**Experimental Report**

1. **Introduction**

Non-small cell lung cancer (NSCLC) is a common form of lung cancer with a high global incidence rate [1]. Despite the availability of various treatment options, including radiotherapy, targeted therapy, and immunotherapy, the overall survival rate remains low, and the prognosis for advanced-stage patients is poor [2]. Both clinical and omics data reveal the high heterogeneity of NSCLC, which makes it difficult for current treatment strategies to meet the demands for precision and personalized therapies. Therefore, there is an urgent need for in-depth analysis of NSCLC using multimodal data, including imaging and clinical data, to explore the relationship between its biological characteristics and clinical indicators.

Recent literature has reported the multifaceted role of radiomics in NSCLC. In predicting treatment outcomes, Gong et al. developed a deep learning model that successfully predicted the probability of brain metastasis in NSCLC patients within three years by integrating radiomics and clinical features [3]. Liu et al. developed a radiomics model based on enhanced CT images to predict the risk of bone metastasis in NSCLC, achieving an AUC of 0.808 in the validation set, indicating that radiomics features can reveal microscopic changes in tumors [4]. Yu et al. constructed a multimodal model combining PET and CT radiomics features to predict distant metastasis after early-stage NSCLC stereotactic body radiotherapy (SBRT), with an AUC of 0.835, outperforming single imaging modalities [5]. Secondly, in assisting NSCLC treatment decision-making, Meng et al. developed a model that combines patient CT images and transcriptomics data to predict NK cell infiltration and clinical prognosis, providing decision support for subsequent immunotherapy [6]. Wang et al. utilized CT radiomics to distinguish between radiation pneumonitis (RP) and immune checkpoint inhibitor-induced pneumonitis (CIP) in NSCLC patients, demonstrating the crucial role of radiomics in distinguishing treatment-related side effects and optimizing treatment strategies [7].

In this study, our contributions are multifaceted. We constructed a CT-based prognostic prediction model for NSCLC after determining the weights of individual models in the developed integrated model. Based on this, using the important features from the model, we further explored a method for patient subtyping through imaging features. This novel approach allows for the subdivision of patients into different biological subtypes based on tumor morphology, texture, and subtle differences in imaging that are difficult to detect directly.

1. **Material and Methods**

**2.1 Data Preprocess**

In this study, CT images and paired clinical data from 422 NSCLC patients were downloaded from The Cancer Imaging Archive (TCIA) database (<https://www.cancerimagingarchive.net/>). CT images and corresponding contour data were collected in DICOM format, containing anatomical and radiotherapy information for each patient. Python software was used to automatically convert DICOM files into NIfTI format for subsequent processing. Specifically, the ImageSeriesReader function from SimpleITK was used to read the DICOM series, which was then converted into 3D NIfTI images and stored in a standardized format (.nii.gz) for further analysis.  
 For each patient, the RTSTRUCT file was parsed to extract regions related to specific tumor areas (GTV). The pydicom library was employed to load the RTSTRUCT file and extract the contour data associated with the target labels. These contour data were then converted into binary masks and registered with the corresponding CT images to ensure proper overlay of the mask on the CT images. The mask generation process involved converting the physical coordinates of the contour data into image indices based on the CT image metadata (such as origin, pixel spacing, and orientation), ensuring accurate alignment of the mask with the CT image.  
 Radiomic features were extracted from the CT images and corresponding ROI masks using the PyRadiomics library. The feature extraction process was standardized with a gray-level width of 3 and a resampling pixel spacing of [1, 1, 1] as one set of features. Additionally, two other sets of features were generated using the "Wavelet" and "LoG" parameters. The RadiomicsFeatureExtractor class was used to extract the features, with all available feature sets enabled to ensure comprehensive feature extraction. The calculated features included shape, texture, and intensity statistics. A total of 321 radiomic features were retained for further analysis.  
 To ensure data accuracy, multiple quality control checks were incorporated at various stages of the pipeline. These checks included verifying the existence of the required DICOM files, confirming that the RTSTRUCT file contained the necessary contour data for the target labels, and ensuring that the generated masks were non-empty. After quality control, a total of 407 samples were successfully processed.

