

RESEARCH

Open Access



A multi-view prognostic model for diffuse large B-cell lymphoma based on kernel canonical correlation analysis and support vector machine

Yanhong Luo^{1,2*}, Yongao Li¹, Zhenhuan Yang¹, Yanbo Zhang^{1,2}, Hongmei Yu^{1,2}, Zhiqiang Zhao³, Kai Yu¹, Yujiao Guo¹, Xueman Wang¹, Na Yang¹, Yan Zhang¹, Tingting Zheng¹ and Jie Zhou^{4*}

Abstract

Background and objective Positron emission tomography/computed tomography (PET/CT) is recommended as the standard imaging modality for diffuse large B-cell lymphoma (DLBCL) staging. However, many studies have neglected the role of patients' prognostic factors with respect to imaging PET/CT of quantitative features. In this paper, a multi-view learning (MVL) model is established to make full use of both clinical and imaging data to predict the prognosis of DLBCL patients and thereby assist doctors in decision-making.

Methods Feature engineering, including feature extraction, feature screening by recursive feature elimination, and dimensionality reduction by principal component analysis, are successively performed on the clinical data and imaging data of the research subjects to obtain the study data. After dividing the data into training and test sets, an instance weighting method is applied to the training data. Subsequently, kernel mapping is performed on the imaging features and clinical features separately, and this kernel mapping is processed in the new kernel feature space using kernel canonical correlation analysis (KCCA). Lastly, model training is performed on the obtained common kernel subspace using a support vector machine (SVM). The final overall model, named SVM-2view-KCCA (SVM-2 K), was compared with three other multi-view models (Ensemble-SVM, Multi-view maximum entropy discrimination, and canonical correlation analysis). The performance of the model was evaluated on the test data with respect to several dichotomous metrics: accuracy, sensitivity, F1 score, the area under the curve (AUC), and G-mean.

Results The SVM model improved AUC by 10.5%, sensitivity by 11.9%, accuracy by 9.8%, F1 score by 9.2%, and G-mean by 7.8% for the DLBCL test data after feature engineering based on dimensionality reduction and instance weighting. In the performance comparison of single-view learning models, the SVM-based integration of clinical and imaging features achieved the best overall performance (AUC = 86.3%, accuracy = 91.6%, sensitivity = 83.2%, F1 = 85.7%, and G-mean = 86.1%). In the comparison of MVL models, SVM-2 K achieved the best overall performance (AUC = 92.1%, accuracy = 96.9%, sensitivity = 90.9%, F1 = 92.8%, and G-mean = 91.4%), and the performance of each MVL model was better than that of the best single-view learning model.

*Correspondence:

Yanhong Luo

sxmulyh@163.com

Jie Zhou

zhoujiezhzhoujie@163.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Conclusions MVL models outperformed single-view learning models. Of the MVL models, the proposed SVM-2 K achieved the best overall performance and could accurately predict patient prognosis.

Keywords Multi-view learning, Kernel canonical correlation analysis, Support vector machine, Diffuse large B-cell lymphoma, Disease prognosis

Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma, accounting for approximately 30%–40% of non-Hodgkin's lymphoma, and has become one of the types of malignant tumor whose incidence is increasing year by year [1]. First-line treatment regimens can lead to complete remission (CR) in 70% of DLBCL patients. Radiomics as a field plays a crucial role in extracting high-dimensional data from medical images, which allows for comprehensive assessment at the molecular level. PET radiomics, in particular, can provide valuable information for understanding tumor heterogeneity, metabolic activity, and treatment response. A recent review highlighted the challenges and promising opportunities of applying radiomics to PET imaging, emphasizing the need for robust and standardized quantitative methods for clinical applications [2]. The best tool for predicting early response and treatment efficacy after first-line induction therapy for DLBCL is positron emission tomography/computed tomography (PET/CT) examination, which has been recommended as the standard imaging modality for lymphoma staging according to international guidelines [3]. Quantitative and semiquantitative features extracted from PET/CT have been proved to have unique prognostic value by a large number of studies [4]. Therefore, it is of great importance to utilize the quantitative and semi-quantitative index data of PET/CT to construct a prediction model to accurately identify DLBCL patients in complete remission and give them timely and effective treatment.

