Journal of Ambient Intelligence and Humanized Computing (2023) 14:15851–15860
https://doi.org/10.1007/s12652-021-03091-2

ORIGINAL RESEARCH

MicroRNA expression classification for pediatric multiple sclerosis identification

Gabriella Casalino[1] · Giovanna Castellano[1] · Arianna Consiglio[2] · Nicoletta Nuzziello[2] · Gennaro Vessio[1]

Received: 17 February 2020 / Accepted: 3 March 2021 / Published online: 19 March 2021

Abstract
MicroRNAs (miRNAs) are a set of short non-coding RNAs that play significant regulatory roles in cells. The study of miRNA data produced by Next-Generation Sequencing techniques can be of valid help for the analysis of multifactorial diseases, such as Multiple Sclerosis (MS). Although extensive studies have been conducted on young adults affected by MS, very little work has been done to investigate the pathogenic mechanisms in pediatric patients, and none from a machine learning perspective. In this work, we report the experimental results of a classification study aimed at evaluating the effectiveness of machine learning methods in automatically distinguishing pediatric MS from healthy children, based on their miRNA expression profiles. Additionally, since Attention Deficit Hyperactivity Disorder (ADHD) shares some cognitive impairments with pediatric MS, we also included patients affected by ADHD in our study. Encouraging results were obtained with an artificial neural network model based on a set of features automatically selected by feature selection algorithms. The results obtained show that models developed on automatically selected features overcome models based on a set of features selected by human experts. Developing an automatic predictive model can support clinicians in early MS diagnosis and provide new insights that can help find novel molecular pathways involved in MS disease.

Keywords MicroRNA expressions · Next-Generation Sequencing · Pediatric Multiple Sclerosis · Bioinformatics · Digital Health · Feature selection · Artificial neural networks · Classification

1 Introduction

Transcriptomics is one of the most important fields of study in Molecular Biology and Bioinformatics. In every living cell, at every moment, the information stored in the DNA is copied into RNA transcripts that are used for all the essential functions of a cell, such as the production of proteins. The copied portions of DNA are called genes, and the frequencies of the gene copies produced in the cell under a given condition are called gene expressions. Next-Generation Sequencing (NGS) techniques allow biologists to decode and quantify the entire gene expression profile of a sample, i.e. the entire set of RNAs present in the sample at a specific time. NGS revolutionised Transcriptomics, as it allowed biologists to discover new genes and RNA transcripts. Before NGS, the gene expression profile was estimated with micorarrays, which are sets of pre-defined probes that produce signals when they recognize known genes. However, these are unable to detect new genes or transcripts containing unknown mutations. On the contrary, NGS allows the study of all of existing RNAs produced in a cell and has promoted the study of novel classes of RNA that play pivotal roles in the cell, such as microRNAs (miRNAs).

MiRNAs are a class of small RNAs that regulate the expression of other longer RNAs and the consequent production of proteins (Bartel 2004). In recent years, research on miRNA-related problems has become a hot field in Bioinformatics mainly because of the miRNA essential biological functions (Huang et al. 2011). The study of miRNA expression changes, in fact, offers the opportunity to identify biomarkers, i.e. molecules predicting the clinical course or

✉ Giovanna Castellano
giovanna.castellano@uniba.it

Gabriella Casalino
gabriella.casalino@uniba.it

Gennaro Vessio
gennaro.vessio@uniba.it

[1] Department of Computer Science, University of Bari “Aldo Moro”, Bari, Italy

[2] Institute for Biomedical Technologies of Bari, Italian National Research Council, Bari, Italy

response to treatments, which can be useful for the (possibily early) diagnosis of complex and multi-factorial diseases, such as Multiple Sclerosis (MS).

Multiple Sclerosis is a demyelinating autoimmune disease of the central nervous system that usually affects young adults (Olsson et al. 2017). The onset during childhood and adolescence is increasingly recognized (Chitnis et al. 2009), along with the demonstration of cognitive deficits in more than one third of these patients (Akbar et al. 2016). Therefore, the study of pediatric MS (PedMS) patients offers a unique opportunity to investigate the pathogenic mechanisms that occur in the early stages of the disease. To this end, the analysis of miRNA expressions can be of great help. Unfortunately, although miRNAs investigations have so far been performed in young adults, the pathogenic mechanisms underlying PedMS are still not fully understood. To this aim, in previous studies (Liguori et al. 2017, 2019), we investigated the transcriptome profile of peripheral blood samples in a cohort of PedMS patients and further validated (with specific laboratory assays) miRNAs with statistically significant increased or decreased expression in PedMS patients compared to healthy pediatric control (HC) subjects.

Bionformatic pipelines developed for miRNA expression analysis usually apply classical statistical tests to look for miRNAs that are differentially expressed between healthy controls and diseased patients (Love et al. 2014; McCarthy et al. 2012). This analysis allows us to isolate noticeable changes in expression; however, it fails to extract more complex interactions among different disease-related miRNAs. Artificial Intelligence techniques, such as machine and deep learning algorithms, can be useful for capturing complex interactions among miRNA expressions and their relationship with the concomitant disease. However, to the best of our knowledge, the use of machine learning techniques to correlate miRNAs with autoimmune diseases has not been studied so far.

