

# Machine Learning Algorithms Significantly Improve the Accuracy of Multi Tumor Biomarker Panel for the Early Detection of Multiple Cancers

Victoria Doseeva, PhD<sup>1</sup>, Hsin-Yao Wang, MD<sup>2</sup>, Chun-Hsien Chen<sup>2</sup>, Richard Scherer<sup>1</sup>, David Schodin, PhD<sup>1</sup>, and Jiming Zhou, PhD<sup>1</sup>

<sup>1</sup>20/20 GeneSystems, Inc., Rockville, MD, and <sup>2</sup>Chang Gung Memorial Hospital at Linkou, Taoyuan City, Taiwan

## ABSTRACT

Early cancer detections significantly improve patient survival rates (5-year survival rate for Stage I cancers 90–95% vs. Stage IV cancers of 1–2%). The probability of developing cancer in one's lifetime is calculated through genetic based analysis. Meanwhile, tumor proteins in blood can indicate the presence of cancer within one year while patients are still asymptomatic. Screening of multi tumor biomarker panels using the blood routinely drawn at/or prior to an annual visit (for testing cholesterol, liver enzymes, etc.) is extremely convenient and appealing for early detection of cancers through screening asymptomatic subjects. In East Asia and beyond, this has been a useful tool in early cancer detection, and millions of individuals undertake these screenings annually. Present approach for the screening tests is to simply adopt test kit manufacturer's cut-off value, resulted in low sensitivities and specificities, which becomes a major obstacle to the general adoption of multi tumor biomarkers panel screening in the United States.

In collaboration with the Chang Gung Memorial Hospital in Taiwan, we achieved optimized sensitivities and specificities by taking advantage of synergies between the biomarkers and clinical parameters and developed OneTest, a machine learning algorithm. OneTest utilizes a unique cohort of 12,622 asymptomatic males and 15,316 asymptomatic females who were tested with a tumor markers panel over a 12-year period in Taiwan. All male and female individuals had complete data on 6 (AFP, CEA, CA19-9, CA15-3, CA125, PSA, SCC, and CYFRA21-1) or 7 (AFP, CEA, CA19-9, CA125, CA15-3, SCC, and CYFRA21-1) tumor markers, respectively. All tumor markers were measured using commercially available IVD kits and instrumentation manufactured by either Roche or Abbott Diagnostics, and were in compliance with the requirements of the College of American Pathologists (CAP) Laboratory Accreditation Program. Outcome data were obtained from a cancer registry to determine whether each patient had received a new diagnosis of malignancy within 1 year of the tumor markers test.

**Statistical analyses:** All 27,938 individuals were randomly allocated to the training (2/3) or testing (1/3) set. All randomizations were performed using Matlab (MathWorks, Natick, MA, USA). Because of the unbalanced nature of the data sets (far greater number of non-cancers vs. true cancers) used in this study, data reprocessing was performed to improve the selection of negative samples using a stratified sampling technique. AFP, CEA, CA19-9, CYFRA21-1, SCC and PSA were determined for all 12622 individuals. A variable selection process was applied to select robust variables from these serum tumor markers to design cancer detection models. The accuracy, sensitivity, specificity, AUC (area under the curve), and Youden index were compared to select the best machine learning models. The Youden index was used as a performance indicator for selecting the variables used in the classifier models in this study. The Youden index, which is among the most widely used performance indicators in biomedical studies, is calculated using the following formula: Youden index = Sensitivity + Specificity – 1.

## INTRODUCTION

Earlier detection of cancer often leads to curative surgery or earlier treatment for patients that would otherwise present with lethal, incurable later stage disease, and significantly improve patient survival rates. According to American Cancer Society, the 5-year survival rate for Stage I Non-small cell lung cancer is 92% vs. Stage IV of 1%.

