FeCO3, constructing the network biomarkers using the inter-feature correlation coefficients

Shenggeng Lin[1,2,*], Yuqi Lin[1,*], Kexin Wu[1,*], Yueying Wang[3], Zixuan Feng[1], Meiyu Duan[1], Shuai Liu[1], Jingxuan Huang[4], Zhao Wang[4], Yusi Fan[4], Lan Huang[1], Fengfeng Zhou[1,#].

[1] College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China.

[2] State Key Laboratory of Microbial Metabolism, and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.

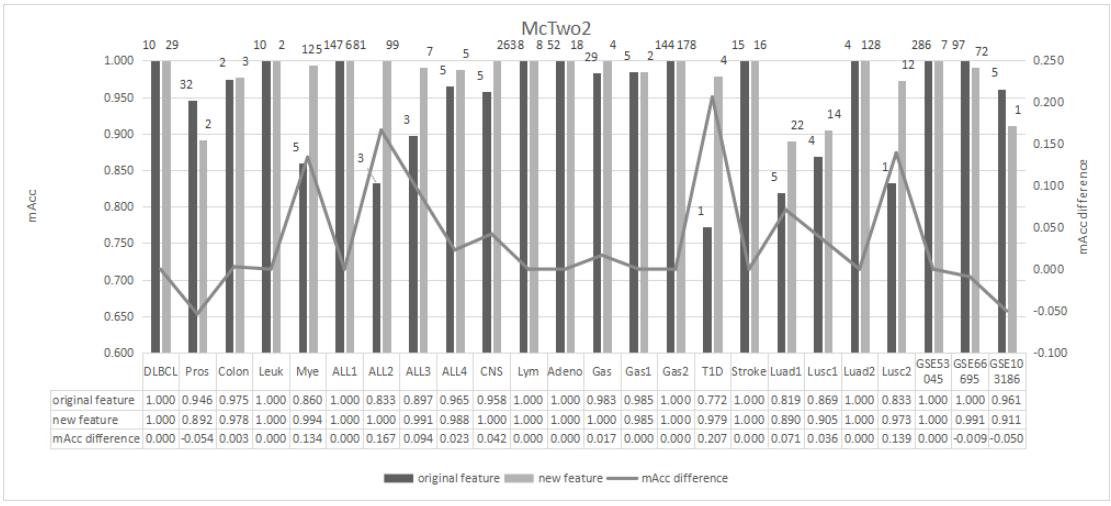
[3] Department of Epidemiology and Biostatistics, School of Public Health, Jilin University, Changchun, Jilin Province, China.

[4] College of Software, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China.

* Contributed equally to this study.

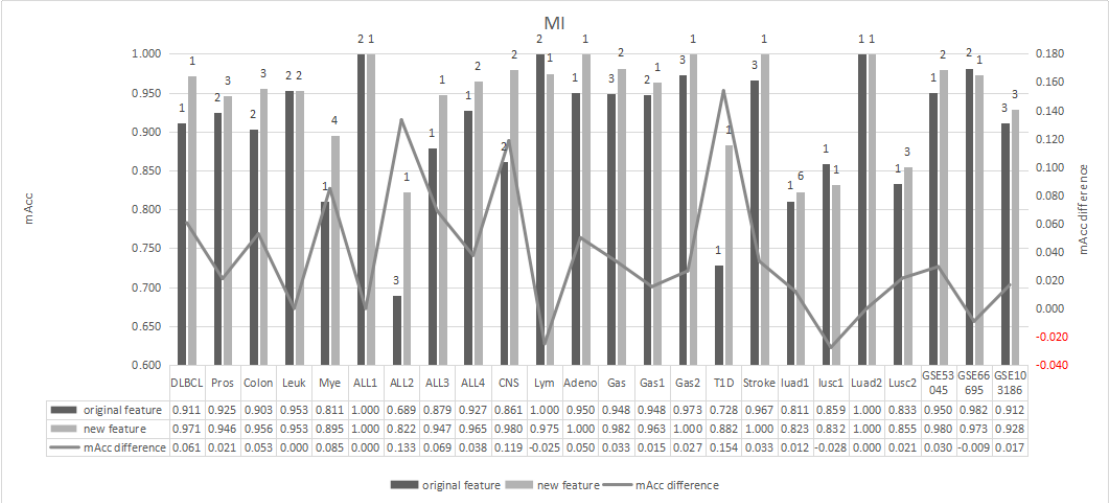
Correspondence may be addressed to Fengfeng Zhou: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn . Lab web site: http://www.healthinformaticslab.org/ . Phone: +86-431-8516-6024. Fax: +86-431-8516-6024.