There are currently three main problems with PET prognostic modeling studies for a wide range of disease types, including DLBCL. First, to date, numerous studies have emerged focusing on prognostic modeling of DLBCL based on PET/CT features. Some of these studies have constructed models using only single-modality PET/CT data, while others have combined PET/CT with clinical data to form multi-modality datasets. However, neither the independent use of features from these diverse sources nor their simple concatenation has fully leveraged the rich information embedded within the data. Data of the same object obtained by different means or from different perspectives is called multi-view data [5]. Multi-view data are characterized by their multi-source, multi-descriptive nature,

polymorphism, and high-dimensional isomorphism. The use of features of different nature from different sources, whether alone or in combination, does not make full use of their information content [6]. Second, because most DLBCL patients have a better prognosis and only a few have a poorer prognosis, there is a category imbalance in the DLBCL patient data. Traditional learning algorithms have substantial bias, which is manifested in the high rate of misrecognition by the classifier for the minority category: the minority samples tend to be misrecognized as the majority, which achieves a higher accuracy but also reduces the sensitivity, and makes the model's performance on the test set much lower [7]. Third, because the high cost of PET/CT increases the difficulty of data collection, existing PET/CT-related prognostic modeling studies for DLBCL patients have small sample sizes, with small numbers (hundreds, dozens, or even fewer) of training instances. Therefore, existing machine learning and deep learning models are generally unable to achieve good performance: models trained with small samples can easily lead to overfitting to small samples and underfitting to the target task [8].

Multi-view learning (MVL) [9] is a solution to the first problem. This type of learning both analyzes the correlation within the same view of the data and finds the differences between the different views, to mine the hidden effective information in the multi-view data and increase the accuracy of the classification results. In addition, MVL can also reduce the feature space dimensions when distinguishing feature views, which avoids the curse of dimensionality to some extent and improves model robustness [5]. Traditional medical research has predominantly relied on single-view data. However, single-view data often suffer from limitations such as insufficient data volume and incomplete information, which can lead to less accurate diagnostic outcomes. To overcome these limitations, multi-view learning has increasingly gained attention in recent years as a promising approach in the medical field. Multi-view learning is capable of simultaneously processing and integrating medical data from different sources, such as clinical and imaging data of DLBCL patients. This approach leverages the complementarity between different views, integrating multiple information sources to overcome the limitations of single-view

data, thereby enabling more robust data analysis and prediction.

To solve the second problem, instances at different locations in the hypothesis space of the classification task have different levels of importance and should be differentiated appropriately [10]. When the data distribution is unbalanced, because a few instances are precious and rare, these few instances should be given higher weights. Therefore, instances of these different types should be given different weights in the classification task.

Finally, for small-sample problems, the current methods for handling small-sample data can be divided into three main categories [11]: increasing the amount of training data, optimizing the process of searching for the optimal model, and reducing the space to be searched by the model. Increasing the amount of training data is insufficient to effectively improve the generalization ability of the model, and it is usually difficult to generalize across multiple datasets. The method of optimizing the process of searching for the optimal model is represented by migration learning [12]; however, this is still unable to solve the small-sample problem well for some domains that have relatively small amounts of data. Reducing the search space needed by the model is a common method to deal with small samples in the medical field; this is represented by principal component analysis (PCA). In embedding learning, to reduce the dimensionality of the feature space, the samples are projected onto a lower-dimensional space in which it is easier to distinguish between the different data categories.

In summary, this study aimed to solve the above three problems and attempted to construct a MVL model that is applicable to small samples with class imbalance, to make full use of both clinical and imaging data to predict the prognosis of DLBCL patients and assist doctors in decision-making.

