Comparison of Machine Learning Classifiers on Integrated Transcriptomic Data

Irene Testa, Giuseppe Prencipe, Corrado Priami, Alina Sîrbu

Department of Computer Science, University of Pisa, Italy

IEEE BigData 2023
6th Special Session on Healthcare Data

Introduction

Background

- ▶ Omics data are a valuable resource for building automatic diagnostic tools
- ▶ Given the limited number of samples in each individual dataset, data integration is necessary
- ▶ A comprehensive comparison of Machine Learning methods on integrated omics data is still missing (currently limited to a single disease or to a few models)

Contribution

- ▶ We compared 7 classifiers on integrated data for 6 diseases, over 14 datasets
- ▶ Models were compared employing different pre-processing techniques, evaluating the effect of feature selection and analyzing the robustness and relevance of the extracted features

Methods

Datasets

Table 1: Datasets information

Name	Platform	Classes (#samples)	#Features
LC1	Affymetrix Human Genome U133 Plus 2.0 Array	LC (60), CTL (60)	54 675
LC2	Affymetrix Human Genome U133 Plus 2.0 Array	SCLC (23), CTL (42)	54 675
PSO1	Affymetrix Human Genome U133 Plus 2.0 Array	PSO (61), CTL (21)	54 675
PSO2	Affymetrix Human Genome U133 Plus 2.0 Array	PSO (58), CTL (64)	54 675
SK1	Affymetrix Human Genome U133 Plus 2.0 Array	PMEL (46), MMEL (12), CTL (16)	44 137
SK2	Affymetrix Human Genome U133A Array	PMEL (31), MMEL (73), CTL (7)	22 215
LK1	Affymetrix Human Genome U133A Array	AML (37), CLL (41), CML (22), MDS (10), B-ALL (17), T-ALL (12)	22 283
LK2	Affymetrix Human Genome U133B Array	AML (37), CLL (41), CML (22), MDS (10), B-ALL (17), T-ALL (13)	22 645
AD1	Illumina HumanHT-12 V3.0 expression beadchip	AD (145), CTL (104)	38 323
AD2	Illumina HumanHT-12 V4.0 expression beadchip	AD (140), CTL (135)	32 049
AD3	Rosetta/Merck Human 44k 1.1 microarray	AD (310), CTL (157)	38 734
AD4	Rosetta/Merck Human 44k 1.1 microarray	AD (129), CTL (101)	39 005
PD1	Invitrogen ProtoArray v5.0	PD (132), CTL (156)	9 480
PD2	Invitrogen ProtoArray v5.0	PD (174), CTL (80)	9 480

CTL	Control	AML	Acute Myeloid Leukemia	AD	Alzheimer's disease
LC	Lung Cancer	CLL	Chronic Lymphocytic Leukemia	PD	Parkinson's disease
SCLC	Small Cell Lung Cancer	CML	Chronic Myeloid Leukemia		
PSO	Psoriasis	MDS	Myelodysplastic Syndrome		
PMEL	Primary Melanoma	B-ALL	B-cell Acute Lymphoblastic Leukemia		
MMEL	Metastatic Melanoma	T-ALL	T-cell Acute Lymphoblastic Leukemia		

Datasets were downloaded from **Gene Expression Omnibus**.

Methods

Pre-processing and Data Integration

Pre-processing

- ▶ $\log_2$ transformation to reduce skewness
- ▶ z-score normalization of each sample to correct for batch effects

Data Integration

- ▶ genes were matched using GenBank accession numbers
- ▶ samples from datasets related to the same disease were simply stacked together, maintaining only the common genes and matching their labels