Supplementary Figure S1



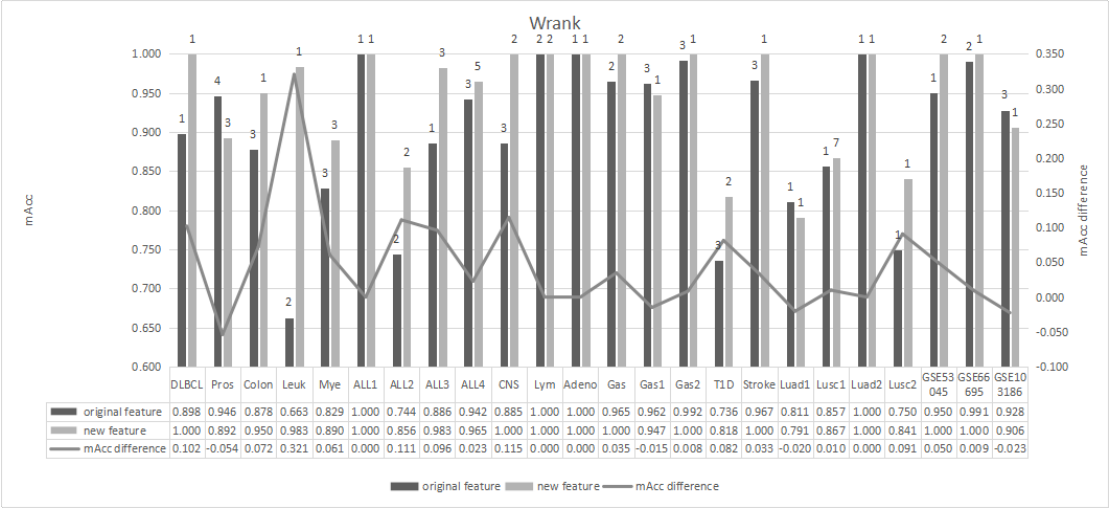
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE5345	GSE6695	GSE103186
original feature	1.000	0.946	0.975	1.000	0.860	1.000	0.833	0.897	0.965	0.958	1.000	1.000	0.983	0.985	1.000	0.772	1.000	0.819	0.869	1.000	0.833	1.000	1.000	0.961
new feature	1.000	0.892	0.978	1.000	0.994	1.000	1.000	0.991	0.988	1.000	1.000	1.000	1.000	0.985	1.000	0.979	1.000	0.890	0.905	1.000	0.973	1.000	0.991	0.911
mAcc difference	0.000	-0.054	0.003	0.000	0.134	0.000	0.167	0.094	0.023	0.042	0.000	0.000	0.017	0.000	0.000	0.207	0.000	0.071	0.036	0.000	0.139	0.000	-0.009	-0.050

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm McTwo2. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



