Paper Title: **CancerEMC: cancer detection from circulating protein biomarkers and mutations in cell-free DNA**

Supplementary Data

Bagging:

Bagging (bootstrap aggregating) is an ensemble meta classifier (Breiman, 1996) that builds a set of training models for classification and each of the training models is obtained from a bootstrap replica of the original training dataset. The instances of those models are selected randomly with replacement. For an imaginary example, the original training set has five instances {1, 2, 3, 4, 5}. After applying bagging method, it builds a set of three training models by random bootstrapping as

Training model 1: {4, 3, 5, 1, 2}

Training model 2: {3, 1, 4, 2, 5}

Training model 3: {5, 2, 4, 1, 3}

Bagging method train the all three-training models by using a base classification learner and find the best training model among the new training models with the minimum classification error Breiman (1996).

AODE:

For cancer detection, CancerEMC method employs this bagging ensemble meta classifier with AODE as base classification learner. In particular, we used the 'base_classifier' parameter of bagging ensemble meta classifier is AODE learning framework (Webb et. al., 2005) that is a variant of Naïve Bayes (NB) learning algorithm. Many approaches have been showed that the weakening features independence assumption can improve performance. Lazy Bayesian Rules (LBR) and Tree Augmented NB (TAN) method both rely on weaker feature independence assumption. However, their computational costs are intensive. On the other hand, model selection has overfitting problem in training dataset that increase the estimation variance. To solve the mentioned limitation of LBR, TAN and NB, the average one dependence estimator (AODE) method has been developed by Webb et. al., (2005). It can avoid the model selection and used 1 dependence classifier. In addition, it uses a threshold *m* where the training data of model is fewer than *m* example of parent features x_i of $\boldsymbol{x}$ feature vector, where x= $\{x_1, x_2, ..., x_n\}$. For CancerEMC, $\boldsymbol{x}$ is the input features vector of multianalyte blood test data for cancer detection. We can propose AODE model for training by using input feature vector x and cancer detection class level y based on the production rule as follows for any feature x_i

$$P(y, \boldsymbol{x}) = P(y, x_i\ )P(\boldsymbol{x}|y, x_i) \tag{7}$$

Equation (7) holds for each feature x_i, it also holds for average over any features set values for one-dependence classifier. Hence,

$$P(y, \boldsymbol{x}) = \frac{\sum_{i:1\leq i\leq n\wedge F(x_i)\geq m} PP(y, x_i\ )P(\boldsymbol{x}|y, x_i)}{|\{i: 1 \leq i \leq n\wedge F(x_i) \geq m\}|} \tag{8}$$

where $F(x_i)$ is the number of training example with feature value x_i. The equation (8) gives a new class estimating technique as classifier that is called Average One-Dependence Estimators (AODE). Thereafter, substituting probability estimate for cancer detection class level y in equation (8), the classifier selects the suitable cancer class as follows:

$$\underset{y}{argmax}\left(\sum_{i:1\leq i\leq n\wedge F(x_i)\geq m}\hat{P}(y,x_i)\prod\nolimits_{j=1}^{n}\hat{P}(x_j|y,x_i)\right) \tag{9}$$

where $\hat{P}$ is the estimated probability. The AODE can be derived for direct class estimating by normalization of numerator of equation (8) for each class.

$$\hat{P}(y,x)=\frac{\sum_{i:1\leq i\leq n\wedge F(x_i)\geq m}\hat{P}(y,x_i)\prod_{j=1}^{n}\hat{P}(x_j|y,x_i)}{\sum_{\acute{y}\in Y}\sum_{i:1\leq i\leq n\wedge F(x_i)\geq m}\hat{P}(\acute{y},x_i)\prod_{j=1}^{n}\hat{P}(x_j|\acute{y},x_i)} \tag{10}$$

Supplementary Figures and Tables:

