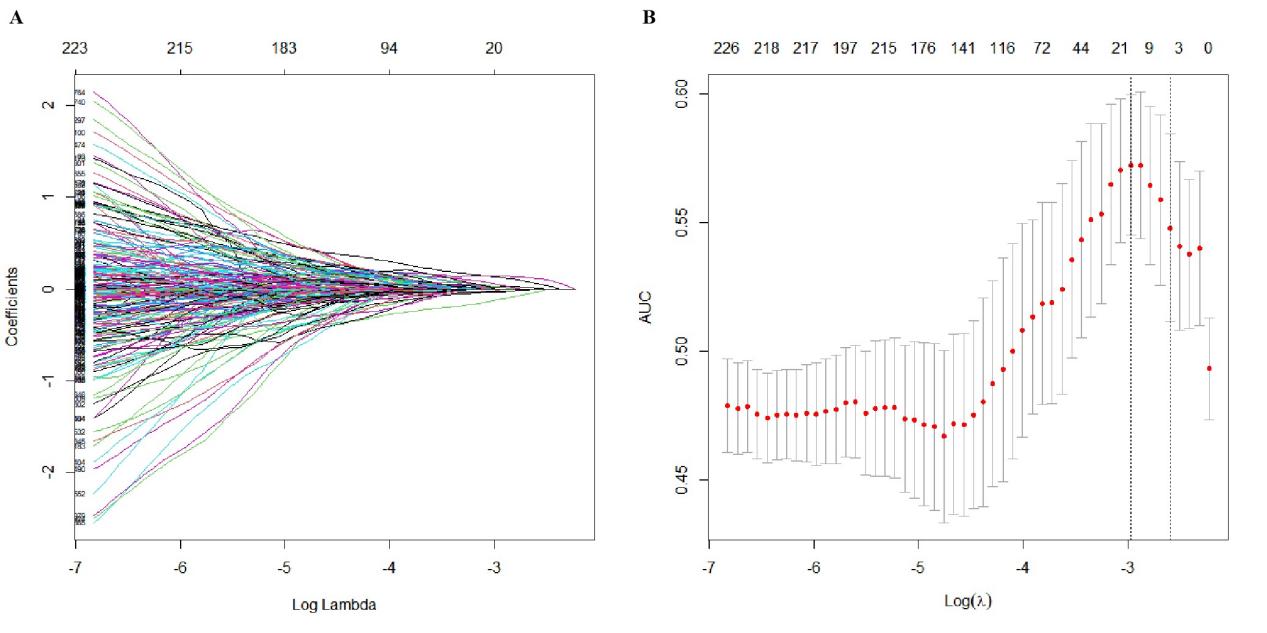


**S1.** Selection of LNM-associated features using RF in building models. (A) Error convergence curve according to the number of trees used in the RF model in training group. (B) Error convergence curve according to the number of variables in the RF model based on 5-fold cross-validation in training group. LNM, lymph node metastasis; RF, randomforest; OOB, out-of-bag; CV, cross-validation.