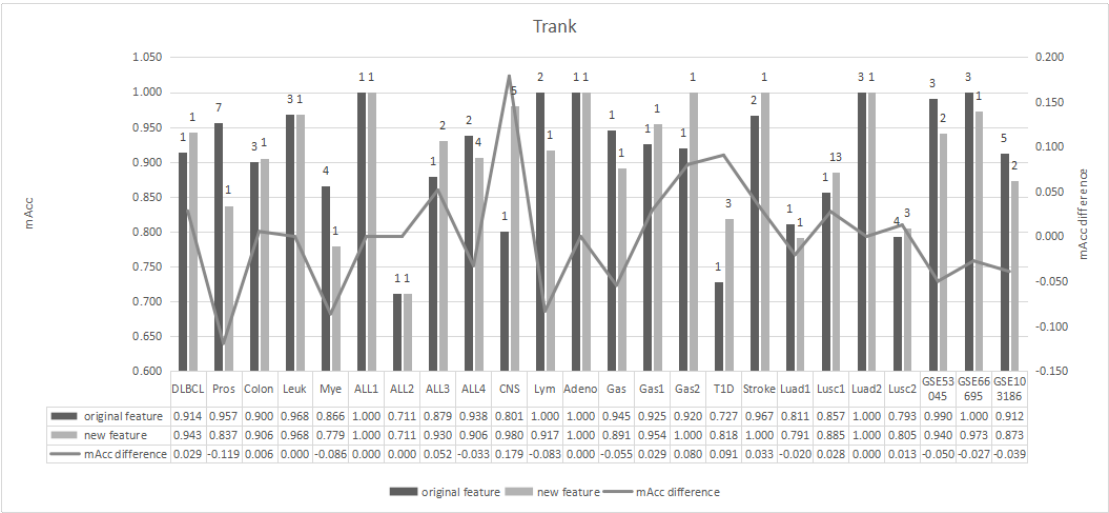
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	luad1	lusc1	Luad2	Lusc2	GSE5345	GSE6695	GSE103186
original feature	0.911	0.925	0.903	0.953	0.811	1.000	0.689	0.879	0.927	0.861	1.000	0.950	0.948	0.948	0.973	0.728	0.967	0.811	0.859	1.000	0.833	0.950	0.982	0.912
new feature	0.971	0.946	0.956	0.953	0.895	1.000	0.822	0.947	0.965	0.980	0.975	1.000	0.982	0.963	1.000	0.882	1.000	0.823	0.832	1.000	0.855	0.980	0.973	0.928
mAcc difference	0.061	0.021	0.053	0.000	0.085	0.000	0.133	0.069	0.038	0.119	-0.025	0.050	0.033	0.015	0.027	0.154	0.033	0.012	-0.028	0.000	0.021	0.030	-0.009	0.017

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm MI. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



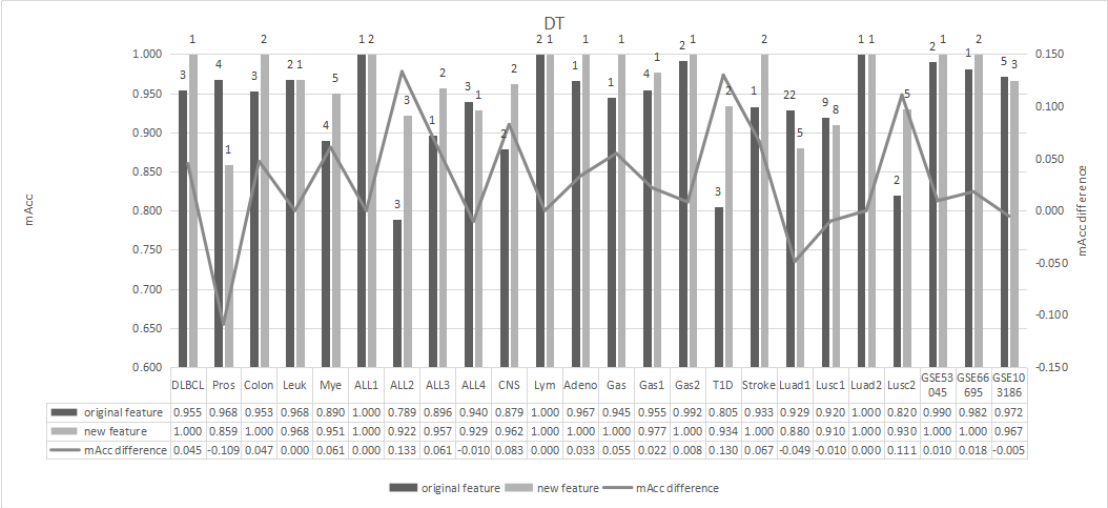
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.898	0.946	0.878	0.663	0.829	1.000	0.744	0.886	0.942	0.885	1.000	1.000	0.965	0.962	0.992	0.736	0.967	0.811	0.857	1.000	0.750	0.950	0.991	0.928
new feature	1.000	0.892	0.950	0.983	0.890	1.000	0.856	0.983	0.965	1.000	1.000	1.000	1.000	0.947	1.000	0.818	1.000	0.791	0.867	1.000	0.841	1.000	1.000	0.906
mAcc difference	0.102	-0.054	0.072	0.321	0.061	0.000	0.111	0.096	0.023	0.115	0.000	0.000	0.035	-0.015	0.008	0.082	0.033	-0.020	0.010	0.000	0.091	0.050	0.009	-0.023

