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= $\{x_1, x_2, \ldots, x_n\}$. 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:

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

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.

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

Supplementary Figures and Tables:

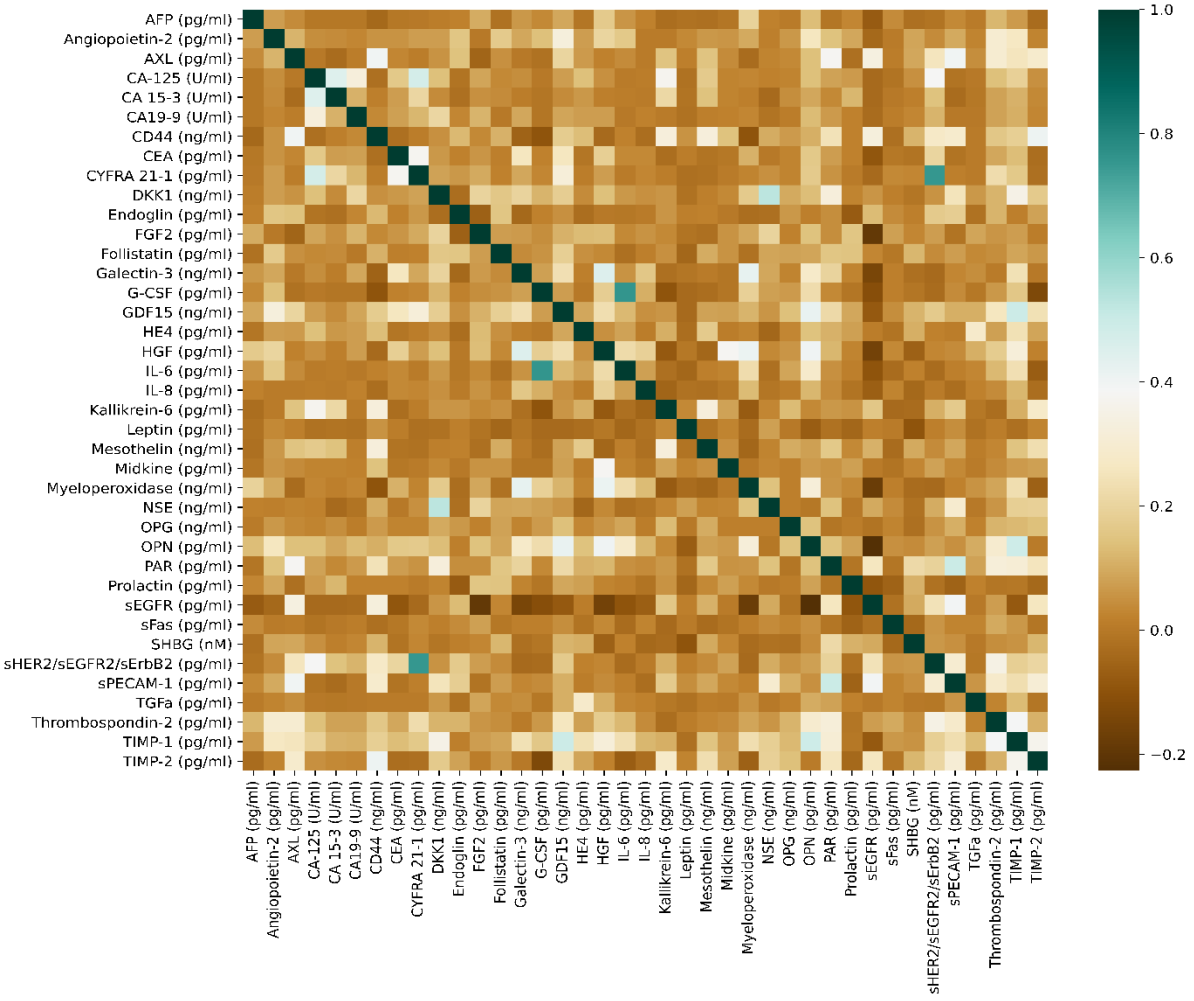

Figure S1: Pearson correlational Heatmap diagram for 39 protein biomarkers in cancer detection.

