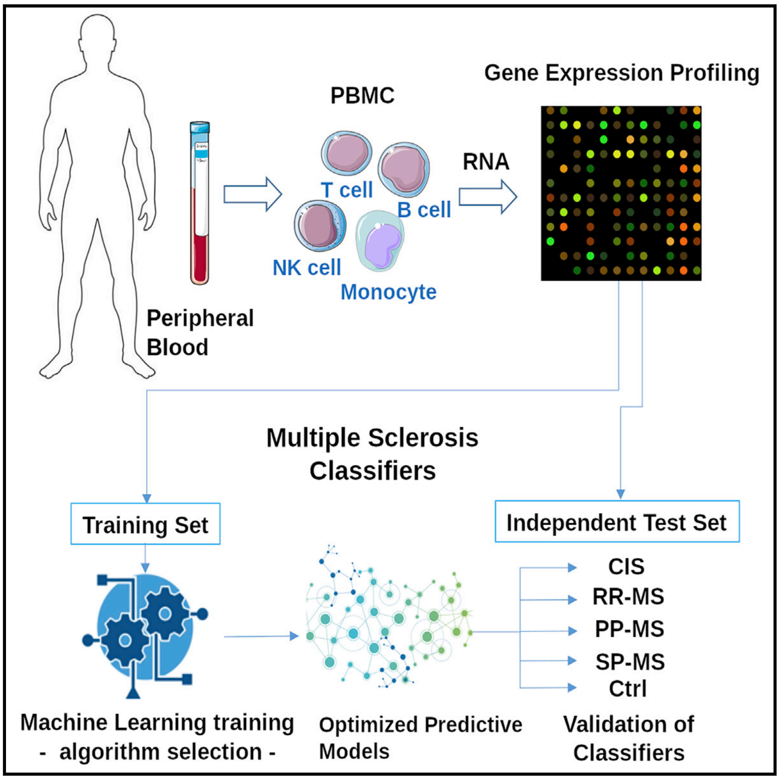
Cell Reports
Medicine

Article

Inferring Multiple Sclerosis Stages from the Blood Transcriptome via Machine Learning

Graphical Abstract


Gene Expression Profiling
PBMC
RNA
T cell
B cell
NK cell
Monocyte
Peripheral Blood
Multiple Sclerosis Classifiers
Training Set
Independent Test Set
CIS
RR-MS
PP-MS
SP-MS
Ctrl
Machine Learning training - algorithm selection -
Optimized Predictive Models
Validation of Classifiers

Highlights

- Generated PBMC transcriptomes from multiple sclerosis and control subjects
- Unbiased machine learning workflow allows algorithm comparison and optimization
- Classifiers built on training cohort have high accuracy in the independent test set
- PBMC transcriptomes identify disease state and stage in multiple sclerosis

Authors

Massimo Acquaviva, Ramesh Menon, Marco Di Dario, ..., Vittorio Martinelli, Giancarlo Comi, Cinthia Farina

Correspondence

farina.cinthia@hsr.it

In Brief

Acquaviva et al. describe the application of machine learning to transcriptional profiles of peripheral immune cells from more than 300 healthy and neurological subjects. Classification models built on the training cohort display high accuracy in the independent test set and identify disease state and stage in multiple sclerosis.

Acquaviva et al., 2020, Cell Reports Medicine *1*, 100053
July 21, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.xcrm.2020.100053

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Inferring Multiple Sclerosis Stages from the Blood Transcriptome via Machine Learning

Massimo Acquaviva,[1] Ramesh Menon,[1] Marco Di Dario,[1] Gloria Dalla Costa,[1] Marzia Romeo,[1] Francesca Sangalli,[1] Bruno Colombo,[1] Lucia Moiola,[1] Vittorio Martinelli,[1] Giancarlo Comi,[1] and Cinthia Farina[1,2,*]
[1]Institute of Experimental Neurology and Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy
[2]Lead Contact
*Correspondence: farina.cinthia@hsr.it

https://doi.org/10.1016/j.xcrm.2020.100053

SUMMARY

Peripheral blood mononuclear cells (PBMCs) bear specific dysregulations in genes and pathways at distinct stages of multiple sclerosis (MS) that may help with classifying MS and non-MS subjects, specifying the early stage of disease, or discriminating among MS courses. Here we describe an unbiased machine learning workflow to build MS stage-specific classifiers based on PBMC transcriptomics profiles from more than 300 individuals, including healthy subjects and patients with clinically isolated syndromes, relapsing-remitting MS, primary or secondary progressive MS, or other neurological disorders. The pipeline, designed to optimize and compare the performance of distinct machine learning algorithms in the training cohort, generates predictive models not influenced by demographic features, such as age and gender, and displays high accuracy in the independent validation cohort. Proper application of machine learning to transcriptional profiles of circulating blood cells may allow identification of disease state and stage in MS.

INTRODUCTION

Multiple sclerosis (MS) is a chronic inflammatory and demyelinating disease of the CNS, characterized by clinical and biological heterogeneity. Generally, the disease starts with a first clinical episode suggestive of MS, classified as clinically isolated syndrome (CIS), which evolves to defined MS in cases of further clinical or neuroradiological activity.[1] Up to 85% of MS patients develop the relapsing-remitting (RR) course of MS, characterized by periodic neurologic deterioration followed by partial or complete remission, and several RR MS subjects eventually evolve to secondary progressive (SP MS, where worsening of neurologic function occurs in the absence of recognizable relapses. Approximately 15 % of MS patients develop the primary progressive (PP) course from onset of the disease.

Despite continuous refinement of diagnostic criteria, the degree of MS misdiagnosis remains quite high, especially for PP MS, because this form manifests MRI findings often overlapping with RR MS or other neurodegenerative and vascular diseases.[2,3] Further, disease prognosis is poorly predictable. For example, the transition from RR MS to SP MS can be defined retrospectively when a sustained period of worsening neurologic impairment has been observed, often resulting in a 2- to 3-year period of diagnostic uncertainty.[4]

Recent advances in artificial intelligence applied to MRI data show promising results in discriminating MS conditions from healthy controls[5,6] but have limited power in classification of single MS stages.[7,8] Because aberrant transcriptional profiles have been described in peripheral blood of patients at distinct MS stages,[9–12] we hypothesized that peripheral blood mononuclear cell (PBMC) transcriptomes contain useful information to build specific classifiers for the different MS forms. We generated and analyzed transcriptomic profiles of PBMCs from more than 300 individuals, including healthy subjects (HCs) and patients with CIS, RR MS, PP MS, SP MS, or other neurological disorders (ONDs), and assembled independent training and validation cohorts. For definition of the optimal classifier, several issues need to be taken into account, including selection of the proper machine learning algorithm and dataset characteristics such as feature and sample numbers. Moreover, collected MS cohorts usually suffer from imbalances among MS stages, with PP MS and RR MS being the rarest and most frequent form, respectively. To overcome these issues, we developed a machine learning workflow, based on nested cross-validation (NCV), able to conduct unbiased optimization and comparison of different algorithms in the training dataset prior to final validation in the independent test set. To limit detrimental effects related to class imbalance and optimize prediction of the class of interest (usually the rarest), each algorithm was specifically fine-tuned in each prediction task. To avoid potential biases introduced by a separate feature selection step,[13] we oriented our choice toward decision tree-based algorithms, which can perform feature selection internally as part of the learning process, such as the powerful and largely employed random forests (RFs).[14] We also explored functional trees (FTs), a promising class of

Table 1. Demographics and Clinical Characteristics of Individuals Included in the Training and Validation Cohorts

	Class	n (Female: Male)	Age (Years)	Expanded Disability Status Scale (EDSS)	Disease Duration (Years)
Training set N = 224	CIS	48 (25:23)	34.2 ± 9.7	1.7 ± 1.0	–
	RR MS	73 (44:29)	38.4 ± 9.3**	1.9 ± 1.2	6.2 ± 6.7
	SP MS	18 (10:8)	53.2 ± 13***	6.8 ± 1.2	25 ± 10.4
	PP MS	25 (16:9)	53.1 ± 11.7***	4.8 ± 1.9	13.9 ± 8.7
	HC	42 (23:19)	32.9 ± 10.2	–	–
	OND	18 (7:11)	43.6 ± 14**	–	–
Independent test set N = 89	CIS	9 (4:5)	34.1 ± 10.7	1.9 ± 0.7	–
	RR MS	35 (22:13)	38.7 ± 10.8**	1.4 ± 0.6	5.2 ± 5
	SP MS	8 (4:4)	48.5 ± 8.8***	6.2 ± 1.5	21.5 ± 4.5
	PP MS	10 (5:5)	49.6 ± 7.3***	6.0 ± 1.6	7.1 ± 4.8
	HC	18 (10:8)	30.8 ± 7.4	–	–
	OND	9 (2:7)	44.0 ± 11**	–	–

The EDSS value (from 0 to 10) indicates disability status. Age, EDSS, and disease duration are given as mean ± SD. Asterisks indicate statistical significance versus HC (**p < 0.01, ***p < 0.001).

decision trees integrated with logistic models,[15,16] and then applied adaptive boosting to FT (ADAboost-FT) because boosting techniques can enhance classification performance in imbalanced datasets.[17,18] Overall, our strategy demonstrated predictive power with an accuracy above 90% for most of the classification tasks.

