An Effective Feature Selection Method for Imbalance Datasets. An Application to the Chemical Toxicity Prediction (Supplemental Material)

Aurelio Antelo Collado,[†] Ramón Carrasco Velar,[†] Nicolás García-Pedrajas,[‡] and Gonzalo Cerruela-García[*,‡]

[†]*University of Informatics Science. Cheminformatic Group. Havana, Cuba.*

[‡]*University of Córdoba. Department of Computing and Numerical Analysis. Campus de Rabanales. Albert Einstein Building. E-14071 Córdoba, Spain.*

E-mail: gcerruela@uco.es

Table S 1: Hyperparameter values of all machine learning models

RF		
Parameter	Value	Description
n_estimators	100	The number of trees in the forest
criterion	gini	The function to measure the quality of a split, gini" for the Gini impurity
min_samples_leaf	1	The minimum number of samples required to be at a leaf node
max_depth	none	None: the nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples
min_samples_split	2	The minimum number of samples required to split an internal node
max_leaf_nodes	none	Grow trees with max_leaf_nodes in best-first fashion. none=unlimited number of leaf nodes
min_impurity_decrease	0	A node will be split if this split induces a decrease of the impurity greater than or equal to this value
min_weight_fraction_leaf	0	The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node
bootstrap	TRUE	Bootstrap samples are used when building trees
SVM		
Parameters	For SVMs, three hyperparameters were set: the kernel type, the C value, and for the Gaussian kernel, the γ value. Thus, we tested a linear kernel with $C \in \{0.1, 1, 10\}$ and a Gaussian kernel with $C \in \{0.1, 1, 10\}$ and $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$. All 21 possible combinations were evaluated	
DT, FAST, FCBF, Undersampling, SMOTE, CSB2.X, MadaBoost.X		
Parameter	No relevant parameters	

Table S 2: Results for DT classifier using ECFP4 representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9084	0.4302	0.6505	0.9639	0.4967	0.6939	0.9619	0.5230	0.6956
DS2	0.9082	0.3946	0.6690	0.9572	0.4186	0.7196	0.9674	0.4252	0.7255
DS3	0.9064	0.2362	0.6219	0.9754	0.2534	0.6659	0.9764	0.2546	0.6616
DS4	0.9127	0.1156	0.5302	0.9895	0.1300	0.5873	0.9865	0.1286	0.5842
DS5	0.9062	0.1258	0.5330	0.8025	0.1680	0.6143	0.8494	0.1635	0.6178
DS6	0.9119	0.2183	0.6007	0.9318	0.2281	0.6441	0.9500	0.2407	0.6557
DS7	0.9174	0.0763	0.5763	0.8535	0.0982	0.6409	0.8512	0.0918	0.6326
DS8	0.9070	0.1657	0.5117	0.8650	0.1471	0.5644	0.9561	0.1488	0.5595
DS9	0.9125	0.0986	0.5945	0.8355	0.1187	0.6503	0.8447	0.1042	0.6314
DS10	0.9150	0.0808	0.5519	0.8207	0.0878	0.5918	0.8285	0.0828	0.5868
DS11	0.9074	0.2279	0.5676	0.9854	0.2438	0.6097	0.9859	0.2402	0.6093
DS12	0.9074	0.1164	0.5196	0.8170	0.1046	0.5895	0.9529	0.1191	0.5625

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9596	0.4243	0.6897	0.9965	0.4053	0.6901	0.9848	0.3620	0.6782
DS2	0.9486	0.4046	0.7408	0.9969	0.4112	0.7303	0.9615	0.3804	0.7162
DS3	0.9773	0.2204	0.5795	0.9611	0.2329	0.6272	0.8824	0.2378	0.6165
DS4	0.9586	0.1274	0.5750	0.9596	0.1388	0.6179	0.9607	0.1246	0.6050
DS5	0.9787	0.1994	0.5115	0.8152	0.1623	0.6069	0.6850	0.1689	0.6131
DS6	0.9582	0.3012	0.6152	0.8953	0.1943	0.6642	0.8934	0.1627	0.6278
DS7	0.9477	0.0623	0.5800	0.8307	0.0961	0.6401	0.8410	0.0791	0.6133
DS8	0.9811	0.1518	0.5601	0.9260	0.1511	0.5825	0.9467	0.1709	0.5713
DS9	0.9596	0.1038	0.6270	0.8377	0.0989	0.6253	0.8426	0.1216	0.6533
DS10	0.9600	0.0712	0.5415	0.8068	0.0735	0.5770	0.8123	0.0901	0.5917
DS11	0.9830	0.2349	0.5967	0.9932	0.2334	0.5987	0.9926	0.2321	0.6057
DS12	0.9682	0.1101	0.5240	0.8854	0.1276	0.6120	0.8791	0.1126	0.5853

