CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

Data Science for Cancer Detection

Early Cancer Detection from Multianalyte Blood Test Results

Ka-Chun Wong and his team

The research is supported by:

Research Grants Council 研究資助局

CityU Department of Computer Science
香港城市大學
City University of Hong Kong

CityU Hong Kong Institute for Data Science
香港城市大學
City University of Hong Kong

Health and Medical Research Fund (HMRF)

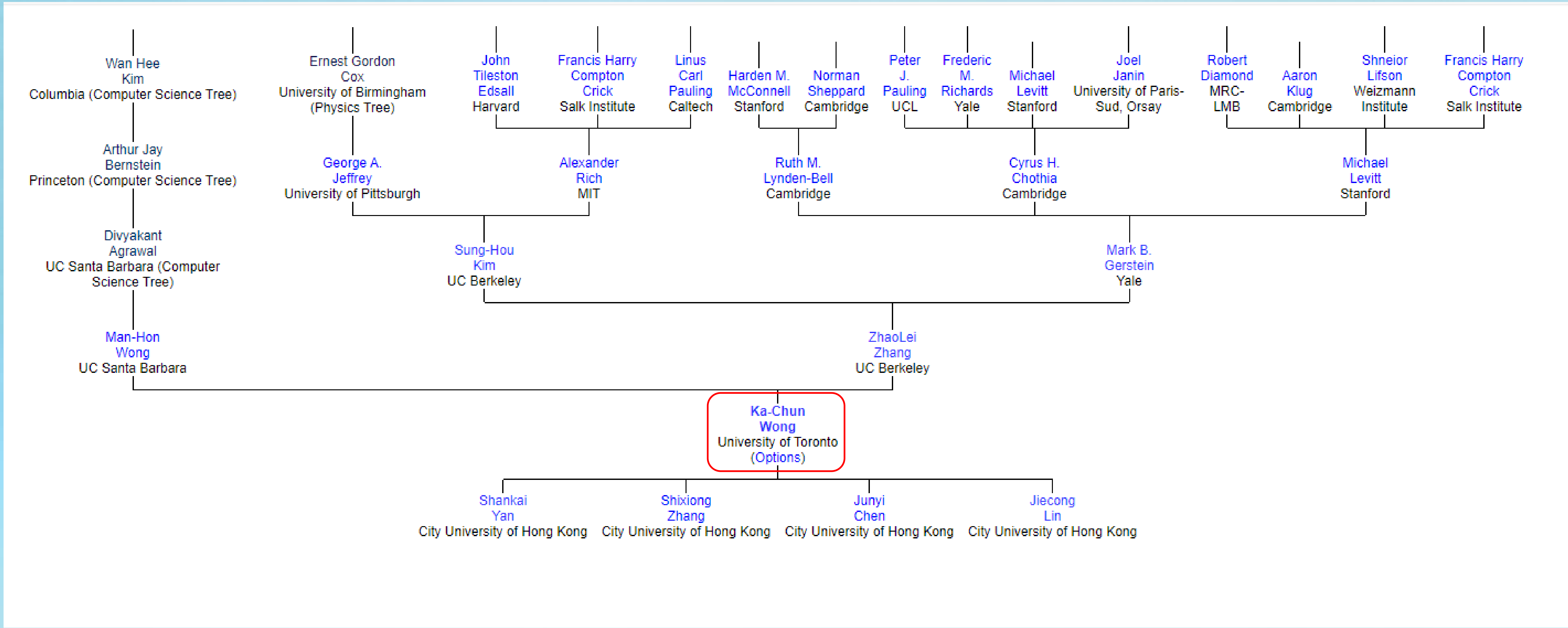
CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

The Speaker is part of The Distinguished Speakers Program which is made possible by

acm
Association for Computing Machinery
Advancing Computing as a Science & Profession

For additional information, please visit http://dsp.acm.org/

Speaker's Academic Family Tree



Wan Hee Kim
Columbia (Computer Science Tree)

Ernest Gordon Cox
University of Birmingham (Physics Tree)

John Tileston Edsall
Harvard

Francis Harry Compton Crick
Salk Institute

Linus Carl Pauling
Caltech

Harden M. McConnell
Stanford

Norman Sheppard
Cambridge

Peter J. Pauling
UCL

Frederic M. Richards
Yale

Michael Levitt
Stanford

Joel Janin
University of Paris-Sud, Orsay

Robert Diamond
MRC-LMB

Aaron Klug
Cambridge

Shneior Lifson
Weizmann Institute

Francis Harry Compton Crick
Salk Institute

Arthur Jay Bernstein
Princeton (Computer Science Tree)

George A. Jeffrey
University of Pittsburgh

Alexander Rich
MIT

Ruth M. Lynden-Bell
Cambridge

Cyrus H. Chothia
Cambridge

Michael Levitt
Stanford

Divyakant Agrawal
UC Santa Barbara (Computer Science Tree)

Sung-Hou Kim
UC Berkeley

Mark B. Gerstein
Yale

Man-Hon Wong
UC Santa Barbara

ZhaoLei Zhang
UC Berkeley

Ka-Chun Wong
University of Toronto
(Options)

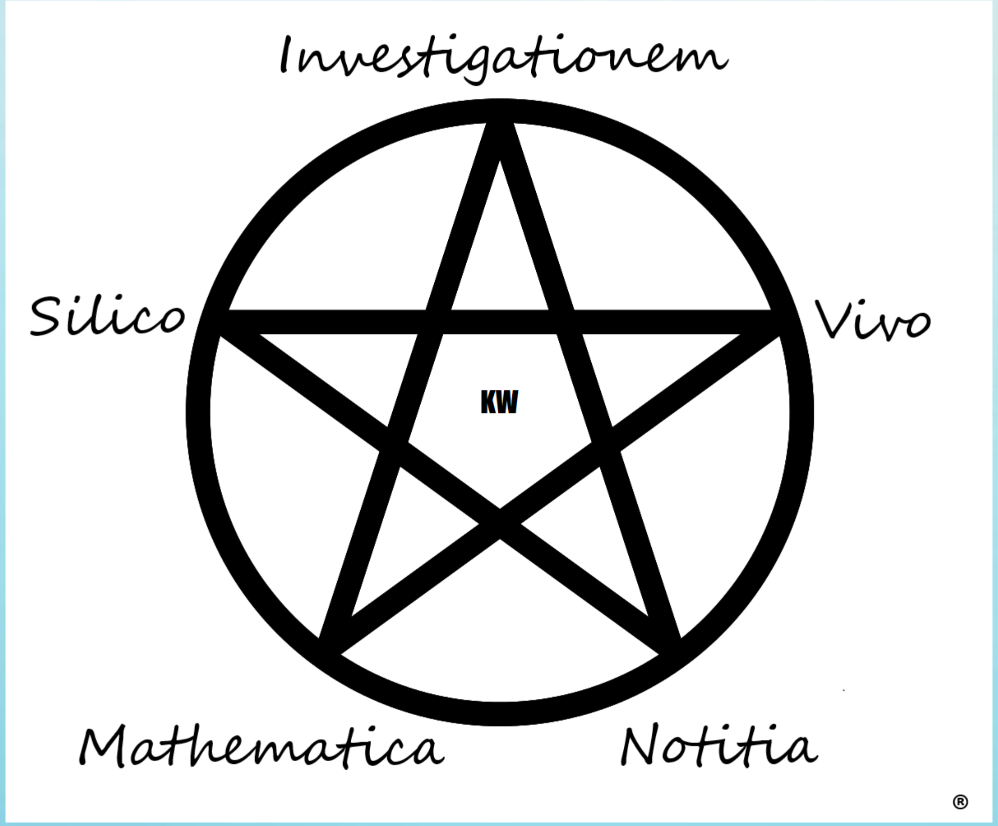
Shankai Yan
City University of Hong Kong

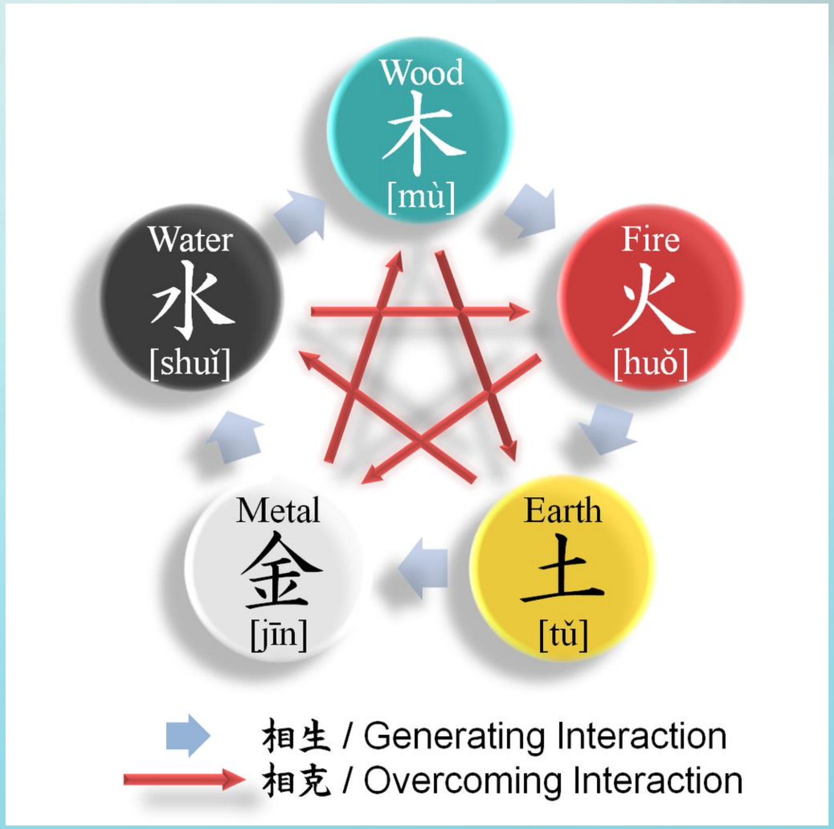
Shixiong Zhang
City University of Hong Kong

