iScience, Volume *26*

Supplemental information

DEcancer: Machine learning framework tailored to liquid biopsy based cancer detection and biomarker signature selection

Andreas Halner, Luke Hankey, Zhu Liang, Francesco Pozzetti, Daniel Szulc, Ella Mi, Geoffrey Liu, Benedikt M Kessler, Junetha Syed, and Peter Jianrui Liu

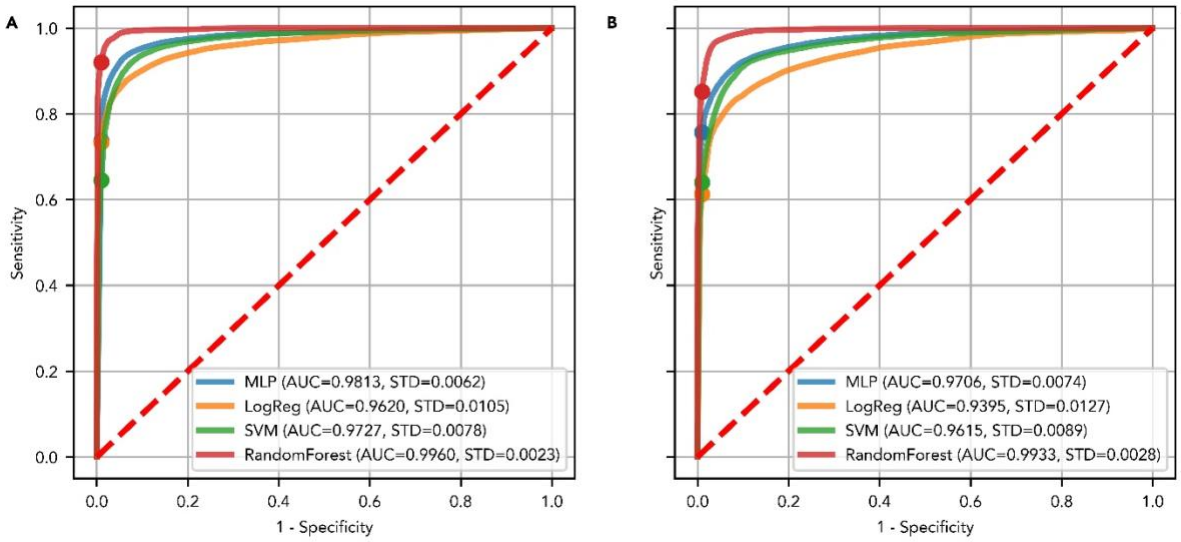

Figure S1 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for distinguishing Cohen *et al*.'s cancer-free patients from patients with one of eight types of cancer. Related to Figure 1. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable model all cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the all cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 97.26% and 97.60%, respectively. B) 28 protein model all cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 96.29% and 96.84%, respectively.

Table S1 | Test set sensitivity of all cancer versus cancer-free DEcancer$_{PDE}$ pipeline according to cancer type and stage for an overall specificity threshold of 99%. Related to Figure 2.

Stage \ Cancer	Lung	Breast	Colorectum	Oesophagus	Liver	Ovary	Pancreas	Stomach	Sensitivity by Stage	
1	1.00	0.83	1.00	1.00	1.00	1.00	0.00	1.00	0.95	
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Sensitivity by Cancer	1.00	0.98	1.00	1.00	1.00	1.00	0.95	1.00	Overall sensitivity	0.99

Table S2 | Test set sensitivity of 28 protein all cancer versus cancer-free DEcancer$_P$ pipeline according to cancer type and stage for an overall specificity threshold of 99%. Related to Figure 2.

Cancer / Stage	Lung	Breast	Colorectum	Oesophagus	Liver	Ovary	Pancreas	Stomach	Sensitivity by Stage	
1	1.00	0.83	1.00	0.00	1.00	1.00	0.00	0.75	0.90	
2	1.00	0.96	1.00	1.00	1.00	1.00	0.71	1.00	0.94	
3	1.00	0.92	0.92	1.00	1.00	1.00	1.00	1.00	0.95	
Sensitivity by Cancer	1.00	0.93	0.97	0.89	1.00	1.00	0.68	0.92	Overall sensitivity	0.94

Table S3 | Cross-validation sensitivity reported by Cohen *et al.* according to cancer type and stage for an overall specificity threshold of 99%. Related to Figure 2.

Stage \ Cancer	Lung	Breast	Colorectum	Oesophagus	Liver	Ovary	Pancreas	Stomach	Sensitivity by Stage	
1	0.43	0.38	0.43	0.20	1.00	0.89	0.25	0.71	0.48	
2	0.67	0.25	0.72	0.86	1.00	1.00	0.73	0.67	0.63	
3	0.74	0.46	0.68	0.45	0.95	1.00	0.83	0.82	0.70	
Sensitivity by Cancer	0.59	0.33	0.65	0.69	0.98	0.98	0.72	0.72	Overall sensitivity	0.62

Table S4 | Augmentation of samples for the DEcancer pipeline according to classification tasks. Balanced has an equal number of samples for both cancer and cancer-free. Cancer is imbalanced favorably and cancer-free is an imbalance in favor of cancer-free individuals. **Related to Figure 6 and STAR Methods.**

