Predicting the risk category of thymoma with machine learning-based Computed Tomography radiomics signatures and their between-imaging phase differences

Zhu Liang [1#], Jiaming Li[3#], YihanTang[3], Yaxuan Zhang[3], Chunyuan Chen[1], Siyuan Li[4], Xuefeng Wang[1], Xinyan Xu[3], Ziye Zhuang[3], Shuyan He[2,5*], Biao Deng[1*]

1. Department of Cardiothoracic surgery, Affiliated Hospital of Guangdong Medical University, Xiashan District, ZhanJiang, Guangdong Province, China

2. Guangzhou Medical University, Panyu District, Guangzhou, Guangdong Province, China

3. Guangdong Medical Universiy, Xiashan District, ZhanJiang, Guangdong Province, China

4. Sun Yat-sen University, Yuexiu District, Guangzhou, Guangdong Province, China

5. Department of Radiology, Guangdong Women and Children Hospital, Guangzhou, China

[*]Corresponding author:

Shuyan He:1012027045@qq.com

Biao Deng: 15760562638@163.com

[#]These authors made an equal contribution to the work

Supplementary Data

Appendix A1: Mathematical Descriptions of Mann-Whitney U Test and LASSO

1. Mann-Whitney U Test

The Mann-Whitney U Test is a non-parametric test used to determine if there are significant differences between two independent samples. Suppose we have two sample groups X1,X2,...,Xn1 and Y1,Y2,...,Yn2Y_1, Y_2, ..., Yn2, where each sample represents imaging features.

The Mann-Whitney U test statistic U is calculated as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where:

- n1 and n2 are the sample sizes of the two groups.
- R1 is the sum of ranks for the first group.
- R2 is the sum of ranks for the second group.

The U statistic is the smaller of U1 and U2:

$$U = min(U_1, U_2)$$

The U statistic is then used to determine the significance of the differences between the groups, thereby selecting imaging features that show statistically significant differences.

2. LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO is a regression analysis method that enhances prediction accuracy and model interpretability through variable selection and regularization. We apply LASSO regression to the feature set filtered by the Mann-Whitney U Test, with the objective function:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathrm{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where:

- yi is the response variable (e.g., tumor risk category).
- xi is the imaging feature vector for the iii-th sample.
- β is the vector of regression coefficients.
- λ is the regularization parameter.
- $\sum_{j=1}^{p} |\beta_j|$is the L1-norm of the coefficients.

By applying LASSO regression, we can further refine the feature selection from the filtered imaging features, identifying the most predictive features while addressing multicollinearity among features. This results in a parsimonious and effective model for predicting thymoma risk.

Relationship between Mann-Whitney U Test and LASSO

Sequential Usage:

- **Mann-Whitney U Test:** Initially used to screen out features that show no significant

differences between groups, thereby reducing the dimensionality of the dataset.

- **LASSO:** Applied to the feature set filtered by the Mann-Whitney U Test to further refine feature selection by imposing penalties on coefficient magnitudes.

Complementary Roles:

- **Mann-Whitney U Test:** Focuses on the statistical significance of individual features.
- **LASSO:** Focuses on selecting a parsimonious model by considering the relationship between features and the response variable, while addressing multicollinearity among features.

Example Configuration: Here's how these methods can be implemented in practice:

1. **Mann-Whitney U Test:**
    - Compute the U statistic and corresponding p-value for each feature.
    - Select features with p-value < 0.05.
2. **LASSO:**
    - Standardize the selected features.
    - Perform LASSO regression with cross-validation to select the optimal λ.
    - Identify features with non-zero coefficients.

Appendix A2: The Benefits and Trade-offs of Ensemble Learning in Medical Image Analysis for Enhanced Thymoma Risk Prediction

Ensemble learning combines multiple machine learning models to improve overall performance, especially in complex tasks such as medical image analysis. Despite requiring longer computation times compared to a single model, ensemble learning is chosen for several key reasons:

1. **Improved Accuracy and Robustness**
    - **Combining Multiple Models**: Ensemble learning methods, such as bagging, boosting, and stacking, leverage the strengths of multiple models. Each model may capture different patterns or relationships within the data, thus enhancing prediction accuracy and robustness.
    - **Reducing Overfitting**: By averaging the predictions of multiple models, ensemble methods help mitigate the risk of overfitting, which is particularly crucial when dealing with small sample sizes and complex model architectures. This ensures better generalization to unseen data.
2. **Enhanced Generalization Ability**
    - **Diverse Perspectives**: Each individual model in the ensemble may focus on different aspects of the data. The combination of these diverse perspectives can lead to better generalization, as the ensemble can capture a broader range of patterns and anomalies.
    - **Error Reduction**: Ensemble methods typically outperform single models by reducing bias and variance. This is achieved through aggregating the predictions of models with different types of errors, thereby improving overall

predictive performance.

