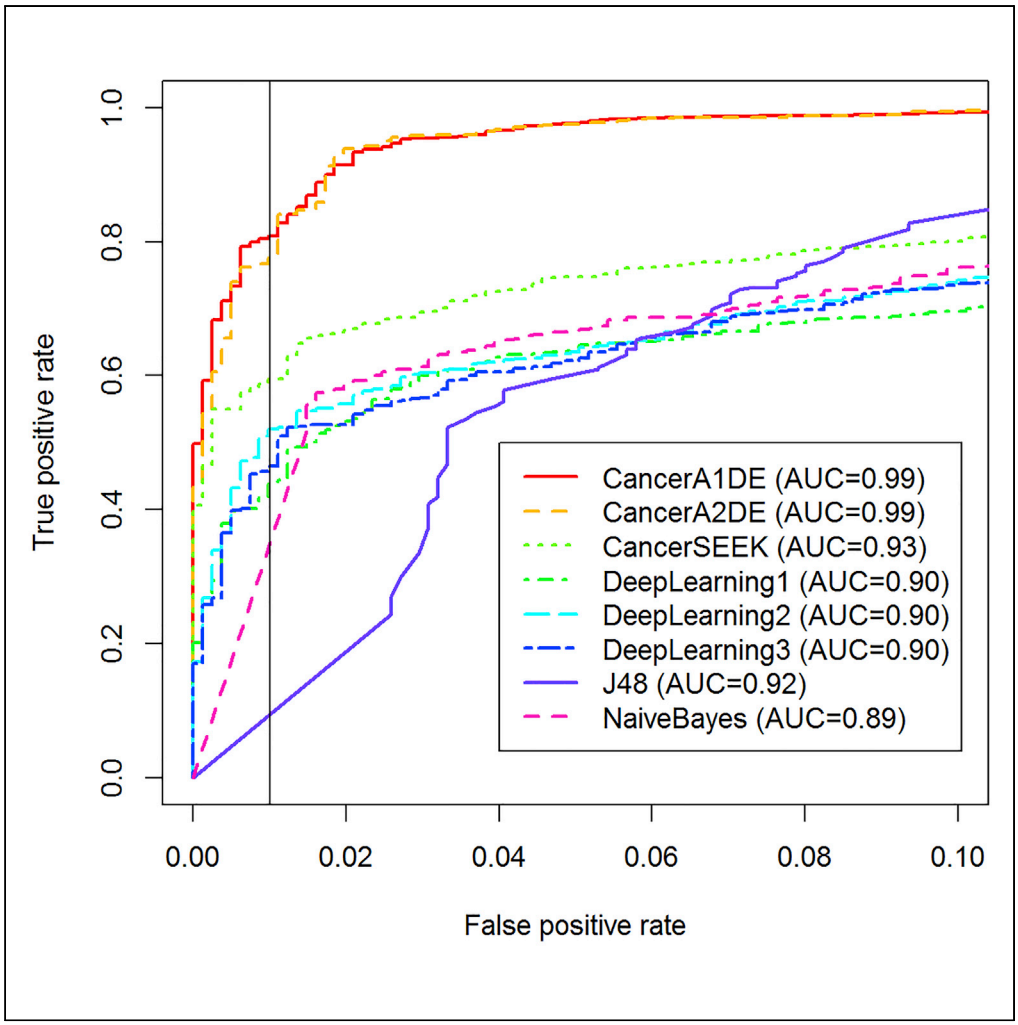
iScience

CellPress

Article

Early Cancer Detection from Multianalyte Blood Test Results



Ka-Chun Wong, Junyi Chen, Jiao Zhang, ..., Qiuzhen Lin, Sam Kwong, Jun Yu

kc.w@cityu.edu.hk

HIGHLIGHTS

We propose an approach (CancerA1DE) to detect early cancers from blood

CancerA1DE doubles the existing sensitivity for the stage I cancer detection

For stage II cancers, it can reach up to 90% across multiple cancer types

The related software is opened and released for future follow-up works

Wong et al., iScience *15*, 332–341
May 31, 2019
https://doi.org/10.1016/j.isci.2019.04.035

iScience

CellPress

Article

Early Cancer Detection from Multianalyte Blood Test Results

Ka-Chun Wong,[1,7,*] Junyi Chen,[1] Jiao Zhang,[1] Jiecong Lin,[1] Shankai Yan,[1] Shxiong Zhang,[1] Xiangtao Li,[2] Cheng Liang,[3] Chengbin Peng,[4] Qiuzhen Lin,[5] Sam Kwong,[1] and Jun Yu[6]

SUMMARY

The early detection of cancers has the potential to save many lives. A recent attempt has been demonstrated successful. However, we note several critical limitations. Given the central importance and broad impact of early cancer detection, we aspire to address those limitations. We explore different supervised learning approaches for multiple cancer type detection and observe significant improvements; for instance, one of our approaches (i.e., CancerA1DE) can double the existing sensitivity from 38% to 77% for the earliest cancer detection (i.e., Stage I) at the 99% specificity level. For Stage II, it can even reach up to about 90% across multiple cancer types. In addition, CancerA1DE can also double the existing sensitivity from 30% to 70% for detecting breast cancers at the 99% specificity level. Data and model analysis are conducted to reveal the underlying reasons. A website is built at http://cancer.cs.cityu.edu.hk/.

INTRODUCTION

Cancers are prevalent across the globe (Torre et al., 2015). Millions of deaths could be found due to various cancer types every year (Chen et al., 2016). Unfortunately, the number of deaths is still projected to be increasing even in developed countries (Rahib et al., 2014). Therefore, it is critical to address those cancers in a timely manner.

In particular, the works in cancer protein marker discovery have been fruitful in the past years (Stoeva et al., 2006); for instance, four analytes (leptin, prolactin, osteopontin, and insulin-like growth factor-II) have been discovered to be predictive in the early detection of ovarian cancers (Mor et al., 2005). Three years later, macrophage inhibitory factor and CA-125 have been proposed on top of the previous four proteins to improve the early detection further (Visintin et al., 2008). The combinatorial expression patterns among HSP-27, GST, Annexin II, and L-FABP are also associated with the lymph node metastasis in colorectal cancer (Pei et al., 2007). Blood and fecal protein markers have also been implicated for colorectal cancer diagnosis (Karl et al., 2008). Multiple markers have also been reported for breast cancers (Harbeck et al., 2014), pancreatic cancers (Takadate et al., 2013), lung cancers (Buszewski et al., 2012), gastric cancers (Rugge et al., 2015), liver cancers (Bertino et al., 2012), esophageal cancers (Napier et al., 2014), and others (Zheng et al., 2005). For a list of well-proved markers, one can refer to the past survey (Polanski and Anderson, 2006).

Given the valuable marker information, it is obvious that one can harness the existing advancement in artificial intelligence to comprehend, digest, and combine those information into a comprehensive and accurate early cancer prediction tool. In particular, there are two major modeling paradigms for such a cancer prediction task under the context of binary classification in machine learning: discriminative learning and generative modeling. For discriminative learning, we seek to explore if we can learn any high-dimensional boundary to discriminate cancer cases from normal cases in the available marker space. For generative modeling, we seek to explore if we can build a model to capture the marker distributions of cancer cases and another model to capture the marker distributions of normal cases. Once the two models are ready, we can compute a sample probability belonging to cancer or normal cases as the early cancer prediction.

A recent method called CancerSEEK takes advantages of multiple protein markers in blood for cancer detection and localization across different cancer types and stages using discriminative learning (Cohen et al., 2018). However, we notice that the CancerSEEK method has several limitations; for instance, its front-line cancer detection component is based on logistic regression, whereby linear assumption on different markers is hardly realistic. Its second-line cancer type localization component is based on random

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR

[2]School of Information Science and Technology, Northeast Normal University, Jilin, China

[3]School of Information Science and Engineering, Shandong Normal University, Shandong, China

[4]Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo, China

[5]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

[6]Institute of Digestive Disease and Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR

[7]Lead Contact

*Correspondence: kc.w@cityu.edu.hk

https://doi.org/10.1016/j.isci.2019.04.035

InfoG	Input Features	Feature Description
1.0389	TGFa (pg/mL)	Circulating Transforming Growth Factor α Concentration in pg/mL
0.8301	HE4 (pg/mL)	Circulating Human Epididymis Protein 4 Concentration in pg/mL
0.6135	sFas (pg/mL)	Circulating soluble Fas Cell Surface Death Receptor Concentration in pg/mL
0.5372	Thrombospondin-2 (pg/mL)	Circulating Thrombospondin-2 Concentration in pg/mL
0.5073	AFP (pg/mL)	Circulating Alpha Fetoprotein Precursor Concentration in pg/mL
0.3759	G-CSF (pg/mL)	Circulating Granulocyte-Colony Stimulating Factor Concentration in pg/mL
0.3633	IL-6 (pg/mL)	Circulating Interleukin-6 Concentration in pg/mL
0.3597	CA-125 (U/mL)	Circulating Cancer Antigen 125 Concentration in U/mL
0.2568	Sex	Patient Gender Information (Male or Female)
0.2352	sHER2/sEGFR2/sErbB2 (pg/mL)	Circulating sHER2/sEGFR2/sErbB2 Concentration in pg/mL
0.2259	TIMP-2 (pg/mL)	Circulating Tissue Inhibitor of Metalloproteinases 2 Concentration in pg/mL
0.2231	CD44 (ng/mL)	Circulating CD44 Concentration in pg/mL
0.183	CA19-9 (U/mL)	Circulating Cancer Antigen 19-9 Concentration in U/mL
0.1805	IL-8 (pg/mL)	Circulating Interleukin-8 Concentration in pg/mL
0.164	CA 15-3 (U/mL)	Circulating Cancer Antigen 15-3 Concentration in U/mL
0.1448	HGF (pg/mL)	Circulating Hepatocyte Growth Factor Concentration in pg/mL
0.1431	OPG (ng/mL)	Circulating Osteopontin Concentration in pg/mL
0.1414	GDF15 (ng/mL)	Circulating Growth Differentiation Factor 15 Concentration in ng/mL
0.1384	Leptin (pg/mL)	Circulating Leptin Concentration in pg/mL
0.1271	Myeloperoxidase (ng/mL)	Circulating Myeloperoxidase Concentration in ng/mL
0.125	Kallikrein-6 (pg/mL)	Circulating Kallikrein-6 Concentration in pg/mL
0.1173	TIMP-1 (pg/mL)	Circulating Tissue Inhibitor of Metalloproteinases 1 Concentration in pg/mL
0.1122	Midkine (pg/mL)	Circulating Midkine Concentration in pg/mL
0.1095	Prolactin (pg/mL)	Circulating Prolactin Concentration in pg/mL
0.1032	Mesothelin (ng/mL)	Circulating Mesothelin Concentration in ng/mL
0.103	Galectin-3 (ng/mL)	Circulating Galectin-3 Concentration in ng/mL
0.096	OPN (pg/mL)	Circulating Osteopontin Concentration in pg/mL
0.0956	NSE (ng/mL)	Circulating Neuron-Specific Enolase Concentration in ng/mL
0.0901	sEGFR (pg/mL)	Circulating Soluble Epidermal Growth Factor Receptor Concentration in pg/mL
0.0901	CEA (pg/mL)	Circulating Carcinoembryonic Antigen Concentration in pg/mL
0.085	AXL (pg/mL)	Circulating AXL Receptor Tyrosine Kinase Concentration in pg/mL
0.0771	sPECAM-1 (pg/mL)	Circulating Soluble Platelet and Endothelial Cell Adhesion Molecule 1 Concentration in pg/mL

Table 1. Feature List for Cancer Type Localization ranked by Information Gain (InfoG)

(Continued on next page)

InfoG	Input Features	Feature Description
0.0637	SHBG (nM)	Circulating Sex Hormone-Binding Globulin Concentration in nM
0.0635	OmegaScore	Omega Score for Mutations in Circulating Cell-Free DNA [Cohen et al. (2018)]
0	Angiopoietin-2 (pg/mL)	Circulating Angiopoietin-2 Concentration in pg/mL
0	DKK1 (ng/mL)	Circulating Dickkopf WNT Signaling Pathway Inhibitor 1 Concentration in ng/mL
0	CYFRA 21-1 (pg/mL)	Circulating Cytokeratin-19 Fragment Concentration in pg/mL
0	PAR (pg/mL)	Circulating Protease-Activated Receptor Concentration in pg/mL
0	Endoglin (pg/mL)	Circulating Endoglin Concentration in pg/mL
0	FGF2 (pg/mL)	Circulating Fibroblast Growth Factor 2 Concentration in pg/mL
0	Follistatin (pg/mL)	Circulating Follistatin Concentration in pg/mL

Table 1. Continued

forest, a modeling known to be difficult for interpretations. From the user perspective, its lack of public Web service also limits its potential impacts.

