Studia i Materiały Informatyki Stosowanej, Tom 15, Nr 2, 2023
str. 18-23

ARTIFICIAL INTELLIGENCE-BASED ANALYSIS OF SMALL DATA SETS IN MEDICINE

Dariusz Mikołajewski [1], Emilia Mikołajewska [2]

[1] *Uniwersytet Kazimierza Wielkiego*
Wydział Informatyki, Wydział Mechatroniki
ul Kopernika 1, 85-074 Bydgoszcz
e-mail: dariusz.mikolajewski@ukw.edu.pl
[1] *Uniwersytet Mikołaja Kopernika w Toruniu, Collegium Medicum im. L. Rydygiera w Bydgoszczy*
Wydział Nauk o Zdrowiu
ul Jagiellońska 13-15, 85-074 Bydgoszcz
e-mail: emiliam@cm.umk.pl

Abstarct: *AI-based computing of small data sets are a step towards edge computing and further personalization of diagnostics, therapy and predictions in clinical practice. However, this still requires many intermediate steps, both in hardware and software. The aim of the work is to assess to what extent current achievements in the area of AI-based small sets analysis constitute the basis for the development of a new group of clinical and programming solutions.*

Keywords: *Artificial intelligence, small data sets, clinical applications*

Analiza małych zbiorów danych medycznych oparta na sztucznej inteligencji

Streszczenie: *Obliczenia małych zbiorów danych w oparciu o sztuczną inteligencję stanowią krok w kierunku obliczeń brzegowych i dalszej personalizacji diagnostyki, terapii i prognoz w praktyce klinicznej. Jednak nadal wymaga to wielu etapów pośrednich, zarówno sprzętowych, jak i programowych. Celem pracy jest ocena, w jakim stopniu dotychczasowe osiągnięcia w obszarze analizy małych zbiorów w oparciu o sztuczną inteligencję stanowią podstawę do opracowania nowej grupy rozwiązań klinicznych i programistycznych.*

Słowa kluczowe: *Sztuczna inteligencja, małe zbiory danych, zastosowania kliniczne*

1. INTRODUCTION

The amount of data collected to date does not translate into an exponential increase in the knowledge we have. To date, research in the area of rare diseases, broadly defined, has rarely involved artificial intelligence (AI) methods and techniques, and in particular data-driven (machine learning - ML) methods, independently extracting rules/mechanisms linking input and output data. Small data sets in medicine for the purposes of AI-based analysis usually range in size from about 20 to several hundred, while medium-sized data sets range from 1-10 thousand, and big data: at least in the order of tens and hundreds of thousands. This was due, on the one hand, to the lack of homogeneity of the groups and, on the other: to the concentration of research on large data sets, so-called big data, often collected and analysed en masse, automatically, as part of routine procedures. In most studies, big data methods and techniques involve group sizes of tens of thousands or more and therefore favour population-based studies, including screening. In addition, it is easier to analyse numerical data, describing the result of the study directly, with clearly marked boundary values (positive, negative), allowing input and output data to be normalised and therefore achieving higher accuracies and lower mean squared error ((R)MSE) with a relatively simple neural network structure, even in the form of a multilayer perceptron (MLP) in traditional neural networks or a convolutional neural network (CNN) in deep learning with accuracies of 70-90%. Only since 2018 research and publications on the use of AI in clinical research on small groups (at the level of a dozen to several hundred patients, the so-called small data set) have started, especially in imaging studies [1-3]. This has been made possible by advanced techniques of image decomposition, analysis and subsequent assembly [4-6]. It seems reasonable to assume

that similar methods can be effective in studies combining laboratory results, patient phenotype and genotype, especially in children in whom other tests (even: screening) may be difficult to perform [7,8].
The aim of the work is to assess to what extent current achievements in the area of AI-based small sets analysis constitute the basis for the development of a new group of clinical and programming solutions.

2. RESEARCH GAPS

The fundamental research gaps identified remain:
- Developing new knowledge: optimal (in light of pre-established criteria: computational complexity, analysis time, accuracy, etc.) pathways for selecting the best method for analysing genetic and other data, including their segmentation and fusion, and segmentation performance metrics [9],
- To develop new knowledge: a pathway for selecting the optimal AI algorithms, methods and techniques (i.e. also fuzzy logic or multifractal analysis), not just ML) best for the analysis of a given condition, with an emphasis on their simplicity and accuracy,
- Guidance for further research on AI in small data set in genetics, neurology, psychiatry, geriatrics, but also rehabilitation and physiotherapy.

