Guest Editorial

Data Science in Smart Healthcare: challenges and opportunities

I. INTRODUCTION

A shift toward a data-driven socio-economic health model is occurring. This is the result of the increased volume, velocity and variety of data collected from the public and private sector in healthcare, and biology in general. In the past five-years, there has been an impressive development of computational intelligence and informatics methods for application to health and biomedical science. However, the effective use of data to address the scale and scope of human health problems has yet to realize its full potential. The barriers limiting the impact of practical application of standard data mining and machine learning methods have been inherent to the characteristics of health data. Besides the volume of the data ('big data'), these are challenging due to their heterogeneity, complexity, variability and dynamic nature. Finally, data management and interpretability of the results have been limited by practical challenges in implementing new and also existing standards across the different health providers and research institutions.

The scope of this Special issue is to discuss some of these challenges and opportunities in health and biological data science, with particular focus on the infrastructure, software, methods and algorithms needed to analyze large datasets in biological and clinical research. After a rigorous review process, 15 articles were selected for publication in this special issue. They are briefly discussed in the following.

II. A BRIEF OVERVIEW OF THE PAPERS IN THIS SPECIAL ISSUE

The first thematic area is the use of deep learning to solve predictive tasks in healthcare and biology. Six papers address this. From the diagnosis of Autism spectrum disorder [2] and mental disorder [6, 7], to the use of autoencoders to cluster similar event logs in latent space [8]; from the automatic assignment of ICD-O3 topography and morphology codes to free-text cancer reports [13] to the red blood cell segmentation and classification from microscopic images [14]. Interestingly, in [14] the authors apply deformable convolution layers to enable freeform deformation of the feature learning process, thus making the network more robust to various cell morphologies and image settings.

Next, the issue of the interpretability of the results obtained in the analysis of health and biological data is addressed. Paper [8] proposes to explain the clusters labels by decoding the corresponding events; paper [13] compares alternative architectures in terms of prediction accuracy and interpretability. In this analysis, an element-wise maximum aggregator performs slightly better than attentive models, offering a way to interpret the classification process. Paper [3] makes use of Extreme Gradient Boosting, Artificial Neural Network Models and Symbolic Regression (SR) to diagnose Parkinson's Disease by monitoring the gait of the patients. The Extreme Gradient Boosting, Artificial Neural Network models were found to outperform Symbolic Regression, although the latter gives more easily readable and interpretable results.

In Paper [1], the authors opt for the use of an expert system based on medical domain knowledge, rather than for machine learning. They make use of a belief rule based method with evidential reasoning to provide a white-box model to analyze radiologist behavior in making diagnosis.

A number of papers address definition and/or extraction of new interesting features. For example, paper [2] introduces the generation of single-volume brain images from the whole-brain image; and paper [4] develops a model for automatic detection of new-onset atrial fibrillation during sepsis from a number of meta-features extracted from electrocardiogram signals. Finally, paper [7] applies self-supervised representation to the training of a deep neural network to recognize distinct cognitive activities in healthy individuals. In this way, the model learns how to encode high-level semantic information, used for discriminating between control subjects and patients with dementia. In paper [5], the authors apply two methods for network reconstruction along with a number of clustering techniques to discover features associated with specific phenotypes.

As papers [2, 3, 4, 6], paper [9] addresses early diagnosis / risk prediction or, more precisely, decision support for patients with Chiari I Malformation, proposing a fully automated method to select the optimal intervention.

Two other applications discussed are those dedicated to the prediction of drug sensitivity [11] and drug repositioning [12]. In [11] the authors predict the sensitivity of cell lines to anti-cancer drugs using different classification algorithms. Algorithmsare then ranked using a reinforcement learning approach. In [12], the authors address the problem of drug repositioning using Non-negative Matrix Tri-Factorization, a method that exploits both data integration and machine learning, to infer novel indications for approved drugs. The authors integrate different heterogeneous data types about drugs and proteins (possible drug targets) by modeling different entities and their relationships as a multi-partite graph. Then they propose a shortest-paths-based method to infer relationships between elements of different type, as well as

between them and other nodes that were not originally linked in the graph.

Paper [10] uses stochastic Markov-chain based methods to model and simulate the kinetics of fluorescence loss that is due to stochastic events of cell division. Fuzzy Self-Tuning Particle Swarm Optimization is used for automatic parameters setting.

Finally, paper [15] presents the state of the art for healthcare data ingestion services on the cloud. Importantly, it addresses the need to facilitate data storage and large-scale analysis and the urgency of considering the issues such as security, availability and disaster recovery and challenges posed by the lack of data standards together with their heterogeneous and sensitive nature.

III. CONCLUDING REMARKS

Overall, this special issue contains several references to the use of artificial intelligence, machine learning and modeling in the health sector. There are examples of integration of heterogeneous and very numerous and complex data and the theme of interpretability is always present.

