Mathematical Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications

Edited by
Monica Bianchini and Maria Lucia Sampoli

Printed Edition of the Special Issue Published in *Mathematics*

www.mdpi.com/journal/mathematics

MDPI

Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications

Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications

Editors

Monica Bianchini
Maria Lucia Sampoli

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors
Monica Bianchini
University of Siena
Italy

Maria Lucia Sampoli
University of Siena
Italy

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Mathematics* (ISSN 2227-7390) (available at: https://www.mdpi.com/journal/mathematics/special_issues/Math_Model_Machine_Learning_Bioinformatics_Data_Science).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-2840-3 (Hbk)
ISBN 978-3-0365-2841-0 (PDF)

Contents

About the Editors

Monica Bianchini , Ph.D., is currently an Associate Professor at the Department of Information Engineering and Mathematics of the University of Siena (Full Professor Abilitation). She received the Laurea cum laude in Mathematics and a Ph.D. degree in Computer Science from the University of Florence, Italy, in 1989 and 1995, respectively. After receiving the Laurea, for two years, she was involved in a joint project of Bull HN Italia and the Department of Mathematics (University of Florence), aimed at designing parallel software for solving differential equations. From 1992 to 1998, she was a Ph.D. student and a Postdoc Fellow with the Computer Science Department of the University of Florence. Since 1999, she has been with the University of Siena. Her main research interests are in the field of machine learning, with emphasis on neural networks for structured data and deep learning, approximation theory, information retrieval, bioinformatics, and image processing. Monica Bianchini has authored more than one hundred papers and has been the editor of books and Special Issues in international journals in her research field. She has been a participant in many research projects focused on machine learning and pattern recognition, founded by both Italian Ministry of Education (MIUR) and University of Siena (PAR scheme), and she has been involved in the organization of several scientific events, including the NATO Advanced Workshop on Limitations and Future Trends in Neural Computation (2001), the 8th AI*IA Conference (2002), GIRPR 2012, the 25th International Symposium on Logic-Based Program Synthesis and Transformation, and the ACM International Conference on Computing Frontiers 2017. Prof. Bianchini served as an Associate Editor for *IEEE Transactions on Neural Networks* (2003–09), *Neurocomputing* (from 2002), *Int. J. of Computers in Healthcare* (from 2010), and *Frontiers in Genomics* (section Computational Genomics). She is a permanent member of the Editorial Board of *IJCNN*, *ICANN*, *ICPRAM*, *ESANN*, *ANNPR*, and *KES*.

Maria Lucia Sampoli, Ph.D., is currently Associate Professor in Numerical Analysis at the Department of Information Engineering and Mathematics of the University of Siena. She graduated in Mathematics with Honors from the University of Florence in 1994, and in 1998 received a PhD in Computational Mathematics and Operative Research from the University of Milan. In the same year she received a postdoc fellowship at the Institute of Applied Geometry of the Technical University of Darmstadt (Germany). In 1999–2000 she was a CNR Senior Fellow at Department of Mathematics of the University of Siena, where she initially was a Research Assistant and later in 2002 became a tenured Assistant Professor in Numerical Analysis. In 2010 she joined the Department of Information Engineering and Mathematics. Her research interests are mainly focused on the use of splines functions in various applications, from shape preserving interpolation and approximation to numerical approximation of elliptic problems (Isogeometic Analysis), quasi-interpolations techniques, Numerical quadrature, and Phythagorean curves and surfaces. Maria Lucia Sampoli has been invited speaker (as a plenary or as a speaker at mini-symposia) at several international conferences as well as involved in the organization of various conferences and workshops. She has been involved in many research projects including as a coordinator for some of them.

Preface to "Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications"

