## Supplementary

**Table S1** Summary of machine learning algorithms

| Model | Summary |
|---|---|
| Elastic net regression | Elastic net regression is an extension of linear regression. It blends and tunes the strength of L1 (Lasso regression) and L2 (Ridge) norm penalties during training process. This combination allows elastic net regression to address the limitations of lasso regression and ridge regression (1) |
| Random forest | Random forest is an ensemble machine learning algorithm widely used for classification and regression tasks, which is based on the idea of the bagging. Multiple decision trees (usually 1000) are established using random subset of data and variables. The final decision is made by aggregating the results of each individual decision tree (2) |
| Support vector machine | Support vector machine is a robust machine learning algorithm used for tasks such as classification, regression, and outlier detection. Its primary objective is to identify an optimal hyperplane that maximizes the margin between all the data points (3) |
| XGBoost and CatBoost | XGBoost and CatBoost are ensemble model of decision trees based on the idea of boosting. It is involves constructing a series of decision tree models, each grown on the residuals of previous tree (4,5) |

XGBoost, extreme gradient boosting machine; CatBoost, categorical boosting.

## References

1. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67:301-20.
2. Breiman L. Random Forests. Mach Learn 2001;45:5-32.
3. Brereton RG, Lloyd GR. Support vector machines for classification and regression. Analyst 2010;135:230-67.
4. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv 2018. arXiv:1810.11363.
5. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York, NY, USA: Association for Computing Machinery; 2016:785-94.

**Table S2** Hyperparameters for the final models

| Model | Hyperparameter |
|---|---|
| OS | |
| Elastic Net Regression | alpha =0, lambda =0.008497534 |
| Random Forest | mtry =6, splitrule =extratrees, min.node.size =1 |
| Support Vector Machines | degree =2, scale =0.1, C =0.1 |
| XGBoost | nrounds =300, max_depth =7, eta =0.05, gamma =1, colsample_bytree =0.5, min_child_weight =1, subsample =0.8 |
| CatBoost | depth =8, learning_rate =0.01, interations =500, l2_leaf_reg =7, rsm =1, border_count =128 |
| OS (One-hot encoding) | |
| Elastic Net Regression | alpha =0.7, lambda =0.01707353 |
| Random Forest | mtry =7, splitrule =gini, min.node.size =1 |
| Support Vector Machines | degree =1, scale =1, C =1 |
| XGBoost | nrounds =300, max_depth =5, eta =0.05, gamma =1, colsample_bytree =0.5, min_child_weight =1, subsample =0.9 |
| CatBoost | depth =8, learning_rate =0.01, interations =500, l2_leaf_reg =2, rsm =1, border_count =128 |
| DSS | |
| Elastic Net Regression | alpha =0, lambda =0.008497534 |
| Random Forest | mtry =3, splitrule =gini, min.node.size =1 |
| Support Vector Machines | degree =2, scale =0.1, C =0.1 |
| XGBoost | nrounds =300, max_depth =3, eta =0.05, gamma =0.25, colsample_bytree =0.5, min_child_weight =1, subsample =0.9 |
| CatBoost | depth =8, learning_rate =0.01, interations =500, l2_leaf_reg =9, rsm =1, border_count =128 |
| DSS (One-hot encoding) | |
| Elastic Net Regression | alpha =0.1, lambda =0.01353048 |
| Random Forest | mtry =5, splitrule =gini, min.node.size =3 |
| Support Vector Machines | degree =2, scale =0.1, C =0.01 |
| XGBoost | nrounds =300, max_depth =3, eta =0.05, gamma =1, colsample_bytree =0.5, min_child_weight =1, subsample =0.9 |
| CatBoost | depth =8, learning_rate =0.01, interations =500, l2_leaf_reg =7, rsm =1, border_count =128 |

OS, overall survival; DSS, disease-specified survival; XGBoost, extreme gradient boosting machine; CatBoost, categorical boosting.