Modeling miRNAs using machine learning methods poses some challenges. Redundant information is usually present in the data; furthermore, not all features are likely to be significant for classification purposes. This affects and sometimes invalidates the predictive modeling process. For this reason, feature selection techniques are commonly used to select a subset of relevant features (Inza et al. 2007). Selecting a subset of the most important features has several advantages: model simplification; shorter training time; mitigation of the overfitting problem; and so on (Tang et al. 2014). Moreover, while large samples are usually required to create accurate predictive models, in biological domains it happens that the number of features is very high, while the number of available observations is quite low. Furthermore, biomedical datasets are often unbalanced, as the number of positive samples (patients with a given disease) is typically less than the number of negative samples. When studying rare pediatric diseases, these problems are exacerbated because it is difficult both to find an adequate number of patients and to recruit healthy children who are allowed to provide blood samples by parents or legal tutors.

The aforementioned limitations require specific data processing techniques capable of reducing the dimensionality of the input space, while balancing the samples under study, before any machine learning algorithm is applied. Moreover, as the results of the analyses must be validated by expert knowledge that should confirm the involvement of selected miRNAs in pathological conditions, intelligent tools that combine human expertise and computational methods for advanced data analysis are needed to develop more reliable predictive models.

In Casalino et al. (2019a) we carried out a preliminary investigation in which an artificial neural network was trained to learn to automatically separate PedMS subjects from healthy children. Specifically, we used a Chi-squared test as feature selection, along with a random oversampling and a hold-out validation scheme to build and evaluate the model. Although promising results were obtained, a fixed scheme was used to validate the classification performance of the predictive model, thus preventing more general conclusions from being drawn. In this paper, we extend our previous analysis (Casalino et al. 2019a) by considering different techniques for feature selection, oversampling and modeling, which are fairly compared for a more robust assessment of predictive performance. This work also integrates a study in which we managed to find some machine learning models capable of detecting the presence of cognitive decline in the PedMS cohort compared to HCs (Casalino et al. 2020).

The dataset under study includes miRNA expressions from children with MS and healthy children, as well as children with Attention Deficit Hyperactivity Disorder (ADHD). In fact, ADHD patients have been observed to share some cognitive impairments with patients with PedMS. Hence, deriving a predictive model capable of distinguishing between the two diseases can be of great help for domain experts. Our work intends to develop a multi-class classification model that is also able to discriminate between ADHD and PedMS, based on miRNA expressions.

The rest of this paper is structured as follows. Section 2 deals with related work. Section 3 describes the experimental setup. Section 4 discusses the results obtained. Section 5 concludes the work.

2 Related work

In recent decades, a digitization process has involved several aspects of daily life such as school, relationships, work, industry, etc. The healthcare sector has not been excluded from this transformation process.

E-Health, medical informatics, health informatics, telemedicine, telehealth and mHealth are different terms used to refer to the same concept: the use of Information and Communication Technology solutions for health, healthcare and well-being.

Digital innovation has brought many benefits to the healthcare sector, becoming increasingly important. The big amount of structured personal health data is a large and valuable resource with many potential benefits. Faster diagnosis, improved monitoring, more effective treatment, reduction of medical errors, increased awareness of one's health and healthcare opportunities are just a few examples of the emerging trend. In recent years, the World Health Organization has used the more flexible expression *digital health* to refer to the application of Artificial Intelligence, Internet of Things, Big Data and Data Analytics to the healthcare sector. These technologies typically rely on huge amounts of different types of data, such as medical images, electronic health records, physiological signals, behavioral data, environmental data and biological data. Automatic techniques are therefore needed to manage this large amount of data in order to extract useful knowledge (Fang et al. 2016).

Several machine learning algorithms have been proposed to extract meaningful knowledge from medical data. Decision tools support clinicians in making faster decisions about patient conditions (Altaf et al. 2017; Shortliffe and Sepúlveda 2018). Diagnostic tools use predictive algorithms to infer the presence or severity of a particular disease. The availability of a wide variety of data has made it possible the application of these technologies to various diseases, such as cancer (Cardillo et al. 2017; Huang et al. 2019), neurodegenerative diseases (Coviello et al. 2020; Diaz et al. 2019; Lella et al. 2019; Vessio 2019), heart disorders (Azar et al. 2016; Powar et al. 2019), diabetes (El-Sappagh et al. 2018), sleep apnea (Mencar et al. 2019), just to mention a few.

In the field of biological data analysis, Bioinformatics has acquired a growing interest due to the recent flow of data from DNA, genomic sequences and functional genomics obtained through the diffusion of NGS technology. The main role of Bioinformatics is to apply IT facilities such as databases and software for biological data management, but the information extraction process requires algorithms capable of managing the complex relationships hidden in such data (Caponetti et al. 2014; Casalino et al. 2019b; Esposito et al. 2019). In particular, the study of gene expression quickly required the expertise of the machine learning community, as it was tested in a huge number of biomedical applications (Di Gangi et al. 2018; Dimauro et al. 2019), thus producing large datasets that require advanced tools to be analysed.

Machine learning has been extensively applied to microarray data, e.g. (Afshar et al. 2018; Hinchcliff et al. 2019; Lancashire et al. 2009; Shipp et al. 2002) and some papers have been presented on NGS data (Leung et al. 2016). Some recent work can be found on the application of (mostly) random forests and artificial neural networks to NGS miRNA data, applied to the search for biomarkers from saliva (Rosato et al. 2019), urine (Ben-Dov et al. 2016), and for melanoma (Torres et al. 2018) or other tumors (Elias et al. 2017; Liao et al. 2018). Some works apply machine learning to microarray data for the study of MS (Acquaviva et al. 2019; Fagone et al. 2019), while an extensive literature search actually provides few results on machine learning applied to NGS data for adult MS (He et al. 2019). To the best of our knowledge, currently the only NGS dataset for transcriptomic analysis of pediatric MS is the one used in this work. This study is the first attempt to apply machine learning approaches to this kind of data.

