Review

SCIENCE PROGRESS

Science Progress
2022, Vol. 105(1) 1–16

DOI: 10.1177/00368504211029777
journals.sagepub.com/home/sci

Machine learning on small size samples: A synthetic knowledge synthesis

Peter Kokol[1], Marko Kokol[2] and Sašo Zagoranski[2]

[1]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia
[2]Semantika, Maribor, Slovenia

Abstract

Machine Learning is an increasingly important technology dealing with the growing complexity of the digitalised world. Despite the fact, that we live in a 'Big data' world where, almost 'everything' is digitally stored, there are many real-world situations, where researchers are still faced with small data samples. The present bibliometric knowledge synthesis study aims to answer the research question 'What is the small data problem in machine learning and how it is solved?' The analysis a positive trend in the number of research publications and substantial growth of the research community, indicating that the research field is reaching maturity. Most productive countries are China, United States and United Kingdom. Despite notable international cooperation, the regional concentration of research literature production in economically more developed countries was observed. Thematic analysis identified four research themes. The themes are concerned with to dimension reduction in complex big data analysis, data augmentation techniques in deep learning, data mining and statistical learning on small datasets.

Keywords

Machine learning, small data sets, knowledge synthesis, bibliometrics

Introduction

Periods of scientific knowledge doubling have become significantly shorter and shorter.[1–3] This phenomenon, combined with information explosion, fast cycles of technological innovations, Open Access and Open Science movements,[4] and new Web/Internet based methods of scholarly communication[5] have immensely increased the complexity and effort needed to synthesis scientific evidence and

Corresponding author:
Peter Kokol, Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška ulica 46, Maribor 2000, Slovenia.
Email: peter.kokol@um.si

knowledge.[6,7] However, above phenomena also resulted in the growing availability of research literature in a digital, machine-readable format.[8]

To solve the emerging complexity of knowledge synthesis and the simultaneous new possibilities offered by digital presentation of scientific literature, Blažun et al.[9] and Kokol et al.[10,11] developed a novel synthetics knowledge synthesis methodology based on the triangulation of (1) distant reading,[12] an approach for understanding the canons of literature not by close (manual) reading, but by using computer based technologies, like text mining and machine learning, (2) bibliometric mapping[13] and (3) content analysis.[14,15] Such triangulation of technologies enables one to combine quantitative and qualitative knowledge synthesis in the manner to extend classical bibliometric analysis of publication metadata with the machine learning supported understanding of patterns, structure and content of publications.[16,17]

One of the increasingly important technologies dealing with the growing complexity of the digitalisation of almost all human activities is the Artificial intelligence, more precisely machine learning.[18–22] Despite the fact, that we live in a 'Big data' world,[23,24] where almost 'everything' is digitally stored, there are many real world situation, where researchers are faced with small data samples.[25–27] Hence, it is quite common, that the database is limited for example by the number of subjects (i.e. patients with rare disseises), the sample is small comparing to number of features like in genetics or biomarkers detection, sampling, there is a lot of noisy or missing data or measurements are extremely expensive or data imbalanced meaning that the size of one class in a data set has very few objects.[28–32]

Using machine learning on small size datasets present a problem, because, in general, the 'power' of machine learning in recognising patterns is proportional to the size of the dataset, the smaller the dataset, less powerful and less accurate are the machine learning algorithms. Despite the commonality of the above problem and various approaches to solve it we didn't found any holistic studies concerned with this important area of the machine learning field. To fill this gap, we used synthetic knowledge synthesis presented above to aggregate current evidence relating to the 'small data set problem' in machine learning. In that manner we can overcome the problem of isolated findings which might be incomplete or might lack possible overlap with other possible solutions. In our analysis, we aimed to extract, synthesis and multidimensionally structure the evidence of as possible complete corpus of scholarship on small data sets. Additionally, the study seeks to identify gaps which may require further research.

Methodology

The study aim is to answer the following research question: *What is the small data problem in machine learning and how it is solved?*

Synthetic knowledge synthesis was performed following the steps below:

1. Harvest the research publications concerning small data sets in machine learning to represent the content to analyse.

2. Condense and code the content using text mining.
3. Analyse the codes using bibliometric mapping and induce the small data set research cluster landscape.
4. Analyse the connections between the codes in individual clusters and map them into sub-categories.
5. Analyse sub-categories to label cluster with themes.
6. Analyse sub-categories to identify research dimensions
7. Cross-tabulate themes and research dimension and identify concepts

The research publications were harvested from the Scopus database, using the advance search using the command *TITLE-ABS-KEY ((‘small database’ OR ‘small dataset’ OR ‘small sample’) AND ‘machine learning’ AND NOT (‘large database’ OR ‘large dataset’ OR ‘large sample’))*, The search was performed on 7th of January, 2021. Following metadata were exported as a CSV formatted corpus file for each publication: Authors, Authors affiliations, Publication Title, Year of publication, Source Title, Abstract and Author Keywords.