**2.2 C-index Prediction Model**

The dataset used in this study was preprocessed and divided into training and testing sets using 5-fold cross-validation. Random Survival Forest (RSF), HistGradientBoostingClassifier (HGB), ExtraTreesClassifier (ET), and GradientBoostingClassifier (GB) were implemented based on the scikit-learn package. Hyperparameters were optimized for each model, and each model was trained on the training set of each fold during cross-validation. A weighted averaging strategy was employed to integrate the predictions from multiple models. Grid search was performed to optimize the performance of the ensemble model by exploring all possible weight combinations. The performance of the models was evaluated using the C-index, which measures the discrimination ability of survival models. The best weight combination, which maximized the C-index, was selected as the final weight for the ensemble model. The optimal weight set was chosen from combinations where the sum of the weights was 1, and the combination with the highest C-index was selected.

Once the optimal weights were determined, the entire dataset was trained using the selected weights. The model was saved using the joblib library, and the integrated risk score was computed by averaging the predictions from the weighted individual models. Additionally, SHAP analysis was conducted to assess the feature importance of the final ensemble model. A custom SHAP interpreter for the ensemble model was created, where the prediction result was the weighted sum of the predictions from the individual models. The SHAP values for the ensemble model were calculated and visualized, with the most important features displayed in a bar chart. To assess the prognostic value of the ensemble model, patients were divided into high-risk and low-risk groups based on the risk scores predicted by the ensemble model, with the median risk score used as the classification threshold. Kaplan-Meier survival curves were generated for the two risk groups, and survival differences between the groups were compared using the Log-rank test. The p-value from the Log-rank test was reported to evaluate the statistical significance of the survival differences. Statistical significance was set at a p-value < 0.05 for evaluating survival differences between high-risk and low-risk groups.

**2.3 Subtype Identification Methods**

In this section, we briefly introduce related algorithms (NMF, DNMF) and the proposed algorithms.

**2.3.1 Non-negative Matrix Factorization Method**

The Non-negative Matrix Factorization (NMF) algorithm is a classical dimensionality reduction method. It can project high-dimensional data into a low-dimensional space. The objective function is given by:

s.t. (1)

where represents the radiomics feature matrix, with and denoting the number of samples and the number of features, respectively. and are the factor matrices, with being the reduced dimension. represents the Frobenius norm. and need to be non-negative.

**2.3.2 Semi-supervised Non-negative Matrix Factorization**

The Semi-supervised Non-negative Matrix Factorization (SNMF) algorithm enforces non-negativity constraints only on the factor matrix, with the objective function as follows:

(2)

**2.3.3 Deep Semi-supervised Non-negative Matrix Factorization**

The Deep Semi-supervised Non-negative Matrix Factorization (DSNMF) algorithm addresses the limitations of previous methods that can only perform shallow factorization. It introduces the factorization into layers and an additional factor. The objective function is given by:

(3)

**2.3.4 CB-DSNMF Algorithm**

While the Deep semi-NMF algorithm allows for deep decomposition, the decomposition process is still linear. Additionally, it does not incorporate information from the underlying network structure. Therefore, we propose the CB-DSNMF algorithm, which incorporates network connectivity information to reflect the interactions among features.

(4)

where represents the connectivity matrix, which reflects the connection strength between nodes and . Using the Laplacian matrix, we can reformulate it as:

(5)

where represents the trace of the matrix, and represents the Laplacian matrix.

(6)

Here, represents the degree matrix, with each element on the diagonal representing the degree of each node in . We further formulate the objective function of the CB-DSNMF algorithm as:

(7)

Here, represents the sigmoid activation function used for non-linear transformation. and are used to control the degree of f-norm constraint of the Laplacian constraint term and S, respectively. represents the junction latent matrix. The obtained after the final iteration of the algorithm is input into the K-means clustering algorithm, and the final output is the clustering result of the algorithm.

