Electronic Supplementary Material (ESI) for Analyst.

Supplementary information

Serum-based differentiation between Multiple Sclerosis and Amyotrophic Lateral Sclerosis by Random Forest classification of FTIR spectra

Youssef El Khoury,[a*] Nicolas Collongues,[b,c] Jérôme De Sèze,[b,c] Vildan Gulsari,[a] Christine Patte-Mensah,[b] Gilles Marcou,[d] Alexandre Varnek,[d] Ayikoé Guy Mensah-Nyagan[b] and Petra Hellwig[a*]

a. Laboratory of Bioelectrochemistry and Spectroscopy, UMR 7140 University of Strasbourg/CNRS, 4 rue Blaise Pascal, 67000 Strasbourg, France.
b. Biopathologie de la Myéline, Neuroprotection et Stratégies Thérapeutiques, INSERM U1119, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg, Bâtiment 3 de la Faculté de Médecine, Strasbourg, France.
c. Department of Neurology, University Hospital of Strasbourg, Strasbourg, France.
d. Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 4 rue Blaise Pascal, 67000 Strasbourg, France.

Table S1: Demographic details of the MS patients participating in the study including the age at which the serum sample was taken, the age of the first symptoms apparition, the MS phenotype and the EDSS score at the time of the serum sampling. PP stands for *Primary Progressive*, SP for *Secondary Progressive* and EDSS for *Expanded Disability Status Scale* (higher scores reflect more severe disabilities).

Patient Number	Age (gender)	First symptoms at age	MS phenotype	EDSS
1	69 (F)	41	PP	6.0
2	48 (F)	24	SP	6.0
3	66 (F)	50	PP	6.5
4	51 (F)	22	SP	6.5
5	38 (F)	21	SP	6.5
6	69 (M)	48	SP	7.5
7	68 (F)	42	PP	7.0
8	66 (F)	21	SP	6.0
9	58 (F)	30	SP	5.0
10	68 (F)	51	SP	4.0
11	52 (M)	46	PP	5.0
12	65 (F)	58	PP	6.0
13	72 (M)	59	PP	6.5
14	60 (M)	30	SP	8.5
15	42 (M)	35	SP	6.5
16	69 (M)	39	PP	7.0
17	67 (F)	42	SP	6.5
18	56 (F)	42	SP	6.0
19	55 (F)	41	SP	6.5
20	59 (F)	32	SP	6.0
21	73 (M)	56	PP	6.5
22	65 (F)	48	PP	6.5
23	61 (F)	36	SP	5.0
24	78 (M)	38	SP	7.5
25	59 (M)	41	SP	6.0
26	61 (F)	41	PP	6.5
27	45 (F)	30	SP	6.0
28	54 (M)	24	SP	7.5
29	65 (F)	42	SP	6.5
30	66 (F)	29	SP	7.5
31	54 (F)	34	PP	6.5
32	59 (M)	39	SP	5.5

Table S2: Demographic details of the ALS patients participating in the study including the age at which the serum sample was taken and the ALSFRS score at the time of the serum sampling. ALSFRS stands for *ALS Functional Rating Scale* (higher scores reflect better autonomy of the patients (i.e. less functional impairment).

Patient Number	Age (gender)	ALSFRS
1	68 (F)	41
2	60 (M)	21
3	69 (F)	44
4	76 (M)	36
5	69 (F)	44
6	81 (M)	25
7	56 (M)	44
8	66 (M)	32
9	61 (M)	35
10	89 (F)	24
11	84 (F)	33
12	65 (F)	42
13	68 (F)	22
14	79 (F)	34
15	71 (F)	28
16	62 (M)	36
17	61 (F)	24
18	72 (M)	25
19	54 (M)	12
20	73 (M)	27
21	51 (F)	35
22	75 (F)	32
23	60 (F)	38
24	70 (F)	45
25	47 (M)	33
26	58 (F)	14
27	46 (M)	37
28	69 (F)	33
29	66 (M)	22
30	68 (F)	24
31	75 (F)	36
32	56 (M)	44

Table S3. Tentative assignment of the signals observed in the sera FTIR spectra along with the major biological molecule contributing to the signal. [1-3]