Bibliometric mapping and text mining were performed using the VOSViewer software (Leiden University, Netherlands). VOSViewer uses text mining to recognise publication terms and then employs the mapping technique called Visualisation of Similarities (VoS), which is based on the co-word analysis, to create bibliometric maps or landscapes.[33] Landscapes are displayed in various ways to present different aspects of the research publications content. In this study the content analysis was performed on the author keywords cluster landscape, due to the fact that previous research showed that authors keywords most concisely present the content authors would like to communicate to the research community.[34] In this type of landscape the VOSviewer merges author keywords which are closely associated into clusters, denoted by the same cluster colours.[35] Using a customised Thesaurus file, we excluded the common terms like study, significance, experiment and eliminated geographical names and time stamps from the analysis.

Results and discussion

The search resulted in 1254 publications written by 3833 authors. Among them there were 687 articles, 500 conference papers, 33 review papers, 17 book chapters and 17 other types of publications. Most productive countries among 78 were China ($n = 439$), United States ($n = 297$), United Kingdom ($n = 91$), Germany ($n = 63$), India ($n = 52$), Canada ($n = 51$) and Spain ($n = 44$), Australia ($n = 42$), Japan ($n = 34$), South Korea ($n = 30$) and France ($n = 30$). Most productive institutions are located in China and United States. Among 2857 institution, China Academy of Sciences is the most productive institution ($n = 52$), followed by Ministry of Education China ($n = 28$), Harvard Medical School, USA ($n = 14$), Harbin Institute of Technology, China ($n = 14$), National Cheng Kung University, China,[13] Shanghai University, China ($n = 13$), Shanghai Jiao Tong University, China ($n = 12$), Georgia Institute of Technology, USA ($n = 12$) and Northwestern

Polytechnical University, China ($n = 11$). The first non China/USA institution is the University of Granada ($n = 8$) on 19th rank. Majority of the productive countries are member of G9 countries, with strong economies and research infrastructure.

Most prolific source titles (journals, proceedings, books) are the Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics ($n = 64$), IEEE Access ($n = 17$), Neurocomputing ($n = 17$), Plos One ($n = 17$), Advances in Intelligent Systems and Computing ($n = 14$), Expert System with application ($n = 14$), ACM International Conference Proceedings Series ($n = 13$) and Proceedings Of SPIE The International Society For Optical Engineering ($n = 13$). The more productive source titles are mainly from the computer science research area and are ranked in the first half of Scimago Journal ranking (Scimago, Elsevier, Netherlands). In average their h-index is between 120 and 380.

First two papers concerning the use of machine learning on small datasets indexed in Scopus were published in 1995.[36,37] After that the publications were rare till 2002, trend starting to rise linearly in 2003, and exponentially in 2016 (Figure 1– lines). In the beginning till the year 2000 all publications were published in journals, after that conference papers started to emerge. Uninterrupted production of review papers started much later in 2016. The number of citations started to grow exponentially in 2005 (Figure 1– bar chart), reaching the peak in 2020.

According to the ratio between articles, reviews and conference paper and the exponential trend in production for the first two we asses that machine learning on small datasets in the third stage of Schneider scientific discipline evolution model.[38] That means that the terminology and methodologies are already highly developed, and that domain specific original knowledge generation is heading toward optimal research productivity.

On the other hand Pestana et al.[39] characterised the scientific discipline maturity as a set of extensive repeated connections between researchers that collaborate on the publication of papers on the same or related topics over time. In our study we analysed those connection with the co-authors networks induced by VOSviewer. The co-author network showed that among countries with the productivity of 10 or more papers ($n = 27$) presented in Figure 2 an elaborate international co-authorship network emerged, confirming that maturity is already reached a high level. The most intensive co-operation is between United States and China, and the 'newest' countries joining the network are Russian federation, South Korea and Brazil. Research collaboration is important because it results in a higher quality and quantity of scientific output.[40]

The content analysis of keyword cluster landscape shown in Figure 3. Resulted in codes, SUB-categories and themes presented in Table 1. The content analysis revealed four prevailing themes. The largest two themes are related to dimension reduction in complex big data analysis and data augmentation techniques in deep learning.

The further analysis of the Table 1, revealed four categories, which we call Research dimensions. Four identified dimensions are: Small data set problem,