RESULTS

Integrated PBMC Transcriptomics Datasets Were Coherent in the Training and Validation Cohorts

To construct optimal gene expression-based classifiers for prediction of MS state and stage, we collected genome-wide transcriptomics profiles of PBMCs derived from 313 individuals (60 HCs, 57 CIS subjects, 108 RR MS subjects, 26 SP MS subjects, 35 PP MS subjects, and 27 OND subjects) and generated with two distinct microarrays (HumanRef-8 v.2 and HumanHT-12 v.4). After batch effect detection and correction procedures (Figure S1), we assessed dataset integrity through clustering analysis of technical and biological replicates. We observed a marked improvement in the number of correctly clustering pairs, passing from 5 to 15 (of 17) after batch correction (Figures S2A and S2B). Similarly, biological replicates tended to pair more reproducibly after correction (Figures S2C and S2D). Principal-component analysis (PCA) of the final global dataset did not show proper separation of subjects based on demographic factors such as age and gender (Figure S3). This final integrated dataset was then divided into the training and independent test sets according to specific clinical and demographic criteria (Table 1). PCA of genome-wide transcriptomics data did not show any clear separation of the different classes (CIS, MS stages, and controls) in the first two principal component (PC) dimensions in the training and independent test sets (Figures 1A and 1C). Nevertheless, a detailed inspection of the training cohort using 3D PCA revealed a coherent clustering structure (Figure 1B) where the CIS centroid was close to RR MS and the centroids for the two progressive forms (SP MS and PP MS) clustered together. This structure was reproduced in the independent test set (Figure 1D). Overall, these results confirmed the robustness of the applied procedure in generating an integrated transcriptomics dataset suitable for MS classifier development.

NCV in the Training Cohort Led to Unbiased Evaluation of Classification Algorithms

To obtain optimal classifiers for MS stages from PBMC transcriptomes, we designed an unbiased framework based on NCV comparing 3 distinct machine learning algorithms—FT, RF, and ADAboost-FT—without use of the independent test set (Figure 2A).

The first test aimed to build a descriptive model differentiating MS (RR+SP+PP) from non-MS (HC+OND) subjects. Precision-recall curves and relative area under the precision-recall curve (AUPRC) indicated a comparably high classification performance of the three algorithms in this task (Figure 3A).

The second test investigated whether the CIS transcriptome carried early-stage-specific alterations that could discriminate CIS from HC, MS, and OND subjects. Here the high degree of class imbalance (48 CIS, 176 “all”), combined with the high heterogeneity of the “all” class, resulted in poor classification performance, especially with the RF algorithm (Figure 3B).

The final three tests investigated the performance of the workflow in distinct classification of PP and/or SP MS from RR MS (Figures 3C–3E). Here, despite the small sample size for progressive cases and, thus, class imbalance with respect to RR MS, classifiers generated reasonably accurate models, and comparison of the AUPRC evidenced a more stable performance of FT-based algorithms than the commonly used RF-based algorithm.

Overall, these NCV experiments on the training dataset confirmed that blood transcriptomes deliver useful information for predictive classification models, whose performance may depend on the algorithm used, the number of samples in each class, and class balance.

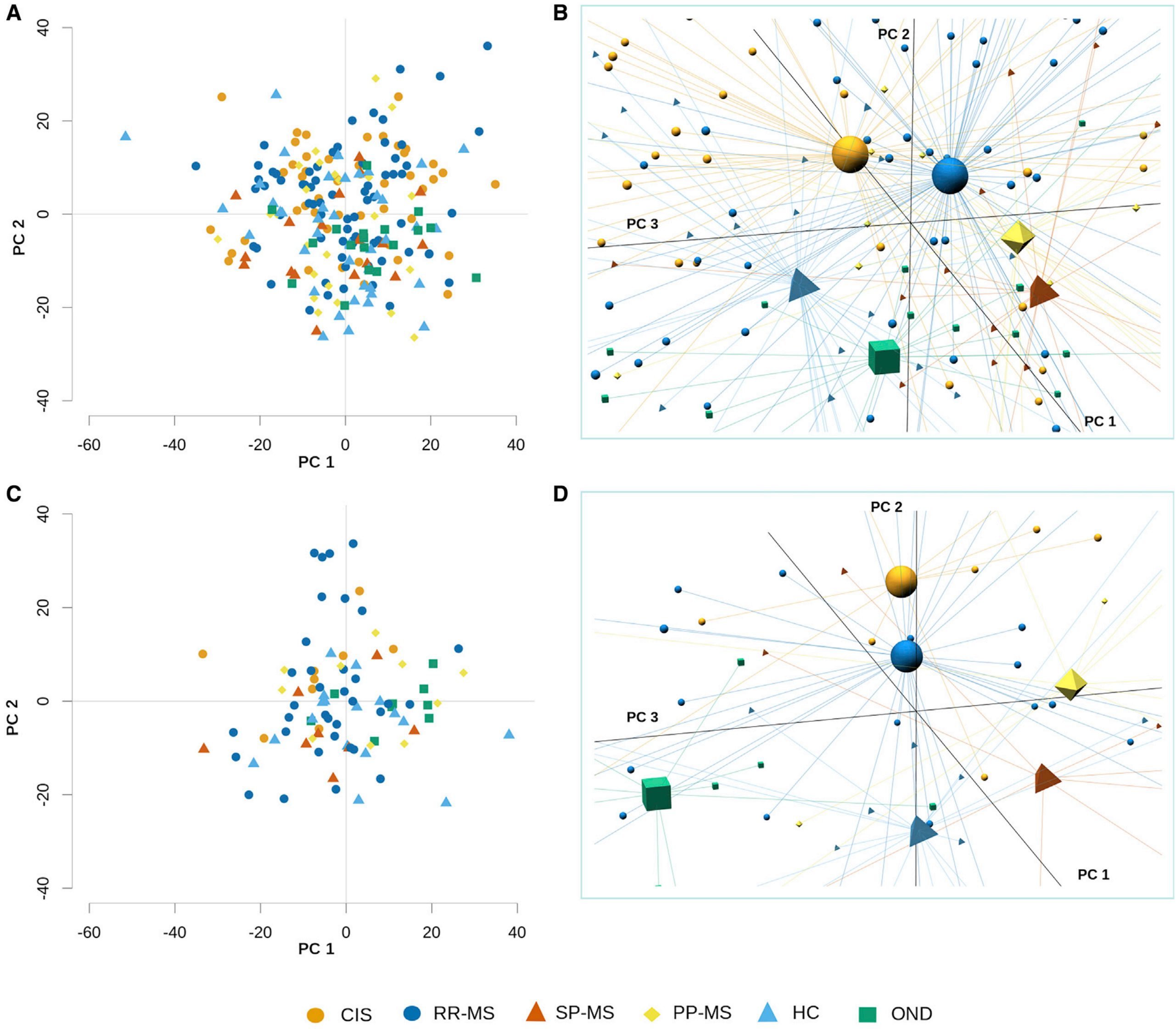


Figure 1. Integrated PBMC Transcriptomics Datasets Were Coherent in the Training and Validation Cohorts

PCA plots of PBMC genome-wide transcriptomes included in the study are shown for the training (top panels) and validation (bottom panels) cohorts.

(A) Distribution of the classes used for classifier development (CIS, clinically defined MS stages, and HC and OND controls) among the first two principal components (PCs) in the training dataset.

(B) Centroids of 3D PC coordinates showing the relative distribution of the different classes in the training dataset.

(C) Distribution of classes among the first two PCs in the independent test set.

(D) Centroids of 3D PC coordinates showing the relative distribution of the different classes in the independent test set.

MS Classifiers Built on the Training Datasets Demonstrated High Accuracy in the Independent Validation Cohorts

For construction and validation of the final classifiers, we retrained and optimized each algorithm on the complete training set and applied each classifier to the independent test set. The workflow used for final validation is shown in Figure 2B. Precision-recall curves and relative AUPRCs of the different algorithms in the independent test set confirmed the NCV results, demonstrating the superior performance of the ADAboost-FT algorithm with respect to single FT and RF in each classification task (Figures 4A, 4B, and 5A–5C).