Material and methods

The first part of the method is feature engineering, feature extraction, and feature screening of clinical and imaging data of the study subjects. In addition, to solve the problem of small samples and imbalance in the distribution of data labels, downscaling and instance weighting are used, and the effects verified using public databases. The second part comprises the construction of the model using kernel canonical correlation analysis (KCCA) and support vector machine (SVM), combining the clinical data and medical imaging data of the study subjects, maximizing the prognostic factors of each aspect of the study subjects, constructing a multi-view machine learning model for classifying DLBCL patient outcomes, and finally evaluating the constructed model by comparing

the performance of the models. A flowchart of the study is shown in Fig. 1.

Data description

Data sources

The data used in this study were obtained from patients diagnosed with DLBCL between December 2010 and December 2020 in the hematology department of a hospital in Shanxi Province, China. Two types of data were obtained for each subject: clinical information (including age, gender, lactate dehydrogenase, type B symptoms, and Ann Arbor staging) and PET/CT imaging data. In this study, we categorized the efficacy of chemotherapy according to whether CR was achieved within eight courses of chemotherapy.

Inclusion and exclusion criteria were used to select the study subjects from the hospital database.

The inclusion criteria were the following: (a) Patients diagnosed with DLBCL between December 2010 and December 2020 in the hematology department of a hospital in Shanxi Province. (b) Patients who had undergone PET/CT scans prior to chemotherapy. (c) Age ≥ 18 years. (d) Patients suitable for standard-of-care first-line chemotherapy. (e) Availability of all clinical, pathology, and imaging data.

The exclusion criteria were the following: (a) Patients with incomplete clinical or imaging data. (b) Patients with concomitant or prior history of other cancer types. (c) Negative baseline PET-CT.

All patients were screened for the inclusion and exclusion criteria, and finally 127 patients were enrolled. A total of 85 cases in the CR group were negatively labeled, and a total of 42 cases in the non-CR group were positively labeled. For the clinical characteristics, relevant clinical indicators involved in oncology were collected and organized according to the clinical practice guidelines in oncology. The relevant variables were extracted: age, gender, tumor stage, treatment options, international prognostic Index (IPI), karnofsky performance status (KPS), white blood cell (WBC), lactate dehydrogenase (LDH), $\beta 2$ -microglobulin ($\beta 2$ -MG), erythrocyte sedimentation rate (ESR), germinal center B-cell (GCB), hepatitis B virus (HBV), BCL-6, Ki-67, and R [13].

For the imaging features, image acquisition, volume of interest (VOI) lesion outlining and correction, voxel point feature measurement, histogram calculation, and correlation matrix calculation were successively employed. The LIFEx software was used to extract semiquantitative features, including the maximum standardized uptake value (SUVmax) and MTV of the PET/CT images [14], and quantitative features. First-order quantitative features include statistical properties of image voxel points, such