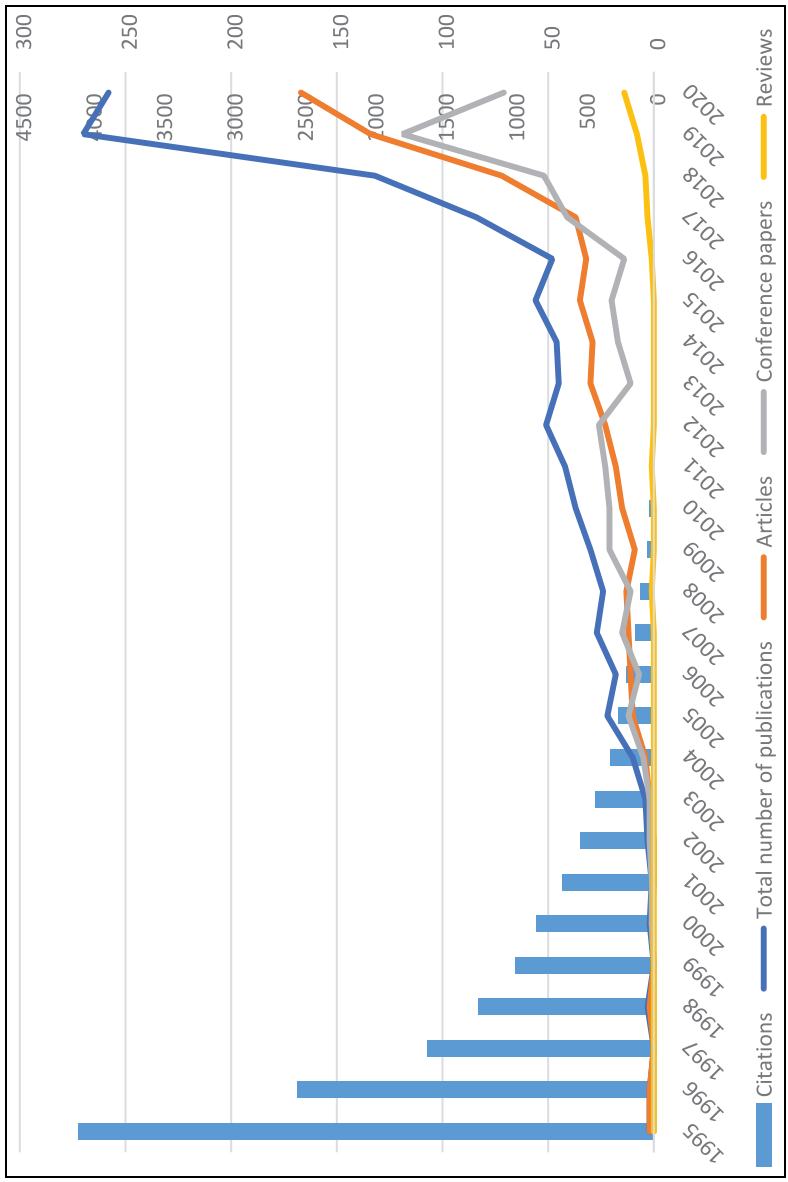

Figure 1. The research literature production dynamics.

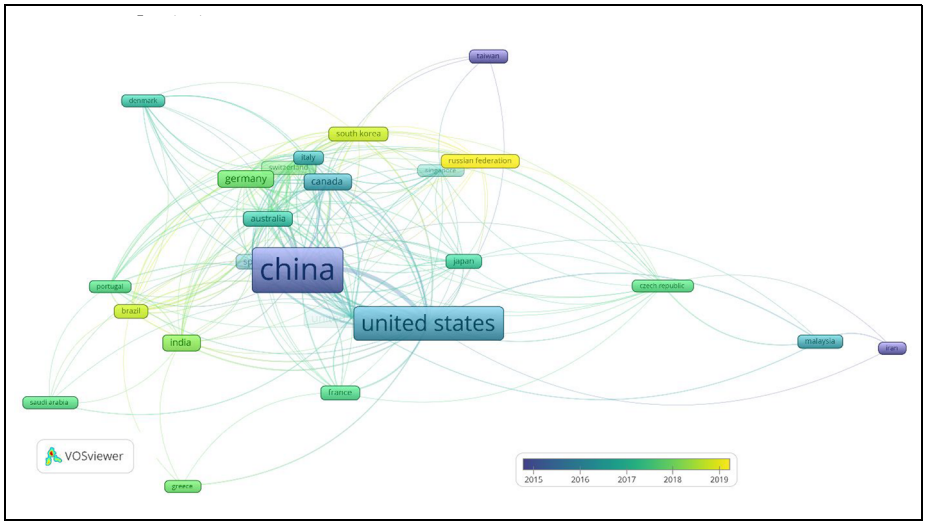

Figure 2. The co-authorship network.

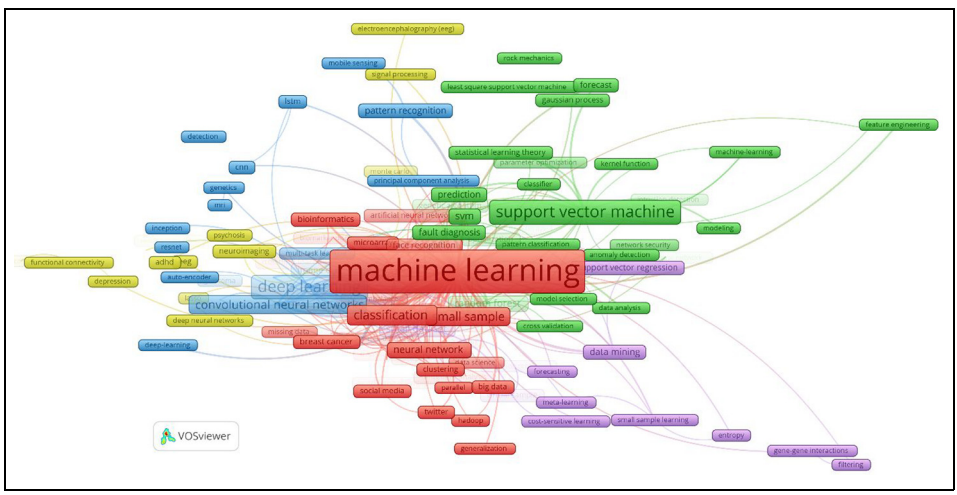

Figure 3. The author keyword cluster landscape.