3 Experimental setting

The main objective of this work was to develop a data science-based framework for PedMS classification. We followed a classic workflow, including data acquisition, model building and model evaluation. In addition, a feature importance analysis was carried out. To cope with the highly dimensional and unbalanced data processed by us, data have been processed in order to obtain a more informative and easy-to-compute representation. In particular, three pre-processing steps were performed, namely normalization, feature selection, and class balancing. These pre-processing steps were part of the model selection process, meaning they were applied only to the training set. Indeed, the a priori application of normalization, feature selection and oversampling to the entire dataset inadvertently introduces a serious bias into the classification workflow that can lead to overly optimistic performance (Hastie et al. 2009). The experimental methodology is described in detail in the following subsections.

3.1 Data

The data used for the present study were produced at the Institute for Biomedical Technologies of the Italian National Research Council (ITB-CNR), by sequencing small RNAs of peripheral blood samples obtained from 47 children. The sequence data files produced were processed with a standard bioinformatic pipeline. Extensive descriptions of this pipeline are reported in previous works (Liguori et al. 2017;

Nuzziello et al. 2019). Sequences were compared to known miRNAs databases, and sequence counts were calculated to estimate miRNA expressions. The resulting dataset includes expressions from the 1287 miRNAs detected in the 47 study participants. Subjects differ based on healthy conditions. In addition to healthy controls, we analyzed some patients with Multiple Sclerosis and others with ADHD. The number of subjects for each class, as well as some demographic characteristics, are reported in Table 1.

More precisely, the dataset consists of 47 rows and 1287 columns. Each row represents a single patient, who can be healthy, or affected by PedMS or ADHD. A column represents a specific miRNA associated with patients: each value is an expression of that specific miRNA for the specific patient. All expression values are numeric and not negative, and there is no missing value in the data. To get an idea of our data, the expression values of the first 15 miRNAs for all study participants are shown in Fig. 1.

It is worth noting that the age ranges between the pathological groups do not overlap. While ADHD is diagnosed during the early school years, the pediatric onset of MS is a rare event and its diagnosis is often retrospective. This is why the clinicians who devised this project decided to also include some teen-aged patients in the recruitment phase.

Table 1 Data summary

Condition	# Samples	Age (avg ± SD)	M/F
HC	20	8.83 ± 3.26	6/14
PedMS	19	15.48 ± 2.74	10/9
ADHD	8	9.78 ± 2.63	8/0

3.2 Model fitting

Since each classification algorithm has its own mechanism for learning a model from data and thus its behaviour can vary significantly depending on the distribution of the data, we compared different state-of-the-art supervised learning algorithms suitable for classification. In particular, we considered random forests, extremely randomized trees and artificial neural networks.

Random forest (RF) is a tree-based method for classification or regression that relies on the concept of *bagging* to build a "forest" of decision trees at training time and provide the majority vote of the classes predicted by the individual trees at test time (Breiman 2001). The bagging procedure consists of the iterative selection of a random sample with replacement from the training set and fitting a decision tree to this sample. Contrary to ordinary bagging, when building a decision tree, RF does not take into account the overall set of features but chooses random subsets. This is to avoid the growth of highly correlated trees. In extremely randomized trees (ET), randomness is pushed one step further, even by randomizing the cut-point choice while splitting each node of a tree (Geurts et al. 2006). This allows the variance of the forest to be further reduced, at the expense of a slight increase in bias. In the present work, for both RF and ET, 500 trees were used to build the forest: this is a common choice. As a splitting criterion for the construction of the trees, we used the popular Gini index. The Gini index or Gini impurity is a measure of the probability of a particular variable being misclassified when it is chosen at random. It is calculated as follows:

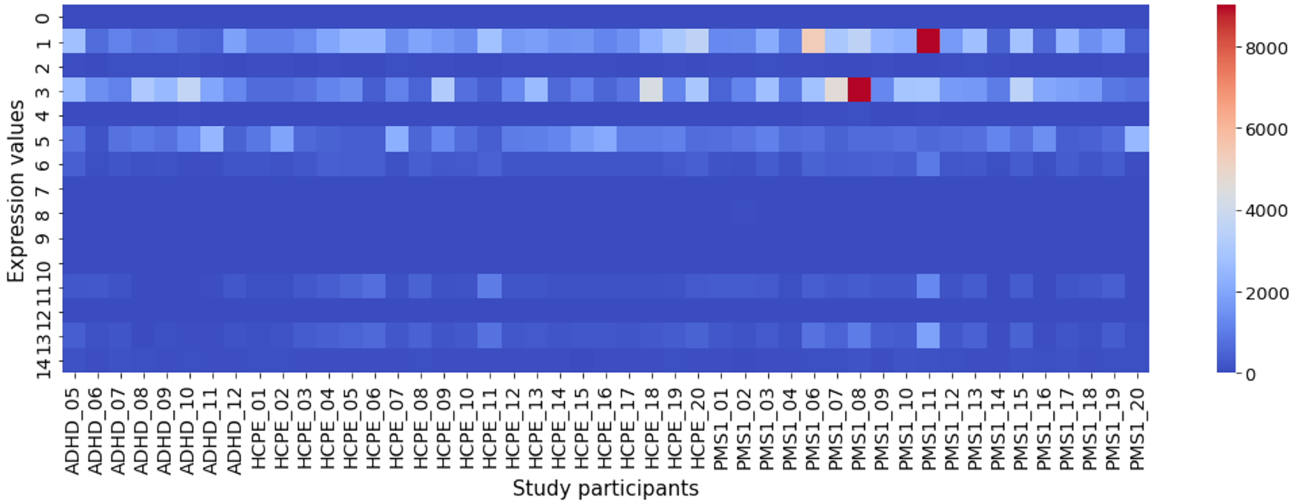

Fig. 1 Heatmap showing the expression values of the first 15 miRNAs of our data for all study participants. Note that in our dataset, healhy subjects are referred to as HCPE, while pediatric subjects with MS as PMS. The cooler rows indicate poorly expressed miRNAs that are likely to be ignored during model building by the feature selection