Table S 3: Results for SVM classifier using ECFP4 representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9084	0.2952	0.6529	0.8896	0.2942	0.7129	0.8494	0.3006	0.7074
DS2	0.9082	0.3234	0.7273	0.9480	0.3646	0.7319	0.8420	0.3637	0.7435
DS3	0.9064	0.3477	0.7181	0.9201	0.3769	0.7725	0.9055	0.3592	0.7608
DS4	0.9127	0.1505	0.6207	0.9299	0.1570	0.6631	0.8695	0.1567	0.6494
DS5	0.9062	0.1920	0.5961	0.7908	0.2378	0.6606	0.8004	0.2164	0.6414
DS6	0.9119	0.1720	0.6182	0.8994	0.1927	0.6823	0.8545	0.2108	0.6649
DS7	0.9174	0.1010	0.6056	0.9432	0.0750	0.6073	0.8623	0.1428	0.6685
DS8	0.9070	0.2361	0.6167	0.8592	0.2718	0.6733	0.8545	0.2397	0.6553
DS9	0.9125	0.1310	0.6292	0.9252	0.1425	0.6696	0.8633	0.1759	0.6964
DS10	0.9150	0.1099	0.5577	0.9170	0.1027	0.5966	0.8285	0.1435	0.6328
DS11	0.9074	0.3375	0.6847	0.9174	0.4079	0.7635	0.8781	0.3706	0.7317
DS12	0.9074	0.1724	0.6297	0.9229	0.1842	0.6804	0.8773	0.1795	0.6686

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9596	0.2149	0.6784	0.8330	0.2021	0.6808	0.7951	0.2285	0.6960
DS2	0.9486	0.2367	0.7438	0.8664	0.2511	0.7668	0.8586	0.2118	0.7467
DS3	0.9773	0.2673	0.6849	0.8148	0.3404	0.7452	0.8287	0.3512	0.7533
DS4	0.9586	0.1297	0.6183	0.8270	0.1636	0.6606	0.8279	0.1573	0.6577
DS5	0.9787	0.1565	0.5484	0.6654	0.2056	0.6382	0.6740	0.2223	0.6468
DS6	0.9582	0.1867	0.6520	0.8002	0.1956	0.6794	0.7582	0.1772	0.6755
DS7	0.9477	0.0662	0.5784	0.8773	0.1372	0.6732	0.8432	0.1089	0.6437
DS8	0.9811	0.1456	0.5838	0.7271	0.2433	0.6565	0.7256	0.2364	0.6501
DS9	0.9596	0.1076	0.6294	0.8584	0.1411	0.6689	0.8150	0.1248	0.6411
DS10	0.9600	0.0733	0.5419	0.7629	0.1211	0.6069	0.7574	0.1280	0.6285
DS11	0.9830	0.3064	0.6817	0.7867	0.3482	0.7234	0.8262	0.3641	0.7340
DS12	0.9682	0.1223	0.5773	0.8252	0.1833	0.6781	0.7926	0.1754	0.6586

Table S 4: Results for RF classifier using ECFP4 representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9084	0.1933	0.6616	0.9639	0.2405	0.6975	0.9619	0.2177	0.7008
DS2	0.9082	0.2783	0.7235	0.9572	0.2781	0.7327	0.9674	0.3007	0.7299
DS3	0.9064	0.3480	0.7200	0.9754	0.3222	0.7324	0.9764	0.2992	0.7209
DS4	0.9127	0.1598	0.6284	0.9895	0.1253	0.5801	0.9865	0.1112	0.5811
DS5	0.9062	0.1614	0.5820	0.8025	0.1936	0.6396	0.8494	0.1806	0.6305
DS6	0.9119	0.1684	0.6303	0.9318	0.2229	0.6928	0.9500	0.2130	0.6749
DS7	0.9174	0.1050	0.6102	0.8535	0.1456	0.6898	0.8512	0.1293	0.6768
DS8	0.9070	0.2196	0.6067	0.8650	0.1775	0.6009	0.9561	0.1825	0.6006
DS9	0.9125	0.1467	0.6458	0.8355	0.1810	0.7146	0.8447	0.1635	0.6971
DS10	0.9150	0.1004	0.5705	0.8207	0.1561	0.6592	0.8285	0.1101	0.6110
DS11	0.9074	0.3635	0.7004	0.9854	0.3129	0.6917	0.9859	0.3174	0.7053
DS12	0.9074	0.1789	0.6340	0.8170	0.1814	0.6550	0.9529	0.1460	0.6128

