Computers in Biology and Medicine 152 (2023) 106373

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Differential diagnosis of systemic lupus erythematosus and Sjögren's syndrome using machine learning and multi-omics data

Jordi Martorell-Marugán [a,b,c,*], Marco Chierici [b], Giuseppe Jurman [b], Marta E. Alarcón-Riquelme [d,e], Pedro Carmona-Sáez [a,c,**]

[a] Department of Statistics and OR, University of Granada, Granada, 18071, Spain
[b] Data Science for Health Research Unit, Fondazione Bruno Kessler, Trento, 38123, Italy
[c] Bioinformatics Unit, GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS Granada, Granada, 18016, Spain
[d] Genetics of Complex Diseases, GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, Andalusian Regional Government, PTS Granada, Granada, 18016, Spain
[e] Unit of Chronic Inflammatory Diseases, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, 17177, Sweden

ARTICLE INFO

Keywords:
Machine learning
Modeling and prediction
Bioinformatics
Clustering
Classification and association rules
Health

ABSTRACT

Systemic lupus erythematosus and primary Sjogren's syndrome are complex systemic autoimmune diseases that are often misdiagnosed. In this article, we demonstrate the potential of machine learning to perform differential diagnosis of these similar pathologies using gene expression and methylation data from 651 individuals. Furthermore, we analyzed the impact of the heterogeneity of these diseases on the performance of the predictive models, discovering that patients assigned to a specific molecular cluster are misclassified more often and affect to the overall performance of the predictive models. In addition, we found that the samples characterized by a high interferon activity are the ones predicted with more accuracy, followed by the samples with high inflammatory activity. Finally, we identified a group of biomarkers that improve the predictions compared to using the whole data and we validated them with external studies from other tissues and technological platforms.

1. Introduction

Systemic autoimmune diseases (SADs) are heterogeneous and complex pathologies whose main hallmark is an immune response to self-antigens, causing tissue injury and failure in different organs. Although these diseases cause different symptoms, they share risk factors [1], clinical manifestations [2], and molecular alterations (e.g., gene expression alterations [3]).

In order to establish a molecular classification of SADs patients, whole blood transcriptome and methylome data have been previously used from patients from seven SADs, including systemic lupus erythematosus (SLE) and primary Sjogren's syndrome (pSjS) [4]. In a previous work, we used these data to demonstrate the existence of four subgroups of SADs patients based on clustering analysis of molecular profiles [4]. The defined clusters were named 'inflammatory', 'lymphoid', 'interferon' and 'undefined' and each one was characterized by different serological, cellular, genetic and clinical features. Overexpression and hypomethylation of genes and CpGs from neutrophil and monocyte-driven modules were characteristic of the inflammatory cluster. T and NK cell functions defined the lymphoid cluster, whereas interferon, viral, and dendritic cell functions defined the interferon cluster. The undefined cluster did not present distinct functional modules compared to healthy controls. B lymphocyte functions were seen in both the lymphoid and interferon clusters, and cell cycle and transcriptional upregulation were connected to the interferon cluster. The undifferentiated patterns of the undefined cluster were explained by the low disease activity found in these samples, concluding that these patients have healthy-like molecular patterns.

SLE is a complex disorder with an autoimmune background, a multifactorial etiology, impairment of several organic systems, a wide spectrum of clinical manifestations, variable prognosis, and an evolving clinical course marked by the occurrence of episodes of active disease and remission [5]. It is interesting to note that the prevalence of SLE is rising across the globe, with rates varying from 40 to >160 per 100,000

* Corresponding author. Department of Statistics and OR, University of Granada, Granada, 18071, Spain.
** Corresponding author. Department of Statistics and OR, University of Granada, Granada, 18071, Spain.
E-mail addresses: jordi.martorell@genyo.es (J. Martorell-Marugán), pcarmona@ugr.es (P. Carmona-Sáez).

https://doi.org/10.1016/j.compbiomed.2022.106373
Received 12 July 2022; Received in revised form 20 October 2022; Accepted 27 November 2022
Available online 28 November 2022

people [6]. On the other hand, pSjS is another autoimmune chronic inflammatory clinical condition that primarily affects the lacrimal and salivary glands and results in a decrease in the salivary and lacrimal flows and, as a result, to symptoms of dry mouth and dry eyes [7].

The heterogeneity of SADs patients makes diagnosis and treatment difficult. In this context, it is sometimes challenging to differentiate between pathologies with overlapping clinical features, like SLE and pSjS. Some symptoms shared by these two diseases are hemolytic anemia, leukopenia, lymphopenia, thrombocytopenia, photosensitivity or fatigue, among others [8–12]. Nevertheless, both pathologies are characterized by the upregulation of the interferon (IFN) signature [13,14], a set of genes regulated by the IFN cytokine.

For these reasons, it is common for pSjS patients to be misdiagnosed, underdiagnosed, or diagnosed at late stages of the disease [15], even after 6–10 years from presenting the first symptoms [16]. Consequently, proper therapeutic strategies are delayed, contributing to evitable tissue damage.

