https://doi.org/10.31891/csit-2025-1-18
UDC 004.9; 004.8

Myroslav HAVRYLIUK
Lviv Polytechnic National University

ENHANCED TWO-STEP AUGMENTATION METHOD FOR ANALYZING SMALL DATASETS IN MEDICAL APPLICATIONS

Despite the enormous possibilities for data collection, situations still often arise where data is scarce. Insufficient data can significantly complicate their effective analysis, since most known approaches require a sufficiently large training sample to obtain accurate predictions. In the field of medicine, the problems of lack of data are quite common for a number of reasons (confidentiality, fragmentation and natural rarity). Accordingly, the development of algorithms that can at least partially eliminate the scarcity of data and demonstrate satisfactory efficiency is relevant. Existing techniques for analyzing small data based on their augmentation can improve the efficiency of traditional methods. However, along with an increase in the number of instances in the sample, the number of features also increases significantly, which can negatively affect the performance of machine learning methods.

In this paper, an improved two-step method was proposed for the intelligent analysis of short high-dimensional data sets based on a generalized regression neural network. A peculiarity of this approach is the avoidance of a multiple increase in the number of features in the augmented sample. The method was used to solve two regression problems: predicting the value of a function and determining the compressive strength of the femur. Both data sets contained less than 100 instances. The optimal parameters were determined using the Dual Annealing optimization algorithm for five distance measures: Euclidean, Chebyshev, Manhattan, Canberra, and cosine. The proposed method showed a significant reduction in errors (such as MAE, RMSE) compared to the traditional GRNN model. The developed technique also surpassed the accuracy of the input doubling method for both solved problems. Along with increasing accuracy, the proposed model also increased the execution time. Therefore, the feasibility of its application depends on the priorities of the problem being solved.

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

Мирослав ГАВРИЛЮК
Національний університет «Львівська політехніка»

ПОКРАЩЕНИЙ ДВОКРОКОВИЙ МЕТОД АУГМЕНТАЦІЇ ДЛЯ АНАЛІЗУ МАЛИХ НАБОРІВ ДАНИХ У МЕДИЧНИХ ЗАСТОСУВАННЯХ

Попри величезні можливості для збору даних, досі нерідко виникають ситуації, де дані є дефіцитними. Недостатня кількість даних може значно ускладнити їх ефективний аналіз, оскільки більшість відомих підходів вимагають достатньо великої тренувальної вибірки для отримання точних передбачень. У галузі медицини проблеми нестачі даних є досить поширеними через низку причин (конфіденційність, фрагментованість та природна рідкісність). Відповідно, актуальною є розробка алгоритмів, що зможуть хоча б частково знівелювати дефіцит даних та продемонструвати задовільну ефективність. Наявні техніки аналізу малих даних, що базуються на їх аугментації, можуть покращити ефективність традиційних методів. Однак, разом зі збільшенням кількості екземплярів у вибірці, кількість ознак також суттєво зростає, що може негативно позначитись на роботі методів машинного навчання.

У цій роботі було запропоновано удосконалений двокроковий метод для інтелектуального аналізу коротких високорозмірних наборів даних на основі нейронної мережі узагальненої регресії. Особливістю цього підходу є уникнення кратного збільшення кількості ознак в аугментованій вибірці. Метод було використано для розв'язання двох регресійних задач: передбачення значення функції та визначення міцності на стиск стегнової кістки. Обидва набори даних містили менше 100 екземплярів. Оптимальні параметри було визначено за допомогою оптимізаційного алгоритму Dual Annealing для п’яти мір відстані: евклідової, Чебишова, мангеттенської, канберрської та косинусової. Запропонований метод показав суттєве зменшення похибок (таких як MAE, RMSE) порівняно з традиційною моделлю GRNN. Також розроблена техніка перевершила точність методу подвоєння входів для обох розв’язуваних задач. Разом із підвищенням точності, запропонована модель також збільшила час виконання. Тому доцільність його застосування залежить від пріоритетів вирішуваної проблеми.

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

Introduction

The current state of information technology development allows us to gain important insights into various processes and make predictions based on this knowledge. The basis of this knowledge is previously collected data. Despite the enormous possibilities of searching and collecting, situations still often arise where data is scarce. Insufficient data can significantly complicate their effective analysis.

