Comparison of predictive models on integrated omics data

Irene Testa✉, Giuseppe Prencipe, Corrado Priami, Alina Sîrbu

Department of Computer Science, University of Pisa, Italy

i.testa@studenti.unipi.it

Università di Pisa

1. Introduction

- Omics datasets often have a limited number of samples, necessitating integration for Machine Learning model training.
- A comprehensive comparison of Machine Learning methods on integrated omics data is still missing (currently limited to a single disease or a few models).

2. Data

- Datasets were selected from NCBI's Gene Expression Omnibus [1] to exhibit a wide range of characteristics (see Table 1).

Name (accession)	Platform	Classes (#samples)	#Features
LC1 (GSE19804)	GPL570	Lung Cancer (60), Control (60)	54 675
LC2 (GSE43346)	GPL570	Small Cell Lung Cancer (23), Control (42)	54 675
PSO1 (GSE14905)	GPL570	Psoriasis (61), Control (21)	54 675
PSO2 (GSE13355)	GPL570	Psoriasis (58), Control (64)	54 675
SK1 (GSE15605)	GPL570	Primary Melanoma (46), Metastatic Melanoma (12), Control (16)	44 137
SK2 (GSE46517)	GPL96	Primary Melanoma (31), Metastatic Melanoma (73), Control (7)	22 215
LK1 (GSE51082)	GPL96	Acute Myeloid Leukemia (37), Chronic Lymphocytic Leukemia (41), Chronic Myeloid Leukemia (22), Myelodysplastic Syndrome (10), Precursor B-cell Acute Lymphoblastic Leukemia (17), T-cell Acute Lymphoblastic Leukemia (12)	22 283
LK2 (GSE51082)	GPL97	Acute Myeloid Leukemia (37), Chronic Lymphocytic Leukemia (41), Chronic Myeloid Leukemia (22), Myelodysplastic Syndrome (10), Precursor B-cell Acute Lymphoblastic Leukemia (17), T-cell Acute Lymphoblastic Leukemia (13)	22 645
AD1 (GSE63060)	GPL6947	Alzheimer's disease (145), Control (104)	38 323
AD2 (GSE63061)	GPL10558	Alzheimer's disease (140), Control (135)	32 049
AD3 (GSE33000)	GPL4372	Alzheimer's disease (310), Control (157)	38 734
AD4 (GSE44770)	GPL4372	Alzheimer's disease (129), Control (101)	39 005
PD1 (GSE62283)	GPL13669	Parkinson's disease (132), Control (156)	9 480
PD2 (GSE29654)	GPL13669	Parkinson's disease (174), Control (80)	9 480

Platform	Type
GPL570	Affymetrix Human Genome U133 Plus 2.0 Array
GPL96	Affymetrix Human Genome U133A Array
GPL97	Affymetrix Human Genome U133B Array
GPL6947	Illumina HumanHT-12 V3.0 expression beadchip
GPL10558	Illumina HumanHT-12 V4.0 expression beadchip
GPL4372	Rosetta/Merck Human 44k 1.1 microarray
GPL13669	Invitrogen ProtoArray v5.0

Table 1: *Datasets details.*

3. Methods

- We compared 7 classifiers: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Random Forest (RF), Extreme Gradient Boosting (XGB), Nearest Centroid (NC) and the Rank Aggregation Classifier (RAC) [2], a classifier similar to NC, but using a ranked representation of features and rank aggregation methods for obtaining the centroids.
- Models were compared using different pre-processing techniques and integration strategies, evaluating also the effect of feature selection with the Recursive Feature Elimination (RFE) algorithm (see Figure 1). The implementation of the evaluation pipeline is publicly available at [3].

