

ORIGINAL RESEARCH ARTICLE

Development of a machine learning-based risk prediction model for lymphoma using data from the 2023 National Health Interview Survey

Supplementary Files

Table S1. Parameter combinations and results of 10 random searches for the least absolute shrinkage and selection operator algorithm

No.	Alpha	Lambda	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.02	0.01	0.69	1.00	0.00	0.03	0.00	0.00
2	0.47	2.20	0.50	1.00	0.00	0.00	0.00	0.00
3	0.54	3.41	0.50	1.00	0.00	0.00	0.00	0.00
4	0.58	0.00	0.69	1.00	0.00	0.04	0.00	0.00
5	0.71	0.35	0.50	1.00	0.00	0.00	0.00	0.00
6	0.85	2.65	0.50	1.00	0.00	0.00	0.00	0.00
7	0.89	0.14	0.50	1.00	0.00	0.00	0.00	0.00
8	0.93	2.48	0.50	1.00	0.00	0.00	0.00	0.00
9	0.96	1.79	0.50	1.00	0.00	0.00	0.00	0.00
10	0.97	0.00	0.69	1.00	0.00	0.04	0.00	0.00

Abbreviations: ROC: Receiver operating characteristics; SD: Standard deviation; Sens: Sensitivity; Spec: Specificity.

Table S2. Parameter combinations of the extreme gradient boosting algorithm

No.	Eta	Max_ depth	Gamma	Colsample_ bytree	Min_child_ weight	Subsample	Nrounds	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.01	8.00	0.00	0.60	1.00	0.80	45.00	0.81	1.00	0.00	0.04	0.00	0.00
2	0.01	8.00	0.00	0.60	1.00	0.80	46.00	0.81	1.00	0.00	0.04	0.00	0.00
3	0.01	8.00	0.00	0.60	1.00	0.80	47.00	0.81	1.00	0.00	0.04	0.00	0.00
4	0.01	8.00	0.00	0.60	1.00	0.80	48.00	0.81	1.00	0.00	0.04	0.00	0.00
5	0.01	8.00	0.00	0.60	1.00	0.80	49.00	0.81	1.00	0.00	0.04	0.00	0.00
6	0.01	8.00	0.00	0.60	1.00	0.80	50.00	0.81	1.00	0.00	0.03	0.00	0.00
7	0.01	8.00	0.00	0.60	1.00	0.80	51.00	0.82	1.00	0.00	0.03	0.00	0.00
8	0.01	8.00	0.00	0.60	1.00	0.80	52.00	0.82	1.00	0.00	0.03	0.00	0.00
9	0.01	8.00	0.00	0.60	1.00	0.80	53.00	0.82	1.00	0.00	0.03	0.00	0.00
10	0.01	8.00	0.00	0.60	1.00	0.80	54.00	0.82	1.00	0.00	0.04	0.00	0.00
11	0.01	8.00	0.00	0.60	1.00	0.80	55.00	0.82	1.00	0.00	0.04	0.00	0.00
12	0.01	8.00	0.00	0.60	1.00	1.00	45.00	0.82	1.00	0.00	0.03	0.00	0.00
13	0.01	8.00	0.00	0.60	1.00	1.00	46.00	0.82	1.00	0.00	0.03	0.00	0.00
14	0.01	8.00	0.00	0.60	1.00	1.00	47.00	0.82	1.00	0.00	0.03	0.00	0.00
15	0.01	8.00	0.00	0.60	1.00	1.00	48.00	0.82	1.00	0.00	0.03	0.00	0.00
16	0.01	8.00	0.00	0.60	1.00	1.00	49.00	0.82	1.00	0.00	0.03	0.00	0.00

(Cont'd...)

Table S2. (Continued)