In the field of medicine, data scarcity problems are quite common [1, 2, 3]. The reasons for such difficulties are quite diverse:

- Confidentiality and security of medical data. Data about patients in healthcare institutions are confidential and require strict adherence to protocols during their exchange and processing [4]. Such features often limit the ability to collect a sufficient amount of data for analysis.

● Fragmentation of data sources. Medical data is often distributed across different healthcare institutions and scientific institutions without a unified information storage system [5]. Such autonomy often does not allow analysts to use the entire volume of available data, even within one country.

● The rarity of certain data. There are diseases and medical conditions that are very rare, and therefore the sample size is a priori small. Also, in such situations, the value of reliable and detailed documentation of each case of such diseases in the context of data collection increases.

The lack of data often creates a problem during their processing, since the vast majority of methods of intellectual analysis require a sufficiently large training sample to obtain accurate predictions. Accordingly, the development of algorithms that can at least partially compensate for the lack of data and demonstrate satisfactory efficiency is relevant.

Related works

At present, there is no single, clear definition of the concept of “small data”. Different researchers use different definitions, depending on the needs of the study. For example, in [6] the author introduces the concept of small data in medicine as personalized data based on the digital footprint of each individual person. He insists on the importance of an individual approach to medical care, and, accordingly, much smaller amounts of data for analysis than usual.

In [7] the authors justify the need to develop a paradigm of small data, as opposed to the concept of big data. Researchers use the definition of “small data” from [6], specifying that the unit for which small data is collected may be not only one specific person, but also, for example, a hospital, a community, etc. Such a concept is quite abstract, while in practice it is often necessary to more specifically define which dataset can be called “small”. For example, in [8] the author considers a small set to be such a set where the number of instances and the number of attributes are comparable. In [9], researchers assess the impact of dataset size on the accuracy of its classification by traditional machine learning methods. The results confirmed the assumption that the representativeness of the dataset distribution relative to a specific subject area can offset the small number of instances in the sample. Taking into account the above, in our study, we consider small datasets to be those where the number of instances is less than 100.

A popular approach among researchers involved in the analysis of small numerical data is augmentation. This class of methods allows you to increase the number of instances, as well as, possibly, attributes, to form a sample whose size is sufficient for processing by traditional machine learning methods. In this context, it makes sense to pay attention to the work [10], where the authors propose a forecasting method that combines data augmentation and the principle of ensemble prediction. For a short dataset containing N observations and k attributes, an augmented dataset with N^2 observations and $2k$ attributes is generated during the implementation of this algorithm. The researchers demonstrated the effectiveness of this method for solving problems in the medical field. The versatility of the method was confirmed by combining training and application procedures with several models.

In [11], a modification of the input doubling method from [10] based on SVR with nonlinear kernels was proposed for predicting urinary calcium concentration in the case of a short dataset. This algorithm demonstrated significantly lower RMSE and MAE errors than the simple model, the baseline method, and popular classical machine learning algorithms.

These methods showed high efficiency when solving the problem set in the work, however, they have a certain drawback in the context of possible processing of high-dimensional datasets (where the ratio of the number of instances to the number of features is less than 100). In the case of processing such a data set, the number of features of the augmented dataset becomes too large, which can negatively affect the performance of machine learning methods.

That is why it is necessary to conduct research in the direction of increasing the number of instances during data augmentation without significantly increasing the number of features. Therefore, this problem is relevant and has potential for research.

The purpose of the work is: to improve the efficiency of analyzing small sets of high-dimensional data in medical applications.

Materials and methods

The basis of our approach in this work is the augmentation of short datasets with the subsequent prediction procedure based on a generalized regression neural network.

The generalized regression neural network (GRNN) was proposed by D. Specht [12] in 1991. It belongs to the class of neural networks that do not require training. The basis of the GRNN prediction procedure is the use of the Gaussian radial basis function.

The topology of this neural network is shown in Fig. 1.