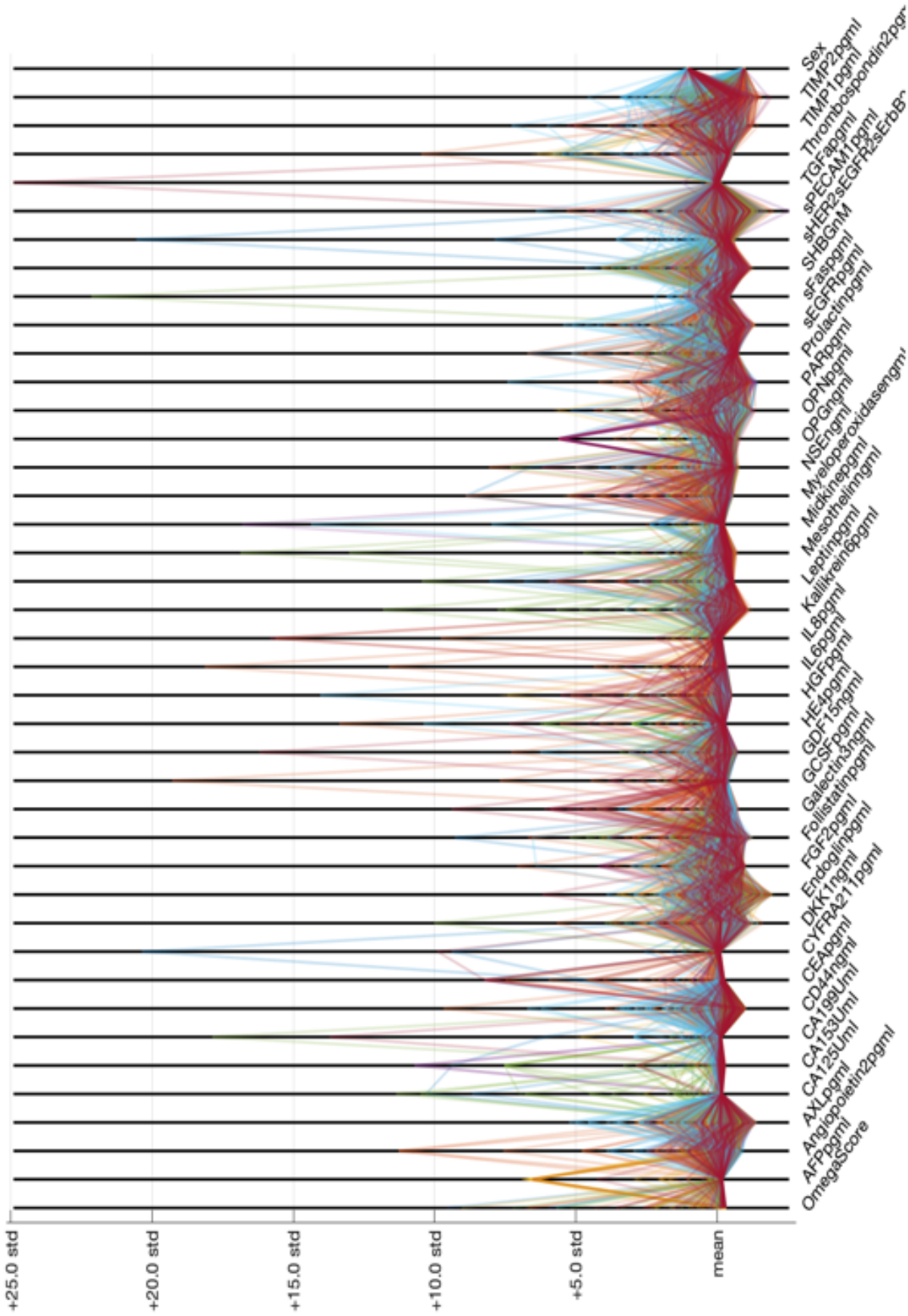

Figure S2: Parallel coordinate plot for features visualization of SD4 dataset for cancer detection