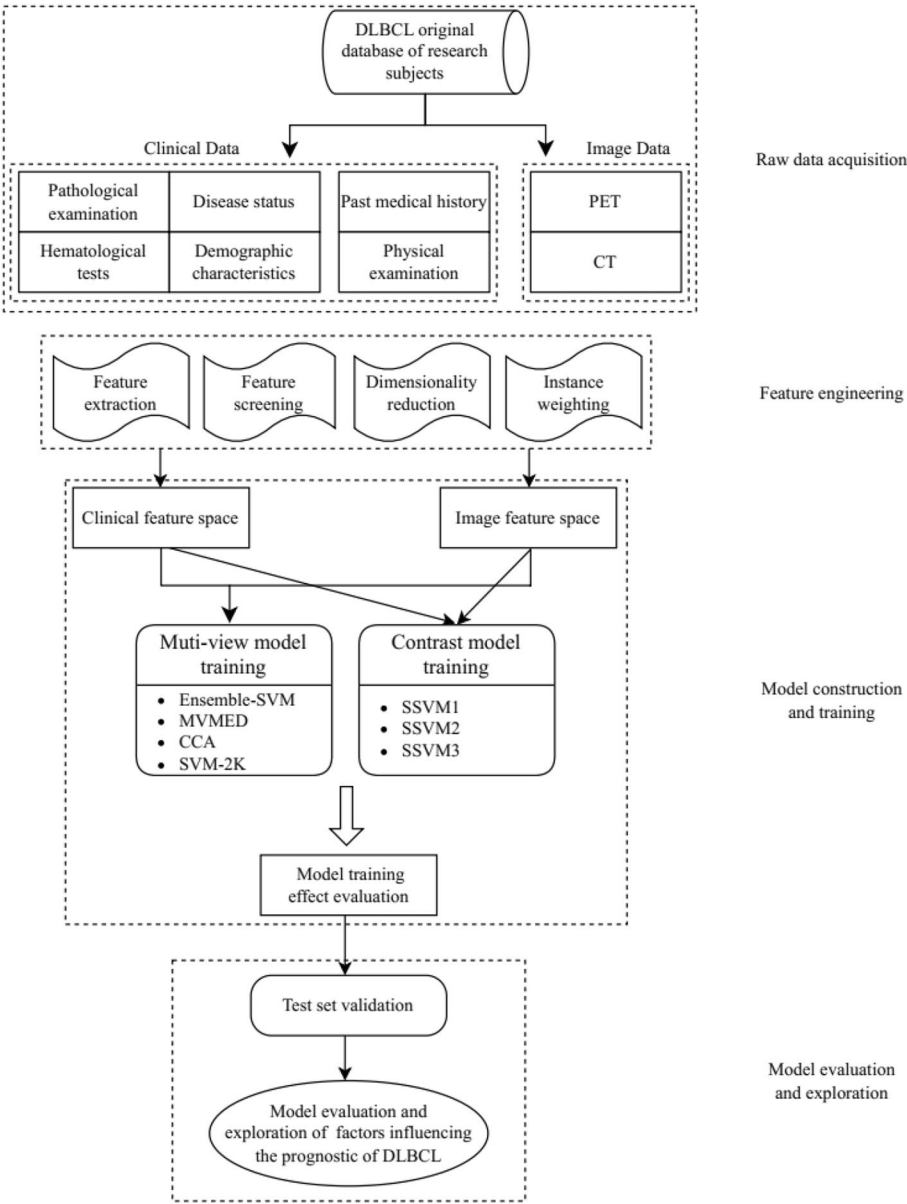


Fig. 1 Illustration of the research frame work

as skewness, kurtosis, and entropy. Second-order quantitative features include homogeneity, intensity, similarity, and contrast, which reflect the image gray level with respect to the direction, adjacency interval, and magnitude of change. Higher-order features include the high gray level emphasis value, gray level inhomogeneity, and stroke length inhomogeneity, which cannot be captured by the human eye [15, 16]. In addition, based on the IPI prognostic scoring criteria and the median age of onset of DLBCL patients in the sample, this study divided all patients into the younger (≤ 60 years old) and older

(> 60 years old) groups using a cutoff value of 60 years old and analyzed the prognostic differences between the two age groups.

¹⁸F-FDG PET/CT

The PET/CT scans were performed using a GE Discovery STE hybrid scanner (USA) for both lymphoma staging and restaging. Imaging data acquisition occurred 60 min following the intravenous administration of ¹⁸F-fluorodeoxyglucose (FDG) at a dose of 4.44–5.55 MBq/kg of body weight. The whole-body ¹⁸F-FDG PET, covering the

area from the head to the mid-thigh, was carried out in 3D mode with 6–8 bed positions, each lasting 2.5 min. CT data was utilized for attenuation correction of PET images. The CT scan settings included a tube voltage of 120 kV, a current of 180 mA, a pitch ratio of 0.938:1, a slice thickness of 3.75 mm, and a rotation time of 0.8 s per round. The PET images were reconstructed using the ordered-subsets expectation maximization (OSEM) algorithm with 2 iterations, 20 subsets, and a 128×128 pixel matrix size.

Public databases

Three public databases (Cleveland, Glass0, Ecoli1) with varying sample sizes and category imbalance rates from the KEEL website (<https://sci2s.ugr.es/keel/imbalanced.php>) were used to validate the effectiveness of feature engineering such as dimensionality reduction and instance weighting in this paper. The basic information about the datasets, including the DLBCL dataset (the subject of this study), is shown in Table 1.