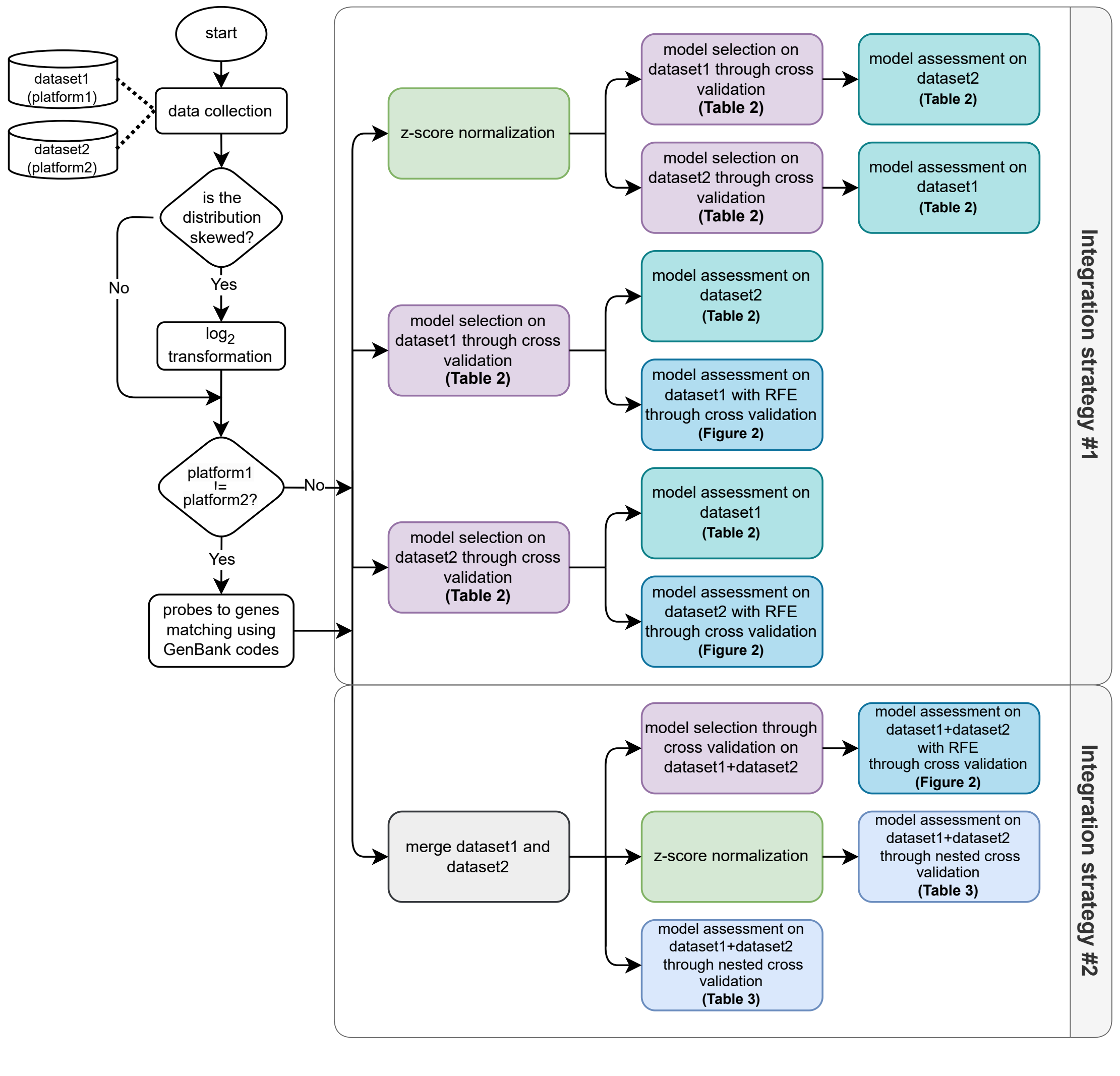
start
dataset1 (platform1)
dataset2 (platform2)
data collection
is the distribution skewed?
No
Yes
log$_2$ transformation
platform1 != platform2?
No
Yes
probes to genes matching using GenBank codes
z-score normalization
model selection on dataset1 through cross validation (Table 2)
model assessment on dataset2 (Table 2)
model selection on dataset2 through cross validation (Table 2)
model assessment on dataset1 (Table 2)
model selection on dataset1 through cross validation (Table 2)
model assessment on dataset2 (Table 2)
model assessment on dataset1 with RFE through cross validation (Figure 2)
model selection on dataset2 through cross validation (Table 2)
model assessment on dataset1 (Table 2)
model assessment on dataset2 with RFE through cross validation (Figure 2)
Integration strategy #1
merge dataset1 and dataset2
model selection through cross validation on dataset1+dataset2
model assessment on dataset1+dataset2 with RFE through cross validation (Figure 2)
z-score normalization
model assessment on dataset1+dataset2 through nested cross validation (Table 3)
model assessment on dataset1+dataset2 through nested cross validation (Table 3)
Integration strategy #2