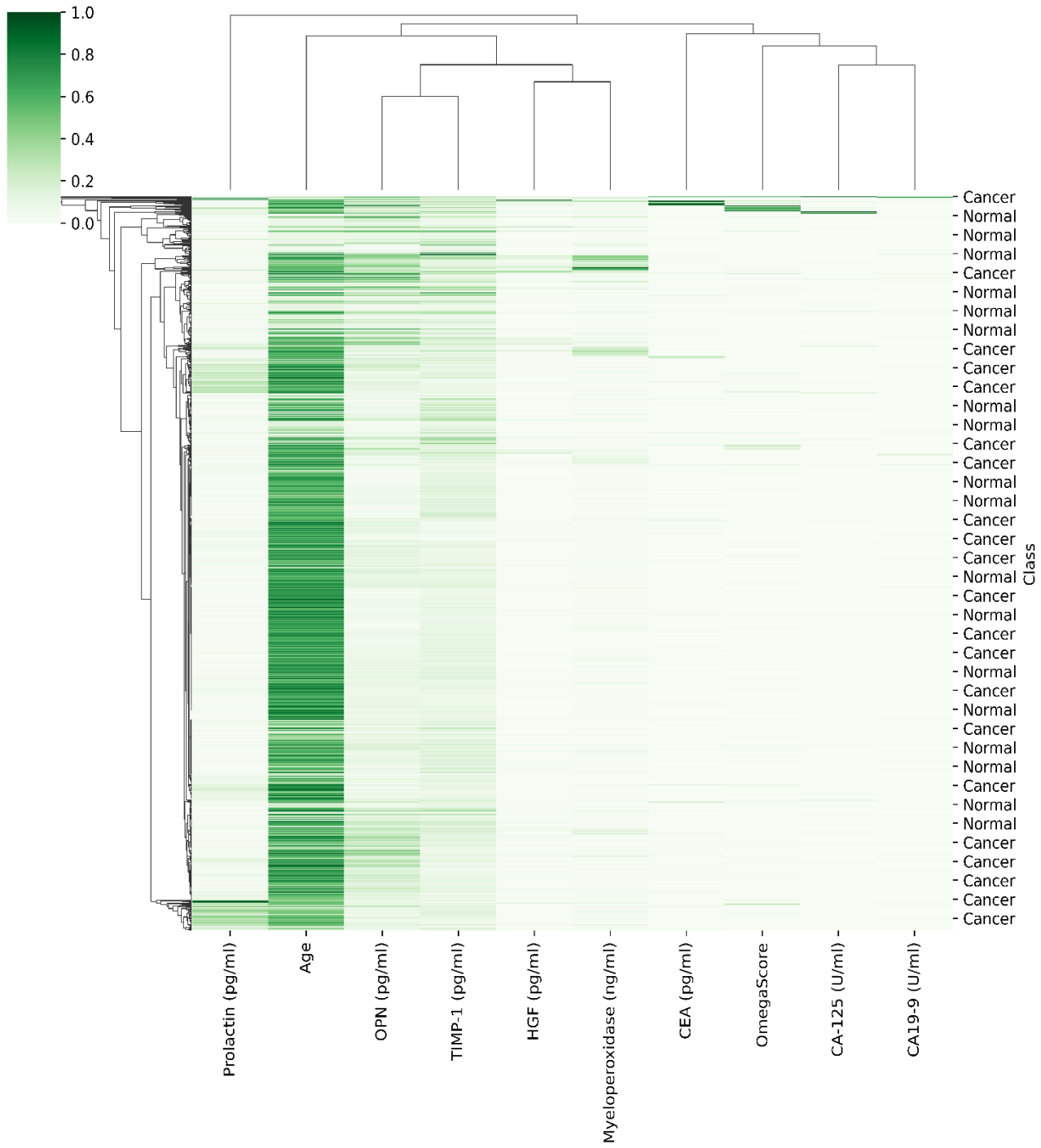

Figure S3: Cluster correlational heatmap diagram for binary cancer classification data of SD1