Feature engineering

Feature extraction

Following *NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines) B-Cell Lymphomas Version 4.2020-August 13, 2020* and *NCCN Guidelines for Patients 2020-Diffuse Large B-Cell Lymphoma* [17], the clinical characteristics were extracted and organized, as shown in Table 2.

The Department of Nuclear Medicine Imaging collected the last FDG-PET/CT of the study subject, as imaging data, before chemotherapy was administered at the hospital. Semiquantitative eigenvalues of the MTV, TLG, and SUV values of the VOI of the tumor lesion of the study subject were obtained by feature measurement. In addition, quantitative eigenvalues of VOI were obtained by radiomics analysis. The third- and higher-order quantitative imaging features in the image were extracted using the gray run length matrix (GLRLM), neighborhood gray difference matrix (NGLDM), and gray zone length matrix (GLZLM), as shown in Table 3.

Table 2 Clinical feature extraction

Features	Grouping values	N (%)
Age	0 = Less than or equal to 60 years old 1 = Older than 60 years old	70 (55) 57 (45)
Gender	0 = Male 1 = Female	60 (48) 67 (52)
Tumor staging	0 = Early III 1 = Late III, IV	41 (32) 86 (68)
Treatment options	1 = chemotherapy 0 = Chemotherapy + Radiotherapy	108 (85) 19 (15)
IPI	0 = Low 1 = High	89 (7) 38 (3)
KPS	0 = Greater than or equal to 80 points 1 = Less than 80 points	83 (66) 44 (34)
WBC	0 = Normal 1 = Abnormal	93 (74) 34 (26)
LDH	0 = Normal 1 = Abnormal	54 (43) 73 (57)
β_2 -MG	0 = Normal 1 = Abnormal	34 (27) 93 (73)
ESR	0 = Normal 1 = Abnormal	57 (45) 70 (55)
GCB	0 = No 1 = Yes	67 (53) 60 (47)
HBV	0 = Negative 1 = Positive	104 (82) 23 (18)
BCL-6	0 = Negative 1 = Positive	95 (75) 32 (25)
Ki-67	0 = Less than or equal to 90% 1 = Greater than 90%	62 (49) 65 (51)
Rituximab (R)	0 = Used 1 = Not used	75 (59) 52 (41)

Feature screening

Recursive feature elimination (RFE) is used in feature selection [18]. In this study, mean decrease of accuracy (MDA) was chosen to measure feature importance. MDA is the average reduction in a model's prediction accuracy on a sample after a feature has been excluded. The larger the MDA, the more important the variable is to the model.

Table 1 Basic dataset information

Data	Sample size	Number of cases in the positive group	Number of cases in the negative group	Category ratios	Number of features
Cleveland	177	13	164	12.62	26
Glass0	214	70	144	2.06	20
Ecoli1	336	77	259	3.36	16
DLBCL	127	42	85	2.02	44

Table 3 PET/CT feature extraction

Indicators	Methods	Features
Semi-quantitative indicators	Measurement	SUVmax, SUVStd, MTV, TLG
Quantitative radiological indicators	Histogram	Skewness, kurtosis, entropy, energy
		Sphericity, compacity volume
	Grayscale symbiosis matrix	Homogeneity, Energy, Contrast, Correlation,
	Grayscale run length matrix	Entropy, dissimilarity
	Neighborhood grayscale difference matrix	SRE/LRE, LGRE/HGRE, SRLGE/SRHGE, LRLGE/LRHGE, GLNU/RLNU, RP
	Grayscale zone length matrix	Coarseness, contrast, busyness
		SZE, LZE, LGZE, HGZE, SZLGE, SZHGE, LZLGE, LZHGE, GLNU, ZLNU, ZP