	Cancer vs cancer-free			Cancer vs other cancers			Cancer vs other cancers and cancer-free		
	Balanced	Cancer	Cancer-free	Balanced	Cancer	Cancer-free	Balanced	Cancer	Cancer-free
Breast	1630	652, 2608	2608, 652	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Colorectum	1920	768, 3072	3072, 768	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Oesophagus	1370	548, 2192	2192, 548	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Liver	1367	547, 2188	2188, 547	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Lung	1462	585, 2340	2340, 585	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Ovary	1382	553, 2212	2212, 553	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Pancreas	1447	579, 2316	2316, 579	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Stomach	1407	563, 2252	2252, 563	1607	643, 2572	2572, 643	2905	1162, 4648	4648, 1162
Pancancer	2905	1162, 4648	4648, 1162	N/A	N/A	N/A	2905	1162, 4648	4648, 1162

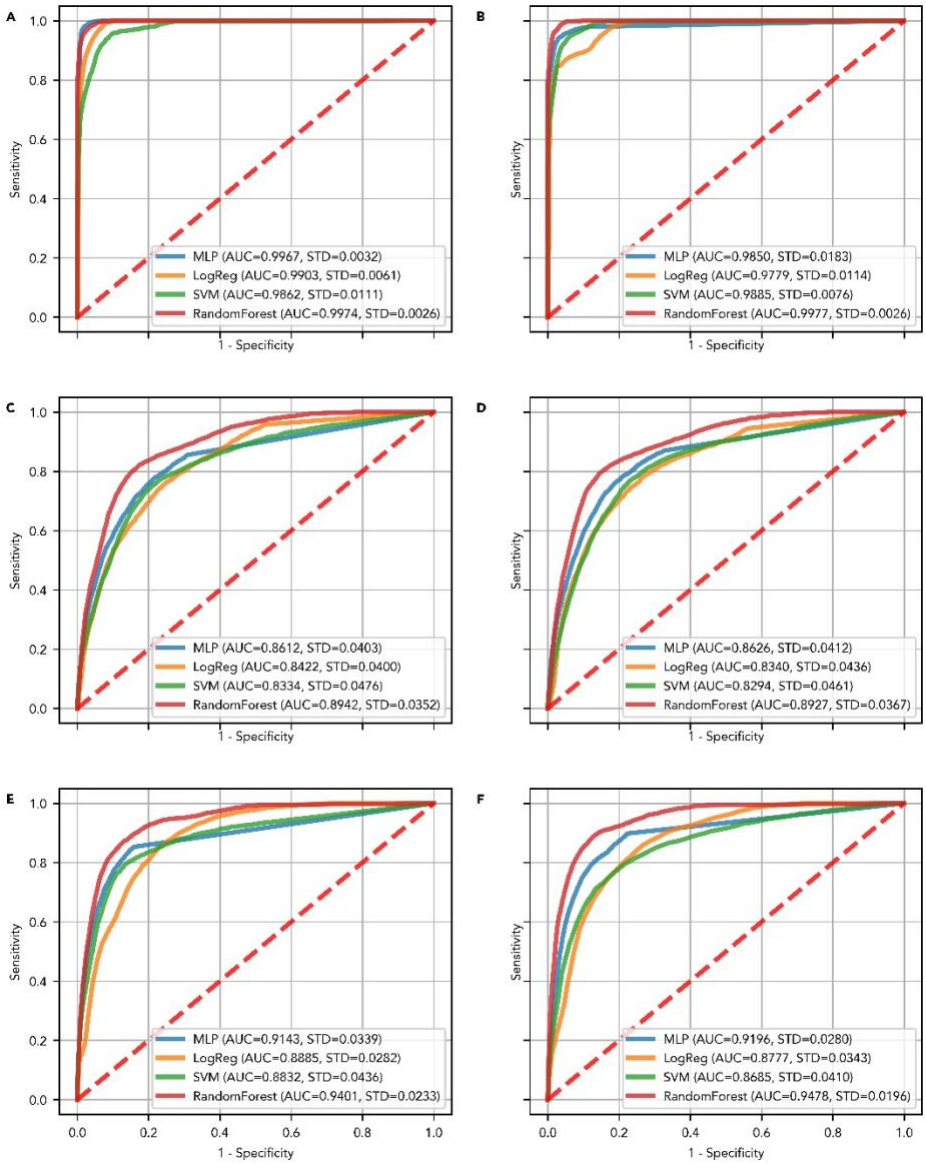

Figure S2 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al*.'s lung cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable lung cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the lung cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 93.41% and 97.88%, respectively. B) 12 protein lung cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 93.65% and 98.01%, respectively. C) Full variable lung cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the lung cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.89. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 80.79% and 83.52%, respectively. D) 39 protein lung cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.89. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 79.62% and 84.57%, respectively. E) Full variable lung cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the lung cancer versus cancer-free classification. The optimal model

is the random forest with an AUC of 0.94. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 85.79% and 87.64%, respectively. F) 22 protein lung cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.95. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.23% and 88.41%, respectively.