Figure 1: *Evaluation pipeline.*

- F1 score was computed considering each class as positive, then averaging the scores by weighting them by the number of positive instances.

3. Results

Training dataset	Test dataset	RAC CV score	RAC Test score	NC CV score	NC Test score	KNN CV score	KNN Test score	SVM CV score	SVM Test score	GNB CV score	GNB Test score	RF CV score	RF Test score	XGB CV score	XGB Test score
LC1	LC2	0.925	0.708	0.925	0.602	*0.908*	**0.865**	0.950	0.501	0.933	*0.185*	0.958	0.217	**0.975**	0.490
LC1*	LC2*	0.925	0.708	0.925	0.773	*0.908*	**0.865**	0.958	0.410	0.933	*0.185*	0.958	0.360	**0.975**	0.677
LC2	LC1	0.985	*0.333*	0.985	*0.333*	**1.000**	*0.333*	**1.000**	*0.333*	0.985	*0.333*	**1.000**	**0.659**	*0.861*	*0.333*
LC2*	LC1*	0.985	*0.333*	0.985	*0.333*	**1.000**	**0.606**	**1.000**	0.369	0.985	0.369	**1.000**	0.352	*0.923*	0.369
PSO1	PSO2	0.916	0.984	*0.884*	**0.992**	0.950	*0.306*	**0.975**	0.361	0.922	*0.306*	**0.975**	0.361	0.962	0.361
PSO1*	PSO2*	0.916	0.984	*0.884*	**0.992**	0.950	0.611	**0.975**	0.876	0.963	*0.306*	**0.975**	0.361	0.950	0.569
PSO2	PSO1	**1.000**	*0.104*	**1.000**	0.512	**1.000**	0.666	**1.000**	0.653	**1.000**	**0.669**	**1.000**	0.635	**1.000**	0.635
PSO2*	PSO1*	**1.000**	*0.104*	**1.000**	0.449	**1.000**	**0.713**	**1.000**	0.627	**1.000**	0.601	**1.000**	0.635	**1.000**	0.635
SK1	SK2	0.801	**0.836**	*0.776*	0.454	0.869	0.204	**0.874**	*0.007*	0.795	0.122	0.788	0.154	0.781	0.564
SK1*	SK2*	0.801	0.836	*0.776*	**0.827**	0.850	*0.313*	**0.890**	0.823	0.781	0.499	0.811	0.702	0.794	0.553
SK2	SK1	0.938	0.535	0.929	0.532	0.938	0.629	**0.955**	0.535	*0.904*	*0.397*	0.936	0.627	0.927	**0.680**
SK2*	SK1*	0.938	0.535	0.929	0.481	0.930	**0.728**	**0.964**	0.634	*0.911*	*0.045*	0.936	0.666	0.918	0.526
LK1	LK2	0.861	0.809	*0.819*	0.595	0.956	**0.986**	**0.993**	0.879	0.940	0.335	0.971	0.440	0.928	*0.331*
LK1*	LK2*	0.861	**0.809**	*0.785*	*0.133*	0.941	*0.133*	**0.993**	0.324	0.921	*0.133*	0.964	0.198	0.971	0.411
LK2	LK1	0.857	0.828	*0.787*	0.649	0.914	0.700	**1.000**	**0.894**	0.900	0.753	0.964	0.642	0.909	*0.560*