No.	Eta	Max_ depth	Gamma	Colsample_ bytree	Min_child_ weight	Subsample	Nrounds	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
17	0.01	8.00	0.00	0.60	1.00	1.00	50.00	0.82	1.00	0.00	0.03	0.00	0.00
18	0.01	8.00	0.00	0.60	1.00	1.00	51.00	0.82	1.00	0.00	0.03	0.00	0.00
19	0.01	8.00	0.00	0.60	1.00	1.00	52.00	0.82	1.00	0.00	0.03	0.00	0.00
20	0.01	8.00	0.00	0.60	1.00	1.00	53.00	0.82	1.00	0.00	0.03	0.00	0.00
21	0.01	8.00	0.00	0.60	1.00	1.00	54.00	0.82	1.00	0.00	0.03	0.00	0.00
22	0.01	8.00	0.00	0.60	1.00	1.00	55.00	0.82	1.00	0.00	0.03	0.00	0.00
23	0.01	8.00	0.00	0.80	1.00	0.80	45.00	0.83	1.00	0.00	0.03	0.00	0.00
24	0.01	8.00	0.00	0.80	1.00	0.80	46.00	0.83	1.00	0.00	0.03	0.00	0.00
25	0.01	8.00	0.00	0.80	1.00	0.80	47.00	0.83	1.00	0.00	0.03	0.00	0.00
26	0.01	8.00	0.00	0.80	1.00	0.80	48.00	0.83	1.00	0.00	0.03	0.00	0.00
27	0.01	8.00	0.00	0.80	1.00	0.80	49.00	0.83	1.00	0.00	0.03	0.00	0.00
28	0.01	8.00	0.00	0.80	1.00	0.80	50.00	0.83	1.00	0.00	0.03	0.00	0.00
29	0.01	8.00	0.00	0.80	1.00	0.80	51.00	0.83	1.00	0.00	0.03	0.00	0.00
30	0.01	8.00	0.00	0.80	1.00	0.80	52.00	0.83	1.00	0.00	0.03	0.00	0.00
31	0.01	8.00	0.00	0.80	1.00	0.80	53.00	0.83	1.00	0.00	0.03	0.00	0.00
32	0.01	8.00	0.00	0.80	1.00	0.80	54.00	0.83	1.00	0.00	0.03	0.00	0.00
33	0.01	8.00	0.00	0.80	1.00	0.80	55.00	0.83	1.00	0.00	0.03	0.00	0.00
34	0.01	8.00	0.00	0.80	1.00	1.00	45.00	0.83	1.00	0.00	0.03	0.00	0.00
35	0.01	8.00	0.00	0.80	1.00	1.00	46.00	0.83	1.00	0.00	0.03	0.00	0.00
36	0.01	8.00	0.00	0.80	1.00	1.00	47.00	0.83	1.00	0.00	0.03	0.00	0.00
37	0.01	8.00	0.00	0.80	1.00	1.00	48.00	0.83	1.00	0.00	0.03	0.00	0.00
38	0.01	8.00	0.00	0.80	1.00	1.00	49.00	0.83	1.00	0.00	0.03	0.00	0.00
39	0.01	8.00	0.00	0.80	1.00	1.00	50.00	0.83	1.00	0.00	0.03	0.00	0.00
40	0.01	8.00	0.00	0.80	1.00	1.00	51.00	0.83	1.00	0.00	0.03	0.00	0.00
41	0.01	8.00	0.00	0.80	1.00	1.00	52.00	0.83	1.00	0.00	0.03	0.00	0.00
42	0.01	8.00	0.00	0.80	1.00	1.00	53.00	0.83	1.00	0.00	0.03	0.00	0.00
43	0.01	8.00	0.00	0.80	1.00	1.00	54.00	0.83	1.00	0.00	0.03	0.00	0.00
44	0.01	8.00	0.00	0.80	1.00	1.00	55.00	0.83	1.00	0.00	0.03	0.00	0.00
45	0.10	8.00	0.00	0.60	1.00	0.80	45.00	0.85	0.99	0.17	0.02	0.00	0.05
46	0.10	8.00	0.00	0.60	1.00	0.80	46.00	0.85	0.99	0.18	0.02	0.00	0.05
47	0.10	8.00	0.00	0.60	1.00	0.80	47.00	0.85	0.99	0.18	0.02	0.00	0.04
48	0.10	8.00	0.00	0.60	1.00	0.80	48.00	0.85	0.99	0.18	0.02	0.00	0.05
49	0.10	8.00	0.00	0.60	1.00	0.80	49.00	0.85	0.99	0.18	0.02	0.00	0.04
50	0.10	8.00	0.00	0.60	1.00	0.80	50.00	0.85	0.99	0.18	0.02	0.00	0.04
51	0.10	8.00	0.00	0.60	1.00	0.80	51.00	0.85	0.99	0.19	0.02	0.00	0.05
52	0.10	8.00	0.00	0.60	1.00	0.80	52.00	0.85	0.99	0.20	0.02	0.00	0.05
53	0.10	8.00	0.00	0.60	1.00	0.80	53.00	0.85	0.99	0.20	0.02	0.00	0.05
54	0.10	8.00	0.00	0.60	1.00	0.80	54.00	0.85	0.99	0.20	0.02	0.00	0.05
55	0.10	8.00	0.00	0.60	1.00	0.80	55.00	0.85	0.99	0.21	0.02	0.00	0.05
56	0.10	8.00	0.00	0.60	1.00	1.00	45.00	0.85	0.99	0.16	0.02	0.00	0.06
57	0.10	8.00	0.00	0.60	1.00	1.00	46.00	0.85	0.99	0.16	0.02	0.00	0.06
58	0.10	8.00	0.00	0.60	1.00	1.00	47.00	0.85	0.99	0.17	0.02	0.00	0.05

(Cont'd...)