In the MS versus non-MS classification task, ADAboost-FT generated a predictive model based on 139 probes (Table S2A), which showed 94.3% sensitivity (recall) and 87.5% precision (Figure 4C). Similar to the NCV results, the CIS classifier (213 probes; Table S2B) showed sub-optimal performance in sensitivity and precision but reached an overall accuracy of 89.9% (Figure 4D). Probes classifying MS from non-MS conditions were mostly distinct from those identifying the CIS stage

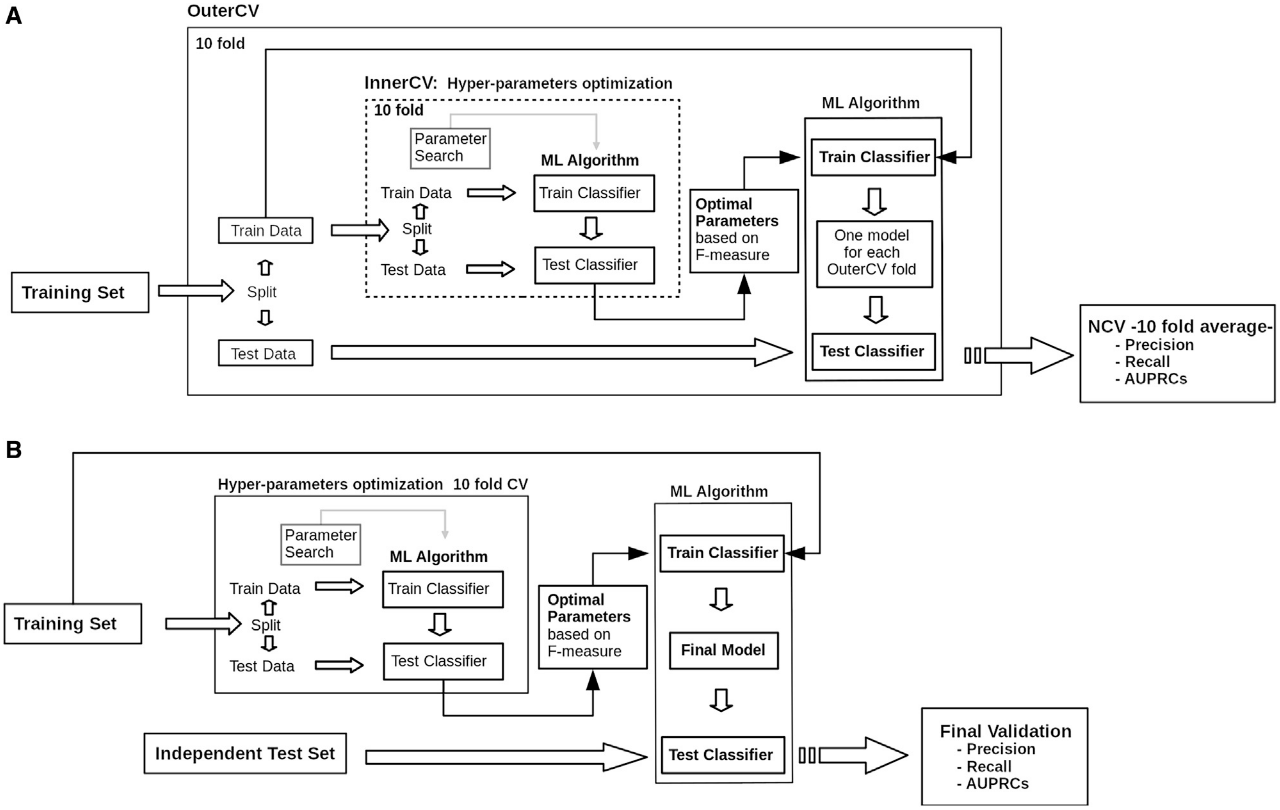


Figure 2. Machine Learning Pipeline and Relative Data Flow

(A) Nested cross-validation (NCV) for algorithm optimization and comparison. At each fold of the outer CV, the whole training data are split in two sub-sampled datasets, one for training and the other for testing. The sub-sampled training set enters the hyper-parameter optimization loop, where it is further split into training and test partitions following a second 10-fold CV (inner CV), where different combinations of hyper-parameters are evaluated (one 10-fold CV for each combination). The combination of hyper-parameters that maximize the F-measure (harmonic mean of precision and recall) of the class of interest is retained as optimal, applied to the algorithm, and tested on the test set of the corresponding outer CV fold. The entire procedure is repeated 10 times, and the averaged performance is collected at the end of the outer CV loop and used for algorithm comparison.

(B) Final optimization and validation of the selected algorithm. The whole training dataset enters the hyper-parameter optimization loop, where it is subjected to an identical search for the combination of hyper-parameters that maximize the F-measure. The optimal hyper-parameters are applied to the algorithm, which is trained on the whole training set and tested on the independent test set for final validation.

(Figure 4E). Interestingly, functional enrichment evidenced transcripts playing a role in common processes, such as regulation of transcription and interferon signaling, with the MS profile presenting additional themes about chromatin remodeling and apoptosis and the CIS signature about cell proliferation and chemotaxis (Figures 4F and 4G; Tables S3A and S3B).

Because significant differences in age distribution were observed between single disease classes and the healthy population (Table 1) and between classes used for construction of each classifier (Table S1), we verified whether classifiers were driven by age. Despite no significant differences in gender ratios being present, we included gender in the analyses as an additional control. PCA plots showed that MS and CIS predictive signatures failed to separate subjects on the basis of gender (Figures S4A and S4B) or age (Figures S4C and S4D). On the other hand, both classifiers generated a clear separation of subjects according to condition (Figures S4E and S4F), with superior performance of the MS signature, as expected from the NCV results.

The progressive MS classifier consisted of 222 probes (Table S2C) that showed 83.3% sensitivity and 93.8% precision in differentiating progressive from RR forms (Figure 5D). The PP MS classifier (266 probes; Table S2D) differentiated PP MS from RR MS with 90% sensitivity and precision (Figure 5E). Finally, the SP MS versus the RR MS classifier (201 probes; Table S2E) showed the highest performance, with 87.5% sensitivity and 100% precision, reaching an overall accuracy of 97.7% (Figure 5F). As expected, the progressive classifier contained probes distinct from those classifying the whole MS condition (Figure 5G). Further, the predictive probes for the single or combined progressive stages showed limited, partial overlap, indicating that classifiers selected specificities for the PP versus SP stages of disease (Figure 5I). Significantly enriched biological themes included cell cycle and T cell activation for the combined (PP+SP) MS signature; protein ubiquitination, cell migration, and fatty acid metabolism for the PP MS profile; and regulation of GTPase activity, locomotor behavior, and blood coagulation in the SP MS signature (Figures 5H, 5J, and 5K; Tables S3C–S3E).

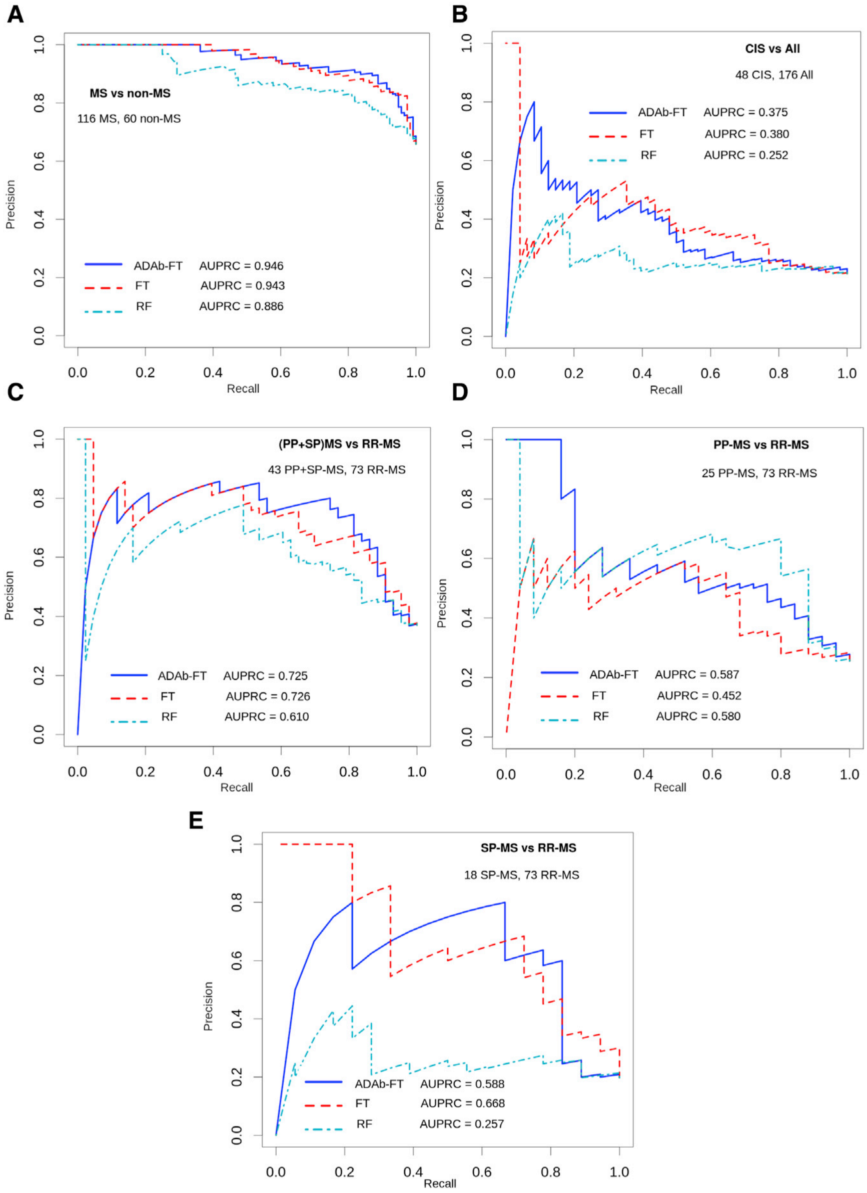


Figure 3. NCV of the Training Cohort Led to Unbiased Evaluation of Classification Algorithms

(A–E) For each of the 5 classification tasks (A, MS vs. non-MS; B, CIS vs. All; C, (PP+SP)MS vs. RR-MS; D, PP-MS vs. RR-MS; E, SP-MS vs. RR-MS), comparison of the performance obtained by the ADAboost-FT, FT, and RF algorithms in the NCV is shown as precision-recall (PR) curves and relative areas under the curves (AUPRCs).