LK2*	LK1*	0.857	**0.828**	*0.762*	0.126	0.942	*0.051*	**1.000**	0.144	0.942	0.134	0.957	0.164	0.935	0.062
AD1	AD2	0.706	0.593	*0.684*	0.640	0.750	**0.655**	**0.830**	0.343	0.685	0.343	0.741	*0.323*	0.816	0.343
AD1*	AD2*	0.706	0.593	*0.684*	**0.621**	0.750	0.608	**0.834**	0.343	0.685	0.579	0.730	*0.323*	0.816	0.486
AD2	AD1	0.625	0.665	*0.621*	**0.693**	0.658	0.680	**0.712**	0.429	0.633	0.625	0.694	*0.246*	0.711	*0.246*
AD2*	AD1*	0.625	0.665	*0.617*	0.695	0.662	0.651	**0.713**	*0.246*	0.633	**0.701**	0.691	0.582	0.696	0.387
AD3	AD4	0.865	0.857	*0.823*	*0.808*	0.889	0.887	**0.957**	**1.000**	0.844	0.830	0.904	**1.000**	0.914	**1.000**
AD3*	AD4*	*0.865*	*0.857*	0.874	0.870	0.901	0.931	**0.946**	0.996	0.882	0.883	0.913	0.987	0.925	**1.000**
AD4	AD3	0.879	*0.875*	0.874	0.882	0.874	0.898	**0.913**	**0.961**	*0.866*	0.877	0.909	0.937	0.896	0.950
AD4*	AD3*	*0.879*	*0.875*	*0.879*	0.876	0.892	0.941	**0.913**	**0.963**	0.883	0.884	**0.913**	0.932	0.904	0.950
PD1	PD2	0.782	*0.288*	0.750	0.431	0.840	0.973	0.903	**1.000**	*0.635*	0.802	0.885	**1.000**	**0.924**	**1.000**
PD1*	PD2*	0.782	0.288	*0.764*	*0.199*	0.827	0.953	**0.920**	**1.000**	0.792	0.795	0.879	**1.000**	0.906	**1.000**
PD2	PD1	**1.000**	**0.726**	*0.934*	0.706	**1.000**	0.669	**1.000**	0.675	**1.000**	0.587	**1.000**	0.572	**1.000**	*0.536*
PD2*	PD1*	**1.000**	**0.726**	*0.988*	0.608	**1.000**	0.614	**1.000**	0.625	**1.000**	0.657	**1.000**	0.572	**1.000**	*0.545*
Mean		0.867	0.653	*0.842*	0.631	0.896	**0.675**	**0.933**	0.612	0.860	*0.512*	0.909	0.558	0.900	0.574
Rank		5	2	7	3	4	1	1	4	6	7	2	6	3	5
Mean*		0.867	**0.653**	*0.847*	0.570	0.897	0.623	**0.936**	0.599	0.879	*0.484*	0.909	0.560	0.908	0.584
Rank*		6	1	7	5	4	2	1	3	5	7	2	6	3	4

Table 2: *F1 scores when training and testing on different datasets ('*' for z-score normalized data). For each classifier, we provide the Cross Validation (CV) score for the best hyper-parameter combination on the training dataset and the score on the test dataset.*