**Table S3** Demographic and clinical characters for patients in the disease specific survival dataset

| Variables | Overall (N=1,203) | Training set (n=842) | Validating set (n=361) | P |
|---|---|---|---|---|
| Age (years), median [IQR] | 60 [11, 88] | 60 [14, 88] | 61 [11, 86] | 0.24 |
| Sex | | | | 0.90 |
| Female | 598 (49.7) | 417 (49.5) | 181 (50.1) | |
| Male | 605 (50.3) | 425 (50.5) | 180 (49.9) | |
| History of other tumors | | | | 0.14 |
| Yes | 353 (29.3) | 236 (28.0) | 117 (32.4) | |
| No | 850 (70.7) | 606 (72.0) | 244 (67.6) | |
| Race | | | | 0.98 |
| White | 826 (68.7) | 579 (68.8) | 247 (68.4) | |
| Black | 158 (13.1) | 111 (13.2) | 47 (13.0) | |
| Other | 219 (18.2) | 152 (18.1) | 67 (18.6) | |
| Tumor size, mm | | | | 0.74 |
| ≤6 | 867 (72.1) | 604 (71.7) | 263 (72.9) | |
| >6 | 336 (27.9) | 238 (28.3) | 98 (27.1) | |
| Masaoka stage | | | | 0.29 |
| I-IIA | 473 (39.3) | 343 (40.7) | 130 (36.0) | |
| IIB | 569 (47.3) | 387 (46.0) | 182 (50.4) | |
| III-IV | 161 (13.4) | 112 (13.3) | 49 (13.6) | |
| Chemotherapy | | | | 0.66 |
| Yes | 295 (24.5) | 203 (24.1) | 92 (25.5) | |
| No | 908 (75.5) | 639 (75.9) | 269 (74.5) | |
| Radiotherapy | | | | 0.08 |
| Yes | 588 (48.9) | 397 (47.1) | 191 (52.9) | |
| No | 615 (51.1) | 445 (52.9) | 170 (47.1) | |
| Surgery type | | | | 0.67 |
| Radical/total resection | 714 (59.4) | 501 (59.5) | 213 (59.0) | |
| Local/partial excision | 455 (37.8) | 315 (37.4) | 140 (38.8) | |
| Debulking | 34 (2.8) | 26 (3.1) | 8 (2.2) | |
| WHO classification | | | | 0.08 |
| Type A | 84 (7.0) | 50 (5.9) | 34 (9.4) | |
| Type AB | 221 (18.4) | 167 (19.8) | 54 (15.0) | |
| Type B1 | 128 (10.6) | 88 (10.5) | 40 (11.1) | |
| Type B2 | 193 (16.0) | 145 (17.2) | 48 (13.3) | |
| Type B3 | 183 (15.2) | 127 (15.1) | 56 (15.5) | |
| Thymic carcinoma | 199 (16.5) | 133 (15.8) | 66 (18.3) | |
| NOS | 195 (16.2) | 132 (15.7) | 63 (17.5) | |
| Number of harvested lymph nodes | | | | 0.86 |
| ≤5 | 366 (30.4) | 260 (30.9) | 106 (29.4) | |
| >5 | 141 (11.7) | 99 (11.8) | 42 (11.6) | |
| No node dissection performed | 696 (57.9) | 483 (57.4) | 213 (59.0) | |
| Lymph node invasion | | | | 0.70 |
| Negative | 459 (38.2) | 327 (38.8) | 132 (36.6) | |
| Positive | 48 (4.0) | 32 (3.8) | 16 (4.4) | |
| No node dissection performed | 696 (57.9) | 483 (57.4) | 213 (59.0) | |
| Lung metastasis | | | | 0.50 |
| Yes | 45 (3.7) | 35 (4.2) | 10 (2.8) | |
| No | 804 (66.8) | 562 (66.7) | 242 (67.0) | |
| Unknown | 354 (29.4) | 245 (29.1) | 109 (30.2) | |

Data were presented as n (%) unless specified. IQR, interquartile range; WHO, World Health Organization; NOS, not otherwise specified.