Table 2: Number of features and samples of the datasets after the integration

Dataset	#Features	Classes (#samples)
LC1+LC2	54 675	Lung Cancer (83), Control (102)
PSO1+PSO2	54 675	Psoriasis (119), Control (85)
SK1+SK2	19 819	Primary Melanoma (77), Metastatic Melanoma (85), Control (23)
LK1+LK2	168	Acute Myeloid Leukemia (74), Chronic Lymphocytic Leukemia (82), Chronic Myeloid Leukemia (44), Myelodys-plastic Syndrome (20), Precursor B-cell Acute Lymphoblastic Leukemia (34), T-cell Acute Lymphoblastic Leukemia (25)
AD1+AD2	22 614	Alzheimer's disease (285), Control (239)
AD3+AD4	38 734	Alzheimer's disease (439), Control (258)
AD1+AD2+AD3+AD4	224	Alzheimer's disease (724), Control (497)
PD1+PD2	9 480	Parkinson's disease (306), Control (236)

Methods
Models

- ▶ Nearest Centroid (NC)
- ▶ K-Nearest Neighbors (KNN)
- ▶ Support Vector Machine (SVM)
- ▶ Gaussian Naive Bayes (GNB)
- ▶ Random Forest (RF)
- ▶ Extreme Gradient Boosting (XGB)
- ▶ **Rank Aggregation Classifier (RAC)**
  - ▶ Uses a ranked representation of features and rank aggregation methods to compute class centroids
  - ▶ Robust to noise

Methods
Rank Aggregation Classifier

Figure 1: Training and Classification of the Rank Aggregation Classifier

Methods
Model evaluation

Measurement of classification performance

- ▶ F1 score weighted by the number of samples per class
- ▶ Area Under ROC curve

Integration strategies

- ▶ Initial dataset merging followed by splitting into train and test folds
    - ▶ two nested loops of 5-fold cross validation (inner loop to select the best hyperparameters; outer loop to evaluate the model with the best hyperparameters on the remaining set of data)
- ▶ Use of one dataset for training and a separate one for testing
    - ▶ 5-fold cross-validation to selected the best hyperparameters on the training dataset
    - ▶ test of the best hyperparameter configuration on the test dataset

Methods
Model evaluation

Feature ranking evaluation

▶ Comparison of SVM, RF, XGB and RAC after having selected the most relevant features using the **Recursive Feature Elimination (RFE)** algorithm

▶ 5-fold cross validation to validate results (the feature elimination process was repeated for each training fold, testing the classifiers with the reduced feature subset on the complementary test fold)

Results

Performance on merged datasets

Table 3: F1 scores from nested cross validation on merged datasets

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	0.960	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

'*' for z-score normalized data
cells colored by score

▶ Good performance albeit slightly reduced compared to individual dataset
▶
▶

Results

Performance on merged datasets

Table 4: F1 scores from nested cross validation on merged datasets

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	0.960	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

'*' for z-score normalized data
cells colored by score on each row independently

▶ Good performance albeit slightly reduced compared to individual dataset
▶ **SVM, RF and XGB perform better than the other classification methods**
▶

Results

Performance on merged datasets

Table 5: F1 scores from nested cross validation on merged datasets

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

'*' for z-score normalized data
cells colored by the difference between z-score normalized and non normalized scores
(blue if performance increases, red if it decreases)

- ▶ Good performance albeit slightly reduced compared to individual dataset
- ▶ SVM, RF and XGB perform better than the other classification methods
- ▶ **The impact of z-score normalization varies depending on the model and on the dataset**

Results

Performance when training and testing on different datasets

Table 6: F1 scores when training and testing on different datasets

Training dataset	Test dataset	RAC		NC		KNN		SVM		GNB		RF		XGB	
		Train score	Test score	Train score	Test score	Train score	Test score	Train score	Test score	Train score	Test score	Train score	Test score	Train score	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	0.657	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

'*' for z-score normalized data
cells colored by score

► Performance varies across models and datasets   ►
►   ►

Results

Performance when training and testing on different datasets

Table 7: F1 scores when training and testing on different datasets

Training dataset	Test dataset	RAC Train score	RAC Test score	NC Train score	NC Test score	KNN Train score	KNN Test score	SVM Train score	SVM Test score	GNB Train score	GNB Test score	RF Train score	RF Test score	XGB Train 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	0.985	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	0.683	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