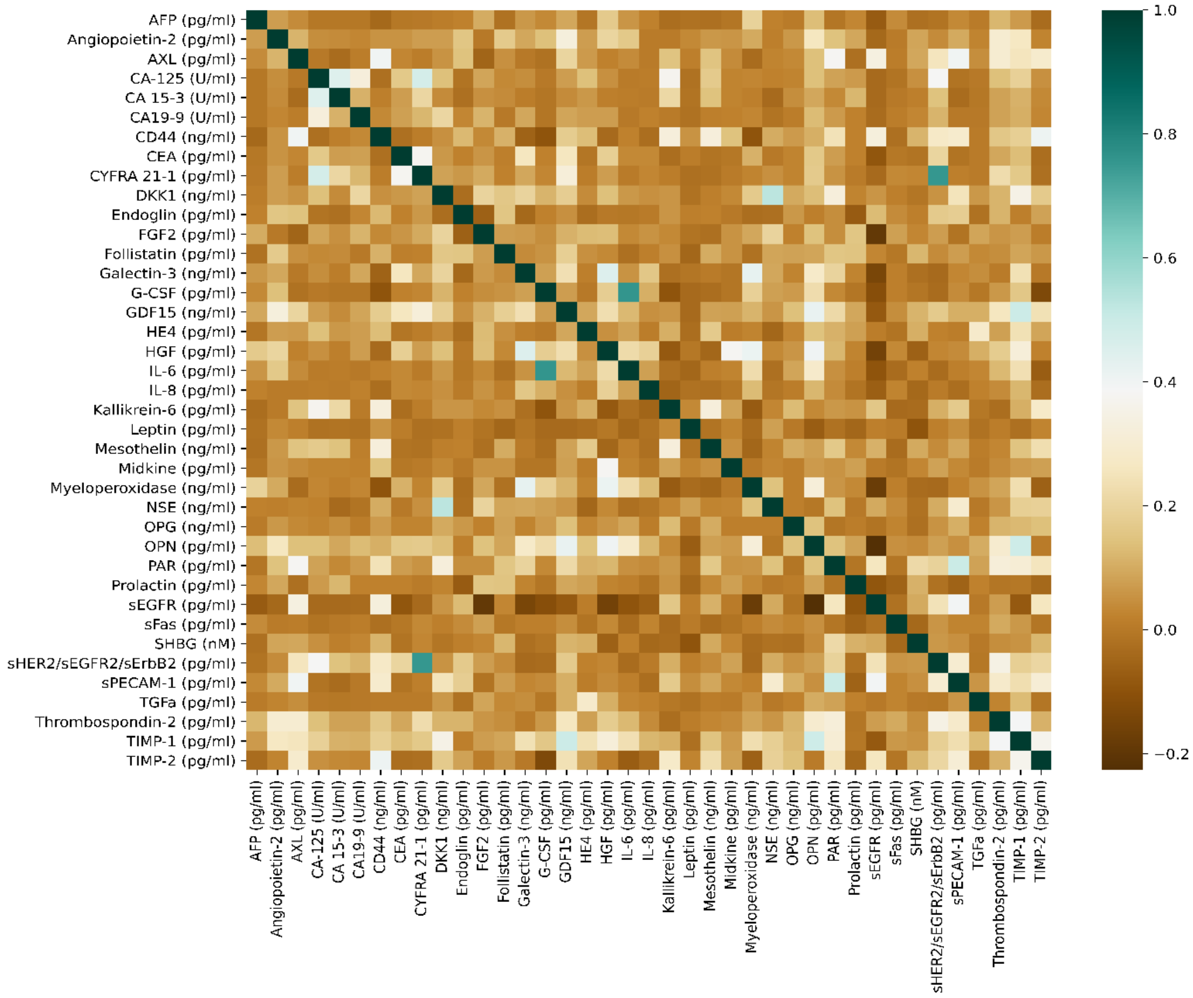

Figure S1: Pearson correlational Heatmap diagram for 39 protein biomarkers in cancer detection.

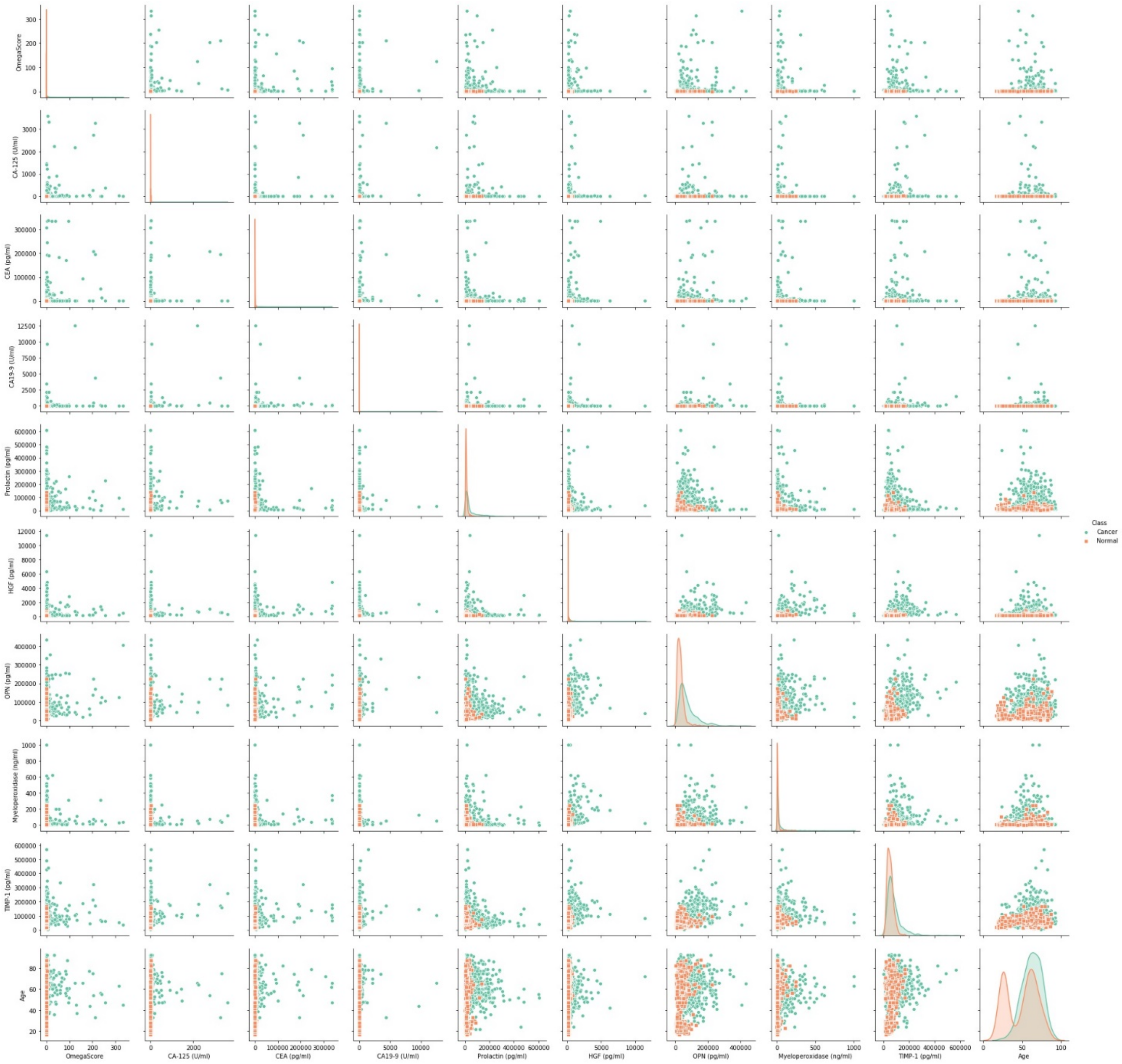

Figure S2: Scatter matrix for features visualization of SD1 dataset for binary cancer detection

