正文

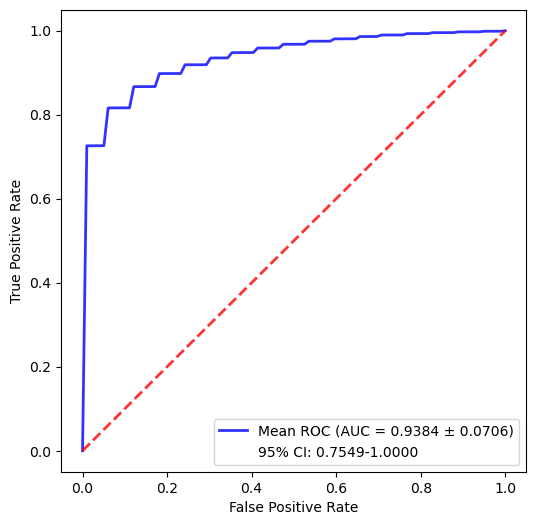
To identify diagnostic biomarkers for distinguishing patients with high and low overall survival time (OS) after chemotherapy, we performed feature selection on 750 lipids in tissue samples and utilized the selected features to construct four predictive models on the training dataset. Subsequently, we evaluated the independent performance of these models and the feature selection method using a testing dataset (Supplementary Figure). Given the complexity of clinical data and lipidomics characteristics, the XGBoost algorithm outperformed the other models and was chosen due to its proficiency in managing intricate feature interactions and its potential for high predictive accuracy. The eXtreme Gradient Boosting Classifier (XGBoost), using a biomarker panel of 9 lipids, enabled the discrimination between low-OS and high-OS groups, achieving an area under the receiver operating characteristic curve of 0.9384 (0.7549–1.0000). Collectively, these findings reveal significant lipid alterations between tissues of patients with different survival risks and demonstrate the potential of lipids for biomarker discovery, which may enhance the diagnosis of small cell lung cancer (SCLC).

Methods--Machine learning for the diagnosis of HCM and prediction of survival outcomes

To distinguish survival risks, we first identified statistical differences using the Mann-Whitney U-test to select a reduced number of diagnostic features for classification on the training dataset. Then five-fold cross-validation was conducted on the training set to obtain stable feature selection results. A P-value threshold of 0.05 was set and we selected features showing significant differences. For the prediction of survival outcomes, 76 participants who underwent chemotherapy were divided into high overall survival rate and low overall survival rate groups based on a commonly used threshold of 24 months (2 years). The patients were then randomly divided into training and testing sets. Features were more strictly selected based on their variable importance, as represented by Shapley values, with a mean Shapley value rank of 20 required as the threshold.

Once feature selection was completed, a prediction model was built using the eXtreme Gradient Boosting (XGBoost) method based on the training dataset. This diagnostic and prediction model was then applied to diagnose patients and predict survival outcomes in the testing set. The XGBoost model was configured with a max\_depth of 6 to balance model complexity and overfitting, a learning rate of 0.1 for gradual updates, and 1000 estimators for sufficient model capacity. XGBoost was performed using the xgboost package (v.2.0.3) and SHAP analysis was performed using the shap package (v.0.42.1) in Python (v.3.9.18).

Figure

1. Receiver operating characteristic (ROC) curve of the XGBoost model in the testing set.  
   
2. precision-recall curve (PRC) of the XGBoost model in the testing set
3. Schematic of the dataset creation and analysis strategy for the distinguish of Survival Rate