Dataset	RAC	NC	KNN	SVM	GNB	RF	XGB
LC1+LC2	0.930	0.795	0.929	**0.968**	*0.604*	0.962	0.930
LC1+LC2*	0.930	0.930	*0.913*	**0.962**	0.919	0.946	0.941
PSO1+PSO2	0.894	*0.612*	0.961	0.975	*0.612*	**0.980**	0.936
PSO1+PSO2*	*0.894*	*0.894*	0.966	**0.985**	*0.894*	0.971	0.946
SK1+SK2	0.677	*0.664*	0.898	**0.935**	0.683	0.892	0.885
SK1+SK2*	0.677	*0.664*	0.914	**0.935**	0.677	0.919	0.913
LK1+LK2	0.829	*0.801*	0.932	**0.996**	0.939	0.967	0.936
LK1+LK2*	0.829	*0.421*	0.920	**0.996**	0.701	0.960	0.939
AD1+AD2	0.596	*0.535*	0.695	**0.744**	*0.535*	0.692	0.692
AD1+AD2*	*0.596*	0.645	0.702	**0.748**	0.624	0.694	0.719
AD3+AD4	0.867	*0.847*	0.944	**0.973**	0.854	0.957	0.967
AD3+AD4*	*0.867*	0.881	0.947	**0.971**	0.883	0.955	0.966
AD1+AD2+AD3+AD4	0.695	*0.557*	0.796	0.792	*0.557*	0.790	**0.808**
AD1+AD2+AD3+AD4*	0.695	0.651	0.799	0.794	*0.557*	0.821	**0.831**
PD1+PD2	0.886	*0.759*	0.927	0.948	0.803	0.937	**0.959**
PD1+PD2*	0.886	0.873	0.916	0.943	*0.843*	**0.946**	0.939
Mean	0.797	*0.696*	0.885	**0.916**	0.698	0.897	0.889
Rank	5	7	4	1	6	2	3
Mean*	0.797	*0.745*	0.885	**0.917**	0.762	0.902	0.899
Rank*	5	7	4	1	6	2	3

Table 3: *F1 scores from nested cross validation on merged datasets ('*' for z-score normalized data).*