Frequency / cm^{-1}	Assignment	Biomolecule
3285	$\upsilon(\mathrm{NH})$	Proteins
2985	$\upsilon_{as}(\mathrm{CH_3})$	Lipids and proteins
2929	$\upsilon_{as}(\mathrm{CH_2})$	Lipids and proteins
2871	$\upsilon_{s}(\mathrm{CH_3})$	Lipids and proteins
2852	$\upsilon_{s}(\mathrm{CH_2})$	Lipids and proteins
1738	$\upsilon(\mathrm{C{=}O})$	Lipids
1640	amide I	Proteins
1535	amide II	Proteins
1453	$\delta_{as}(\mathrm{CH})_3$	Lipids and proteins
1397	$\delta_{s}(\mathrm{CH})_3$	Lipids and proteins
1307	amide III	Proteins
1240	$\upsilon_{as}(\mathrm{PO_2^-})$	DNA
1169	$\upsilon(\mathrm{C–O})$	Glycoproteins and carbohydrates
1117	$\upsilon(\mathrm{C–O})$	RNA
1031	$\upsilon(\mathrm{CH_2OH})$, υ $\mathrm{PO_2^-}$)	Carbohydrates and DNA
929	deoxyribose	Left-Handed DNA (Z-form)
834	deoxyribose	DNA (B and Z forms)
734	δ(ring)	DNA

How to download and use the Random Forest classifier model:

The model is available as a ZIP file on the website of *Analyst* as a supplementary information. Note: In order to use the model, the second derivatives of the spectra need to be saved in the ARFF format. The frequencies designated by *“attributes”* in the model need to be strictly respected.

Step 1:

In Weka, use the “SimpleGLI” tool in the interface “Weka GUI Chooser”

Step 2:

Type the following command line by replacing file.arff » by the file containing the new data and by replacing “file.model by the name of the model’s file.

```
java weka.classifiers.trees.RandomForest -T file.arff -L file.model --classifications "weka.classifiers.evaluation.output.prediction.PlainText"
```

Note: Please be careful to change the directory.

Figure S1: Percentage of explained variance with the first ten principal components as well as the corresponding error bars. A total of 91.7% of variance is explained.

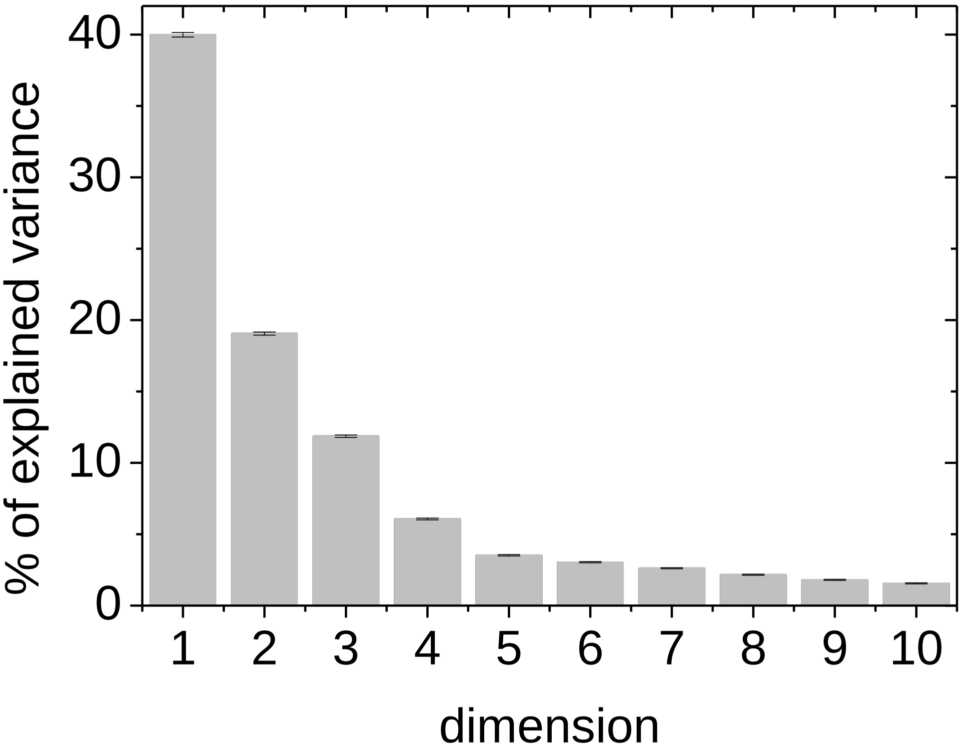


Figure S2: loadings of the first three principal components PC1, PC2 and PC3.