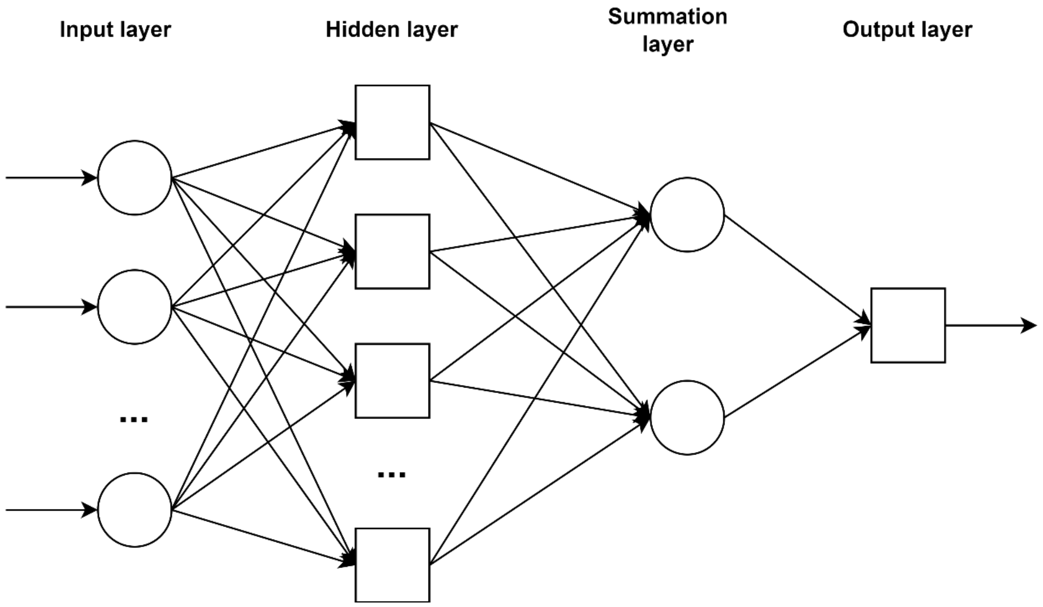

Fig. 1. GRNN topology

The structure of the GRNN consists of 4 layers: input, hidden, summation and output.

- The values of the input layer are the values of the attributes of the test vector transmitted by the network.
- In the hidden layer of the neural network, the Gaussian radial basis functions are calculated between the test vector and each of the support vectors of the set:

$$H_i = \exp(-\frac{(R_i)^2}{2\sigma^2}) \tag{1}$$

where R_i is the distance (Euclidean or other) between the test vector and the *i*th vector of the support sample; σ is the smoothing factor.

- In the summation layer, the sum of the values of the radial basis functions and the sum of the products of the values of the radial basis functions and the values of the corresponding target attribute of the reference sample vectors are calculated:

$$NS = \sum_{i=1}^{n} y_i H_i \tag{2}$$

$$DS = \sum_{i=1}^{n} H_i \tag{3}$$

- The output value of GRNN is the result of dividing the values calculated in the previous layer:

$$Y = \frac{NS}{DS} \tag{4}$$

Among the advantages of using GRNN compared to other neural networks, we can highlight:

- The absence of a training procedure, which allows making predictions in real time;
- Quite high speed of the network in the case of a short support sample [13];
- The presence of only two parameters for tuning - the metric by which the distances between vectors are calculated and the smoothing factor σ.

The main disadvantage of this neural network is the increase in the number of neurons in the hidden layer with an increase in the support sample, which leads to a decrease in the prediction speed. Thus, given the peculiarities of the work of GRNN, this neural network has good potential for processing short datasets, where there is a small number of support instances.

As already mentioned, in work [10] the authors propose a method for analyzing short datasets, where the number of features in the augmented dataset increases to *2k* (*k* - the number of attributes in the initial dataset). In our work, we propose a new improved method that combines the advantages of the approach from [10] with improvements that ensure efficient processing of small high-dimensional data.

The main features of the developed approach are:

- using the differences of the corresponding attributes for each pair of vectors of the initial sample as attributes of the augmented sample;

- using the predictions made by the classical GRNN model as features;
- the dimensionality of the augmented sample is k+2, which eliminates the problem of a multiple increase in the number of features during augmentation (which is a significant difference from the method from [10]).

Like many machine learning methods, our proposed method operates through training and application (prediction) procedures. The main stages of the training procedure of our proposed method are:

1. Obtaining predictions for each of the training instances using GRNN.
2. Creating an augmented training sample of data from the initial training dataset and the predictions obtained in stage 1.