Junyi Chen
City University of Hong Kong

Jiecong Lin
City University of Hong Kong

Speaker’s Group Research in **Five Elements**





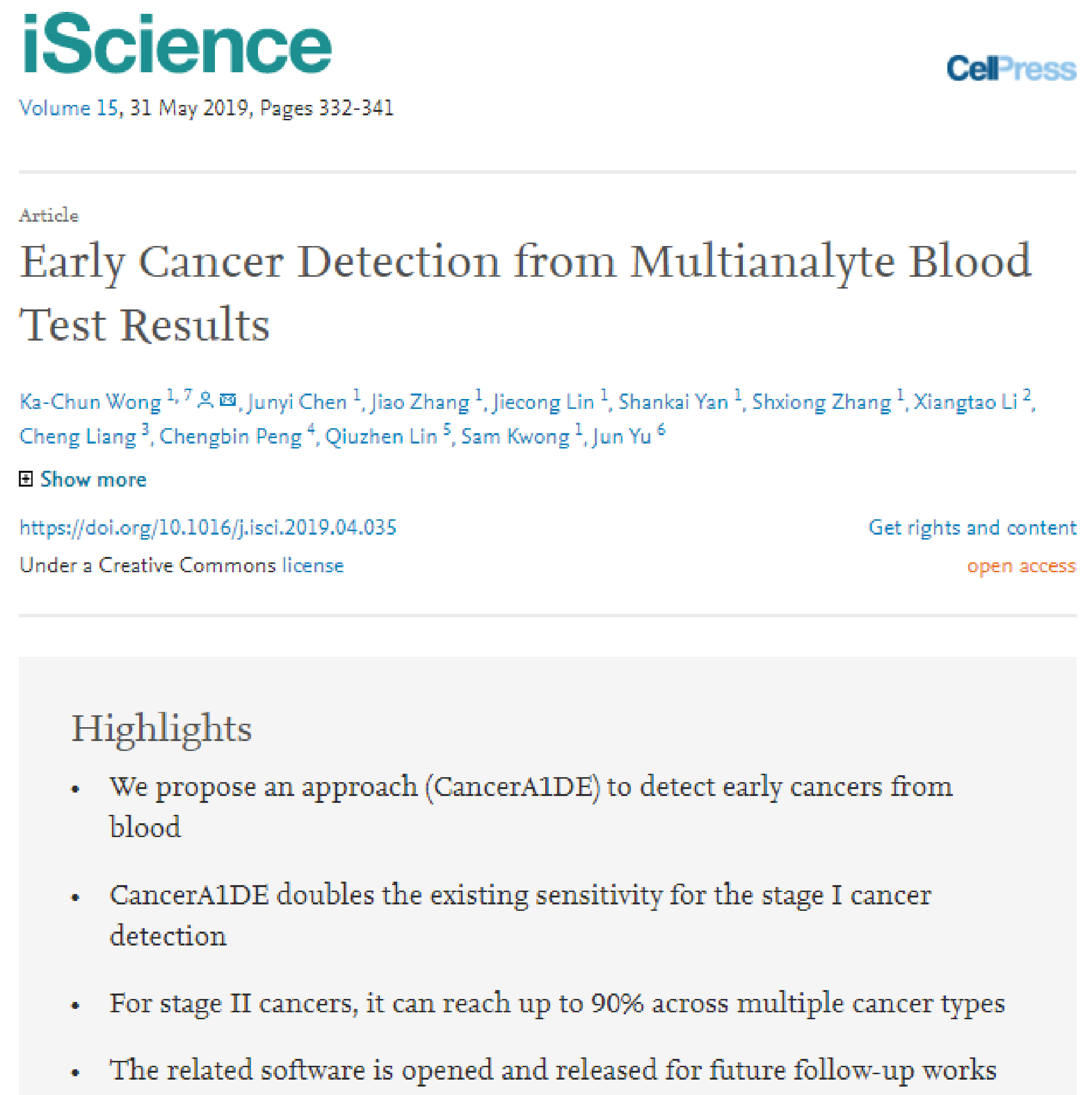
https://en.wikipedia.org/wiki/Wuxing_(Chinese_philosophy)#/media/File:Wu_Xing.png

Outline

- Motivation
- Background
- Data Science Analysis
- Proposed Approach
- Proposed Results
- Demo and Conclusion

Reference

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.



iScience

CellPress

Volume 15, 31 May 2019, Pages 332-341

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], Jun Yu[6]

⊞ Show more

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

Get rights and content

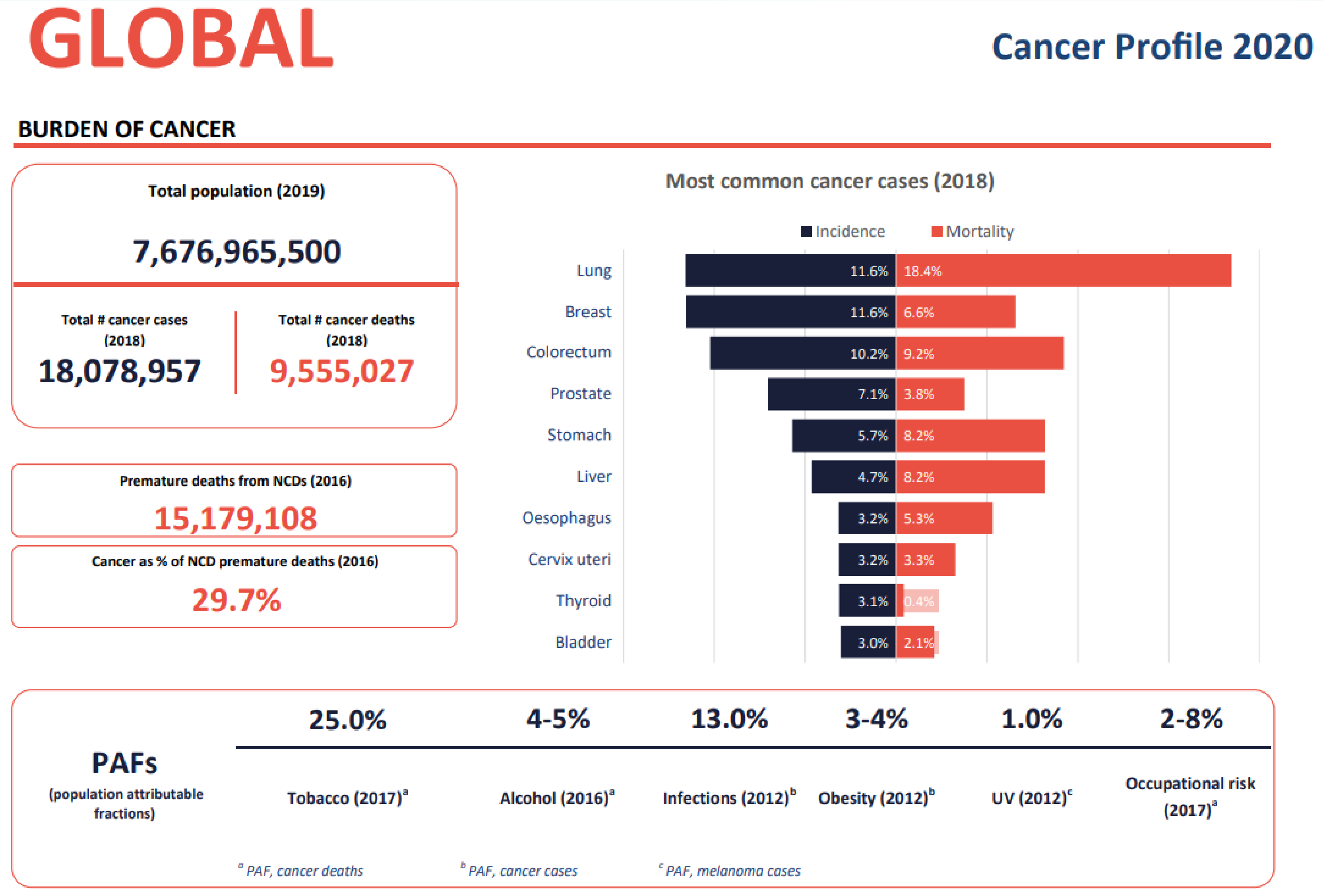
Highlights

- 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, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Motivation

World Health Organization



GLOBAL

Cancer Profile 2020

BURDEN OF CANCER

Total population (2019)

7,676,965,500

Total # cancer cases (2018)

18,078,957

Total # cancer deaths (2018)

9,555,027

Premature deaths from NCDs (2016)

15,179,108

Cancer as % of NCD premature deaths (2016)