**2.3.5 Implementation of Comparison Algorithms**

In this study, the results obtained from the proposed algorithm were compared with those from Kmeans, NMF and consensus clustering based on three metrics (see Section 2.3.6). The Kmeans algorithm is implemented by the scikit-learn package. The consensus clustering algorithm was implemented using the "ConsensusClusterPlus" package in R. The analysis aimed to classify samples into a predefined number of clusters, with a maximum cluster number set to 4. The clustering algorithm used was PAM (Partitioning Around Medoids), and the distance between samples was calculated using Euclidean distance. The clustering process was repeated 50 times to ensure the stability of the results, with 80% of the samples (pItem = 0.8) and 100% of the features (pFeature = 1) randomly selected in each iteration. The NMF algorithm was performed using the nmf function from the NMF package, with the cluster rank (range 2 to 10) set, and the "brunet" method chosen for matrix factorization. This analysis was repeated 10 times to ensure the stability of the results. All of the above analyses were performed with the seed value to ensure reproducibility.

**2.3.6 Clustering performance evaluation method**

Silhouette Score is used to evaluate the compactness and separation of clusters. Its value ranges from -1 to 1. The larger the value, the better the clustering effect. For a given data point is the average distance from point to all other points within the same cluster (intra-cluster distance). is the average distance from point to all points in the nearest cluster (inter-cluster distance). The formula for calculating the Silhouette Score is as follows:

(8)

The Silhouette Score for the entire dataset is the average Silhouette Score of all points. Davies-Bouldin Score evaluates clustering quality by calculating the average similarity ratio of each cluster with its most similar cluster. A lower score indicates better clustering quality. For two clusters and , let be the average intra-cluster distance of cluster and be the average intra-cluster distance of cluster . is the distance between the centroids of clusters and . The formula for calculating the Davies-Bouldin Score is as follows:

(9)

For each cluster , find the maximum :

(10)

The final Davies-Bouldin Score is the average of all clusters:

(11)

where is the total number of clusters.

Calinski-Harabasz Score measures the compactness and separation of clusters. A higher score indicates a better clustering effect. Let be the total number of clusters, the total number of data points, the inter-cluster dispersion matrix, and the intra-cluster dispersion matrix. The formula for calculating the Calinski-Harabasz Score is as follows:

(12)

1. **Results**

**3.1 Results of Prognostic Model Construction**

In this study, we employed an ensemble model incorporating the RSF, HGB, ET, and GB classifiers to comprehensively analyze and identify key prognostic factors. Model performance was evaluated through cross-validation, using the C-index as the evaluation metric. Figure 1A presents the top features derived from SHAP analysis along with their corresponding weights, where SHAP values quantify each feature's contribution to risk prediction. SHAP analysis indicated that features such as "glcm\_InverseVariance" and "firstorder\_10Percentile" played significant roles in the model's predictions. Final model weight distribution is as follows: 'RSF: 0.0,' HGB: 0.1, 'ET' : 0.0, 'GB: 0.9.

The weights of individual models were optimized using a weighted averaging method to maximize the C-index, achieving a final optimal C-index value of 0.5682, as demonstrated by the results. Additionally, Figure 1B displays the Kaplan-Meier (KM) survival curves for the high-risk and low-risk groups based on risk values provided by the ensemble model. The results revealed a significant difference in survival rates between the high-risk and low-risk groups, supporting the prognostic value of the model. The log-rank test yielded a p-value of 6.97e-10, indicating that risk stratification based on our ensemble model effectively predicts patient survival outcomes.