Table S2. (Continued)

No.	Eta	Max_ depth	Gamma	Colsample_ bytree	Min_child_ weight	Subsample	Nrounds	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
59	0.10	8.00	0.00	0.60	1.00	1.00	48.00	0.85	0.99	0.18	0.02	0.00	0.05
60	0.10	8.00	0.00	0.60	1.00	1.00	49.00	0.85	0.99	0.18	0.02	0.00	0.05
61	0.10	8.00	0.00	0.60	1.00	1.00	50.00	0.85	0.99	0.19	0.02	0.00	0.04
62	0.10	8.00	0.00	0.60	1.00	1.00	51.00	0.85	0.99	0.19	0.02	0.01	0.04
63	0.10	8.00	0.00	0.60	1.00	1.00	52.00	0.85	0.99	0.20	0.02	0.01	0.05
64	0.10	8.00	0.00	0.60	1.00	1.00	53.00	0.85	0.99	0.21	0.02	0.01	0.05
65	0.10	8.00	0.00	0.60	1.00	1.00	54.00	0.85	0.99	0.21	0.02	0.01	0.05
66	0.10	8.00	0.00	0.60	1.00	1.00	55.00	0.85	0.99	0.22	0.02	0.01	0.06
67	0.10	8.00	0.00	0.80	1.00	0.80	45.00	0.85	0.99	0.24	0.02	0.00	0.03
68	0.10	8.00	0.00	0.80	1.00	0.80	46.00	0.86	0.99	0.24	0.02	0.00	0.03
69	0.10	8.00	0.00	0.80	1.00	0.80	47.00	0.86	0.99	0.25	0.02	0.00	0.04
70	0.10	8.00	0.00	0.80	1.00	0.80	48.00	0.86	0.99	0.24	0.02	0.00	0.04
71	0.10	8.00	0.00	0.80	1.00	0.80	49.00	0.86	0.99	0.25	0.02	0.00	0.04
72	0.10	8.00	0.00	0.80	1.00	0.80	50.00	0.86	0.99	0.24	0.02	0.00	0.04
73	0.10	8.00	0.00	0.80	1.00	0.80	51.00	0.86	0.99	0.25	0.02	0.00	0.04
74	0.10	8.00	0.00	0.80	1.00	0.80	52.00	0.86	0.99	0.25	0.02	0.01	0.04
75	0.10	8.00	0.00	0.80	1.00	0.80	53.00	0.86	0.99	0.25	0.02	0.00	0.04
76	0.10	8.00	0.00	0.80	1.00	0.80	54.00	0.86	0.98	0.25	0.02	0.00	0.04
77	0.10	8.00	0.00	0.80	1.00	0.80	55.00	0.86	0.98	0.26	0.02	0.00	0.04
78	0.10	8.00	0.00	0.80	1.00	1.00	45.00	0.86	0.99	0.26	0.02	0.00	0.05
79	0.10	8.00	0.00	0.80	1.00	1.00	46.00	0.86	0.99	0.27	0.02	0.00	0.05
80	0.10	8.00	0.00	0.80	1.00	1.00	47.00	0.86	0.99	0.27	0.02	0.00	0.05
81	0.10	8.00	0.00	0.80	1.00	1.00	48.00	0.86	0.99	0.27	0.02	0.00	0.05
82	0.10	8.00	0.00	0.80	1.00	1.00	49.00	0.86	0.99	0.27	0.02	0.00	0.05
83	0.10	8.00	0.00	0.80	1.00	1.00	50.00	0.86	0.99	0.27	0.02	0.00	0.05
84	0.10	8.00	0.00	0.80	1.00	1.00	51.00	0.86	0.99	0.27	0.02	0.00	0.06
85	0.10	8.00	0.00	0.80	1.00	1.00	52.00	0.86	0.99	0.27	0.02	0.00	0.06
86	0.10	8.00	0.00	0.80	1.00	1.00	53.00	0.86	0.99	0.26	0.02	0.00	0.05
87	0.10	8.00	0.00	0.80	1.00	1.00	54.00	0.86	0.99	0.27	0.02	0.00	0.05
88	0.10	8.00	0.00	0.80	1.00	1.00	55.00	0.86	0.99	0.26	0.02	0.00	0.06

Abbreviations: ROC: Receiver operating characteristics; SD: Standard deviation; Sens: Sensitivity; Spec: Specificity.