With the enormous amount of data flowing from a variety of real-world problems, Artificial Intelligence (AI), and in particular Machine Learning (ML) and Deep Learning (DL) techniques, have powered new achievements in many complex applications, that were prohibitive with deterministic approaches. These advances, which are based on a multidisciplinary research framework involving computer science, numerical analysis and statistics, come from research efforts in both industry and academia and are particularly well suited to address complex problems in data science and, more specifically, in biotechnological and medical applications. While these methods have proven to be astounding in performance, they still suffer from a sort of opacity, meaning that their produced results, though correct and quickly obtained, are difficult to be interpreted or explained, a fundamental drawback especially for those problems where people's lives are at stake. Therefore, a deeper understanding of the fundamental principles of machine and deep learning methods is mandatory in order to evidence both their advantages and limitations. Though a variety of different approaches exists to face the problem of explainability—ranging from methods that try to understand on which features an AI model bases its prediction, to the construction of ad hoc architectures which some logical knowledge can be extracted from—we think that a viable alternative is that of a *contamination* between mathematical modeling and machine learning, in the belief that the insertion of the equations deriving from the physical world in the data-driven models can greatly enrich the information content of the sampled data, allowing us to simulate very complex phenomena, with drastically reduced calculation times and interpretable solutions. The application of such hybrid techniques to structured data, such as time series or graphs, however, opens to other interesting challenges, aimed at determining whether these techniques are able to generalize to different problems, how much the data structure (for instance, sampling from a subspace or manifold) affects the method, and how to choose appropriate (hyper-)parameters to ensure a good fit, while still avoiding overfitting.

This Special Issue brings together researchers from different disciplinary fields, who focus on building theoretical foundations and presenting cutting-edge applications for deep learning applications. We hope this issue will serve as a hint for researchers to reflect on fundamental issues in a wide variety of fields, including pure and applied mathematics, statistics, computer science and engineering, to join forces to integrate different approaches suitable for solving complex problems, quickly, reliably and understandably for human experts.

The contributions collected can be divided into two main categories, relating, respectively, to the application of mathematical models/DL techniques to the study of biological macrosystems and to the automatic analysis/prediction of medical data for the prognosis of human diseases.

For the first category, the paper titled "Machine Learning Techniques Applied to Predict Tropospheric Ozone in a Semi-Arid Climate Region" describes a comparative evaluation of a large class of statistical modeling methods for classifying high or low ozone concentration levels. Indeed, ground-level ozone exposure has led to a significant increase in environmental risks, since it adversely affects not only human health but also some delicate plants and vegetation.

Additionally, in the paper "Interactions Obtained from Basic Mechanistic Principles: Prey Herds and Predators", four different predator–prey–herd models are presented, that are derived assuming that the prey gathers in herds, that the predator can be specialist—i.e., it feeds on only one species—or generalist—i.e., it feeds on multiple resources—and considering two functional responses, the herd-linear and herd-Holling type II functional responses. The paper aims at deriving their mathematical formulation from the individual-level state transitions, and compare the models' dynamics in terms of equilibria, stability and bifurcation diagrams. The predator–prey–herd antagonistic behavior has been widely observed in population ecology, especially in aquatic species and insects, and has been proven to deeply affect niche expansion and speciation.

To the second category belongs the manuscript "Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques", which presents a nonlinear approach for identifying combinations of CpG DNA methylation data as biomarkers for Alzheimer disease (AD). Indeed, the possibility of having techniques that can determine earlier if an individual has AD is becoming increasingly important, especially after the FDA approval of the first drug for AD treatment (there were drugs before it targeting some of the effects of the illness, but not the actual illness itself). Such an early diagnosis will be possible soon, thanks to non-invasive medical tests to capture methylation data, simply based on blood.

Two contributions, namely "Visual Sequential Search Test Analysis: An Algorithmic Approach" and "A Mixed Statistical and Machine Learning Approach for the Analysis of Multimodal Trail Making Test Data", are devoted to the automatic analysis of Trail Making Test (TMT) data. TMT is a popular neuropsychological test, commonly used in clinical settings as a diagnostic tool for the evaluation of some frontal functions, that provides qualitative information on high order mental activities, including speed of processing, mental flexibility, visual spatial orientation, working memory and executive functions. Such data are preprocessed in the form of sequences and treated with an algorithmic approach based on the episode matching method, or in the form of scan-path images, that can be processed via DL and clustering methods, for distinguishing patients affected by the extrapyramidal syndrome and by chronic pain from healthy subjects. A statistical analysis, based on the blinking rate and on the pupil size, is also carried out, to help classifying different pathologies.

Finally, the paper "A Multi-Stage GAN for Multi-Organ Chest X-ray Image Generation and Segmentation" proposes a deep learning approach to the generation of realistic synthetic images—particularly useful in medical applications where the scarcity of data often prevents the use of DL architectures—that can be employed to train a segmentation network. Segmentation is, in fact, the preventive step for automatic image analysis and classification, and has proven fundamental, for instance, in order to diagnose COVID-19 based on lung damage.

Monica Bianchini, Maria Lucia Sampoli
Editors