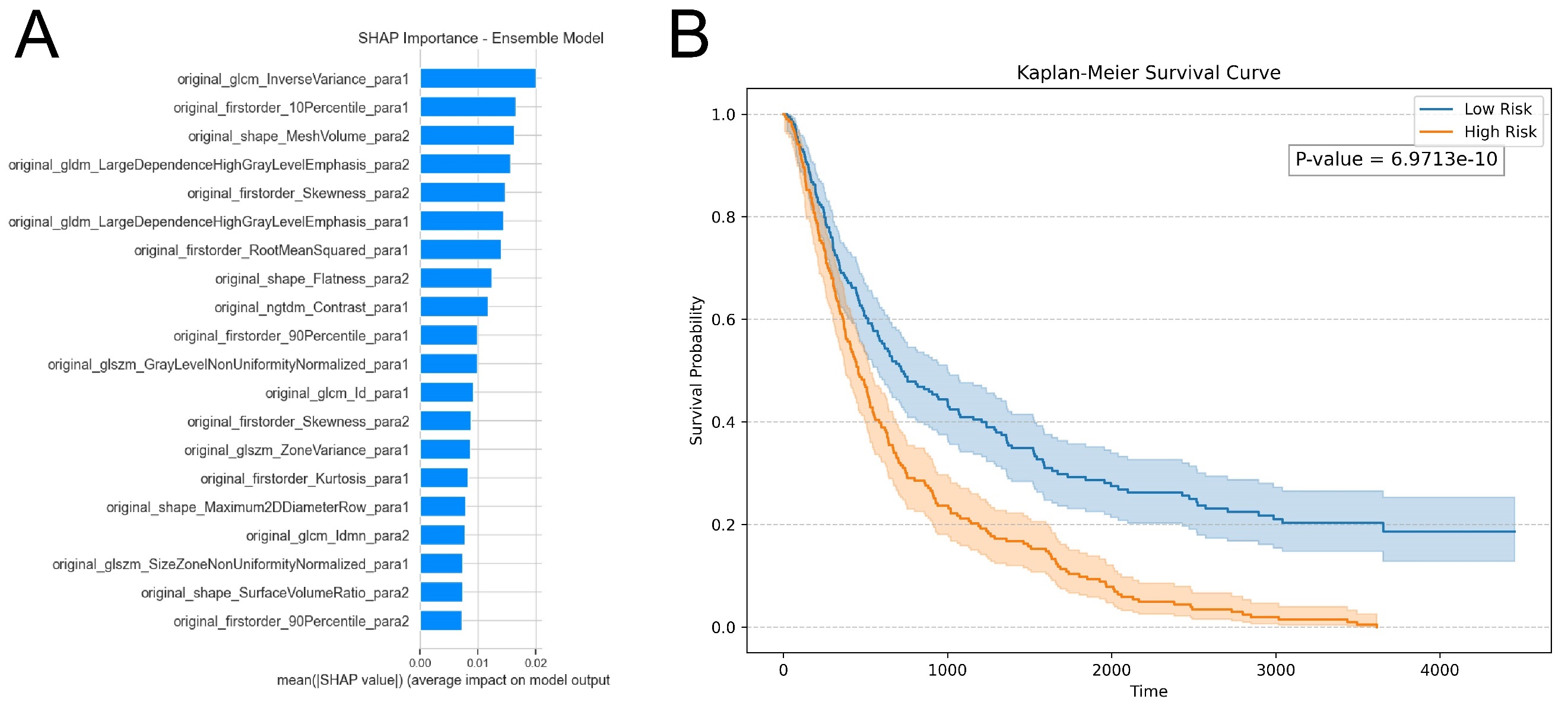


Figure 1 Prognostic model analysis results. (A) Top features and their corresponding weights obtained from the ensemble model using the SHAP package. (B) Kaplan-Meier survival curves for high- and low-risk groups based on risk scores provided by the ensemble model.

**3.2 Subtype Identification Results**

In this section, we propose a CB-DSNMF algorithm-based approach for the subtype identification of NSCLC and compare its clustering performance to other algorithms, including consensus clustering, NMF, and K-means. Clustering analyses were performed on NSCLC data, and results were visualized using two-dimensional t-SNE plots. The performance of different algorithms was compared across varying cluster numbers (k = 2 to 4). For the consensus clustering method, t-SNE visualization illustrated the spatial distribution of samples at different cluster numbers (k = 2, 3, 4) (Figure 2A-C). As the number of clusters increased, the compactness of sample clustering gradually decreased, and partial sample overlap was observed, indicating certain limitations of consensus clustering in identifying NSCLC subtypes. The clustering results based on the NMF method are shown in Figure 2D-F. At k = 2, the sample distribution was relatively compact; however, when k increased to 3 and 4, the t-SNE spatial distribution became more dispersed, suggesting that the clustering performance of NMF deteriorates with a larger number of clusters. Figure 2G-I displays the t-SNE distribution for the K-means method at k = 2, 3, and 4. K-means achieved good clustering performance at k = 2, but as the cluster number increased, the clustering boundaries became less distinct, and the sample separability declined.