A number of tumor markers have been identified for different types of cancer; however, to date no single tumor marker (except for PSA) has demonstrated high levels of specificity or sensitivity in the general early detection of cancer. To address the limitations of individual biomarkers, numerous academic and industry scientific studies have demonstrated that the combination of established biomarkers into optimized multi-biomarker panels provide increased diagnostic accuracy, making such panels much improved cancer detection tool. While in the United States only PSA, for the early detection of prostate cancer in men over 50 years, has received FDA approval for use in screening for cancer in a broad population; each day tens of thousands of individuals in East Asia undertake multi tumor biomarker panels screening using the blood routinely drawn at or prior to annual "Health Check-Ups" at hospital and specialty centers. The procedures are safe, convenient, and economical with end-user fees ranging between \$20-\$30 per biomarker.

Present approach for multi tumor biomarker panels screening is to simply adopt test kit manufacturer's reference values, resulted in low sensitivities and specificities, which becomes a major obstacle to the general adoption of multi tumor biomarker panels screening in the United States. To overcome this obstacle, scientists at 20/20 GeneSystem, teamed with the Taiwan Chang Gung Memorial Hospital, developed OneTest, the world's first A.I. powered (machine learning algorithms) multi-cancer early screening platform, through a 12-year large clinical studies involving 27,938 asymptomatic individuals. OneTest significantly improves screening sensitivities of present single threshold by 200% to 500%.

## MATERIALS & METHODS

**Patient cohort:** 12,622 asymptomatic males and 15,316 asymptomatic females were tested with a tumor marker panel over a 12-year period in Taiwan. All male and female individuals had complete data on 6 (AFP, CEA, CA19-9, CA15-3, CA125, PSA, SCC, and CYFRA21-1) or 7 (AFP, CEA, CA19-9, CA125, CA15-3, SCC, and CYFRA21-1) tumor markers, respectively. All tumor markers were measured using commercially available IVD kits and instrumentation manufactured by either Roche or Abbott Diagnostics, and were in compliance with the requirements of the College of American Pathologists (CAP) Laboratory Accreditation Program. Outcome data were obtained from a cancer registry to determine whether each patient had received a new diagnosis of malignancy within 1 year of the tumor markers test.

**Statistical analyses:** All 27,938 individuals were randomly allocated to the training (2/3) or testing (1/3) set. All randomizations were performed using Matlab (MathWorks, Natick, MA, USA). Because of the unbalanced nature of the data sets (far greater number of non-cancers vs. true cancers) used in this study, data reprocessing was performed to improve the selection of negative samples using a stratified sampling technique. AFP, CEA, CA19-9, CYFRA21-1, SCC and PSA were determined for all 12622 individuals. A variable selection process was applied to select robust variables from these serum tumor markers to design cancer detection models. The accuracy, sensitivity, specificity, AUC (area under the curve), and Youden index were compared to select the best machine learning models. The Youden index was used as a performance indicator for selecting the variables used in the classifier models in this study. The Youden index, which is among the most widely used performance indicators in biomedical studies, is calculated using the following formula: Youden index = Sensitivity + Specificity – 1.

**Statistical Algorithms for Cancer Screening:** Multiple cancer screening models based on above serum tumor markers were designed using machine learning methods, namely: SVM, kNN, MLR., Sequential Minimal Optimization (SMO), J48 decision tree, Neighborhood-Based Clustering Algorithm (NBC), Library for Support Vector Machines LibSVM, Ensemble Vote Classifier (LibSVM, LR, NBC), and Multilayer Perceptron (MLP).

## RESULTS

To design cancer detection models using machine learning methods, 6 biomarkers (AFP, CEA, CA19-9, CYFRA21-1, PSA and SCC) were evaluated. Accordingly, 63 combinations of tumor markers were evaluated using the Youden index to select an appropriate combination of variables for constructing effective cancer classification models with the highest AUC and/or Youden Index. ROC curves and AUC values were used to assess the performance of the various machine learning methods for cancer prediction. The AUC values for all various machine learning methods that integrated multiple biomarkers outperformed the individual biomarker AUC values, as previously published.