To address those limitations altogether, we seek to explore different approaches to solve the multiple cancer type detection problem. A public Web server with open programs is also provided for scientific reproducibility and impact at http://cancer.cs.cityu.edu.hk/.

RESULTS

Data Collection

We have collected the multianalyte blood test data from Cohen et al. (2018). Those data have been processed according to the supplementary guideline provided, resulting in two datasets.

The first dataset has 1,817 patient blood test records, which are designed and adopted to build models to detect cancers as the front-line detector in a binary manner (i.e., cancer or normal). Therefore, to be scalable and economical, it has the minimal number of input feature information involving eight circulating protein marker concentrations and one cell-free DNA mutation score (OmegaScore) as listed in Table S1.

The second dataset has 626 patient blood test records, which are designed and adopted to build models to localize cancer types as the second-line diagnosis (i.e., Breast, Colorectum, Upper GI, Liver, Lung, Ovary, or Pancreas). Therefore, its input feature set covers the previous nine features and includes additional 31 protein markers and patient gender as listed in Table 1.

To visualize the first and second datasets, we have adopted a series of dimensional reduction techniques to project the datasets onto two-dimensional spaces as visualized in Figures S1 and S9, respectively. Linear discriminant analysis was also conducted as depicted in Figure S2. Unfortunately, it can be observed that the datasets are not easily separated even on the full datasets, necessitating advanced algorithmic development under the cross-validations with isolated separations of training and testing data in subsequent sections.

Model Descriptions

To build cancer detection models from the aforementioned datasets, we consider it as a supervised learning task from the machine learning perspective. Therefore, we have selected a range of multiclass supervised learning algorithms: AODE (Webb et al., 2005), deep learning (Angermueller et al., 2016), decision tree (Bhargava et al., 2013), and naive Bayes (NB) (Lewis, 1998). We note that logistic regression and random forest have already been adopted and encapsulated in CancerSEEK (Cohen et al., 2018).

Model Parameter Setting

For AODE, we have adopted A1DE (Webb et al., 2005) and A2DE (Webb et al., 2012) as our classifiers (namely, CancerA1DE and CancerA2DE). The minimum description length (MDL) principle is adopted

for continuous marker feature discretization (see the Supplemental Information). For deep learning, since the problem here is a very standard supervised learning task with well-crafted input features, we do not need to increase any model complexity for deep feature learning. Therefore, we have adopted the deep feedforward neural networks with one hidden layer, two hidden layers, and three hidden layers (namely, DeepLearning1, DeepLearning2, and DeepLearning3, respectively) (Angermueller et al., 2016). The remaining training setting follows the default settings of WEKA (Hall et al., 2009). For decision tree, J48 tree building method is adopted (Bhargava et al., 2013). For NB, a Gaussian distribution is assumed for each continuous feature under each class (Lewis, 1998). The remaining parameter setting follows the default parameter values of WEKA (Hall et al., 2009).

CancerA1DE Modeling

CancerA1DE has demonstrated the best performance in the subsequent sections. Therefore, we briefly describe CancerA1DE in this section. CancerA1DE is based on the A1DE framework (Webb et al., 2005), which is a variant of NB. Weakening the attribute (or feature) independence assumption on the NB classifier has been proved to achieve significant improvement by various approaches. Nevertheless, the techniques such as the Lazy Bayesian Rules and Tree Augmented Naive Bayes weaken the assumption with the cost of intensive computational power because the optimal subset selection procedure from all attributes or parent attributes could be computationally intensive. In addition, the model selection tends to over-fit the training data and thus increase the estimation variance. To speed it up, Aggregating One-Dependence Estimators (A1DE) was developed; A1DE is designed to avoid any model selection by enumerating all possible 1-dependence classifiers in each of which there is an attribute as the parent of all others (Webb et al., 2005). The mathematical details can be found in the Supplemental Information.

The advantage of A1DE is its relatively low complexities as tabulated in Table S2. We can observe that, although A1DE has significant performance gain over the original NB, its training time complexity is nether quadratic to the number of samples nor cubic to the number of feature attributes. Its training time complexity is just $O(tn^2)$ where t is the number of samples and n is the number of feature attributes. Such a property guarantees that it can scale with the number of samples in a linear manner. Its quadratic complexity to the number of feature attributes (i.e., markers here) should not be a big problem since the number of blood test markers is usually finite and limited. In addition, its model formulation is incremental; it means that the CancerA1DE model can be easily updated with new blood marker samples, unlike random forest (i.e. CancerSEEK).

Cancer Detection Results

As ranked in Table S1, it is not surprising that the cancer antigen markers are the most informative features for cancer detection. To act as a control, we have also trained random forests on the features using the Python scikit-learn package (Pedregosa et al., 2011) and explored the feature ranking based on three different measurements: decrease in purity, decrease in accuracy, and recursive feature elimination. The feature ranking results are illustrated in Figure S3. Contrary to the information gain ranking on Table S1, the cancer antigen markers are no longer the top predictive features. Instead, we observe the opposite trend for the purity and accuracy measurements; such a phenomenon exemplifies the underlying complex behavior for cancer detection. It also necessitates our subsequent machine learning approaches.

To explore those features (also known as protein biomarkers) further, we have computed their correlation matrix as visualized in Figure S4. Congruent with our general belief, the cancer antigen markers are positively correlated with statistical significance ($p < 0.001$). Interestingly, it can be observed that TIMP-1, Myeloperoxidase (or MPO), OPN (or SPP1), and HGF form a positively correlated feature cluster. To explore their relationships, we have mapped the protein markers into STRING network analysis (Szklarczyk et al., 2016) as demonstrated in Figure S5. From the figure, it is now clear that those four proteins did demonstrate different levels of evidences for their interactions; it justifies their positive correlations. On the other hand, we also observe an enriched pathway from the STRING results: PI3K-Akt signaling pathway (ID: 04151) with the supporting proteins (HGF, PRL, and SPP1) at FPR = 0.0243; it indicates that such a feature subset serves as proxy measurements for that pathway in the following cancer detection tasks.

Given those features, we have benchmarked different supervised learning models on the 1,817-patient dataset for cancer detection (i.e., cancer or normal) under 10-fold cross-validations (i.e., 10 randomly divided held-out data subsets for testing, whereas the remaining are for training in 10 rounds) as visualized in Figure 1.

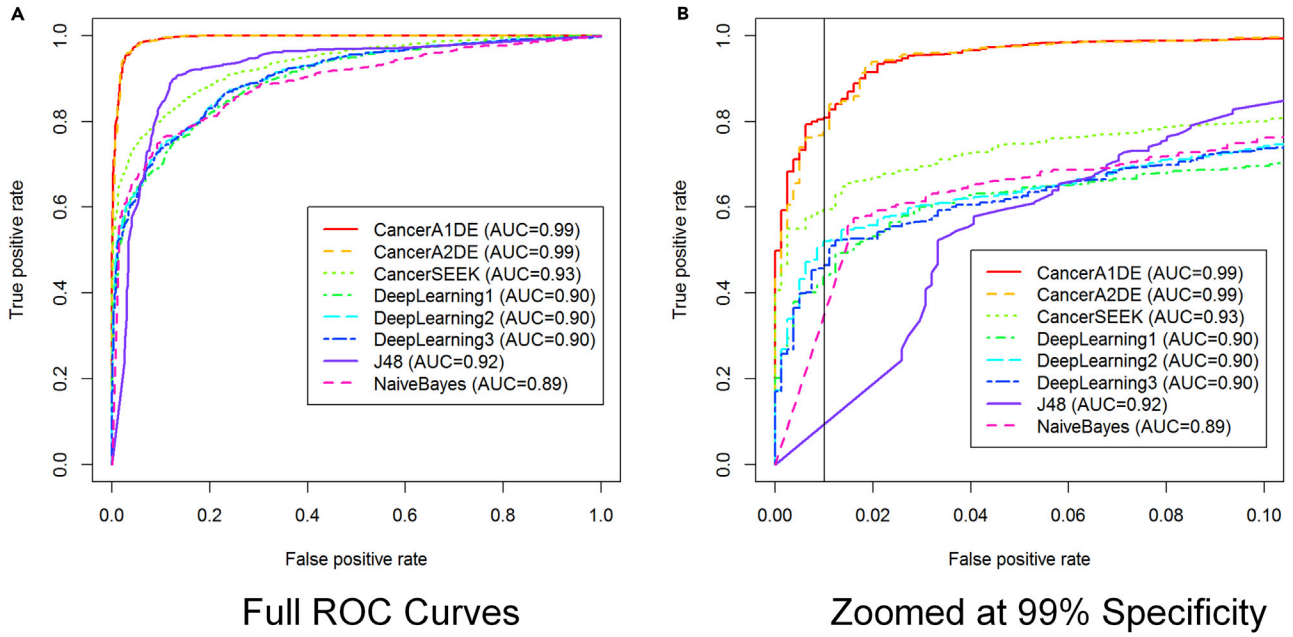


Figure 1. Receiver Operating Characteristic (ROC) curves for Cancer Detection

Different methods have different colors and line styles. The curves are generated under 10-fold cross-validations. The vertical black line on the right panel is drawn at the 99% specificity level.

(A) Full Scale ROC Curves.

(B) ROC Curves Zoomed to FPR<=0.1.

Interestingly, it can be observed that our proposed CancerA1DE and CancerA2DE outperformed CancerSEEK by a significant margin in terms of the Area Under Curve (AUC) values (0.99 vs 0.93). If the figure is zoomed to the regions with false-positive rates less than 0.1, such a performance gain is even pronounced. At the 99% specificity level, our CancerA1DE and CancerA2DE can even achieve around 80% sensitivity, which is higher than that of CancerSEEK by 20% (Cohen et al., 2018). We attribute such a performance gain to two reasons: (1) the generative modeling approach undertaken by CancerA1DE and CancerA2DE and (2) the feature discretization of CancerA1DE and CancerA2DE based on the MDL principle (Kononenko, 1995). Reason (1) is obvious in that generative modeling can generalize itself well over the discriminative modeling, which is prone to over-fitting, whereas reason (2) is not intuitive but insightful. Its feature discretization performance suggests that we should focus on the protein marker concentration intervals rather than on the actual protein marker concentration magnitudes. As depicted in Figure S6, we can observe that the input features are well separated into different groups based on MDL. In particular, it is interesting that the feature groups are associated with different levels of misclassification risks, which can well inform CancerA1DE and CancerA2DE to make cancer detection decisions in a probabilistically generative manner.

To confirm its performance further, we have conducted a performance sensitivity analysis based on different withheld data amount settings. Specifically, we randomly divided the dataset into two subsets: training set and testing set in different proportions as tabulated in the Table S3. From the table, we can observe that our CancerA1DE and CancerA2DE models can reach the AUC performance of 0.98 as soon as it has about 20% data for model training (i.e., 363 patient samples). Such a rapid learning characteristic is important as patient sample costs are substantial. Its underlying Bayesian generative modeling also prevents it from over-fitting.

In the context of cancer detection, its early detection performance is the most important task in the clinical setting. Therefore, we proceed to wonder how early different methods can detect cancers for timely follow-ups. On the other hand, we would also like to ensure that the false-positive rate can be minimized. Therefore, we limit our cancer detections to the 99% specificity level where the detected proportions of cancers with different stages are depicted in Figure 2.