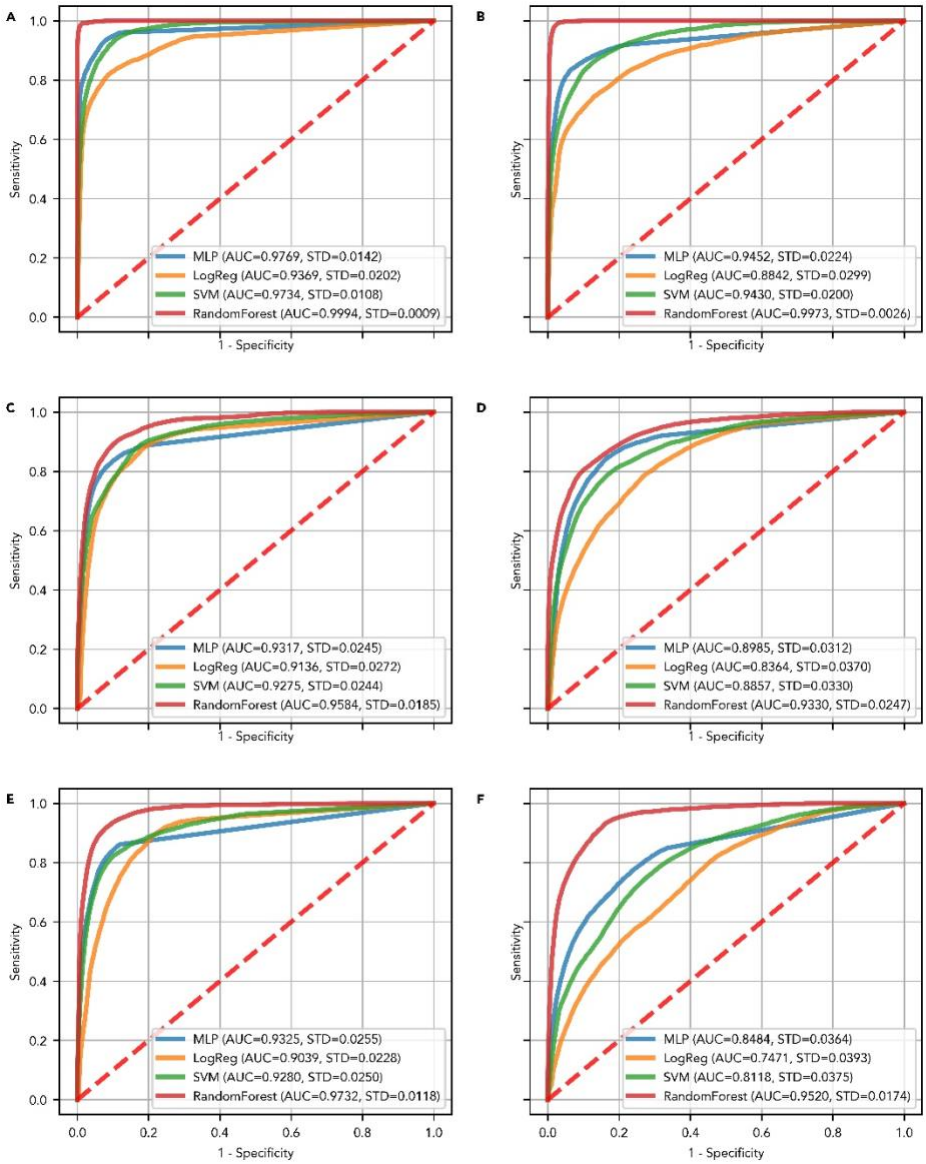

Figure S3 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s breast cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable breast cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the breast cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 96.50% and 99.35%, respectively. B) 27 protein breast cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 96.65% and 98.25%, respectively. C) Full variable breast cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the breast cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.96. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 88.55% and 90.82%, respectively. D) 29 protein breast cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.93. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 84.07% and 87.82%, respectively. E) Full variable breast cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the breast cancer versus cancer-free classification. The optimal model

is the random forest with an AUC of 0.97. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 91.74% and 92.07%, respectively. F) 26 protein breast cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.95. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 90.33% and 87.00%, respectively.