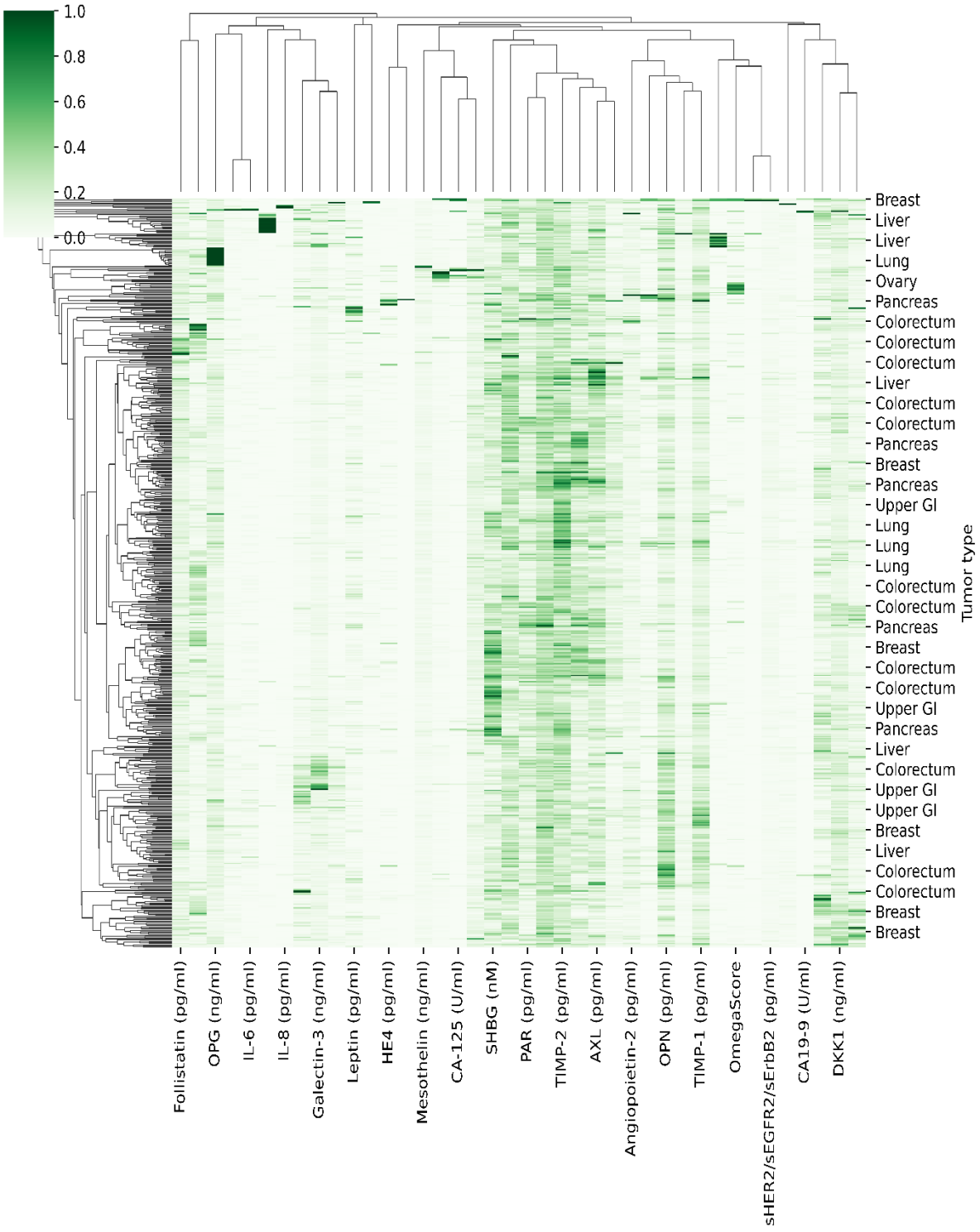

Figure S4: Cluster correlational heatmap diagram of SD4 for cancer localization types classification

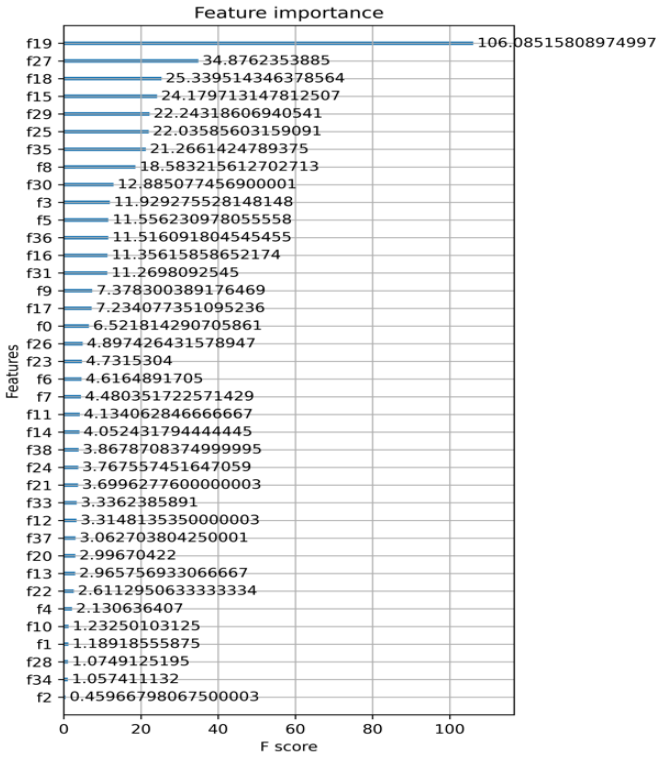

Figure S5: XGBoost Protein Bio-marker Importance Bar Chart with the average gain across all splits for Binary cancer detection