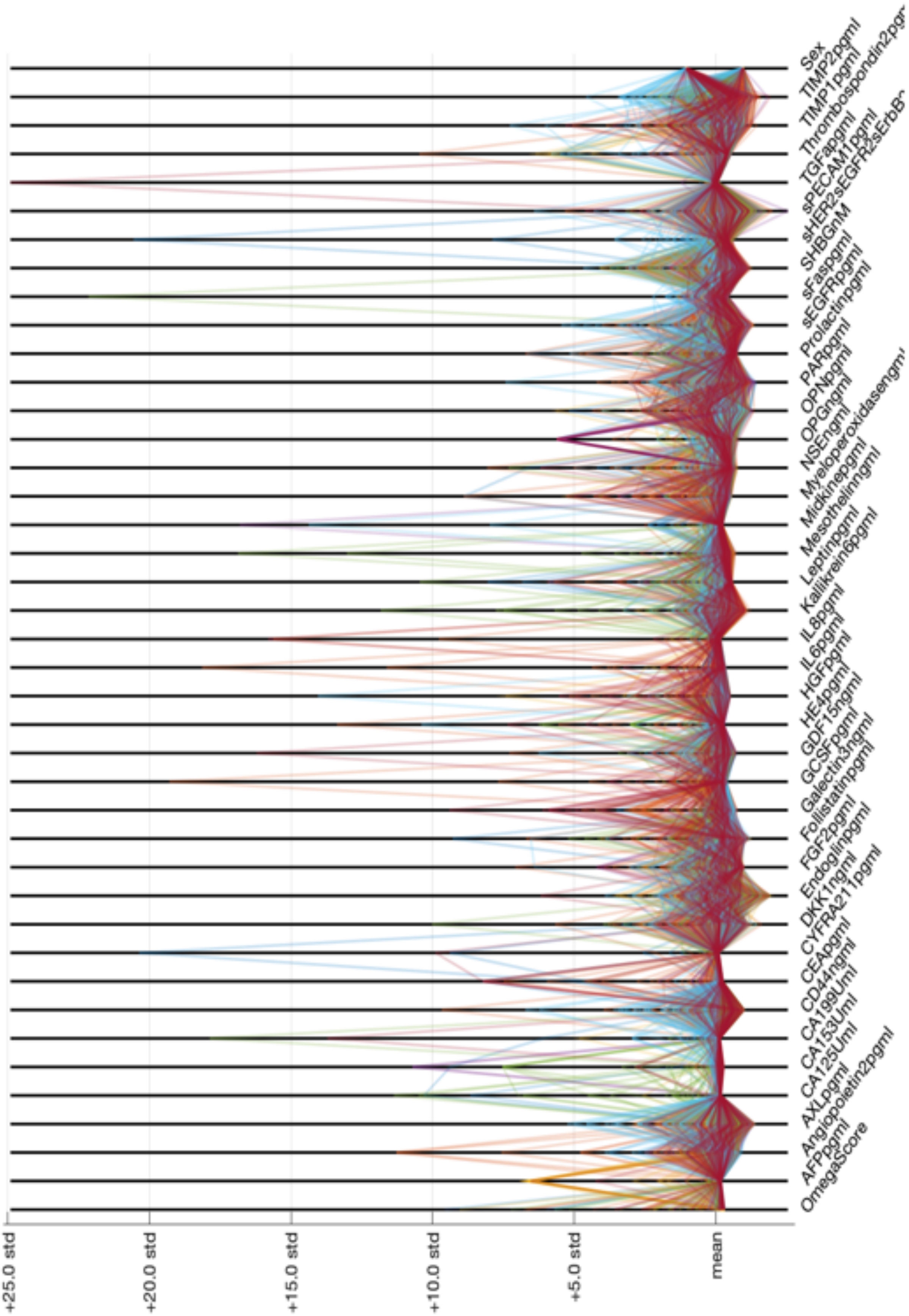

Figure S3: Parallel coordinate plot for features visualization of SD4 dataset for cancer detection