$$Gini = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where $k = 1, \dots, K$ are the different classes and $\hat{p}_{mk}$ is the proportion of examples labeled with the class k at node m. It is worth pointing out that entropy is a good alternative to the Gini index: they are both measures of impurity. We found that when applied to our data these metrics turn out to be interchangeable and give the same results. We preferred to use the Gini index as it is faster to compute than entropy, which is computationally heavier due to the logarithm in the equation.

As an artificial neural network architecture, we used a classic multi-layer perceptron (MLP). An MLP is a feedforward neural network capable of learning a non-linear function approximator either for classification or regression (Murphy 2018). Contrary to the traditional logistic regression algorithm, which is based on a single weighted linear combination between the input layer and the output layer, an MLP has one or more non-linear (*hidden*) layers, which learn to represent the initial input with more abstract features. In the present work, we considered an MLP with two hidden layers, each with 32 hidden units. During our experiments, we found that slightly reducing or increasing the number of hidden units provides lower performance. Likewise, the use of more hidden layers has a negative impact on the classification performance, given the very large number of parameters to be optimized with respect to the very limited number of samples. As an activation function, we used the commonly used ReLU. Since the classification task is not binary, the output layer performs a softmax activation:

$$softmax(z)_i = \frac{\exp(z_i)}{\sum_{l=1}^{k} \exp(z_l)},$$

where K is the number of classes and z_i is the i-th element of the softmax input, corresponding to class i. The output is a vector of the probabilities of a sample x to belonging to each class: the prediction provided by the network is the class with the highest probability. The network tries to optimize the cross-entropy loss function:

$$\mathcal{L} = -\sum_{k=1}^{K} y_{i,k} \log(p_{i,k}),$$

where y is a binary indicator that evaluates 1 if the class label k is the correct prediction for sample i, 0 otherwise; while $p_{i,k}$ is the predicted probability that i is of class k. The loss function was minimized through backpropagation using the Limited-memory BFGS algorithm (Liu and Nocedal 1989). This is an optimization algorithm in the family of quasi-Newton methods that is known to perform well when, as in this case, the training data is small (Morales and Nocedal 2011).

3.3 Normalization

Before feeding the classification algorithms with data, they underwent a normalization, feature selection and oversampling process. The activity of miRNAs and their relative expression can have effect on different scales. This could hamper the quality of the predictive model derived from the data. To mitigate this effect, we applied a common standardization so that all features values were bounded in the interval $[-1, 1]$. From each value x_i in each feature x, a standardized value $\hat{x}_i$ is obtained by removing the mean value μ and dividing it by its standard deviation σ:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}.$$

3.4 Feature selection

As mentioned above, the number of features in our dataset is disproportionately greater than the number of subjects. To significantly reduce the dimensionality of the data without loosing too much information, we propose to apply a feature selection method. It is worth noting that feature selection techniques are generally preferred to feature extraction techniques in this context, as they preserve initial information. In fact, in medical applications this is a necessary condition to obtain the interpretability of the results.

For classification purposes, we have fixed the feature selection strategy to be used and we have chosen one based on a linear support vector machine (SVM) (Scholkopf and Smola 2001) classifier. Linear models penalized with a regularization term, such as the ℓ_1 or ℓ_2 norm, have sparse solutions, as many of their coefficients are zeroed or reduced to very small values. When the goal is to reduce the dimensionality to use the reduced feature space with another classifier, linear models can be used to select the non-zero or higher coefficients. To select features based on their importance weights, we fit a linear SVM, with ℓ_2 penalty and regularization parameter C equals to 1, and dynamically kept the top-ranked features. It is worth noting that ℓ_1 is better for making feature vectors sparse instead of ℓ_2. We made this choice for performance reasons, as using ℓ_1 produces lower results. One reason for this could be that many small values in our feature vectors add up and provide some useful information for classification.

Additionally, because each feature selection technique can produce slightly different results than alternative techniques, we also used two other feature selection strategies to perform a feature importance analysis. More specifically, we used a recursive feature elimination (RFE) with the same linear SVM classifier and a selection based on the feature importance given by an ET model, with the same parameters previously described. As for RFE, the method removes the least significant features in iterations (Yan and Zhang 2015). The process is computed iteratively until all features are removed from the feature set, then the final output is a ranked feature list. Regarding the tree-based method, for each tree, the feature importance was calculated as the decrease of node impurity weighted by the expected fraction of the samples reaching that node. For the overall forest, the normalized feature importance were simply summed. Overlapping the three feature rankings produces a feature set that represents a more robust selection of the most important features that describe the examples in the dataset.

3.5 Class balancing

As discussed above, the dataset exhibits class imbalance (see Table 1), as the number of patients with ADHD is significantly lower than the number of subjects belonging to the other two classes. Class imbalance should be avoided as it may mislead the classification results. Common methods for balancing a dataset are undersampling and oversampling (Theodoridis and Koutroumbas 2008). In undersampling, a subset of samples is removed from the over-represented class. In oversampling, new samples are generated from the under-represented class. Undersampling is not suitable for our dataset, as the total number of samples is small. For this reason, we applied oversampling.