Table S3. Optimal parameter combinations of different kernel functions

No.	C	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.11	0.48	1.00	0.00	0.05	0.00	0.00
2	0.67	0.53	1.00	0.00	0.06	0.00	0.00
3	0.94	0.50	1.00	0.00	0.05	0.00	0.00
4	1.71	0.46	1.00	0.00	0.05	0.00	0.00
5	5.96	0.46	1.00	0.00	0.06	0.00	0.00
6	6.41	0.52	1.00	0.00	0.08	0.00	0.00
7	12.68	0.53	1.00	0.00	0.05	0.00	0.00
8	16.35	0.54	1.00	0.00	0.03	0.00	0.00
9	16.75	0.50	1.00	0.00	0.04	0.00	0.00
10	22.08	0.47	1.00	0.00	0.03	0.00	0.00

Abbreviations: ROC: Receiver operating characteristics; SD: Standard deviation; Sens: Sensitivity; Spec: Specificity.

Table S4. Parameter combinations and results of 10 random searches for the random forest algorithm

Mtry	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	0.60	1.00	0.00	0.01	0.00	0.00
2	0.63	1.00	0.00	0.02	0.00	0.00
3	0.64	1.00	0.00	0.03	0.00	0.00
4	0.65	1.00	0.02	0.02	0.00	0.02
5	0.65	0.99	0.04	0.02	0.00	0.03
6	0.66	0.99	0.08	0.02	0.00	0.02
7	0.68	0.99	0.13	0.03	0.00	0.02
8	0.67	0.99	0.15	0.02	0.00	0.02
9	0.68	0.99	0.17	0.02	0.01	0.01
10	0.68	0.98	0.19	0.03	0.01	0.02
11	0.69	0.98	0.19	0.02	0.01	0.02
12	0.68	0.98	0.22	0.03	0.01	0.03
13	0.68	0.98	0.21	0.02	0.01	0.02
14	0.69	0.98	0.22	0.02	0.01	0.03
15	0.69	0.98	0.22	0.03	0.01	0.03
16	0.69	0.98	0.21	0.02	0.01	0.03
17	0.68	0.98	0.21	0.03	0.01	0.03
18	0.68	0.98	0.21	0.03	0.01	0.03

Abbreviations: ROC: Receiver operating characteristics; SD: Standard deviation; Sens: Sensitivity; Spec: Specificity.

Table S5. Performance evaluation of the four machine learning models after hyperparameter tuning

Model	Sensitivity	Specificity	Precision	F1	Accuracy	AUC
RF	0.99	0.24	0.92	0.95	0.91	0.71
SVM	0.99	0.26	0.92	0.95	0.91	0.75
XGBoost	0.99	0.12	0.91	0.95	0.90	0.82
LASSO	1.00	0.00	0.90	0.95	0.90	0.70

Abbreviations: AUC: Area under the curve; LASSO: Least absolute shrinkage and selection operator; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

Table S6. Performance metrics and confidence intervals of four machine learning models

Model	Sensitivity	Sensitivity_ CI	Specificity	Specificity_ CI	Precision	PPV	PPV_CI	NPV	NPV_CI	F1	Accuracy	Accuracy_ CI	AUC
RF	0.99	[0.954, 0.997]	0.25	[0.237, 0.264]	0.92	0.25	[0.21, 0.294]	0.99	[0.985, 0.992]	0.95	0.91	[0.905, 0.922]	0.70
SVM	0.99	[0.954, 0.997]	0.25	[0.237, 0.264]	0.92	0.25	[0.21, 0.294]	0.99	[0.985, 0.992]	0.95	0.91	[0.905, 0.922]	0.76
XGBoost	0.99	[0.948, 0.999]	0.17	[0.157, 0.18]	0.91	0.17	[0.135, 0.208]	0.99	[0.99, 0.995]	0.95	0.91	[0.9, 0.918]	0.86
LASSO	1.00	[0, 1]	0.00	[0, 0.001]	0.90	0.00	[0, 0.007]	1.00	[0.999, 1]	0.95	0.90	[0.889, 0.907]	0.70

Abbreviations: AUC: Area under the curve; CI: Confidence interval; LASSO: Least absolute shrinkage and selection operator; NPV: Negative predictive value; PPV: Positive predictive value; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

Table S7. The ranking of the importance score of variables

Variable	Importance
Obese	100.00
Breast cancer–no	83.32
Healthy weight	77.02
Hypertension–no	69.13
Non-Hispanic White	67.24
45–64 years	65.74
≥65 years	65.16
Health status–fair	58.18
Non-Hispanic Black/African American	54.73
Health status–very good	54.29
Overweight	52.96
25–44 years	49.35
Health status–good	48.40
Health status–poor	46.29
Non-Hispanic Asian	15.65
Other single and multiple races	9.40
Non-Hispanic AIAN	6.23
Non-Hispanic AIAN any other group	5.84

Abbreviation: AIAN: American Indian or Alaska Native.