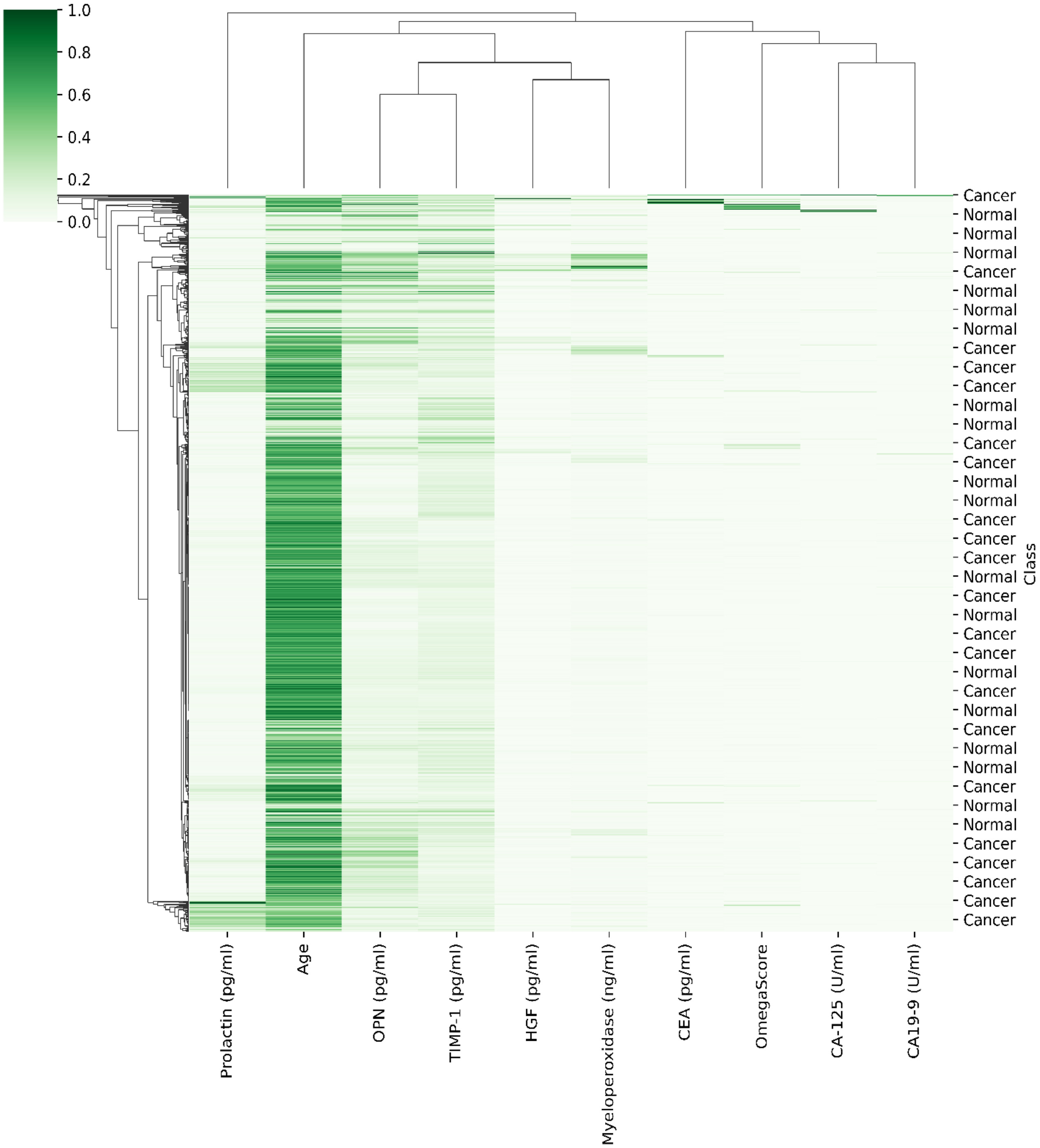

Figure S4: Cluster correlational heatmap diagram for binary cancer classification data of SD1