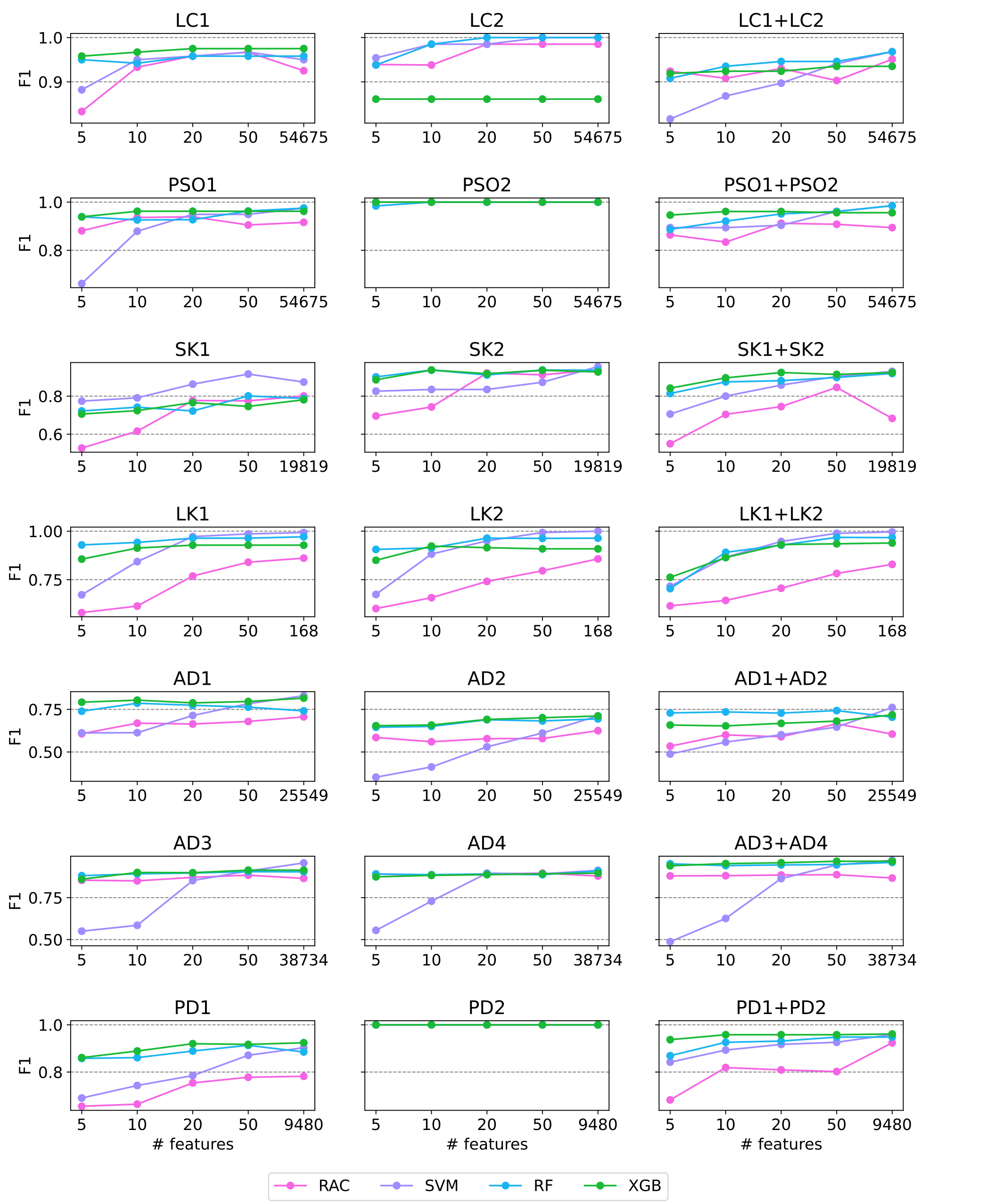

Figure 2: *F1 scores after feature selection on individual and merged datasets. Results are displayed for 5, 10, 20, and 50 features as well as without feature selection (using all features).*

Dataset	RAC	RAC RFE	SVM	SVM RFE	RF	RF RFE	XGB	XGB RFE
PSO1	5/0	5/0	5/0	4/1	2/6	2/3	1/4	1/4
PSO2	6/0	5/0	5/0	5/0	26/11	4/1	1/0	1/4
PSO1+PSO2	5/0	5/0	5/0	5/0	5/0	5/0	3/2	3/2

Table 4: *Known/Unknown genes in the literature related to Psoriasis among the genes with the highest 5 importance scores (including ties) or selected by RFE (setting the number of features to select to 5). To assess the correlation between specific genes and Psoriasis we searched for "[gene name] AND Psoriasis" on PubMed [4].*