Previous efforts have been made to distinguish SLE and pSjS patients from the clinical point of view [2,15], using methylation data [17], metabolomics data [18], or salivary protein biomarkers [19]. However, to the best of our knowledge, gene expression data has not been used to perform a differential diagnosis between SLE, pSjS and healthy controls. Given that measuring gene expression has become an affordable approach, we studied its potential as a diagnostic tool for these patients. Furthermore, although methylation data has been used previously to perform pairwise predictions (SLE vs. healthy, pSjS vs. healthy, and SLE vs. pSjS) [17], a multiclass classifier to distinguish the samples from these 3 groups simultaneously was not proposed. Nevertheless, we consider that a multiclass predictor may be helpful in the clinical context. Importantly, some specific dysregulation in these diseases may be observed at the gene expression and DNA methylation level, but not with other data types such as metabolomics. For instance, the IFN signature, which plays a crucial role in the development and progression of SLE and pSjS [13,14], may be directly assessed with gene expression data and indirectly with methylation data, since promoters of IFN-regulated genes (e.g., IFI44L) are differentially methylated [20]. Therefore, these are very valuable data modalities to study and predict the clinical diagnosis in this context.

In this study, we applied machine learning (ML) methodologies to classify SLE and pSjS patients, as well as healthy controls. Operatively, we designed an analysis pipeline to train a eXtreme Gradient Boosting (XGBoost) multiclass predictor for each type of data. Given the heterogeneity of these diseases, we also studied whether the performance of the XGBoost models varies according to the molecular clusters previously defined for these patients [4]. Furthermore, we obtained subsets of genes and cytosine-phosphate-guanine (CpGs) that improve the performance of the predictors and we characterized the molecular implications of these features into the studied diseases. In addition, we evaluated the predictivity of these features as biomarkers for the labels in independent datasets. Finally, we constructed specific models and feature selections for the group of patients with healthy-like molecular patterns.

2. Methods

2.1. Data

The data used in this study was produced for a previous work [4], which generated multi-omics data from whole blood samples obtained from patients with seven different SADs and healthy controls. In detail, we used matched expression and methylation data from 213 SLE patients, 181 pSjS patients and 257 healthy controls. Table 1 shows the distribution of the SLE and pSjS samples according to their assigned molecular cluster.

Illumina HiSeq2500 and Illumina Beadchip 450k technologies were used to measure the transcriptome and methylome, respectively, in each sample. For the transcriptome experiment, messenger ribonucleic acid (mRNA) was extracted and sequenced. These sequenced reads were aligned to the reference human genome and the number of reads aligned to each gene (raw counts) were extracted. These raw counts are a measure of genes expression since they are proportional to the amount of mRNA in the cells. For the methylome analysis, the intensity of the methylated and unmethylated probes in the Beadchip arrays was measured and, after appropriate data processing, M-values for each analyzed CpG was obtained. Further information about the cohorts, experimental procedures and data preprocessing is available from the previous article [4].

Table 1
Distribution of the patients samples in the four molecular clusters.

Cluster	SLE samples	pSjS samples	Total
Lymphoid	28	42	70
Inflammatory	45	23	68
Interferon	80	70	150
Undefined	60	46	106

2.2. Data processing

In order to discard uninformative features, we filtered out those genes with less than 5 counts in at least half of the samples of each group. Furthermore, we discarded those genes and CpG sites with a coefficient of variation lower than 0.4 and 0.1 respectively, retaining information for 8413 genes and 253,903 CpGs. We adjusted the expression and methylation data for the technical variables pool and sample plate, respectively, using the ComBat method [21] implemented in the sva R package. We normalized the adjusted expression values with the trimmed mean of M-values (TMM) normalization [22] implemented in the NOISeq R package [23]. We used the Uniform Manifold Approximation and Projection (UMAP) method [24] for dimension reduction to explore the gene expression data. Finally, given the number of features, we performed differential expression and methylation analyses to reduce the computational costs of the ML models. For that aim, we used the limma package [25], to compare SLE vs. Healthy, pSjS vs. Healthy and SLE vs. pSjS groups. After ranking each feature by the obtained P-value, we selected the top 2000 genes and 3000 CpGs from each comparison, obtaining a final list of 3.681 genes and 7.596 CpGs after removing duplicated features.

2.3. Model fitting and evaluation

In the first place, we used the Python library lazypredict to test the performance of 29 ML models on both the expression and the methylation data. XGBoost was the model with the highest Matthews Correlation Coefficient (MCC) for both types of data (Supplementary Table 1) and was selected for further analyses.

XGBoost is a ML algorithm that use gradient boosting and ensemble learning to combine multiple decision trees for classification and regression tasks [26]. This approach minimizes a loss function to add new trees to the model. XGBoost has become a broadly used ML method due to its good performance and scalability.

For each data type, we split the data randomly into a training set and a test set following a 80/20 partition. We standardized the training sets with the StandardScaler function of the scikit-learn Python library [27], which calculates Z-scores subtracting the mean and dividing by the standard deviation. We used the parameters learnt from the training sets (means and standard deviations) to scale the test sets. We used the training set to optimize the hyperparameters for the XGBoost classifier through a 10-fold cross-validation (CV) approach. Specifically, we optimized the learning rate, maximum tree depth, gamma, alpha, lambda, minimum child weight, subsample ratio of the training instance and subsample ratio of columns hyperparameters. We used the softmax

loss function for the multiclass classification task. Given the high number of possible combinations of hyperparameter values, we performed a grid search with 100 random combinations using scikit-learn's RandomizedSearchCV function and we selected the hyperparameter combinations with the best mean MCC on the internal validation sets.

Once the best fitting hyperparameters were selected, we trained a XGBoost model with those hyperparameters and the training set. We used that model to predict the labels of the test set and we evaluated the performance for this prediction. This workflow is shown in Fig. 1. The whole process was repeated 10 times to avoid biases due to the training and test set splitting of the first step, following the FDA-SEQC guidelines for reproducibility. We used the Python library scikit-learn [27] to apply all the described ML methodologies.

