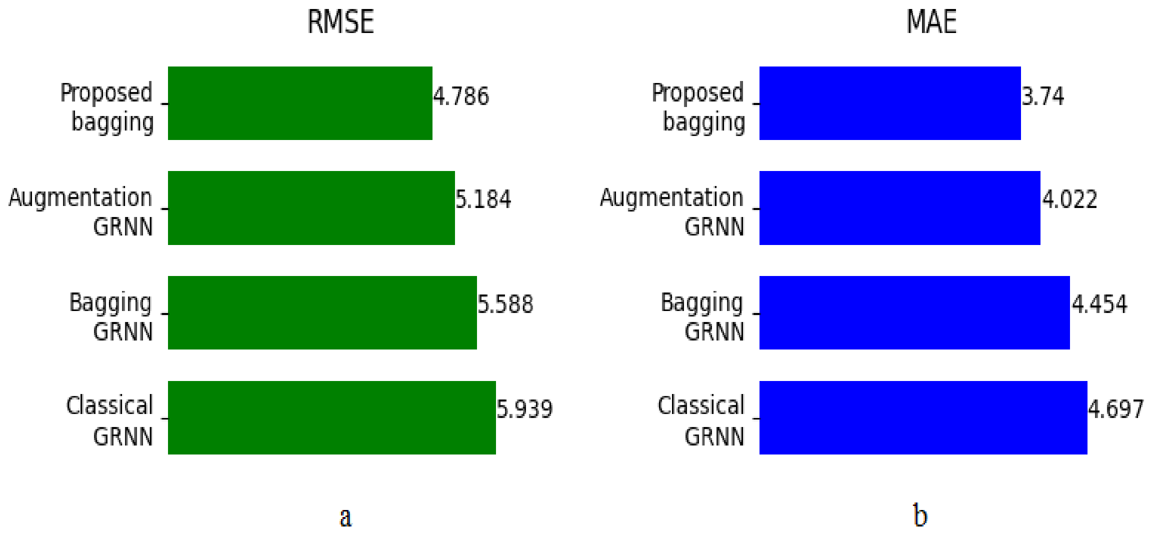

Fig. 3 Visualization of errors for dataset 1: a- RMSE; b-MAE

Figure 4 visualizes the comparison of the obtained performance metrics for the second dataset. The developed ensemble also outperformed all baseline models in terms of the calculated metrics. For example, the proposed approach improved RMSE (5.01%) and MAE (7.09%) compared to the augmentation method from [12].