The clustering results of the CB-DSNMF algorithm are presented in Figure 2J-L. Compared to other algorithms, CB-DSNMF demonstrated a clearer clustering distribution in the t-SNE visualization space. Notably, at k = 2 and 3, the sample boundaries were well-defined, and the clustering was more compact, highlighting the superior performance of this method in identifying NSCLC subtypes. To quantitatively compare the clustering performance of different algorithms, we utilized the silhouette score, Davies-Bouldin index, and Calinski-Harabasz index for evaluation (Figure 2M-O). The results showed that the CB-DSNMF method outperformed the other algorithms across all three clustering performance metrics, achieving the highest silhouette score, the lowest Davies-Bouldin index, and the highest Calinski-Harabasz index. These findings further validate the superior performance of the CB-DSNMF algorithm in NSCLC subtype identification.

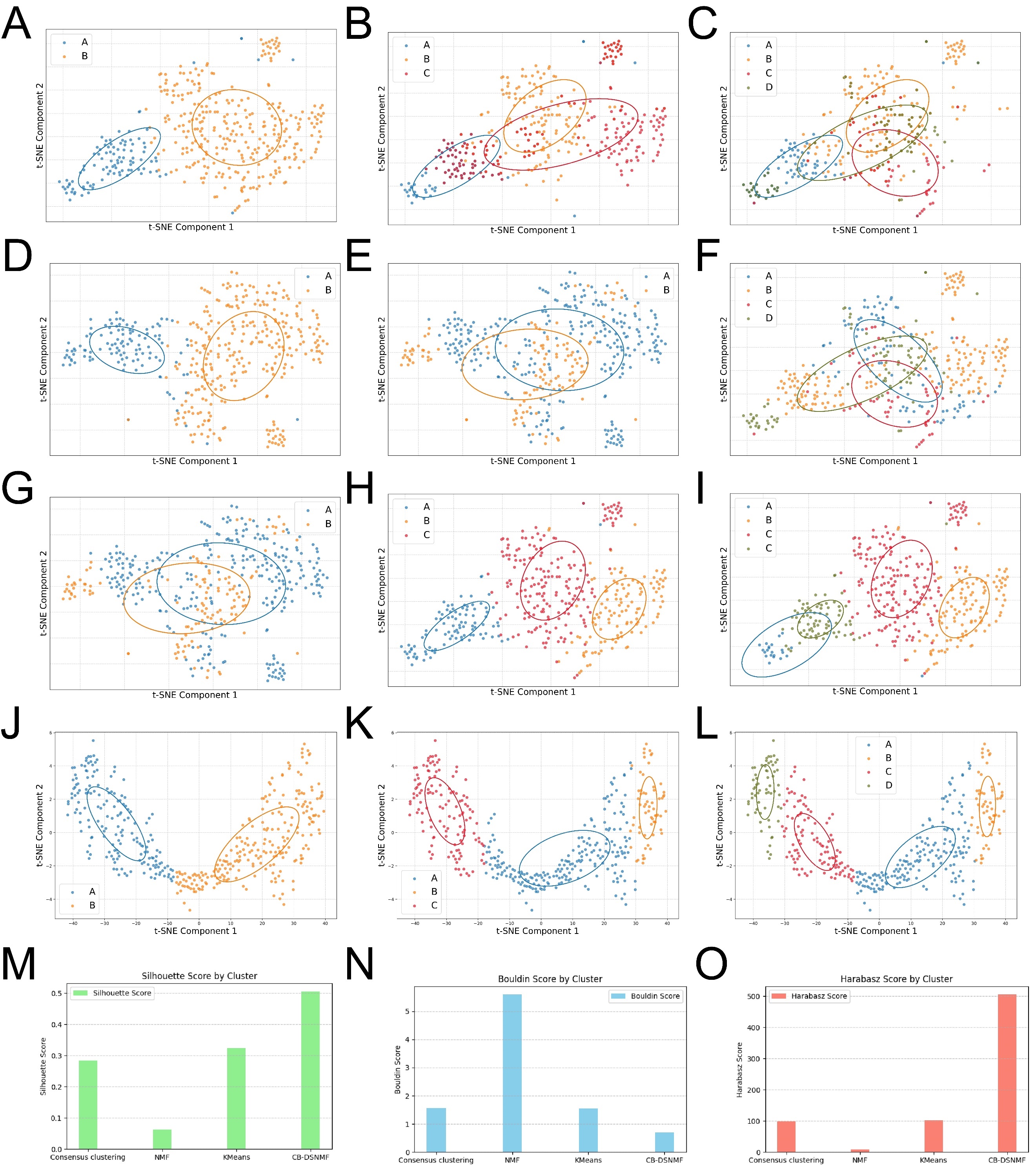


Figure 2 Results of NSCLC subtype identification based on the CB-DSNMF algorithm and comparative algorithms. (A-C) Distribution of clusters in the t-SNE two-dimensional space using the consensus clustering method for cluster numbers 2-4. (D-F) Distribution of clusters in the t-SNE two-dimensional space using the NMF method for cluster numbers 2-4. (G-I) Distribution of clusters in the t-SNE two-dimensional space using the K-means method for cluster numbers 2-4. (J-L) Distribution of clusters in the t-SNE two-dimensional space using the CB-DSNMF method for cluster numbers 2-4. (M-O) Clustering performance evaluation of different algorithms based on silhouette, Davies-Bouldin, and Calinski-Harabasz scores.

1. **Discussion**