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9596	0.2303	0.6949	0.9965	0.3858	0.6905	0.9848	0.3245	0.6723
DS2	0.9486	0.2564	0.7611	0.9969	0.3763	0.7306	0.9615	0.3295	0.7456
DS3	0.9773	0.2778	0.6879	0.9611	0.2745	0.6890	0.8824	0.2861	0.6611
DS4	0.9586	0.1489	0.6410	0.9596	0.1366	0.6192	0.9607	0.1385	0.6167
DS5	0.9787	0.1640	0.5438	0.8152	0.1844	0.6244	0.6850	0.1930	0.6331
DS6	0.9582	0.1943	0.6505	0.8953	0.2147	0.6807	0.8934	0.1873	0.6449
DS7	0.9477	0.0808	0.5880	0.8307	0.1124	0.6554	0.8410	0.0912	0.6324
DS8	0.9811	0.1534	0.5817	0.9260	0.1604	0.5845	0.9467	0.1855	0.6006
DS9	0.9596	0.1174	0.6424	0.8377	0.1493	0.6795	0.8426	0.1404	0.6749
DS10	0.9600	0.0782	0.5363	0.8068	0.1063	0.6101	0.8123	0.1122	0.6100
DS11	0.9830	0.3096	0.6821	0.9932	0.2734	0.6266	0.9926	0.2783	0.6625
DS12	0.9682	0.1266	0.5816	0.8854	0.1472	0.6296	0.8791	0.1457	0.6137

Table S 5: Results for DT classifier using GSFrag representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9991	0.3605	0.6499	0.9933	0.2667	0.7033	0.9900	0.3123	0.7159
DS2	0.9991	0.3480	0.7165	0.9877	0.3384	0.7570	0.9954	0.3822	0.7543
DS3	0.9302	0.2883	0.6756	0.9615	0.2911	0.7057	0.9492	0.3103	0.7259
DS4	0.9337	0.1704	0.6427	0.9603	0.1796	0.6861	0.9506	0.1676	0.6704
DS5	0.9988	0.1873	0.5948	0.9907	0.2058	0.6314	0.9185	0.2257	0.6436
DS6	0.9793	0.1251	0.5921	0.8861	0.1582	0.6618	0.9520	0.1749	0.6679
DS7	0.9098	0.0879	0.5981	0.8698	0.0873	0.6267	0.9093	0.0887	0.6257
DS8	0.9981	0.1738	0.5802	0.9009	0.1988	0.6296	0.9677	0.1959	0.6282
DS9	0.9123	0.0939	0.5772	0.8764	0.1097	0.6379	0.8564	0.1229	0.6535
DS10	0.9105	0.0892	0.5503	0.8497	0.0926	0.5964	0.8963	0.1237	0.6246
DS11	0.9986	0.2725	0.6259	0.9947	0.2705	0.6750	0.9865	0.2742	0.6722
DS12	0.9812	0.1598	0.6279	0.8578	0.1977	0.6937	0.9250	0.1734	0.6709

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9946	0.3615	0.7106	0.9931	0.2931	0.7175	0.9858	0.3587	0.7190
DS2	0.9947	0.3705	0.7442	0.9851	0.3533	0.7602	0.9678	0.3380	0.7584
DS3	0.9937	0.2244	0.6265	0.9100	0.3151	0.7276	0.9387	0.3199	0.7327
DS4	0.9938	0.1837	0.6900	0.9851	0.1801	0.6876	0.9824	0.1931	0.7028
DS5	0.9935	0.2023	0.6311	0.9351	0.2113	0.6241	0.9685	0.1952	0.6246
DS6	0.9942	0.1123	0.5885	0.9515	0.1968	0.6678	0.9402	0.2053	0.6706
DS7	0.9937	0.0468	0.5354	0.9504	0.0736	0.6066	0.9450	0.0887	0.6242
DS8	0.9933	0.2035	0.6301	0.8984	0.2148	0.6403	0.9446	0.2035	0.6338
DS9	0.9937	0.0975	0.6128	0.9200	0.1005	0.6269	0.9144	0.1224	0.6561
DS10	0.9937	0.0709	0.5039	0.9195	0.0961	0.6003	0.8996	0.1135	0.6094
DS11	0.9926	0.2897	0.6749	0.9786	0.3080	0.6930	0.9357	0.2917	0.6853
DS12	0.9917	0.1801	0.6776	0.9571	0.1771	0.6759	0.9582	0.2039	0.7014

Table S 6: Results for SVM classifier using GSFrag representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9991	0.3438	0.6502	0.8619	0.3128	0.7291	0.9292	0.3310	0.7126
DS2	0.9991	0.3491	0.7183	0.8692	0.3522	0.7522	0.9587	0.3591	0.7669
DS3	0.9302	0.3324	0.7038	0.8513	0.3803	0.7691	0.8822	0.3802	0.7709
DS4	0.9337	0.1865	0.6604	0.8838	0.2051	0.7162	0.8652	0.2000	0.7043
DS5	0.9988	0.1780	0.5834	0.8540	0.2430	0.6581	0.8402	0.2277	0.6500
DS6	0.9793	0.1353	0.6037	0.8634	0.1904	0.6848	0.8726	0.1987	0.6823
DS7	0.9098	0.1111	0.6226	0.8545	0.1021	0.6359	0.8170	0.1236	0.6722
DS8	0.9981	0.1718	0.5799	0.8573	0.2488	0.6590	0.8610	0.2535	0.6667
DS9	0.9123	0.1150	0.5985	0.8295	0.1288	0.6538	0.8141	0.1194	0.6431
DS10	0.9105	0.1221	0.5853	0.7965	0.1450	0.6292	0.7815	0.1408	0.6160
DS11	0.9986	0.2558	0.6284	0.8768	0.3923	0.7507	0.9032	0.3588	0.7329
DS12	0.9812	0.1789	0.6424	0.8380	0.2098	0.6930	0.8459	0.2164	0.7017