PCA plots based on the progressive signatures did not evidence any major effect of gender (Figures S5A, S5B, and S6A) but showed modest separation according to age (Figures S5C, S5D, and S6B), as expected from the large difference in age distribution between RR MS and progressive MS (Table S1). This age-related effect disappeared almost completely when the progressive signatures were applied to all subjects, including CIS and controls (Figures S5E, S5F, and S6C). These results indicate that the initial age-dependent separation by the progressive classifiers was mainly caused by the large superposition of progressive MS classes with the older age group rather than by a separate effect of age. Most importantly, PCA plots based on progressive MS signatures clearly discriminated between the progressive and the RR forms of MS (Figures S5G, S5H, and S6D).

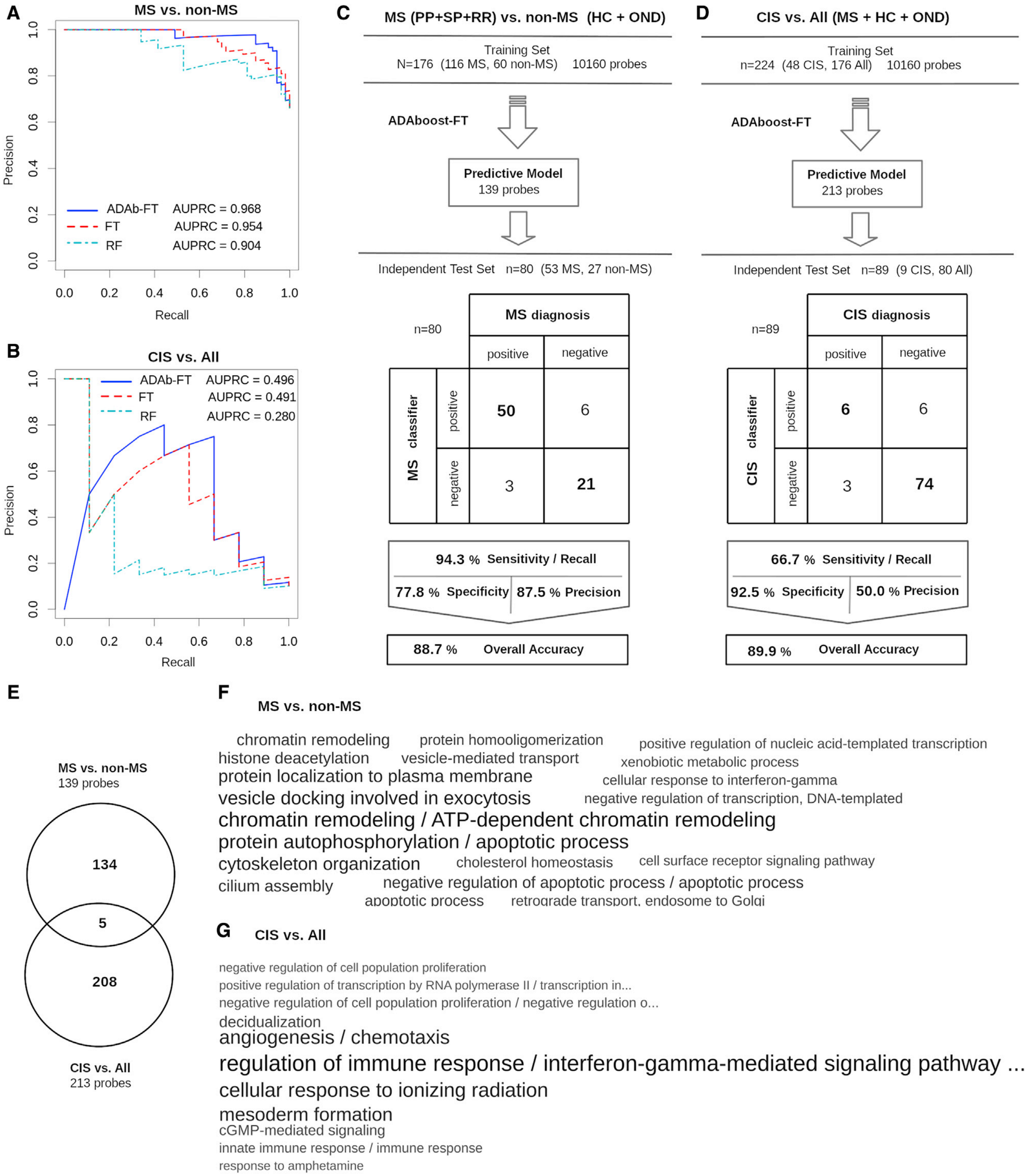
Altogether, these results confirmed that machine learning classifiers detect transcriptomic variations specifically linked to MS disease state and stage but not to demographic factors.

DISCUSSION

Here we described a machine learning workflow based on NCV for unbiased fine-tuning and comparison of different algorithms that led to selection of MS stage-specific PBMC signatures with high predictive power directly from genome-wide transcriptomics.

Although the majority of machine learning applications in MS research focus on classification based on MRI data, alone[5,6,8,19] or in combination with clinical data,[7] a few examples show promising results when considering biological dimensions, such as the cerebrospinal fluid (CSF) proteome[20] or blood metabolome.[21] Here we hypothesized that biological information derived from PBMC transcriptomes could facilitate disease classification. We oriented our choice toward PBMCs because our priority was to develop classifiers that are easy to apply in clinical settings while providing a global view of MS-related gene expression changes in all PBMCs. Although several studies have explored gene expression patterns from blood in MS using traditional statistical analyses,[9–11] only a couple of reports have attempted to apply machine learning to blood transcriptomics and were limited to discrimination between the RR MS form and controls.[22] In this context, our study, which also included PBMC transcriptomes from the progressive forms of MS, responds to the unmet clinical need of potential predictive biomarkers for distinct MS courses.

Gene expression datasets pose a great challenge to classification algorithms because high dimensionality may lead to models overfitting the training data and with poor performance on new "unseen" samples. Moreover, MS datasets are subjected to class imbalance, adding further complexity to the learning process. To select the optimal classifier, we developed an NCV workflow performing unbiased fine-tuning and



n=80		MS diagnosis positive	MS diagnosis negative
MS classifier	positive	50	6
	negative	3	21

94.3 % Sensitivity / Recall
77.8 % Specificity 87.5 % Precision
88.7 % Overall Accuracy

n=89		CIS diagnosis positive	CIS diagnosis negative
CIS classifier	positive	6	6
	negative	3	74

66.7 % Sensitivity / Recall
92.5 % Specificity 50.0 % Precision
89.9 % Overall Accuracy