The main stages of the application procedure are:

1. Obtaining predictions for the test instance using GRNN.
2. Creating an augmented test sample of data from the initial training dataset, the test instance, and the predictions obtained in stage 1 (and in stage 1 of the training mode).
3. Obtaining the final prediction using the augmented test sample and GRNN.

So we describe these procedures in more detail.

The augmented training sample A_{Train} is constructed as follows. We denote the input feature matrix of the initial training set of data, consisting of *N* instances with *k* attributes, as X_{Train} , and the target attribute of the training sample as a column vector y_{Train} . Using the basic model described above (with smoothing factor σ$_1$), we obtain a column vector of basic predictions p_{Train} (where every element is the prediction for a corresponding vector from X_{Train}).

We propose to calculate the difference of the corresponding attributes for each pair of vectors of the training sample and use these differences as attributes in the augmented training set. We denote the matrix of the above-described pairwise attribute differences as F_{Train} . It can be calculated as follows:

$$F_{Train} = X_{Train} \otimes j_N + j_N \otimes (-X_{Train}) \tag{5}$$

where ⊗ is the Kronecker product operation; j_N is all-ones column vector of size *N*.

The two additional attributes in the augmented training set are the predictions for each of the vectors in the pair made by the base model. We denote these features as l_{Train} and r_{Train} :

$$l_{Train} = p_{Train} \otimes j_N \tag{6}$$

$$r_{Train} = j_N \otimes p_{Train} \tag{7}$$

Thus, the matrix of the augmented training set A_{Train} has the following form:

$$A_{Train} = \left[F_{Train} \quad l_{Train} \quad r_{Train} \right] \tag{8}$$

We propose to calculate the artificial target attribute t_{Train} as the difference between the target attributes for each pair of training sample vectors:

$$t_{Train} = y_{Train} \otimes j_N + j_N \otimes (-y_{Train}) \tag{9}$$

After creating the augmented training sample and the artificial target attribute, the base model is trained using them. Since, in our study, we use GRNN, which does not have an explicit training procedure, the next step is to apply the method.

Suppose we need to predict the value of the target attribute y_{Test} for the test vector x_{Test} . Similar to the training procedure, using the base model, we can calculate the predictions p_{Test} for the test vector. Similar to the training mode, we propose to use the differences between the corresponding attributes of the test vector and each of the vectors of the initial training sample as attributes in the augmented test sample:

$$F_{Test} = x_{Test} \otimes j_N + (-X_{Train}) \tag{10}$$

The two additional attributes in the augmented test set are the predictions for each of the vectors in the pair made by the base model. We denote these attributes as l_{Test} and r_{Test} :

$$l_{Test} = p_{Test} \otimes j_N \tag{11}$$

$$r_{Test} = j_N \otimes p_{Train} \tag{12}$$

So, the augmented test sample matrix A_{Test} has the following form:

$$A_{Test} = \left[F_{Test} \quad l_{Test} \quad r_{Test} \right] \tag{13}$$

After performing the prediction procedure for A_{Test} by the base model(with smoothing factor σ₂), we will obtain a column vector of the artificial target attribute t_{Pred} . We propose to calculate the predicted value for the test vector using the ensemble principle as the average value among all predictions of the target attribute *y*:

$$y_{Pred} = \frac{1}{N} (t_{Pred} + y_{Train})^T j_N \tag{14}$$

For the study, we used two small, high-dimensional datasets – one artificial and one real – to test the effectiveness of our method.

The first dataset was generated using the function proposed by S. Haykin in [14]:

$$f(a,b) = (1-a^2) + 2(1-b)^2 \tag{15}$$

This feature is useful for creating artificial datasets for the purpose of modeling regression problems. The artificial sample we used consists of 30 instances, each of which has 2 input features and one target attribute.

The second dataset from [15] contains real-world data on certain femoral bone parameters of patients suffering from osteoarthritis. It contains 34 observations on six features:

- Patients' age;
- Patients' gender;
- Structure model index;
- Tissue porosity;
- Trabecular thickness factor;
- Compressive strength (in MPa) – target attribute.

Bone fragility is one of the common problems in medicine, so predicting bone compressive strength is quite important task.

Modeling, results, and comparison.