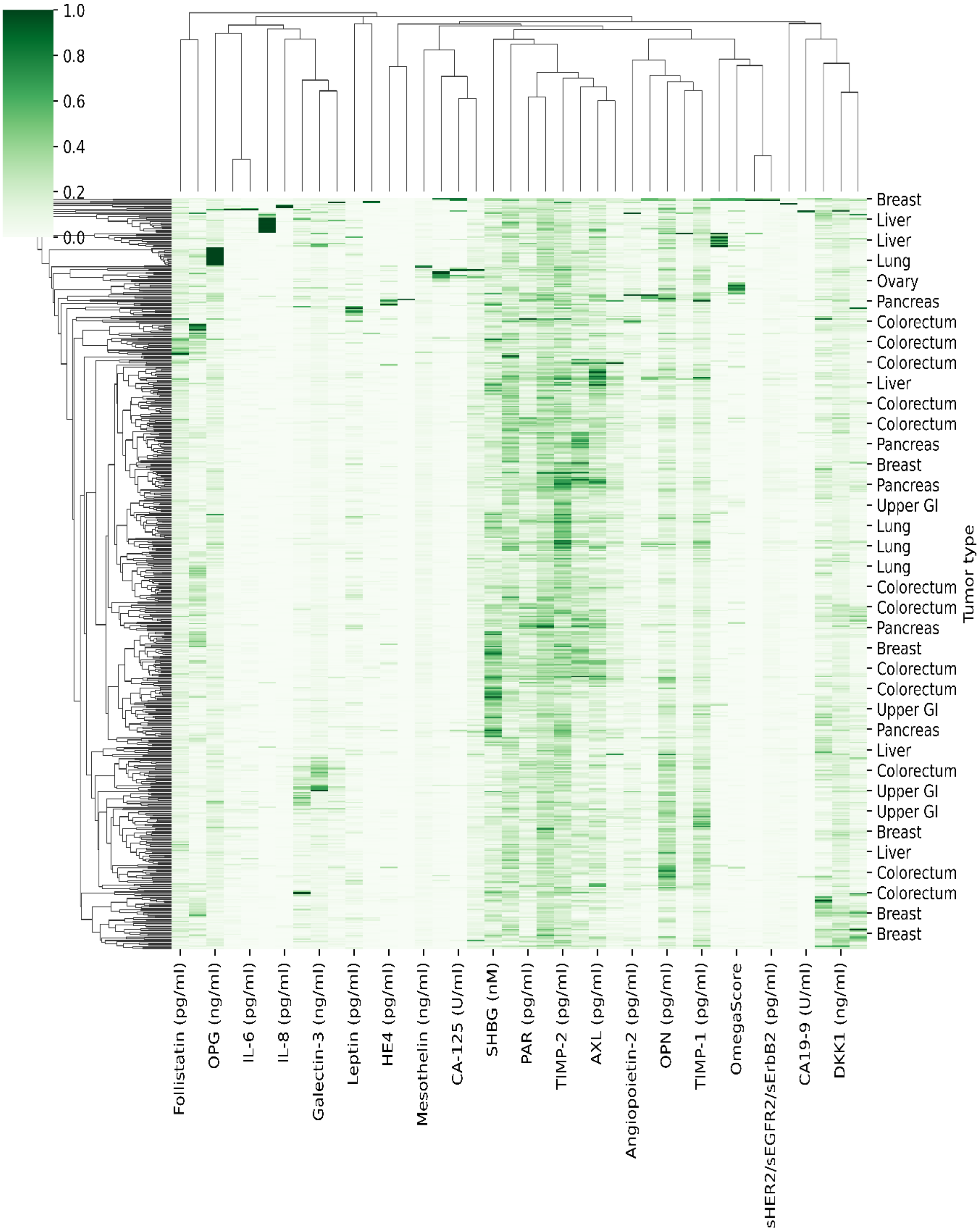

Figure S5: Cluster correlational heatmap diagram of SD4 for cancer localization types classification

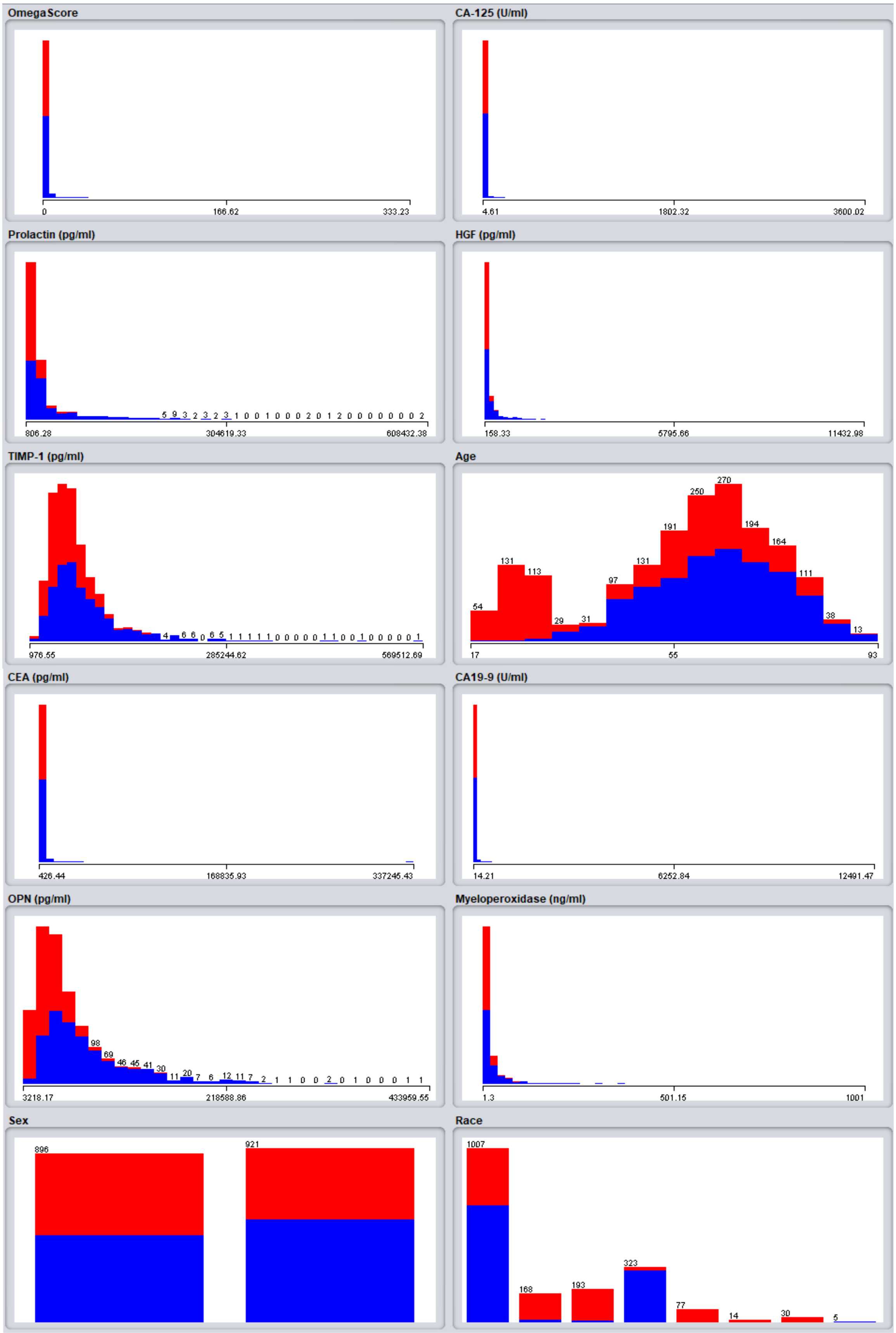

Figure S6: Feature Histogram for sub-dataset SD1 for binary cancer detection (Blue color represent the Cancer class and red color for healthy Normal class)