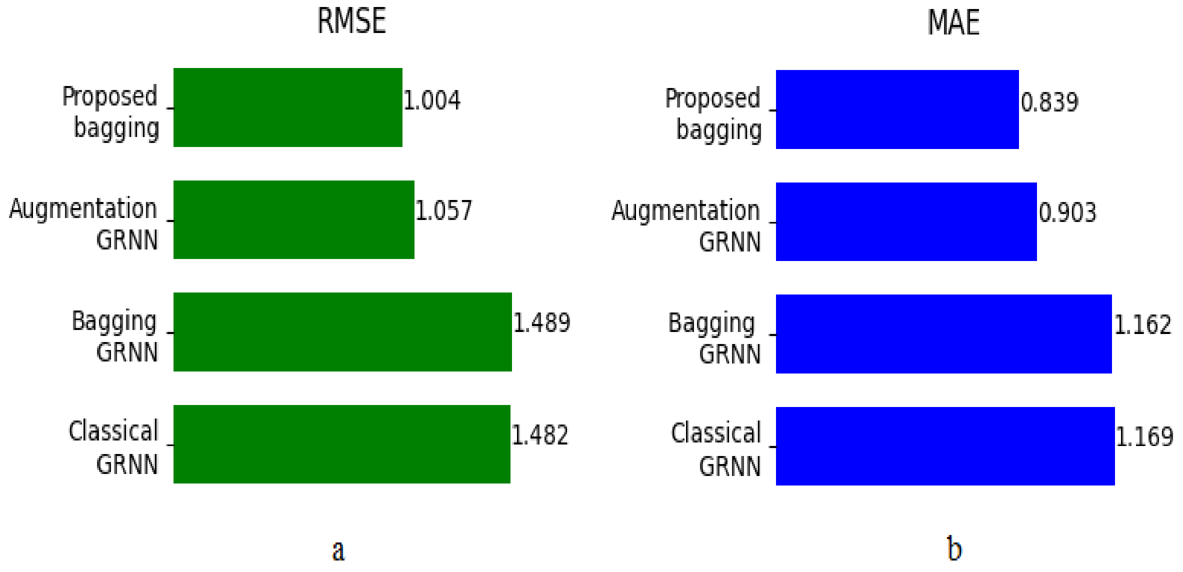

Fig. 4 Visualization of errors for dataset 2: a- RMSE; b-MAE

We also recorded the execution time of each of the evaluated models. The comparison results were visualized in Fig. 5. As expected, the augmentation technique significantly increases the duration of the algorithm. The use of bagging also requires more time due to the use of multiple models.

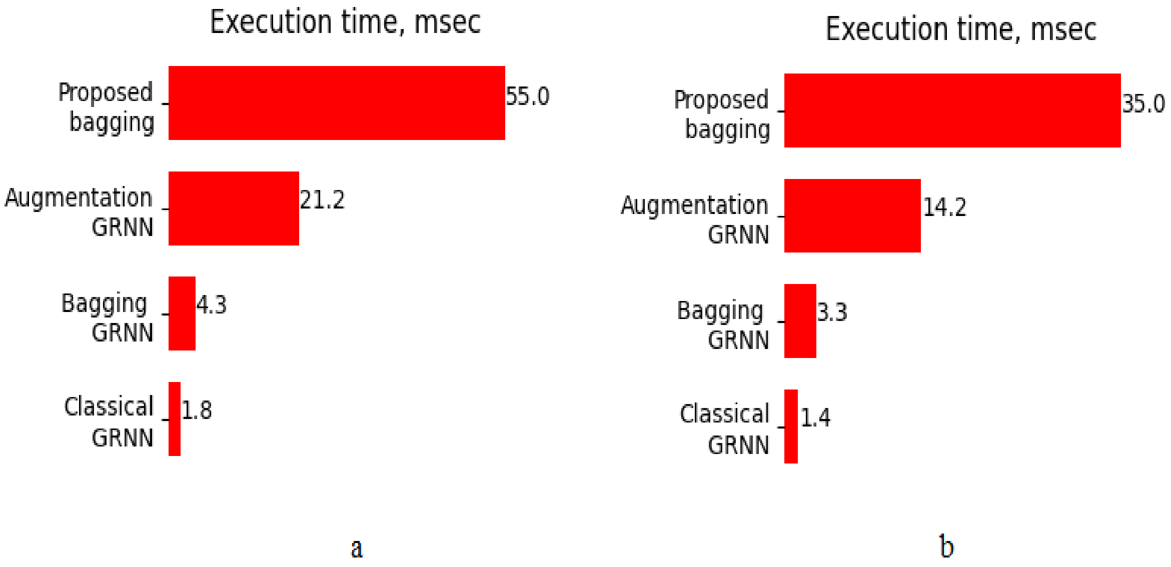

Fig. 5 Execution time comparison: a- for the obd; b- for the body_fat

The obtained results of the comparison of metrics demonstrate that the proposed augmentation-bagging approach outperforms the basic methods in terms of prediction accuracy for both tasks. The improvement in the quality of the analysis, in particular the reduction of the mean absolute error (MAE) and the root mean square error (RMSE), can be explained by a combination of two key factors: expanding the support sample through augmentation and reducing the variance of the predictions through the use of an ensemble of models.

Bagging provides averaging of the results of many independent models. This reduces the impact of random errors of individual models and makes the prediction more stable. The use of augmented data at the stage of forming the support subsamples increases the diversity of the ensemble, which is one of the key factors in increasing the accuracy. In medical tasks, where even a small increase in accuracy can be critical for practical application, such a combination of methods provides particularly significant advantages.

At the same time, the results show that the increase in accuracy is accompanied by an increase in the execution time of the algorithm. The main reason for this is the need to train more models within the ensemble, as well as the increase in the amount of training data through augmentation. In addition, the increase in the number of models in the ensemble leads to a proportional increase in duration.