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9946	0.3094	0.7229	0.8722	0.3102	0.7113	0.9156	0.3355	0.7259
DS2	0.9947	0.3222	0.7457	0.9158	0.3080	0.7386	0.9118	0.3444	0.7559
DS3	0.9937	0.2516	0.6851	0.9167	0.3717	0.7610	0.9318	0.3832	0.7703
DS4	0.9938	0.1772	0.6872	0.8905	0.2224	0.7247	0.9158	0.2018	0.6993
DS5	0.9935	0.1924	0.6247	0.8815	0.2002	0.6335	0.8768	0.2294	0.6515
DS6	0.9942	0.1182	0.5921	0.8903	0.2534	0.7038	0.8942	0.2144	0.6909
DS7	0.9937	0.0695	0.5823	0.8893	0.0927	0.6238	0.8900	0.1050	0.6433
DS8	0.9933	0.1952	0.6273	0.8761	0.2464	0.6620	0.8822	0.2566	0.6696
DS9	0.9937	0.0797	0.5960	0.8807	0.1168	0.6314	0.8815	0.1349	0.6612
DS10	0.9937	0.0839	0.5500	0.8780	0.1502	0.6264	0.8782	0.1377	0.6267
DS11	0.9926	0.2733	0.6759	0.8845	0.3308	0.7173	0.8889	0.3490	0.7281
DS12	0.9917	0.1796	0.6772	0.8714	0.2415	0.7251	0.9156	0.2222	0.7102

Table S 7: Results for RF classifier using GSFrag representation

Dataset	FAST			CSB2.FAST			MadaBoost.FAST		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9991	0.3554	0.6465	0.9933	0.2225	0.7158	0.9900	0.2085	0.7102
DS2	0.9991	0.3116	0.7009	0.9877	0.2534	0.7431	0.9954	0.2277	0.7395
DS3	0.9302	0.2947	0.6875	0.9615	0.3065	0.7181	0.9492	0.3343	0.7456
DS4	0.9337	0.1842	0.6676	0.9603	0.1766	0.6905	0.9506	0.1825	0.6956
DS5	0.9988	0.1704	0.5825	0.9907	0.1875	0.6283	0.9185	0.1931	0.6391
DS6	0.9793	0.1140	0.5912	0.8861	0.2003	0.7014	0.9520	0.1689	0.6741
DS7	0.9098	0.0826	0.5888	0.8698	0.0900	0.6311	0.9093	0.0716	0.6046
DS8	0.9981	0.1450	0.5647	0.9009	0.2204	0.6470	0.9677	0.1951	0.6302
DS9	0.9123	0.0918	0.5846	0.8764	0.1133	0.6422	0.8564	0.1188	0.6486
DS10	0.9105	0.0792	0.5516	0.8497	0.1272	0.6329	0.8963	0.1277	0.6290
DS11	0.9986	0.2533	0.6288	0.9947	0.2910	0.6918	0.9865	0.2780	0.6845
DS12	0.9812	0.1321	0.5996	0.8578	0.2100	0.7037	0.9250	0.1800	0.6782

Dataset	FCBF			CSB2.FCBF			MadaBoost.FCBF		
	r	MCC	G-Mean	r	MCC	G-Mean	r	MCC	G-Mean
DS1	0.9946	0.2644	0.7214	0.9931	0.2481	0.7099	0.9858	0.2585	0.7374
DS2	0.9947	0.2184	0.7260	0.9851	0.2359	0.7501	0.9678	0.2247	0.7308
DS3	0.9937	0.2326	0.6720	0.9100	0.3676	0.7660	0.9387	0.3676	0.7654
DS4	0.9938	0.1756	0.6882	0.9851	0.1628	0.6745	0.9824	0.1809	0.6932
DS5	0.9935	0.1818	0.6111	0.9351	0.1905	0.6356	0.9685	0.1770	0.6157
DS6	0.9942	0.1022	0.5930	0.9515	0.1904	0.6951	0.9402	0.1976	0.6928
DS7	0.9937	0.0622	0.5797	0.9504	0.0963	0.6396	0.9450	0.1012	0.6464
DS8	0.9933	0.1707	0.6070	0.8984	0.2367	0.6577	0.9446	0.2240	0.6482
DS9	0.9937	0.0782	0.5942	0.9200	0.1236	0.6563	0.9144	0.1475	0.6828
DS10	0.9937	0.0723	0.5382	0.9195	0.1205	0.6238	0.8996	0.1361	0.6382
DS11	0.9926	0.2745	0.6743	0.9786	0.3235	0.7101	0.9357	0.3403	0.7201
DS12	0.9917	0.1409	0.6372	0.9571	0.2026	0.6981	0.9582	0.2223	0.7172