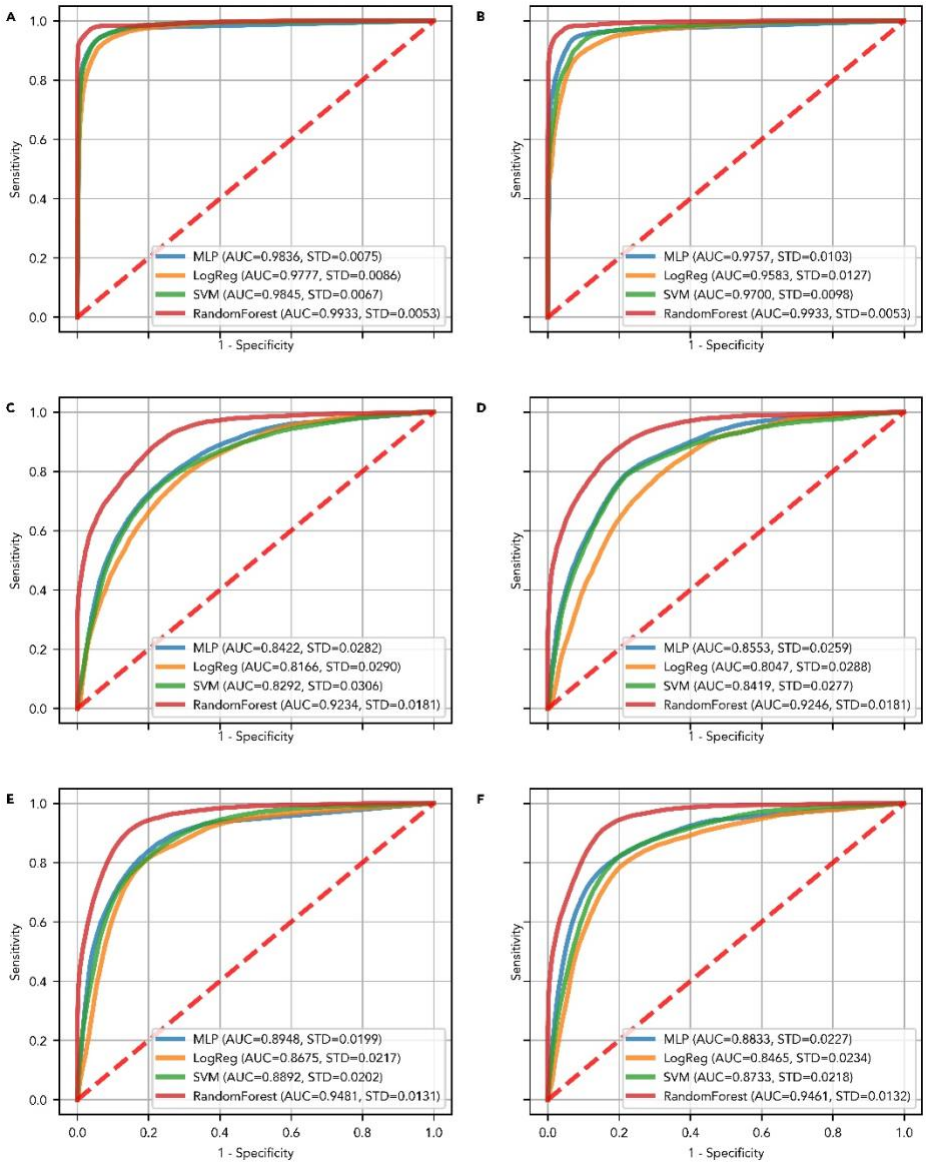

Figure S4 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s colorectal cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable colorectal cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the colorectal cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 95.55% and 98.02%, respectively. B) 22 protein colorectal cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 95.47% and 97.42%, respectively. C) Full variable colorectal cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the colorectal cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.92. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.80% and 82.32%, respectively. D) 22 protein colorectal cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.92. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.08% and 83.93%, respectively. E) Full variable colorectal cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the colorectal cancer versus

cancer-free classification. The optimal model is the random forest with an AUC of 0.95. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 90.00% and 86.90%, respectively. F) 35 protein colorectal cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.95. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 90.86% and 85.47%, respectively.

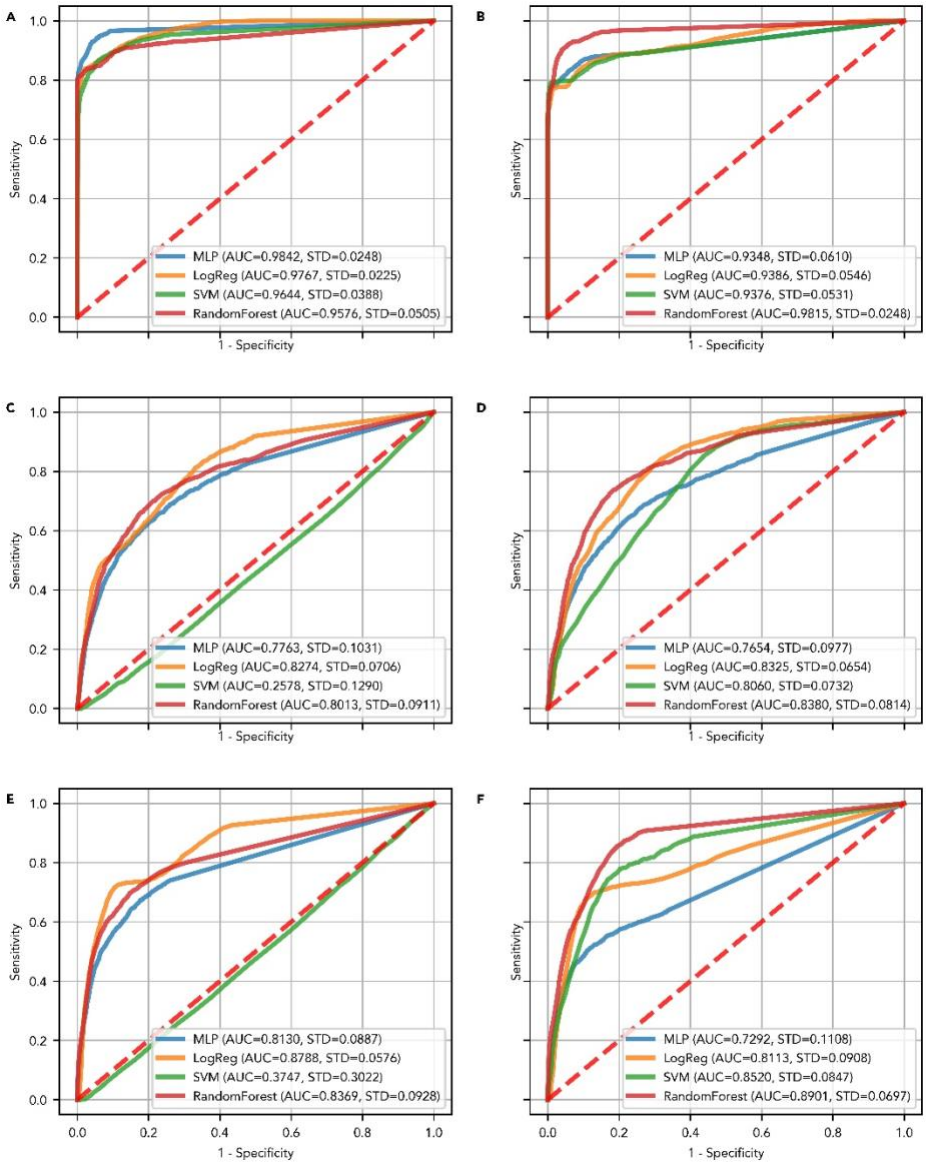

Figure S5 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al*.'s oesophageal cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable oesophageal cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the oesophageal cancer versus cancer-free classification. The optimal model is the MLP with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 82.19% and 96.97%, respectively. B) 8 protein oesophageal cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 81.53% and 95.37%, respectively. C) Full variable oesophageal cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the oesophageal cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.80. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 70.52% and 75.14%, respectively. D) 23 protein oesophageal cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.84. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 68.02% and 81.86%, respectively. E) Full variable oesophageal cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the

oesophageal cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.84. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 71.12% and 82.21%, respectively. F) 15 protein oesophageal cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.89. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 74.80% and 84.14%, respectively.