MTV Metabolic Tumor Volume, TLG Total Lesion Glycolysis, GLCM Gray Level Coevolution Matrix, GLRLM Gray Level Run Length Matrix, SRE/LRE Short/Long Run Emphasis, LGRE/HGRE Low/High Gray Run Emphasis, SRLGE/SRHGE Short Run Low/High Gray Emphasis, LRLGE/LRHGE Long Run Low/High Gray Emphasis, GLNU/RLNU Gray Level Non-Uniformity/ Run Length Non-Uniformity, NGLDM Neighborhood Gray Difference Matrix, GLZLM Gray Level Zone Length Matrix, LGZE/HGZE Low/high Gray Zone Emphasis, SZLGE/SZHGE Short Zone Low/High Gray Emphasis, LZLGE/LZHGE Long Zone Low/High Gray Emphasis, GLNU/ZLNU Gray Level Non-Uniformity/ Zone length Non-Uniformity, ZP Zone Percentage

Dimensionality reduction

Because LDA is prone to overfitting for small samples with labeled information, PCA was chosen in this study for reducing the dimensionality of data [19].

Instance weighting

To solve the problem of sample label class imbalance, the proposed method adds sample weights during the training process. The location importance weight is denoted by LI . To calculate the LI of an instance, the instance first searches for the k nearest neighbors among all training instances, the number of neighbors with the same label as the instance is recorded as N_s , and the value of LI is defined as follows:

$$LI(x) = \begin{cases} 1, & \frac{N_s}{k} = 0 \text{ or } 1 \\ 1 + \frac{N_s}{k}, & \text{else} \end{cases}$$

Although LI assigns different weights to instances at different locations, it does not consider the imbalance in the classification task. In general, an imbalance occurs when the number of instances of one class (the minority class) is much less than the number of instances of another class (the majority class). The minority class instances are more valuable than those in the majority class. Therefore, the minority class instances should be given higher weights. Suppose that the number of instances of the minority class is N_{min} and the number of instances of the majority class is N_{maj} . The imbalance cost IC of input instance x is defined as

$$IC(x) = \begin{cases} \frac{N_{maj}}{N_{min}}, & x \in \text{minority class} \\ 1, & x \in \text{majority class} \end{cases}$$

The weights of the input instances x are then defined as follows:

$$W_{(x)} = LI_{(x)} \times IC_{(x)}$$

Model construction and training

Construction of the proposed model

Single-view model

Single-view1: Training using SVM (SSVM1) only within the image feature view.

Single-view2: Training using SVM (SSVM2) only within the clinical feature view.

Single-view3: Training using SVM (SSVM3) within the combined clinical and image views, and comparison with decision tree (C5.0), logistic regression classification (logistic), and multi-layer perceptron (MLP).

Multi-view model Ensemble-SVM (ESVM): An SVM classifier is trained for the clinical view and imaging view separately, and their classification labels are determined by a voting method.

Multi-view maximum entropy discrimination (MVMED): The joint distribution and common boundary γ of the parameters of the two-view MED classifiers, using clinical features and image features, are used to satisfy the consensus principle.

Canonical correlation analysis (CCA): The common feature space is learned by maximizing the correlation between two views (clinical and imaging) to obtain the projection matrix, and the classifier is trained on the joint space matrix for label prediction.

SVM-2 K: Combines KCCA with SVM. The SVM is trained on the joint feature space comprising clinical and imaging features. On the basis of CCA, relaxation variables and constraints are added to satisfy the principles of similarity and complementarity, map different view features to different kernel spaces before obtaining the common kernel subspace, and train the weights and thresholds of the SVM to obtain classification results.

In addition to the proposed SVM-2 K model, single-view and multi-view models were constructed in this study for performance comparison. These included CCA, ESVM [20], MVMED [21], decision tree [22], logistic classifier [23], MLP [24], and single-core SVM.

Model training

Following the random stratified sampling principle, the entire dataset was randomly divided into five copies before training, of which four copies were used as the training set and the remaining copy was used as the test set. In addition, to reduce the variation due to dataset partitioning, the dataset partitioning and evaluation were repeated 100 times and the final evaluation was based on the mean of the 100 results.

Model assessment

Because this study was dichotomous, values of accuracy, sensitivity, F-measure, AUC, and G-mean were used as evaluation metrics to evaluate classification performance, as shown in Table 4.