Machine learning algorithms, Small-data pre-processing technique and Application Area. Cross tabulation of themes and research dimension resulted in a taxonomy presented in Table 2. The entries in the table presents the most popular concepts in each taxonomical entity.

Table 1. Small size sample research themes.

Theme	Colour	More frequent codes	Prevailing sub-categories
Dimension reduction in complex big data analysis	Red	Machine learning (339), Classification (51), Feature selection (50), Artificial neural network (28), Neural network (19), Natural language processing (16), Ensemble learning (14), Bioinformatics (13), Decision tree (12), Breast cancer (11), Machine learning algorithms (10), Computer vision (10), Microarray (9), Small datasets (9), Text classification (9), Big data (9), Dimensionality reduction (9), Social media (7),	Machine learning algorithms in bioinformatics, Classification on small samples with feature selection, Ensemble learning for solving under-sampling in personalised medicine, Solving missing data in breast cancer using deep neural networks, Dimensionality reduction. with feature extraction in cancer classification, Natural language processing in social media and text analysis
Data augmentation techniques in deep learning for pattern recognition and classification	Blue and yellow	Deep learning (89), Convolutional neural networks (45), Transfer learning (38), Pattern recognition (14), Image classification (14), data augmentation (12), Generative adversarial networks (8), LSTM (6), Resnet – Residual neural network ($n = 4$)	Deep and transfer learning using in image classification using data augmentation and synthetic data, complex neural networks, Pattern recognition in mobile sensing using Principal component analysis and LTSM, Resnet and Desnet based deep learning using auto encoder on case of glaucoma,
Support vector machines, Random forest and Genetic algorithms in statistical learning	Green	Support vector machines (109), Random forest (24), Prediction (18), Fault diagnosing (18), Forecast (10), Genetic algorithms (8), Regression (8), Gaussian process (7), Optimisation (7), Time series (6), Statistical learning theory (6)	Support vector machines and random forests in prediction, forecasting and fault diagnosis, Genetic algorithms in optimisation and time series analysis
Data mining on small datasets with support vectors regression	Violet	Small datasets (2), Data mining (15), Support vector regression (11), Virtual sample (7)	Small datasets augmentation with virtual samples, Data mining with support vector regression

Numbers in parenthesis present the number of papers in which an author keyword occurred

Table 2. Taxonomy of themes and research dimensions categories.

Theme	Small data size problem	Machine learning algorithms	Small – data pre-processing technique	Application area
Dimension reduction in complex big data analysis	Missing data (3), Unbalanced data (8), Under sampling (3)	Bagging (9), Bayesian networks (7), Boosting (6), Decision trees and forests (43), deep neural network (12), Ensemble learning (21), manifold learning (5), neural network (30), naïve Bayes (2), semi supervised learning (17), supervised learning (19), unsupervised learning (4), FMRI (7)	Dimensionality reduction (9), Feature extraction (12), engineering (4), fusion (2), Lasso (6), Linear and multiple discriminant analysis (13), PCA (16), autoencoder (10)	Image analysis (18), genetics and bioinformatics (27), personalised and precision medicine (6) neuroimaging (7), mental health (25)
Data augmentation for deep learning in pattern recognition and classification		Convolutional neural network (42), Dense CNN (3), Generative adversarial networks (8), LSTM (6), Resnet (4), transfer learning (40), semi-supervised learning (17)	Data augmentation (12), Domain adaption (5), data fusion (3), synthetic data (3), autoencoder (10)	Image classification (22), genetics (3), mobile sensing (10), predictive modelling (6), semantic segmentation (3)
Statistical based machine learning	Small sample (98)	Support vector machines (89). Least square SVM (10), genetic algorithms, random forest (24), Particle swarm (27), support vector regression (11)	Feature engineering (3), meta-learning (5), Monte Carlo (3), Small sample learning (5), virtual sample (11)	Forecasting (13), Fault diagnosing (18), regression (19), prediction (26), pattern recognition (17), rock mechanics (5), security 7), computer vision (10), data mining (15)
Data mining on small datasets	Big data (9)			Text and social media analysis (21), Precision medicine (3), Natural language processing (16), Sentiment analysis (9), clustering (8)

Numbers in parenthesis present the number of publications.