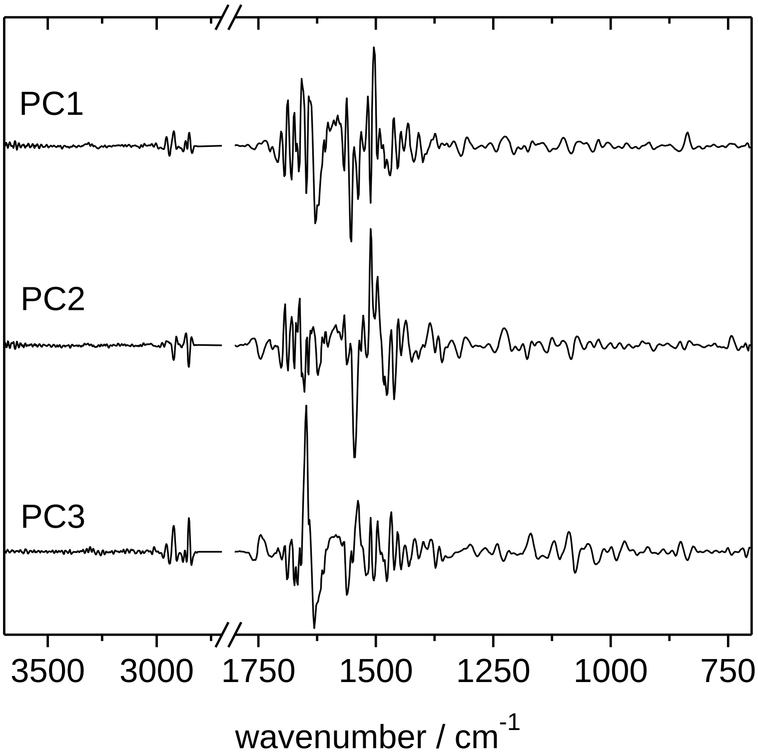


Figure S3: 2D scatterplot of the loadings of all the samples on PC1 and PC2 (A), PC2 and PC3 (B) and PC1 and PC3 (C).Black disks (Healthy controls), red triangles (MS) and green squares (MS) loadings of the first three principal components PC1, PC2 and PC3.