In this regard, the choice towards the use of traditional machine learning and artificial intelligence approaches rather than deep learning, seems to be driven by the number of available examples and the need of model transparency. However, there are also interesting examples of seeking for transparency with deep learning methods.

It is also evident that the definition of features and meta-features of clinical interest on the one hand increases the interpretability of the model, and on the other hand, seems to facilitate the learning process. Therefore, there is an urgent need to consider technological solutions, such as the cloud, capable of coping with the large amount of clinical data currently produced, meanwhile considering the representation of these data.

We hope that the clinical data modeling and prediction works presented in this special issue can inspire and offer application examples, and potential solutions, to scientists working in the expanding field of data science for healthcare.

ACKNOWLEDGEMENT

The guest editors thank all those who helped make this special issue possible, especially the Editor-in-Chief, JBHI editorial office, and the authors and reviewers of the contributions.

References

[1] L. Chang, C. Fu, Z. Wu, W. Liu, S. Yang, "Data-driven analysis of radiologists' behavior for diagnosing thyroid nodules," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2969322.

[2] M. R. Ahmed, Y. Zhang, Y. Liu, H. Liao, "Single Volume Image Generator and Deep Learning-based ASD Classification," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2998603.

[3] J. Hughes, S. Houghten, J. A. Brown, "Models of Parkinson's Disease Patient Gait," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2019.2961808.

[4] S. Bashar, M. B. Hossain, E. Ding, A. Walkey, D. McManus, K. Chon, "Atrial Fibrillation Detection during Sepsis: Study on MIMIC III ICU Data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2995139.

[5] J. Kramer, L. Boone, T. Clifford, J. Bruce, J. D. Matta, "Analysis of Medical Data Using Community Detection on Inferred Networks," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3003827.

[6] J. Kacur, J. Polec, E. Smolejova, A. Heretik, "An Analysis of Eye-Tracking Features and Modelling Methods for Free-Viewed Standard Stimulus: Application for Schizophrenia Detection," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3002097.

[7] K. Mengoudi, D. Ravi', K. Yong, S. Primativo, I. Pavisic, E. Brotherhood, K. Lu, J. M. Schott, S. J. Crutch, D. C. Alexander, "Augmenting Dementia Cognitive Assessment with Instruction-less Eye-tracking Tests," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3004686.

[8] H. DeOliveira, V. Augusto, B. Jouaneton, L. Lamarsalle, M. Prodel, X. Xie, "Automatic and Explainable Labeling of Medical Event Logs with Autoencoding," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3021790.

[9] L. Mesin, F. Mokabberi, C. F. Carlino, "Automated morphological measurements of brain structures and identification of optimal surgical intervention for Chiari I malformation," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3016886.

[10] M. Nobile, E. Nisoli, T. Vlachou, S. Spolaor, P. Cazzaniga, G. Mauri, P. G. Pelicci, D. Besozzi, "cuProCell: GPU-accelerated analysis of cell proliferation with flow cytometry data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3005423.

[11] S.Daoud, A. Mdhaffar, M. Jmaiel, B. Freisleben, "Q-Rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3004663.

[12] G. Ceddia, P. Pinoli, S. Ceri, M. Masseroli, "Matrix Factorization-based Technique for Drug Repurposing Predictions," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2991763.

[13] S. Martina, L. Ventura, P. Frasconi, "Classification of cancer pathology reports: a large-scale comparative study," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3005016.

[14] M. Zhang, X. Li, M. Xu, Q. Li, "Automated Semantic Segmentation of Red Blood Cells for Sickle Cell Disease," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3000484.

[15] R. Ranchal, P. Bastide, X. Wang, A. Gkoulalas-Divanis, M. Mehra, S. Bakthavachalam, H. Lei, A. Mohindra, "Disrupting Healthcare Silos: Addressing Data Volume, Velocity and Variety with a Cloud-Native Healthcare Data Ingestion Service," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3001518.

BARBARA DI CAMILLO
Department of Information Engineering
University of Padova, Italy
(barbara.dicamillo@unipd.it)

GIUSEPPE NICOSIA
Department of Biochemistry
Cambridge Systems Biology Centre
University of Cambridge, United Kingdom
(gn263@cam.ac.uk)
Department of Biomedical & Biotechnological Sciences
School of Medicine - University of Catania, Italy
(giuseppe.nicosia@unict.it)

FRANCESCA BUFFA
Computational Biology and Integrative Genomics Lab
Department of Oncology, Medical Sciences Division
University of Oxford, United Kingdom
(francesca.buffa@oncology.ox.ac.uk)

BENNY LO
Department of Surgery and Cancer/ The Hamlyn Centre
Imperial College London, United Kingdom
benny.lo@imperial.ac.uk