29.7%

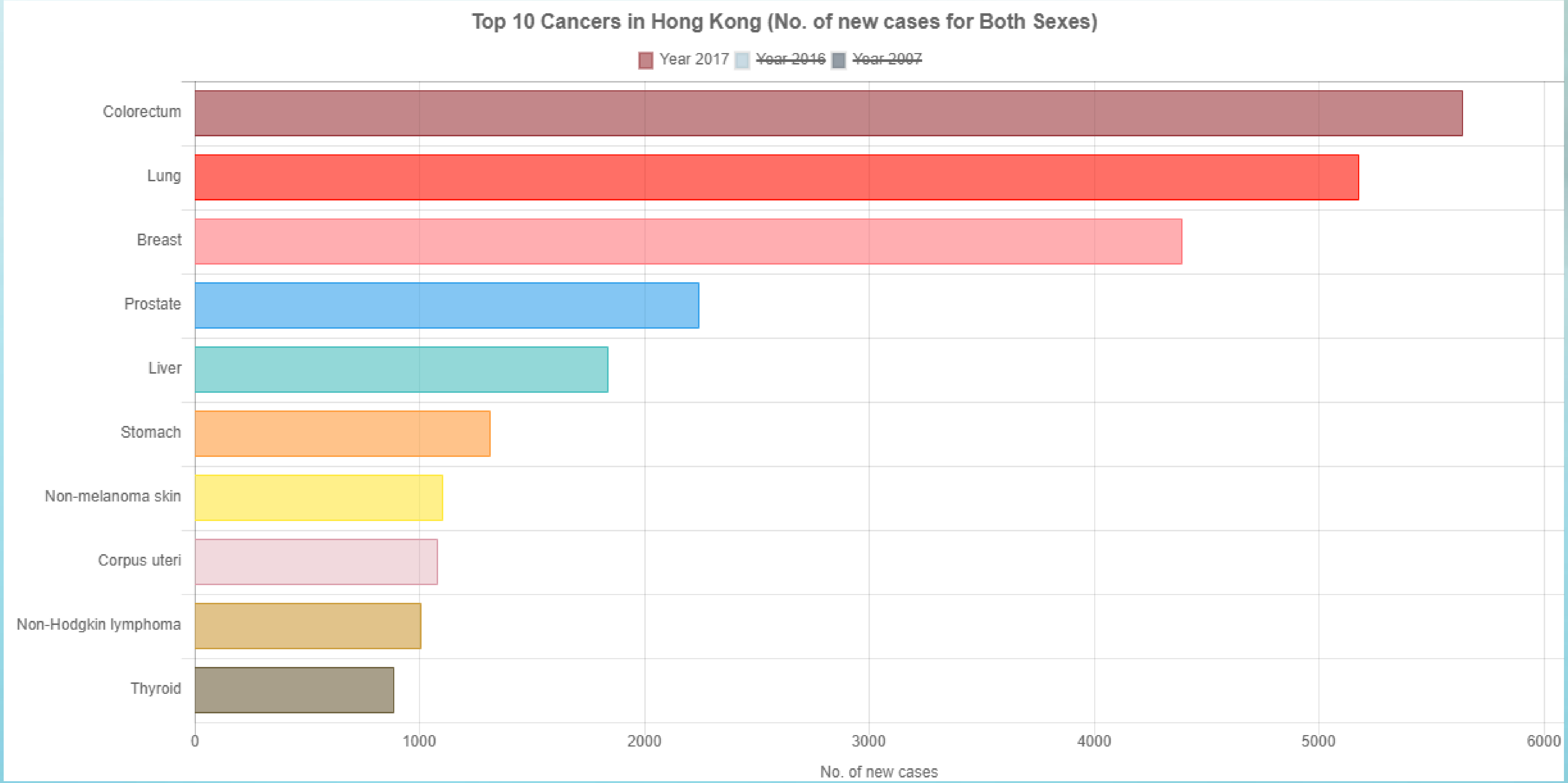
Most common cancer cases (2018)

■ Incidence
■ Mortality
Lung 11.6% 18.4%
Breast 11.6% 6.6%
Colorectum 10.2% 9.2%
Prostate 7.1% 3.8%
Stomach 5.7% 8.2%
Liver 4.7% 8.2%
Oesophagus 3.2% 5.3%
Cervix uteri 3.2% 3.3%
Thyroid 3.1% 0.4%
Bladder 3.0% 2.1%

PAFs (population attributable fractions)	Tobacco (2017)[a]	Alcohol (2016)[a]	Infections (2012)[b]	Obesity (2012)[b]	UV (2012)[c]	Occupational risk (2017)[a]
	25.0%	4-5%	13.0%	3-4%	1.0%	2-8%

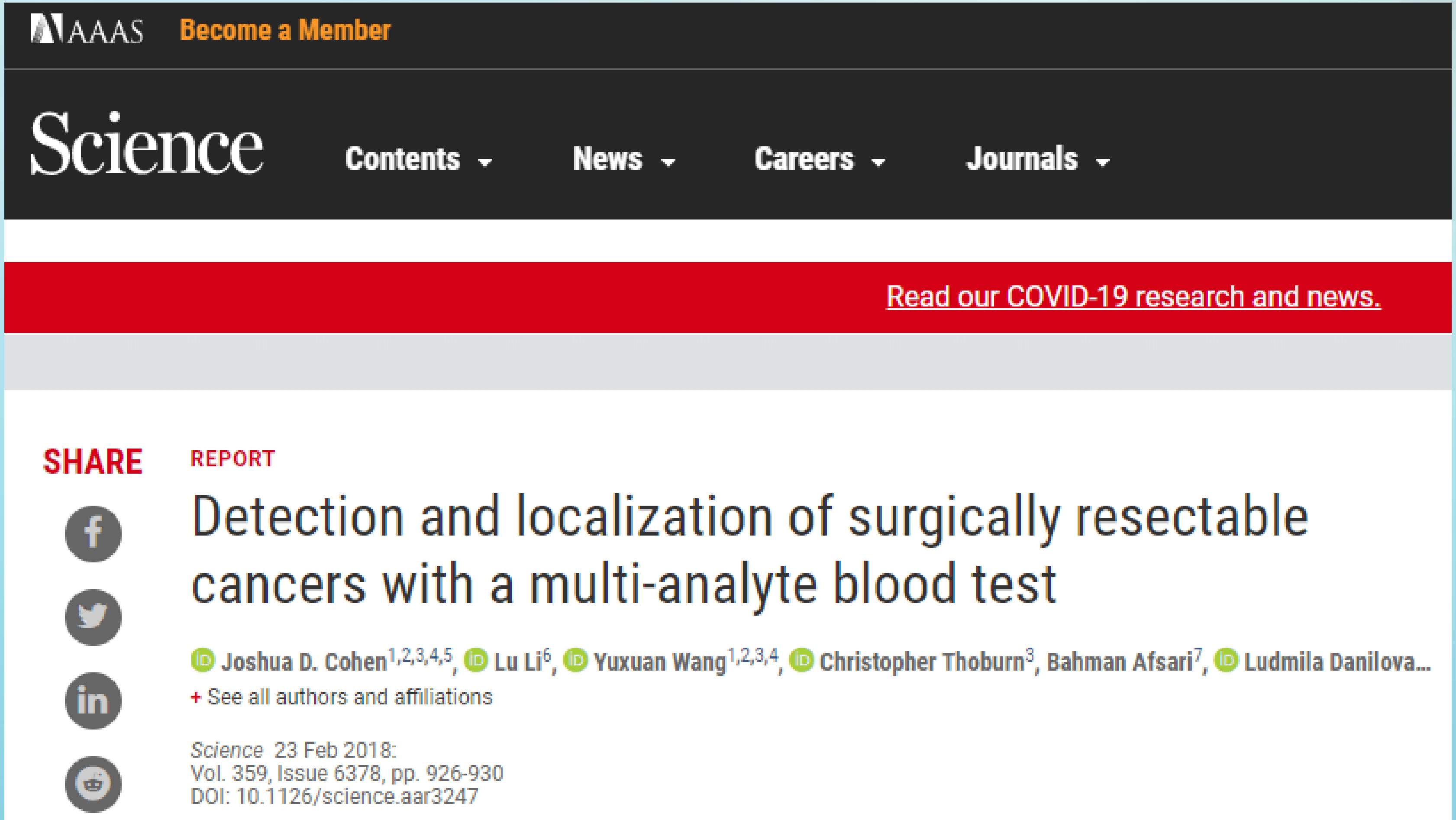
[a] *PAF, cancer deaths* [b] *PAF, cancer cases* [c] *PAF, melanoma cases*

Motivation



https://www3.ha.org.hk/cancereg/top10incidence.html

Background



REPORT

Detection and localization of surgically resectable cancers with a multi-analyte blood test

Joshua D. Cohen[1,2,3,4,5], Lu Li[6], Yuxuan Wang[1,2,3,4], Christopher Thoburn[3], Bahman Afsari[7], Ludmila Danilova...

+ See all authors and affiliations

Science 23 Feb 2018:
Vol. 359, Issue 6378, pp. 926-930
DOI: 10.1126/science.aar3247

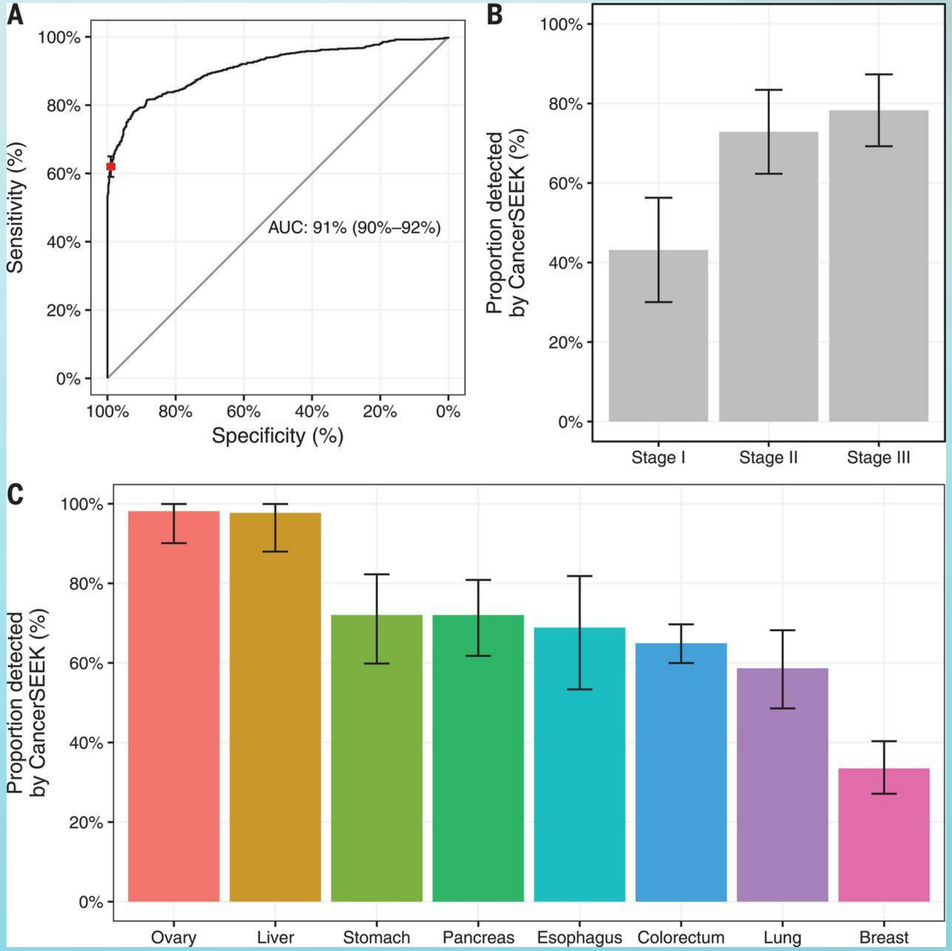
Background

Performance of CancerSEEK [1]