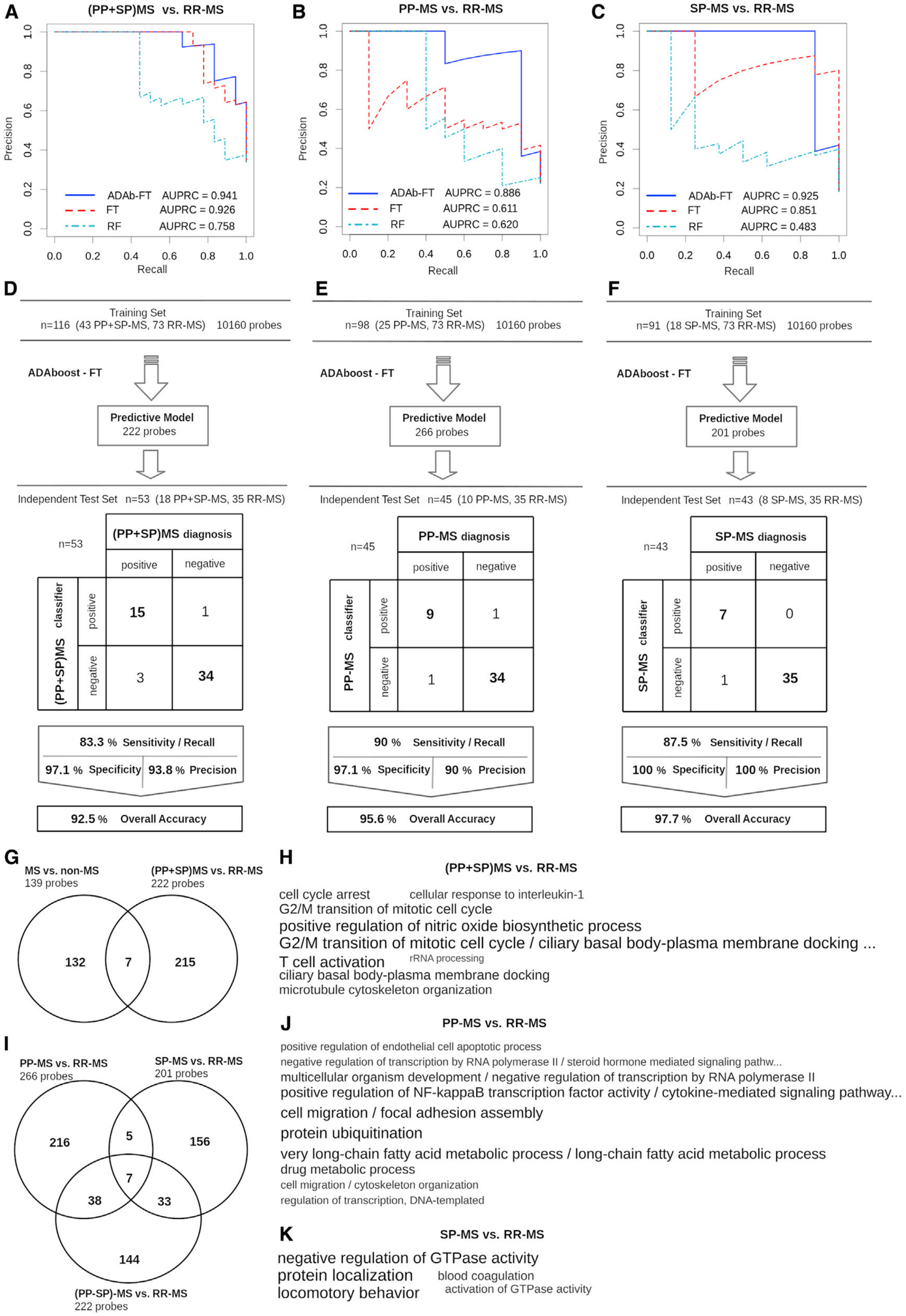
Figure 4. Construction, Test, and Biological Contents of MS versus Non-MS and CIS versus "All" Classifiers

(A and B) Comparison of the performance of the different algorithms on the independent test set for MS versus non-MS (A) and CIS versus "all" (B) classifiers.

(C and D) ADAboost-FT models generated on the complete training set and classification results on the independent test set for the MS vs. non-MS (C) and CIS vs. All (D) classifiers.

(E) Venn diagram of probes identified by the two classifiers.

(F and G) TagCloud of enriched gene ontology (GO) biological processes in MS versus non-MS (F) and CIS versus All (G) classifiers. The height and width of TagCloud terms are proportional to the level of significance.



n=53		(PP+SP)MS diagnosis positive	(PP+SP)MS diagnosis negative
(PP+SP)MS classifier	positive	15	1
	negative	3	34

83.3 % Sensitivity / Recall
97.1 % Specificity | 93.8 % Precision
92.5 % Overall Accuracy

n=45		PP-MS diagnosis positive	PP-MS diagnosis negative
PP-MS classifier	positive	9	1
	negative	1	34

90 % Sensitivity / Recall
97.1 % Specificity | 90 % Precision
95.6 % Overall Accuracy

n=43		SP-MS diagnosis positive	SP-MS diagnosis negative
SP-MS classifier	positive	7	0
	negative	1	35

87.5 % Sensitivity / Recall
100 % Specificity | 100 % Precision
97.7 % Overall Accuracy

G
MS vs. non-MS
139 probes
(PP+SP)MS vs. RR-MS
222 probes
132
7
215
H
(PP+SP)MS vs. RR-MS
cell cycle arrest
cellular response to interleukin-1
G2/M transition of mitotic cell cycle
positive regulation of nitric oxide biosynthetic process
G2/M transition of mitotic cell cycle / ciliary basal body-plasma membrane docking ...
T cell activation
rRNA processing
ciliary basal body-plasma membrane docking
microtubule cytoskeleton organization
I
PP-MS vs. RR-MS
266 probes
SP-MS vs. RR-MS
201 probes
216
5
156
7
38
33
144
(PP-SP)-MS vs. RR-MS
222 probes
J
PP-MS vs. RR-MS
positive regulation of endothelial cell apoptotic process
negative regulation of transcription by RNA polymerase II / steroid hormone mediated signaling pathw...
multicellular organism development / negative regulation of transcription by RNA polymerase II
positive regulation of NF-kappaB transcription factor activity / cytokine-mediated signaling pathway...
cell migration / focal adhesion assembly
protein ubiquitination
very long-chain fatty acid metabolic process / long-chain fatty acid metabolic process
drug metabolic process
cell migration / cytoskeleton organization
regulation of transcription, DNA-templated
K
SP-MS vs. RR-MS
negative regulation of GTPase activity
protein localization
blood coagulation
locomotory behavior
activation of GTPase activity

(legend on next page)

comparison of three algorithms based on a decision tree paradigm. The NCV results were confirmed by validation on the independent test set, where FT-based algorithms showed superior performance to the largely employed RF. In particular, ADAboost-FT displayed the best predictive power in all classification tasks, in spite of class imbalance. The weaker classification power of the algorithms in NCV experiments compared with the validation data may be related to the pessimistic bias that can occur in NCV (and cross-validation [CV] in general), where classifiers are trained on sub-samples of the training data, and that is more evident in imbalanced datasets.[23] For this reason, it is expected that, if overfitting is successfully avoided, then classifiers display better performance when trained on the complete training set and tested on the validation cohort. Classifiers using deep learning applied to MRI showed impressive performance (99% accuracy) in discriminating MS subjects from HCs in recent studies.[5,6] However, none of them included an OND cohort in the control group, limiting classifier ability in differential diagnosis among distinct neurological conditions. Further, although employing valid techniques to limit overfitting and estimate classification performance, these reports did not perform any validation on independent cohorts. Our MS versus non-MS classifier identified a signature capable of discriminating the MS condition from HC and OND subjects with high sensitivity and precision. Functional enrichment showed that processes involved in chromatin remodeling and apoptosis may have major roles in definition of the MS state. Recent work reported epigenetic reprogramming during MS.[24,25] Moreover, dysregulation of apoptotic genes has been described in PBMCs of MS patients.[26,27]

Because of the high degree of intra-class variability combined with class imbalance, CIS versus “all” represented the most challenging prediction task, and although the classifier demonstrated some improvement when trained on the whole training set and tested on the validation cohort, the performance remained considerably lower than that of others, indicating the lack of a strong signature for the very early phase of disease. Although not optimal, the resulting classifier achieved a performance close to MRI-based classifiers comparing CIS with RR MS.[8,19] Interestingly, the CIS signature showed significant enrichment for transcripts involved in interferon signaling, confirming previous findings of dysregulation of interferon-related transcripts throughout all MS stages, including CIS,[11] possibly providing insights regarding the benefits observed with early interferon beta treatment of CIS patients.[28]

The progressive MS versus RR MS classifier showed highly accurate prediction results and was able to extract a transcriptional signature specific for the progressive (PP+SP) MS stage. Interestingly, Barbour et al.[20] developed an RF classifier based on CSF proteins, obtaining a classification performance similar to our results in MS versus non-MS and progressive MS versus RR MS. However, our model has the advantage of analyzing peripheral blood, which is easier and less invasive to collect and resample than CSF. The progressive MS signature was distinct from the more general MS signature derived by the MS versus the non-MS classifier and showed enrichment for specific processes such as T cell activation and nitric oxide biosynthesis. The activation and autoproliferation of brain-homing autoreactive T cells has been described recently for PBMCs of RR MS patients,[29] but the potential role of this mechanism in progressive MS has not yet been investigated. Moreover, nitric oxide is known to be involved in MS, and its metabolites are generally increased in serum, CSF, and urine of MS patients.[30] Because PP MS and RR MS share many pathological findings at presentation, and the correct diagnosis of PP MS is often delayed,[2,3] we trained PP MS versus the RR MS model to identify stage-specific transcripts that could facilitate early discrimination between the two courses of disease. Our PP MS versus RR MS classifier showed impressive performance in the independent test set, reaching 90% of precision and recall with an overall accuracy of 95.6%. The SPMS versus RR MS classifier was built to generate reliable predictive models of disease progression from the RR to the SP stage. The extremely high precision (100%) reached by our SP MS model in the independent test set indicates that the classifier may help to more precisely identify the transition from RR MS to SP MS, which often represents a clinical challenge.[4] By taking into account the effects of class imbalances, precision and recall are better evaluation metrics than accuracy when focusing on small positive classes. For example, in our PP MS versus RR MS classifier, where there is a majority of negative samples (RR MS) compared with the positive class (PP MS), misclassification of one real PP MS as RR MS (false negative) and one RR MS as PP MS (false positive) lowered the precision and recall values from 100% to 90%, with little effect on overall accuracy. Similarly, in the SP MS versus RR MS classifier, where the positive class is even smaller, the only one false negative dropped the recall value to 87.5%. Because classifier hyperparameters were optimized on the F-measure, which is the harmonic mean between precision and recall, the misclassification rates in the independent test set reflect the best compromise between precision and recall and represent the true limits of our classifiers. Thus, the high degree of precision and recall shown by PP MS and SP MS classifiers in the independent validation cohort make them suitable for application in clinical practice and may offer clinicians a rationale for appropriate treatment planning or switching.