The results of each classifier can be classified into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Various evaluation metrics can then be computed according to the model output.

Results

Validation of feature effects on public datasets

This section aims to verify the effectiveness of the feature engineering method using PCA dimensionality reduction

with instance weighting on public datasets to improve the model performance. Three public datasets, Cleveland, Glass0 and Ecoli1, are used to process the data before and after feature engineering, and the SVM model is used for training and testing. The effect of feature engineering is evaluated by comparing AUC, accuracy (ACC), sensitivity (SEN), F1 value and G-mean.

On the Cleveland dataset, the feature-engineered SVM model improves the AUC by 12% (95%CI:0.111~0.128), the ACC by 10.1% (95%CI:0.093~0.109), the SEN by 20.5% (95%CI:0.194~0.211), and the F1 value and the G-mean by 10.4% (95%CI:0.099~0.116) and 14.1% (95%CI:0.134~0.153), respectively (Fig. 2), the difference was statistically significant ($p<0.001$). These results indicate that PCA downscaling and instance weighting effectively improve the model's ability to discriminate between minority class samples.

On the Glass0 dataset, the SVM improved the AUC by 8.3% (95%CI:0.074~0.091), the ACC by 9.1% (95%CI:0.083~0.099), the SEN by 13.2% (95%CI:0.121~0.138), and the F1 value and the G-mean by 8.4% (95%CI:0.079~0.096) and 10.9%(95%CI:0.102~0.121), respectively (Fig. 3), the difference was statistically significant ($p<0.001$). This result shows that the overall performance of the model can be improved by feature engineering.

On the Ecoli1 dataset, feature engineering improved AUC by 6.2% (95%CI:0.053~0.070), ACC by 4.1% (95%CI:0.033~0.049), SEN by 7.2% (95%CI:0.061~0.079), and F1 value and G-mean by 5.4% (95%CI:0.046~0.063) and 6.9%(95%CI:0.062~0.081), respectively (Fig. 4), the difference was statistically significant ($p<0.001$). These improvements demonstrate that feature engineering has a significant gain effect on small sample datasets.

Feature engineering results for the DLBCL study dataset

The aim of this part is to improve the performance of classification models for DLBCL patient datasets through feature engineering and to identify the most important clinical and imaging features. The clinical and imaging data of DLBCL patients were feature extracted and

Table 4 Model evaluation metrics

Classification model evaluation metrics	Significance	Formula
Accuracy (ACC)	Measures the proportion of the samples that are correctly classified	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity (SEN)	The recall rate, which measures the proportion of the samples that are correctly classified	$Sensitivity = \frac{TP}{TP+FN}$
F1 score (F1)	Harmonic mean of precision and recall	$F = \frac{2 \times Precision \times Recall}{Precision+Recall}$
Area under the curve (AUC)	Area under the ROC curve with false positive rate (FPR) as the horizontal axis and true positive rate (TPR) as the vertical axis	$FPR = \frac{FP}{FP+TN}$ $TPR = \frac{TP}{TP+FN}$
G-mean	Geometric mean of the classification accuracy of the minority class and that of the majority class. It is used to evaluate the performance of the model after using the unbalanced data of the class	$G - means = \sqrt{TPR \times TNR}$