**Table S4** Variables importance after one-hot encoding

| Variables | Contribution to the ROC curves (%) | | | | | |
|---|---|---|---|---|---|---|
| | OS | | | DSS | | |
| | ELR | RF | CatBoost | ELR | RF | CatBoost |
| Age | 100.00 | 157.45 | 100.00 | 72.25 | 52.8 | 100.00 |
| Sex | | | | | | |
| Female | 2.05 | 17.25 | 11.09 | 13.96 | 8.37 | 22.71 |
| Male | 2.05 | 17.36 | 14.27 | 13.77 | 8.20 | 20.10 |
| History of other tumors | | | | | | |
| Yes | 4.34 | 17.65 | 17.21 | 14.10 | 8.44 | 26.95 |
| No | 4.46 | 17.76 | 20.55 | 14.31 | 8.42 | 31.18 |
| Race | | | | | | |
| White | 3.11 | 16.22 | 10.10 | 6.86 | 7.38 | 13.41 |
| Black | 13.24 | 12.76 | 3.69 | 5.22 | 5.36 | 4.89 |
| Other | 16.14 | 11.79 | 5.92 | 0.00 | 5.55 | 5.17 |
| Tumor size, mm | | | | | | |
| ≤6 | 19.71 | 14.75 | 18.18 | 23.38 | 7.50 | 16.04 |
| >6 | 19.54 | 14.05 | 18.69 | 23.14 | 7.20 | 20.60 |
| Masaoka stage | | | | | | |
| I-IIA | 29.03 | 17.33 | 19.75 | 51.98 | 10.84 | 45.65 |
| IIB | 1.10 | 14.93 | 9.12 | 18.52 | 7.54 | 22.95 |
| III-IV | 39.47 | 25.32 | 22.14 | 27.10 | 11.55 | 18.25 |
| Chemotherapy | | | | | | |
| Yes | 26.76 | 18.82 | 16.77 | 41.09 | 16.72 | 36.22 |
| No | 26.97 | 19.15 | 22.56 | 40.60 | 16.87 | 48.39 |
| Radiotherapy | | | | | | |
| Yes | 17.71 | 17.75 | 23.43 | 23.81 | 9.10 | 27.36 |
| No | 17.86 | 18.00 | 26.00 | 23.57 | 9.23 | 31.66 |
| Surgery type | | | | | | |
| Radical/total resection | 11.26 | 20.77 | 24.93 | 14.28 | 9.10 | 27.32 |
| Local/partial excision | 5.28 | 17.77 | 13.17 | 12.84 | 8.70 | 26.37 |
| Debulking | 18.30 | 7.73 | 0.63 | 0.00 | 2.78 | 0.00 |
| WHO classification | | | | | | |
| Type A | 19.43 | 10.57 | 4.83 | 49.53 | 2.57 | 6.05 |
| Type AB | 14.17 | 16.04 | 11.55 | 87.88 | 8.06 | 37.54 |
| Type B1 | 4.07 | 10.51 | 0.55 | 23.54 | 6.08 | 11.90 |
| Type B2 | 6.97 | 17.56 | 14.88 | 14.63 | 7.22 | 15.47 |
| Type B3 | 9.73 | 12.57 | 2.49 | 0.00 | 6.74 | 8.18 |
| Thymic carcinoma | 47.88 | 31.93 | 28.41 | 100.00 | 43.63 | 86.37 |
| NOS | 14.17 | 13.01 | 7.36 | 20.75 | 5.07 | 5.36 |
| Number of harvested lymph nodes | | | | | | |
| ≤5 | 1.40 | 16.08 | 9.60 | 24.04 | 7.04 | 20.42 |
| >5 | 15.55 | 13.40 | 6.10 | 15.05 | 7.40 | 10.30 |
| No node dissection performed | 8.89 | 16.92 | 10.87 | 0.00 | 6.74 | 11.66 |
| Lymph node invasion | | | | | | |
| Negative | 15.55 | 16.56 | 13.78 | 0.00 | 7.57 | 11.86 |
| Positive | 28.87 | 16.45 | 5.45 | 35.84 | 14.21 | 16.74 |
| No node dissection performed | 3.61 | 15.82 | 10.87 | 0.00 | 7.27 | 13.76 |
| Lung metastasis | | | | | | |
| Yes | 13.49 | 7.66 | 0.00 | 29.54 | 9.48 | 6.36 |
| No | 5.57 | 16.68 | 5.87 | 24.57 | 9.10 | 26.12 |
| Unknown | 0.00 | 15.62 | 8.42 | 4.92 | 7.19 | 10.17 |

Ensemble models had the best performance in the one-hot encoding dataset. This table presents the variables importance of individual models that serve as the component of ensemble models. ROC, receiver operation characteristic; OS, overall survival; DSS, disease-specific survival; ELR, elastic net regularized logistic regression; RF, random forest; WHO, World Health Organization. CatBoost, categorical boosting; NOS, not otherwise specified.