Table S 8: Multiple test results in terms of reduction

DT Classifier							
	FAST	CSB2.FAST	MadaBoost.FAST		FCBF	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9363	0.9136	0.9337	Mean	0.9793	0.9287	0.9193
Ranks	1.7500	2.2917	1.9583	Ranks	1.3333	2.2083	2.4583
Nemenyi CD	0.6767			Nemenyi CD	0.6767		
Win/draw/loss		8/0/16	10/0/14	Win/draw/loss		4/0/20	4/0/20
Wilcoxon p-value		0.0865	0.7533	Wilcoxon p-value		0.0004	0.0001
R+/R-		90.0/210.0	139.0/161.0	R+/R-		27.0/273.0	16.0/284.0
Wilcoxon test		No	No	Wilcoxon test		YES(-)	YES(-)
Ranks difference		-0.5417	-0.2083	Ranks difference		-0.8750	-1.1250
Nememyi test		No	No	Nememyi test		YES(-)	YES(-)
SVM Classifier							
	FAST	CSB2.FAST	MadaBoost.FAST		FCBF	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9363	0.8791	0.8607	Mean	0.9793	0.8455	0.8452
Ranks	1.3333	2.0000	2.6667	Ranks	1.0000	2.5833	2.4167
Nemenyi CD	0.6767			Nemenyi CD	0.6767		
Win/draw/loss		8/0/16	0/0/24	Win/draw/loss		0/0/24	0/0/24
Wilcoxon p-value		0.0025	0.0000	Wilcoxon p-value		0.00002	1.8E-05
R+/R-		44.0/256.0	0.0/300.0	R+/R-		0.0/300.0	0.0/300.0
Wilcoxon test		YES(-)	YES(-)	Wilcoxon test		YES(-)	YES(-)
Ranks difference		-0.6667	-1.3333	Ranks difference		-1.5833	-1.4167
Nememyi test		No	YES(-)	Nememyi test		YES(-)	YES(-)
RF Classifier							
	FAST	CSB2.FAST	MadaBoost.FAST		FCBF	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9363	0.9136	0.9337	Mean	0.9793	0.9287	0.9193
Ranks	1.7500	2.2917	1.9583	Ranks	1.3333	2.2083	2.4583
Nemenyi CD	0.6767			Nemenyi CD	0.6767		
Win/draw/loss		8/0/16	10/0/14	Win/draw/loss		4/0/20	4/0/20
Wilcoxon p-value		0.0865	0.7533	Wilcoxon p-value		0.0004	0.0001
R+/R-		90.0/210.0	139.0/161.0	R+/R-		27.0/273.0	16.0/284.0
Wilcoxon test		No	No	Wilcoxon test		YES(-)	YES(-)
Ranks difference		-0.5417	-0.2083	Ranks difference		-0.8750	-1.1250
Nememyi test		No	No	Nememyi test		YES(-)	YES(-)

Table S 9: Results for CDK descriptors

DT

	FAST		CSB2.FAST		MadaBoost.FAST		FCBF		CSB2.FCBF		MadaBoost.FCBF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.8743	0.5328	0.9516	0.5560	0.8998	0.5669	0.9682	0.5485	0.9433	0.5557	0.9844	0.5557
DS2	0.8179	0.3436	0.8968	0.4508	0.8948	0.4584	0.9712	0.2731	0.8566	0.4827	0.7132	0.4831
DS3	0.8722	0.2817	0.8698	0.3148	0.8909	0.3127	0.9779	0.3310	0.9794	0.3347	0.9833	0.3348
DS4	0.9002	0.0100	0.8095	0.0713	0.8898	0.0263	0.9741	0.0781	0.9342	0.1259	0.9292	0.1845
DS5	0.8706	0.1554	0.9139	0.1653	0.9159	0.1653	0.9731	0.1625	0.9572	0.1691	0.9592	0.1662
DS6	0.8719	0.2333	0.9655	0.2483	0.9606	0.2482	0.9798	0.2482	0.9709	0.2482	0.9734	0.2483
DS7	0.8710	0.0120	0.9112	0.0354	0.8799	0.0354	0.9767	0.0126	0.9553	0.0354	0.9643	0.0354
DS8	0.9919	0.0535	0.8567	0.2306	0.8530	0.2200	0.9702	0.2754	0.9035	0.3465	0.8692	0.3454
DS9	0.8723	0.0534	0.8792	0.1084	0.8515	0.1084	0.9767	0.0283	0.8881	0.1084	0.9248	0.0799
DS10	0.8710	0.1853	0.9097	0.2144	0.9186	0.2497	0.9722	0.1774	0.9246	0.2388	0.9285	0.2388
DS11	0.9232	0.4501	0.8419	0.5502	0.8586	0.5229	0.9768	0.3914	0.8970	0.5345	0.9045	0.5334
DS12	0.8995	0.1520	0.8771	0.2163	0.8900	0.2062	0.9746	0.0932	0.9060	0.2414	0.9184	0.2029