**S2.** The selected features of RF model based on Mean Decrease Accuracy value.

|  |  |
| --- | --- |
| Features | Mean Decrease Accuracy |
| logsigma70mm3D\_GLCM\_DifferenceVariance | 3.442338719 |
| logsigma100mm3D\_GLCM\_Idn | 3.294362441 |
| logsigma50mm3D\_GLRLM\_LongRunEmphasis | 2.907933971 |
| logsigma100mm3D\_GLCM\_Imc1 | 2.848503234 |
| logsigma50mm3D\_GLCM\_JointEntropy | 2.700642897 |
| logsigma50mm3D\_GLCM\_Idn | 2.680285674 |
| logsigma40mm3D\_GLDM\_DependenceNonUniformityNormalized | 2.649375099 |
| squareroot\_GLSZM\_SmallAreaLowGrayLevelEmphasis | 2.561607438 |
| logsigma30mm3D\_GLSZM\_SizeZoneNonUniformity | 2.515200473 |
| age | 2.469885086 |
| waveletLL\_GLSZM\_SmallAreaHighGrayLevelEmphasis | 2.415871739 |
| waveletLL\_GLCM\_SumEntropy | 2.403451695 |
| logsigma30mm3D\_GLSZM\_LargeAreaHighGrayLevelEmphasis | 2.388607208 |
| logsigma60mm3D\_GLRLM\_GrayLevelVariance | 2.385591031 |
| logsigma50mm3D\_GLRLM\_RunVariance | 2.385295934 |
| logsigma50mm3D\_GLCM\_SumEntropy | 2.381945955 |
| logsigma90mm3D\_GLCM\_Imc2 | 2.37975256 |
| waveletLH\_GLSZM\_SizeZoneNonUniformity | 2.351606132 |
| logarithm\_firstorder\_Median | 2.344617418 |
| logsigma60mm3D\_GLCM\_Idm | 2.340143257 |
| exponential\_firstorder\_Variance | 2.334027244 |
| waveletHH\_GLRLM\_LongRunLowGrayLevelEmphasis | 2.313033086 |
| square\_GLRLM\_ShortRunLowGrayLevelEmphasis | 2.310312281 |
| logsigma20mm3D\_GLRLM\_LongRunHighGrayLevelEmphasis | 2.300114726 |
| logsigma50mm3D\_GLRLM\_LongRunHighGrayLevelEmphasis | 2.258299248 |
| ANTITGAB | 2.257590595 |
| logsigma40mm3D\_GLRLM\_LongRunHighGrayLevelEmphasis | 2.236490899 |
| logsigma40mm3D\_GLDM\_DependenceVariance | 2.222274931 |
| squareroot\_firstorder\_Range | 2.210056901 |
| waveletHL\_firstorder\_Entropy | 2.20493006 |
| logsigma30mm3D\_GLSZM\_SmallAreaEmphasis | 2.166425562 |
| logsigma100mm3D\_GLDM\_SmallDependenceEmphasis | 2.161927299 |
| logsigma50mm3D\_GLRLM\_HighGrayLevelRunEmphasis | 2.149818249 |
| logsigma40mm3D\_GLSZM\_ZonePercentage | 2.146638329 |
| square\_GLDM\_DependenceNonUniformity | 2.115328511 |
| waveletLH\_GLDM\_DependenceNonUniformity | 2.113110565 |
| logsigma70mm3D\_GLCM\_Correlation | 2.085284956 |
| logsigma20mm3D\_firstorder\_Mean | 2.062195706 |
| logarithm\_GLCM\_Imc1 | 2.058154579 |
| logsigma70mm3D\_GLSZM\_LargeAreaEmphasis | 2.05195445 |
| waveletLL\_GLSZM\_GrayLevelVariance | 2.050627697 |
| logsigma40mm3D\_GLCM\_Imc1 | 2.027466223 |
| logsigma60mm3D\_GLCM\_MCC | 2.026707168 |
| waveletLL\_GLCM\_Correlation | 2.021495861 |
| originalshape2D\_Maximum\_Diameter | 2.015011187 |
| logsigma90mm3D\_firstorder\_InterquartileRange | 2.012767863 |
| original\_shape2D\_Elongation | 2.006399302 |
| squareroot\_GLCM\_SumEntropy | 1.99906682 |
| logsigma60mm3D\_firstorder\_Median | 1.985804489 |
| logsigma60mm3D\_firstorder\_10Percentile | 1.984271372 |
| logsigma80mm3D\_GLCM\_SumEntropy | 1.982301874 |
| logsigma20mm3D\_GLRLM\_RunEntropy | 1.980886173 |
| logarithm\_firstorder\_Mean | 1.964682756 |
| logsigma90mm3D\_GLSZM\_GrayLevelNonUniformity | 1.960297904 |
| logsigma40mm3D\_firstorder\_TotalEnergy | 1.957262258 |
| original\_firstorder\_TotalEnergy | 1.952041398 |
| original\_GLRLM\_RunLengthNonUniformityNormalized | 1.950707237 |
| squareroot\_GLSZM\_HighGrayLevelZoneEmphasis | 1.946934304 |
| waveletLH\_GLCM\_JointEnergy | 1.938152843 |
| waveletLH\_GLRLM\_RunVariance | 1.932361515 |
| waveletHH\_GLRLM\_LongRunEmphasis | 1.931604791 |
| waveletLL\_GLCM\_Imc2 | 1.923805337 |
| logsigma80mm3D\_firstorder\_RobustMeanAbsoluteDeviation | 1.919619001 |
| logsigma20mm3D\_firstorder\_10Percentile | 1.915172116 |
| logsigma60mm3D\_GLRLM\_RunPercentage | 1.911094116 |
| waveletHL\_firstorder\_Median | 1.899101544 |
| logsigma10mm3D\_GLSZM\_LargeAreaEmphasis | 1.898739837 |
| logsigma30mm3D\_GLDM\_DependenceNonUniformityNormalized | 1.894202084 |
| logsigma20mm3D\_firstorder\_TotalEnergy | 1.884635295 |
| squareroot\_GLRLM\_ShortRunLowGrayLevelEmphasis | 1.882161421 |
| logsigma70mm3D\_GLCM\_JointEntropy | 1.879656003 |
| exponential\_GLSZM\_LargeAreaEmphasis | 1.878771427 |
| logsigma60mm3D\_GLRLM\_LongRunLowGrayLevelEmphasis | 1.875265523 |
| logsigma10mm3D\_GLCM\_ClusterTendency | 1.87116524 |
| logsigma80mm3D\_firstorder\_Maximum | 1.865183301 |
| logsigma100mm3D\_GLCM\_Idmn | 1.864642734 |
| logsigma80mm3D\_GLRLM\_RunLengthNonUniformityNormalized | 1.852966822 |
| logsigma40mm3D\_GLRLM\_LongRunEmphasis | 1.835899894 |
| logarithm\_GLSZM\_GrayLevelNonUniformityNormalized | 1.834274717 |
| logsigma80mm3D\_GLCM\_JointEntropy | 1.833734197 |
| logsigma30mm3D\_GLCM\_InverseVariance | 1.829201456 |
| waveletLH\_GLRLM\_LongRunLowGrayLevelEmphasis | 1.825475796 |
| waveletLH\_firstorder\_Maximum | 1.82280215 |
| logsigma60mm3D\_firstorder\_90Percentile | 1.821766788 |
| logsigma50mm3D\_GLDM\_LargeDependenceEmphasis | 1.817627623 |
| squareroot\_GLCM\_SumAverage | 1.814586705 |
| exponential\_GLDM\_DependenceNonUniformity | 1.814045565 |
| original\_GLSZM\_LargeAreaHighGrayLevelEmphasis | 1.810612012 |
| waveletHL\_firstorder\_RobustMeanAbsoluteDeviation | 1.802476149 |
| logsigma30mm3D\_GLCM\_SumSquares | 1.800811751 |
| logsigma10mm3D\_GLSZM\_GrayLevelVariance | 1.798159556 |
| waveletLL\_GLSZM\_HighGrayLevelZoneEmphasis | 1.785759743 |
| gradient\_firstorder\_Energy | 1.782318805 |
| logsigma50mm3D\_GLCM\_DifferenceVariance | 1.770452141 |
| original\_GLSZM\_SmallAreaEmphasis | 1.761187711 |
| logarithm\_firstorder\_InterquartileRange | 1.760621506 |
| logsigma90mm3D\_GLRLM\_ShortRunHighGrayLevelEmphasis | 1.760161341 |
| waveletHL\_GLSZM\_SmallAreaHighGrayLevelEmphasis | 1.753846028 |
| original\_GLRLM\_GrayLevelNonUniformityNormalized | 1.749520871 |
| logsigma80mm3D\_GLCM\_InverseVariance | 1.744447008 |