There are many oversampling techniques available; among these we have chosen two very popular strategies, i.e. random oversampling and SMOTE (Chawla et al. 2002). Random oversampling is a naïve strategy in which new samples are simply generated by random sampling with replacement of currently available samples. SMOTE, on the other hand, generates synthetic data points using the k-nearest neighbor algorithm. Given a sample x, a new sample x' is generated considering its k nearest neighbors. Then, one of the nearest neighbors x_z is selected and a sample is created as follows:

$$x' = x_i + \lambda(x_z - x_i),$$

where $\lambda \in [0, 1]$ is a random number. In this work, we set $k = 5$. This setting was due to the small size of the dataset which constrains the use of small values for this parameter, otherwise the algorithm is unable to find k neighbors of a sample in the feature space. Lower values result in slightly lower performance.

3.6 Validation

The classification performance was validated with a 5-fold cross-validation. This scheme is generally preferred when dealing with small datasets, as in our case. With this scheme, the set of examples is divided into five folds: one fold is treated as a test set; the remaining folds form the training set. The whole procedure is repeated 5 times, until each fold is used once as a test set. It is worth noting that the splitting was stratified by diagnosis so that each fold contained approximately the same number of subjects from each diagnostic group. We also experimented with leave-one-out, where only one example is selected as a test instance each time, yielding very similar results. We preferred to use cross-validation for computational purposes, as it is much less computationally expensive than leave-one-out.

4 Results

The results of two experiments are reported below. In the first experiment, we compared the three classification models, i.e. RF, ET and MLP, on the original dataset without oversampling. Then, we chose the best model to evaluate which oversampling technique was capable of providing better performance. In the second experiment, we performed a feature importance analysis. First, we compared the features automatically selected by the methods we used with the features selected by the domain experts. Finally, we compared the best model, based on the automatically selected features, to a model based on the features that matched the domain expertise.

Classification results are reported in terms of well-known metrics: accuracy, precision and recall. Accuracy is the fraction of correctly classified instances relative to the overall dataset:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative predictions, respectively. Precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP}.$$

Intuitively, precision is the model's ability not to label a negative sample as positive. Similarly, recall is calculated as follows:

$$Recall = \frac{TP}{TP + FN}.$$

Table 2 Classification performance of the three supervised learning algorithms without oversampling

Class	RF (Accuracy = 0.72)		ET (Accuracy = 0.74)		MLP (Accuracy = 0.79)	
	Precision	Recall	Precision	Recall	Precision	Recall
HC	0.78	0.90	0.83	0.95	0.90	0.90
PedMS	0.67	0.74	0.71	0.79	0.76	0.84
ADHD	0.67	0.25	0.33	0.12	0.50	0.38

Intuitively, recall is the model's ability to find all the positive instances. It is worth noting that in the following are the mean values for each classification metric, averaged over all iterations of the cross-validation scheme.

4.1 Classification performance

Table 2 reports the results obtained from the three classification models on the original dataset without oversampling. As can be seen, the best overall accuracy was obtained by MLP, achieving a value of 0.79 in the correct classification of the three classes. By looking at the precision and recall value for each class, it can be observed that the three models were collectively able to accurately detect healthy control subjects. Conversely, as regards diseased patients, the performance obtained suggest that all models are deceived by possibly overlapping pathological patterns and tend to mistakenly categorize subjects with ADHD as subjects with PedMS. However, it should be noted that this may have happened due to the few ADHD samples.

To overcome this bias, we chose the best algorithm, namely MLP, and replicated the same classification by applying a random oversampling or SMOTE. The results are shown in Table 3. Applying random oversampling does not improve the recognition accuracy of the ADHD class, but it also reduces performance on the other two classes. It appears that duplicating the same data does not help to find meaningful patterns in ADHD, but only introduces additional variance in the data. Instead, performance improves using SMOTE: both the prediction accuracy of PedMS class and that of the ADHD class increase, although the latter continues to show unsatisfactory results.

4.2 Comparison with features selected by domain experts

Table 4 shows the 40 top-ranked features, among the initial 1287 features, selected as an overlap between the feature rankings of the methods described in Section 3.4. Table 5, on the other hand, reports a subset of the 42 most important features proposed by biology experts. This subset contains some results of the differential expression analysis previously performed on this dataset and a list of miRNAs known to be involved in the disease under consideration. We only kept the top 40 features of the ranked list produced from our feature selection process to make a fairer comparison to the 42 important features previously selected by the domain experts. It can be noted that the set of features selected by domain experts is partially overlapping with the set of automatically selected features. Specifically, the two lists share nine common miRNAs (let-7i-3p, miR-125a-5p, miR-128-3p, miR-130b-3p, miR-221-3p, miR-484, miR-501-3p, miR-652-3pm, miR-942-5p), a complementary miRNA (let-7b-3p instead of 5p) and two variants (miR-30c-2-3p and miR-30d-3p instead of miR-30e-3p), for a total of 12 miRNAs already known to be involved in pediatric MS. The list of the remaining 28 miRNAs was presented to the domain experts and they found that nearly all of them (26 out of 28) have a role in neurodegeneration and other neuronal functions such as neuron genesis, development, projection and death. In addition, significant upregulation of miR-744 was observed in peripheral blood mononuclear cells of treatment-naïve MS patients compared to controls and the expression level of miR-939 was lower in plasma samples from MS patients compared to controls (Søndergaard et al. 2013), while in a very recent study miR-4286 was found to be over-expressed in lesions of the gray matter of MS (Fritsche et al. 2019).

