BITS :: Call for Abstracts 2021 - Poster

Type	Poster
Session	Gene regulation, transcriptomics and epigenomics
Title	Gene expression analysis of pediatric Multiple Sclerosis using Machine Learning
All Authors	Casalino G(1), Castellano G(1), Consiglio A(2), Nuzziello N(2), Vessio G(1).

Affiliation

(1) Department of Computer Science, University of Bari Aldo Moro, 70125 Bari, Italy
(2) Institute for Biomedical Technologies of Bari, CNR, Bari

Motivation

Multiple Sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system that usually affects young adults [1]. However, onset in childhood and adolescence is increasingly recognized by researchers, accounting for 3-5% of cases of MS [2]. Improvements in diagnostic tools and/or greater sensitivity to early signs, in fact, have contributed to a better recognition of MS in the very early ages; while, in the past, it had mostly been monitored retrospectively. Genetic predisposition, environmental factors, and lifestyle appear to contribute significantly to the overall risk of developing MS [3]; however, very few studies have investigated the “environmentally naïve” genetic load of pediatric MS (PedMS). Studying the transcriptomic involvement of PedMS can help elucidate the pathogenic mechanisms underlying MS in its early stages, which are not fully understood yet. The bioinformatic pipeline usually developed for differential gene expression analysis applies traditional statistical tests to search for genes that are differentially expressed between healthy controls and diseased patients. This analysis allows biologists to isolate evident changes in expression; however, it may fail to find more complex, nonlinear interactions among disease-related genes.

Methods

Machine learning techniques are increasingly used in the biological and health domain, mainly due to their ability to find complex relationships and nonlinear interactions between input and output data (e.g., [4, 5, 6]). In our research, we studied the predictive potential of gene expression features for the development of a computerized predictive model to distinguish the pathological sample from the healthy control group. Moreover, since we also collected data from a small sample of patients with Attention Deficit Hyperactivity Disorder (ADHD), and it was observed that these patients share some cognitive impairments with PedMS subjects, we also included this cohort, resulting in a multi-class classification problem. Deriving a predictive model that can also discriminate between the two diseases can be of great help for experts in this sector. The experiments were conducted on RNA-Seq data produced from whole blood samples of a mixed cohort of 47 subjects: 20 healthy controls (from now on HCs); 19 patients with PedMS; and 8 children with ADHD. It is worth noting that, in order to improve the generalization capacity of the models, data were subjected to a filtering, normalization and feature selection process to obtain a more informative and easy-to-compute representation. To get reliable prediction performance estimates, normalization and feature selection were “nested” within the validation scheme, which was a 5-fold cross-validation.

Results

We compared the results provided by state-of-the-art classification algorithms: Logistic Regression (LR); Decision Tree (DT); Random Forest (RF); Support Vector Machine (SVM); and Multi-layer Perceptron (MLP). While the more complex methods, particularly RF, SVM and MLP, gave the worst results, mainly because they overfit the small sample of data, the best accuracy, 0.83, was achieved by LR. Feeding classification algorithms with gene expressions is an effective strategy for discriminating PedMS patients from the HC individuals. Instead, an acceptable classification of ADHD patients was never achieved. This may be due to the very small size and imbalance of this experimental group. However, it should be noted that LR incorrectly classified some ADHD as PedMS (and not HCs) indicating that the model may have been misled by the possible overlap of pathological patterns between the two conditions.

Info

All authors contributed equally to this work.

References

1. Lassmann, H., Bradl, M.: Multiple sclerosis: experimental models and reality. Acta neuropathologica 133(2), 223–244 (2017)
2. Yeh, E.A., Chitnis, T., Krupp, L., Ness, J., Chabas, D., Kuntz, N., Waubant, E., of Pediatric Multiple Sclerosis Centers of Excellence, U.N., et al.: Pediatric multiple sclerosis. Nature Reviews Neurology 5(11),

621 (2009)
3. Waubant, E., Lucas, R., Mowry, E., Graves, J., Olsson, T., Alfredsson, L., Langer-Gould, A.: Environmental and genetic risk factors for MS: an integrated review. Annals of clinical and translational neurology 6(9),1905–1922 (2019)
4. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics 3(2), 119–131 (2016)
5. Chicco, D.: Ten quick tips for machine learning in computational biology. Bio Data mining 10(1), 35 (2017)
6. Oriol, J.D.V., Vallejo, E.E., Estrada, K., Pena, J.G.T.: Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. BMC bioinformatics 20(1), 1–17(2019)

Figure	
-	
Availability	-

Corresponding Author

Name, Surname	Arianna, Consiglio
Email	arianna.consiglio@ba.itb.cnr.it
Submitted on	30.04.2021