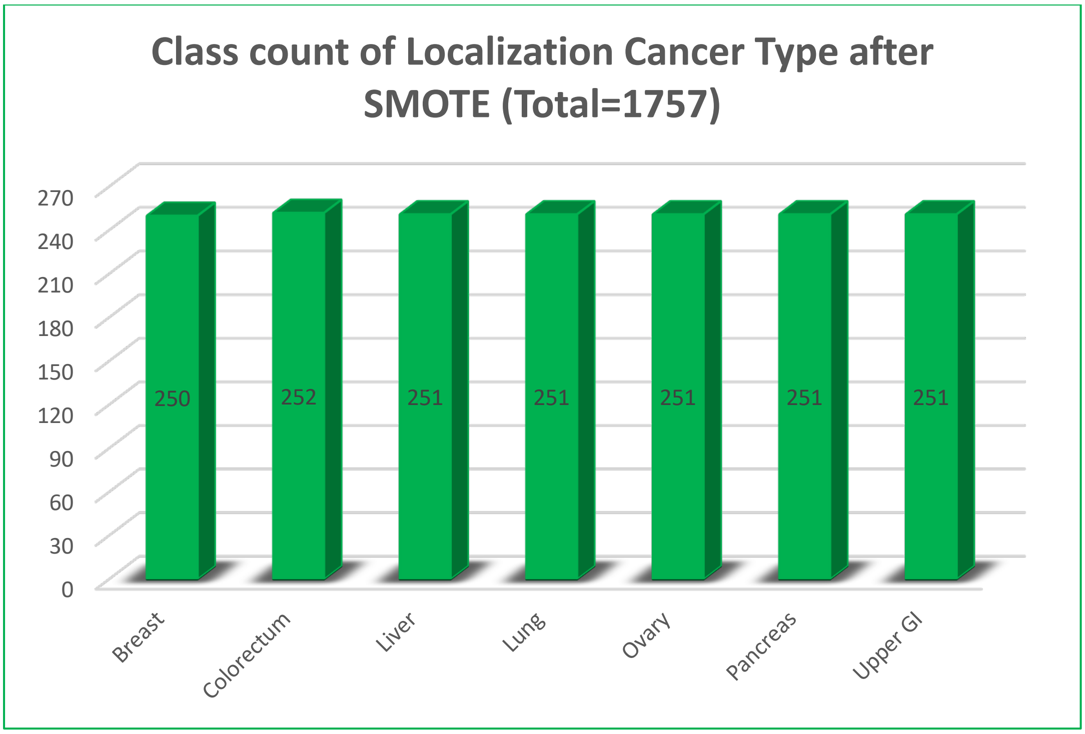

Figure S7: Instance number of each class in localize cancer detection dataset SD4 after 500% SMOTE.