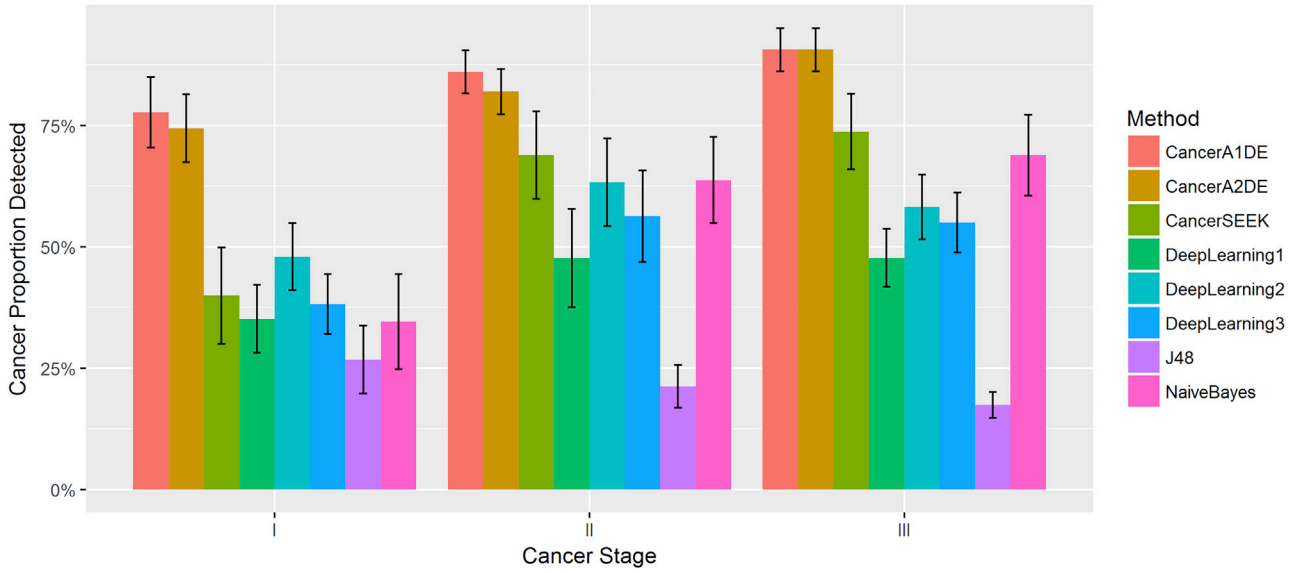


Figure 2. Proportion of Detected Cancers with Different Stages at the 99% Specificity Level

Each color represents a method, and the horizontal axis has been ordered by cancer stages. Each bar represents the median sensitivity of each method on each cancer stage with standard errors.

We can observe that our CancerA1DE and CancerA2DE can detect close to 77% of the early-stage cancers (i.e., Stage I), whereas the others are well below 50%. Such a phenomenon is clinically critical since early cancer treatment intervention can often lead to increased patient survival rates (Miller et al., 2016).

To investigate whether any of the methods is biased to specific cancer types, we have also plotted the detected proportion of different cancer types at the 99% specificity level in Figure 3.

Clearly, it can be observed that both CancerA1DE and CancerA2DE have broadly competitive edges over the other state-of-the-art methods across different cancer types. It can achieve at least 60% sensitivity for all tested cancer types at the 99% specificity level. In particular, CancerA1DE and CancerA2DE can double the detected proportions of breast cancer, which is the second most common cancer diagnosed among women in the United States (DeSantis et al., 2014).

To provide further insights, we have plotted the conditional probabilities $\widehat{P}(x_j|y, x_i)$ of CancerA1DE for all possible combinations. An example is depicted in Figure S7, whereas the others are exhaustively enumerated in Figure S8. From those figures, it can be observed that the conditional probabilities are highly specific to normal samples (see the spikes and the monotonically decreasing trends in the figures). This implies that the CancerA1DE modeling places strong pattern recognition emphasis on filtering out the normal samples; therefore, it explains why the CancerA1DE can achieve the highly sensitive detection performance at the 99% specificity level as previously demonstrated.

Cancer Type Localization Results

As ranked in Table 1 and Figure S10, it is interesting to observe that the previous nine features in cancer detection are no longer the top features for cancer type localization. This necessitates our motivation that we have to include additional features (i.e., additional 31 protein markers and patient gender) for cancer type localization.

However, we also need to ensure the nonredundancy of those features. Therefore, we have computed the feature correlation matrix as heatmap in Figure S11. It can be observed that the protein marker features are not highly correlated to each other. Nonetheless, if we zoom in specific regions, we did observe a few weak feature communities here. Therefore, similar to the previous section, we have mapped the protein marker names into STRING network analysis (Szklarczyk et al., 2016) in Figure S12. Unfortunately, most of the interactions are based on text mining, which is not conclusive. Nonetheless, based on the evidence strength,

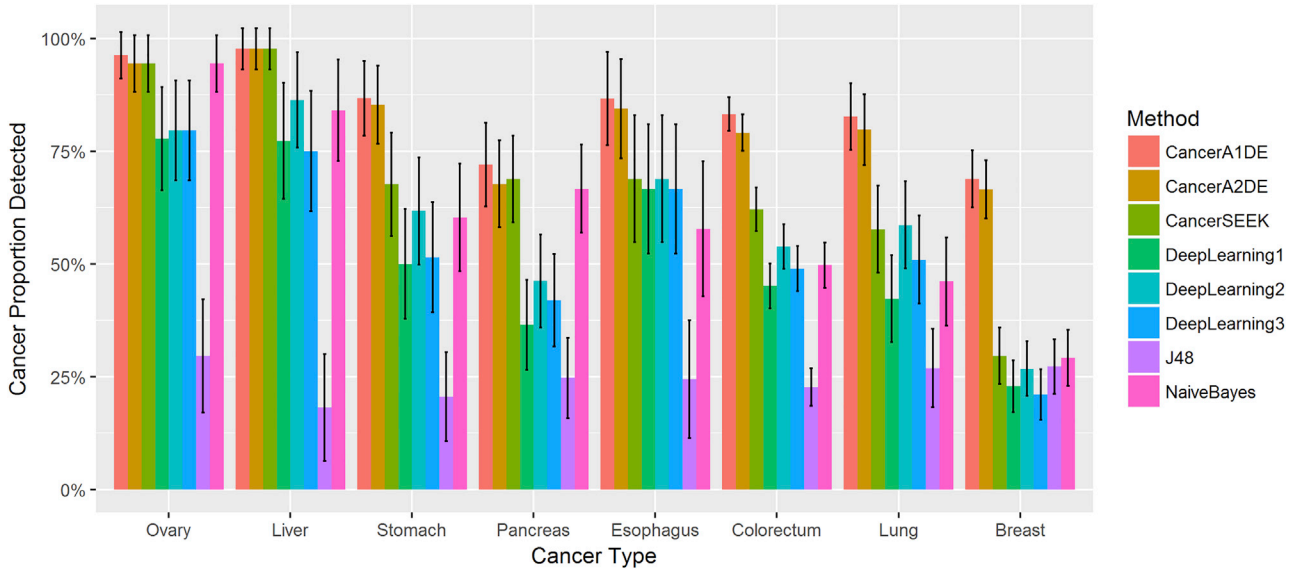


Figure 3. Detected Proportions of Different Cancer Types at the 99% Specificity Level

Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

we can still observe few existing communities that are congruent with our observations from the correlation heatmap. Different from the previous cancer detection task, several pathways are enriched as listed in Table S4; this indicates the comprehensiveness and diversity of the protein markers for the following cancer type localization tasks.

The previous methods are also adopted here for fair benchmark comparisons. In particular, the multiclass supervised learning setting is experimented on the 626-patient dataset for cancer type localization (i.e., Breast, Colorectum, Upper GI, Liver, Lung, Ovary, and Pancreas) under 10-fold cross-validations as visualized in Figure S13.

It can be observed that both CancerA1DE and CancerA2DE can well maintain the receiver operating characteristic curves above the diagonal baseline, whereas the others suffer from several drawbacks; for instance, CancerSEEK appears to be polarized into two extreme performance groups; this indicates that CancerSEEK, which is based on random forest, could have bias towards specific cancer type differentiations, consistent with the observation to be made from Figure 4. On the other hand, the deep learning methods cannot scale to full performance once the specificity level is relaxed. Interestingly, the NB classifier demonstrates surprising classification performance across few cancer types. Congruent with CancerA1DE and CancerA2DE, such a phenomenon reflects that multiclass generative modeling is more suitable than multiclass discriminative modeling for the cancer type localization here.

To compare the methods in a more realistic setting than the previous curves, we seek to validate the methods for the top-one predictions. In other words, for each patient record, we allow each method to predict and localize its cancer type once only. Under the 10-fold cross-validation, the results are visualized in Figure 4.

Congruent with the ROC curves, CancerA1DE and CancerA2DE demonstrate stable performance across different cancer types. Although CancerSEEK can score well on colorectal and ovarian cancers, its random forest has sacrificed its performance on other cancer types. In addition, we also observe that most methods cannot perform well on the liver cancer localization. Such a performance deviation can be attributed to the scarce data availability issue here.

To investigate the reasons further, we have adopted Learning Vector Quantization to perform feature importance ranking under the one-class-versus-all setting with 10-fold cross-validations. The complete

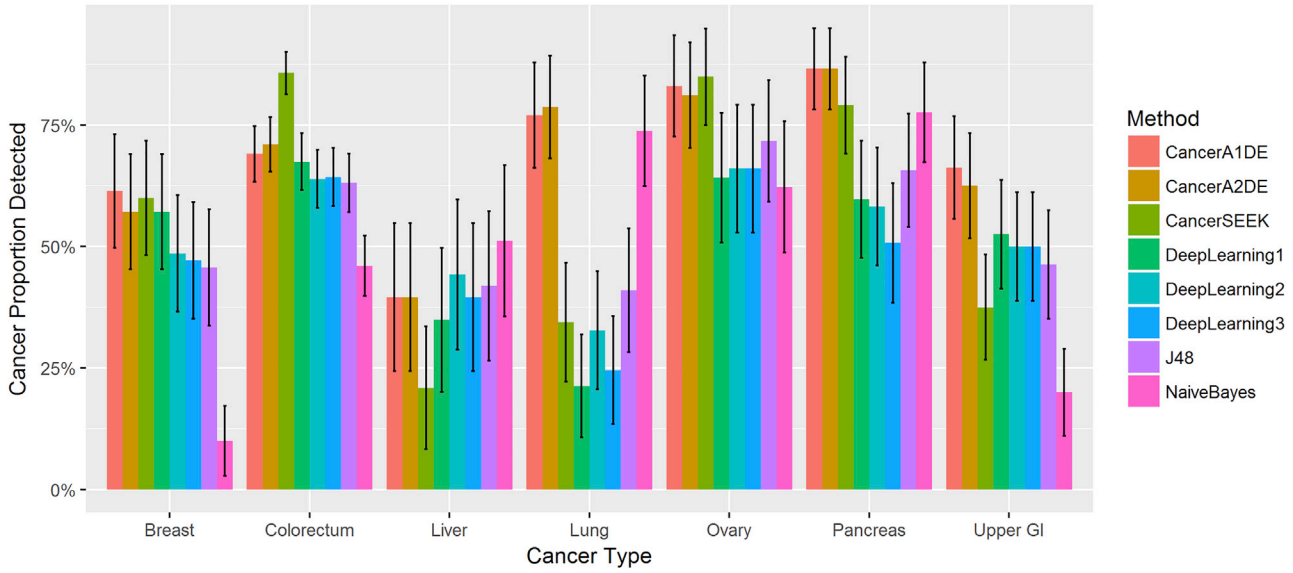


Figure 4. Localized Proportions of Different Cancer Types using the Top One Prediction Approach

Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

results are listed in Figure S14. To summarize the results, we have performed hierarchical clustering and heatmap visualization on the feature importance results as depicted in Figure 5. Interestingly, we observe that the cancer types can be clustered into three groups. The ovarian cancers and pancreatic cancers form independent groups; it explains why most methods can perform well on those two cancers since their important features are well isolated from the others. In contrast, the other five cancer types share significant portions of important features. This forms a machine learning trap that discriminative learning algorithms (e.g., the random forest adopted by CancerSEEK) could be biased towards the cancer type that has the largest data samples (e.g., colorectal cancer here as demonstrated in Figure 4).

DISCUSSION

In this study, we have explored different approaches for multiple cancer type detection from multianalyte blood test results. With eight circulating protein markers and one circulating DNA mutation score, the best approach (CancerA1DE) can outperform the existing approach (CancerSEEK) for cancer detection at the 99% specificity level.

Nonetheless, such an observation is based only on the available 1,817 patient blood samples. It is undeniable that different approaches can result in different performance on different datasets. This is further complicated by the fact that we have diverse cancer types within the 1,817 patient blood samples. The main competitiveness of our approach is that it can take into account the nonlinear relationships between the available markers, although the only cost is the slightly increased computational complexity. However, we conceive such a cost as manageable under the current computing environment (e.g., it takes only less than 1 sec for the CancerA1DE model building on our desktop computer using the Weka library).

