Supplementary Table 3. Comparison of model performance between setting 5 and others (settings 1–4)

Machine learning mode	Setting for comparison	Mean test set AUC	*P*-value (Wilcoxon rank-sum test)
Logistic regression	Setting 1	0.654	0.0006
	Setting 2	0.635	0.0001
	Setting 3	0.636	0.0001
	Setting 4	0.680	0.0128
	Setting 5	0.750	1
Random forest	Setting 1	0.603	<0.0001
	Setting 2	0.615	<0.0001
	Setting 3	0.670	0.0008
	Setting 4	0.761	0.4009
	Setting 5	0.748	1
Support vector machine	Setting 1	0.648	0.0003
	Setting 2	0.646	<0.0001
	Setting 3	0.656	0.0014
	Setting 4	0.724	0.2426
	Setting 5	0.757	1
Deep neural network	Setting 1	0.630	0.0010
	Setting 2	0.658	0.0025
	Setting 3	0.576	<0.0001
	Setting 4	0.681	0.1513
	Setting 5	0.709	1

For each model, *P*-value was calculated using the Wilcoxon rank-sum test to compare the predictive power between setting 5 and each of the other settings.