(A) ROC curve for CancerSEEK. The red point on the curve indicates the test's average performance (62%) at >99% specificity. Error bars represent 95% confidence intervals for sensitivity and specificity at this particular point. The median performance among the eight cancer types assessed was 70%.

(B) Sensitivity of CancerSEEK by stage. Bars represent the median sensitivity of the eight cancer types, and error bars represent standard errors of the median.

(C) Sensitivity of CancerSEEK by tumor type. Error bars represent 95% confidence intervals.



[1] Cohen, Joshua D., et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test." *Science* 359.6378 (2018): 926-930.

Background

However, we note three limitations of CancerSEEK [1]:

1. Its front-line cancer detection component is based on logistic regression, whereby linear assumption on different markers is hardly realistic.
2. Its second-line cancer type localization component is based on random forest, a modeling known to be difficult for interpretations.
3. From the user perspective, its lack of public Web service also limits its potential impacts.

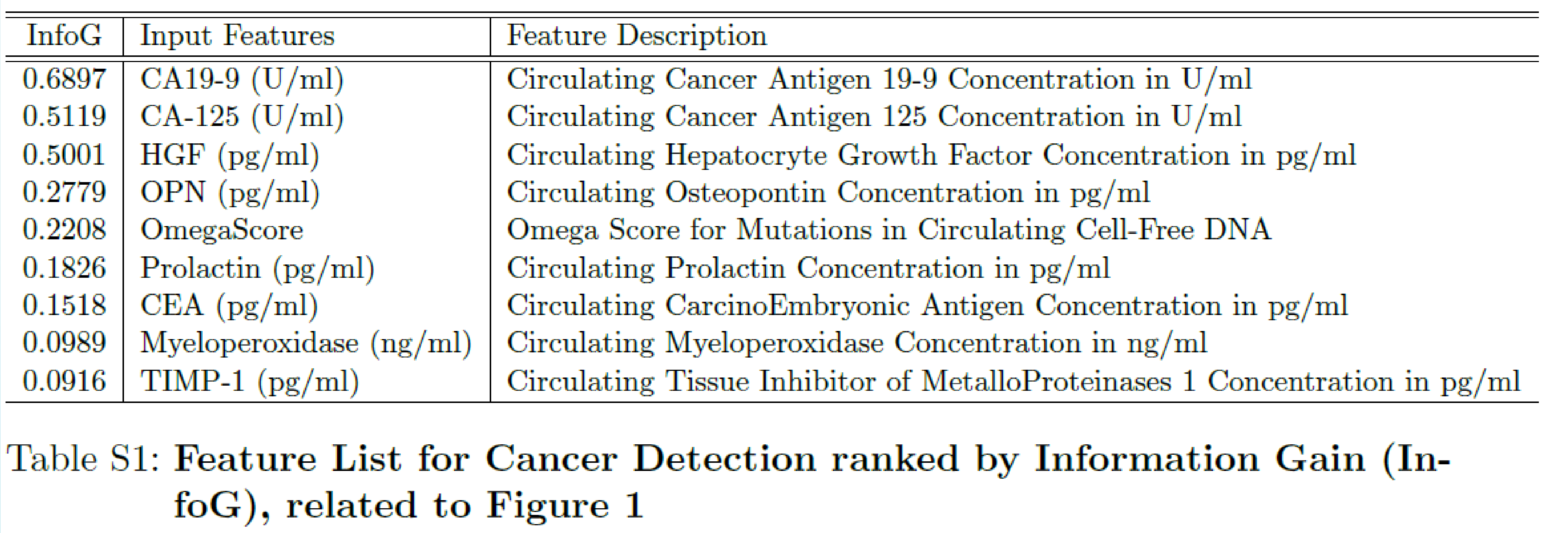
[1] Cohen, Joshua D., et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test." *Science* 359.6378 (2018): 926-930.

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 the following table.
- 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 the later slides.

First Dataset

Feature Ranking for Binary Cancer Detection



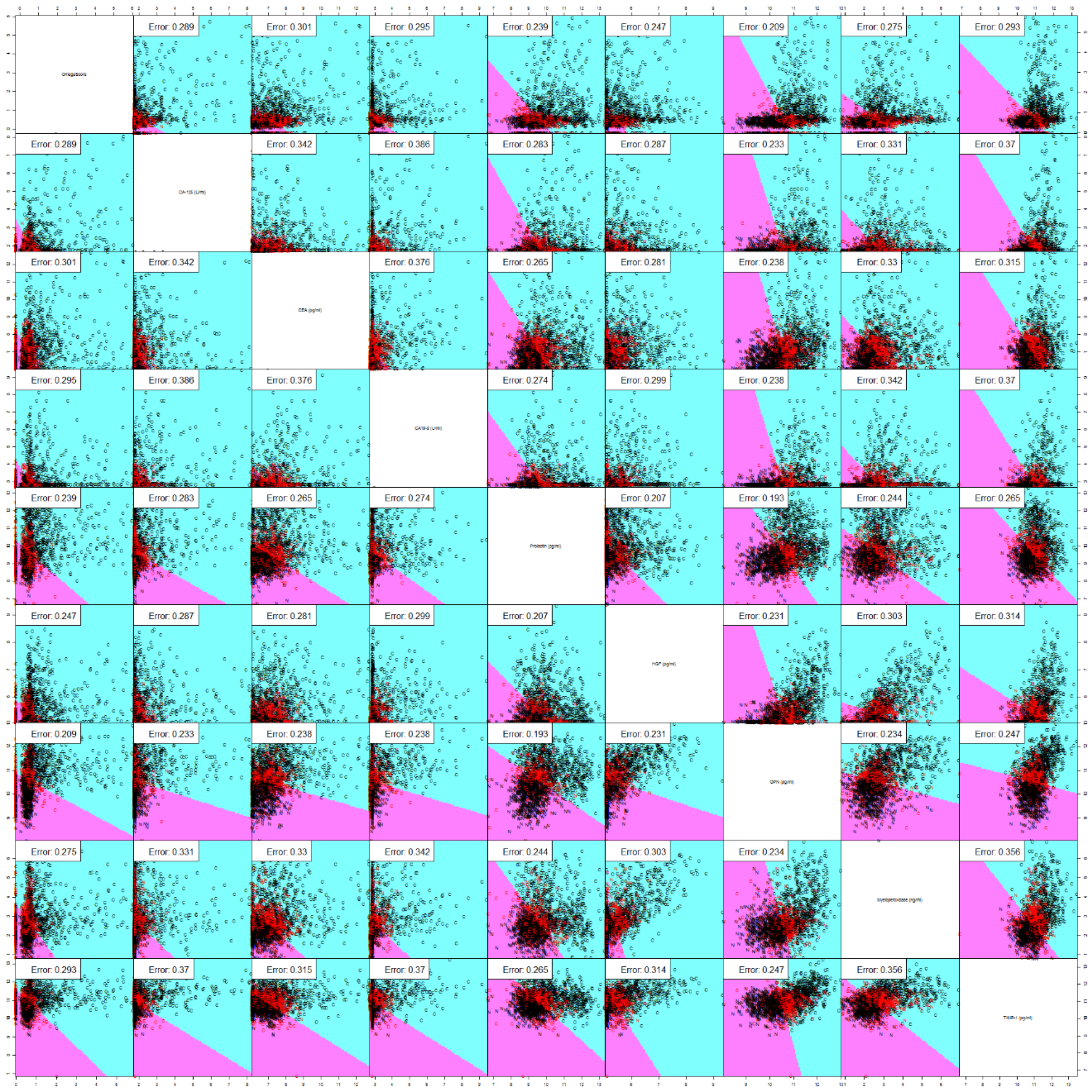
InfoG	Input Features	Feature Description
0.6897	CA19-9 (U/ml)	Circulating Cancer Antigen 19-9 Concentration in U/ml
0.5119	CA-125 (U/ml)	Circulating Cancer Antigen 125 Concentration in U/ml
0.5001	HGF (pg/ml)	Circulating Hepatocryte Growth Factor Concentration in pg/ml
0.2779	OPN (pg/ml)	Circulating 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**