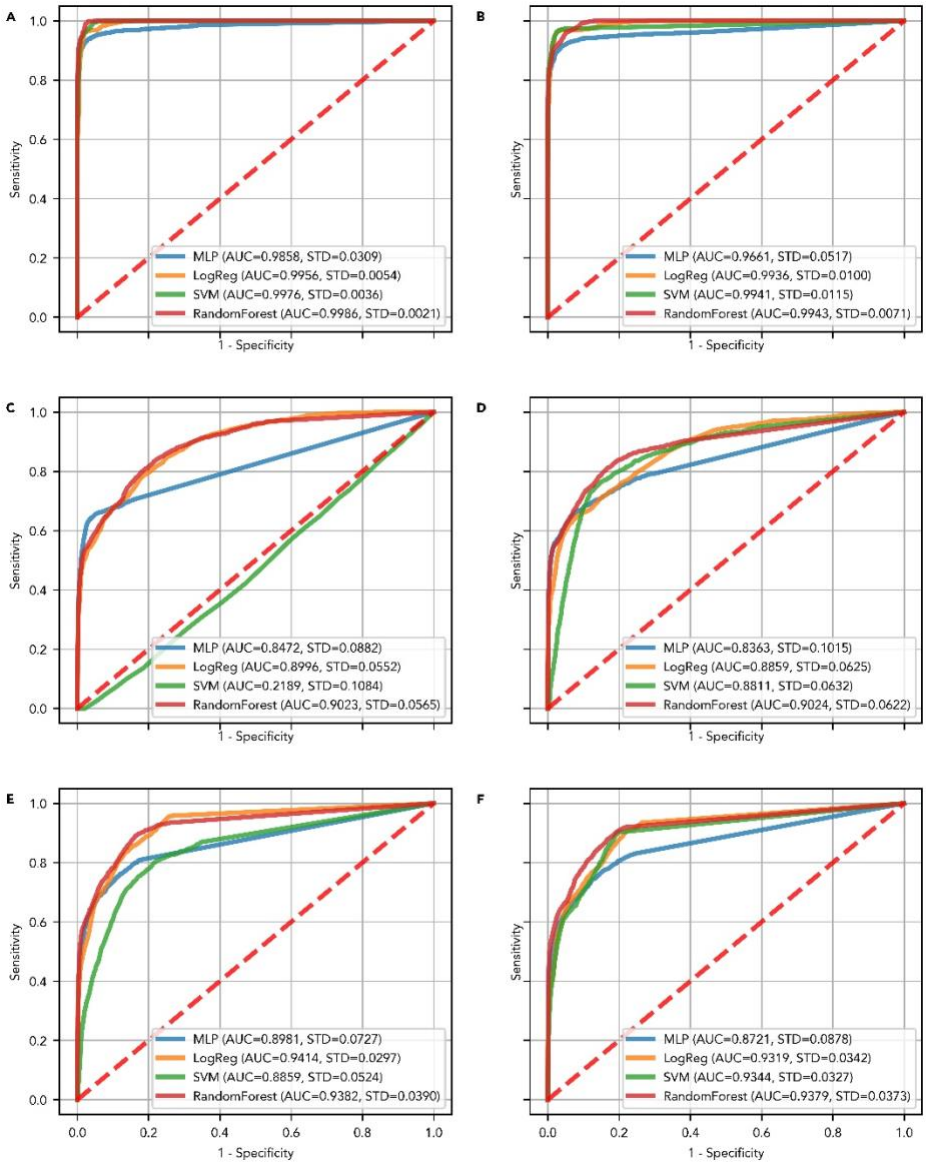

Figure S6 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s liver cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable liver cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the liver cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.81% and 99.15%, respectively. B) 23 protein liver cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.81% and 97.02%, respectively. C) Full variable liver cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the liver cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.90. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 73.14% and 84.59%, respectively. D) 8 protein liver cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.90. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 72.18% and 86.87%, respectively. E) Full variable liver cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the liver cancer versus cancer-free classification. The optimal model

is the logistic regression with l2 penalty with an AUC of 0.94. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 80.38% and 85.06%, respectively. F) 19 protein liver cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.94. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 77.13% and 87.89%, respectively.

