Electronic Supplimentary Information

Noninvasive diagnostic test for lung cancer using biospectroscopy and variable selection techniques in saliva samples

Camilo L.M. Morais[1,2], Kássio M.G. Lima[1], Andrew W Dickinson[3], Tarek Saba[3], Thomas Bongers[3], Maneesh N Singh[4,5], Francis L Martin[3,4,*], Danielle Bury[3,*]

[1] *Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil*

[2] *Center for Education, Science and Technology of the Inhamuns Region, State University of Ceará, Tauá 63660-000, Brazil*

[3] *Department of Cellular Pathology, Blackpool Teaching Hospitals NHS Foundation Trust, Whinney Heys Road, Blackpool FY3 8NR, UK*

[4] *Biocel UK Ltd., Hull HU10 6TS, UK*

[5] *Chesterfield Royal Hospital, Chesterfield Road, Calow, Chesterfield S44 5BL, UK*

*** Corresponding authors:** francis.martin2@nhs.net (F.L.M.); danielle.bury@nhs.net (D.B.)

Table S1: Age and gender for the different groups (lung cancer and controls). *P*-values calculated based on an ANOVA test.

Patient Characteristics	Lung cancer (LC) (N = 56)	Controls (N = 1888)	*P*-value
Age			8×10^{-5}
Mean ± SD (range)	69.8 ± 5.2 (57–76)	67.1 ± 5.1 (56–77)	
Gender, **n/N** (%)			0.5894
Female	21 (37.5)	776 (41.1)	
Male	35 (62.5)	1112 (58.9)	
Others comorbidities, n/N (%)			0.0141
Prostate cancer	--	30 (1.589)	0.3420
Breast cancer	--	29 (1.536)	0.3503
Bladder cancer	--	14 (0.741)	0.5181
Squamous cell carcinoma (skin)	--	13 (0.689)	0.5335
Colorectal cancer	--	7 (0.371)	0.6482
Melanoma	--	4 (0.212)	0.7304
Cervical cancer	--	3 (0.159)	0.7654
CLL	--	3 (0.159)	0.7654
Ovarian cancer	--	2 (0.106)	0.8076
Endometrial cancer	--	2 (0.106)	0.8076
CMML	--	2 (0.106)	0.8076
Other cancers	--	7 (0.371)	0.6482
Other systematic diseases	--	68 (3.602)	0.1484
Healthy controls	--	1704 (90.254)	8×10^{-10}

COPD: Chronic obstructive pulmonary disease; CLL: Chronic lymphocytic leukemia; CMML: Chronic myelomonocytic leukaemia.

Table S2: Classification performance by LDA- and QDA-based models to distinguish between lung cancer and controls in the training set (70% of data).

				TRAINING		
MODEL	PARAMETERS	ACCURACY	SENSITIVITY	SPECIFICITY	F-SCORE	G-SCORE
PCA-LDA	10 PCs (97.64% explained variance)	0.700	0.667	0.701	0.683	0.683
PCA-QDA	10 PCs (97.64% explained variance)	0.964	0.692	0.972	0.809	0.820
SPA-LDA	4 selected wavenumbers	0.737	0.513	0.744	0.607	0.618
SPA-QDA	4 selected wavenumbers	0.979	0.923	0.980	0.951	0.951
GA-LDA	3 selected wavenumbers	0.737	0.590	0.741	0.657	0.661
GA-QDA	3 selected wavenumbers	0.979	0.923	0.981	0.951	0.952

Table S3: Classification performance by LDA- and QDA-based models to distinguish between lung cancer and controls in the test set (30% of data).

		TEST				
MODEL	PARAMETERS	ACCURACY	SENSITIVITY	SPECIFICITY	F-SCORE	G-SCORE
PCA-LDA	10 PCs (97.64% explained variance)	0.789	0.353	0.802	0.490	0.532
PCA-QDA	10 PCs (97.64% explained variance)	0.985	0.941	0.986	0.963	0.963
SPA-LDA	4 selected wavenumbers	0.793	0.471	0.802	0.593	0.614
SPA-QDA	4 selected wavenumbers	0.988	0.941	0.989	0.965	0.965
GA-LDA	3 selected wavenumbers	0.935	0.294	0.954	0.450	0.530
GA-QDA	3 selected wavenumbers	0.991	1.000	0.991	0.996	0.996