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

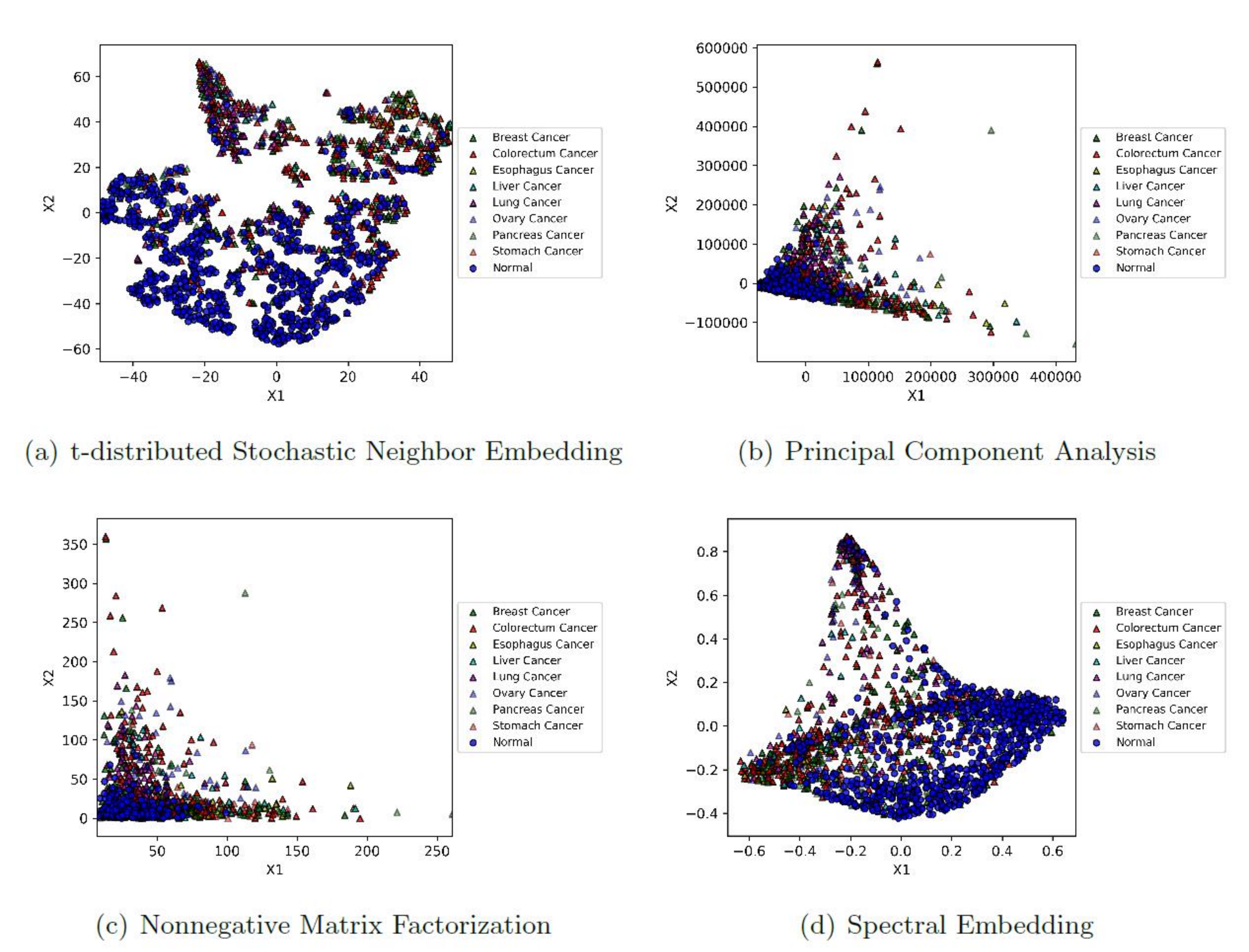
First Dataset

Linear Discriminant Analysis



First Dataset

Dimensional Reduction



(a) t-distributed Stochastic Neighbor Embedding

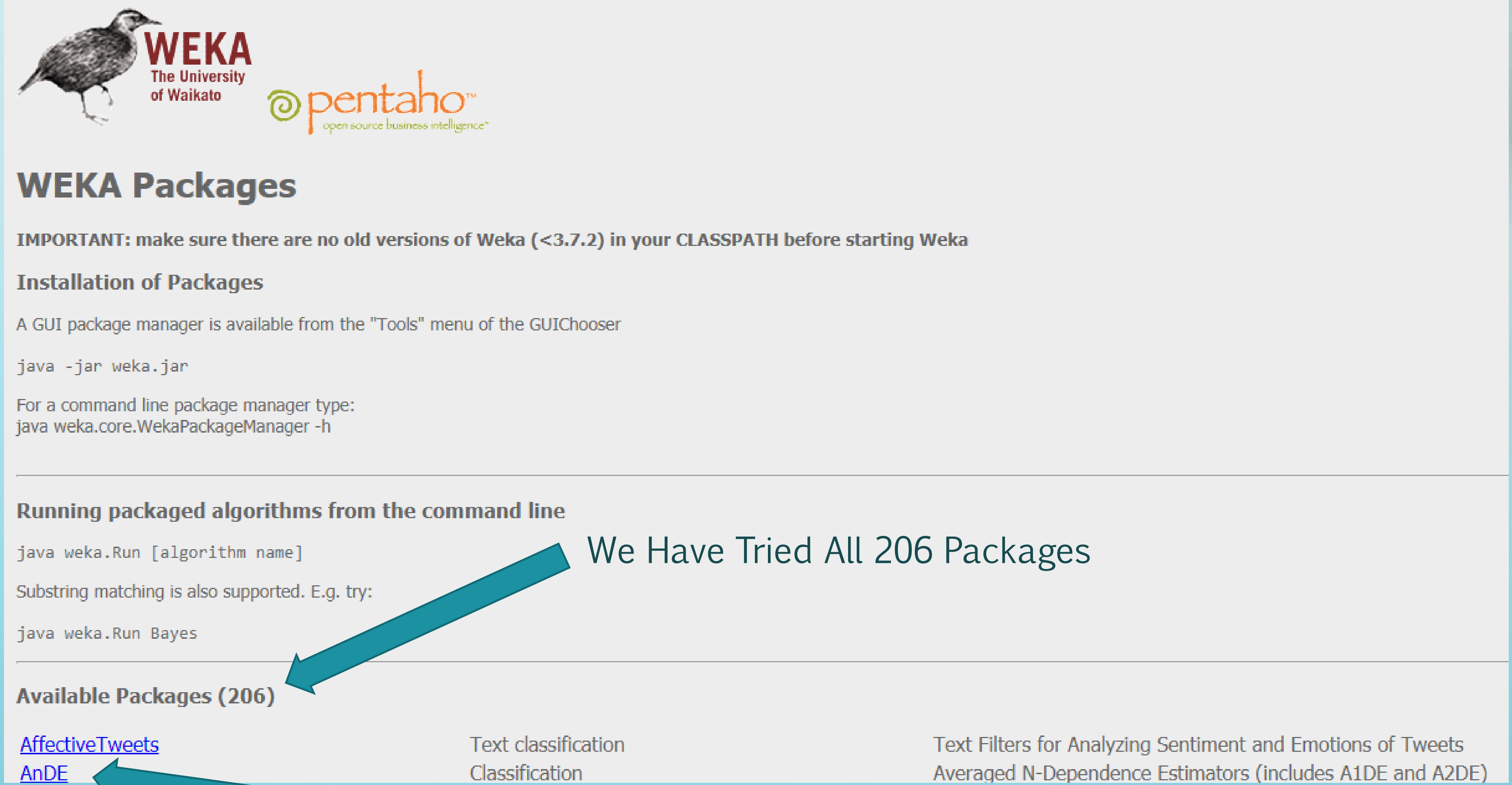
(b) Principal Component Analysis

(c) Nonnegative Matrix Factorization

(d) Spectral Embedding

First Dataset

Proposed Approach – A1DE



WEKA Packages

IMPORTANT: make sure there are no old versions of Weka (<3.7.2) in your CLASSPATH before starting Weka

Installation of Packages

A GUI package manager is available from the "Tools" menu of the GUIChooser

```
java -jar weka.jar
```

For a command line package manager type:
java weka.core.WekaPackageManager -h

Running packaged algorithms from the command line

```
java weka.Run [algorithm name]
```

Substring matching is also supported. E.g. try:

```
java weka.Run Bayes
```

Available Packages (206)