Figure 5. Construction, Test, and Biological Contents of Progressive MS versus RR MS Classifiers

(A–C) Comparison of the performance of the different algorithms on the independent test set for progressive MS versus RR MS (A), PP MS versus RR MS (B), and SP MS versus RR MS (C) classifiers.

(D–F) ADAboost-FT models generated on the complete training set and classification results on the independent test set for the (PP+SP)MS vs. RR-MS (D), PP-MS vs. RR-MS (E), SP-MS vs. RR-MS (F) classifiers.

(G) Venn diagram of probes identified by MS versus non-MS and progressive versus RR MS classifiers.

(H) TagCloud of enriched GO biological processes in the progressive MS versus RR MS classifier.

(I) Venn diagram of probes identified by progressive MS versus RR MS, PP MS versus RR MS, and SP MS versus RR MS classifiers.

(J and K) TagClouds of enriched GO biological processes in PP MS versus RR MS (J) or SP MS versus RR MS (K) classifiers. In (J), only the top 10 terms are shown. The height and width of TagCloud terms are proportional to the level of significance.

Altogether, these results were superior to PP MS and SP MS classifiers based on MRI connectivity data[8] and to classifiers that used combinations of clinical data and MRI metabolic features.[7]

The rate of overlap among transcripts identified by our workflow in the three progressive MS prediction tasks was low, indicating that classifiers were able to capture distinctive signatures. Although (PP+SP) MS versus RR MS preferentially captured commonalities between the two progressive forms, PP MS versus RR MS and SP MS versus RR MS tasks were able to identify transcripts specific to each progressive stage. Distinction of PP MS from SP MS is highly debated in the literature, with several studies suggesting complete biological overlap between the two forms when analyzing MRI imaging,[31] neuropathology,[32] or the CSF proteome.[20] On the other hand, significant differences have been described in the blood gene expression profiles of the two forms.[10,12] The highly specific signatures derived by PP MS versus RR MS and SP MS versus RR MS classifiers evidenced enrichment for distinctive processes that deserve further investigation. For example, our results indicate that dysregulation of processes related to protein ubiquitination may play a central role in PP MS but not in SP MS. Interestingly, the ubiquitin-proteasome system is tightly related to antigen presentation and has been linked previously to neuro-inflammatory processes and MS pathogenesis.[33] Similarly, the SP MS signature showed enriched classes distinct from PP MS, including blood coagulation. Interestingly, altered plasma levels of specific coagulation factors have been recently described in SP MS but not in PP MS patients in comparison with healthy controls.[34]

Classifier results were largely independent of age and gender factors and therefore specifically linked to blood transcriptomics dysregulations underpinning MS state and stage. However, we cannot exclude that age- or gender-related MS signatures can be obtained by appropriately orienting the experimental design, as already demonstrated for gender in the RR form of disease by our group.[25] The main limitation of the present study is the relatively low number of progressive MS samples. However, this limitation is compensated for by the high number of RR MS samples, which enhance the level of significance of classifier results. Nevertheless, this imbalance between progressive and RR cases mimics clinical scenarios, offering more realistic training and test sets for MS classifier development. The current study can thus be considered an advanced pilot study that can be extended to larger cohorts, new constantly evolving algorithms (i.e., deep learning), and more complex inquires (i.e., analysis of different subsets of immune cells) in future investigations.

In conclusion, the present study justifies application of artificial intelligence methods to blood transcriptomes and offers a robust pipeline for generation of classifiers supporting MS diagnosis and prognosis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Human subjects
- METHOD DETAILS
  - ○ PBMC isolation and RNA extraction
  - ○ Microarray experiment and data processing
  - ○ Machine learning pipeline
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Microarray data processing
  - ○ Machine learning pipeline
  - ○ Functional annotation

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.xcrm.2020.100053.

ACKNOWLEDGMENTS

We are grateful to Enzo Medico and Daniela Cantarella for support regarding the microarray experiments, all participants who donated blood for this study, and Amici Centro Sclerosi Multipla (ACESM). This study was funded by Merck-Serono and the Italian Ministry of Health (RF-2011-02349698).

AUTHOR CONTRIBUTIONS

M.A. conducted transcriptomics and statistical analyses, designed the machine learning pipeline, and wrote the paper. R.M. conducted the preliminary analysis on dataset integration and classifier development. M.D.D. prepared PBMCs and extracted and processed RNA for microarray experiments. G.D.C., M.R., F.S., B.C., L.M., and V.M. enrolled the patients in the study and provided clinical information. G.C. critically discussed the project and the results. C.F. conceived and designed the experiments, coordinated the study, discussed the results, and wrote the paper.

DECLARATION OF INTERESTS

F.S. and L.M. report personal fees from Merck Serono, Novartis, Biogen, Sanofi-Genzyme, Teva, and Roche outside of the submitted work. M.R. reports personal fees from Merck-Serono and Genzyme and non-financial support from Novartis, Teva Pharmaceutical Industries, and Almirall outside of the submitted work. V.M. reports honoraria for speaking and/or for consultancy and support for travel expenses and participation in congresses from Biogen, Merck, Novartis, Roche, Genzyme, Almirall, and Teva. G.C. reports personal fees from Novartis, Teva Pharmaceutical Industries Ltd., Teva Italia Srl, Sanofi Genzyme, Genzyme Corporation, Genzyme Europe, Merck KGaA, Merck Serono SpA, Celgene Group, Biogen Idec, Biogen Italia Srl, F. Hoffman-La Roche, Roche SpA, Almirall SpA, Forward Pharma, Medday, and Excemed outside of the submitted work. C.F. reports grants from Merck-Serono and the Italian Ministry of Health during the conduct of the study and grants from Teva, Novartis, and Fondazione Italiana Sclerosi Multipla outside of the submitted work.

Received: December 13, 2019
Revised: May 5, 2020
Accepted: June 22, 2020
Published: July 21, 2020

REFERENCES

1. Miller, D.H., Chard, D.T., and Ciccarelli, O. (2012). Clinically isolated syndromes. Lancet Neurol. *11*, 157–169.

2. Solomon, A.J., Naismith, R.T., and Cross, A.H. (2019). Misdiagnosis of multiple sclerosis: Impact of the 2017 McDonald criteria on clinical practice. Neurology *92*, 26–33.

3. Rice, C.M., Cottrell, D., Wilkins, A., and Scolding, N.J. (2013). Primary progressive multiple sclerosis: progress and challenges. J. Neurol. Neurosurg. Psychiatry *84*, 1100–1106.

4. Katz Sand, I., Krieger, S., Farrell, C., and Miller, A.E. (2014). Diagnostic uncertainty during the transition to secondary progressive multiple sclerosis. Mult. Scler. *20*, 1654–1657.

5. Marzullo, A., Kocevar, G., Stamile, C., Durand-Dubief, F., Terracina, G., Calimeri, F., and Sappey-Marinier, D. (2019). Classification of Multiple Sclerosis Clinical Profiles via Graph Convolutional Neural Networks. Front. Neurosci. *13*, 594.

6. Wang, S.H., Tang, C., Sun, J., Yang, J., Huang, C., Phillips, P., and Zhang, Y.D. (2018). Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. Front. Neurosci. *12*, 818.

7. Ion-Mărgineanu, A., Kocevar, G., Stamile, C., Sima, D.M., Durand-Dubief, F., Van Huffel, S., and Sappey-Marinier, D. (2017). Machine Learning Approach for Classifying Multiple Sclerosis Courses by Combining Clinical Data with Lesion Loads and Magnetic Resonance Metabolic Features. Front. Neurosci. *11*, 398.

8. Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., and Sappey-Marinier, D. (2016). Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses. Front. Neurosci. *10*, 478.

9. Koch, M.W., Ilnytskyy, Y., Golubov, A., Metz, L.M., Yong, V.W., and Kovalchuk, O. (2018). Global transcriptome profiling of mild relapsing-remitting versus primary progressive multiple sclerosis. Eur. J. Neurol. *25*, 651–658.

10. Gandhi, K.S., McKay, F.C., Cox, M., Riveros, C., Armstrong, N., Heard, R.N., Vucic, S., Williams, D.W., Stankovich, J., Brown, M., et al.; ANZgene Multiple Sclerosis Genetics Consortium (2010). The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. Hum. Mol. Genet. *19*, 2134–2143.

11. Srinivasan, S., Severa, M., Rizzo, F., Menon, R., Brini, E., Mechelli, R., Martinelli, V., Hertzog, P., Salvetti, M., Furlan, R., et al. (2017). Transcriptional dysregulation of Interferome in experimental and human Multiple Sclerosis. Sci. Rep. *7*, 8981.

12. Srinivasan, S., Di Dario, M., Russo, A., Menon, R., Brini, E., Romeo, M., Sangalli, F., Costa, G.D., Rodegher, M., Radaelli, M., et al. (2017). Dysregulation of MS risk genes and pathways at distinct stages of disease. Neurol. Neuroimmunol. Neuroinflamm. *4*, e337.

13. Ambroise, C., and McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA *99*, 6562–6566.

14. Hastie, T., Tibshirani, R., and Friedman, J. (2009). Random Forests. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York, NY: Springer New York), pp. 587–604.

15. Gama, J. (2004). Functional Trees. Mach. Learning *55*, 219–250.

16. Landwehr, N., Hall, M., and Frank, E. (2005). Logistic Model Trees. Mach. Learning *59*, 161–205.

17. Seiffert, C., Khoshgoftaar, T., Van Hulse, J., and Napolitano, A. (2008). Building Useful Models from Imbalanced Data with Sampling and Boosting. AAAI Press. In Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference, FLAIRS-21, pp. 306–311.

18. Hastie, T., Friedman, J., and Tibshirani, R. (2001). Boosting and Additive Trees. In The Elements of Statistical Learning: Data Mining, Inference, and Prediction, T. Hastie, J. Friedman, and R. Tibshirani, eds. (New York, NY: Springer New York), pp. 299–345.

19. Muthuraman, M., Fleischer, V., Kolber, P., Luessi, F., Zipp, F., and Groppa, S. (2016). Structural Brain Network Characteristics Can Differentiate CIS from Early RRMS. Front. Neurosci. *10*, 14.

20. Barbour, C., Kosa, P., Komori, M., Tanigawa, M., Masvekar, R., Wu, T., Johnson, K., Douvaras, P., Fossati, V., Herbst, R., et al. (2017). Molecular-based diagnosis of multiple sclerosis and its progressive stage. Ann. Neurol. *82*, 795–812.

21. Andersen, S.L., Briggs, F.B.S., Winnike, J.H., Natanzon, Y., Maichle, S., Knagge, K.J., Newby, L.K., and Gregory, S.G. (2019). Metabolome-based signature of disease pathology in MS. Mult. Scler. Relat. Disord. *31*, 12–21.

22. Gurevich, M., Miron, G., and Achiron, A. (2015). Optimizing multiple sclerosis diagnosis: gene expression and genomic association. Ann. Clin. Transl. Neurol. *2*, 271–277.

23. Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv, arXiv:1811:12808. https://arxiv.org/abs/1811.12808.

24. Kular, L., Liu, Y., Ruhrmann, S., Zheleznyakova, G., Marabita, F., Gomez-Cabrero, D., James, T., Ewing, E., Linden, M., Gornikiewicz, B., et al. (2018). DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. Nat. Commun. *9*, 2397.

25. Menon, R., Di Dario, M., Cordiglieri, C., Musio, S., La Mantia, L., Milanese, C., Di Stefano, A.L., Crabbio, M., Franciotta, D., Bergamaschi, R., et al. (2012). Gender-based blood transcriptomes and interactomes in multiple sclerosis: involvement of SP1 dependent gene transcription. J. Autoimmun. *38*, J144–J155.

26. Gomes, A.C., Jönsson, G., Mjörnheim, S., Olsson, T., Hillert, J., and Grandien, A. (2003). Upregulation of the apoptosis regulators cFLIP, CD95 and CD95 ligand in peripheral blood mononuclear cells in relapsing-remitting multiple sclerosis. J. Neuroimmunol. *135*, 126–134.

27. Severa, M., Rizzo, F., Srinivasan, S., Di Dario, M., Giacomini, E., Buscarinu, M.C., Cruciani, M., Etna, M.P., Sandini, S., Mechelli, R., et al. (2019). A cell type-specific transcriptomic approach to map B cell and monocyte type I interferon-linked pathogenic signatures in Multiple Sclerosis. J. Autoimmun. *101*, 1–16.

28. Edan, G., Kappos, L., Montalbán, X., Polman, C.H., Freedman, M.S., Hartung, H.P., Miller, D., Barkhof, F., Herrmann, J., Lanius, V., et al.; BENEFIT Study Group (2014). Long-term impact of interferon beta-1b in patients with CIS: 8-year follow-up of BENEFIT. J. Neurol. Neurosurg. Psychiatry *85*, 1183–1189.

29. Jelcic, I., Al Nimer, F., Wang, J., Lentsch, V., Planas, R., Jelcic, I., Madjovski, A., Ruhrmann, S., Faigle, W., Frauenknecht, K., et al. (2018). Memory B Cells Activate Brain-Homing, Autoreactive CD4^{+} T Cells in Multiple Sclerosis. Cell *175*, 85–100.e23.

30. Smith, K.J., and Lassmann, H. (2002). The role of nitric oxide in multiple sclerosis. Lancet Neurol. *1*, 232–241.

31. Kuchling, J., Ramien, C., Bozin, I., Dörr, J., Harms, L., Rosche, B., Niendorf, T., Paul, F., Sinnecker, T., and Wuerfel, J. (2014). Identical lesion morphology in primary progressive and relapsing-remitting MS–an ultrahigh field MRI study. Mult. Scler. *20*, 1866–1871.

32. Frischer, J.M., Weigand, S.D., Guo, Y., Kale, N., Parisi, J.E., Pirko, I., Mandrekar, J., Bramow, S., Metz, I., Brück, W., et al. (2015). Clinical and pathological insights into the dynamic nature of the white matter multiple sclerosis plaque. Ann. Neurol. *78*, 710–721.

33. Limanaqi, F., Biagioni, F., Gaglione, A., Busceti, C.L., and Fornai, F. (2019). A Sentinel in the Crosstalk Between the Nervous and Immune System: The (Immuno)-Proteasome. Front. Immunol. *10*, 628.

34. Göbel, K., Kraft, P., Pankratz, S., Gross, C.C., Korsukewitz, C., Kwiecien, R., Mesters, R., Kehrel, B.E., Wiendl, H., Kleinschnitz, C., and Meuth, S.G. (2016). Prothrombin and factor X are elevated in multiple sclerosis patients. Ann. Neurol. *80*, 946–951.

35. McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., et al. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann. Neurol. *50*, 121–127.

36. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1137–1143.

37. Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., and Napolitano, A. (2008). Resampling or Reweighting: A Comparison of Boosting Implementations. In 2008 20th IEEE International Conference on Tools with Artificial Intelligence, pp. 445–451.

38. Sumner, M., Frank, E., and Hall, M. (2005). Speeding Up Logistic Model Tree Induction. In Knowledge Discovery in Databases: PKDD 2005, A.M. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, eds. (Springer), pp. 675–683.

39. Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE *10*, e0118432.

40. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

41. Yi, H., Raman, A.T., Zhang, H., Allen, G.I., and Liu, Z. (2018). Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinformatics *34*, 1141–1147.

42. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

43. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, eds. (Springer), pp. 319–326.

44. Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J. (2016). Data Mining: Practical Machine Learning Tools and Techniques*Fourth Edition* (Morgan Kaufmann).

45. Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics *31*, 2595–2597.

46. Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. *40*, W478-83.