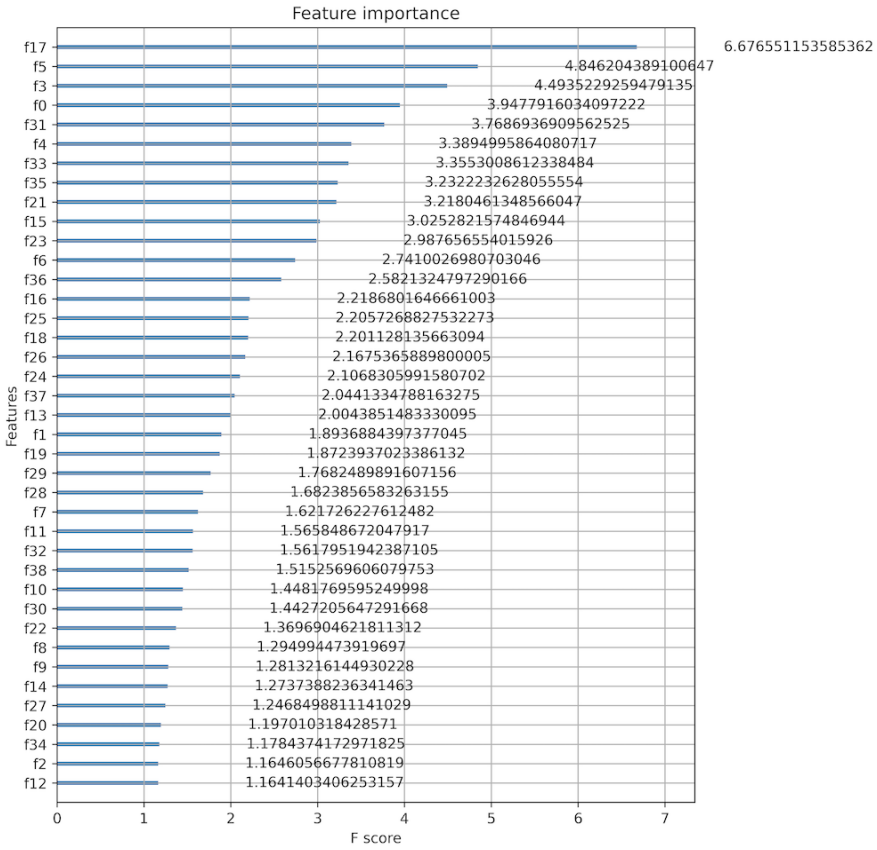

Figure S6: XGBoost Protein Bio-marker Importance Bar Chart with the average gain across all splits for Localize cancer detection

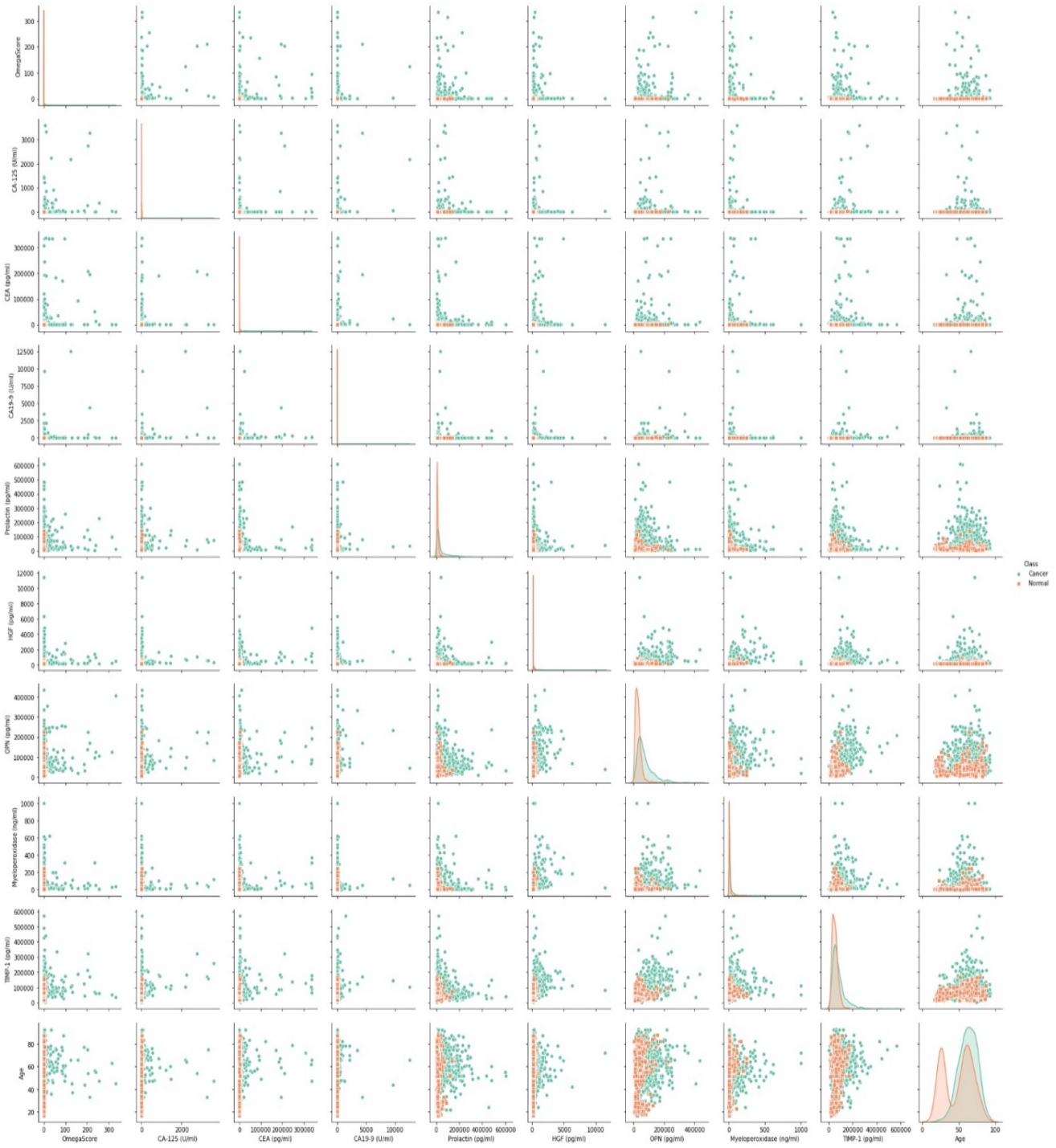

Figure S7: Scatter matrix with histogram for features visualization of SD1 dataset for binary cancer detection