The most frequently reported difficulties causing the small data problem are the small size of the dataset, high/low dimensionality of datasets and unbalanced data. Small size datasets can cause problems when machine learning is applied in material sciences,[41] engineering[42,43] and various omics fields[44] due to the high cost of sampling; differentiating between autistic and non-autistic patients[29] and diagnosing rare diseases[45] due to the small number of available patients, or unavailability of other subjects for example PhD students in prediction of their grades;[46] and new pandemic prediction due fact that samples are scare when pandemic occurs and there is lack of medical knowledge.[47] Similar reasons can also lead to too high or low dimensionality of the datasets. The first case can occur even if the sample size is not small, however the ratio between number of features and the sample size is large, like for example in particle physics[48] or bioinformatics.[49] The second case can occur when the number of instances is not problematic, however the number of features is very small, like for example in characterisation of high-entropy alloys.[50] Unbalanced data present a long standing problem in machine learning and still remains a challenge in various applications like face recognition, credit scoring, fault diagnosing and anomaly detection where mayor class has much more instances than one or more of remaining classes.[51–53]

The most used machine learning algorithms used on small datasets are support vector machines,[54–56] decision trees/forests,[57,58] convolutional neural networks[59–61] and transfer learning.[62,63]

Most frequently employed data pre-processing techniques to overcame the small size problem are linear and nonlinear Principal component analysis[43,64,65] Discriminant analysis,[46,66,67] Data augmentation,[47,68,69] Virtual sample,[70–72] Feature extraction[50,73] and Auto-encoder.[43,74]

Most affected areas are Bioinformatics,[44,75,76] image classification and analysis,[68,69,77] fault diagnosing,[78,79] forecasting and prediction,[47,80] social media analysis[81,82] and health care[83,84]

Our analysis also revealed that small datasets may result from hardware requirements to cope with limited processing power or small storage size of devices like for example Raspberry PI.[85] Some research also shows that high quality small sample can be better than a low quality large sample in the case of statistical machine learning.[86]

Strengths and limitations

The main strength of the study is that it is the first bibliometrics and content analysis of the small dataset research. One of the limitation is that the analysis was limited to publications indexed in Scopus only, however Scopus is indexing the largest and most complete set of information titles, thus assuring the integrity of the data source. Additionally, the analysis also included qualitative components, which might have introduced slight bias to the results of our study.

Conclusion

Our bibliometric study showed the positive trend in the number of research publications concerning the use of small dataset and substantial growth of research community dealing with the small dataset problem, indicating that the research field is moving toward the higher maturity levels. Despite notable international cooperation, regional concentration of research literature production in economically more developed countries was observed. The content analysis showed that the small data sample challenge is recently mainly tackled with more complex machine learning approaches like Deep learning and Support vector machines. The main hot-topic areas of application are medicine/health care, material sciences and economy.

The results of study present a multi-dimensional facet and science landscape of the small datasets problem, which can help small data sample community to solve theoretical and practical challenges. Machine learning researchers and practitioners can use the study results to improve their understanding of the area and can catalyse their further knowledge development. On the other hand, it can inform novice researchers, interested readers or research mangers and evaluators without specific knowledge and help them to develop a perspective on the most important small dataset research dimensions. Finally, the study output can serve as a guide to further research and a starting point to more formal knowledge synthesis endeavours like systematic reviews and meta-analyses.

Our analysis reveals that in the world where everything should be automated and digitalised the possible future of small data analysis lies in the slogan 'the future of big data is small data'. It is speculated that on 1000 big data sets there are millions of small data sets, especially in fields where extensive manual data preprocessing have to be made, that is, contract reviews or financial audits automation, where manual and extremely expensive annotation performed by layers or accountants are needed. Hence, the future might be in the so called transfer machine learning, where learning could be generalised on data sets from various fields and many different small data sets might become a big data set.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Peter Kokol https://orcid.org/0000-0003-4073-6488

References

1. Bornmann L, de Moya Anegón F and Leydesdorff L. Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS One* 2010; 5(10): e13327.
2. McDeavitt JT. Medical education: toy airplane or stone flywheel? http://wingofzock.org/2014/12/23/medical-education-toy-airplane-or-stone-flywheel/ (2014, accessed 29 Apr 2017).
3. R. Buckminster Fuller. *Critical path*. London, UK: Century Hutchinson Ltd, 1983.
4. Moore SA. Revisiting "the 1990s debutante": Scholar-led publishing and the prehistory of the open access movement. *J Assoc Inf Sci Technol* 2020; 7(7): 856–866.
5. Fathalla S, Vahdati S, Lange C, et al. Analysing scholarly communication metadata of computer science events. In: Kamps J, Tsakonas G, Manolopoulos Y and et al (eds) *Research and advanced technology for digital libraries*. Cham: Springer International Publishing, 2017, pp.342–54. (Lecture Notes in Computer Science).
6. Kokol P. Meta approaches in knowledge synthesis in nursing: a bibliometric analysis. *Nurs Outlook*. In press.
7. Tricco AC, Garritty CM, Boulos L, et al. Rapid review methods more challenging during COVID-19: commentary with a focus on 8 knowledge synthesis steps. *J Clin Epidemiol* 2020; 126: 177–183.
8. Burda D and Teuteberg F. Sustaining accessibility of information through digital preservation: A literature review. *J Inf Sci* 2013; 39(4): 442–458.
9. Blažun Vošner H, Železnik D, Kokol P, et al. Trends in nursing ethics research: mapping the literature production. *Nurs Ethics* 2017; 24(8): 892–907.
10. Kokol P, Zavrsnik J, Turcin M, et al. Enhancing the role of academic librarians in conducting scoping reviews. Library Philosophy and Practice (e-journal), https://digitalcommons.unl.edu/libphilprac/4293 (2020, 17 April 2021).
11. Kokol P, Zagoranski S and Kokol M. Software development with scrum: a bibliometric analysis and profile. *Library Philosophy and Practice (e-journal)*. https://digitalcommons.unl.edu/libphilprac/4705 (2020).
12. Moretti F. *Distant reading*. London: Verso, 2013.
13. Noyons E. Bibliometric mapping of science in a science policy context. *Scientometrics* 2001; 50(1): 83–98.
14. Kyngäs H, Mikkonen K and Kääriäinen M. *The application of content analysis in nursing science research*. Berlin, Germany: Springer, 2020.
15. Lindgren B-M, Lundman B and Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. *Int J Nurs Stud* 2020; 108: 103632.
16. Atanassova I, Bertin M and Mayr P. Editorial: mining scientific papers: NLP-enhanced bibliometrics. *Front Res Metr Anal* 2019; 4: 2.
17. Haddaway NR, Collins AM, Coughlin D, et al. The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLoS One* 2015; 10(9): e0138237.
18. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. *Eur J Intern Med* 2018; 48: e13–e14.
19. Grant K, McParland A, Mehta S, et al. Artificial intelligence in emergency medicine: surmountable barriers with revolutionary potential. *Ann Emerg Med* 2020; 75(6): 721–726.