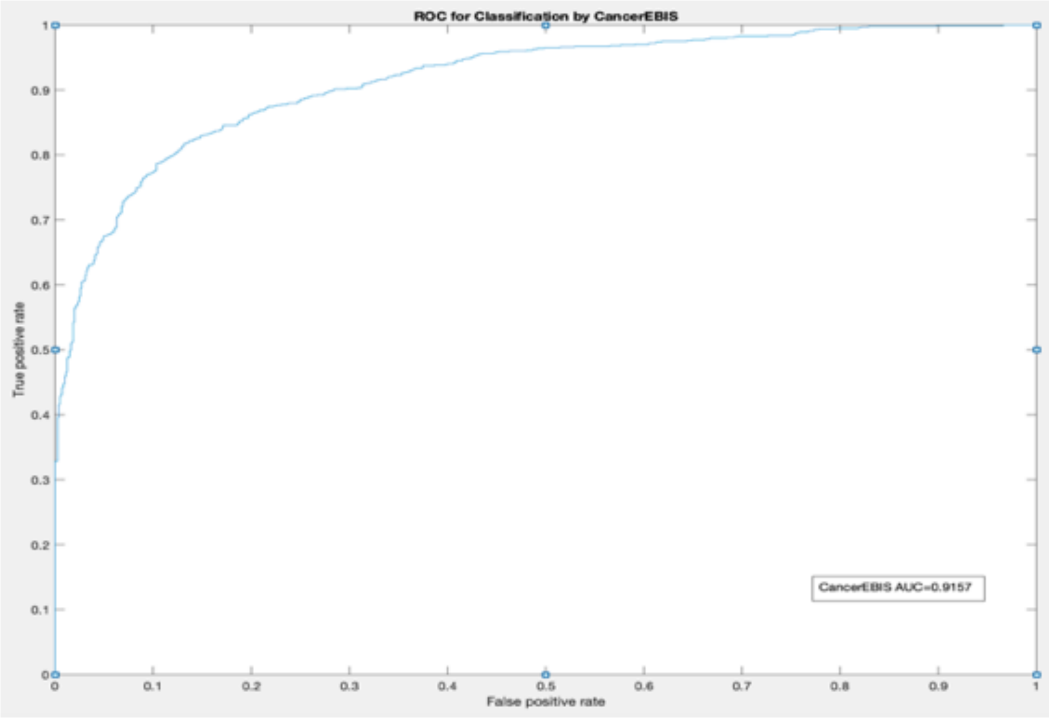

Figure S8: ROC curve and AUC for CancerEBIS method for binary Cancer detection

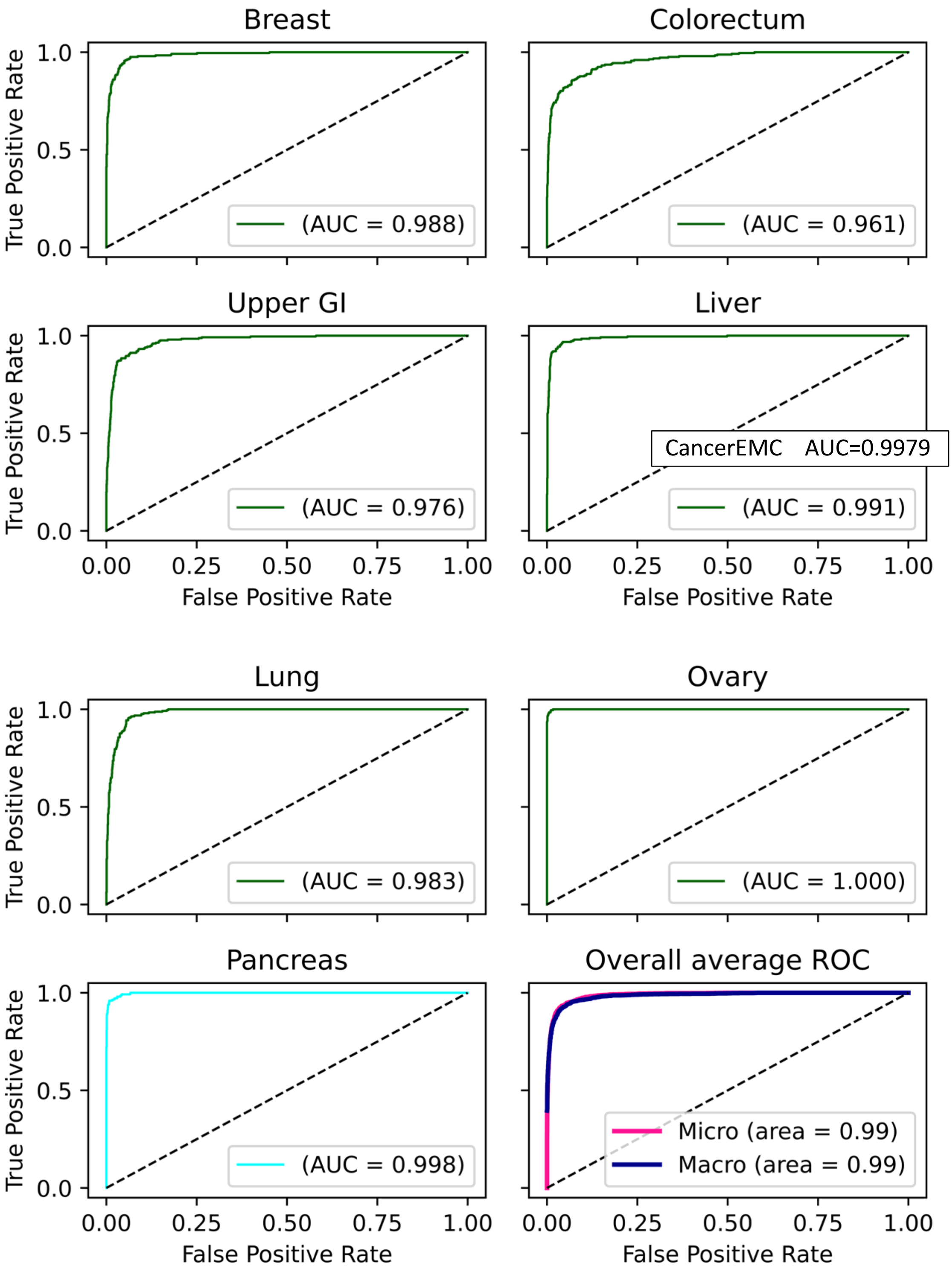

Figure S9: ROC curve analysis of different localize cancer types detection of CancerEMC.