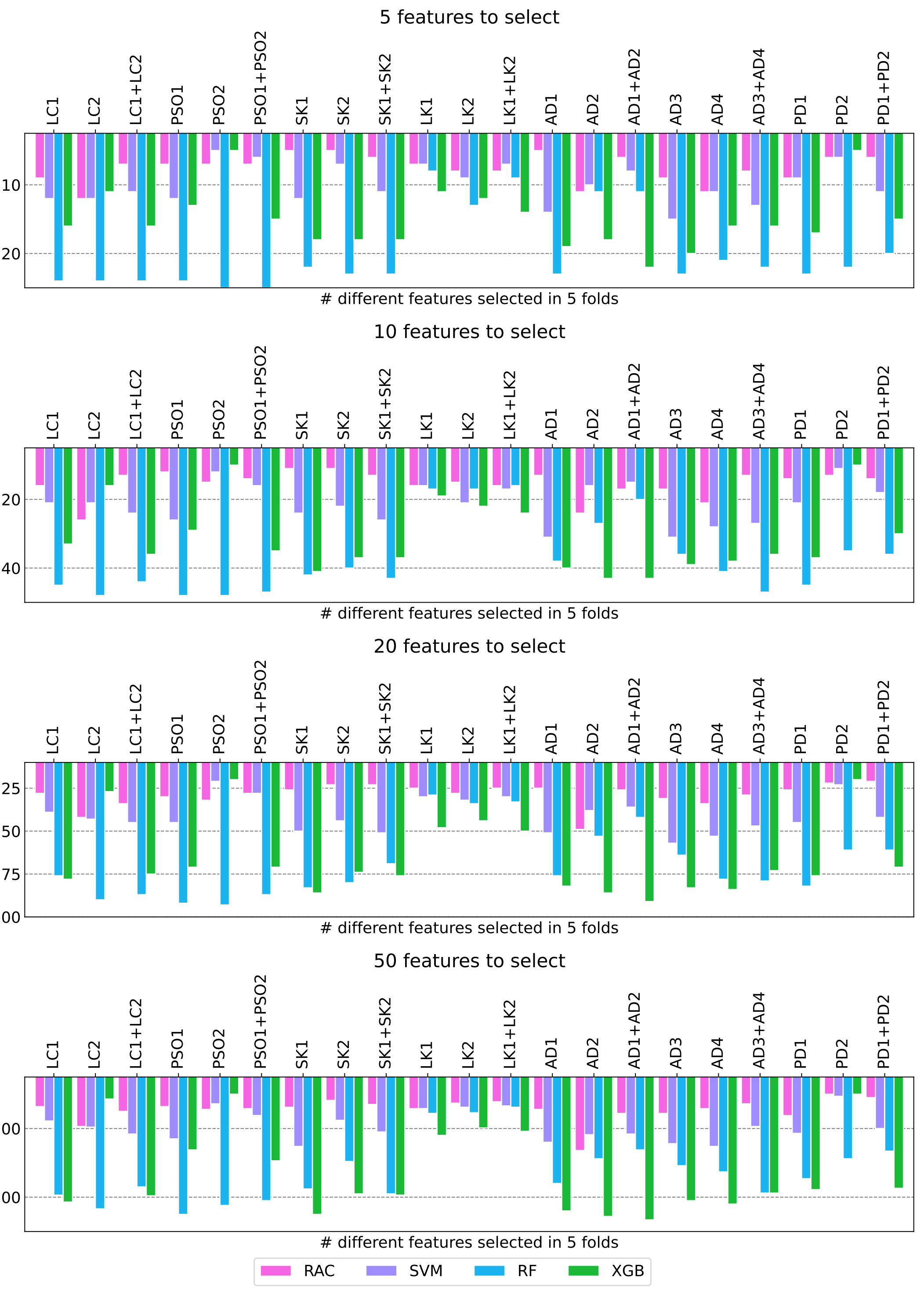

Figure 3: *Feature selection robustness: total number of features selected from the five folds of cross validation. Ideally the total number should be close to the number of features to select for each fold.*

4. Conclusions

- Training on a dataset and testing on another: poor performance, no method stands out.
- Merging datasets: good performance albeit slightly reduced compared to individual datasets; SVM is the top classifier, followed closely by RF.
- Z-score normalization: improves performance on merged dataset, degrades test performance when training and testing on separate datasets.
- RFE typically does not improve the performance of classifiers. As opposed to RAC and SVM, RF and XGB still perform well with fewer features.
- RFE with XGB and RF is not robust to the change of the underlying samples (it tends to select different features over different folds).
- All the features selected by RAC and SVM are supported by the literature; some features selected by RF and XGB are not. The fraction of features without support decreases when moving on to the integrated dataset.

References

[1] https://www.ncbi.nlm.nih.gov/geo.

[2] https://github.com/iretes/rac.

[3] https://anonymous.4open.science/r/geo-classification.

[4] https://pubmed.ncbi.nlm.nih.gov.

Acknowledgments

This work was financed by the European Commission through the HORIZON project "AInCP: clinical validation of Artificial Intelligence for providing a personalized motor clinical profile assessment and rehabilitation of upper limb in children with unilateral Cerebral Palsy" (grant agreement no. 101057309) and by the European Union through the "NextGenerationEU" project, in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), CUP: B83C22003920001.