For male, the SVM (SMO, PolyKernel, no normalization) model that combined all 6 biomarkers (AFP, CEA, CA19-9, CYFRA21-1, PSA and SCC) and age attained the highest Youden Index (0.63) (Table 1). However, the highest AUC was achieved for Ridge Logistic Regression model that incorporated the same variables - 6 biomarkers and age (Table 1).

Classifier	Accuracy	AUC	Sensitivity	Specificity	Youden Idx
LibSVM (RBF)	64.94%	0.695	0.742	0.648	0.390
SMO (PolyKernel)	80.87%	0.816	0.823	0.808	0.631
KNN (k=15)	75.90%	0.839	0.790	0.759	0.549
J48 Decision Tree	85.64%	0.760	0.484	0.862	0.346
NBC	96.79%	0.826	0.210	0.979	0.189
Logistic Regression (Simple)	76.87%	0.870	0.823	0.768	0.591
Ridge Logistic Regression	80.44%	0.874	0.823	0.804	0.627
Vote (LibSVM, LR, NBC)	82.91%	0.839	0.677	0.831	0.508
MLP	68.70%	0.868	0.871	0.684	0.555

Table 1. Comparison of Various Methods for Cancer Screening (Male) using model that includes all 6 biomarkers (AFP, CEA, CA19-9, CYFRA21-1, PSA and SCC) and age.

- As a result of Leave one out analysis, the Logistic Regression model that included 5 tumor markers (without SCC) and age slightly outperformed SMO model (6 biomarkers and age) resulting in slightly higher AUC (0.872) and similar Youden Index (0.66). The best performing cancer screening models for males were selected (Table 4, Fig.1).

Male Panel	Algorithm	Biomarkers	AUC	SE (%)	SP (%)	Youden Idx
6 Biomarkers + Age	SMO	AFP, CEA, CA19-9, CYFRA21-1, PSA, and SCC	0.82	82.3	80.8	0.63
5 Biomarkers + Age	MLR	AFP, CEA, CA19-9, CYFRA21-1, and PSA	0.87	83.9	82.5	0.66
Individual Biomarkers (Single threshold method)	None	AFP, CEA, CA19-9, CYFRA21-1, PSA, and SCC	0.65	40.3	88.8	0.29

Table 4. Performance of best cancer screening algorithms for males.

- While a pan-cancer test with high specificity and sensitivity yields great promise in identifying cancers earlier while still in a premetastatic stage, there is clearly a need for follow-up in these patients to identify the cancer type and location. A balanced sensitivity and specificity are achieved when the Top three most likely affected organ systems are reported (Fig.2). To a large extent the accuracies/ sensitivities best reflect both the number of overall cases of a given cancer type in the dataset (i.e. Gastro-Intestinal (GI) and Genitourinary (GU) cancers vs. dermatological cancers) as well the nature of the biomarkers (e.g. PSA is specific for prostate and therefore GU).