20. Jordan MI and Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015; 349(6245): 255–260.
21. Mohri M, Rostamizadeh A and Talwalkar A. *Foundations of machine learning*. 2nd ed. Cambridge, MA: MIT Press, 2018. p.505.
22. Alpaydin E. *Machine learning: The new AI*. Cambridge, MA: MIT Press, 2016, p.225.
23. Grover V. Do we need to understand the world to know it? Knowledge in a big data world. *J Glob Inf Technol Manage* 2020; 23(1): 1–4.
24. Anderson KM. Embrace the challenges: software engineering in a big data world. In: *2015 IEEE/ACM 1st international workshop on big data software engineering*, Florence, Italy, 23–May 2015, pp.19–25. New York: IEEE.
25. Shu J, Xu Z and Meng D. Small sample learning in big data era. *arXiv: 180804572* [cs, stat], http://arxiv.org/abs/1808.04572 (2018, accessed 9 May 2021).
26. Hoyle RH. *Statistical strategies for small sample research*. Thousand Oaks, CA: SAGE, 1999, p.394.
27. Hand DJ, Daly F, McConway K, et al. *A handbook of small data sets*. Boca Raton, FL: CRC Press, 1993, p.476.
28. Thomas RM, Bruin W, Zhutovsky P, et al. Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders. In: *Machine learning: methods and applications to brain disorders*.Cambridge, MA: Academic Press, 2019. pp.249–266.
29. Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size. *PLoS One* 2019; 14(11): e0224365.
30. Li X, Deng S, Wang S, et al. Review of small data learning methods. In: *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, Tokyo, Japan, 23–27 July 2018, pp.106–109. New York: IEEE.
31. Hekler EB, Klasnja P, Chevance G, et al. Why we need a small data paradigm. *BMC Med* 2019; 17(1): 133.
32. Bornschein J, Visin F and Osindero S. Small data, big decisions: model selection in the small-data regime. In: *International conference on machine learning*, London, UK, 26 December 2020, pp.1035–1044. PMLR. London, UK: MLR Press. http://proceedings.mlr.press/v119/bornschein20a.html (accessed 9 May 2021).
33. van Eck NJ and Waltman L. Visualizing bibliometric networks. In: Ding Y, Rousseau R and Wolfram D (eds) *Measuring scholarly impact: methods and practice*. Cham: Springer International Publishing, 2014, pp.285–320.
34. Železnik D, Blažun Vošner H, et al. A bibliometric analysis of the Journal of Advanced Nursing, 1976–2015. *J Adv Nurs* 2017; 73(10): 2407–2419.
35. van Eck NJ and Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010; 84(2): 523–538.
36. Forsström JJ, Irjala K, Selén G, et al. Using data preprocessing and single layer perceptron to analyze laboratory data. *Scand J Clin Lab Invest* 1995; 55(S222): 75–81.
37. Reich Y and Travitzky N. Machine learning of material behaviour knowledge from empirical data. *Mater Des* 1995; 16(5): 251–259.
38. Shneider AM. Four stages of a scientific discipline; four types of scientist. *Trends Biochem Sci* 2009; 34(5): 217–223.
39. Pestana MH, Sánchez AV and Moutinho L. The network science approach in determining the intellectual structure, emerging trends and future research opportunities – an application to senior tourism research. *Tour Manage Perspect* 2019; 31: 370–382.