'*' for z-score normalized data
cells colored by score on each row independently

- Performance varies across models and datasets
- No method stands out

Performance when training and testing on different datasets

Table 8: F1 scores when training and testing on different datasets

Training dataset	Test dataset	RAC Train score	RAC Test score	NC Train score	NC Test score	KNN Train score	KNN Test score	SVM Train score	SVM Test score	GNB Train score	GNB Test score	RF Train score	RF Test score	XGB Train 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	0.702	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	0.946	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	0.988	0.708	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

'*' for z-score normalized data
cells colored by the difference between train scores and test scores
(blue if test performance increases, red if it decreases)

- Performance varies across models and datasets
- No method stands out

- Often test scores are significantly lower than train scores

Results

Performance when training and testing on different datasets

Table 9: F1 scores when training and testing on different datasets

Training dataset	Test dataset	RAC Train score	RAC Test score	NC Train score	NC Test score	KNN Train score	KNN Test score	SVM Train score	SVM Test score	GNB Train score	GNB Test score	RF Train score	RF Test score	XGB Train 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	0.879	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	3	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	4	3	

'*' for z-score normalized data
cells colored by the difference between z-score normalized and non normalized scores
(blue if performance increases, red if it decreases)

- Performance varies across models and datasets
- No method stands out

- Often test scores are significantly lower than train scores
- Again, the impact of z-score normalization varies depending on the model and on the dataset
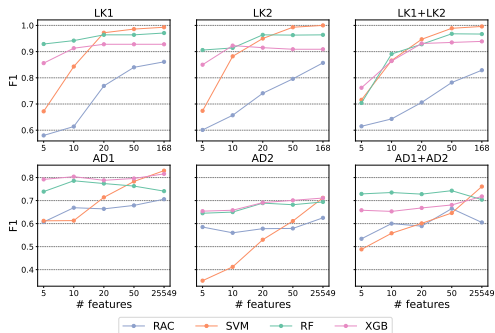
Results
The effect of feature selection

Figure 2: F1 scores when varying the number of features selected by RFE

▶ RFE typically does not improve the performance of classifiers
▶ Using 50 features the performance is already quite close to that obtained when using all features
▶ As opposed to RAC and SVM, RF and XGB still perform well with fewer features
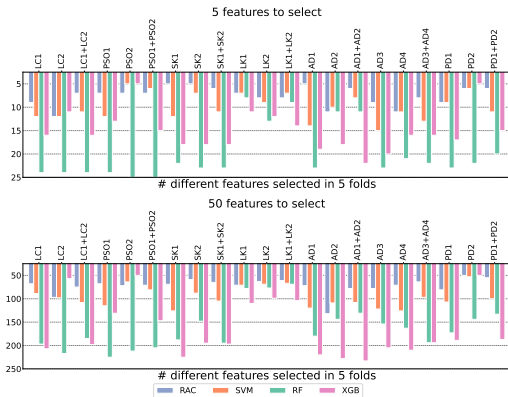
Results

Robustness in feature ranking

Figure 3: Total number of features selected by RFE on the five folds of cross validation

▶ RFE with XGB and RF is not robust to the change of the underlying samples (it tends to select different features over different folds)

Results

Feature relevance

Table 10: Number of genes in the literature related to Psoriasis among the 5 genes selected by RFE

Dataset	RAC	SVM	RF	XGB
PSO1	5	5	2	2
PSO2	5	5	4	1
PSO1+PSO2	5	5	4	3

- ▶ All the features selected by RAC and SVM are supported by the literature
- ▶ The features selected by XGB are those less supported by the literature
- ▶ The number of features without support decreases when moving on to the integrated dataset

Conclusion

- ▶ Training on one dataset and testing on another may not produce a good classification
  - ▶ The high variability between models indicates that we cannot a-priori choose a best model
- ▶ Merging multiple datasets produces stable results
  - ▶ Re-training models as new data becomes available is needed to insure good performance on new data
  - ▶ SVM is the clear best classifier choice, followed closely by RF and XGB
- ▶ The features selected by RF and XGB vary highly from one cross validation fold to another, indicating limited robustness to changes in the underlying samples

Thank you!

Irene Testa
i.testa@studenti.unipi.it

Scan the QR code to visit the GitHub repository