We applied the developed method to the two datasets described above. The following metrics were used to evaluate the model's performance:

- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean squared error (RMSE)
- Median error (MedE)
- Execution time (ExT)

For modeling, Python tools and its libraries were used: numpy, scipy, sklearn. Before the application procedure, the augmented sample was scaled using MaxAbsScaler. Optimal parameters σ_{1opt} and σ_{2opt} were selected on the interval [0.001;5] using the Dual Annealing optimization algorithm [16] for five measures of the distance. To evaluate the values of the metrics, 5-fold cross-validation was performed. The results of the experiments for the first dataset are given in Table 1, for the second dataset – in Table 2.

Table 1

Modeling results for the first dataset

Distance	σ_{1opt}	σ_{2opt}	**ExT, msec**	**MAPE**	**RMSE**	**MAE**	**MSE**	**MedE**
Cityblock	0,942	0,172	41,4±5,5	0,216±0.114	0,384±0.145	0,278±0.081	0,169±0.120	0,191±0.102
Euclidean	1,090	0,064	37,0±1,5	0,268±0.154	0,420±0.164	0,299±0.122	0,203±0.145	0,160±0.130
Chebyshev	0,891	0,043	20,6±0,6	0,267±0.137	0,423±0.154	0,311±0.111	0,203±0.136	0,191±0.134
Canberra	0,669	0,093	22,8±2,6	0,225±0.092	0,431±0.160	0,320±0.100	0,212±0.142	0,231±0.075
Cosine	4,205	0,001	26,4±4,5	0,382±0.125	0,653±0.132	0,522±0.101	0,444±0.193	0,407±0.111

Table 2

Modeling results for the second dataset

Distance	σ_{1opt}	σ_{2opt}	**ExT, msec**	**MAPE**	**RMSE**	**MAE**	**MSE**	**MedE**
Euclidean	0,153	0,059	25,0±1.4	0,395±0.173	5,184±1.366	4,022±1.169	28,742±13.690	3,090±1.592
Cityblock	0,183	0,293	24,2±1.5	0,434±0.189	5,412±1.395	4,201±1.206	31,237±14.784	3,133±1.575
Canberra	0,016	1,287	31,3±3.7	0,542±0.199	5,724±0.736	4,702±0.596	33,308±8.270	4,462±0.985
Chebyshev	0,983	0,116	25,7±2.9	0,497±0.158	5,733±0.948	4,543±0.755	33,761±10.554	3,655±1.343
Cosine	1,827	0,015	25,2±2.0	0,512±0.159	5,863±0.947	4,739±0.675	35,268±10.621	4,417±1.118

We compared the main results of the proposed method for both datasets with the performance of the simple GRNN model, as well as the input doubling method from [10], which inspired us. Overall, our proposed method outperformed the aforementioned models for both artificial and real datasets.

Fig. 2 visualizes the comparison of RMSE and MAE for the first (artificial) dataset. As can be seen, the use of augmentation methods reduced the errors compared to the classical GRNN. It is also worth noting that the proposed method demonstrates 25.4% lower RMSE and 32.7% lower MAE than the method from [10].