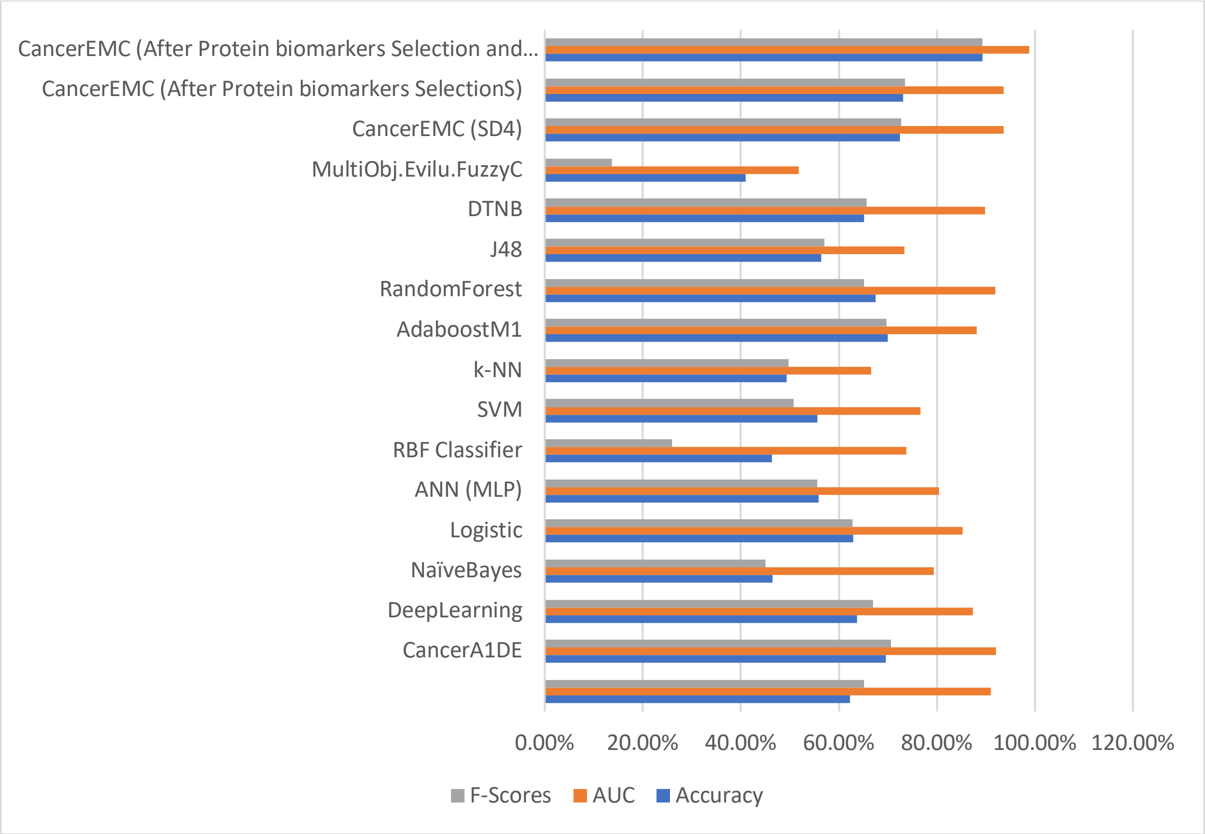

Figure S10: Effects on CancerEMC cancer localization results by Oversampling (SMOTE) on SD4

Sl.	Features	InfoGainRatio	CfsSubsetEval (10 folds evaluation)
1	CA19-9 (U/ml)	0.6897	100%
2	CA-125 (U/ml)	0.5119	100%
3	HGF (pg/ml)	0.5001	100%
4	OPN (pg/ml)	0.2779	100%
5	OmegaScore	0.2208	100%
6	Prolactin (pg/ml)	0.1826	100%
7	CEA (pg/ml)	0.1518	70%
8	Myeloperoxidase	0.0989	0%
9	TIMP-1 (pg/ml)	0.0916	0%

Table S1: Features selection evaluation for CancerEBIS framework from sub-dataset1 (SD1) for binary cancer classification

ID	Features	Input Range for reference	References	Input transform of first instance to belief degree of referential values {VH, H, M, L, VL}
(1)	(2)	(3)	(4)	(5)
A1	CA19-9	A1>=356	VH	If input A1=16.452
	(U/ml)	356>A1>=82	H	Then
		82>A1>=53	M	A1{0, 0, 0, 0.0502, 0.9498}
		53>A1>=25	L	
		25>A1>=0	VL	
A2	CA-125	A2>=164	VH	If A2=5.09
	(U/ml)	164>A2>=41	H	Then
		41>A2>=25	M	A2 {0, 0, 0, 0.0090, 0.9910}
		25>A2>=15	L	
		15>A2>=0	VL	
A3	HGF (pg/ml)	A3>=748	VH	If A3=377.26
		748>A3>=430	H	Then
		430>A3>=249	M	A3 {0, 0.5071, 0.4929, 0, 0}
		249>A3>=175	L	
		175>A3>=0	VL	
A4	OPN (pg/ml)	A4>=103329	VH	If A4=56516.58
		103329>A4>=76176	H	Then
		76176>A4>=56295	M	A4 {0, 0.0111, 0.9889, 0, 0}
		56295>A4>=36874	L	
		36874>A4>=0	VL	
A5	Omega	A5>=7.1	VH	If A5=2.96
	Score	7.1>A5>=5.6	H	Then
		5.6>A5>=4.2	M	A5 {0, 0, 0.2706, 0.7294, 0}
		4.2>A5>=2.5	L	
		2.5>A5>=0	VL	
A6	Prolactin (PRL)	A6>=82164	VH	If A6=11606.6
	(pg/ml)	82164>A6>=48095	H	Then
		48095>A6>=32314	M	
		32314>A6>=22548	L	A6{0, 0, 0, 0, 1.0000
		22548>A6>=0	VL	

Table S2: Input features of CancerEBIS Framework with input range from sub-dataset1 (SD1) for binary cancer classification

Methods	Accuracy	Sensitivity	AUC	Rank
CancerSEEK	77.71%	60.39%	0.930	10
CancerA1DE	96.64%	97.40%	0.991	3
DeepLearning	82.05%	75%	0.916	8
NaïveBayes	77.10%	60.80%	0.889	11
SVM	80.18%	70.20%	0.814	9
k-NN	76.82%	76.82%	0.762	12
AdaboostM1	94.55%	95.6%	0.982	5
RandomForest	91.90%	93.6%	0.913	6
J48	89.21%	89.60%	0.913	7
DTNB	95.54%	97.10%	0.990	4
MultiObj.Evilu.FuzzyC	76.49%	74.60%	0.767	13
CancerEMC (SD1)	97.91%	98.40%	0.9989	2
CancerEMC (After protein biomarker selection from SD2)	**99.17%**	**99.60%**	**0.999**	**1**

Table S3: Comparisons of evaluation metrics for all methods for binary classification

Class Index	Oversampling Size	No. of instances	Accuracy	AUC	F-Score
0	0%	626	73.16%	0.936	0.735
1	258.%	806	78.66%	0.96	0.787
3	215%	977	81.88%	0.965	0.816
4	486%	1185	84.56%	0.977	0.845
5	313%	1375	86.55%	0.979	0.865
6	375%	1573	87.60%	0.986	0.875
7	276%	1757	89.36%	0.989	0.893

Table S4: Median results of cancer localization after applying different size of SMOTE in CancerEMC methods