47. Kolde, R., and Vilo, J. (2015). GOsummaries: an R Package for Visual Functional Annotation of Experimental Data. F1000Res. *4*, 574.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
PBMC from MS subjects	IRCCS San Raffaele Scientific Institute, Milan, Italy	https://www.hsr.it:443/
PBMC from healthy subjects	IRCCS San Raffaele Scientific Institute, Milan, Italy	https://www.hsr.it:443/
PBMC from other neurological disease subjects	IRCCS San Raffaele Scientific Institute, Milan, Italy	https://www.hsr.it:443/
Critical Commercial Assays		
Lymphoprep density gradient medium	STEMCELL Technologies Inc	Cat# 07801
Trypan Blue	Sigma-Aldrich	Cat#t8154
TRI Reagent Solution	ThermoFisher Scientific	Cat#AM9738
TotalPrep RNA Amplification Kit	Ambion	Cat#AMIL1791
HumanRef-8 v2 expression beadchip	Illumina Inc.	N/A
humanht-12 v4 expression beadchip	Illumina Inc.	Cat#BD-901-1001
Deposited Data		
Raw and processed microarray data	GEO database	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136411
Software and Algorithms		
BeadStudio	Illumina Inc.	https://www.illumina.com/
limma	Ritchie et al.[40]	http://www.bioconductor.org/packages/release/bioc/html/limma.html
DASC	Yi et al.[41]	https://github.com/HaidYi/DASC
ComBat	Johnson et al.[42]	https://rdrr.io/bioc/sva/man/ComBat.html
pca3d	https://cran.r-project.org/web/packages/pca3d/pca3d.pdf	https://cran.r-project.org/web/packages/pca3d/index.html
KNIME	Berthold et al.[43]	https://www.knime.com/
Weka 3.7	Witten et al.[44]	https://hub.knime.com/knime/extensions/org.knime.features.ext.weka_3.7/latest

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for data should be directed to and will be fulfilled by the Lead Contact, Dr. Cinthia Farina (farina.cinthia@hsr.it), Head of Immunobiology of Neurological Disorders Lab, Institute of Experimental Neurology (INSpe) and Division of Neuroscience, San Raffaele Scientific Institute Building Dibit 2-San Gabriele,Via Olgettina, 58 20132 Milan – Italy.

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Raw and processed microarray datasets were deposited at GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136411). Accession is currently private and can become available upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Clinical investigations were conducted according to the principles expressed in the Declaration of Helsinki and after approval of the Ethics Committee of San Raffaele Hospital. Written informed consent was obtained from all participants. MS subjects were diagnosed according to McDonald criteria[35] and were clinically stable at the time of blood sampling. Moreover they were not suffering from any other acute or chronic inflammatory/autoimmune diseases, had not started any immunomodulatory therapy for MS yet or were treatment-free during at least one year before sampling. They had not been treated with steroids during at least one month prior to their enrollment in the study. All healthy subjects had no acute or chronic inflammatory or autoimmune disorders. The study included a total of 313 individuals (172 females and 141 males, with a mean age of 41.7 y.), comprising of 60 healthy controls (HC), 57 subjects with CIS, 169 clinically defined MS cases and 27 OND cases. The MS cohort contained 108 RR-MS, 26 secondary SP-MS and 35 PP-MS cases. The OND cohort included both inflammatory and non-inflammatory neurological diseases (4 stroke, 2 Parkinson disease, 2 spastic paraparesis, 2 primary progressive aphasia, 2 transient global amnesia, 2 migraine with aura, 1 autoimmune encephalitis, 1 chronic cerebrovascular disease, 1 chronic cephalgia, 1 asthenia associated with thyroid disfunction, 1 myelopathy, 1 herniated disk, 1 medullary meningioma, 1 obstructive hydrocephalus, 1 progressive cerebellar atrophy, 1 brain tumor, 1 CNS vasculitis. 176 subjects (39 HC, 46 CIS, 23 PP-MS, 47 RR-MS, 21 SP-MS) out of 313 were included in a previously published study by our group,[11] 5 subjects were sampled twice to evaluate biological variability. Additional 137 subjects (21 HC, 11 CIS, 12 PP-MS, 61 RR-MS, 5 SP-MS, 27 OND) were recruited for this study. The clinical and demographical characteristics of all individuals included in the present study are summarized in Table 1. Peripheral blood was drawn between 9 and 12 a.m.

METHOD DETAILS

PBMC isolation and RNA extraction

PBMC were isolated using a discontinuous density gradient (Lymphoprep, STEMCELL Technologies Inc). Viable cells were counted by Trypan Blue (Sigma-Aldrich) exclusion. Then total RNA was extracted using TRI-Reagent (ThermoFisher Scientific) and stored at −80°C. Quantification and quality control of RNA were performed on Bioanalyzer 2100 (Agilent).

Microarray experiment and data processing

PBMC transcriptomes relative to 176 subjects and 10 biological or technical replicates were conducted using HumanRef-8 v2 arrays and previously published.[11,12] PBMC transcriptomes relative to additional 137 subjects and 13 technical replicates were generated with Illumina HumanHT-12 v4 arrays (Illumina, Netherlands). Reverse transcription and biotinylated cRNA synthesis were performed using the Illumina TotalPrep RNA Amplification Kit (Ambion), according to the manufacturer's protocol. Array hybridization, washing, staining and scanning in the Beadstation 500 (Illumina Inc) were performed according to standard Illumina protocols. Microarray raw data originating from Illumina v2 and v4 arrays were pre-processed, normalized independently and then merged into a single dataset according to the common probes. Raw intensities were background subtracted and filtered according to detection p values ($p < 0.05$ in at least 20% of samples) and then normalized using quantile normalization. Pre-processed data were log2 transformed and inspected for the presence of hidden batch effects. Hierarchical clustering of technical and biological replicates and principal component analysis (PCA) were used to assess dataset integrity following batch correction procedure. Replicates were then removed from further analysis. Generation of training and independent test sets was carefully conducted in order to avoid clinical, age and gender biases between the two datasets (Table 1).

Machine learning pipeline

For unbiased comparison of distinct learning algorithms we developed a nested cross-validation (NCV) workflow (Figure 2A) followed by final validation on the independent test set (Figure 2B).[23] Using NCV we compared three algorithms based on decision trees: Functional Trees (FT), Adaptive Boosting applied to FT (ADAboost-FT) and Random Forests (RF). Fine-tuning of the specific hyper-parameters of each algorithm was performed automatically in an inner cross-validation loop (innerCV) nested inside an outer cross-validation loop (outerCV), which was used for the proper estimation of the predictive model. To preserve class ratio in each split of the training data, a ten-fold stratified CV was applied to both inner and outer loops.[36] Hyper-parameters used for ADAboost were the number of iterations ($I = 2$ to 10; step = 1) and the learning rate ($h = 0.2$ to 0.9; step = 0.1). Resampling instead of reweighting was applied to ADAboost as it showed superior performance in imbalanced data.[37] Hyper-parameters used for RF were the number of trees ($nt = 500$ to 2000; step = 500) and the number of selected features ($mtry = 50$ to 300; step = 50; where sqrt of nTot features = 100). The combination of hyper-parameters that maximized the F-measure (harmonic mean of precision and recall) of the class of interest was retained as optimal and applied to the algorithm. We applied a stratified sampling method to RF to mitigate class imbalance issues. FT was applied with the default parametrization except for the number of iterations that was set on CV and for error minimization set on probabilities. To reduce computation time of ADAboost we applied a small weight trimming ($w = 0.01$) to the associated FT base learner for all tasks except CIS versus All, where the high heterogeneity of both classes prompted us to avoid weight trimming. To further reduce heterogeneity issues in CIS versus All, we employed a CV stratification based on all the different subclasses instead of the simple binary stratification used for all the other tasks. Moreover, in MS versus non-MS and CIS versus All tasks, the larger dataset dimensions imposed the use of the AIC criterion in FT base learner in order to achieve reasonable time reduction in ADAboost execution.[38] Area under the precision-recall curve (AUPRC) was used to evaluate candidate models, as it constitutes a more reliable estimator than classic AUROC in imbalanced datasets.[39]

QUANTIFICATION AND STATISTICAL ANALYSIS

Microarray data processing

The raw gene expression intensities were exported from the Beadstation using BeadStudio software (Illumina Inc) and further processed in R-Bioconductor. Microarray raw data were pre-processed and normalized using limma package.[40] DASC package was used for the detection of hidden batch effects and batch correction was performed using Combat.[41,42] PCA and cluster analysis were performed using respectively pca3d package and hclust function in R. Regarding cluster analysis, Spearman correlation was used to compute the distance matrix and clustering was performed using the single-linkage metrics. Distance (1-correlation) was used to represent clustering results. Statistical significance of differences in age and gender-ratios was assessed by t test and chi-square test, respectively, using p.value $\leq$ 0.05 as a cut-off.

Machine learning pipeline

Machine learning pipeline was assembled using KNIME, integrated with Weka 3.7.[43,44] FT and ADAboost (RealAdaBoost) were implemented using Weka libraries, while RF was implemented using the tree ensemble node of KNIME. Details about algorithm optimization are described in the previous section. Precision-recall curves and AUPRC were generated using R package PRROC.[45]

Functional annotation

Functional enrichment analysis of predictive gene signatures was performed in GeneCodis3 using Gene Ontology Biological Process categories and genes on HumanRef-8 v2 array as background.[46] Enrichment p.values were adjusted using Benjamini-Hochberg correction and an adjusted p.value $\leq$ 0.05 was used as significance cut-off. Tag-clouds of enriched terms were generated using GO summaries package in R.[47]