To evaluate the performance of the models, we calculated the accuracy (defined as the fraction of correct predictions), the mean precision for the three classes (1) weighted by the number of samples of each class, and the weighted F1 score (2).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Being TP the number of true positives and FP the number of false positives.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2}$$

Where recall is calculated following (3).

$$Recall = \frac{TP}{TP + TN} \tag{3}$$

Being TN the true negatives.

We also calculated the MCC [28,29], which is a balanced measure of accuracy and precision that can be used to evaluate multiclass models. MCC maximum value is 1, which indicates a perfect prediction. The MCC for K classes is calculated following (4) [30].

$$MCC = \frac{c \times s - \sum_{k}^{K} p_k \times t_k}{\sqrt{\left(s^2 - \sum_{k}^{K} p_k^2\right) \times \left(s^2 - \sum_{k}^{K} t_k^2\right)}} \tag{4}$$

Where t_k is the amount of k labelled samples in the data; p_k is the number of times that class k was predicted; c is the number of samples correctly predicted and s is the total number of samples.

2.4. Variable selection and functional analysis

For feature selection, we excluded the data from the patients belonging to the undefined cluster, which have a healthy-like molecular pattern. For each of the 10 rounds of the model fitting and evaluation (Fig. 1), we ranked the features of the model (i.e., genes and CpGs) based on their Gini importances [31] in the XGBoost models fitted with the training sets. Then, we obtained the mean position of each feature for the 10 folds and we sorted the features accordingly. We repeated the model fitting and evaluation with increasing subsets of top features (from 10 to 1000). We calculated the mean MCC for each subset along the 10 rounds and we selected the subset with the maximum mean MCC (i.e., the top 90 genes for expression and the top 900 CpGs for methylation).

We annotated the genes and CpGs with biomaRt [32,33] and IlluminaHumanMethylation450kanno.ilmn12.hg19 R packages respectively. We used the GeneCodis tool [34,35] to perform the functional analysis of the selected features, using as input the ENSEMBL and CpG probe identifiers and selecting Reactome as annotation. We considered as significant those pathways with a False Discovery Rate (FDR) < 0.05.

2.5. External data processing and analysis

With the aim of testing the selected biomarkers utility for the

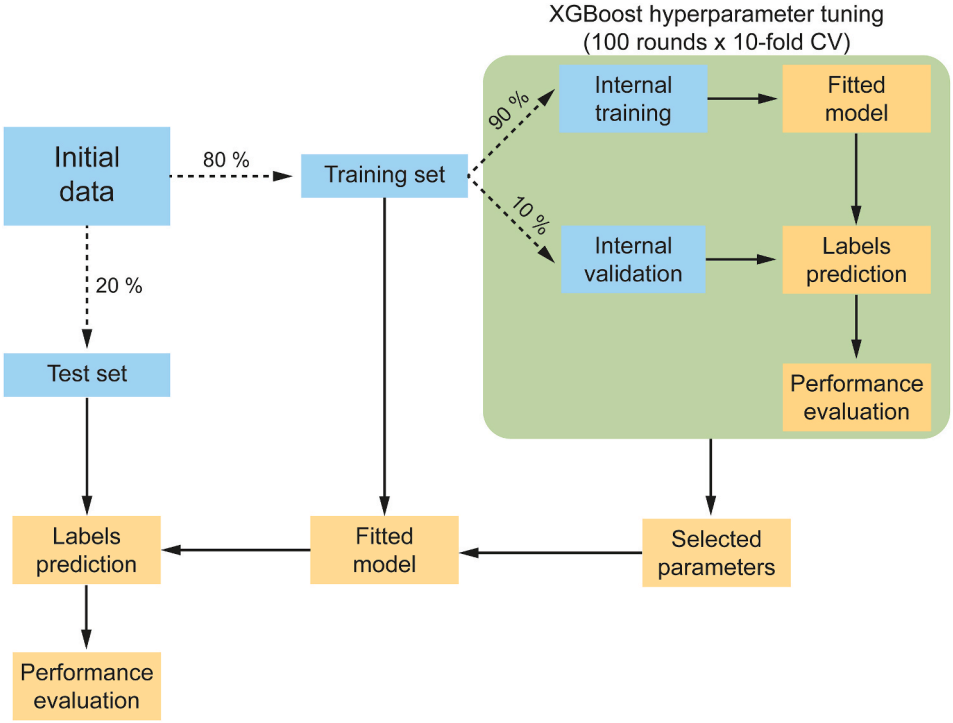

Fig. 1. Data analysis workflow. The initial data was split into training and test sets. Training sets were used to tune the XGBoost model with 100 rounds of 10-fold CV. A XGBoost model with these optimized hyperparameters was trained with the training set. This model was used to predict the labels of the test set and to evaluate the performance with the results of the prediction. This outer data splitting was repeated 10 times.