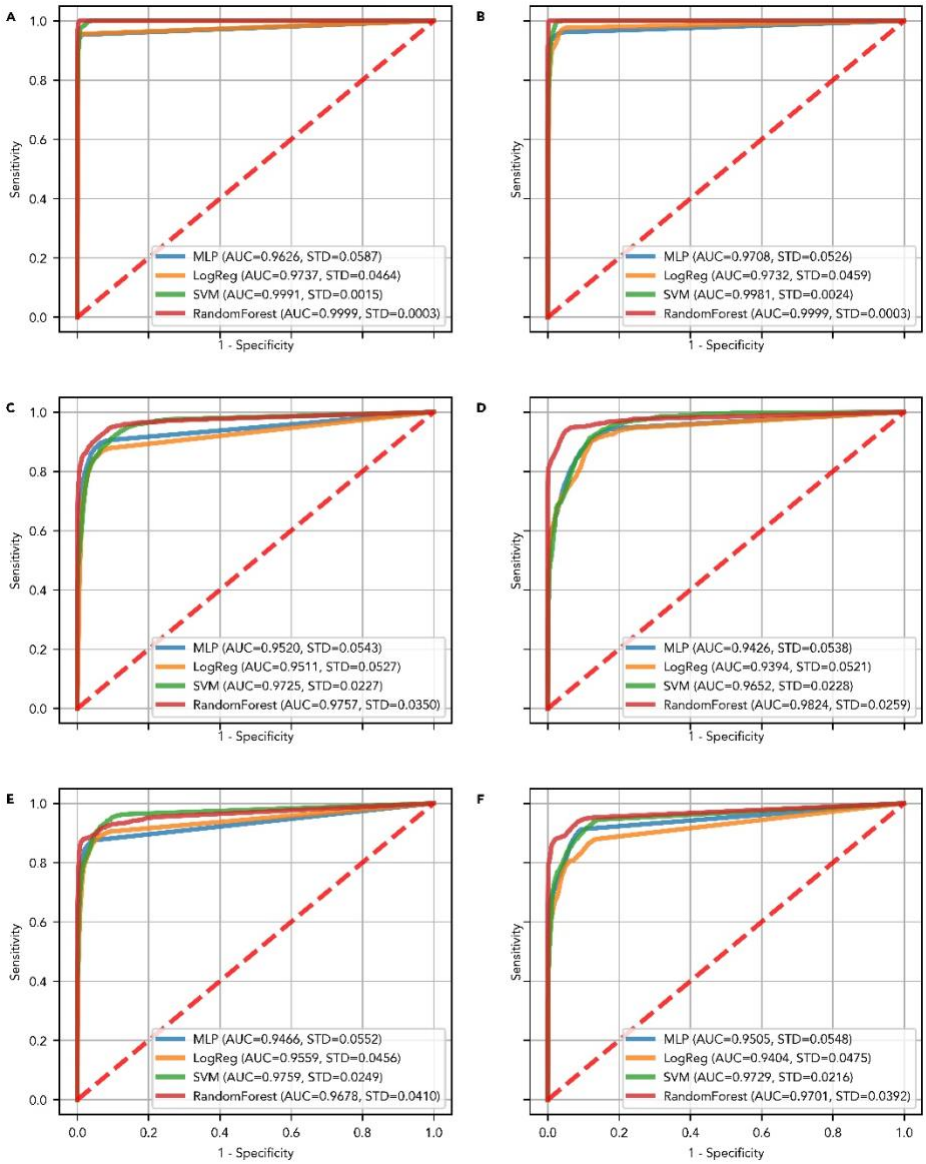

Figure S7 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s ovarian cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable ovarian cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the ovarian cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.44% and 99.94%, respectively. B) 15 protein ovarian cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 1.00. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.71% and 99.93%, respectively. C) Full variable ovarian cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the ovarian cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.14% and 95.97%, respectively. D) 13 protein ovarian cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.47% and 96.75%, respectively. E) Full variable ovarian cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the ovarian cancer versus cancer-free classification. The

optimal model is the SVM with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.45% and 95.23%, respectively. F) 12 protein ovarian cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the SVM with an AUC of 0.97. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 83.54% and 93.38%, respectively.