In this study, we successfully developed a robust prognostic model for non-small cell lung cancer (NSCLC) using an ensemble approach that integrates multiple survival models, including Random Survival Forest (RSF), Histogram-Based Gradient Boosting (HGB), Extra Trees (ET), and Gradient Boosting (GB). The model demonstrated strong performance with an optimal C-index of 0.5682, indicating its potential clinical utility for risk prediction. SHAP analysis identified features such as *glcm\_InverseVariance* and *firstorder\_10Percentile* as key prognostic factors. These findings align with previous studies that have associated similar features with tumor metastasis and prognosis. For instance, Cheng et al. constructed a radiomics signature based on *firstorder\_10Percentile* and six additional radiomic features derived from CT imaging to distinguish between lymphoma and benign splenic lesions [8]. Similarly, in a radiomics prediction model for colorectal liver metastases, Vincenza Granata et al. identified *glcm\_InverseVariance* as the most significant predictor [9]. Kaplan-Meier survival curves further validated the model’s effectiveness by stratifying patients into high-risk and low-risk groups based on their risk scores.

In addition, we proposed a novel CB-DSNMF algorithm and applied it to cluster NSCLC data, comparing its performance against conventional clustering algorithms, including consensus clustering, Non-negative Matrix Factorization (NMF), and K-means. The results demonstrated that CB-DSNMF outperformed the traditional methods in terms of clustering accuracy and performance evaluation, further underscoring its potential in identifying NSCLC subtypes. While consensus clustering, NMF, and K-means displayed clustering capability, their performance declined as the number of clusters increased, likely due to their limitations in handling high-dimensional gene expression data, which is susceptible to noise and sparsity. In contrast, CB-DSNMF addressed these challenges through improved non-negative matrix factorization and the incorporation of robust clustering strategies, resulting in more accurate and reliable clustering outcomes.