SVM

	FAST		CSB2.FAST		MadaBoost.FAST		FCBF		CSB2.FCBF		MadaBoost.FCBF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.8743	0.6297	0.7086	0.6724	0.8303	0.6701	0.9682	0.5463	0.7976	0.6537	0.8333	0.6239
DS2	0.8179	0.6183	0.7851	0.7156	0.8347	0.6770	0.9712	0.3321	0.8099	0.5696	0.8337	0.5301
DS3	0.8722	0.6047	0.8260	0.6485	0.8197	0.6282	0.9779	0.5060	0.8324	0.6199	0.8570	0.6425
DS4	0.9002	0.4420	0.7377	0.5152	0.7845	0.4918	0.9741	0.3077	0.8219	0.4978	0.8549	0.4915
DS5	0.8706	0.4844	0.6592	0.5025	0.7378	0.5107	0.9731	0.3644	0.8065	0.5225	0.8418	0.5036
DS6	0.8719	0.5229	0.7833	0.6076	0.7798	0.5511	0.9798	0.3637	0.8177	0.5767	0.8488	0.5880
DS7	0.8710	0.3275	0.7474	0.2960	0.7261	0.2898	0.9767	0.2705	0.8333	0.3276	0.8675	0.2952
DS8	0.9919	0.2723	0.8150	0.5687	0.7985	0.5634	0.9702	0.3335	0.7854	0.5374	0.8167	0.5456
DS9	0.8723	0.4059	0.7089	0.4271	0.7163	0.4600	0.9767	0.1770	0.8040	0.4173	0.8277	0.4156
DS10	0.8710	0.3540	0.7400	0.4072	0.7851	0.4095	0.9722	0.1611	0.8045	0.3805	0.8169	0.3799
DS11	0.9232	0.6775	0.8601	0.7357	0.8535	0.7321	0.9768	0.5858	0.8657	0.7208	0.8571	0.7284
DS12	0.8995	0.3549	0.7443	0.4371	0.7667	0.4179	0.9746	0.1933	0.8030	0.3682	0.8383	0.4054

RF

	FAST		CSB2.FAST		MadaBoost.FAST		FCBF		CSB2.FCBF		MadaBoost.FCBF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.8743	0.6233	0.9516	0.6545	0.8998	0.6580	0.9682	0.5649	0.9433	0.5807	0.9844	0.5491
DS2	0.8179	0.6218	0.8968	0.6714	0.8948	0.6575	0.9712	0.6058	0.8566	0.6175	0.7132	0.6499
DS3	0.8722	0.5353	0.8698	0.5623	0.8909	0.5532	0.9779	0.5701	0.9794	0.4816	0.9833	0.5166
DS4	0.9002	0.2941	0.8095	0.3426	0.8898	0.3227	0.9741	0.3596	0.9342	0.3492	0.9292	0.3749
DS5	0.8706	0.4228	0.9139	0.4357	0.9159	0.4543	0.9731	0.3958	0.9572	0.3506	0.9592	0.4165
DS6	0.8719	0.4690	0.9655	0.4971	0.9606	0.5048	0.9798	0.4426	0.9709	0.4721	0.9734	0.4235
DS7	0.8710	0.0332	0.9112	0.0354	0.8799	0.0354	0.9767	0.3074	0.9553	0.2450	0.9643	0.2962
DS8	0.9919	0.3154	0.8567	0.4000	0.8530	0.4076	0.9702	0.4406	0.9035	0.4409	0.8692	0.4391
DS9	0.8723	0.1706	0.8792	0.2061	0.8515	0.1177	0.9767	0.2634	0.8881	0.2646	0.9248	0.2616
DS10	0.8710	0.2442	0.9097	0.2472	0.9186	0.2451	0.9722	0.3118	0.9246	0.3070	0.9285	0.2808
DS11	0.9232	0.6110	0.8419	0.6555	0.8586	0.6572	0.9768	0.6438	0.8970	0.7052	0.9045	0.7030
DS12	0.8995	0.2611	0.8771	0.2396	0.8900	0.2738	0.9746	0.3018	0.9060	0.2956	0.9184	0.3016