40. Sahoo J and Pati B. Highly cited articles in knowledge management: a study of the content and authorship trend. *Int J Inf Dissem Technol* 2018; 8(4): 196–200.
41. Zhang Y and Ling C. A strategy to apply machine learning to small datasets in materials science. *NPJ Comput Mater* 2018; 4(1): 1–8.
42. Babič M, Kokol P, Belič I, et al. Using of genetic programming in engineering. *Electrotechn Rev* 2014; 81(3): 143–147.
43. Feng S, Zhou H and Dong H. Using deep neural network with small dataset to predict material defects. *Mater Des* 2019; 162: 300–310.
44. Ko S, Choi J and Ahn J. GVES: machine learning model for identification of prognostic genes with a small dataset. *Sci Rep* 2021; 11(1): 439.
45. Spiga O, Cicaloni V, Fiorini C, et al. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet J Rare Dis* 2020; 15(1): 1–10.
46. Abu Zohair LM. Prediction of Student's performance by modelling small dataset size. *Int J Educ Technol High Educ* 2019; 16(1): 27.
47. Fong SJ, Li G, Dey N, et al. Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. *IJIMAI* 2020; 6(1): 132.
48. Komiske PT, Metodiev EM, Nachman B, et al. Learning to classify from impure samples with high-dimensional data. *Phys Rev D* 2018; 98(1): 011502.
49. Zhang Y, Zhu R, Chen Z, et al. Evaluating and selecting features via information theoretic lower bounds of feature inner correlations for high-dimensional data. *Eur J Oper Res* 2021; 290(1): 235–247.
50. Dai D, Xu T, Wei X, et al. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Comput Mater Sci* 2020; 175: 109618.
51. He Z, Shao H, Cheng J, et al. Support tensor machine with dynamic penalty factors and its application to the fault diagnosis of rotating machinery with unbalanced data. *Mech Syst Signal Process* 2020; 141: 106441.
52. Liang XW, Jiang AP, Li T, et al. LR-SMOTE — an improved unbalanced data set oversampling based on K-means and SVM. *Knowl Based Syst* 2020; 196: 105845.
53. Marceau L, Qiu L, Vandewiele N, et al. A comparison of deep learning performances with other machine learning algorithms on credit scoring unbalanced data. *arXiv:190712363* [cs, stat], http://arxiv.org/abs/1907.12363 (2020, accessed 18 February 2021).
54. Alvarsson J, Lampa S, Schaal W, et al. Large-scale ligand-based predictive modelling using support vector machines. *J Cheminform* 2016; 8(1): 39.
55. Cao J, Fang Z, Qu G, et al. An accurate traffic classification model based on support vector machines. *Int J Netw Manage* 2017; 27(1): e1962.
56. Razzak I, Zafar K, Imran M, et al. Randomized nonlinear one-class support vector machines with bounded loss function to detect of outliers for large scale IoT data. *Future Gener Comput Syst* 2020; 112: 715–723.
57. Zhang Z, Song Y, Cui H, et al. Topological analysis and gaussian decision tree: effective representation and classification of biosignals of small sample size. *IEEE Trans Biomed Eng* 2017; 64(9): 2288–2299.
58. Shaikhina T, Lowe D, Daga S, et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control* 2019; 52: 456–462.

59. Liu S and Deng W. Very deep convolutional neural network based image classification using small training sample size. In: *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, 3–6 November 2016, pp.730–734. New York: IEEE.
60. Yamashita R, Nishio M, Do RKG, et al. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018; 9(4): 611–629.
61. Guo F, Ng M, Goubran M, et al. Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: a continuous kernel cut approach. *Med Image Anal* 2020; 61: 101636.
62. Hall LO, Paul R, Goldgof DB, et al. Finding Covid-19 from chest X-rays using deep learning on a small dataset. arXiv:200402060 [cs, eess], http://arxiv.org/abs/2004.02060 (2020, accessed 2021 February 18).
63. Tits N, El Haddad K and Dutoit T. Exploring transfer learning for low resource emotional TTS. In: Bi Y, Bhatia R and Kapoor S (eds) *Intelligent systems and applications*. Cham: Springer International Publishing, 2020, pp.52–60. (Advances in Intelligent Systems and Computing).
64. Jalali A, Mallipeddi R and Lee M. Sensitive deep convolutional neural network for face recognition at large standoffs with small dataset. *Expert Syst Appl* 2017; 87: 304–315.
65. Athanasopoulou L, Papacharalampopoulos A and Stavropoulos P. Context awareness system in the use phase of a smart mobility platform: a vision system for a light-weight approach. *Procedia CIRP* 2020; 88: 560–564.
66. Li W, Zhao Y, Chen X, et al. Detecting Alzheimer's disease on small dataset: a knowledge transfer perspective. *IEEE J Biomed Health Inf* 2019; 23(3): 1234–1242.
67. Silitonga P, Bustamam A, Muradi H, et al. Comparison of dengue predictive models developed using artificial neural network and discriminant analysis with small dataset. *Appl Sci* 2021; 11(3): 943.
68. Han D, Liu Q and Fan W. A new image classification method using CNN transfer learning and web data augmentation. *Expert Syst Appl* 2018; 95: 43–56.
69. Hagos MT and Kant S. Transfer learning based detection of diabetic retinopathy from small dataset. arXiv:190507203 [cs], http://arxiv.org/abs/1905.07203 (2019, accessed 19 February 2021).
70. Gong H-F, Chen Z-S, Zhu Q-X, et al. A Monte Carlo and PSO based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: an empirical study of petrochemical industries. *Appl Energy* 2017; 197: 405–415.
71. MacAllister A, Kohl A and Winer E. Using high-fidelity meta-models to improve performance of small dataset trained Bayesian Networks. *Expert Syst Appl* 2020; 139: 112830.
72. Zhu Q-X, Chen Z-S, Zhang X-H, et al. Dealing with small sample size problems in process industry using virtual sample generation: a Kriging-based approach. *Soft Comput* 2020; 24(9): 6889–6902.
73. Kumar I, Dogra K, Utreja C, et al. A comparative study of supervised machine learning algorithms for stock market trend prediction. In: *2018 second international conference on inventive communication and computational technologies (ICICCT)*, Coimbatore, India, 20–21 April 2018, pp.1003–1007. New York: IEEE.
74. Pei Z, Jiang H, Li X, et al. Data augmentation for rolling bearing fault diagnosis using an enhanced few-shot Wasserstein auto-encoder with meta-learning. *Meas Sci Technol* 32(8): 084007.