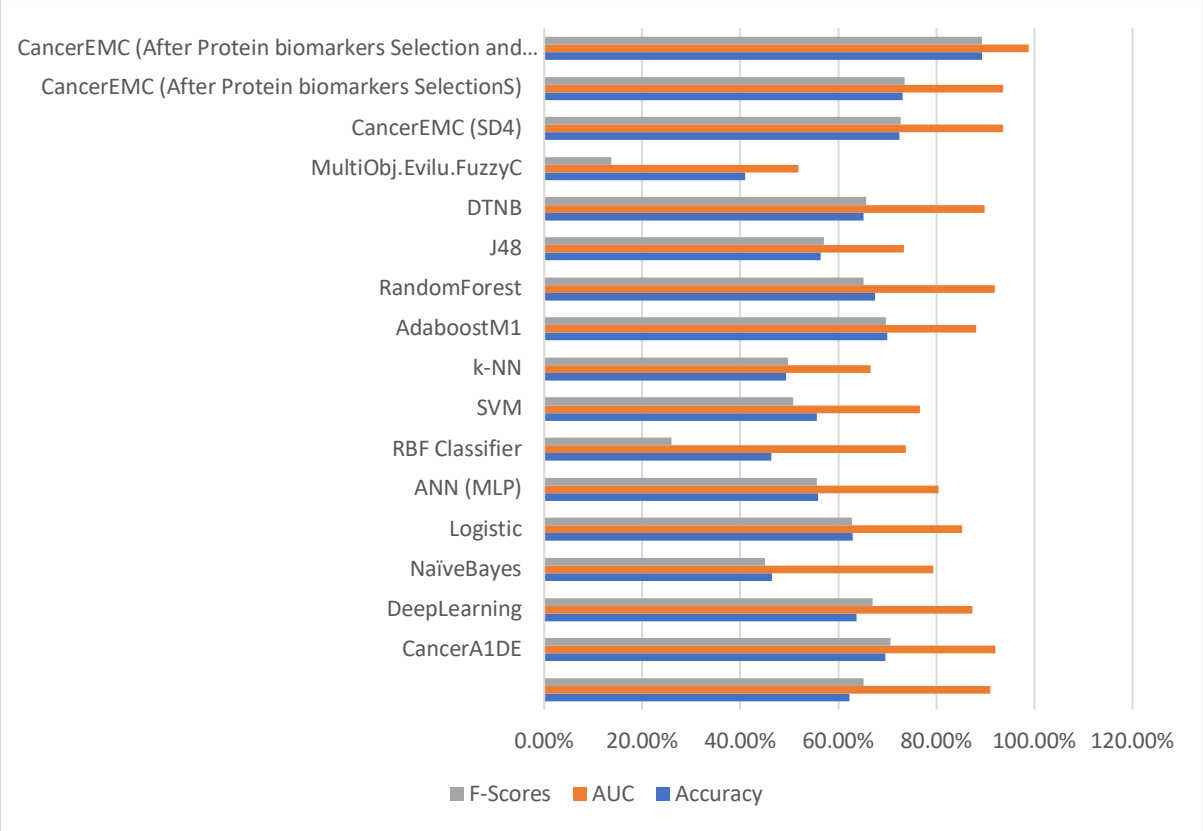

Figure S9: Effects on CancerEMC cancer localization results by Oversampling (SMOTE) on SD4