Table 3 Classification performance of MLP with random oversampling and SMOTE

Class	Random (Accuracy = 0.74)		SMOTE (Accuracy = 0.81)	
	Precision	Recall	Precision	Recall
HC	0.85	0.85	0.90	0.90
PedMS	0.71	0.79	0.80	0.84
ADHD	0.50	0.38	0.57	0.50

Finally, to complete our evaluation we compared the best model obtained with the automatically selected features in combination with SMOTE with the same model but trained on the subset of 42 features based on domain knowledge: results are reported in Table 6. We found that the model obtained by training the neural network on the features proposed by the experts achieves an overall classification accuracy of 0.79, which is slightly lower than the overall accuracy achieved by the model obtained with our selection of features, i.e. 0.81. Interestingly, the model using the features selected by the experts shows perfect recall for the healthy control class, while showing lower prediction accuracy in patient classification. In particular, performance for the ADHD class

Table 4 List of 40 miRNAs selected by our feature selection procedure

let-7b-3p	let-7i-3p	miR-125a-5p	miR-128-3p	miR-130b-3p
miR-1468-5p	miR-221-3p	miR-30c-2-3p	miR-30d-3p	miR-3150a-3p
miR-3200-5p	miR-3667-3p	miR-381-3p	miR-3909	miR-4286
miR-4489	miR-4731-3p	miR-4731-5p	miR-4746-5p	miR-484
miR-501-3p	miR-5096	miR-543	miR-5695	miR-582-3p
miR-619-5p	miR-652-3p	miR-6770-3p	miR-6801-3p	miR-6816-3p
miR-6820-3p	miR-6868-3p	miR-6883-3p	miR-6884-3p	miR-7110-3p
miR-744-5p	miR-758-3p	miR-877-5p	miR-939-5p	miR-942-5p

Table 5 List of 42 miRNAs selected by domain experts

let-7a-5p	let-7b-5p	let-7i-3p	let-7i-5p	miR-10a-5p
miR-125a-5p	miR-128-3p	miR-1304-3p	miR-1307-3p	miR-130b-3p
miR-140-3p	miR-144-5p	miR-148b-3p	miR-151a-3p	miR-151b
miR-15b-5p	miR-16-2-3p	miR-16-5p	miR-181a-2-3p	miR-181a-5p
miR-182-5p	miR-185-3p	miR-185-5p	miR-21-5p	miR-221-3p
miR-25-3p	miR-26a-5p	miR-26b-3p	miR-26b-5p	miR-27b-3p
miR-28-3p	miR-29a-3p	miR-30e-3p	miR-30e-5p	miR-320a
miR-3605-3p	miR-484	miR-501-3p	miR-652-3p	miR-6842-3p
miR-942-5p	miR-99b-5p			

Table 6 Comparison of the classification performance obtained by the proposed MLP with the features automatically selected and SMOTE and those obtained by the same MLP with features based on domain expertise and SMOTE

Class	Proposed (accuracy = 0.81)		Experts (accuracy = 0.79)	
	Precision	Recall	Precision	Recall
HC	0.90	0.90	0.91	1.00
PedMS	0.80	0.84	0.78	0.74
ADHD	0.57	0.50	0.43	0.38

drops further. This can be explained considering that the experts selected their features by an analysis of miRNAs that are expressed differently between HC and PedMS, while no experience was gained on patients with ADHD.

5 Conclusions

In this paper, we presented a classification study of pediatric patients with Multiple Sclerosis, based on their miRNA expressions obtained with NGS technology. A first contribution consisted in proposing a neural network model for the automatic discrimination of PedMS from healthy controls and children with ADHD. Furthermore, a comparison with a classification model built on the 42 features proposed by the domain experts confirmed the effectiveness of the proposed method in correctly identifying significant features that led to promising classification results. The model developed was more accurate in detecting the healthy control group. Specificity above sensitivity indicates that the decision support system is better at detecting the absence of disease in the healthy population than at detecting the presence of disease in the pathological group. These results suggest that a screening test based on our tool may be able to correctly exclude disease from the healthy population, so it may be useful for ruling in disease when a positive response is obtained. As for the diseased classes, the proposed model is confusing. This behaviour may be justified by the fact that ADHD patients share some cognitive impairments with PedMS patients, so there may be significant overlap between the molecular pathways of these two categories. 27% of PedMS patients, in fact, tend to show cognitive symptoms similar to those with ADHD (Weisbrot et al. 2014).

Of course, the results obtained strongly depend on the available dataset which is quite small. Indeed, the size of the current dataset has been limited by the rarity of the PedMS onset and includes all the patients recruited over a 3-year period. In fact, as previously mentioned, the pediatric onset of MS is a rare event and its diagnosis is often retrospective. Unfortunately, collecting a large sample of data in clinical settings is a time-consuming and expensive process involving many aspects, including privacy issues, which are even more delicate when dealing with child patients, as is our case. It would be useful to repeat the experiments proposed in this work with a larger dataset. Also, as future work, it would be interesting to consider other feature engineering or representation learning strategies to achieve the goal of model creation and compare performance with the method proposed here.