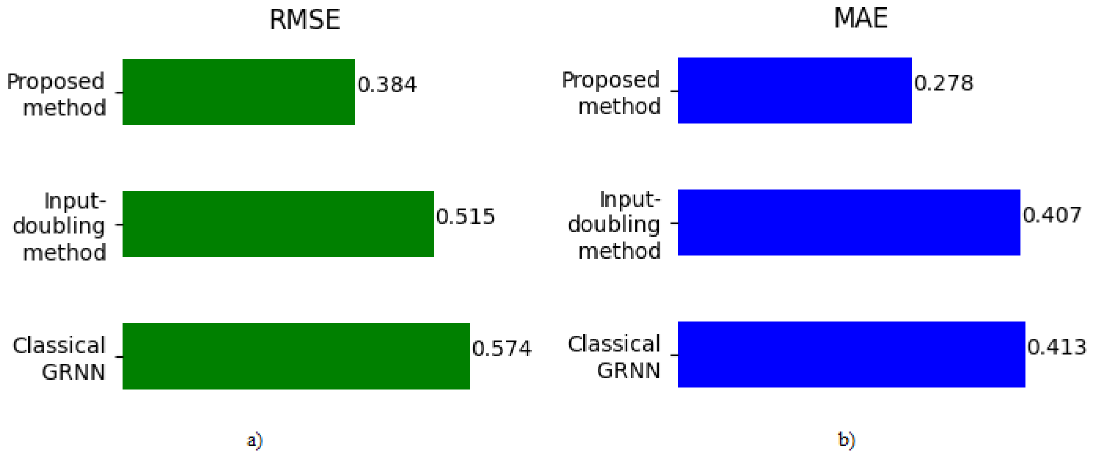

Fig. 2 Comparison of errors for the first dataset: a- RMSE; b-MAE

Fig. 3 visualizes the comparison of regression metrics of the used models for the second dataset. As can be seen, the proposed method showed 8.8% lower RMSE and 14.1% lower MAE than the method from [10]. The developed algorithm also demonstrated lower RMSE and MAE values than the classical GRNN.

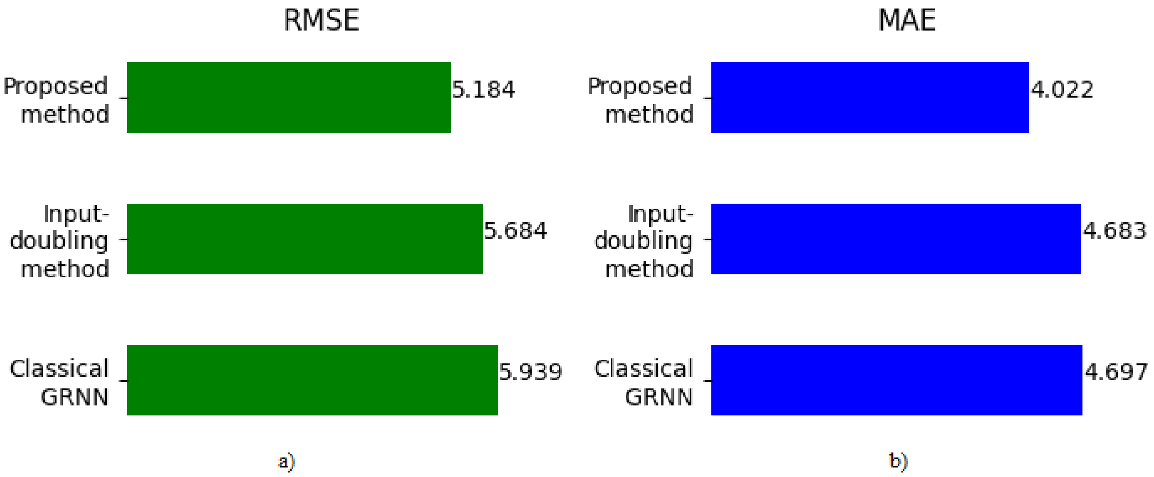

Fig. 3 Comparison of errors for the second dataset: a- RMSE; b-MAE

Another important indicator is the execution time of the algorithm. Fig. 4 shows a visualization of the comparison of the studied models in this aspect.