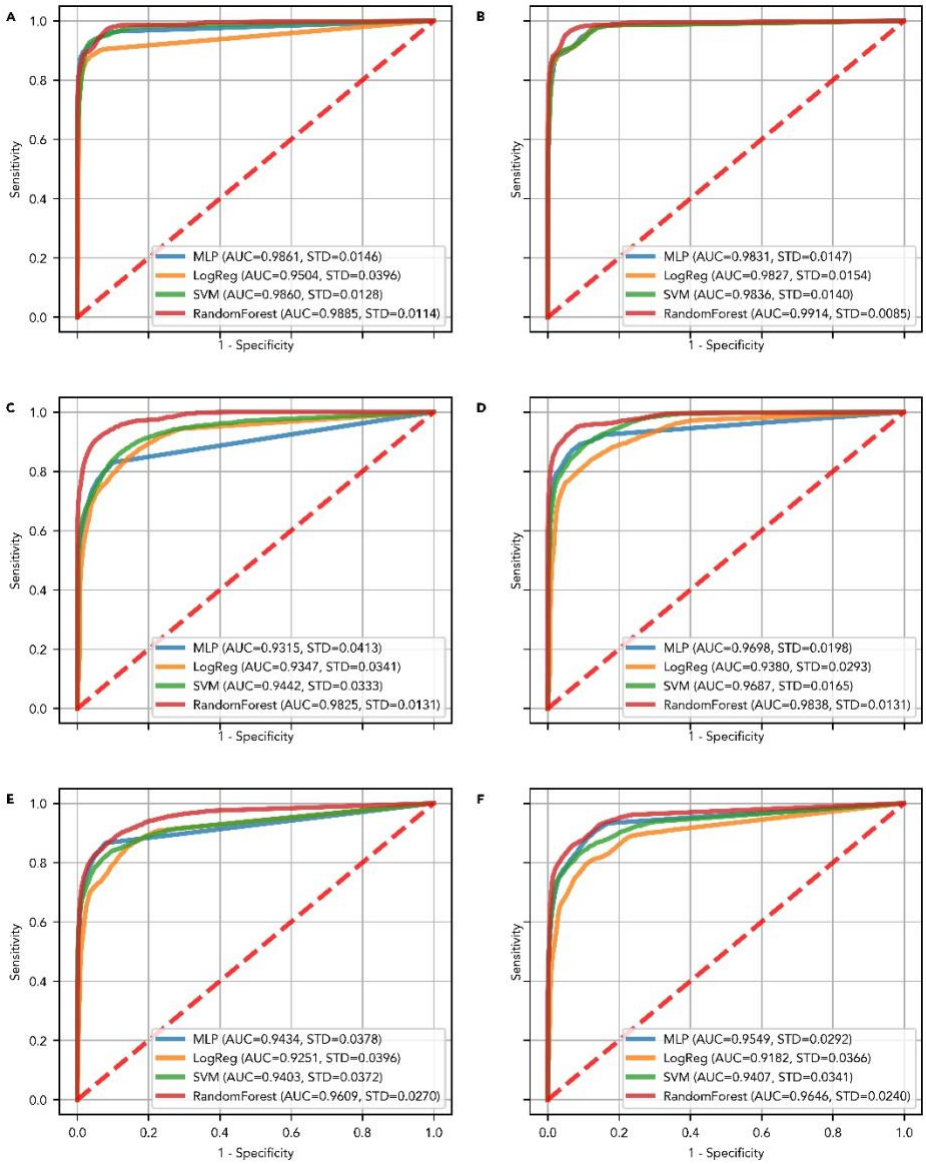

Figure S8 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s pancreatic cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable pancreatic cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the pancreatic cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 90.79% and 95.20%, respectively. B) 8 protein pancreatic cancer versus cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 91.04% and 96.03%, respectively. C) Full variable pancreatic cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the pancreatic cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 87.99% and 95.13%, respectively. D) 14 protein pancreatic cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.98. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 88.32% and 95.77%, respectively. E) Full variable pancreatic cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the pancreatic cancer versus

cancer-free classification. The optimal model is the random forest with an AUC of 0.96. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 85.48% and 91.60%, respectively. F) 9 protein pancreatic cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.96. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 86.75% and 91.52%, respectively.

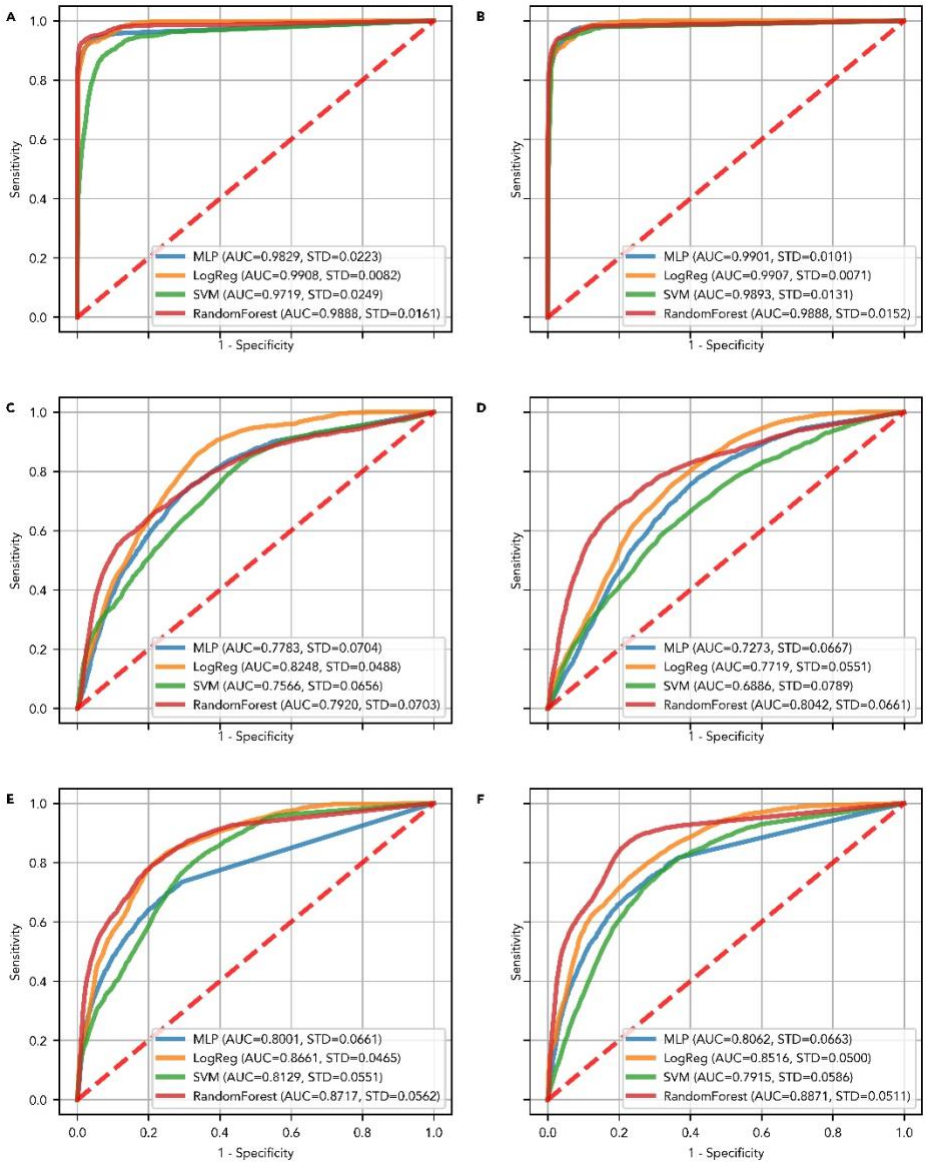

Figure S9 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Cohen *et al.*'s gastric cancer patients. Related to Figure 3. ROC curves are shown for four different classifier models (MLP = multilayer perceptron; LogReg = logistic regression with l2 penalty; SVM = support vector machine; random forest) based on average performance across Monte Carlo cross validation. A) Full variable stomach cancer versus cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the stomach cancer versus cancer-free classification. The optimal model is the logistic regression with l2 penalty with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 88.31% and 95.83%, respectively. B) 17 protein stomach cancer versus cancer-free validation ROC curves. The optimal model is the logistic regression with l2 penalty with an AUC of 0.99. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 89.15% and 95.20%, respectively. C) Full variable stomach cancer versus other cancers validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and ethnicity for the stomach cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.79. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 81.13% and 70.24%, respectively. D) 19 protein stomach cancer versus other cancers validation ROC curves. The optimal model is the random forest with an AUC of 0.80. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 67.26% and 80.54%, respectively. E) Full variable stomach cancer versus other cancers or cancer-free validation ROC curves. The full variable model uses 39 proteins, omega score representing DNA mutations, age, sex and