This allow, in the future, to expand the repertoire of available analyses and markers to include computational markers derived indirectly from the results of ongoing studies, but calculated in the extraction of relationships/mechanisms between input and output variables. Virtual screening of patients, including based on plausible data sequences, will become possible, speeding up the computation and shortening the search process - it will become cheaper and more accessible, and the resulting knowledge will be based on a solid, Essential research tools will include:
- Building and comparing simple algorithms for classification and/or prediction in the scientific Matlab 2023b environment, with the possibility of transferring the results of the best ones to Python and R in the Visual Studio Code environment,
- Building and comparing more complex algorithms in Python, e.g. based on XGBoost,
- Automating the best solutions using the ML.NET package in C# in MS Visual Studio 2022.

AI support of pre-clinical and clinical research can here extract new diagnostic pathways, but this is already beyond the scope of this study.

3. CURRENT SOLUTIONS

To date, few working models for small data sets have been developed. Analyses of small datasets are used to search databases for spectrally aligned sequences (e.g. for proteomic studies), i.e. to identify peptide sequences. This approach does not provide confidence or a probability assessment of the exact location (necessary for ana) when several possible sites are available. Localisation is absolutely required for further molecular biology analysis of cellular PTM function in vitro and in vivo. Therefore, we have developed PTMProphet, a free and open-source software tool integrated into the Trans-Proteomic Pipeline, which re-analyses identified spectra from any search engine for which pepXML output is available, to provide localisation confidence to enable appropriate further characterisation of biological events. Localisation of any type of mass modification (e.g. phosphorylation) is supported. PTMProphet uses Bayesian mixed models to calculate probabilities for each site/peptide spectrum match where a PTM is identified. These probabilities can be combined to calculate a global false localisation rate at any threshold to guide further analysis. We describe the PTMProphet tool, its basic algorithms and demonstrate its performance on synthetic peptide reference datasets, one previously published small dataset, one new larger dataset, and a previously published phosphoenriched data set [1]. A novel deep reinforcement learning model for single hormone (insulin) and dual hormone (insulin and glucagon) delivery in people with diabetes. their delivery are developed through double Q-learning with augmented recurrent neural networks based on the FDA-approved UVA/Padova Type 1 simulator. The generalized population model was personalized with a small set of patient-specific data, allowing accuracy to improve from 77, 6% to 80.9% for single-hormone control and up to 85.6% for dual-hormone control [2]. Lipid transfer inhibitor protein (LTIP) is an important regulator of cholesteryl ester transfer protein function, hence the development of an immunoassay for the quantitative determination of LTIP in plasma with different lipid content is very important. A negative association between plasma triglyceride (TG) levels and LTIP was confirmed based on a small data set, visible only in men. The mechanisms underlying this specific response require

further investigation [3]. Training an accurate generalized model on data consisting of multi-orientation cardiac MRI images is possible using a 3D deep learning method combining Transformers and U-Net based on a very small training dataset (150 cases: 90 cases for training and 60 cases for testing) from three different suppliers. Each set included two phases of the cardiac cycle and three movie sequences, and Dice and HD metrics were used to evaluate the segmentation performance. High segmentation accuracy, excellent correlation and consistency of function assessment were obtained, which confirmed that the tested solution is a fast and effective method for examining heart MRI and heart diseases [4].

The main limitation of deep learning is its computational complexity, which translates into the required high-performance computing resources and long training time. From the above For this reason, most of the described solutions require sending data via the cloud and processing it on remote supercomputers. Using ready-made models, e.g. on mobile devices, requires training them on external computers and transferring them to mobile devices, which makes it difficult to train them outside of technical breaks. Network delays, lack of standard data formats, and cybersecurity (including patient data privacy) are therefore significant technical problems to solve. A solution is to individually train several pre-trained general machine learning models at the same time, each of them trained on its own data, without sharing it, and then aggregate them in a consensus model. Using explainable ethical AI to infer research results avoids the "black box" perception of neural network models. The use of simpler algorithms and edge computing makes this task easier compared to existing cloud models. This is supported by advanced high-throughput feature analysis techniques for faster and more accurate extraction of diagnostic or predictive information based on interpretation related to the biological mechanism of the injury/disease. Acquiring larger data sets to create deeper networks in this context only makes a lot of indirect sense: centralizing them can be a burden. Big data does not always mean new knowledge, especially in relation to rare diseases and personalized diagnostics [6]. There is hope in genome relatedness matrix (GRM) approaches that allow for improved social lineage approaches [7], Urinary P concentration (Pu) and creatinine in cows were analysed using a linear mixed model on a small dataset [8**].** The Bayesian model in medical diagnostics can be used to solve the problem of finding the globally optimal logical combination of classifiers (e.g. subcellular localisation of proteins based on estimates of their sensitivity and specificity compared to estimates made using the gold standard). For all models, running times were acceptable and results were accurate. The methods are suitable for both small and large datasets, for solving a wide range of bioinformatics classification problems, and are robust to dependencies between classifiers [9].