Table S 10: Results for SMOTE method

GSFrag

	FAST.SMOTE						FCBF.SMOTE					
	DT		SVM		RF		DT		SVM		RF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.9095	0.7018	0.9095	0.6534	0.9095	0.6634	0.9921	0.7126	0.9921	0.6483	0.9921	0.6517
DS2	0.9478	0.6793	0.9478	0.7184	0.9478	0.7090	0.9916	0.7597	0.9916	0.7121	0.9916	0.7097
DS3	0.9102	0.6102	0.9102	0.7005	0.9102	0.6204	0.9930	0.5954	0.9930	0.5335	0.9930	0.4687
DS4	0.9095	0.3192	0.9095	0.5429	0.9095	0.4922	0.9921	0.1766	0.9921	0.3642	0.9921	0.3808
DS5	0.9107	0.3808	0.9107	0.5750	0.9107	0.5389	0.9931	0.2830	0.9931	0.4174	0.9931	0.3926
DS6	0.9134	0.1895	0.9134	0.5893	0.9134	0.5667	0.9924	0.0120	0.9924	0.4564	0.9924	0.4394
DS7	0.9095	0.0986	0.9095	0.3115	0.9095	0.2461	0.9917	0.0100	0.9917	0.0333	0.9917	0.0666
DS8	0.9097	0.6046	0.9097	0.5907	0.9097	0.5335	0.9926	0.5243	0.9926	0.5234	0.9926	0.4592
DS9	0.9095	0.0276	0.9095	0.4452	0.9095	0.3935	0.9919	0.0277	0.9919	0.2164	0.9919	0.2064
DS10	0.9095	0.2717	0.9095	0.4301	0.9095	0.3903	0.9917	0.1647	0.9917	0.2563	0.9917	0.2721
DS11	0.9098	0.6913	0.9098	0.7380	0.9098	0.6830	0.9912	0.6682	0.9912	0.6292	0.9912	0.5431
DS12	0.9111	0.3042	0.9111	0.4571	0.9111	0.4283	0.9902	0.2878	0.9902	0.3996	0.9902	0.3783

ECFP4

	FAST.SMOTE						FCBF.SMOTE					
	DT		SVM		RF		DT		SVM		RF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.9062	0.6221	0.9062	0.6687	0.9062	0.6704	0.9514	0.6342	0.9514	0.6234	0.9514	0.6247
DS2	0.9062	0.6219	0.9062	0.6611	0.9062	0.7053	0.9520	0.7051	0.9520	0.6500	0.9520	0.6431
DS3	0.9062	0.4877	0.9062	0.6347	0.9062	0.5097	0.9566	0.4558	0.9566	0.5384	0.9566	0.4968
DS4	0.9062	0.1136	0.9062	0.4806	0.9062	0.3966	0.9496	0.0364	0.9496	0.4222	0.9496	0.3880
DS5	0.9062	0.4250	0.9062	0.4778	0.9062	0.4481	0.9572	0.3660	0.9572	0.4111	0.9572	0.4056
DS6	0.9062	0.5806	0.9062	0.5397	0.9062	0.5420	0.9514	0.5036	0.9514	0.5005	0.9514	0.4886
DS7	0.9062	0.0930	0.9062	0.3854	0.9062	0.2318	0.9521	0.0930	0.9521	0.2852	0.9521	0.2073
DS8	0.9062	0.2972	0.9062	0.5254	0.9062	0.4203	0.9631	0.2324	0.9631	0.4139	0.9631	0.3919
DS9	0.9062	0.2510	0.9062	0.3579	0.9062	0.3030	0.9514	0.2144	0.9514	0.3821	0.9514	0.3475
DS10	0.9062	0.2947	0.9062	0.3758	0.9062	0.2642	0.9553	0.1898	0.9553	0.2650	0.9553	0.2620
DS11	0.9062	0.5457	0.9062	0.6338	0.9062	0.5773	0.9572	0.5317	0.9572	0.5918	0.9572	0.5516
DS12	0.9062	0.3431	0.9062	0.4445	0.9062	0.3592	0.9498	0.3396	0.9498	0.3790	0.9498	0.3516

CDK descriptors

	FAST.SMOTE						FCBF.SMOTE					
	DT		SVM		RF		DT		SVM		RF	
Dataset	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean	r	G-Mean
DS1	0.8729	0.6493	0.8729	0.6605	0.8729	0.6773	0.9545	0.5650	0.9545	0.6588	0.9545	0.6377
DS2	0.8710	0.6712	0.8710	0.7026	0.8710	0.7290	0.9479	0.6523	0.9479	0.6981	0.9479	0.6598
DS3	0.8722	0.6058	0.8722	0.6854	0.8722	0.6483	0.9582	0.5745	0.9582	0.6087	0.9582	0.5900
DS4	0.8703	0.2462	0.8703	0.5115	0.8703	0.4953	0.9521	0.1037	0.9521	0.5042	0.9521	0.4567
DS5	0.8706	0.5168	0.8706	0.5655	0.8706	0.5461	0.9607	0.2366	0.9607	0.5439	0.9607	0.4711
DS6	0.8719	0.4134	0.8719	0.6359	0.8719	0.6162	0.9562	0.2546	0.9562	0.5613	0.9562	0.5245
DS7	0.8710	0.0500	0.8710	0.3126	0.8710	0.2851	0.9524	0.0492	0.9524	0.4039	0.9524	0.2012
DS8	0.8712	0.6131	0.8712	0.5865	0.8712	0.5683	0.9601	0.5782	0.9601	0.5354	0.9601	0.4810
DS9	0.8713	0.0285	0.8713	0.4759	0.8713	0.3572	0.9525	0.0200	0.9525	0.4282	0.9525	0.3150
DS10	0.8710	0.3639	0.8710	0.4387	0.8710	0.3515	0.9514	0.3214	0.9514	0.4314	0.9514	0.3694
DS11	0.8712	0.7219	0.8712	0.7610	0.8712	0.7611	0.9611	0.6886	0.9611	0.6805	0.9611	0.6793
DS12	0.8706	0.3582	0.8706	0.4870	0.8706	0.4246	0.9502	0.3704	0.9502	0.4650	0.9502	0.3526