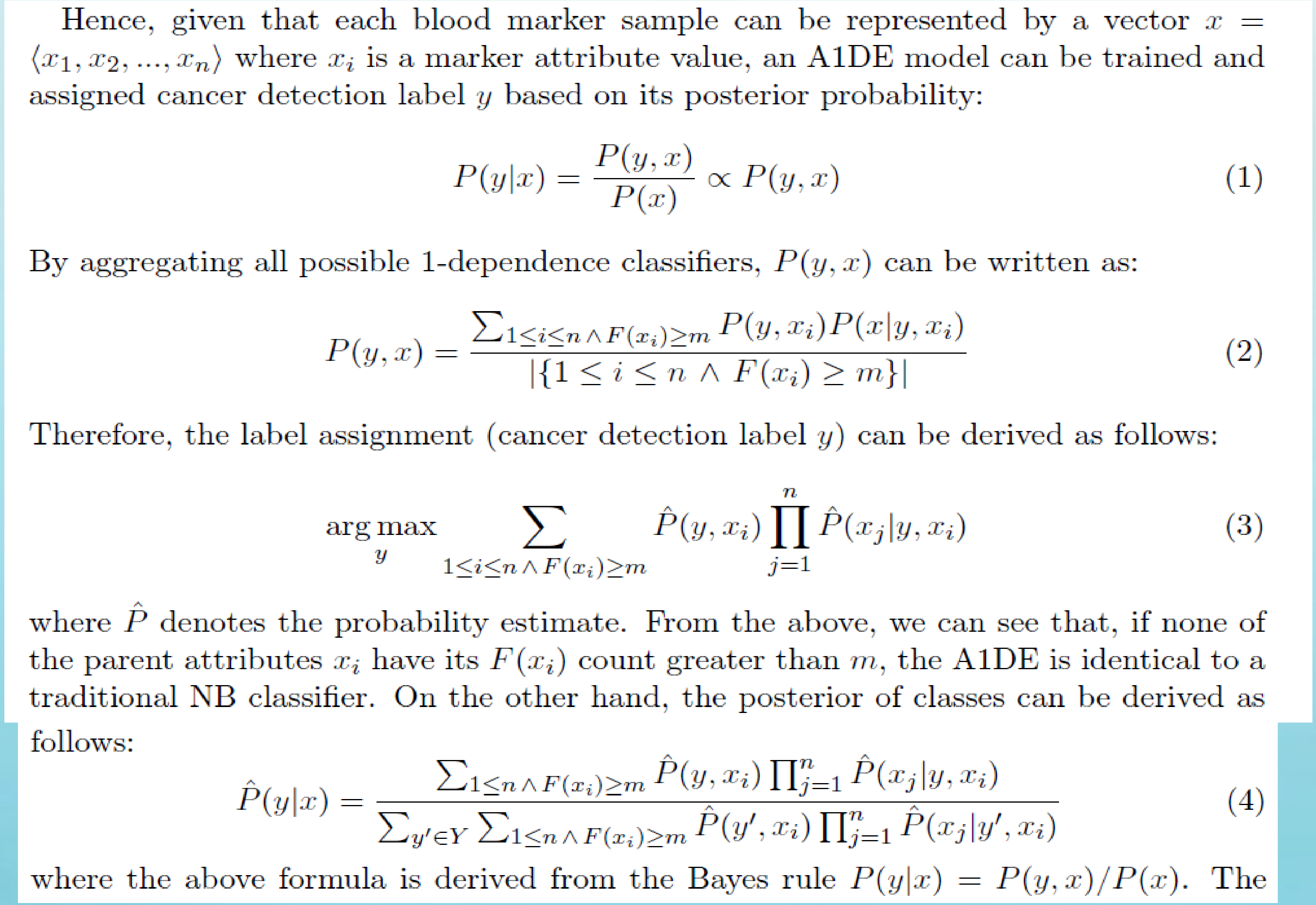
AffectiveTweets	Text classification	Text Filters for Analyzing Sentiment and Emotions of Tweets
AnDE	Classification	Averaged N-Dependence Estimators (includes A1DE and A2DE)

We Have Tried All 206 Packages

It is the best for this task.

First Dataset

Proposed Approach – A1DE



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:

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

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

$$P(y,x) = \frac{\sum_{1\leq i\leq n \wedge F(x_i)\geq m} P(y,x_i)P(x|y,x_i)}{|\{1 \leq i \leq n \wedge F(x_i) \geq m\}|} \qquad (2)$$

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

$$\arg\max_{y} \sum_{1\leq i\leq n \wedge F(x_i)\geq m} \hat{P}(y,x_i) \prod_{j=1}^{n} \hat{P}(x_j|y,x_i) \qquad (3)$$

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:

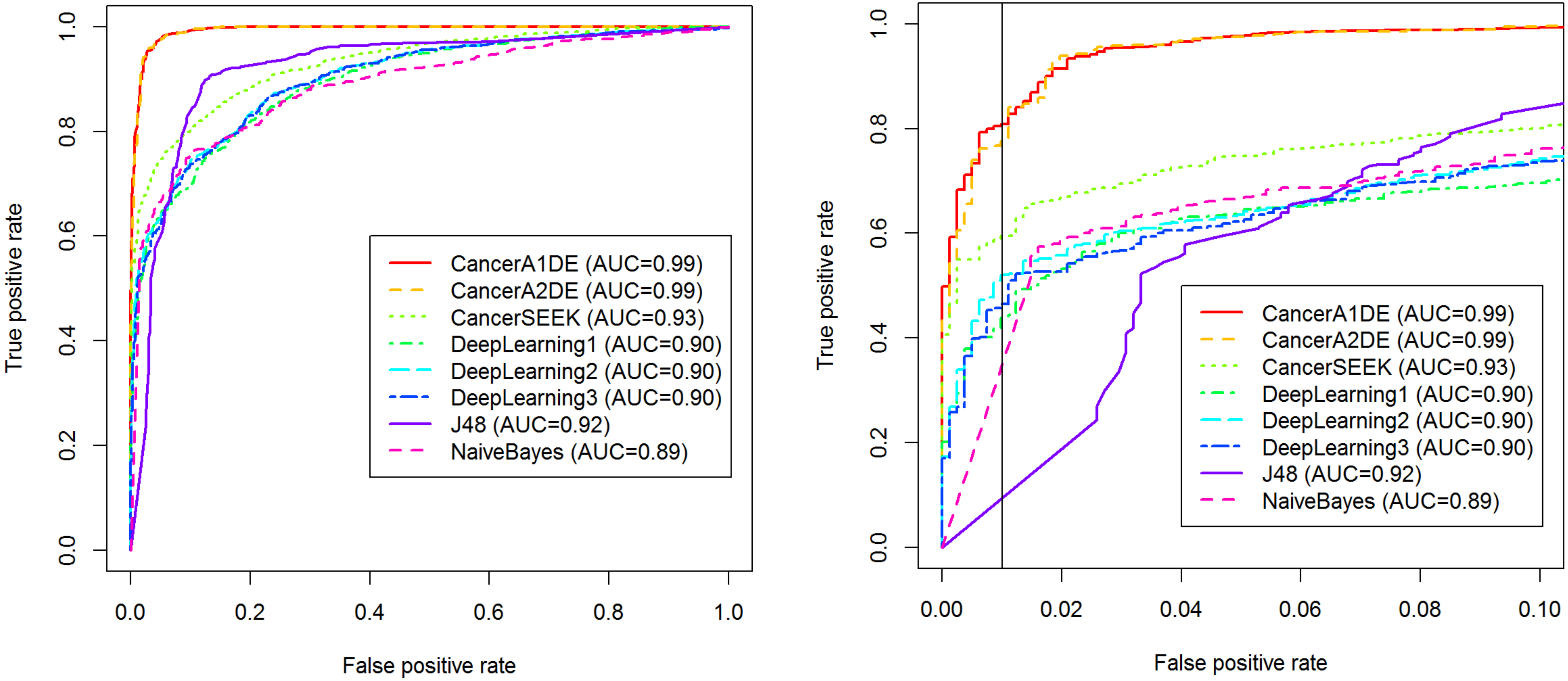
$$\hat{P}(y|x) = \frac{\sum_{1\leq n \wedge F(x_i)\geq m} \hat{P}(y,x_i) \prod_{j=1}^{n} \hat{P}(x_j|y,x_i)}{\sum_{y'\in Y} \sum_{1\leq n \wedge F(x_i)\geq m} \hat{P}(y',x_i) \prod_{j=1}^{n} \hat{P}(x_j|y',x_i)} \qquad (4)$$

where the above formula is derived from the Bayes rule $P(y|x) = P(y,x)/P(x)$. The

*with Minimum Description Length (MDL) feature discretization

First Dataset

ROC Curves for Binary Cancer Detection



Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

First Dataset

Sensitivities (Recalls) for Binary Cancer Detection

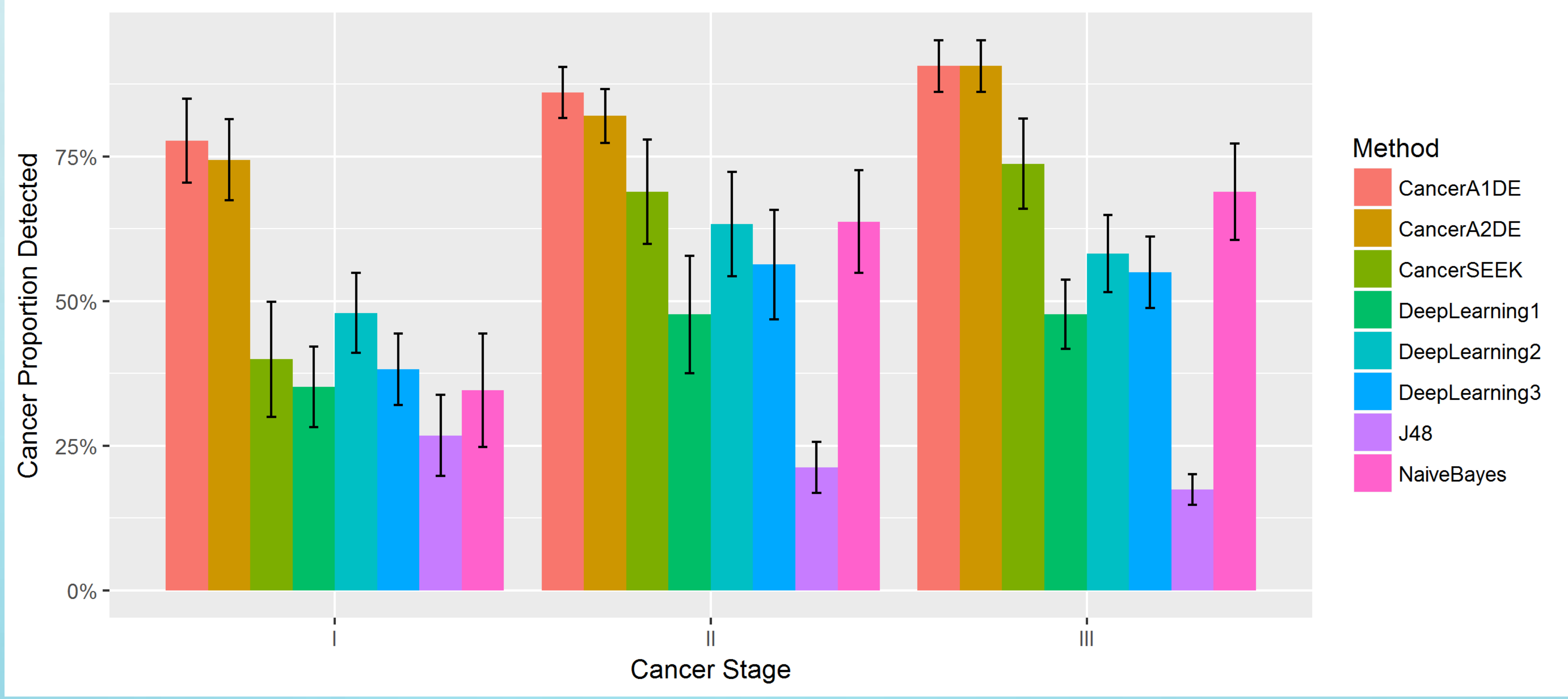


Figure 2. Proportion of Detected Cancers with Different Stages at the 99% Specificity Level Each color represents a method, and the horizontal axis has been ordered by cancer stages. Each bar represents the median sensitivity of each method on each cancer stage with standard errors.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