ethnicity for the stomach cancer versus cancer-free classification. The optimal model is the random forest with an AUC of 0.87. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 75.17% and 81.85%, respectively. F) 25 protein stomach cancer versus other cancers or cancer-free validation ROC curves. The optimal model is the random forest with an AUC of 0.89. The optimal sensitivity and specificity corresponding to the top left corner of ROC curve is 79.16% and 81.90%, respectively.

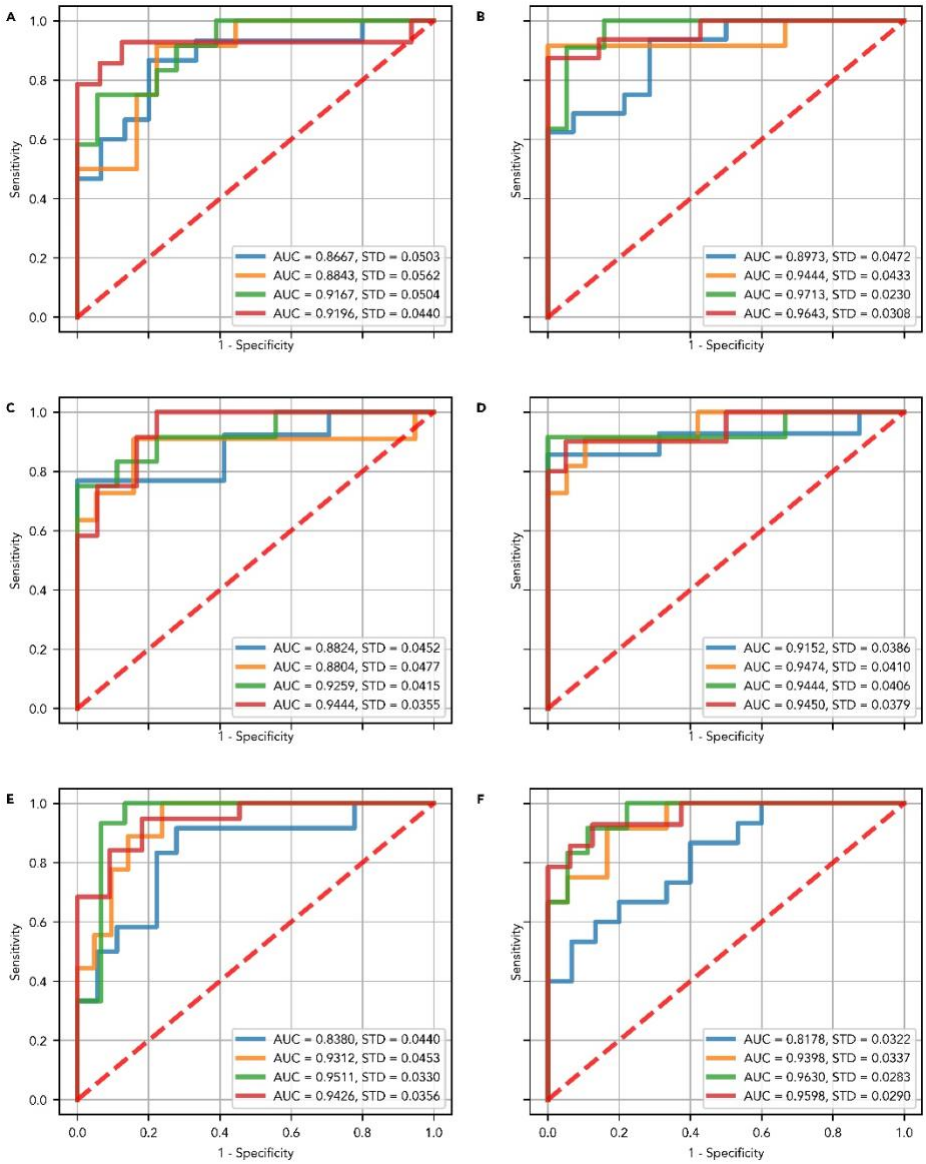

Figure S10 | Validation set receiver operating characteristic (ROC) curves and area under receiver operating characteristic curve (AUC) to select classifier models for detecting Blume *et al.*'s lung cancer patients for each spion or depleted plasma. Related to Figure 5. For each spion or depleted plasma, the optimal subset of proteins used by classifier models to distinguish between cancer-free and non-lung cancer samples is indicated. Classifier models used in each panel are coloured as: Blue = Multilayer perceptron; Orange = Logistic regression with l2 penalty; Green = Support vector machine; Red = Random forest. (A) Depleted plasma in which the optimal protein set included 30 proteins and the best classifier was the random forest (AUC 0.97); (B) SP003 in which the optimal protein set included 32 proteins and the best classifier model was the support vector machine (AUC 0.93); (C) SP006 in which the optimal protein set included 26 proteins and the best classifier was the random forest (AUC 0.94); (D) SP007 in which the optimal protein set included 14 proteins and the best classifier was the logistic regression with l2 penalty (AUC 0.95); (E) SP333 in which the optimal protein set included 36 proteins and the best classifier was the support vector machine (AUC 0.95); (F) SP339 in which the optimal protein set included 43 proteins and the best classifier was the random forest (AUC 0.96).