On the other hand, for cancer type localization with additional 31 circulating protein markers and gender information, we observed diverse performance behavior among different approaches. Such performance discrepancy is attributed to several reasons ranging from the cancer-type-specific feature importance to the supervised learning methodology paradigm (i.e., generative or discriminative modeling) as revealed. In particular, we observe the resurge of the generative Bayesian approach (CancerA1DE), which has been demonstrated to be more successful than the state-of-the-art methods such as random forest (CancerSEEK) and deep learning in this study.

With the advancement of edge computing devices (e.g., mobile phones) and biotechnology (e.g., biosensor devices), we envision that the integration between them can provide one-stop blood tests

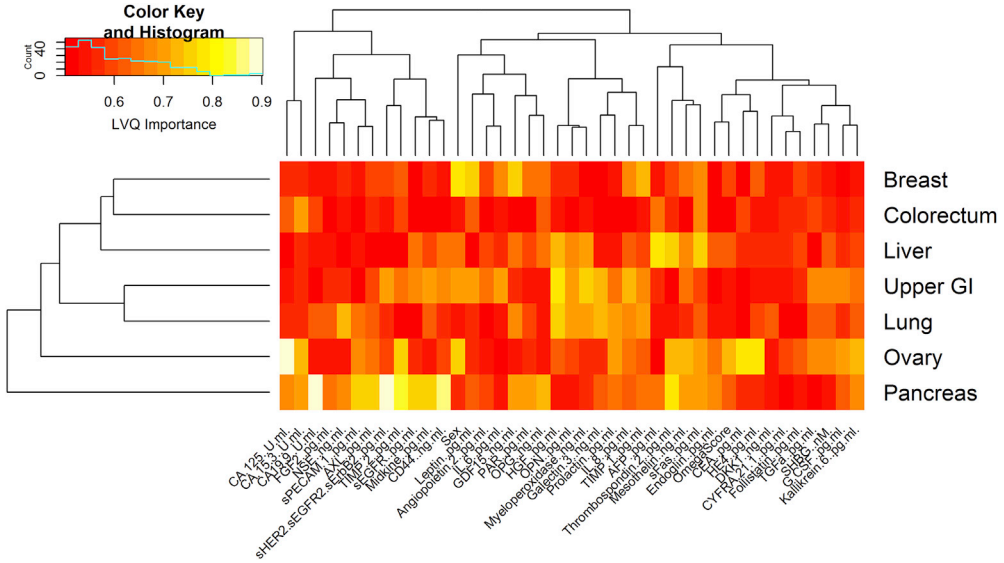


Figure 5. Feature Importance Heatmap for Cancer Type Localization under One-Class-versus-Others Setting

The feature rankings are measured based on the Learning Vector Quantization (LVQ) building under Python caret package (Bischl et al., 2016). Ten-fold cross-validations are run to compute the feature importance values. After that, the function "heatmap.2" in R language is adopted with the default setting to cluster and visualize the feature importance values. Further details can be found in Figure S14.

accessible to patients. Our current study can provide a reference to deploy a sensitive and robust blood test model for routine cancer screening at the software level.

In the future, it can be interesting if additional marker features can be explored; however, the related costs are substantial on actual patient cohorts. Comprehensive genomics and pathway analysis on the newly arisen markers are also needed for providing mechanistic insights before incorporating into the current study. Continuous assessment on the proposed approaches is also required for extensive uses. Therefore, with the public Web server and open software programs for scientific reproducibility, we believe that this study can serve as a useful platform for future studies on early cancer detection from blood.

Limitation of Study

The current study is limited by the absence of a truly independent validation set.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.04.035.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments. They would also like to thank Prashant Sridhar and Ajay Rajnikanth for their English proofreading. The work described in this paper was substantially supported by three grants from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 21200816], [CityU 11203217], and [CityU 11200218]. We acknowledge the donation support of the Titan Xp GPU from the Nvidia Corporation.

AUTHOR CONTRIBUTIONS

K.-C.W. conceived and designed the study; K.-C.W., X.L., C.L., C.P., Q.L, S.K., and J.Y. supervised the study; K.-C.W., J.C., J.Z., J.L., S.Y., and S.Z. implemented the study; K.-C.W., J.C., X.L., and J.Y. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 21, 2018
Revised: March 18, 2019
Accepted: April 29, 2019
Published: May 31, 2019

REFERENCES

Angermueller, C., Parnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. *12*, 878.

Bertino, G., Ardiri, A., Malaguarnera, M., Malaguarnera, G., Bertino, N., and Calvagno, G.S. (2012). Hepatocellualar carcinoma serum markers. In Seminars in Oncology, *Vol 39*, A.T. Fojo, ed., Seminars in Oncology (Elsevier), pp. 410–433.

Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng. *3*.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z.M. (2016). mlr: Machine learning in r. J. Mach. Learn. Res. *17*, 1–5.

Buszewski, B., Ligor, T., Jezierski, T., Wenda-Piesik, A., Walczak, M., and Rudnicka, J. (2012). Identification of volatile lung cancer markers by gas chromatography–mass spectrometry: comparison with discrimination by canines. Anal. Bioanal. Chem. *404*, 141–146.

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. (2016). Cancer statistics in China, 2015. CA Cancer J. Clin. *66*, 115–132.

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science *359*, 926–930.

DeSantis, C., Ma, J., Bryan, L., and Jemal, A. (2014). Breast cancer statistics, 2013. CA Cancer J. Clin. *64*, 52–62.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The weka data mining software: an update. ACM SIGKDD Explor. Newslett. *11*, 10–18.

Harbeck, N., Sotlar, K., Wuerstlein, R., and Doisneau-Sixou, S. (2014). Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. Cancer Treat. Rev. *40*, 434–444.

Karl, J., Wild, N., Tacke, M., Andres, H., Garczarek, U., Rollinger, W., and Zolg, W. (2008). Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. Clin. Gastroenterol. Hepatol. *6*, 1122–1128.

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *14th International Joint Conference on Artificial Intelligence*, pages 1034–1040.

Lewis, D.D. (1998). Naive (Bayes) at forty: the independence assumption in information retrieval. In European Conference on Machine Learning, C. Nédellec and C. Rouveirol, eds. (Springer), pp. 4–15.

Miller, K.D., Siegel, R.L., Lin, C.C., Mariotto, A.B., Kramer, J.L., Rowland, J.H., Stein, K.D., Alteri, R., and Jemal, A. (2016). Cancer treatment and survivorship statistics, 2016. CA Cancer J. Clin. *66*, 271–289.

Mor, G., Visintin, I., Lai, Y., Zhao, H., Schwartz, P., Rutherford, T., Yue, L., Bray-Ward, P., and Ward, D.C. (2005). Serum protein markers for early detection of ovarian cancer. Proc. Natl. Acad. Sci. U S A *102*, 7677–7682.

Napier, K.J., Scheerer, M., and Misra, S. (2014). Esophageal cancer: a review of epidemiology, pathogenesis, staging workup and treatment modalities. World J. Gastrointest. Oncol. *6*, 112.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.

Pei, H., Zhu, H., Zeng, S., Li, Y., Yang, H., Shen, L., Chen, J., Zeng, L., Fan, J., Li, X., et al. (2007). Proteome analysis and tissue microarray for profiling protein markers associated with lymph node metastasis in colorectal cancer. J. Proteome Res. *6*, 2495–2501.

Polanski, M., and Anderson, N.L. (2006). A list of candidate cancer biomarkers for targeted proteomics. Biomarker Insights *1*, 1–48.

Rahib, L., Smith, B.D., Aizenberg, R., Rosenzweig, A.B., Fleshman, J.M., and Matrisian, L.M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res. *74*, 2913–2921.

Rugge, M., Fassan, M., and Graham, D.Y. (2015). Epidemiology of gastric cancer. In Gastric Cancer, V.E. Strong, ed. (Springer), pp. 23–34.

Stoeva, S.I., Lee, J.-S., Smith, J.E., Rosen, S.T., and Mirkin, C.A. (2006). Multiplexed detection of protein cancer markers with biobarcoded nanoparticle probes. J. Am. Chem. Soc. *128*, 8378–8379.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2016). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Takadate, T., Onogawa, T., Fukuda, T., Motoi, F., Suzuki, T., Fujii, K., Kihara, M., Mikami, S., Bando, Y., Maeda, S., et al. (2013). Novel prognostic protein markers of resectable pancreatic cancer identified by coupled shotgun and targeted proteomics using formalin-fixed paraffin-embedded tissues. Int. J. Cancer *132*, 1368–1382.

Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. CA Cancer J. Clin. *65*, 87–108.

Visintin, I., Feng, Z., Longton, G., Ward, D.C., Alvero, A.B., Lai, Y., Tenthorey, J., Leiser, A., Flores-Saaib, R., Yu, H., et al. (2008). Diagnostic markers for early detection of ovarian cancer. Clin. Cancer Res. *14*, 1065–1072.

Webb, G.I., Boughton, J.R., and Wang, Z. (2005). Not so naive Bayes: Aggregating one-dependence estimators. Mach. Learn. *58*, 5–24.

Webb, G.I., Boughton, J.R., Zheng, F., Ting, K.M., and Salem, H. (2012). Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive bayesian classification. Mach. Learn. *86*, 233–272.

Zheng, G., Patolsky, F., Cui, Y., Wang, W.U., and Lieber, C.M. (2005). Multiplexed electrical detection of cancer markers with nanowire sensor arrays. Nat. Biotechnol. *23*, 1294.

ISCI, Volume *15*

Supplemental Information

Early Cancer Detection from Multianalyte Blood Test Results

Ka-Chun Wong, Junyi Chen, Jiao Zhang, Jiecong Lin, Shankai Yan, Shxiong Zhang, Xiangtao Li, Cheng Liang, Chengbin Peng, Qiuzhen Lin, Sam Kwong, and Jun Yu

S1 Supplemental Information

S1.1 Transparent Methods

The software programs and related models are opened and released for scientific reproducibility. An open-access webserver is also built and provided in the following web address: `http://cancer.cs.cityu.edu.hk/`

S1.1.1 CancerA1DE Modeling

CancerA1DE has demonstrated the best performance in the subsequent sections. Therefore, we briefly describe CancerA1DE in this section. CancerA1DE is based on the A1DE framework (Webb et al., 2005) which is a variant of Naive Bayes (NB). Weakening the attribute (or feature) independence assumption on the naive Bayes classifier has been proved to achieve significant improvement by various approaches. Nevertheless, the techniques such as the Lazy Bayesian Rules (LBR) and Tree Augmented Naive Bayes (TAN) weaken the assumption with the cost of intensive computational power because the optimal subset selection procedure from all attributes or parent attributes $p(x_i)$ could be computationally intensive. In addition, the model selection tends to over-fit the training data and thus increase the estimation variance. To speed it up, Aggregating One-Dependence Estimators (A1DE) is developed; A1DE is designed to avoid any model selection by enumerating all possible 1-dependence classifiers in each of which there is an attribute as the parent of all others. Moreover, A1DE introduces a threshold m such that the models are discarded if the training data contain less than m examples of each parent attribute value x_i noted as $F(x_i) < m$ for computational efficiency.

Hence, given that each blood marker sample can be represented by a vector $x = \langle x_1, x_2, ..., x_n \rangle$ where x_i is a marker attribute value, an A1DE model can be trained and assigned cancer detection label y based on its posterior probability:

$$P(y|x) = \frac{P(y,x)}{P(x)} \propto P(y,x) \tag{1}$$

By aggregating all possible 1-dependence classifiers, $P(y,x)$ can be written as:

$$P(y,x) = \frac{\sum_{1 \le i \le n \wedge F(x_i) \ge m} P(y,x_i) P(x|y,x_i)}{|\{1 \le i \le n \wedge F(x_i) \ge m\}|} \tag{2}$$

Therefore, the label assignment (cancer detection label y) can be derived as follows:

$$\arg\max_y \sum_{1 \le i \le n \wedge F(x_i) \ge m} \hat{P}(y,x_i) \prod_{j=1}^{n} \hat{P}(x_j|y,x_i) \tag{3}$$

where $\hat{P}$ denotes the probability estimate. From the above, we can see that, if none of the parent attributes x_i have its $F(x_i)$ count greater than m, the A1DE is identical to a traditional NB classifier. On the other hand, the posterior of classes can be derived as

follows:

$$\hat{P}(y|x) = \frac{\sum_{1 \leq n \wedge F(x_i) \geq m} \hat{P}(y, x_i) \prod_{j=1}^{n} \hat{P}(x_j|y, x_i)}{\sum_{y' \in Y} \sum_{1 \leq n \wedge F(x_i) \geq m} \hat{P}(y', x_i) \prod_{j=1}^{n} \hat{P}(x_j|y', x_i)} \tag{4}$$

where the above formula is derived from the Bayes rule $P(y|x) = P(y, x)/P(x)$. The advantage of A1DE is its relative low complexities as tabulated in Table S2. We can observe that, although A1DE has significant performance gain over the original NB, its training time complexity is nether quadratic to the number of samples nor cubic to the number of attributes. Its training time complexity is just $O(tn^2)$ where t is the number of samples and n is the number of attributes. Such a property guarantees that it can scale with the number of samples in a linear manner. Its quadratic complexity to the number of attributes (i.e. markers here) should not be a big problem since the number of blood test markers is usually finite and limited. In addition, its model formulation is incremental; it means that the CancerA1DE model can be easily updated with new blood marker samples, unlike random forest.