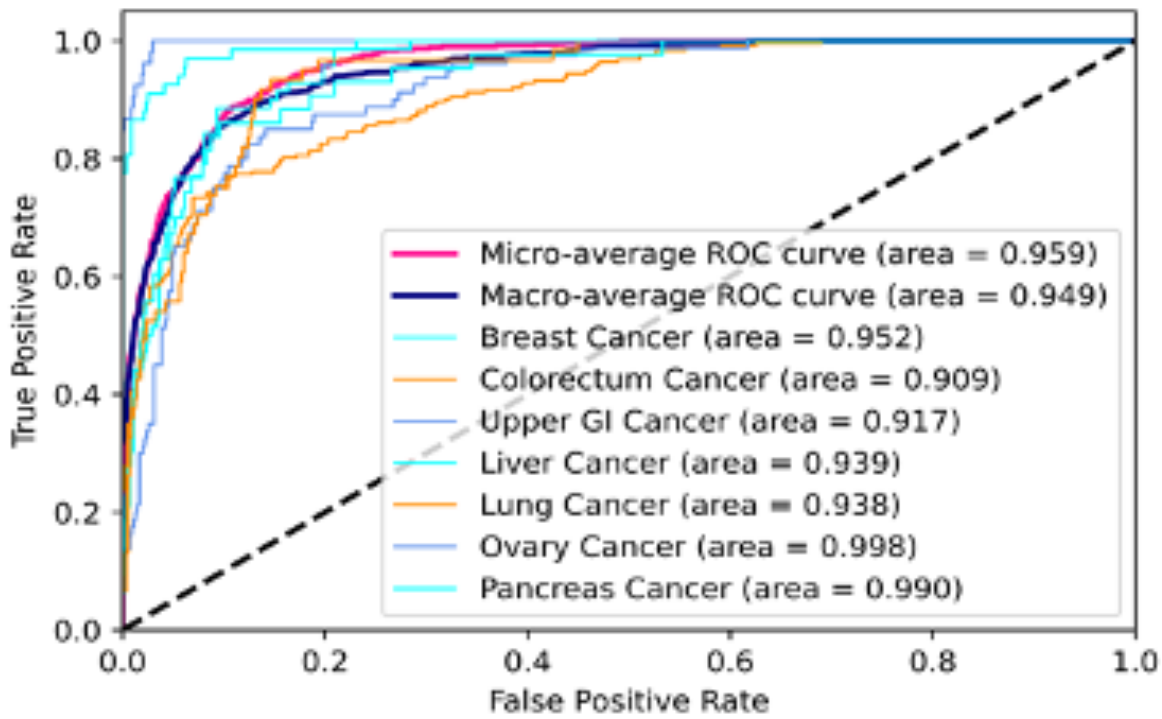

Figure S10: ROC curves of localize cancer detection from before Oversampling on SD4