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm Wrank. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



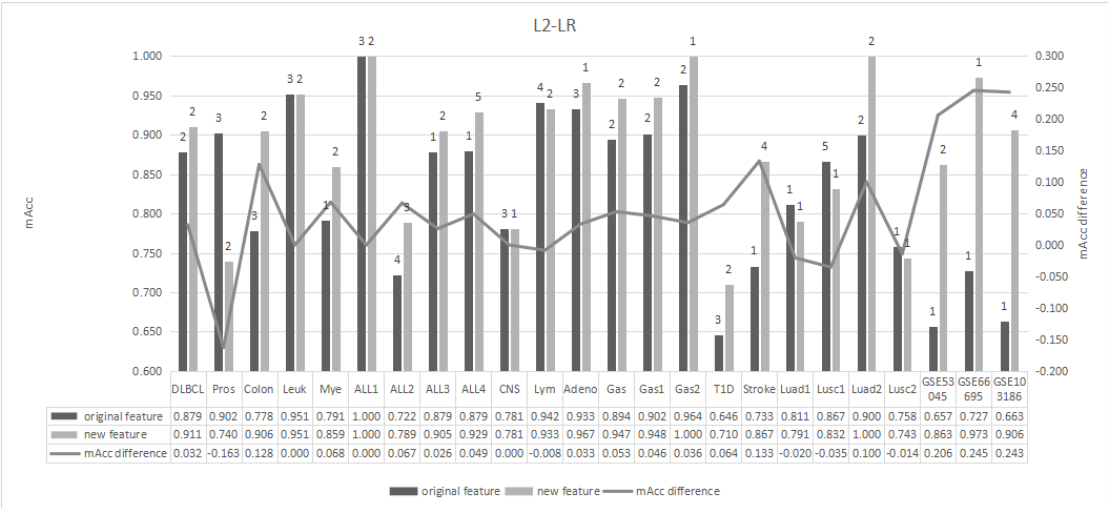
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.914	0.957	0.900	0.968	0.866	1.000	0.711	0.879	0.938	0.801	1.000	1.000	0.945	0.925	0.920	0.727	0.967	0.811	0.857	1.000	0.793	0.990	1.000	0.912
new feature	0.943	0.837	0.906	0.968	0.779	1.000	0.711	0.930	0.906	0.980	0.917	1.000	0.891	0.954	1.000	0.818	1.000	0.791	0.885	1.000	0.805	0.940	0.973	0.873
mAcc difference	0.029	-0.119	0.006	0.000	-0.086	0.000	0.000	0.052	-0.033	0.179	-0.083	0.000	-0.055	0.029	0.080	0.091	0.033	-0.020	0.028	0.000	0.013	-0.050	-0.027	-0.039

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm Trank. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