S1.1.2 Minimum Description Length (MDL) Principle

The minimum description length (MDL) is a principle defined to be the minimum size of information to specify all samples. Under the principle, the supervised attribute (or marker feature in this study) discretization problem can be modeled as a sender and receiver problem: sender maintains the function of choosing the shortest description of proper class labels while the receiver determines class labels for examples (Fayyad and Irani, 1993). Given a set of samples S, we test the attribute A and use the threshold T to generate a binary partition π_T. Assuming that the partition can either be accepted or rejected, HT represents a hypothesis that the partition is accepted while NT is a null hypothesis that π_T is rejected. Hence, HT suggests the classifier assigns all samples whose attribute value $A < T$ to the same attribute (or feature) interval while NT suggests that all samples belongs to the same attribute (or feature) interval regardless of any examination.

The null hypothesis (NT) For the null hypothesis NT, the sender encodes shortest class labels in code length (e.g. Huffman tree) for every sample in the sample set S. Assuming that the average code length is l and the number of classes is k, the sender sends $(|S| \cdot l)$ bits to the receiver while the dictionary of encoding has the size of $(k \cdot l)$. Therefore, the cost measured in bits is:

$$|S| \cdot l + k \cdot l \tag{5}$$

The hypothesis (HT) For the hypothesis HT, it is corresponding to the partition job where pi_T separates the set S into S_1 and S_2 with the code lengths l_1 and l_2 respectively. Since the code is encoded, the cut value can be determined with the cost of $log_2(|S|-1)$. Hence, transmitting HT has the cost measured in bits of:

$$log_2(|S| - 1) + |S_1| \cdot l_1 + |S_2| \cdot l_2 \tag{6}$$

Considering each partitioned subset can be one of the $2^k - 1$ subsets under k classes, the transmission of the encoding dictionary has the cost measured in bits of:

$$\left[\sum_{k_i=1}^{k-1} \binom{k}{k_i} 2^{k_i}\right] + 2^k - 1 = 3^k - 2 \tag{7}$$

After all, the cost of HT measured in bits is:

$$log_2(|S| - 1) + |S_1| \cdot l_1 + |S_2| \cdot l_2 + log_2(3^k - 2) + |S_1| \cdot k_1 + |S_2| \cdot k_2 \tag{8}$$

Decision Rule Accepting a partition π_T on the set S, S will be separated into S_1 and S_2 with k_1 and k_2 distinguish class labels respectively. The information gain of the partition is:

$$IG(\pi_T; S) = Ent(S) - \frac{|S_1|}{|S|} Ent(S_1) - \frac{|S_2|}{|S|} Ent(S_2) \tag{9}$$

To accept a partition π_T on the set S, the cost of null hypothesis should be larger than the cost of hypothesis.

$$Cost(NT) > Cost(HT)$$

$$\begin{aligned} Cost(NT) &= N \cdot Ent(S) + k \cdot Ent(S) \\ Cost(HT) &= log_2(|S| - 1) + |S_1| \cdot Ent(S_1) + |S_2| \cdot Ent(S_2) \\ &\quad + log_2(3^k - 2) + Ent(S_1) \cdot k_1 + Ent(S_2) \cdot k_2 \end{aligned} \tag{10}$$

Derived from the above equations, the partition acceptance criterion should be:

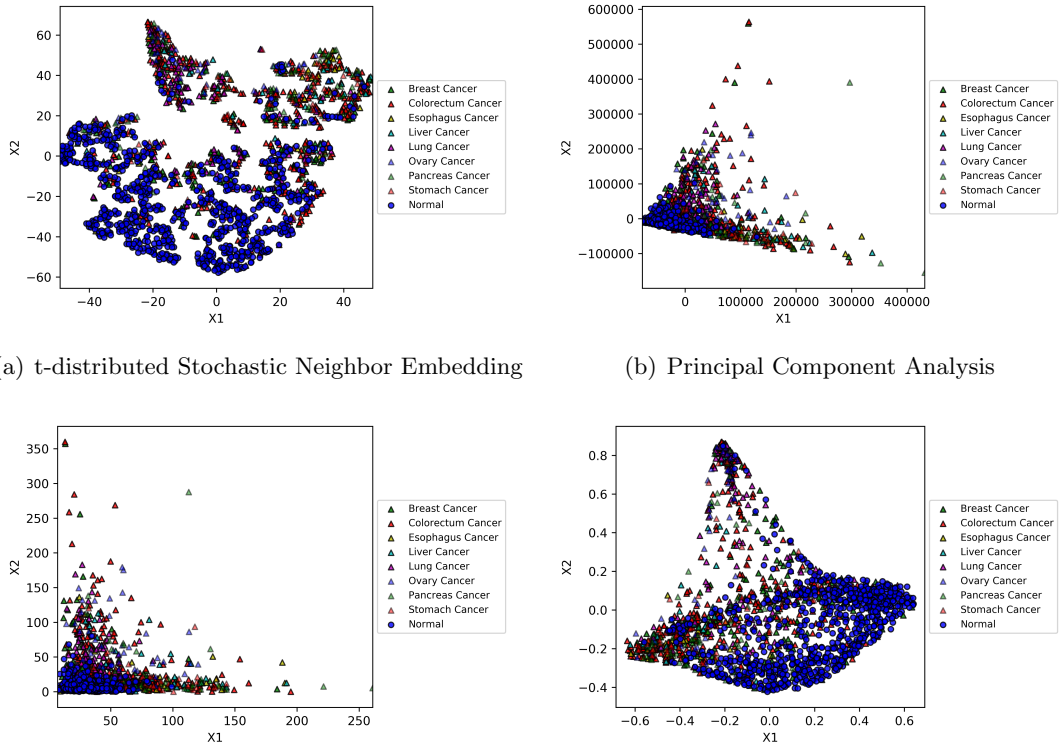
$$IG(\pi_T; S) - \frac{log_2(|S| - 1)}{N} > log_2(3^k - 2) - [kEnt(S) - k_1 Ent(S_1) - k_2 Ent(S_2)] \tag{11}$$

Based on the above MDL decision rules, we can partition each attribute (or feature) into intervals for Bayesian learning in the study.

S1.2 Supplemental Figures and Tables

InfoG	Input Features	Feature Description
0.6897	CA19-9 (U/ml)	Circulating Cancer Antigen 19-9 Concentration in U/ml
0.5119	CA-125 (U/ml)	Circulating Cancer Antigen 125 Concentration in U/ml
0.5001	HGF (pg/ml)	Circulating Hepatocryte Growth Factor Concentration in pg/ml
0.2779	OPN (pg/ml)	Circulating Osteopontin Concentration in pg/ml
0.2208	OmegaScore	Omega Score for Mutations in Circulating Cell-Free DNA
0.1826	Prolactin (pg/ml)	Circulating Prolactin Concentration in pg/ml
0.1518	CEA (pg/ml)	Circulating CarcinoEmbryonic Antigen Concentration in pg/ml
0.0989	Myeloperoxidase (ng/ml)	Circulating Myeloperoxidase Concentration in ng/ml
0.0916	TIMP-1 (pg/ml)	Circulating Tissue Inhibitor of MetalloProteinases 1 Concentration in pg/ml

Table S1: **Feature List for Cancer Detection ranked by Information Gain (InfoG), related to Figure 1**



(a) t-distributed Stochastic Neighbor Embedding

(b) Principal Component Analysis

(c) Nonnegative Matrix Factorization

(d) Spectral Embedding

Figure S1: **Dataset Visualization for Cancer Detection, related to Figure 1** All figures are drawn using the captioned method with Python scikit-learn package and its default setting (Pedregosa et al., 2011).

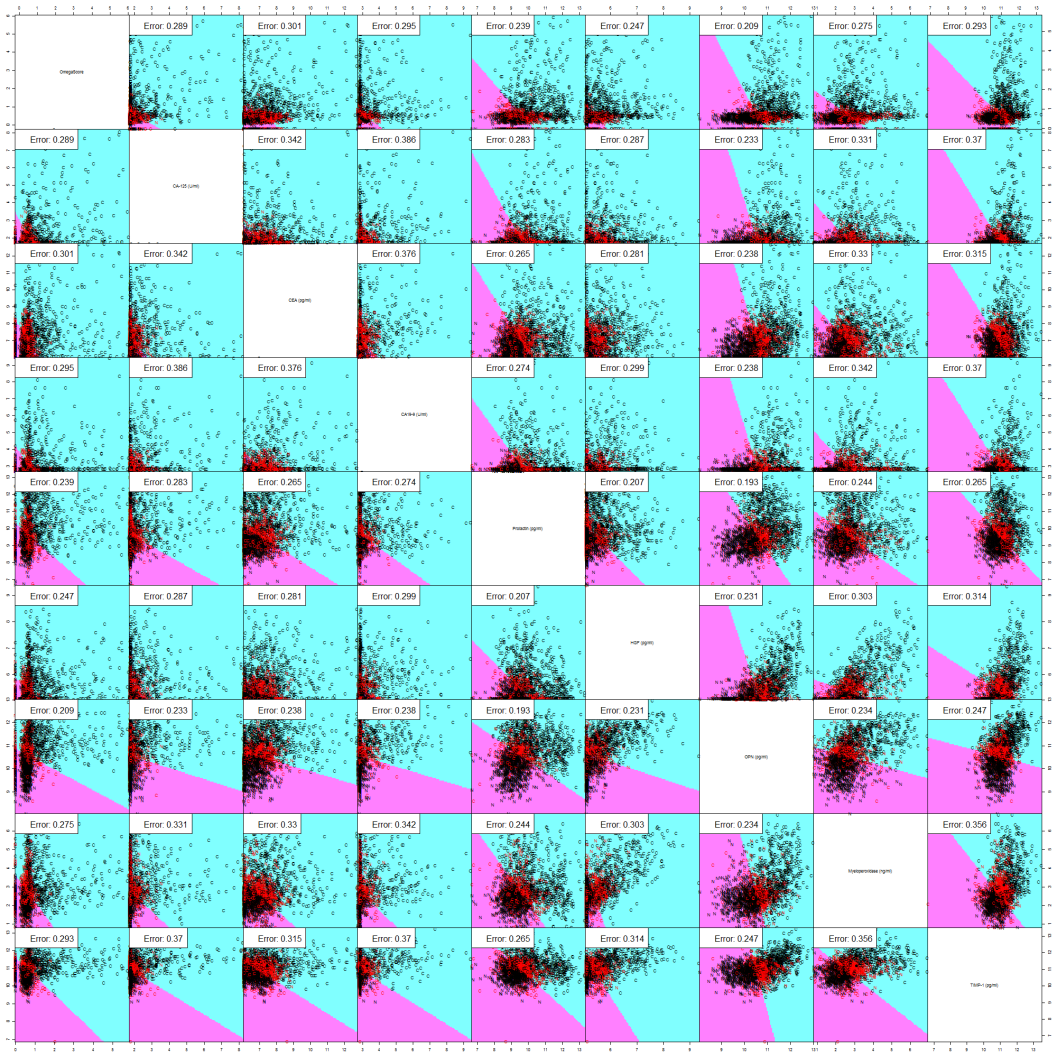
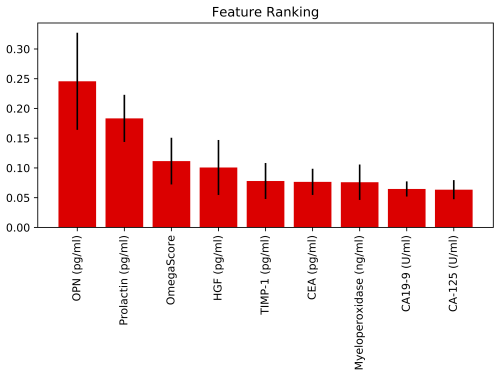
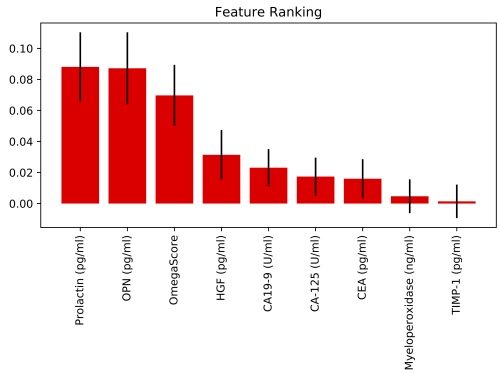


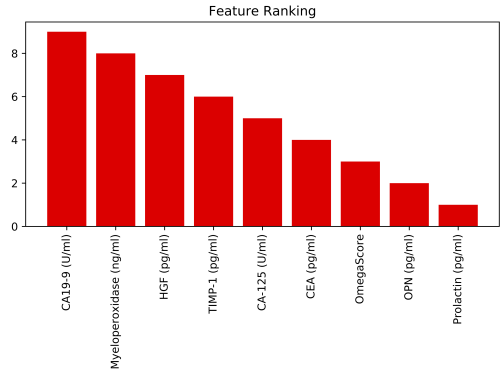
Figure S2: **Linear Discriminant Analysis (LDA) for Cancer Detection, related to Figure 1**
The plot is drawn using the default setting of the klaR package in R. The two colors correspond to the linearly discriminated regions. The red fonts denote the wrongly classified samples.