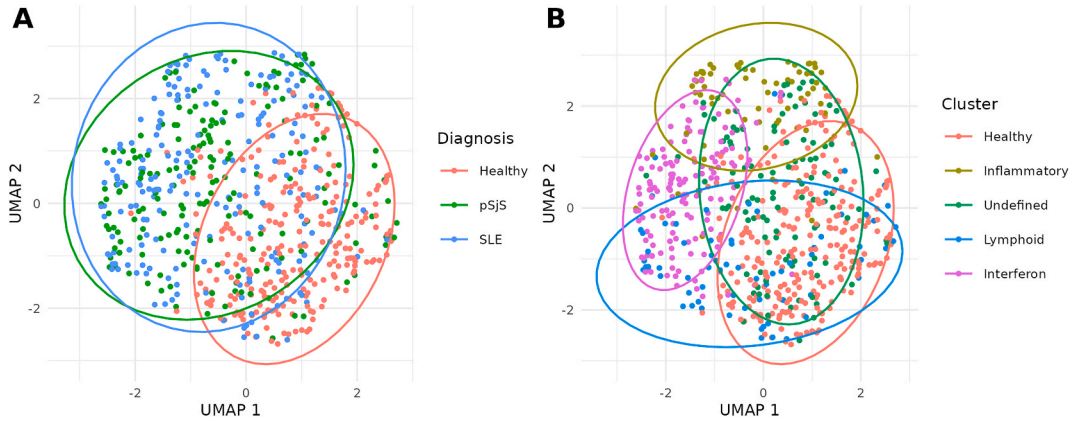

Fig. 2. UMAP plots of the gene expression data with samples colored by their clinical diagnosis (A) and by their previously assigned molecular cluster (B).

classification of healthy and diseased samples, we used public gene expression and DNA methylation datasets available in the Gene Expression Omnibus (GEO) [36]. For the gene expression dataset (GEO ID: GSE108497), we downloaded the processed data from the ADEx database [37]. Data for 62 of the 90 genes were available for this dataset. For the DNA methylation dataset (GEO ID: GSE166373), we used the GEOquery R package [38] to download the raw idat files. We processed these raw data with the minfi package [39] applying the normal-exponential out-of-band (Noob) normalization [40]. Data for the 900 selected CpGs were available for this dataset.

We followed the same analytical pipeline described previously, with the exception that we did not perform a batch effect correction for these data. We assessed the performance of the models with the accuracy and MCC metrics. Being this a binary classification, the MCC formula (4) reduces to the simpler form (5):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

The code to process and analyze the validation data is available at https://github.com/GENyO-BioInformatics/SLE_pSjS_classifier.

3. Results and discussion

3.1. Diagnosis capacity depends on the patients' molecular profile

In the first place, we represented the UMAP plots to explore how samples are grouped according to their gene expression profiles (Fig. 2). As can be observed, SLE and pSjS samples overlap completely, and samples from both diseases overlap partially with the healthy controls (Fig. 1A). These results illustrate the challenge of classifying these diseases. On the other hand, if samples are colored by their molecular cluster assigned in our previous work [4], it can be observed that the inflammatory and interferon clusters are almost perfectly separated from the healthy samples, while the lymphoid and, especially, the undefined clusters overlap with the healthy samples (Fig. 2B).

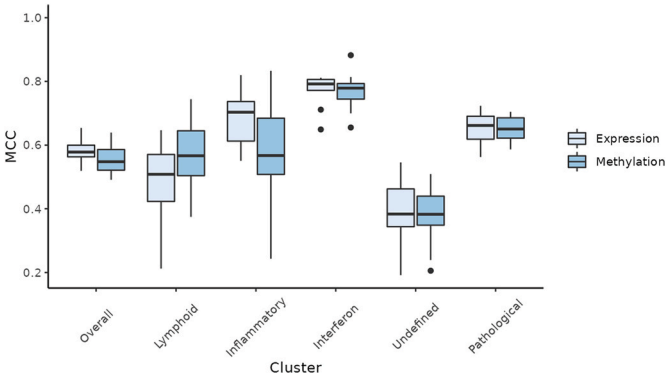

Fig. 3. MCC scores for each molecular cluster with XGBoost models based on expression and methylation data. For each cluster and data type there are 10 MCCs corresponding to the 10 rounds of training-test splitting.

Next, we used lazypredict library to calculate the mean MCCs obtained by different ML models after 10 rounds of training/test splitting (Supplementary Table 1). The model with the highest mean MCC in expression and methylation data was XGBoost, which was selected to perform the classifications.

We trained and evaluated XGBoost models to classify the three

Table 2
Accuracy, precision, F1 and MCC scores for the test sets of each group of samples.

Group	Accuracy		Precision		F1		MCC	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Expression								
Overall	0.7221	0.0284	0.7162	0.0313	0.7077	0.0237	0.5791	0.0409
Lymphoid	0.8515	0.0433	0.5674	0.1668	0.5010	0.0887	0.4705	0.1466
Inflammatory	0.9077	0.0370	0.6680	0.1622	0.6272	0.1025	0.6802	0.0900
Interferon	0.8805	0.0262	0.8014	0.0646	0.7877	0.0598	0.7717	0.0523
Undefined	0.7630	0.0323	0.5990	0.1353	0.5243	0.0857	0.3840	0.1075
Pathological	0.7780	0.0398	0.7525	0.0399	0.7351	0.0364	0.6538	0.0551
Methylation								
Overall	0.7053	0.0330	0.6952	0.0346	0.6865	0.0342	0.5546	0.0484
Lymphoid	0.8667	0.0473	0.6540	0.1588	0.5858	0.1047	0.5709	0.1135
Inflammatory	0.8815	0.0441	0.6502	0.1961	0.5760	0.1289	0.5855	0.1668
Interferon	0.8780	0.0359	0.8045	0.0390	0.7801	0.0538	0.7692	0.0622
Undefined	0.7644	0.0308	0.6168	0.1558	0.4971	0.0754	0.3767	0.0985
Pathological	0.7780	0.0296	0.7566	0.0310	0.7395	0.0303	0.6522	0.0401

classes of our cohort (SLE, pSjS and healthy) considering all the samples ('overall' group). Accuracy, weighted precision and weighted F1 scores for our predictors are reported in Table 2. However, it has been reported that these metrics may be inflated [41]. For that reason, we also calculated the MCCs and we interpreted the results based on these scores.