Table S 11: Nemenyi test results in terms of *G-Mean* including GSFrag, ECFP4, MACCS, and CDK descriptors

DT Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.4666	0.4082	0.5205	0.5193	Mean	0.4826	0.3582	0.5284	0.5279
Ranks	3.3889	2.9444	1.7361	1.9306	Ranks	3.1111	3.0556	1.9444	1.8889
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	0.4444	1.6528	1.4583	Ranks difference	✓	0.0556	1.1667	1.2222
Nememyi test		No	YES(+)	YES(+)	Nememyi test		No	YES(+)	YES(+)
Ranks difference		✓	1.2083	1.0139	Ranks difference		✓	1.1111	1.1667
Nememyi test			YES(+)	YES(+)	Nememyi test			YES(+)	YES(+)
SVM Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.5813	0.5489	0.6411	0.6373	Mean	0.5396	0.4770	0.6286	0.6282
Ranks	3.3889	2.9722	1.6111	2.0278	Ranks	3.2778	3.2222	1.8056	1.6944
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	0.4167	1.7778	1.3611	Ranks difference	✓	0.0556	1.4722	1.5833
Nememyi test		No	YES(+)	YES(+)	Nememyi test		No	YES(+)	YES(+)
Ranks difference		✓	1.3611	0.9444	Ranks difference		✓	1.4167	1.5278
Nememyi test			YES(+)	YES(+)	Nememyi test			YES(+)	YES(+)
RF Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.5475	0.5043	0.5883	0.5808	Mean	0.5678	0.4407	0.5874	0.5908
Ranks	3.2222	2.8889	1.8194	2.0694	Ranks	2.5833	3.1389	2.0833	2.1944
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	0.3333	1.4028	1.1528	Ranks difference	✓	-0.5556	0.5000	0.3889
Nememyi test		No	YES(+)	YES(+)	Nememyi test		No	No	No
Ranks difference		✓	1.0694	0.8194	Ranks difference		✓	1.0556	0.9444
Nememyi test			YES(+)	YES(+)	Nememyi test			YES(+)	YES(+)

Table S 12: Nemenyi test results in terms of reduction r including GSFrag, ECFP4, MACCS, and CDK descriptors

DT Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9197	0.8970	0.9058	0.9198	Mean	0.9776	0.9669	0.9279	0.9199
Ranks	2.0000	3.1667	2.6111	2.2222	Ranks	1.3611	2.4167	3.1111	3.1111
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	-1.1667	-0.6111	-0.2222	Ranks difference	✓	-1.0556	-1.7500	-1.7500
Nememyi test		YES(-)	No	No	Nememyi test		YES(-)	YES(-)	YES(-)
Ranks difference		✓	0.5556	0.9444	Ranks difference		✓	-0.6944	-0.6944
Nememyi test			No	YES(+)	Nememyi test			No	No
SVM Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9197	0.8970	0.8393	0.8358	Mean	0.9776	0.9669	0.8354	0.8439
Ranks	1.3611	2.1667	2.9722	3.5000	Ranks	1.0556	1.9444	3.6944	3.3056
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	-0.8056	-1.6111	-2.1389	Ranks difference	✓	-0.8889	-2.6389	-2.2500
Nememyi test		YES(-)	YES(-)	YES(-)	Nememyi test		YES(-)	YES(-)	YES(-)
Ranks difference		✓	-0.8056	-1.3333	Ranks difference		✓	-1.75	-1.3611
Nememyi test			YES(-)	YES(-)	Nememyi test			YES(-)	YES(-)
RF Classifier									
	FAST	FAST.SMOTE	CSB2.FAST	MadaBoost.FAST		FCBF	FCBF.SMOTE	CSB2.FCBF	MadaBoost.FCBF
Mean	0.9197	0.8970	0.9058	0.9198	Mean	0.9776	0.9669	0.9279	0.9199
Ranks	2.0000	3.1667	2.6111	2.2222	Ranks	1.3611	2.4167	3.1111	3.1111
Nemenyi CD	0.7817				Nemenyi CD	0.7817			
Ranks difference	✓	-1.1667	-0.6111	-0.2222	Ranks difference	✓	-1.0556	-1.7500	-1.7500
Nememyi test		YES(-)	No	No	Nememyi test		YES(-)	YES(-)	YES(-)
Ranks difference		✓	0.5556	0.9444	Ranks difference		✓	-0.6944	-0.6944
Nememyi test			No	YES(+)	Nememyi test			No	No