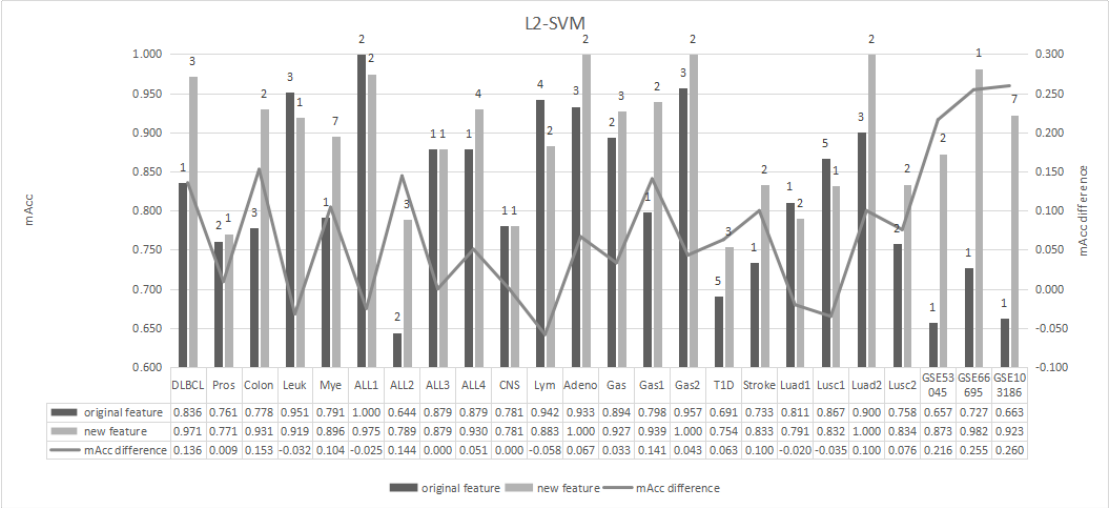
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.955	0.968	0.953	0.968	0.890	1.000	0.789	0.896	0.940	0.879	1.000	0.967	0.945	0.955	0.992	0.805	0.933	0.929	0.920	1.000	0.820	0.990	0.982	0.972
new feature	1.000	0.859	1.000	0.968	0.951	1.000	0.922	0.957	0.929	0.962	1.000	1.000	1.000	0.977	1.000	0.934	1.000	0.880	0.910	1.000	0.930	1.000	1.000	0.967
mAcc difference	0.045	-0.109	0.047	0.000	0.061	0.000	0.133	0.061	-0.010	0.083	0.000	0.033	0.055	0.022	0.008	0.130	0.067	-0.049	-0.010	0.000	0.111	0.010	0.018	-0.005

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm DT. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



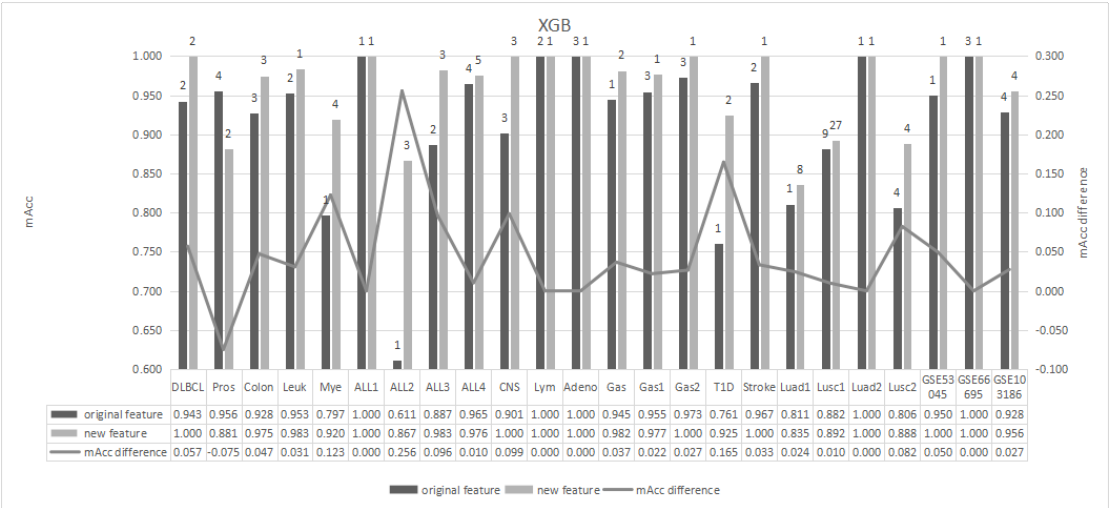
	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.879	0.902	0.778	0.951	0.791	1.000	0.722	0.879	0.879	0.781	0.942	0.933	0.894	0.902	0.964	0.646	0.733	0.811	0.867	0.900	0.758	0.657	0.727	0.663
new feature	0.911	0.740	0.906	0.951	0.859	1.000	0.789	0.905	0.929	0.781	0.933	0.967	0.947	0.948	1.000	0.710	0.867	0.791	0.832	1.000	0.743	0.863	0.973	0.906
mAcc difference	0.032	-0.163	0.128	0.000	0.068	0.000	0.067	0.026	0.049	0.000	-0.008	0.033	0.053	0.046	0.036	0.064	0.133	-0.020	-0.035	0.100	-0.014	0.206	0.245	0.243