As a final observation, we report that this work was the first attempt to create a classification model useful to support experts in analyzing miRNA data obtained from pediatric patients. This model can be used as a tool to distinguish the three classes of diagnosis in a fully automatic way, uncovering hidden relationships among miRNAs that cannot be derived from a classic differential expression analysis. Furthermore, the 40 miRNAs automatically selected by the proposed feature selection method can be further analyzed to derive other biological observations, such as an assessment of the genes that are regulated by those miRNAs and an analysis of the molecular pathways involved in the activation of the target genes, both for the study of pediatric Multiple Sclerosis and for novel investigations about ADHD. To fully assess the importance of the selected features, a thorough biological investigation is required since most of the functions of miRNAs are still unknown. This work represents the first step towards the development of an intelligent system capable of supporting the expert in the analysis of miRNA expressions for the early diagnosis of pediatric Multiple Sclerosis. To this end, further work is underway to combine miRNA expression data with other patient clinical data in order to obtain more powerful diagnostic support tools.

Acknowledgements The molecular data for this investigation derive from a fully supported grant (cod. 2014/R/10) from Fondazione Italiana Sclerosi Multipla (FISM). This work was partially supported by INdAM GNCS within the research project “Computational Intelligence methods for Digital Health”. Ga.C., Gi.C., A.C. and G.V. are members of the INdAM GNCS research group.

Author Contributions All authors contributed equally to this work.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

References

Acquaviva M, Menon R, Di Dario M, Dalla Costa G, Romeo M, Sangalli F, Colombo B, Moiola L, Martinelli V, Comi G et al (2019) Design of an unbiased machine learning workflow to predict multiple sclerosis staging from blood transcriptome. Mult Scler J 25:908–908

Afshar S, Afshar S, Warden E, Manochehri H, Saidijam M (2018) Application of artificial neural network in miRNA biomarker selection and precise diagnosis of colorectal cancer. Iran Biomed J 23(3):175–183

Akbar N, Till C, Sled JG, Binns MA, Doesburg SM, Aubert-Broche B et al (2016) Altered resting-state functional connectivity in cognitively preserved pediatric-onset ms patients and relationship to structural damage and cognitive performance. Mult Scler J 22(6):792–800

Altaf W, Shahbaz M, Guergachi A (2017) Applications of association rule mining in health informatics: a survey. Artif Intell Rev 47(3):313–340

Azar AT, Kumar SS, Inbarani HH, Hassanien AE (2016) Pessimistic multi-granulation rough set-based classification for heart valve disease diagnosis. Int J Model Identif Control 26(1):42–51

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2):281–297

Ben-Dov IZ, Whalen VM, Goilav B, Max KE, Tuschl T (2016) Cell and microvesicle urine microRNA deep sequencing profiles from healthy individuals: observations with potential impact on biomarker studies. PloS One 11(1):e0147249

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Caponetti L, Castellano G, Basile MT, Corsini V (2014) Fuzzy mathematical morphology for biological image segmentation. Appl Intell 41(1):117–127

Cardillo FA, Masulli F, Rovetta S (2017) Automatic approaches for CE-MRI examination of the breast: a survey. In: 2017 IEEE international conference on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData), IEEE, pp 147–154

Casalino G, Castellano G, Consiglio A, Liguori M, Nuzziello N, Primiceri D (2019a) A predictive model for microRNA expressions in pediatric multiple sclerosis detection. In Torra V, Narukawa Y, Pasi G, Viviani M (eds) Modeling decisions for artificial intelligence. Springer International Publishing, Cham., pp 177–188

Casalino G, Coluccia M, Pati ML, Pannunzio A, Vacca A, Scilimati A, Perrone MG (2019b) Intelligent microarray data analysis through non-negative matrix factorization to study human multiple myeloma cell lines. Appl Sci 9(24):5552

Casalino G, Vessio G, Consiglio A (2020) Evaluation of cognitive impairment in pediatric multiple sclerosis with machine learning: an exploratory study of miRNA expressions. In: 2020 IEEE conference on evolving and adaptive intelligent systems (EAIS), IEEE, pp 1–6

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chitnis T, Glanz B, Jaffin S, Healy B (2009) Demographics of pediatric-onset multiple sclerosis in an MS center population from the Northeastern United States. Mult Scler J 15(5):627–631

Coviello G, Avitabile G, Florio A (2020) A synchronized multi-unit wireless platform for long-term activity monitoring. Electronics 9(7):1118

Di Gangi M, Bosco GL, Rizzo R (2018) Deep learning architectures for prediction of nucleosome positioning from sequences data. BMC Bioinform 19(14):418

Diaz M, Ferrer MA, Impedovo D, Pirlo G, Vessio G (2019) Dynamically enhanced static handwriting representation for Parkinson’s disease detection. Pattern Recogn Lett 128:204–210

Dimauro G, Colagrande P, Carlucci R, Ventura M, Bevilacqua V, Caivano D (2019) CRISPRLearner: a deep learning-based system to predict CRISPR/Cas9 sgRNA on-target cleavage efficiency. Electronics 8(12):1478

El-Sappagh S, Alonso JM, Ali F, Ali A, Jang J, Kwak K (2018) An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access 6:37371–37394

Elias KM, Fendler W, Stawiski K, Fiascone SJ, Vitonis AF, Berkowitz RS, Frendl G, Konstantinopoulos P et al (2017) Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer. Elife 6:e28932

Esposito F, Gillis N, Del Buono N (2019) Orthogonal joint sparse NMF for microarray data analysis. J Math Biol 1–25