3. **Stability and Reliability**
    - **Mitigating Model Variability**: Single models can exhibit high variability, especially with limited data. Ensembles provide more stable and reliable predictions by alleviating the variability of individual models.
    - **Robustness to Noise**: Ensemble models are generally more robust to noise and outliers in the data. By leveraging multiple models, they can more effectively filter out noise and focus on the true underlying patterns.

Trade-Offs

While ensemble learning indeed requires longer training and prediction times compared to single models, the gains in accuracy, robustness, generalization ability, and reliability often outweigh the computational costs. In medical image analysis, accurate and reliable predictions are crucial, making this trade-off usually worthwhile.

Supporting Sources

1. **Improved Accuracy and Robustness**: Ensemble methods significantly enhance the performance of machine learning models by combining multiple predictions, capturing a wider range of patterns, and reducing the risk of overfitting.
2. **Enhanced Generalization Ability**: Ensemble learning methods are known for their superior ability to generalize to unseen data, achieved by utilizing the diverse perspectives of individual models.
3. **Stability and Reliability**: The stability and reliability of ensemble models make them suitable for tasks requiring high accuracy and robustness, such as medical image analysis.

By employing ensemble learning, the proposed method aims to achieve higher levels of performance and reliability, which are critical for predicting thymoma risk based on complex medical imaging data.

Appendix A3: Ensemble Learning Process and Parameter Settings

In the proposed approach, XGBoost is used to aggregate predictions from base models. The base models include Multilayer Perceptron (MLP), Random Forest (RF), and XGBoost. Below are detailed explanations regarding the number of base models, parameter settings for each base model, and the ensemble learning process:

Number of Base Models There are three base models:

1. Multilayer Perceptron (MLP)
2. Random Forest (RF)
3. XGBoost

Parameter Settings for Base Models

1. Multilayer Perceptron (MLP):

o Number of layers and neurons: For example, 3 layers with 64, 128, and 64 neurons respectively.

o Activation function: ReLU (Rectified Linear Unit).

o Optimizer: Adam.

o Learning rate: 0.001.

- Number of epochs: 100.
- Batch size: 32.

2. Random Forest (RF):
   - Number of trees (n_estimators): 100.
   - Maximum depth (max_depth): None (unlimited depth).
   - Minimum samples for split (min_samples_split): 2.
   - Minimum samples per leaf (min_samples_leaf): 1.
   - Feature selection criterion (max_features): sqrt (square root of features for each split).
3. XGBoost:
   - Number of trees (n_estimators): 100.
   - Learning rate: 0.1.
   - Maximum depth (max_depth): 6.
   - Subsample ratio of the training instance (subsample): 0.8.
   - Subsample ratio of columns when constructing each tree (colsample_bytree): 0.8.
   - Regularization parameter (lambda): 1.

The above parameters are initial settings, which may vary during parameter optimization for each model.

Supplementary Figures

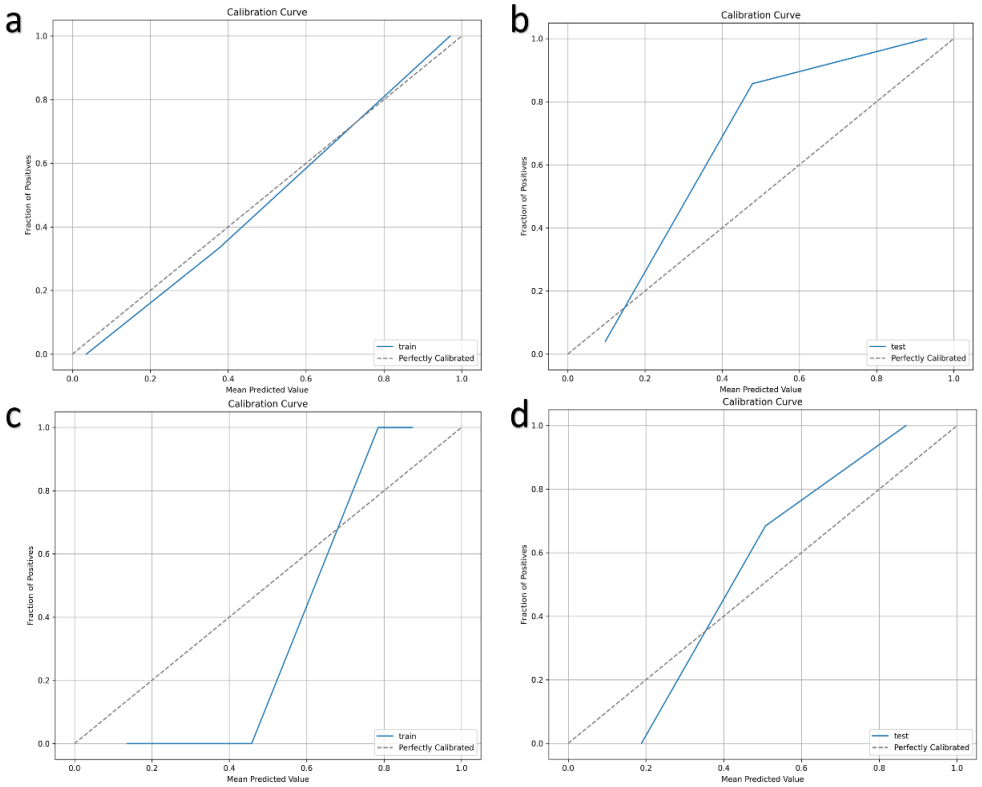

Figure S1: Pearson correlation heatmap for all the features and parameters. There was no linear correlation between the parameters and treatment response.