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm L2-LR. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.836	0.761	0.778	0.951	0.791	1.000	0.644	0.879	0.879	0.781	0.942	0.933	0.894	0.798	0.957	0.691	0.733	0.811	0.867	0.900	0.758	0.657	0.727	0.663
new feature	0.971	0.771	0.931	0.919	0.896	0.975	0.789	0.879	0.930	0.781	0.883	1.000	0.927	0.939	1.000	0.754	0.833	0.791	0.832	1.000	0.834	0.873	0.982	0.923
mAcc difference	0.136	0.009	0.153	-0.032	0.104	-0.025	0.144	0.000	0.051	0.000	-0.058	0.067	0.033	0.141	0.043	0.063	0.100	-0.020	-0.035	0.100	0.076	0.216	0.255	0.260

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm L2-SVM. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms



	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	Luad1	Lusc1	Luad2	Lusc2	GSE53045	GSE6695	GSE103186
original feature	0.943	0.956	0.928	0.953	0.797	1.000	0.611	0.887	0.965	0.901	1.000	1.000	0.945	0.955	0.973	0.761	0.967	0.811	0.882	1.000	0.806	0.950	1.000	0.928
new feature	1.000	0.881	0.975	0.983	0.920	1.000	0.867	0.983	0.976	1.000	1.000	1.000	0.982	0.977	1.000	0.925	1.000	0.835	0.892	1.000	0.888	1.000	1.000	0.956
mAcc difference	0.057	-0.075	0.047	0.031	0.123	0.000	0.256	0.096	0.010	0.099	0.000	0.000	0.037	0.022	0.027	0.165	0.033	0.024	0.010	0.000	0.082	0.050	0.000	0.027

Comparison of the classification metric mAcc and the number of features between the original features and the FeCO3 features using the feature selection algorithm XGB. The series “original” gives the data of the original features, and the series “FeCO3” illustrates the status of the FeCO3 features. The series “Improvement” is the mAcc value of the FeCO3 features minus the mAcc value of the original features. The classification accuracy is calculated using the 5-fold cross validation strategy. The data was calculated using the feature selection algorithms

References

Cherkassky, V. The nature of statistical learning theory~. *IEEE Trans Neural Netw* 1997;8(6):1564.

Ge, R., *et al.* McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics* 2016;17:142.

Geva, S. and Sitte, J. Adaptive nearest neighbor pattern classification. *IEEE Trans Neural Netw* 1991;2(2):318-322.

Guo, P., *et al.* Gene expression profile based classification models of psoriasis. *Genomics* 2014;103(1):48-55.

Haridas, V., *et al.* TRANK, a novel cytokine that activates NF-kappa B and c-Jun N-terminal kinase. *J Immunol* 1998;161(1):1-6.

Li, Y., *et al.* Association of miR-155 and Angiotensin Receptor Type 1 Polymorphisms with the Risk of Ischemic Stroke in a Chinese Population. *DNA Cell Biol* 2019.

Liu, W.M., *et al.* Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 2002;18(12):1593-1599.

Mortazavi, A. and Moattar, M.H. Robust Feature Selection from Microarray Data Based on Cooperative Game Theory and Qualitative Mutual Information. *Adv Bioinformatics* 2016;2016:1058305.

Nguyen, D.H. and Patrick, J.D. Supervised machine learning and active learning in classification of radiology reports. *J Am Med Inform Assoc* 2014;21(5):893-901.

Pal, M. Random forest classifier for remote sensing classification. *International journal of remote sensing* 2005;26(1):217-222.

Pregibon, D. Logistic regression diagnostics. *The Annals of Statistics* 1981;9(4):705-724.

Senders, J.T., *et al.* Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. *JCO Clin Cancer Inform* 2019;3:1-9.

Waldmann, P. On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction. *Front Genet* 2019;10:899.

Wang, Y., *et al.* In Silico Prediction of Human Intravenous Pharmacokinetic Parameters with Improved Accuracy. *J Chem Inf Model* 2019;59(9):3968-3980.

Ye, Y., *et al.* RIFS: a randomly restarted incremental feature selection algorithm. *Sci Rep* 2017;7(1):13013.

Yu, X., *et al.* Individual-specific edge-network analysis for disease prediction. *Nucleic Acids Res* 2017;45(20):e170.