First Dataset

Sensitivities (Recalls) for Binary Cancer Detection

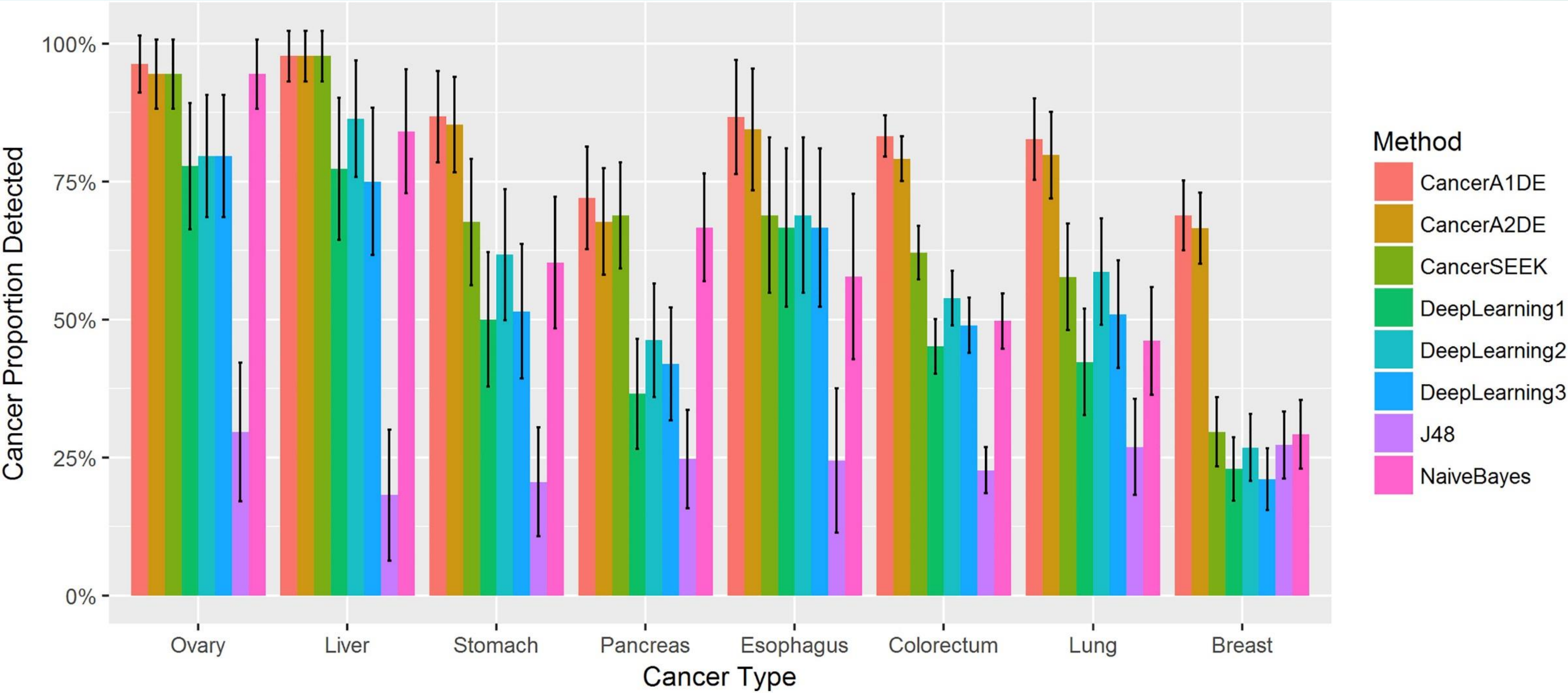
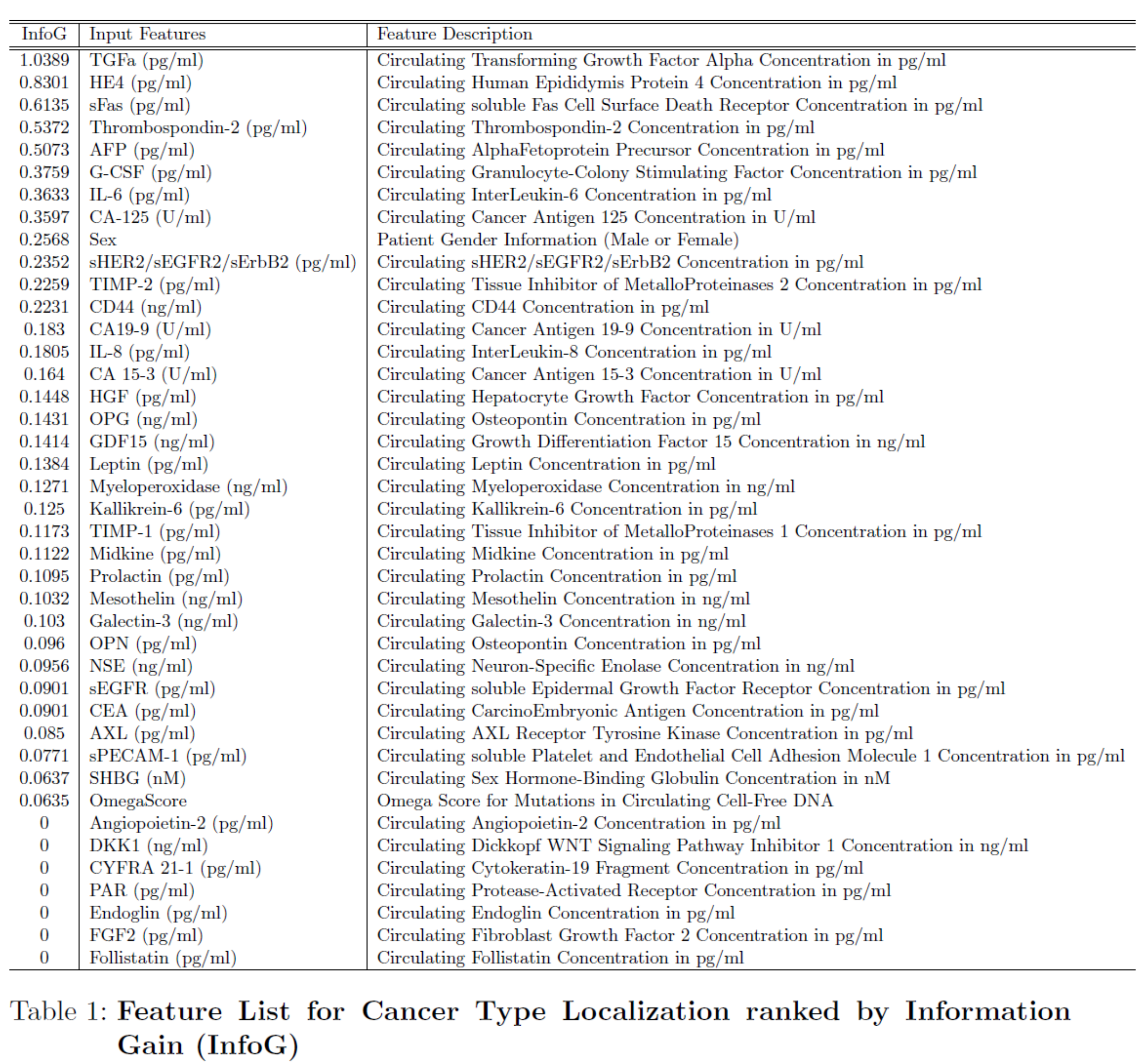


Figure 3. Detected Proportions of Different Cancer Types at the 99% Specificity Level
Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Feature Ranking for Multiple Cancer Classification



InfoG	Input Features	Feature Description
1.0389	TGFa (pg/ml)	Circulating Transforming Growth Factor Alpha 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 AlphaFetoprotein 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 Hepatocryte 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
0.0637	SHBG (nM)	Circulating Sex Hormone-Binding Globulin Concentration in nM
0.0635	OmegaScore	Omega Score for Mutations in Circulating Cell-Free DNA
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: **Feature List for Cancer Type Localization ranked by Information Gain (InfoG)**

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Sensitivities (Recalls) for Multiple Cancer Classification