Our first multiclass predictor model, which used the complete datasets, achieved a test set MCC of 0.5791 ± 0.0409 and 0.5546 ± 0.0484 for expression and methylation data, respectively (Table 2).

In previous work, we demonstrated that SADs patients may be classified in four molecular clusters with different characteristics [4]. For this reason, we wondered whether the performance of our models varies between the different groups of samples. To answer this question, we repeated the analyses for each individual cluster of samples (i.e., lymphoid, inflammatory, interferon and undefined clusters).

Using the expression data, these models showed a differential performance among clusters (Table 2, Fig. 3), with remarkable low performance for the undefined cluster (MCC = 0.3840 ± 0.1075). This cluster of patients was previously associated with low disease activity and showed a healthy-like molecular pattern [4], so they were expected to be poorly predicted. At the other extreme, the interferon cluster showed the best prediction results (MCC = 0.7717 ± 0.0523), followed by the inflammatory cluster results were close (MCC = 0.6802 ± 0.0900). These results are also in accordance with the characterization of these two molecular clusters, since they were associated with the most extreme manifestations (e.g., nephritis and thrombosis for the interferon cluster, and fibrosis complications for the inflammatory cluster) [4]. On the other hand, the moderate performance for the lymphoid cluster (MCC = 0.4705 ± 0.1466) is also coherent with the mild symptoms associated with this cluster (e.g., gastrointestinal manifestations) [4].

We followed the same methodology to construct predictor models based on methylation data, observing the same trend regarding the differential prediction capacity in the different molecular clusters (Table 2, Fig. 3). Furthermore, we decided to repeat the overall analysis excluding the samples belonging to the undefined cluster, given that they represent healthy-like molecular profiles and they were probably affecting the overall model accuracy. We defined this new group of samples as the 'pathological' group, which includes the samples from the inflammatory, lymphoid and interferon clusters. The pathological group comprises the molecular heterogeneity found for active SLE and pSjS patients and our predictor showed a middle performance between the individual pathological clusters (MCC = 0.6538 ± 0.0551 and 0.6522 ± 0.0401 for expression and methylation data respectively).

3.2. Variable selection improves the predictors performance

Following a feature selection strategy (see Methods) we identified the most informative subsets of genes and CpGs, which could be used as biomarkers to differentiate SLE, pSjS and healthy controls. In this way, we proposed a set of 90 genes for gene expression (Supplementary Table 2) and 900 CpGs for DNA methylation (Supplementary Table 3) that may be used to distinguish SLE and pSjS, both from healthy controls and between them. For gene expression, the mean MCC for the pathological cluster increased from 0.6538 to 0.7310 ± 0.0312 using the subset of 90 genes. For DNA methylation, the MCC increased from 0.6522 to 0.6717 ± 0.0572.

We performed an enrichment analysis to get insight into the biological pathways in which the selected biomarkers are involved. We discovered that the most enriched pathways are related to relevant processes in SADs, such as interferon signaling and interleukins signaling (Supplementary Table 4). We obtained similar results with the selected CpGs methylation sites (Supplementary Table 5, Fig. 4).

In addition, we constructed an integrated model concatenating the selected genes and CpGs in a new dataset containing gene expression and DNA methylation data. Using these data, we obtained a mean MCC of 0.7159 ± 0.0387, what is lower than using only the gene expression, but higher than using only the DNA methylation.

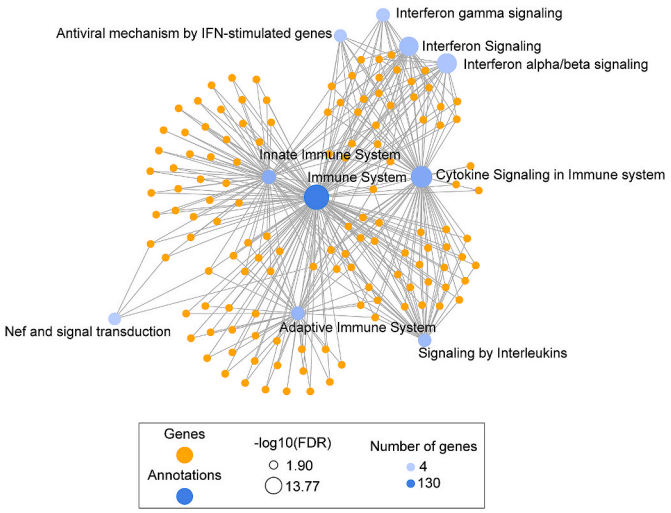

Fig. 4. Network with the top 10 most significant pathways for the enrichment analysis of the methylation biomarkers. Reactome pathways (blue circles) are connected to the corresponding genes (orange circles). The circles size is proportional to the statistical significance and the color intensity to the number of biomarkers.

Furthermore, we tested the utility of selecting these features for predictive tasks on external public datasets available in GEO [27]. For gene expression, we used a dataset with 325 SLE patients and 187 healthy controls (GEO ID: GSE108497), which used the Illumina HumanHT-12 V4.0 expression beadchip technology to measure gene expression in whole blood samples. Following the same workflow that we used to prepare diagnostic models for our data, we obtained predictive models with a mean accuracy of 0.8496 ± 0.0365 and a mean MCC of 0.6712 ± 0.0836. For DNA methylation, we used a dataset with 64 pSjS patients and 67 healthy controls (GEO ID: GSE166373), which used the Illumina HumanMethylation450 BeadChip and the Illumina Infinium MethylationEPIC platforms to measure the DNA methylation levels in labial salivary gland samples. For this dataset, the predictive models achieved a mean accuracy of 0.7926 ± 0.0894 and a mean MCC of 0.5958 ± 0.1764. These results suggest that the selected biomarkers may be useful for the diagnosis of SLE and pSjS patients, even for data from different technologies and tissues than the ones used in our cohort.