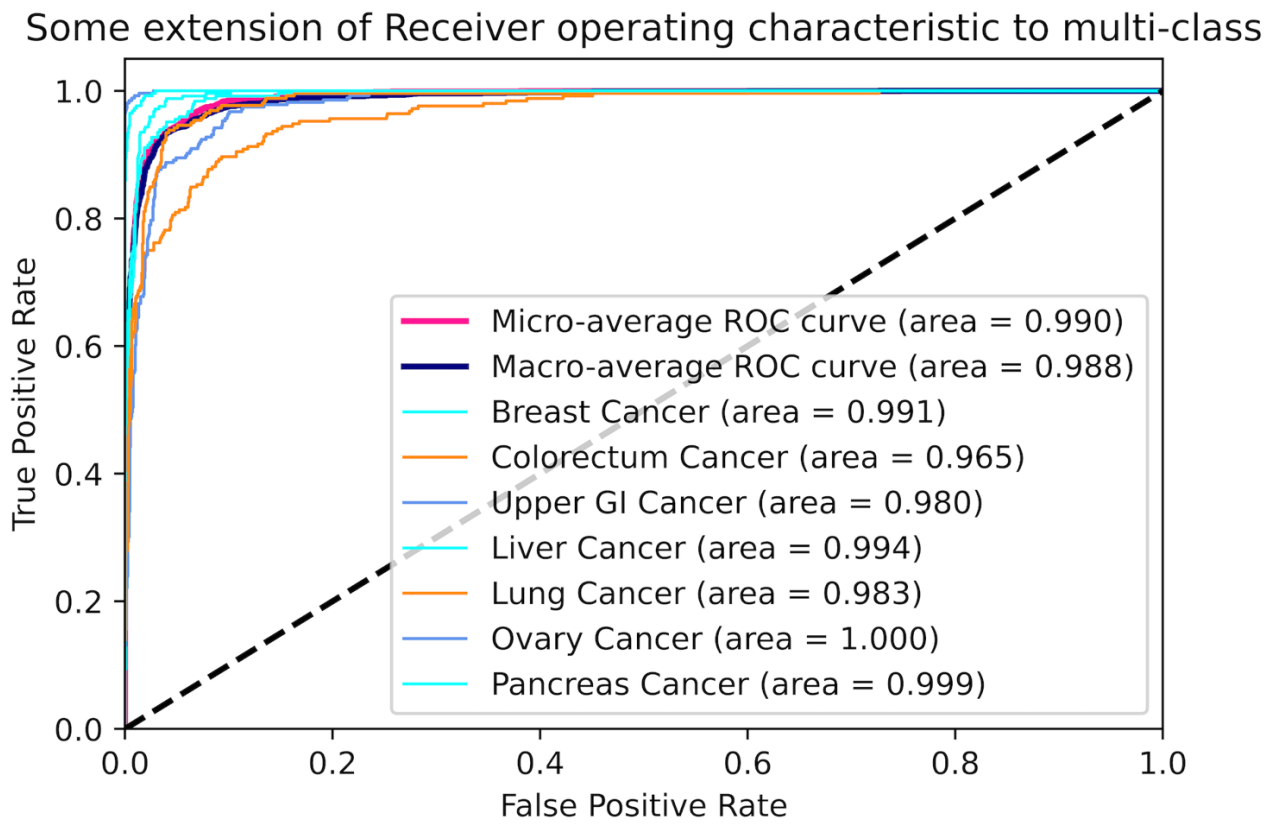

Figure S11: ROC curves of localize cancer detection from after ADASYN Oversampling on SD4.

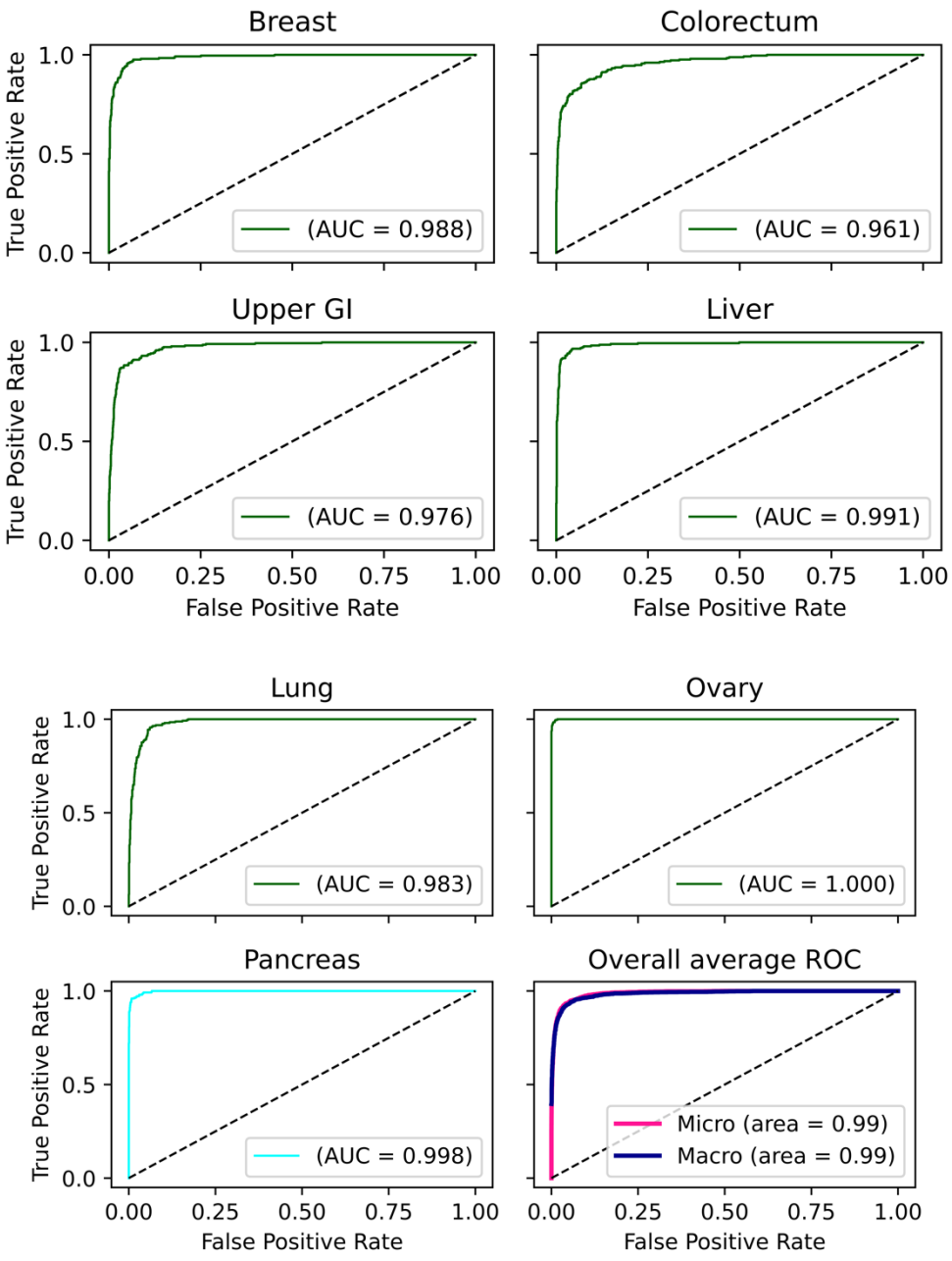

Figure S12: Individual ROC curves of localize cancer detection from after ADASYN Oversampling on SD4.