4. DISCUSSION

Researchers use deep learning for modeling tasks on small data sets, often relying on decomposition and convolutional neural networks (CNN), a type of deep learning that is particularly effective at classifying images while ensuring repeatability. AI allows both the automation of human tasks (view recognition, image segmentation, assessment of standardized structural and functional parameters of the heart) and advanced identification of patterns in image data (discovering new associations, phenotypes, predicting results and supporting clinical decisions).

4.1. Limitations

Although deep learning models have high performance, they are prone to learning errors and poor generalization. Other limitations of AI-based models using small data sets in medicine include: The solution is to validate deep learning algorithms on sets other than the main training set. Key challenges associated with using AI on small data sets in the medical field are following:

1. - Low statistical significance because without special processing methods (e.g. fragmentation into overlapping elements) small data sets may not provide statistically significant results, and the results themselves may be more susceptible to random variation, making it difficult to draw reliable conclusions;

- Limited feature space, as small datasets may not cover the full spectrum of factors relevant to health status;

2. - Bias and variability as small datasets may contain inadvertent errors or be influenced by variability that is not representative of the wider population, which can lead to biased predictions;
3. - Limited generalizability because small datasets may not represent the diversity and complexity of the entire patient population;
4. - Higher risk of overfitting.

Addressing the aforementioned limitations requires both interdisciplinary collaboration, data sharing and the development of standardised small data sets in clinical practice. Removing limitations and discovering new

opportunities can significantly increase the use and effectiveness of artificial intelligence in healthcare.

4.2. Directions for further research

Research on artificial intelligence-based small data analysis in medicine is an emerging and key area of research. At the initial stage of development of this group of methods, it is proposed to make data and models available to the scientific community whenever possible. Such comparative evaluations provide useful additional information, as suboptimal training data sets may affect the generated algorithm and require optimal adaptation of the algorithm and confirmation of the results.

Research into the analysis of small data sets based on artificial intelligence in medicine is a growing and crucial field. Exploring new possibilities in this area could significantly increase the application and effectiveness of artificial intelligence in healthcare, especially for rare conditions. Here are some directions for further research:

- Investigate and develop robust data augmentation techniques to artificially expand small datasets by, for example, generating additional training samples, semi-automatically and automatically applying transformations, simulations or other methods to existing data;
- Explore transfer learning and pre-training methods in which models pre-trained on larger datasets are fine-tuned on smaller medical datasets, helping to apply knowledge gained from wider datasets and improve the accuracy achieved in specific clinical tasks on small groups of patients;
- Explore Bayesian methods that include uncertainty estimation useful for small datasets where uncertainty in predictions is critical for informed decision-making;
- Explore semi-supervised and self-supervised learning approaches that can use both labelled and unlabelled data;
- Exploring ensemble learning techniques that combine predictions from multiple models, yielding improved generalisation generalizability and robustness to data-limited drawbacks;
- Implement a strategy to intelligently select the most informative samples for prioritising hybrid data analysis and improving model performance with a limited number of labelled samples;
- Investigate domain adaptation techniques such that the accuracy of the model in the target domain (based on a small medical dataset) can be improved by using knowledge from a related source domain (based on a larger dataset from a similar domain);
- Benefit from previous experience in analysing small and medium-sized data sets with distinct, characteristic features, e.g. from gait analysis or hand movements after stroke [10,11] or repetitive movement patterns in patients with autism spectrum disorders [12,13];
- Implement a human-in-the-loop approach in which AI systems work directly with medical experts, making more effective use of clinicians' knowledge and experience in learning and improving the reliability of AI predictions;
- Automation and/or optimisation of e-health procedures in a manner consistent with current Industry 4.0 or Industry 5.0 (personalised mass production) paradigms [14-18].

By exploring these directions, researchers can contribute to the development of more robust, reliable and ethical applications of artificial intelligence in medicine, even when working with small data sets. The next stage in the development of this group of solutions is edge computing: the basic process of analyzing data (including images) as close to the end device as possible, with reduced delays, threats to data security and the costs of managing large data sets [19-22]. However, this requires further simplification of network architectures and placing AI algorithms directly in diagnostic devices and even in patients' clothing and equipment. This could reduce research time and provide a range of data analyzes without having to upload it to the cloud. Hence, a general machine learning model can be trained on data specific to a single patient to make these models increasingly individualized [23-24].

5. CONCLUSIONS

AI-based computing of small data sets are a step towards edge computing and further personalization of diagnostics, therapy and predictions in clinical practice. However, this still requires many intermediate steps, both in hardware and software.

Literatura

5. Shteynberg D.D, Deutsch E.W., Campbell D.S, Hoopmann M.R., Kusebauch U., Lee D., Mendoza, L. Midha M.K., Sun Z., Whetton A.D., Moritz R.L. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J Proteome Res. 2019; 18(12), 4262-4272. doi: 10.1021/acs.jproteome.9b00205.
6. Zhu T., Li K., Herrero P., Georgiou P. Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An In Silico Validation. IEEE J Biomed Health Inform. 2021; 25(4), 1223-1232. doi: 10.1109/JBHI.2020.3014556.