3.3. The undefined cluster has a specific gene expression signature

As previously commented, the undefined cluster contains a group of patients with molecular patterns very similar to healthy controls. For that reason, it is very challenging to classify these samples with expression and methylation data. Nevertheless, we tried to improve the results with specific models and feature selection for this subset of samples. For that aim, we performed the feature selection process described in Section 3.2 for this group, selecting 50 genes for gene expression (Supplementary Table 6) and 640 CpGs for DNA methylation (Supplementary Table 7). The mean MCCs were 0.5746 ± 0.0811 and 0.4914 ± 0.1079 for expression and methylation data respectively. Although the performance without feature selection was similar for expression and methylation (Table 2, Fig. 3), the results are better with expression after feature selection. Interestingly, only 14 of the 50 genes (28%) overlap with the 90 selected genes for the pathological group, while the majority of CpGs (624 of 640, 97.5%) overlap with the selected CpGs for the pathological group. These results may indicate that the undefined cluster has specific gene expression alterations that may help to diagnose these difficult samples, while the alterations at DNA methylation level are similar to the other groups and are not as useful to perform this classification.

4. Conclusions

SLE and pSjS are two SADs with some overlapping symptoms. High throughput technologies such as RNA-Seq and microarrays may be valuable tools for the differential diagnosis of these pathologies. In this work, we demonstrated that ML methodologies can predict the disease status of each patient from expression and methylation data, although the prediction capacity depends on the molecular background of the patients. Furthermore, we selected subsets of features that improve the predictions and have important roles in the pathological mechanisms of SADs. We validated these features with external datasets, demonstrating their capability to diagnose each disease from both expression and methylation data. Finally, we obtained specific sets of biomarkers that may be useful to classify the samples from the undefined cluster, which have a healthy-like molecular pattern. Although some previous works used different types of data to perform differential diagnosis of SLE and pSjS, as far as we know the present study appears to be the first work describing multiclass classifiers to perform this task from expression and methylation data. In our opinion, the results of this work demonstrate the potential use of transcriptome and methylome data to perform differential diagnoses of SADs using ML approaches.

Declaration of competing intrest

None Declared.

Acknowledgments

This work was funded by grant PID2020-119032RB-I00 funded by MCIN/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (grants P20_00335 and B-CTS-40-UGR20), EU/EFPIA Innovative Medicines Initiative Joint Undertaking PRECISESADS (115565) and the Gustaf den Ve:80-års fond for MEAR. JMM is funded by European Union – NextGenerationEU, Ministerio de Universidades (Spain's Government) and Recovery, Transformation and Resilience Plan, through a call from the University of Granada. JMM was partially funded by EMBO (Grant ASTF 8692). Competing interest: The authors declare that they have no conflicting interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106373.

References

[1] S. Jörg, D.A. Grohme, M. Erzler, M. Binsfeld, A. Haghikia, D.N. Müller, R.A. Linker, M. Kleinewietfeld, Environmental factors in autoimmune diseases and their role in multiple sclerosis, Cell. Mol. Life Sci. 73 (2016) 4611–4622, https://doi.org/10.1007/s00018-016-2311-1.

[2] F. Assan, R. Seror, X. Mariette, G. Nocturne, New 2019 SLE EULAR/ACR classification criteria are valuable for distinguishing patients with SLE from patients with pSS, Ann. Rheum. Dis. (2019), https://doi.org/10.1136/annrheumdis-2019-216222 annrheumdis-2019-216222.

[3] D. Toro-Domínguez, P. Carmona-Sáez, M.E. Alarcón-Riquelme, Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis, Arthritis Res. Ther. 16 (2014) 489, https://doi.org/10.1186/s13075-014-0489-x.