Thus, the proposed method demonstrates a trade-off between increasing accuracy and increasing computational costs. In many medical applications, where prediction accuracy takes priority over computational speed, this compromise is justified. However, for tasks where time is a critical factor, it may be advisable to use simpler algorithms.

CONCLUSIONS

In this paper, a new method for intellectual analysis of short high-dimensional datasets for solving regression problems based on a combination of ensemble principles and augmentation was proposed, and its algorithm and results are described in detail. Two medical problems were solved using the approach: predicting the level of bone fragility in patients with osteoarthritis and the percentage of body fat.

According to the results of comparing the main performance metrics of the developed approach and the baseline models, the proposed method demonstrated superiority for both problems. The proposed bagging method can be used in cases where the amount of available data is limited and traditional models do not provide the required accuracy. Further research may include a combination of other ensemble strategies and the described data augmentation principles to further improve efficiency.

References

1. Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80(4), 463-475.

2. Hekler, E. B., Klasnja, P., Chevance, G., Golaszewski, N. M., Lewis, D., & Sim, I. (2019). Why we need a small data paradigm. BMC medicine, 17, 1-9. https://doi.org/10.1186/s12916-019-1366-x .

3. Estrin, D. (2014). Small data, where n= me. Communications of the ACM, 57(4), 32-34. https://doi.org/10.1145/2580944 .

4. Dumyn, I., Basystiuk, O., Dumyn, A. (2025). Graph-based approaches for multimodal medical data processing. Proceedings of the 7th International Conference on Informatics & Data-Driven Medicine Birmingham, United Kingdom, November 14-16, 2024, pp. 349-362.

5. Izonin, I., Tkachenko, R., Yemets, K., & Havryliuk, M. (2024). An interpretable ensemble structure with a non iterative training algorithm to improve the predictive accuracy of healthcare data analysis. Scientific Reports, 14(1), 12947. https://doi.org/10.1038/s41598-024-61776-y .

6. Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, 45(1), 1-40. https://doi.org/10.1145/2379776.2379786 .

7. Branco, P., Torgo, L., & Ribeiro, R. P. (2018). Rebagg: Resampled bagging for imbalanced regression. *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2018, pp. 67-81.

8. Stocksieker, S., Pommeret, D., & Charpentier, A. (2023). Data augmentation for imbalanced regression. arXiv preprint arXiv:2302.09288. https://doi.org/10.48550/arXiv.2302.09288 .

9. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28(1), 92-122. https://doi.org/10.1007/s10618-012-0295-5 .

10. Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, 74, 1-26. https://doi.org/10.18637/jss.v074.i11 .

11. Izonin, I., & Tkachenko, R. (2022). Universal intraensemble method using nonlinear AI techniques for regression modeling of small medical data sets. In *Cognitive and Soft Computing Techniques for the Analysis of Healthcare Data* (pp. 123-150). Academic Press. https://doi.org/10.1016/B978-0-323-85751-2.00002-5 .

12. Havryliuk, M. (2025). Enhanced two-step augmentation method for analyzing small datasets in medical applications. *Computer systems and information technologies*, (1), 156-162. https://doi.org/10.31891/csit-2025-1-18

13. Specht, D. F. (1991). A general regression neural network. IEEE transactions on neural networks, 2(6), 568-576.

14. Perilli, E., Baleani, M., Öhman, C., Baruffaldi, F., & Viceconti, M. (2007). Structural parameters and mechanical strength of cancellous bone in the femoral head in osteoarthritis do not depend on age. Bone, 41(5), 760-768. https://doi.org/10.1016/j.bone.2007.07.014

15. Body Fat Percentage of Women. Available online: https://www.kaggle.com/datasets/vishweshsalodkar/body-fat-percentage (accessed on 20 July 2025).

16. Nykoniuk, M., Basystiuk, O., Shakhovska, N., & Melnykova, N. (2025). Multimodal Data Fusion for Depression Detection Approach. Computation, 13(1), 9. https://doi.org/10.3390/computation13010009

17. Yemets, K. (2024). Time series forecasting model for solving cold start problem via temporal fusion transformer. Computer systems and information technologies, 1, 57-64. https://doi.org/10.31891/csit-2024-1-7