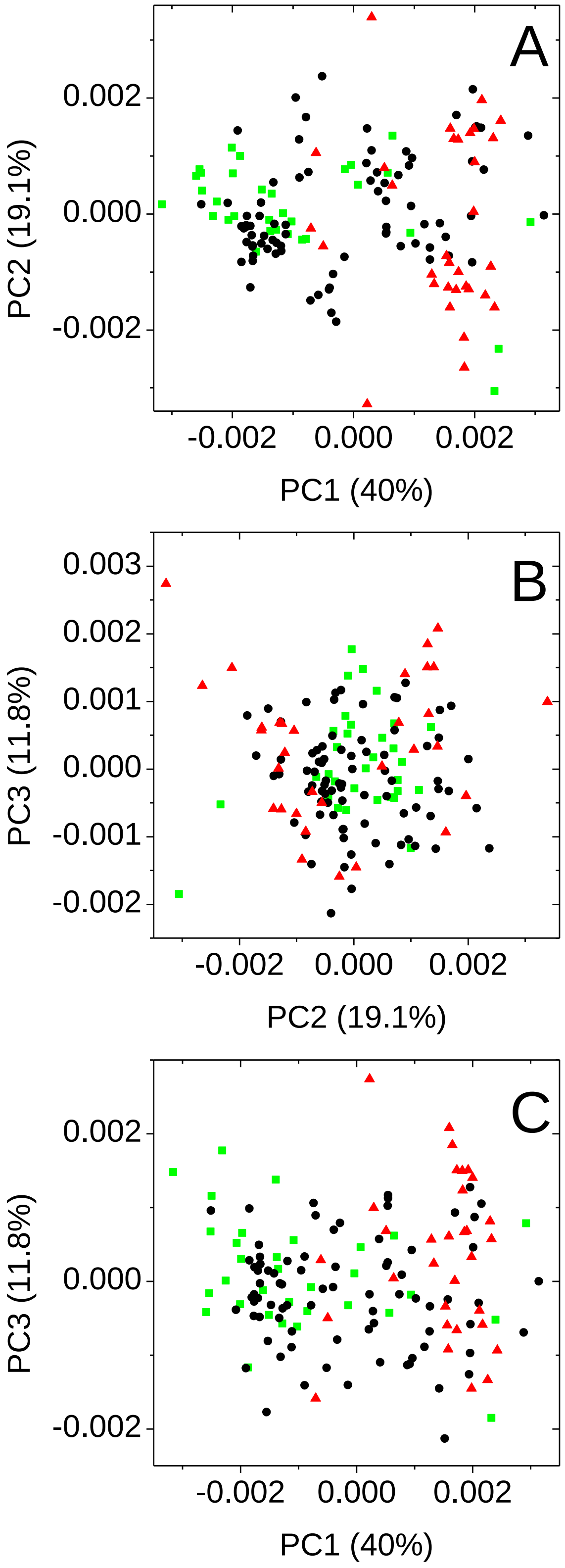


Figure S4: ROC curves of HC (black), M0 (red) and ALS (green).

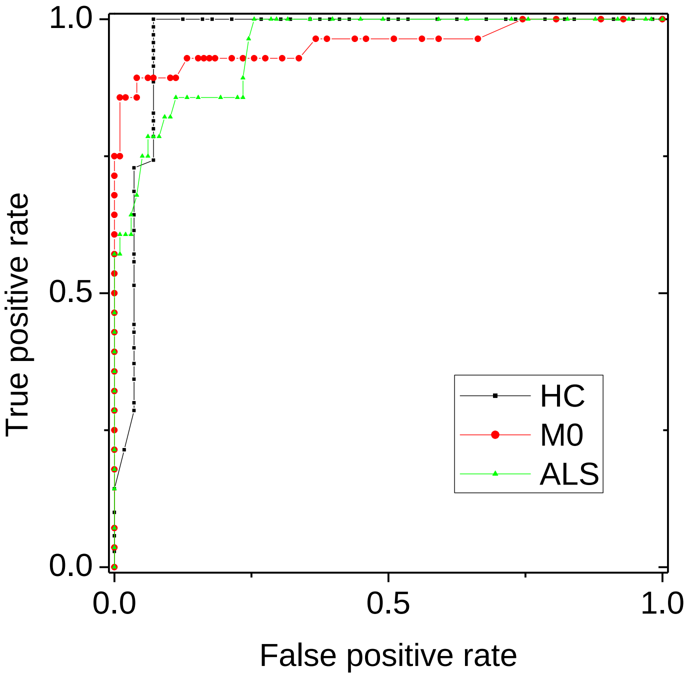
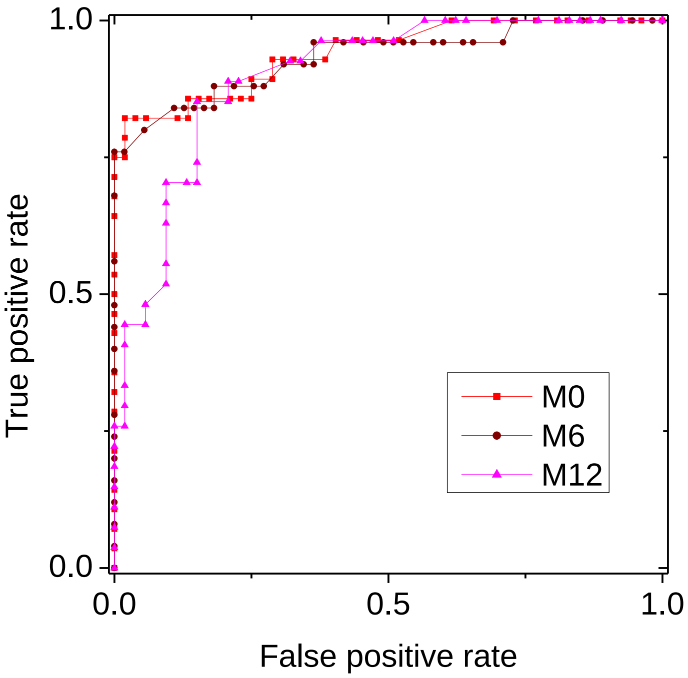


Figure S5. ROC curves of M0 (red) M6 (wine) and M12 (magenta).



References

1. A. Barth, *Biochim. Biophys. Acta*, 2007, **1767**, 1073-1101.
2. I. U. Rehman, Z. Movasaghi and S. Rehman, *Vibrational spectroscopy for tissue analysis*, CRC Press, Boca Raton, 1st Edition edn., 2012.
3. M. Ghomi, R. Letellier, J. Liquier and E. Taillandier, *Intl. J. Biochem.*, 1990, **22**, 691-699.