[4] G. Barturen, S. Babaei, F. Català-Moll, M. Martínez-Bueno, Z. Makowska, J. Martorell-Marugán, P. Carmona-Sáez, D. Toro-Domínguez, E. Carnero-Montoro, M. Teruel, M. Kerick, M. Acosta-Herrera, L. Le Lann, C. Jamin, J. Rodríguez-Ubreva, A. García-Gómez, J. Kageyama, A. Buttgereit, S. Hayat, J. Mueller, R. Lesche, M. Hernandez-Fuentes, M. Juarez, T. Rowley, I. White, C. Marañón, T. Gomes Anjos, N. Varela, R. Aguilar-Quesada, F.J. Garrancho, A. López-Berrio, M. Rodriguez Maresca, H. Navarro-Linares, I. Almeida, N. Azevedo, M. Brandão, A. Campar, R. Faria, F. Farinha, A. Marinho, E. Neves, A. Tavares, C. Vasconcelos, E. Trombetta, G. Montanelli, B. Vigone, D. Alvarez-Errico, T. Li, D. Thiagaran, R. Blanco Alonso, A. Corrales Martínez, F. Genre, R. López Mejías, M.A. Gonzalez-Gay, S. Remuzgo, B. Ubilla Garcia, R. Cervera, G. Espinosa, I. Rodríguez-Pintó, E. De Langhe, J. Cremer, R. Lories, D. Belz, N. Hunzelmann, N. Baerlecken, K. Kniesch, T. Witte, M. Lehner, G. Stummvoll, M. Zauner, M.A. Aguirre-Zamorano, N. Barbarroja, M.C. Castro-Villegas, E. Collantes-Estevez, E. de Ramon, I. Díaz Quintero, A. Escudero-Contreras, M.C. Fernández Roldán, Y. Jiménez Gómez, I. Jiménez Moleón, R. Lopez-Pedrera, R. Ortega-Castro, N. Ortego, E. Raya, C. Artusi, M. Gerosa, P.L. Meroni, T. Schioppo, A. De Groof, J. Ducreux, B. Lauwerys, A.-L. Maudoux, D. Cornec, V. Devauchelle-Pensec, S. Jousse-Joulin, P.-E. Jouve, B. Rouvière, A. Saraux, Q. Simon, M. Alvarez, C. Chizzolini, A. Dufour, D. Wynar, A. Balog, M. Bocskai, M. Deák, S. Dulic, G. Kádár, L. Kovács, Q. Cheng, V. Gerl, F. Hiepe, L. Khodadadi, S. Thiel, E. de Rinaldis, S. Rao, R.J. Benschop, C. Chamberlain, E.R. Dow, Y. Ioannou, L. Laigle, J. Marovac, J. Wojcik, Y. Renaudineau, M.O. Borghi, J. Frostegård, J. Martín, L. Beretta, E. Ballestar, F. McDonald, J.-O. Pers, M.E. Alarcón-Riquelme, Integrative analysis reveals a molecular stratification of systemic autoimmune diseases, Arthritis Rheumatol. 73 (2021) 1073–1085, https://doi.org/10.1002/art.41610.

[5] M. Di Battista, E. Marcucci, E. Elefante, A. Tripoli, G. Governato, D. Zucchi, C. Tani, A. Alunno, One year in review 2018: systemic lupus erythematosus, Clin. Exp. Rheumatol. 36 (2018) 763–777.

[6] M.J. Lewis, A.S. Jawad, The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus, Rheumatology 56 (2017) i67–i77, https://doi.org/10.1093/rheumatology/kew399.

[7] K. Psianou, I. Panagoulias, A.D. Papanastasiou, A.-L. de Lastic, M. Rodi, P. I. Spantidea, S.E. Degn, P. Georgiou, A. Mouzaki, Clinical and immunological parameters of Sjögren's syndrome, Autoimmun. Rev. 17 (2018) 1053–1064, https://doi.org/10.1016/j.autrev.2018.05.005.

[8] X. Mariette, L.A. Criswell, Primary Sjögren's syndrome, N. Engl. J. Med. 378 (2018) 931–939, https://doi.org/10.1056/NEJMcp1702514.

[9] D.D. Gladman, D. Ibañez, M.B. Urowitz, Systemic lupus erythematosus disease activity index 2000, J. Rheumatol. 29 (2002) 288–291.

[10] R. Seror, P. Ravaud, S.J. Bowman, G. Baron, A. Tzioufas, E. Theander, J.-E. Gottenberg, H. Bootsma, X. Mariette, C. Vitali, EULAR Sjögren's Task Force, EULAR Sjogren's syndrome disease activity index: development of a consensus systemic disease activity index for primary Sjogren's syndrome, Ann. Rheum. Dis. 69 (2010) 1103–1109, https://doi.org/10.1136/ard.2009.110619.

[11] M. Petri, A.-M. Orbai, G.S. Alarcón, C. Gordon, J.T. Merrill, P.R. Fortin, I.N. Bruce, D. Isenberg, D.J. Wallace, O. Nived, G. Sturfelt, R. Ramsey-Goldman, S.-C. Bae, J. G. Hanly, J. Sánchez-Guerrero, A. Clarke, C. Aranow, S. Manzi, M. Urowitz, D. Gladman, K. Kalunian, M. Costner, V.P. Werth, A. Zoma, S. Bernatsky, G. Ruiz-Irastorza, M.A. Khamashta, S. Jacobsen, J.P. Buyon, P. Maddison, M.A. Dooley, R. F. van Vollenhoven, E. Ginzler, T. Stoll, C. Peschken, J.L. Jorizzo, J.P. Callen, S. S. Lim, B.J. Fessler, M. Inanc, D.L. Kamen, A. Rahman, K. Steinsson, A.G. Franks, L. Sigler, S. Hameed, H. Fang, N. Pham, R. Brey, M.H. Weisman, G. McGwin, L. S. Magder, Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus, Arthritis Rheum. 64 (2012) 2677–2686, https://doi.org/10.1002/art.34473.

[12] M.C. Hochberg, Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus, Arthritis Rheum. 40 (1997) 1725, https://doi.org/10.1002/art.1780400928.

[13] S. Bezalel, K.M. Guri, D. Elbirt, I. Asher, Z.M. Sthoeger, Type I interferon signature in systemic lupus erythematosus, Isr. Med. Assoc. J. 16 (2014) 246–249.

[14] C.Q. Nguyen, A.B. Peck, The interferon-signature of Sjögren's syndrome: how unique biomarkers can identify underlying inflammatory and immunopathological mechanisms of specific diseases, Front. Immunol. 4 (2013) 142, https://doi.org/10.3389/fimmu.2013.00142.

[15] A. Rasmussen, L. Radfar, D. Lewis, K. Grundahl, D.U. Stone, C.E. Kaufman, N. L. Rhodus, B. Segal, D.J. Wallace, M.H. Weisman, S. Venuturupalli, B.T. Kurien, C. J. Lessard, K.L. Sivils, R.H. Scofield, Previous diagnosis of Sjögren's Syndrome as rheumatoid arthritis or systemic lupus erythematosus, Rheumatology 55 (2016) 1195–1201, https://doi.org/10.1093/rheumatology/kew023.