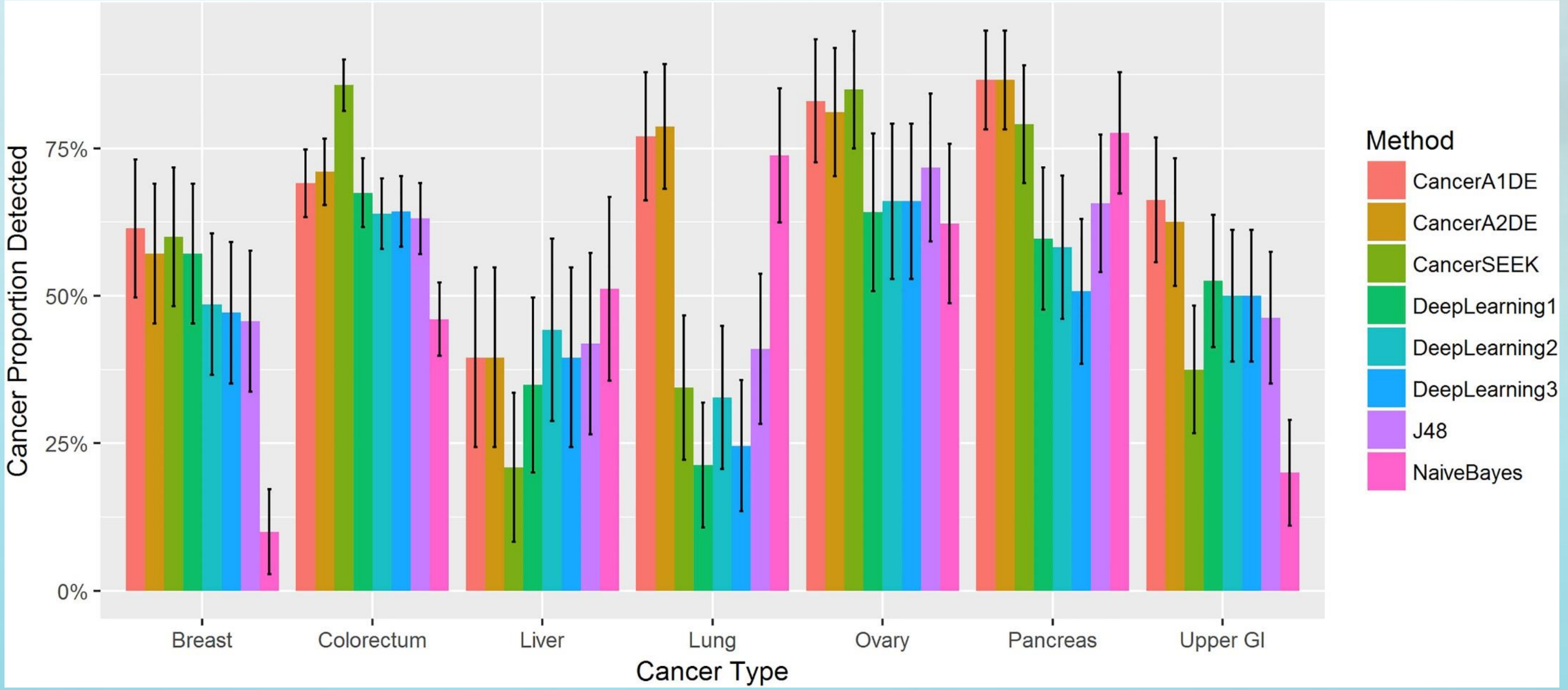


Figure 4. Localized Proportions of Different Cancer Types using the Top One Prediction Approach
Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Feature Importance Bi-Clustering

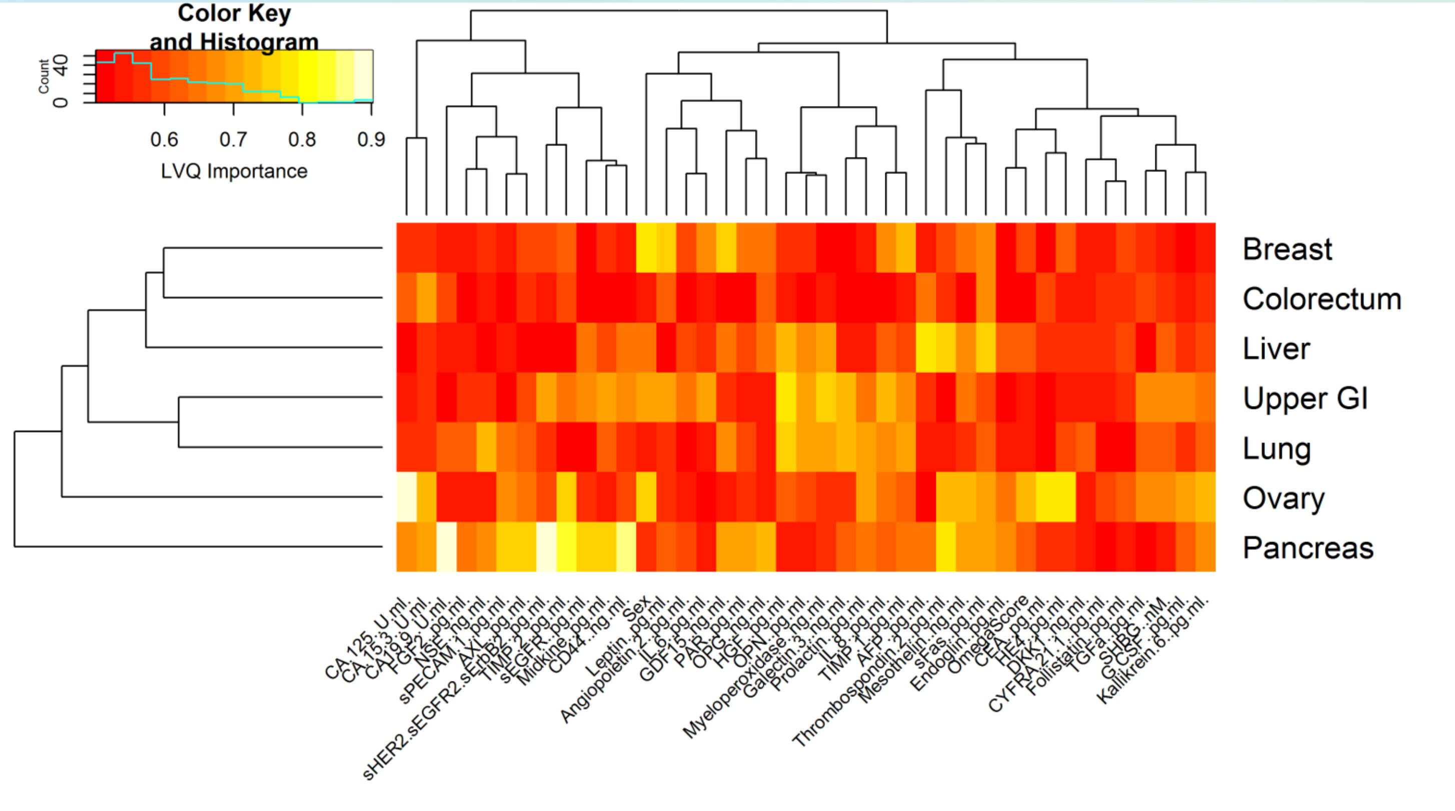


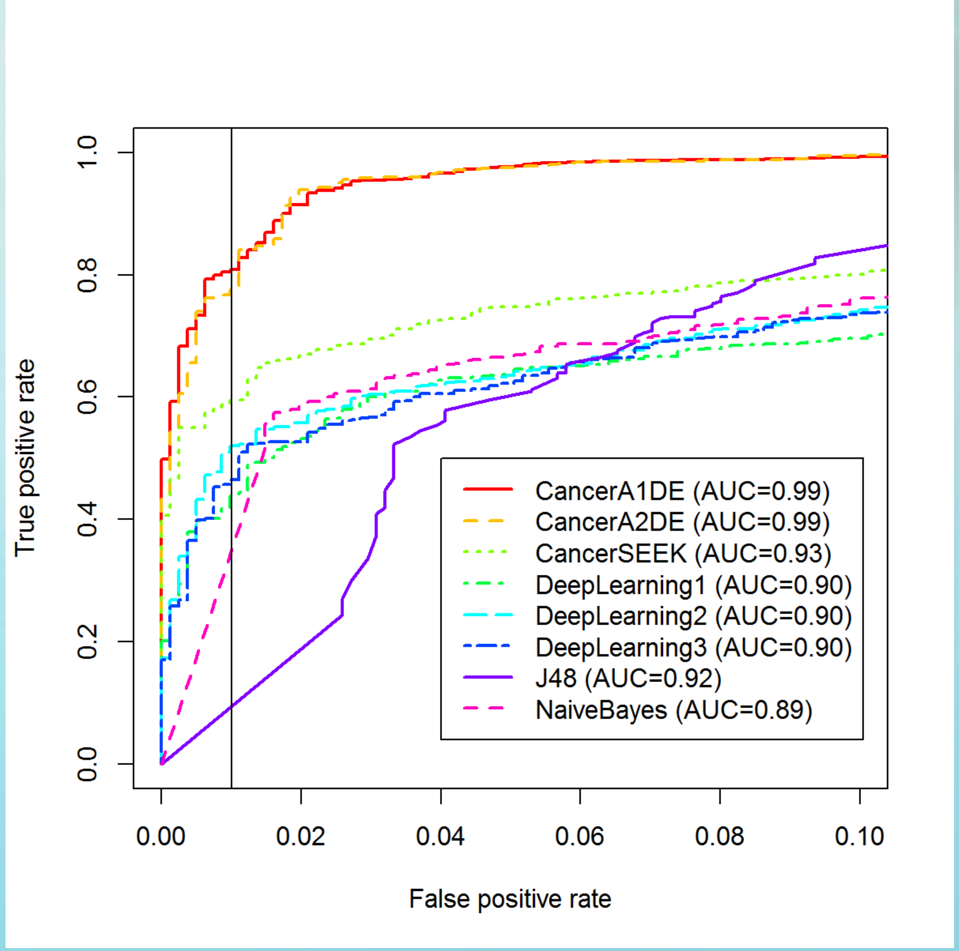
Figure 5. Feature Importance Heatmap for Cancer Type Localization under One-Class-versus-Others Setting The feature rankings are measured based on the Learning Vector Quantization (LVQ) building under Python caret package (Bischl et al., 2016). Ten-fold cross-validations are run to compute the feature 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.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Conclusion

Summary

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/.



Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.