In summary, the proposed prognostic model enhances robustness by integrating multiple classifiers, reducing the risk of overfitting, and improving generalizability across diverse patient populations. Future work will focus on external validation of the model in independent cohorts to confirm its predictive performance and clinical applicability. Despite the demonstrated advantages of CB-DSNMF in NSCLC subtype identification, several limitations remain. First, the sample size in this study is relatively small, necessitating validation in larger cohorts. Second, the practical application of the CB-DSNMF method to clinical datasets requires further optimization to address the complexities of diverse datasets. Moreover, future studies could integrate multi-omics data, such as DNA methylation and proteomics, to improve the accuracy and stability of subtype identification.

1. **Conclusion**

This study explores the potential of radiomics, a non-invasive approach, for evaluating the prognosis and identifying subtypes of NSCLC. This provides clinicians with a non-invasive tool to assist in treatment decision-making and patient management.

**Reference**

[1] I. Odintsov, L.M. Sholl, Prognostic and predictive biomarkers in non-small cell lung carcinoma, Pathology 56(2) (2024) 192-204.

[2] L.E.L. Hendriks, J. Remon, C. Faivre-Finn, M.C. Garassino, J.V. Heymach, K.M. Kerr, D.S.W. Tan, G. Veronesi, M. Reck, Non-small-cell lung cancer, Nature reviews. Disease primers 10(1) (2024) 71.

[3] J. Gong, T. Wang, Z. Wang, X. Chu, T. Hu, M. Li, W. Peng, F. Feng, T. Tong, Y. Gu, Enhancing brain metastasis prediction in non-small cell lung cancer: a deep learning-based segmentation and CT radiomics-based ensemble learning model, Cancer imaging : the official publication of the International Cancer Imaging Society 24(1) (2024) 1.

[4] Z. Liu, R. Yin, W. Ma, Z. Li, Y. Guo, H. Wu, Y. Lin, V.P. Chekhonin, K. Peltzer, H. Li, M. Mao, X. Jian, C. Zhang, Bone metastasis prediction in non-small-cell lung cancer: primary CT-based radiomics signature and clinical feature, BMC medical imaging 24(1) (2024) 203.

[5] L. Yu, Z. Zhang, H. Yi, J. Wang, J. Li, X. Wang, H. Bai, H. Ge, X. Zheng, J. Ni, H. Qi, Y. Guan, W. Xu, Z. Zhu, L. Xing, A. Dekker, L. Wee, A. Traverso, Z. Ye, Z. Yuan, A PET/CT radiomics model for predicting distant metastasis in early-stage non-small cell lung cancer patients treated with stereotactic body radiotherapy: a multicentric study, Radiation oncology (London, England) 19(1) (2024) 10.

[6] X. Meng, H. Xu, Y. Liang, M. Liang, W. Song, B. Zhou, J. Shi, M. Du, Y. Gao, Enhanced CT-based radiomics model to predict natural killer cell infiltration and clinical prognosis in non-small cell lung cancer, Frontiers in immunology 14 (2023) 1334886.

[7] M.D. Peiliang Wang, M.M. Yikun Li, M.M. Mengyu Zhao, M.D. Jinming Yu, M.D. Feifei Teng, Distinguishing immune checkpoint inhibitor-related pneumonitis from radiation pneumonitis by CT radiomics features in non-small cell lung cancer, International immunopharmacology 128 (2024) 111489.

[8] J.A. Cheng, Y.C. Lin, Y. Lin, R.C. Wu, H.Y. Lu, L.Y. Yang, H.J. Chiang, Y.H. Juan, Y.C. Lai, G. Lin, Machine Learning Radiomics Signature for Differentiating Lymphoma versus Benign Splenomegaly on CT, Diagnostics (Basel, Switzerland) 13(24) (2023).

[9] V. Granata, R. Fusco, S.V. Setola, M.C. Brunese, A. Di Mauro, A. Avallone, A. Ottaiano, N. Normanno, A. Petrillo, F. Izzo, Machine learning and radiomics analysis by computed tomography in colorectal liver metastases patients for RAS mutational status prediction, La Radiologia medica 129(7) (2024) 957-966.