75. Giansanti V, Castelli M, Beretta S, et al. Comparing deep and machine learning approaches in bioinformatics: a miRNA-target prediction case study. In: Rodrigues JMF, Cardoso PJS, Monteiro J, et al (eds) *Computational science – ICCS 2019*. Cham: Springer International Publishing, 2019, pp.31–44. (Lecture Notes in Computer Science).
76. Khatun MstS, Shoombuatong W, Hasan MdM and Kurata H. Evolution of sequence-based bioinformatics tools for protein-protein interaction prediction. *Curr Genomics* 2020; 21(6): 454–463.
77. Yadav SS and Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data* 2019; 6(1): 113.
78. Saufi SR, Ahmad ZAB, Leong MS, et al. Gearbox fault diagnosis using a deep learning model with limited data sample. *IEEE Trans Ind Inf* 2020; 16(10): 6263–6271.
79. Wen L, Li X and Gao L. A transfer convolutional neural network for fault diagnosis based on ResNet-50. *Neural Comput Appl* 2020; 32(10): 6111–6124.
80. Croda RMC, Romero DEG and Morales SOC. Sales prediction through neural networks for a small dataset. *IJIMAI* 2019; 5(4): 35–41.
81. Khan JY, Khondaker MTI, Iqbal A, et al. A benchmark study on machine learning methods for fake news detection. arXiv:190504749 [cs, stat], http://arxiv.org/abs/1905.04749 (2019, accessed 19 February 2021).
82. Renault T. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digit Finance* 2020; 2(1): 1–13.
83. Rajpurkar P, Park A, Irvin J, et al. AppendiXNet: deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci Rep* 2020; 10(1): 3958.
84. Subirana B, Hueto F, Rajasekaran P, et al. Hi Sigma, do I have the Coronavirus? Call for a new artificial intelligence approach to support health care professionals dealing with the COVID-19 pandemic. *arXiv:200406510* [cs]. http://arxiv.org/abs/2004.06510 (2020 accessed 2021 February 19).
85. Rodríguez-Rodríguez I, Rodríguez J-V, Chatzigiannakis I, et al. On the possibility of predicting glycaemia 'on the fly' with constrained IoT devices in type 1 diabetes mellitus patients. *Sensors (Basel)* 2019; 19(20):4538.
86. Faraway JJ and Augustin NH. When small data beats big data. *Stat Probab Lett* 2018; 136: 142–145.

Author biographies

Peter Kokol joined the Department of Computer Science in University of Maribor as an assistant researcher in 1982 where he is now a Full Professor and the head of Laboratory of System Design. From 1997 to 2013 he was the head of Research Institute at the Faculty of Health Sciences, University of Maribor, from February 2001 the dean for research and from October 2007 till October 2012 the dean. Since January 2002 to 2008 he was the Director of the independent Centre for Interdisciplinary and Multidisciplinary Studies and Research. He has written over 700 technical and research papers published in recognised international journals and major conferences and co-authored some textbooks. He was the general or program chair of major conferences, had numerous invited presentations and won several best papers awards. He has more than 3750 citations in Google Scholar, more than 1500 in SCOPUS and more than 1100 in WOS. His main research interests are data mining, machine

learning, intelligent systems. software engineering, Information systems design, knowledge synthesis, bibliometrics and health informatics.. He has acted as coordinator and principal investigator in more than 40 international and national research projects. He is a member of ACM, IEEE and ASIS and some IMIA technical committees. He was the President of the IEEE Committee on Computational Medicine.

Marko Kokol is a PhD student at the Faculty of Electrical Engineering and Computer Science at the University of Maribor and has more than 15 years of experience in the software & database development industry ranging in a wide variety of fields spanning from database design, security software architecture and intelligent and decision systems. He holds a bachelor's degree in computer science from the University of Maribor has a number of Professional Certifications.

Sašo Zagoranski is a PhD student at the Faculty of Electrical Engineering and Computer Science at the University of Maribor and has more than 15 years of experience in the software & database development industry ranging in a wide variety of fields spanning from database design, security software architecture to natural user interfaces using gestures and touch. He holds the bachelor's degree in computer science from the University of Ljubljana.