Supplement Tables

Table S1. Image- and patient-level evaluation metrics of random forest (RF).

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.79	0.76	0.85	0.85	0.85
Precision	0.81	0.76	0.95	0.95	0.91
Recall	0.58	0.54	0.73	0.73	0.77
F1-score	0.68	0.64	0.83	0.83	0.83
AUROC	0.85	0.80	0.94	0.89	0.91
p-value	ref	<0.01	ref	0.01	0.6

F2048 represents a 2048-dimension feature vector from the feature extractor and BM5 represents five key biomarkers. The p-value denotes the result of the DeLong test, which compares the performance differences between different feature sets. AUROC, area under the receiver-operating characteristic curve.

Table S2. Image- and patient-level evaluation metrics of linear discriminant analysis (LDA).

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.67	0.63	0.65	0.62	0.87
Precision	0.58	0.53	0.70	0.65	0.88
Recall	0.54	0.48	0.54	0.50	0.85

Chang T-W, *et al. BMJ Health Care Inform* 2025; 32:e101192. doi: 10.1136/bmjhci-2024-101192

F1-score	0.56	0.50	0.61	0.57	0.86
AUROC	0.69	0.61	0.67	0.65	0.93
p-value	ref	<0.01	ref	0.60	<0.01

Table S3. Image- and patient-level evaluation metrics of support vector machine (SVM).

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.90	0.78	0.85	0.77	0.87
Precision	0.96	0.75	0.95	0.77	0.91
Recall	0.77	0.66	0.73	0.77	0.81
F1-score	0.85	0.70	0.83	0.77	0.86
AUROC	0.97	0.82	0.92	0.89	0.91
p-value	ref	<0.01	ref	0.59	0.68

Table S4. Image- and patient-level evaluation metrics of LightGBM.

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.89	0.77	0.85	0.77	0.90
Precision	0.93	0.73	0.85	0.77	1
Recall	0.77	0.63	0.85	0.77	0.81
F1-score	0.84	0.68	0.85	0.77	0.89

Chang T-W, *et al. BMJ Health Care Inform* 2025; 32:e101192. doi: 10.1136/bmjhci-2024-101192

AUROC	0.93	0.81	0.91	0.85	0.95
p-value	ref	<0.01	ref	0.19	0.16

Table S5. Image- and patient-level evaluation metrics of adaptive boosting (AdaBoost).

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.82	0.77	0.79	0.77	0.83
Precision	0.82	0.73	0.80	0.77	0.90
Recall	0.68	0.64	0.77	0.77	0.73
F1-score	0.75	0.69	0.78	0.77	0.81
AUROC	0.87	0.81	0.85	0.86	0.90
p-value	ref	<0.01	ref	0.89	0.15

Table S6. Image- and patient-level evaluation metrics of gradient boosting.

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.88	0.76	0.90	0.77	0.85
Precision	0.96	0.71	0.96	0.77	0.88
Recall	0.71	0.64	0.85	0.77	0.81
F1-score	0.82	0.67	0.90	0.77	0.84
AUROC	0.95	0.80	0.94	0.84	0.92

p-value	ref	<0.01	ref	0.06	0.43

Table S7. Image- and patient-level evaluation metrics of CatBoost.

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.88	0.73	0.87	0.75	0.85
Precision	0.93	0.66	0.88	0.76	0.91
Recall	0.76	0.62	0.85	0.73	0.77
F1-score	0.83	0.64	0.86	0.75	0.83
AUROC	0.94	0.80	0.93	0.85	0.92
p-value	ref	<0.01	ref	0.08	0.75

Table S8. Image- and patient-level evaluation metrics of stacking 6 models.

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.88	0.77	0.87	0.77	0.88
Precision	0.93	0.73	0.88	0.77	0.92
Recall	0.75	0.65	0.85	0.77	0.85
F1-score	0.83	0.69	0.86	0.77	0.88
AUROC	0.88	0.82	0.93	0.85	0.95
p-value	ref	0.01	ref	0.14	0.33

This method uses a stacking ensemble of 6 models: XGBoost, RF, SVM, logistic regression, LightGBM, and gradient boosting. The final estimator is trained using

Chang T-W, *et al. BMJ Health Care Inform* 2025; 32:e101192. doi: 10.1136/bmjhci-2024-101192

logistic regression. The hyperparameters for each model are the same as those used in the individual model training and may vary based on grid search results for different feature sets (BM5, F2048, BM+F2048).

Table S9. Image- and patient-level evaluation metrics of stacking 4 models.

	Image level		Patient level		
	F2048+BM5	F2048	F2048+BM5	F2048	BM5
Accuracy	0.88	0.77	0.88	0.77	0.88
Precision	0.93	0.74	0.92	0.77	0.92
Recall	0.75	0.64	0.85	0.77	0.85
F1-score	0.83	0.68	0.88	0.77	0.88
AUROC	0.92	0.82	0.94	0.86	0.95
p-value	ref	<0.01	ref	0.07	0.92

This method uses a stacking ensemble of 4 models: XGBoost, logistic regression, gradient boosting, and Catboost. The final estimator is trained using Logistic Regression. The hyperparameters for each model are the same as those used in the individual model training and may vary based on grid search results for different feature sets (BM, F2048, BM+F2048).