Fagone P, Mazzon E, Mammana S, Di Marco R, Spinasanta F, Basile MS, Petralia MC, Bramanti P, Nicoletti F, Mangano K (2019) Identification of CD4+ T cell biomarkers for predicting the response of patients with relapsing-remitting multiple sclerosis to natalizumab treatment. Mol Med Rep 20(1):678–684

Fang R, Pouyanfar S, Yang Y, Chen S-C, Iyengar S (2016) Computational health informatics in the big data age: a survey. ACM Comput Surveys (CSUR) 49(1):12

Fritsche L, Teuber-Hanselmann S, Soub D, Harnisch K, Mairinger F, Junker A (2019) MicroRNA profiles of MS gray matter lesions identify modulators of the synaptic protein synaptotagmin-7. Brain Pathol

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

He K, Huang S, Qian X (2019) Early detection and risk assessment for chronic disease with irregular longitudinal data analysis. J Biomed Inform 103231

Hinchcliff ME, Frech TM, Wood TA, Huang C-C, Lee J, Aren K, Ryan JJ, Wilson B, Beussink-Nelson L, Whitfield ML et al (2017) Machine learning of the cardiac phenome and skin transcriptome to categorize heart disease in systemic sclerosis. bioRxiv, p 213678

Huang S, Yang J, Fong S, Zhao Q (2019). Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett

Huang Y, Shen XJ, Zou Q, Wang SP, Tang SM, Zhang GZ (2011) Biological functions of microRNAs: a review. J Physiol Biochem 67(1):129–139

Inza I, Larraaga P, Saeys Y (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Lancashire LJ, Lemetre C, Ball GR (2009) An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform 10(3):315–329

Lella E, Amoroso N, Diacono D, Lombardi A, Maggipinto T, Monaco A, Bellotti R, Tangaro S (2019) Communicability characterization of structural DWI subcortical networks in Alzheimer's disease. Entropy 21(5):475

Leung MK, Delong A, Alipanahi B, Frey BJ (2016) Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE 104(1):176–197

Liao Z, Li D, Wang X, Li L, Zou Q (2018) Cancer diagnosis through IsomiR expression with machine learning method. Curr Bioinform 13(1):57–63

Liguori M, Nuzziello N, Licciulli F, Consiglio A, Simone M, Viterbo RG et al (2017) Combined microRNA and mRNA expression analysis in pediatric multiple sclerosis: an integrated approach to uncover novel pathogenic mechanisms of the disease. Hum Mol Genet 27(1):66–79

Liguori M, Nuzziello N, Simone M, Amoroso N, Viterbo RG, Tangaro S, Consiglio A, Giordano P, Bellotti R, Trojano M (2019) Association between miRNAs expression and cognitive performances of pediatric multiple sclerosis patients: a pilot study. Brain Behav e01199

Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Program 45(1–3):503–528

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucl Acids Res 40(10):4288–4297

Mencar C, Gallo C, Mantero M, Tarsia P, Carpagnano GE, Foschino Barbaro MP, Lacedonia D (2019) Application of machine learning to predict obstructive sleep apnea syndrome severity. Health Inf J 26(1):298–317

Morales JL, Nocedal J (2011) Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization." ACM Trans Math Softw (TOMS) 38(1):7–1

Murphy KP (2018) Machine learning: a probabilistic aerspective (adaptive computation and machine learning series). The MIT Press, London, UK

Nuzziello N, Craig F, Simone M, Consiglio A, Licciulli F, Margari L, Grillo G, Liuni S, Liguori M (2019) Integrated analysis of microRNA and mRNA expression profiles: An attempt to disentangle the complex interaction network in attention deficit hyperactivity disorder. Brain Sci 9(10):288

Olsson T, Barcellos LF, Alfredsson L (2017) Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol 13(1):25

Powar A, Shilvant S, Pawar V, Parab V, Shetgaonkar P, Aswale S (2019). Data mining & artificial intelligence techniques for prediction of heart disorders: a survey. In 2019 international conference on vision towards emerging trends in communication and networking (ViTECoN), IEEE, pp 1–7

Rosato AJ, Chen X, Tanaka Y, Farrer LA, Kranzler HR, Nunez YZ, Henderson DC, Gelernter J, Zhang H (2019) Salivary microRNAs identified by small RNA sequencing and machine learning as potential biomarkers of alcohol dependence. Epigenomics 11(7):739–749

Scholkopf B, Smola AJ (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS et al (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1):68

Shortliffe EH, Sepúlveda MJ (2018) Clinical decision support in the era of artificial intelligence. JAMA 320(21):2199–2200

Søndergaard HB, Hesse D, Krakauer M, Sørensen PS, Sellebjerg F (2013) Differential microrna expression in blood in multiple sclerosis. Mult Scler J 19(14):1849–1857

Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review. Algorithms and applications, data classification, p 37

Theodoridis S, Koutroumbas K et al (2008) Pattern recognition. IEEE Trans Neural Netw 19(2):376

Torres R, Lang UE, Hejna M, Shelton SJ, Joseph NM, Shain AH, Yeh I, Wei ML, Oldham MC, Bastian BC et al (2018) A machine-learning classifier trained with microRNA ratios to distinguish melanomas from nevi. bioRxiv, p 507400

Vessio G (2019) Dynamic handwriting analysis for neurodegenerative disease assessment: a literary review. Appl Sci 9(21):4666

Weisbrot D, Charvet L, Serafin D, Milazzo M, Preston T, Cleary R, Moadel T et al (2014) Psychiatric diagnoses and cognitive impairment in pediatric multiple sclerosis. Mult Scler J 20(5):588–593

Yan K, Zhang D (2015) Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sens Actuat B Chem 212:353–363

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.