[16] R. Manthorpe, K. Asmussen, P. Oxholm, Primary Sjögren's syndrome: diagnostic criteria, clinical features, and disease activity, J. Rheumatol. Suppl. 50 (1997) 8–11.

[17] J. Imgenberg-Kreuz, J.C. Almlöf, D. Leonard, C. Sjöwall, A.-C. Syvänen, L. Rönnblom, J.K. Sandling, G. Nordmark, Shared and unique patterns of DNA methylation in systemic lupus erythematosus and primary Sjögren's syndrome, Front. Immunol. 10 (2019), https://doi.org/10.3389/fimmu.2019.01686.

[18] A.A. Bengtsson, J. Trygg, D.M. Wuttge, G. Sturfelt, E. Theander, M. Donten, T. Moritz, C.-J. Sennbro, F. Torell, C. Lood, I. Surowiec, S. Rännar, T. Lundstedt, Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis, PLoS One 11 (2016), e0159384, https://doi.org/10.1371/journal.pone.0159384.

[19] S. Hu, K. Gao, R. Pollard, M. Arellano, H. Zhou, L. Zhang, D. Elashoff, C. G. Kallenberg, A. Vissink, D.T. Wong, Preclinical validation of salivary biomarkers for primary Sjögren's syndrome, Arthritis Care Res. 62 (2010) 1633–1638, https://doi.org/10.1002/acr.20289.

[20] M. Zhao, Y. Zhou, B. Zhu, M. Wan, T. Jiang, Q. Tan, Y. Liu, J. Jiang, S. Luo, Y. Tan, H. Wu, P. Renauer, M. Del Mar Ayala Gutiérrez, M.J. Castillo Palma, R. Ortega Castro, C. Fernández-Roldán, E. Raya, R. Faria, C. Carvalho, M.E. Alarcón-Riquelme, Z. Xiang, J. Chen, F. Li, G. Ling, H. Zhao, X. Liao, Y. Lin, A.H. Sawalha, Q. Lu, IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus, Ann. Rheum. Dis. 75 (2016) 1998–2006, https://doi.org/10.1136/annrheumdis-2015-208410.

[21] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127, https://doi.org/10.1093/biostatistics/kxj037.

[22] M.D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data, Genome Biol. 11 (2010) R25, https://doi.org/10.1186/gb-2010-11-3-r25.
[23] S. Tarazona, P. Furió-Tarí, D. Turrà, A.D. Pietro, M.J. Nueda, A. Ferrer, A. Conesa, Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package, Nucleic Acids Res. 43 (2015) e140, https://doi.org/10.1093/nar/gkv711.
[24] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold approximation and projection, J. Open Source Software 3 (2018) 861, https://doi.org/10.21105/joss.00861.
[25] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47, https://doi.org/10.1093/nar/gkv007.
[26] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, https://doi.org/10.1145/2939672.2939785.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[28] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta Protein Struct. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.
[29] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424, https://doi.org/10.1093/bioinformatics/16.5.412.
[30] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, Comput. Biol. Chem. 28 (2004) 367–374, https://doi.org/10.1016/j.compbiolchem.2004.09.006.
[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[32] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber, BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis, Bioinformatics 21 (2005) 3439–3440, https://doi.org/10.1093/bioinformatics/bti525.
[33] S. Durinck, P.T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, Nat. Protoc. 4 (2009) 1184–1191, https://doi.org/10.1038/nprot.2009.97.
[34] P. Carmona-Saez, M. Chagoyen, F. Tirado, J.M. Carazo, A. Pascual-Montano, GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists, Genome Biol. 8 (2007) R3, https://doi.org/10.1186/gb-2007-8-1-r3.
[35] A. Garcia-Moreno, R. López-Domínguez, J.A. Villatoro-García, A. Ramirez-Mena, E. Aparicio-Puerta, M. Hackenberg, A. Pascual-Montano, P. Carmona-Saez, Functional enrichment analysis of regulatory elements, Biomedicines 10 (2022) 590, https://doi.org/10.3390/biomedicines10030590.
[36] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (2002) 207–210.
[37] J. Martorell-Marugán, R. López-Domínguez, A. García-Moreno, D. Toro-Domínguez, J.A. Villatoro-García, G. Barturen, A. Martín-Gómez, K. Troule, G. Gómez-López, F. Al-Shahrour, V. González-Rumayor, M. Peña-Chilet, J. Dopazo, J. Sáez-Rodríguez, M.E. Alarcón-Riquelme, P. Carmona-Sáez, A comprehensive database for integrated analysis of omics data in autoimmune diseases, BMC Bioinf. 22 (2021) 343, https://doi.org/10.1186/s12859-021-04268-4.
[38] S. Davis, P.S. Meltzer, GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (2007) 1846–1847, https://doi.org/10.1093/bioinformatics/btm254.
[39] M.J. Aryee, A.E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A.P. Feinberg, K. D. Hansen, R.A. Irizarry, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, Bioinformatics 30 (2014) 1363–1369, https://doi.org/10.1093/bioinformatics/btu049.
[40] T.J. Triche, D.J. Weisenberger, D. Van Den Berg, P.W. Laird, K.D. Siegmund, Low-level processing of Illumina Infinium DNA methylation BeadArrays, Nucleic Acids Res. 41 (2013), https://doi.org/10.1093/nar/gkt090 e90-e90.
[41] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.