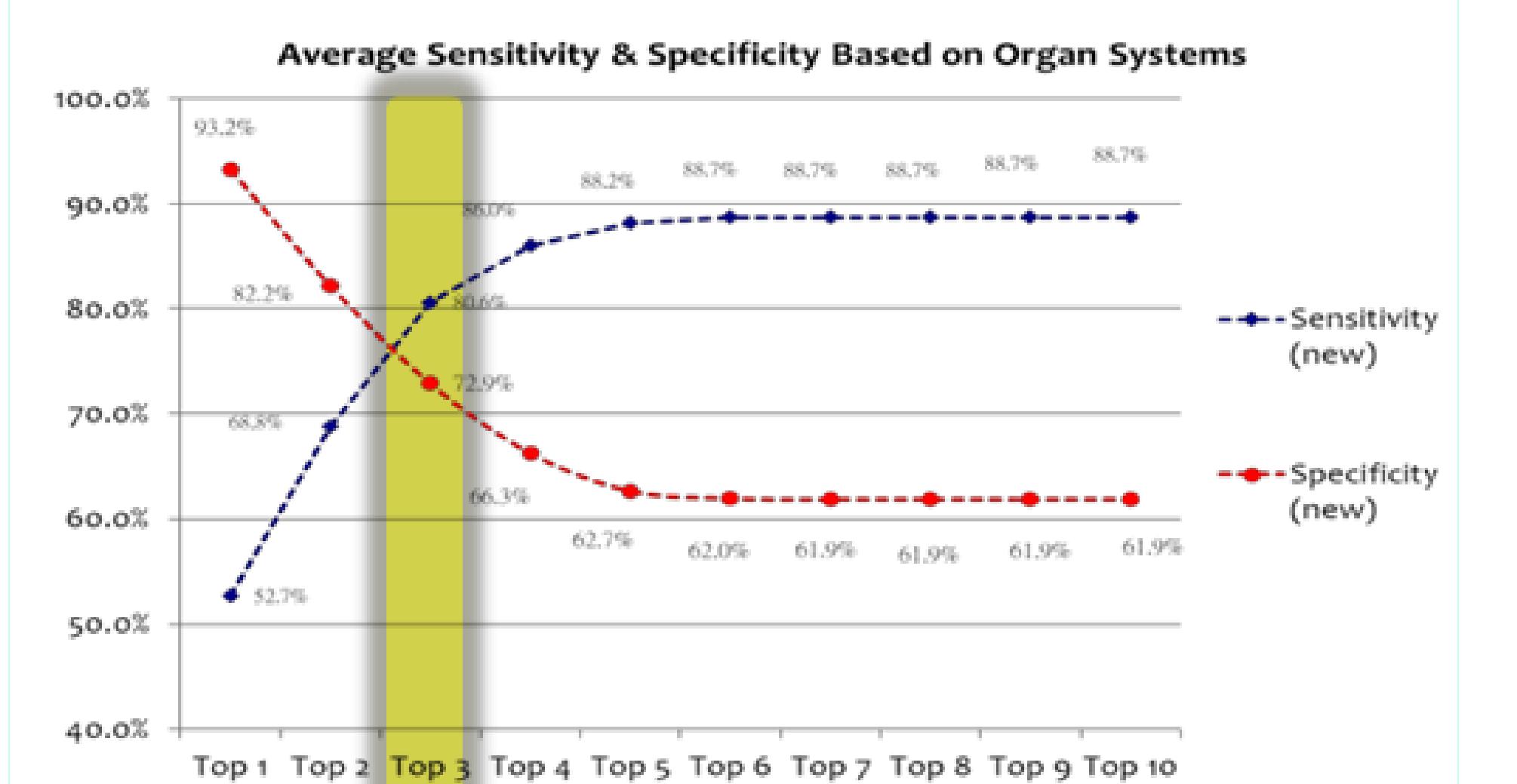


Figure 2. Performance of Top-N Male Cancer Model Based on Organ Systems.