7. Morton R.E., Gnizak H.M., Greene D.J., Cho K.H., Paromov V.M. Lipid transfer inhibitor protein (apolipoprotein F) concentration in normolipidemic and hyperlipidemic subjects. J Lipid Res. 2008; 49(1), 127-35. doi: 10.1194/jlr.M700258-JLR200.
8. Wang J., Zhang N., Wang S., Liang W., Zhao H., Xia W., Zhu J., Zhang Y., Zhang W., Chai S. AI approach to biventricular function assessment in cine-MRI: an ultra-small training dataset and multivendor study. Phys Med Biol. 2023; doi: 10.1088/1361-6560/ad0903.
9. Lima J.A.C., Venkatesh B.A. Building Confidence in AI-Interpreted CMR. JACC Cardiovasc Imaging. 2022; 15(3), 428-430. doi: 10.1016/j.jcmg.2021.10.008.
10. Sengupta P.P., Chandrashekhar Y., Imaging With Deep Learning: Sharpening the Cutting Edge. JACC Cardiovasc Imaging. 2022; 15(3), 547-549. doi: 10.1016/j.jcmg.2022.02.001.
11. Perrier C., Delahaie B., Charmantier A. Heritability estimates from genomewide relatedness matrices in wild populations: Application to a passerine, using a small sample size. Mol Ecol Resour. 2018; 18(4), 838-853. doi: 10.1111/1755-0998.12886.
12. Løvendahl P., Sehested J. Short communication: Individual cow variation in urinary excretion of phosphorus. J Dairy Sci. 2016; 99(6), 4580-4585. doi: 10.3168/jds.2015-10338.
13. Keith J.M., Davey C.M., Boyd S.E. A Bayesian method for comparing and combining binary classifiers in the absence of a gold standard. BMC Bioinformatics. 2012; 13, 179. doi: 10.1186/1471-2105-13-179.
14. Prokopowicz P., Mikołajewski D., Tyburek K., Mikołajewska E. Computational gait analysis for post-stroke rehabilitation purposes using fuzzy numbers, fractal dimension and neural networks. Bulletin of the Polish Academy of Sciences: Technical Sciences, 2020, 68 (2), 191-198, DOI: 10.24425/bpasts.2020.13184.
15. Mikołajewska E., Prokopowicz, P., Mikolajewski D. Computational gait analysis using fuzzy logic for everyday clinical purposes–preliminary findings. Bio-Algorithms and Med-Systems 2017, 13 (1), 37-42.
16. Duch W., Nowak W., Meller J., Osiński G., Dobosz K., Mikołajewski D., Wójcik G.M. Computational approach to understanding autism spectrum disorders. Computer Science 2014, 13 (2), 47-47.
17. Duch W., Nowak W., Meller J., Osiński G., Dobosz K., Mikołajewski D. Consciousness and attention in autism spectrum disorders. Proceedings of Cracow Grid Workshop 2010, 202-211.
18. Macko M., Szczepański Z., Mikołajewski D., Mikołajewska E., Listopadzki S. The method of artificial organs fabrication based on reverse engineering in medicine. Proceedings of the 13th International Scientific Conference: Computer Aided Engineering Springer 2017, 353-365.
19. Mikołajewska E., Mikołajewski D. Roboty rehabilitacyjne. Rehabil. Prakt 2010, 4, 49-53.
20. Mikołajewska E., Mikołajewski D. Zastosowania automatyki i robotyki w wózkach dla niepełnosprawnych i egzoszkieletach medycznych. Pomiary Automatyka Robotyka 2011, 15(5), 58-63.
21. Rojek I., Mikołajewski D., Macko M., Szczepański Z., Dostatni E. Optimization of extrusion-based 3D printing process using neural networks for sustainable development. Materials 2021, 14 (11), 2737.
22. Rojek I., Mikołajewski D., Kotlarz P., Macko M., Kopowski J. Intelligent system supporting technological process planning for machining and 3D printing. Bulletin of the Polish Academy of Sciences. Technical Sciences 2021, 69 (2), e136722, DOI: 10.24425/bpasts.2021.136722.
23. Galas K., Efektywność klasyfikacji mrugnięcia z wykorzystaniem wybranych sieci neuronowych. Studia i Materiały Informatyki Stosowanej 2021, 13(1), 11-16.
24. Piszcz A., BCI w VR: imersja sposobem na sprawniejsze wykorzystywanie interfejsu mózg-komputer. Studia i Materiały Informatyki Stosowanej 2021, 13(1), 5-10.