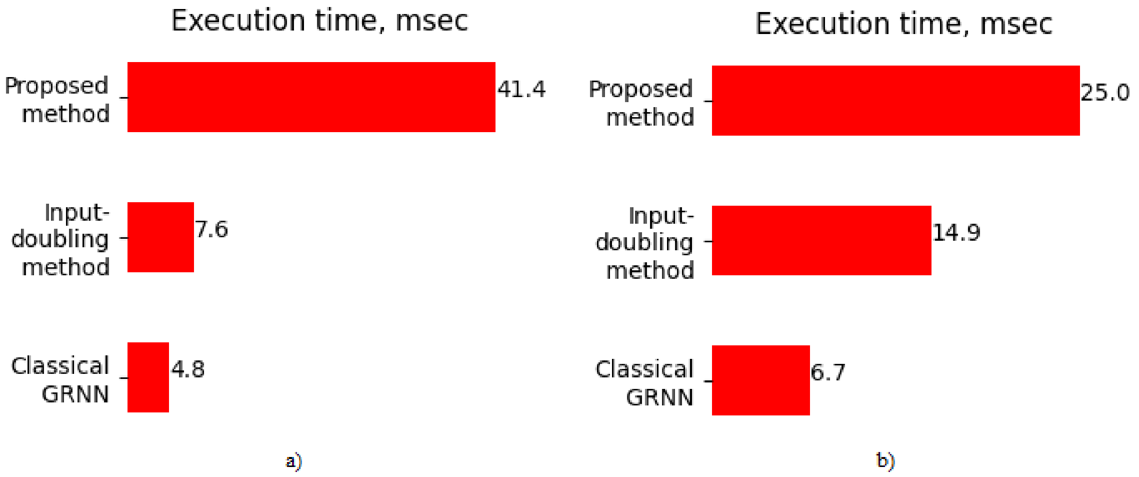

Fig. 4 Execution time comparison: a- for the first dataset; b- for the second dataset

The comparison demonstrated certain advantages and disadvantages of the developed method compared to the classical GRNN. On the one hand, the proposed augmentation algorithm increased the accuracy of the simple model. On the other hand, the execution time increased several times (depending on the task). From the point of view of comparison with the input doubling method, the developed algorithm significantly increased the prediction accuracy for the described tasks, while increasing the execution time. Obviously, the increase in duration is a consequence of using the classical GRNN model to obtain basic predictions in the first step of the proposed method.

From these results, we can conclude that the selection of a specific prediction model for practical purposes should be carried out depending on the priorities and specifics of the situation. In cases where prediction accuracy is key, it is necessary to use the algorithm with the smallest error. However, in situations where efficiency can be neglected in favor of execution speed, simpler models may be suitable.

Conclusions

In this paper, an improved method for intelligent analysis of short data sets based on a generalized regression neural network was proposed. The algorithm was applied to solve the problem of determining the compressive strength of bone. The optimal parameters were selected: distance metric and smoothing factor. The method was also tested on an artificial dataset.

According to the results of comparing the method with the classical GRNN and the basic method, it was found that the developed algorithm outperforms them in accuracy for both tasks. Along with improvement in accuracy, the proposed technique also increases execution time. Therefore, the appropriateness of its application depends on the priorities of the problem being solved. Further research may concern reducing the duration of the algorithm execution while maintaining satisfactory accuracy.

Acknowledgment. This research was funded by the European Union (through EURIZON H2020 project, grant agreement 871072).

References

1. 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

2. 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

3. 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

4. Tolstyak, Y., Chopyak, V., & Havryliuk, M. (2023). An investigation of the primary immunosuppressive therapy's influence on kidney transplant survival at one month after transplantation. *Transplant Immunology*, 78, 101832. https://doi.org/10.1016/j.trim.2023.101832

5. Izonin, I. (2023). An unsupervised-supervised ensemble technology with non-iterative training algorithm for small biomedical data analysis. *Computer Systems and Information Technologies*, 4, 67–74. https://doi.org/10.31891/csit-2023-4-9

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

7. 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

8. Andonie, R. (2010). Extreme data mining: Inference from small datasets. *International Journal of Computers, Communication & Control*. 5(3), 280-291. https://doi.org/10.15837/ijccc.2010.3.2481

9. Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., ... & Kurdi, H. (2021). Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, 11(2), 796. https://doi.org/10.3390/app11020796

10. 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

11. Izonin, I., Tkachenko, R., Shakhovska, N., & Lotoshynska, N. (2021). The additive input-doubling method based on the SVR with nonlinear kernels: Small data approach. *Symmetry*, 13(4), 612. https://doi.org/10.3390/sym13040612

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

13. 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

14. S. S. Haykin, Neural networks and learning machines, 3. ed. New York Munich: Prentice-Hall, 2009.

15. 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

16. Yemets, K. (2024). Methods for forecasting time series with strong seasonality based on transformers. *Herald of Khmelnytskyi National University. Technical sciences*, 333(2), 131-134. https://doi.org/10.31891/2307-5732-2024-333-2-20

Myroslav Havryliuk **Мирослав Гаврилюк**	Ph.D. student (Computer Science), at the Department of Artificial Intelligence, Lviv Polytechnic National University https://orcid.org/0000-0001-5259-7564 e-mail: myroslav.a.havryliuk@lpnu.ua	Аспірант кафедри систем штучного інтелекту Національного університету «Львівська політехніка»