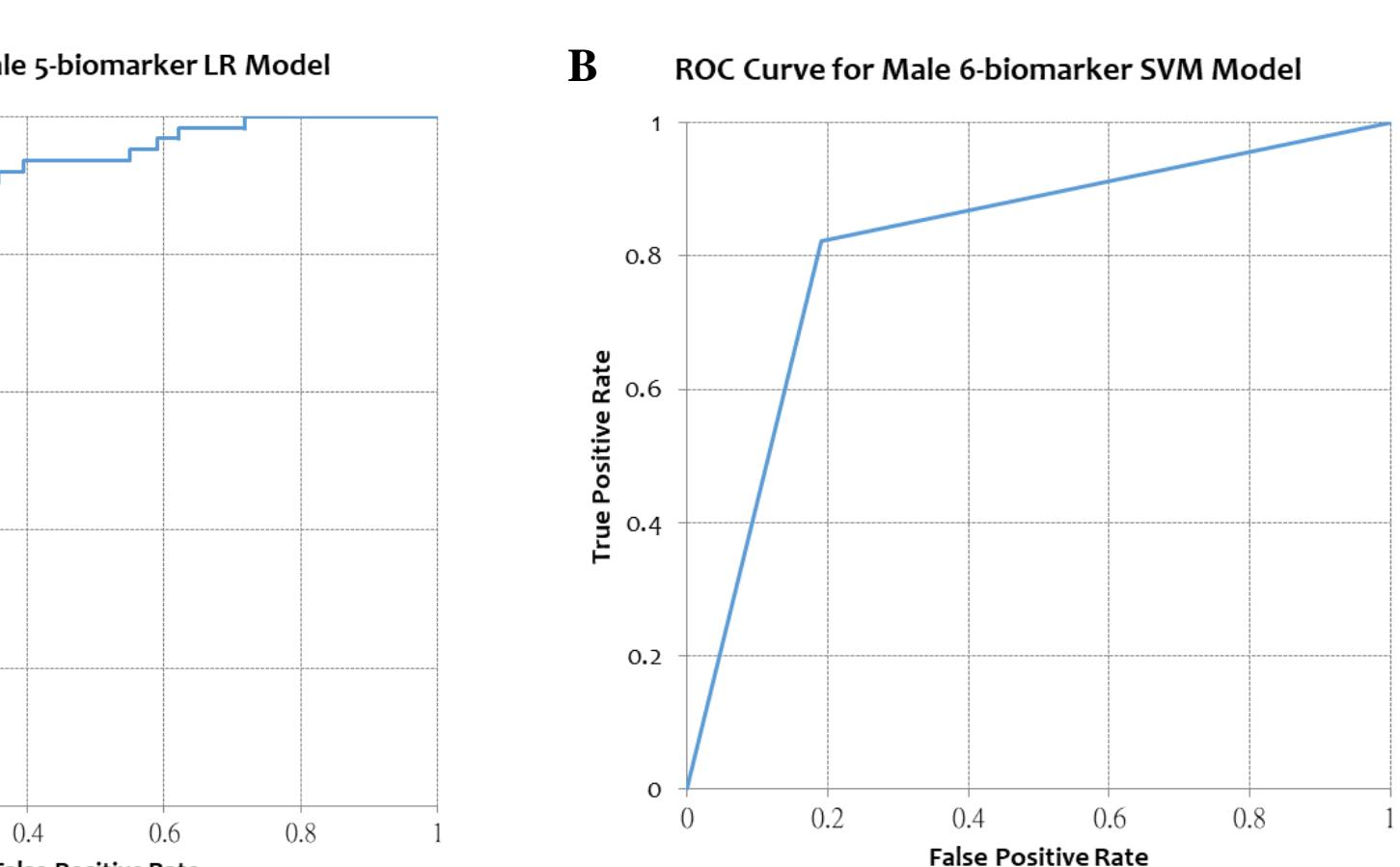


Figure 1. ROC Curves of the best Machine Learning Models (Males): A. Ridge Logistic Regression (AUC 0.872, Youden Index 0.66). B. SVM model (AUC 0.816, Youden Index 0.631).

- For females, the sensitivity and specificity of the machine learning SVM model were not as high as those for the male model (Table 5). However, the performance of the best ML model for females (BST) was also greatly improved over the single threshold method (Youden Index 0.53 vs 0.03, respectively). The ML algorithms are amenable to periodic review and redefinition. With a larger data set acquisition by combining the US and Asian cohorts, we expect to achieve a substantial improvement in the accuracy of the OneTest algorithm for females by leveraging additional data and expanding the number of clinical factor predictors.

Female Panel	Algorithm	Biomarkers	AUC	SE (%)	SP (%)	Youden Idx
6 Biomarkers + Age	BST	AFP, CEA, CA19-9, CYFRA21-1, CA15-3, CA125, and SCC	0.79	75.0	74.4	0.49
5 Biomarkers + Age	BST	AFP, CEA, CA19-9, CYFRA21-1, CA15-3, and CA125	0.71	78.2	75.0	0.53
Individual Biomarkers (Single threshold method)	None	AFP, CEA, CA19-9, CYFRA21-1, PSA, and SCC	0.51	11.5	91.5	0.03

Table 5. Performance of best cancer screening algorithms for females.