(a) Decrease in Purity



(b) Decrease in Accuracy



(c) Recursive Feature Elimination

Figure S3: **Feature Rankings for Cancer Detection, related to Figure 1** The feature rankings are measured based on the random forest building under Python scikit-learn package(Pedregosa et al., 2011). Each bar represents one feature. For purity and accuracy, 5-fold cross-validations are run on the random forests of 250 Gini decision trees for 300 times to give the means and standard deviations as visualized on the error bars.

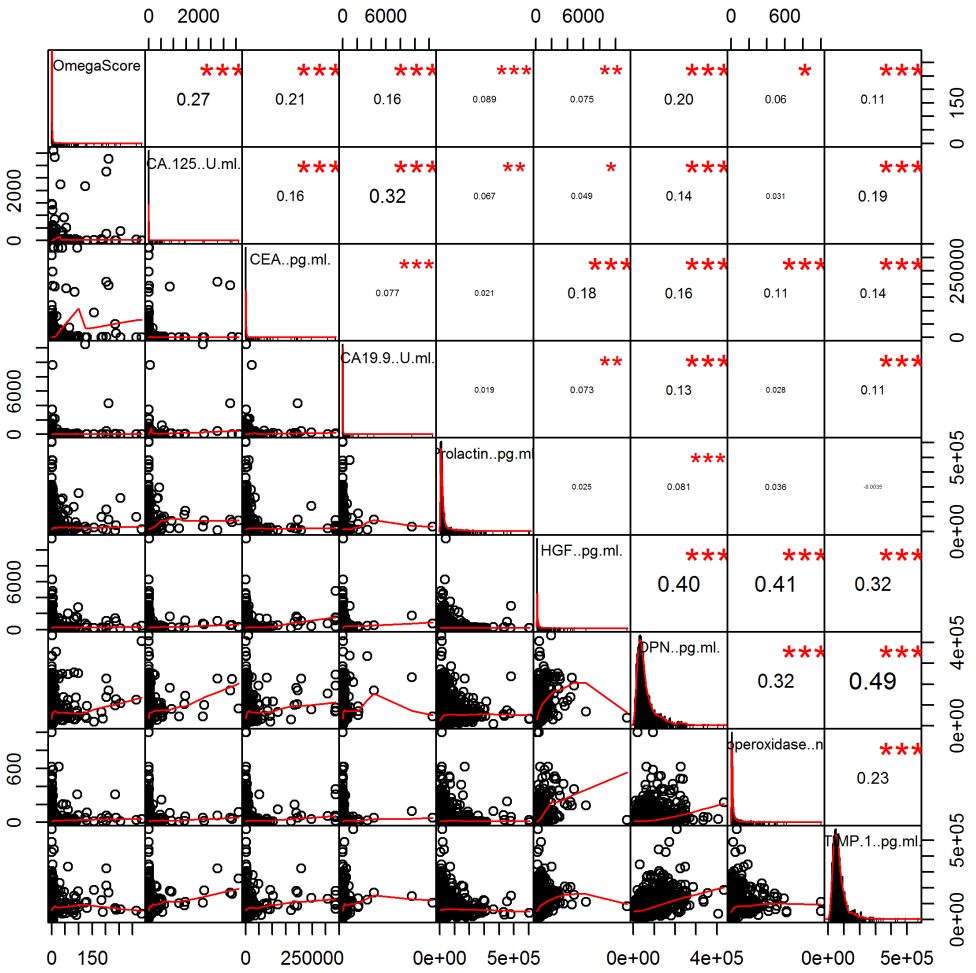


Figure S4: **Feature Correlation Matrix for Cancer Detection, related to Figure 1**

The matrix is drawn using the 'PerformanceAnalytics' package in R with its default setting. The upper triangle tabulates the Pearson correlation values under different pairing scenarios. P-values are also computed where each significance level is associated to a symbol where *** denotes 0.001; ** denotes 0.01; * denotes 0.05.

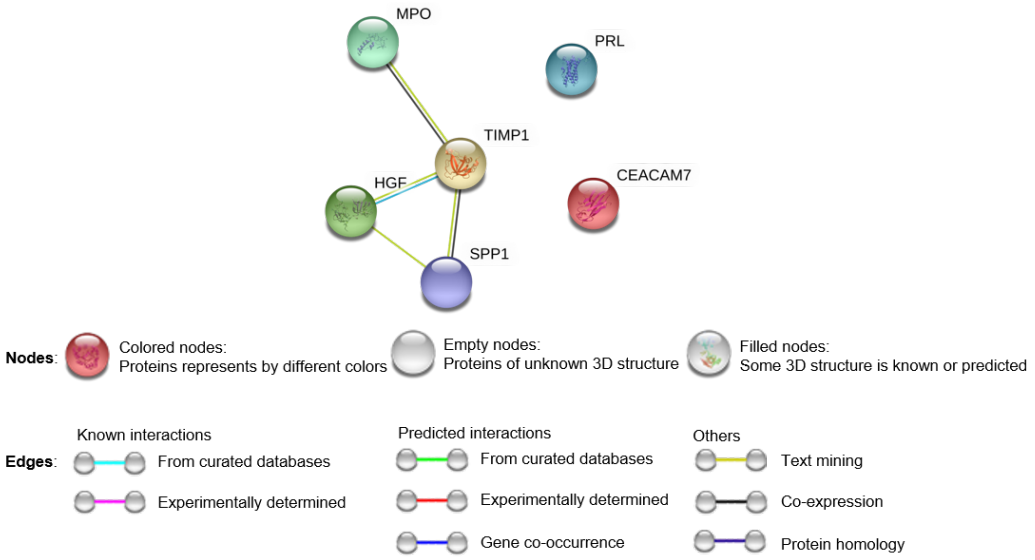


Figure S5: **Protein-Protein Interaction Network for Cancer Detection, related to Figure 1**
Only the protein markers which can be precisely mapped onto STRING network analysis without any ambiguity are shown here (Szklarczyk et al., 2016).

Table S2: **Comparative Computational Complexity Analysis on Naive Bayes Variants, related to Figure 1**
A1DE is the model underlying CacnerA1DE; NB denotes Naive Bayes; LBR denotes Lazy Bayesian Rule; TAN denotes Tree Augmented Naive Bayes; SP-TAN denotes SuperParent TAN.

Complexity	A1DE	NB	LBR	TAN	SP-TAN
Training time complexity	$O(tn^2)$	$O(tn)$	$O(tn)$	$O(n^2t + kn^2t^2 + n^2logn)$	$O(tkn^3)$
Prediction time complexity	$O(kn^2)$	$O(kn)$	$O(tkn^3)$	$O(kn)$	$O(knv^2)$
Training space complexity	$O(k(nv)^2)$	$O(knv)$	$O(tn)$	$O(k(nv)^2)$	$O(k(nv)^2 + tn)$
Prediction space complexity	$O(k(nv)^2)$	$O(knv)$	$O(tn)$	$O(knv^2)$	$O(knv^2)$

Note:
k is the number of classes
n is the number of attributes
v is the mean number of attributes values
t is the number of training samples

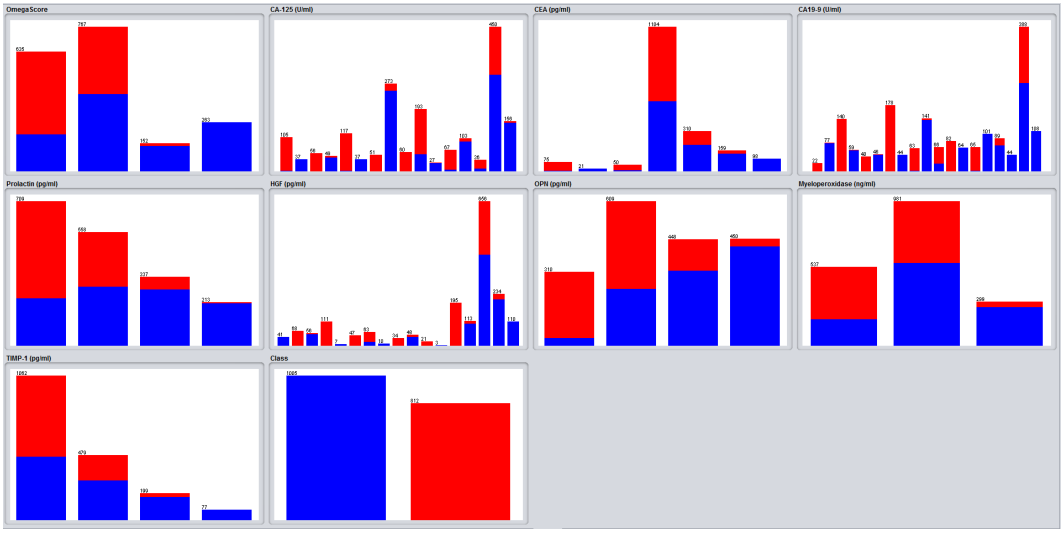


Figure S6: **Cancer Detection Feature Histograms (vertical axis) after Feature Discretization (horizontal axis), related to Figure 1**
The feature discretization is based on the Minimum Description Length (MDL) principle using JAVA Weka (Kononenko, 1995). The red colour denotes cancer while the blue colour denotes healthy patient sample.