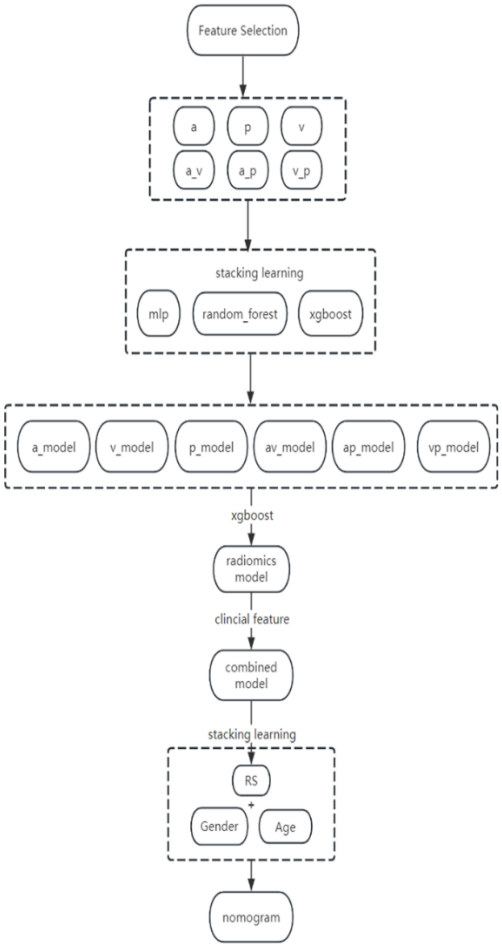

Figure S2: Process of establishing the eight models.

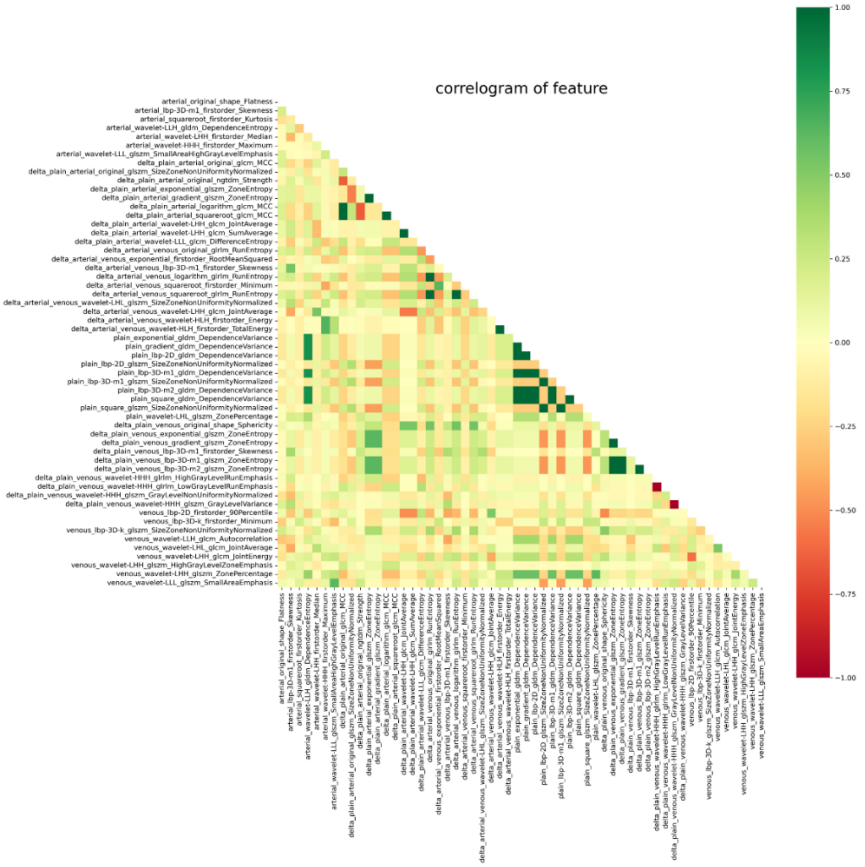

Figure S3: Model calibration curves. (a), (b): Combined radiomic model in the training (a) and test sets (b). (c), (d): Nomogram in the training (c) and test sets (d).