Footnote: RF, randomforest; GLSZM, gray-level size-zone matrix; GLCM, gray-level co-occurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level runlength matrix; LL, low-pass/low-pass; LH, low-pass/high-pass; HL, high-pass/low-pass; HH, high-pass/high-pass; ANTITGAB, anti-thyroglobulin antibodies.



**S3.** Selection of LNM-associated features using LASSO regression in building models. (A) The coefficient profiles of 1804 features against the log Lambda (λ) when building models. (B) Tuning parameter (λ) selection in the LASSO logistic model based on 7-fold cross-validation. LNM, lymph node metastasis; LASSO, the least absolute shrinkage and selection operator; AUC, area under the curve.

**S4.** The coefficients of selected features of LASSO logistic model.

|  |  |
| --- | --- |
| Features | Coefficients |
| size | -0.3511604 |
| sex | 0.02778425 |
| composition | -0.1209118 |
| echogenicfoci | 0.1278081 |
| TT3 | 0.05224362 |
| Logarithm\_firstorder\_Kurtosis | 0.000599087 |
| Logarithm\_firstorder\_RobustMeanAbsoluteDeviation | 0.1411578 |
| Square\_GLRLM\_RunLengthNonUniformityNormalized | 0.01784644 |
| Exponential\_GLRLM\_RunLengthNonUniformityNormalized | 1.92759E-16 |
| Gradient\_GLRLM\_RunLengthNonUniformityNormalized | 2.40948E-17 |
| WaveletLL\_GLSZM\_GrayLevelVariance | -0.03316834 |
| WaveletLL\_GLSZM\_HighGrayLevelZoneEmphasis | 2.13783E-14 |
| WaveletLL\_GLSZM\_LargeAreaEmphasis | 0.02457482 |
| WaveletLL\_GLSZM\_SizeZoneNonUniformity | -1.20231E-14 |
| Square\_GLSZM\_SmallAreaEmphasis | 0.00189928 |
| Square\_GLSZM\_SmallAreaHighGrayLevelEmphasis | -0.01868737 |
| Square\_GLSZM\_SmallAreaLowGrayLevelEmphasis | -1.1082E-15 |
| Square\_GLSZM\_ZoneEntropy | -1.64821E-16 |
| Exponential\_GLSZM\_SmallAreaHighGrayLevelEmphasis | -5.03261E-18 |
| Exponential\_GLSZM\_SmallAreaLowGrayLevelEmphasis | -1.50978E-17 |
| Gradient\_GLSZM\_SmallAreaHighGrayLevelEmphasis | -5.03261E-18 |
| logsigma20mm3D\_GLSZM\_SmallAreaEmphasis | -0.02484 |
| logsigma90mm3D\_GLSZM\_SizeZoneNonUniformityNormalized | 0.000232344 |

Footnote: LASSO, the least absolute shrinkage and selection operator; GLSZM, gray-level size-zone matrix; GLRLM, gray-level runlength matrix; LL, low-pass/low-pass; TT3, total triiodothyronine 3.

RF train group 5折

**S5.** Error convergence curve according to the number of variables in the RF-RF model based on 5-fold cross-validation in training group. RF, randomforest; CV, cross-validation.