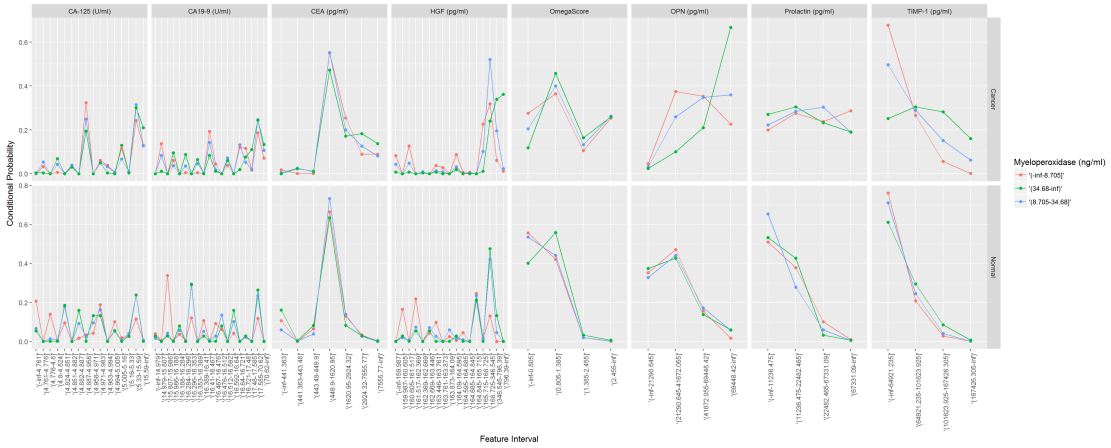
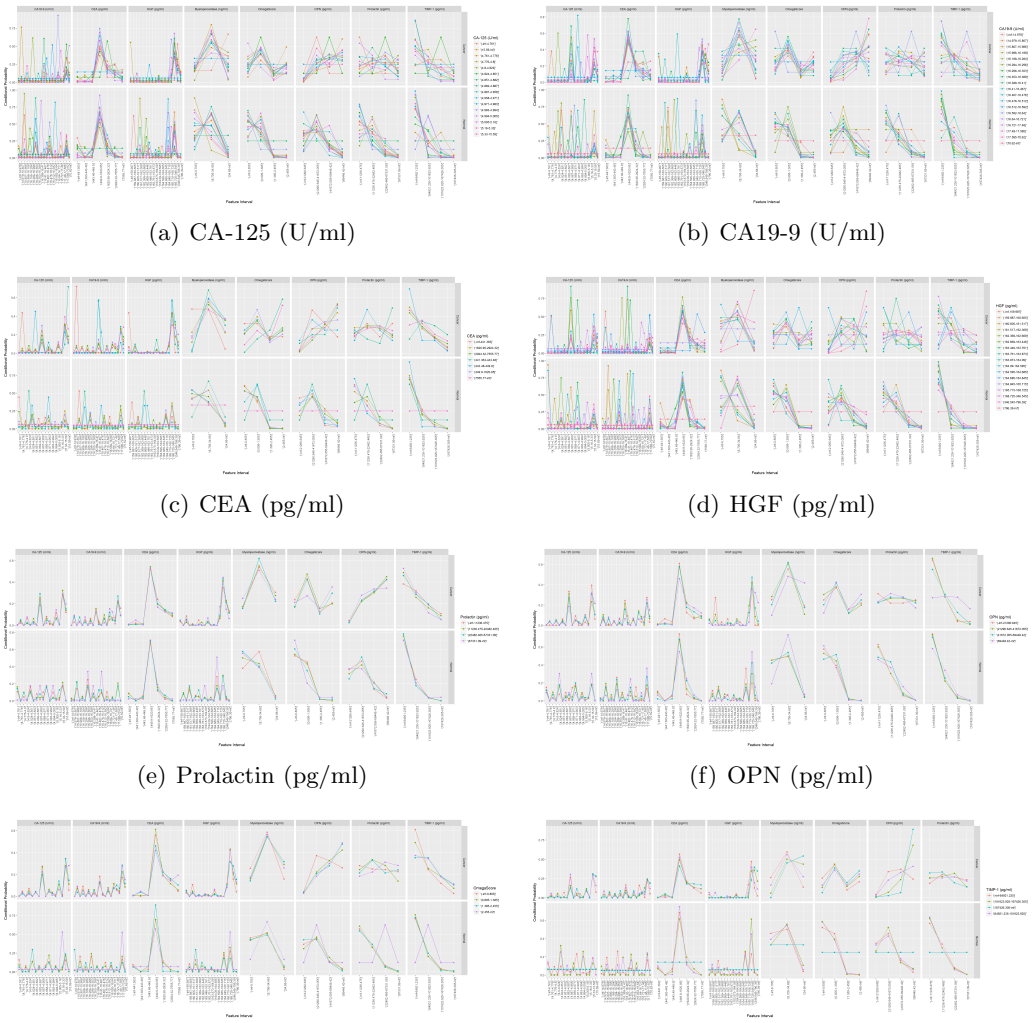


Figure S7: **Conditional Probability Plots ($\hat{P}(x_j|y, x_i)$) for CancerA1DE when the Parent Feature x_i refers to Myeloperoxidase (ng/ml), related to Figure 2**
The colours denote different parent feature values as demonstrated in the right legends. The horizontal axis denotes the child feature values (x_j) which conditional probabilities are scaled across the vertical axis grouped by cancer or normal (y).



(a) CA-125 (U/ml)

(b) CA19-9 (U/ml)

(c) CEA (pg/ml)

(d) HGF (pg/ml)

(e) Prolactin (pg/ml)

(f) OPN (pg/ml)

(g) OmegaScore

(h) TIMP-1 (pg/ml)

Figure S8: **Conditional Probability Plots ($\hat{P}(x_j|y, x_i)$) for CancerA1DE, related to Figure 2**
The colors denote the parent feature values (x_i) as demonstrated in the right legends. The horizontal axis denotes the child feature values (x_j) which conditional probabilities are scaled across the vertical axis grouped by cancer or normal (y).

Table S3: **CancerA1DE and CancerA2DE Performance Sensitivity Analysis, related to Figure 1**

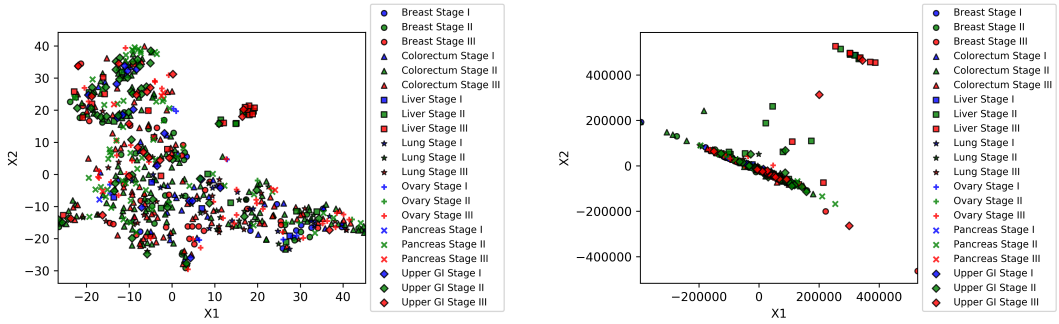
The analysis is based on independent training and testing data in different size proportions.

Training %	Testing %	CancerA1DE ROC AUC	CancerA2DE ROC AUC
1	99	0.694	0.694
5	95	0.892	0.892
10	90	0.926	0.924
20	80	0.983	0.983
30	70	0.983	0.984
40	60	0.987	0.987
50	50	0.989	0.989
60	40	0.989	0.989
70	30	0.991	0.991
80	20	0.991	0.991
90	10	0.994	0.994

Pathway ID	Pathway Description	False Discovery Rate	Matching Proteins
5144	Malaria	2.14E-06	CSF3,HGF,IL6,IL8,THBS2
4151	PI3K-Akt signaling pathway	3.97E-06	ANGPT2,CSF3,FGF2,HGF,IL6,PRL,SPP1,THBS2
4060	Cytokine-cytokine receptor interaction	9.58E-06	CSF3,HGF,IL6,IL8,LEP,PRL,TNFRSF11B
4066	HIF-1 signaling pathway	0.00123	ANGPT2,ENO2,IL6,TIMP1
4630	Jak-STAT signaling pathway	0.00476	CSF3,IL6,LEP,PRL
5200	Pathways in cancer	0.00528	FGF2,HGF,IL6,IL8,TGFA
4512	ECM-receptor interaction	0.0103	CD44,SPP1,THBS2
4640	Hematopoietic cell lineage	0.0103	CD44,CSF3,IL6
4620	Toll-like receptor signaling pathway	0.0151	IL6,IL8,SPP1
4932	Non-alcoholic fatty liver disease (NAFLD)	0.0405	IL6,IL8,LEP

Table S4: **Enriched Pathway List for Cancer Type Localization, related to Figure 4**

It is computed from the network in Figure S12 using STRING network analysis (Szklarczyk et al., 2016).



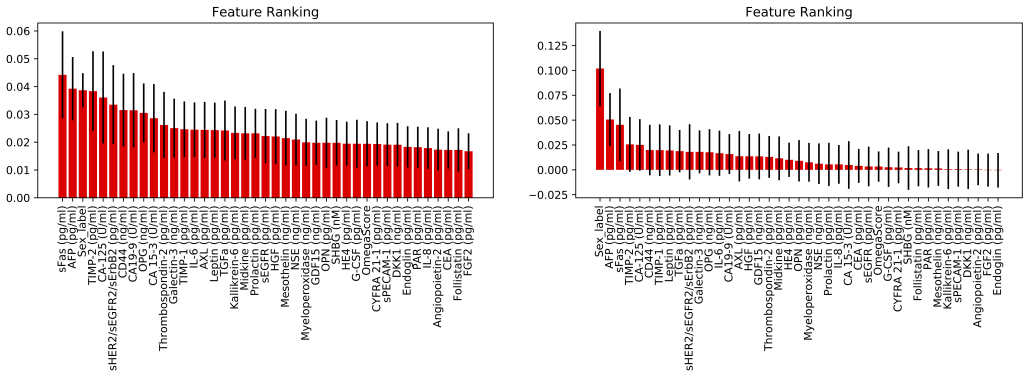
(a) t-distributed Stochastic Neighbor Embedding

(b) Principal Component Analysis

(c) Nonnegative Matrix Factorization

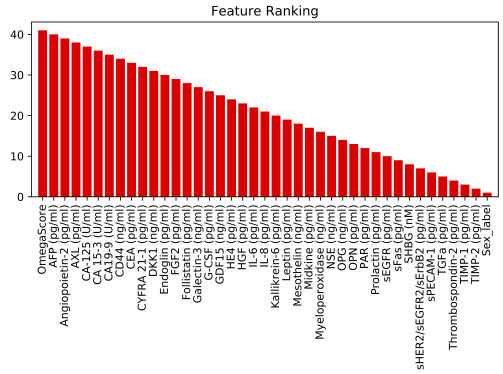
(d) Spectral Embedding

Figure S9: **Dataset Visualization for Cancer Type Localization, related to Table 1**
All figures are drawn using the captioned method with Python scikit-learn package and its default setting (Pedregosa et al., 2011).



(a) Decrease in Purity

(b) Decrease in Accuracy



(c) recursive feature elimination

Figure S10: **Feature Rankings for Cancer Type Localization, Table 1**
The feature rankings are measured based on the random forest building under Python scikit-learn package(Pedregosa et al., 2011). Each bar represents one feature. For purity and accuracy, 5-fold cross-validations are run on the random forests of 250 Gini decision trees for 300 times to give the means and standard deviations as visualized on the error bars.

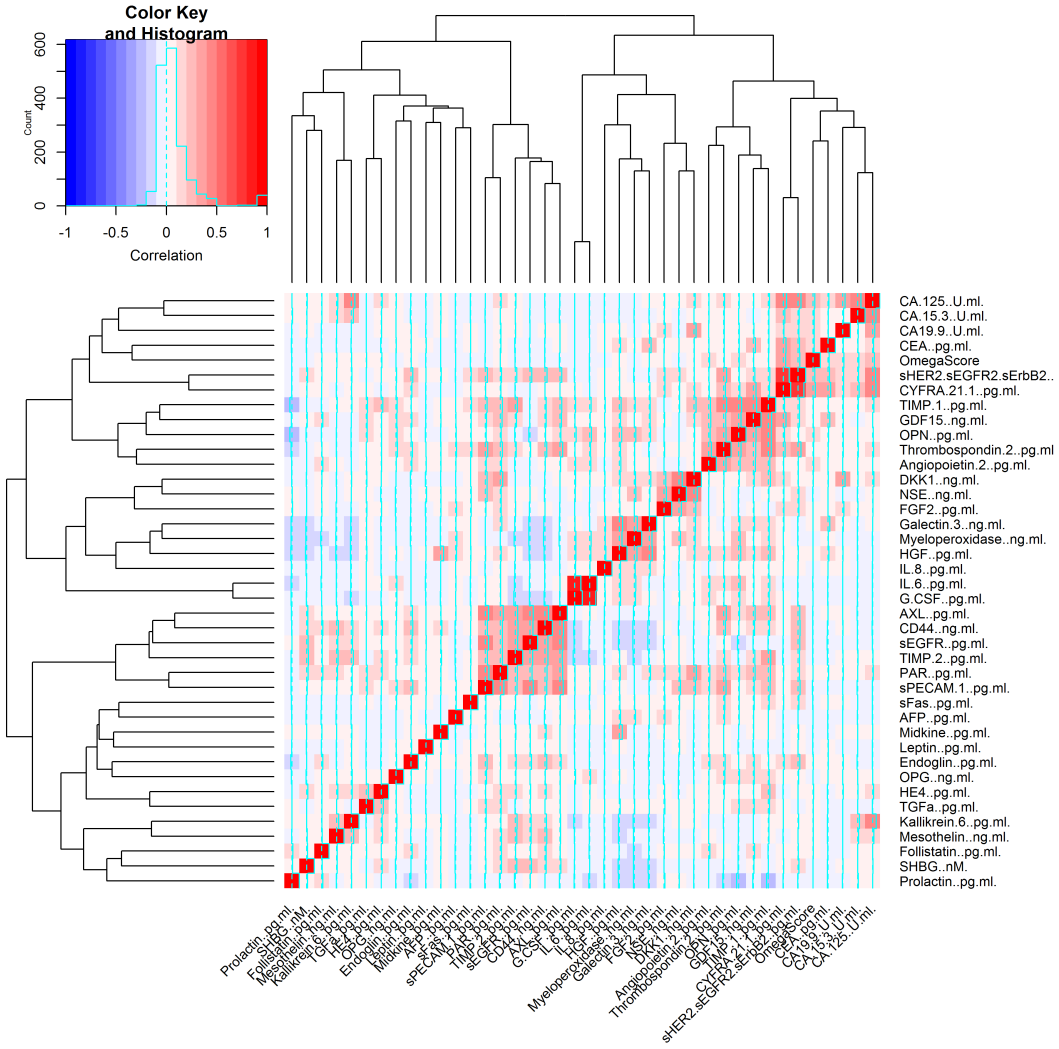


Figure S11: **Feature Correlation Matrix for Cancer Type Localization, related to Table 1**

The heatmap is drawn using the 'gplots' package in R with its default setting under the function 'heatmap.2'. In particular, the correlations values are computed using the Pearson correlation approach. The features are order by the corresponding hierarchical clustering on each axis.

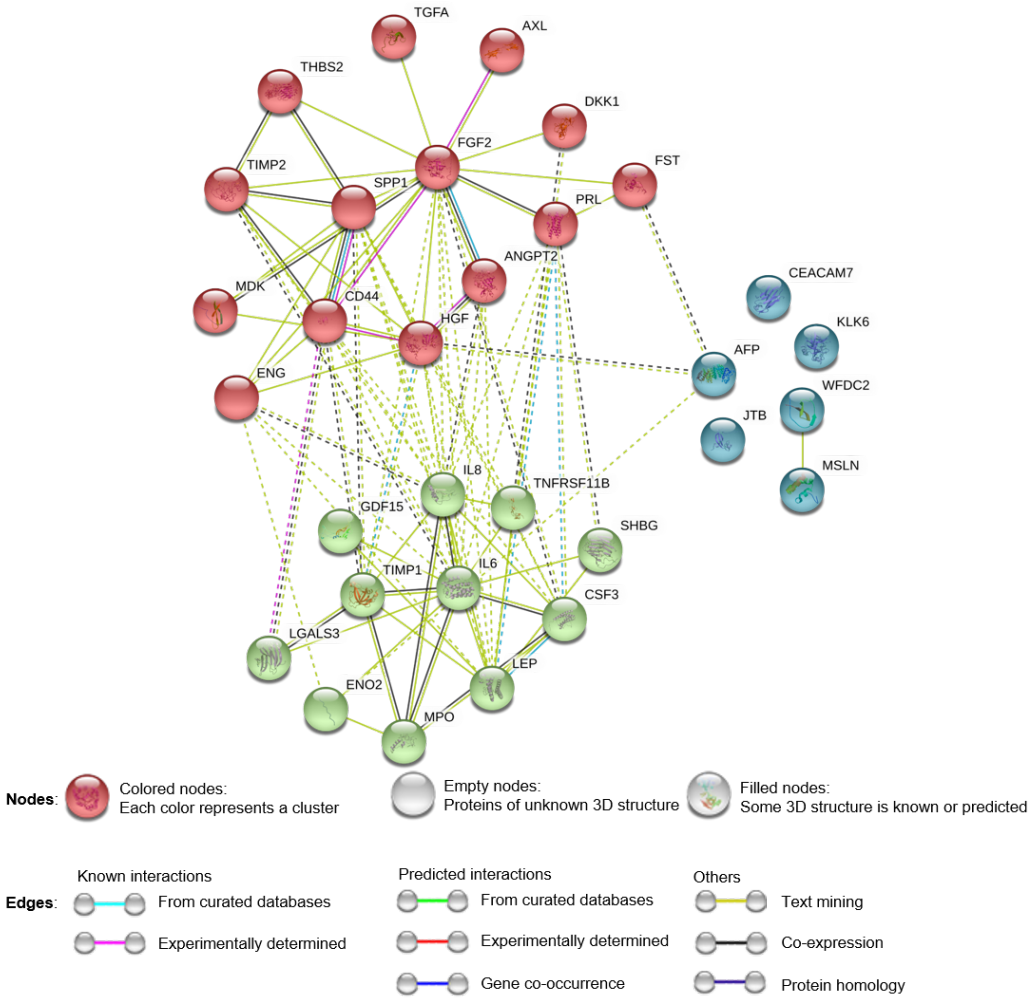
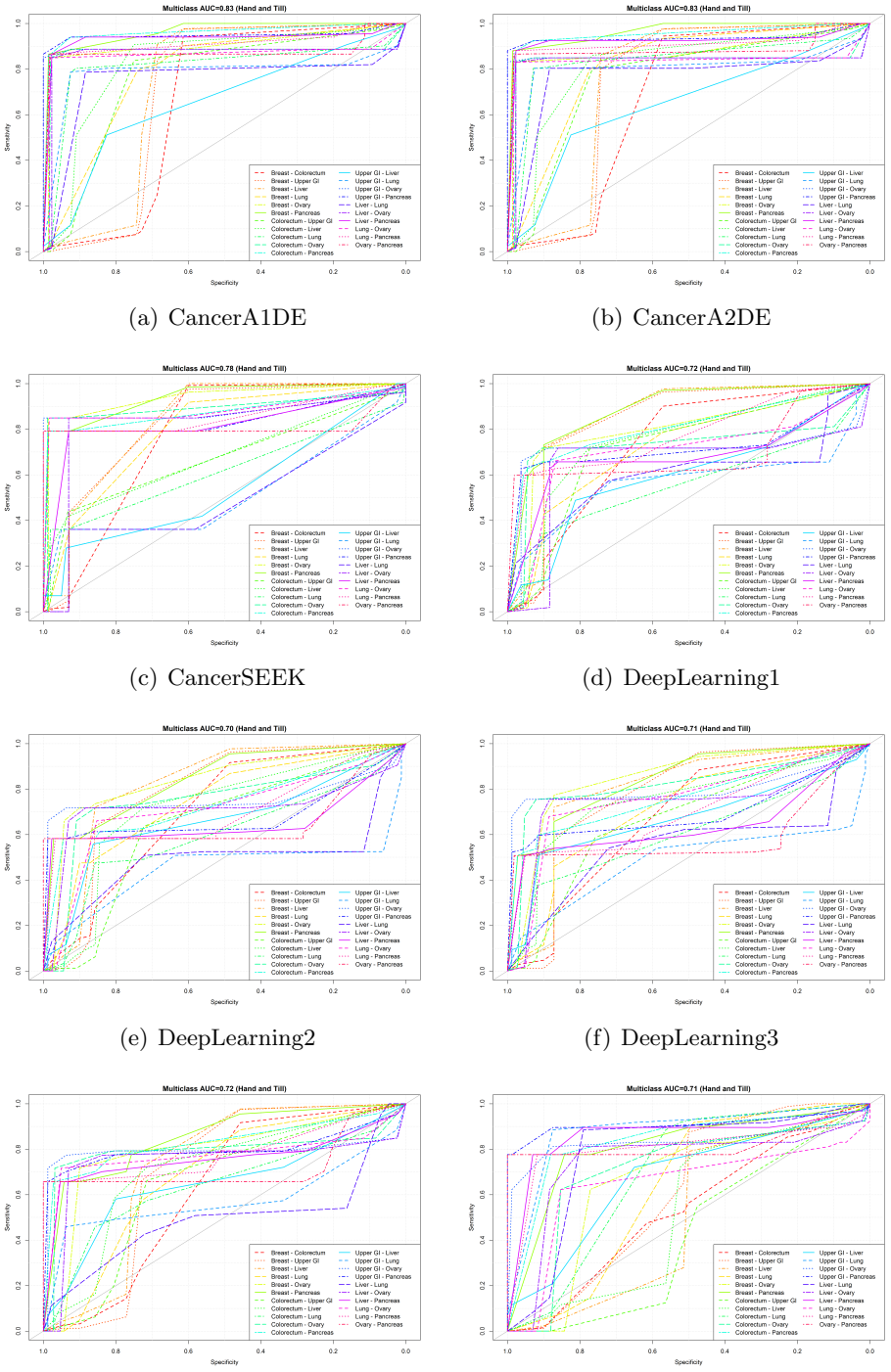


Figure S12: **Protein-Protein Interaction Network for Cancer Type Localization, related to Figure 4**
Only the protein markers which can be precisely mapped onto STRING network analysis without any ambiguity are shown here (Szklarczyk et al., 2016). K-means community detection has been applied onto the network with 3 communities.



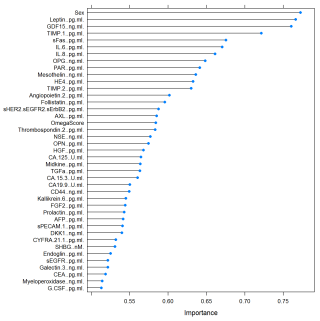
(a) CancerA1DE (b) CancerA2DE

(c) CancerSEEK (d) DeepLearning1

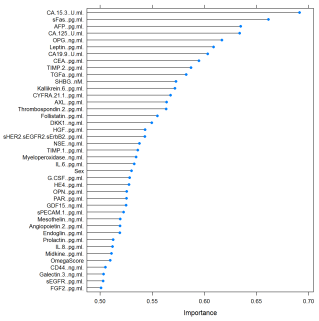
(e) DeepLearning2 (f) DeepLearning3

(g) J48 (h) Naive Bayes

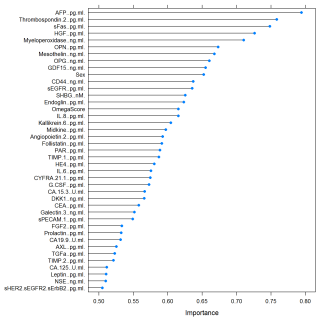
Figure S13: **Receiver Operating Characteristic (ROC) curves for Cancer Type Localizations, related to Figure 4**
The curves are generated under 10-fold cross-validations using the R package 'pROC' with the multiclass setting according to Hand and Till (Hand and Till, 2001). Each plot represents a method.



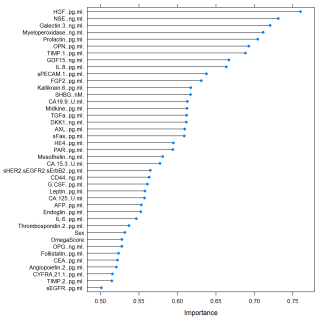
(a) Breast



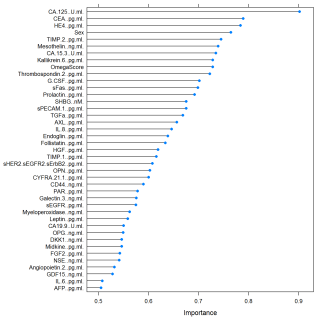
(b) Colorectum



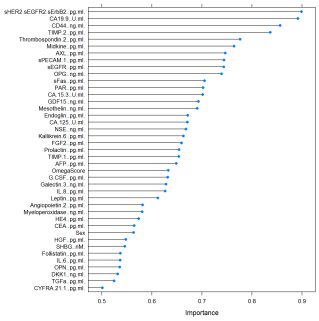
(c) Liver



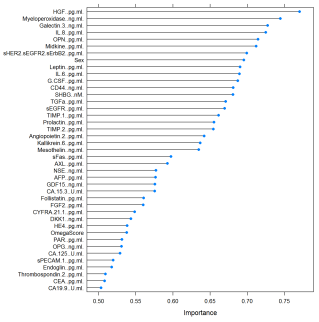
(d) Lung



(e) Ovary



(f) Pancreas



(g) Upper GI

Figure S14: **Feature Importance Analysis for Cancer Type Localization under One-Class-versus-Others Setting, related to Table 1**
The feature rankings are measured based on the Learning Vector Quantization (LVQ) building under Python caret package (Bischl et al., 2016). 10-fold cross-validations are run compute the feature importance values.