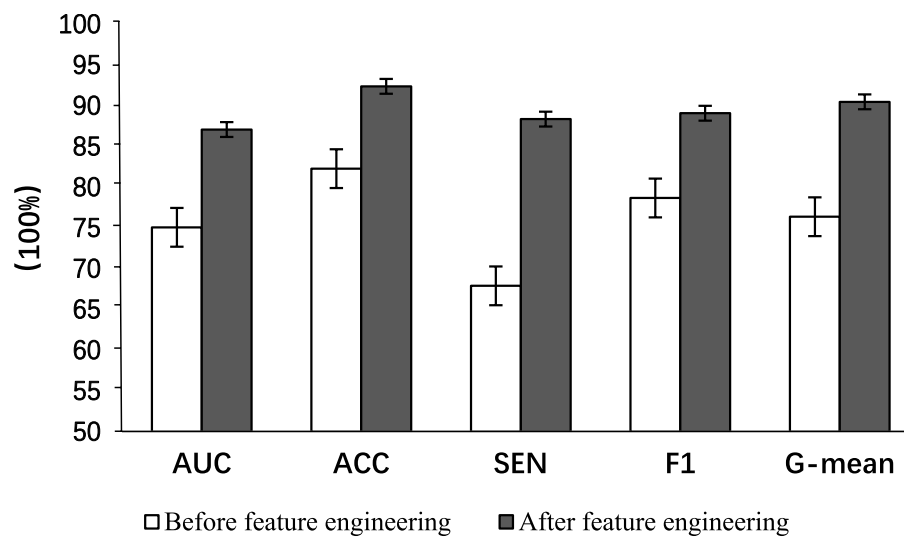


Fig. 2 Comparison of classification performance of SVM before and after feature engineering on Cleveland dataset

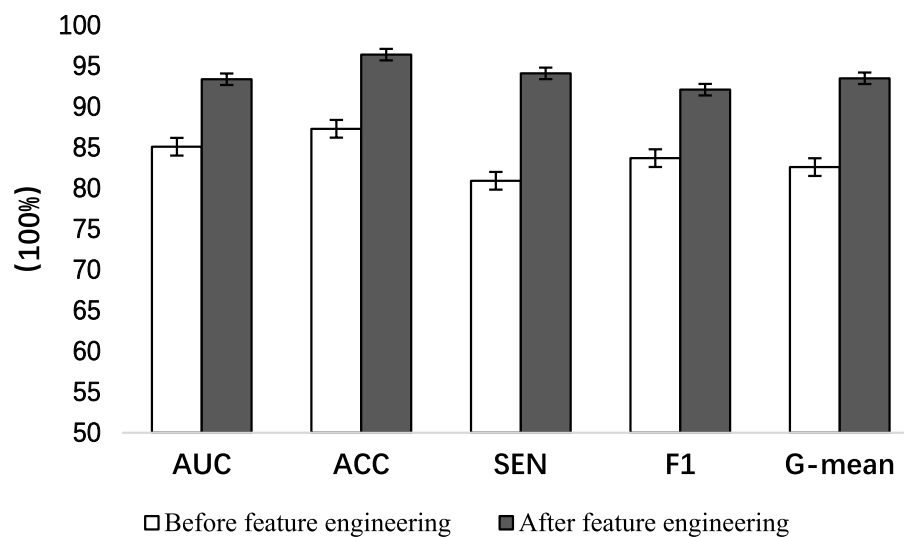


Fig. 3 Comparison of classification performance of SVM before and after feature engineering on Glass0 dataset

screened using SVM recursive feature elimination (RFE), and a total of 13 features with the best performance in cross-validation were finally obtained using ACC as a measure (Accuracy=0.863). On this basis, feature engineering was performed using PCA dimensionality reduction with instance weighting, and its impact on model performance was analyzed.

Among the optimal features, 7 were from clinical data and 6 were from imaging data. ipi score (MDA=14.16) and tumor volume (MDA=11.21) were the most important features (Fig. 5). after PCA downscaling

and instance weighting, the AUC of the SVM model was improved by 10.5% (95%CI:0.097~0.109), the ACC was improved by 9.8% (95%CI:0.091~0.100), the SEN was improved by 11.9% (95%CI:0.112~0.122), and the F1 value and G-means improved by 9.2% (95%CI:0.082~0.095) and 7.8% (95%CI:0.067~0.083), respectively (Table 5), the difference was statistically significant ($p < 0.001$). These results indicate that feature engineering significantly improves the generalization ability and prediction accuracy of the model.

Model performance comparison

Performance of each model on DLBCL training set

Compare the performance of different single-view and multi-view learning models on the DLBCL training set to evaluate the advantages of SSVM3 and SVM-2 K models. Multiple single-view and multi-view models are used for the DLBCL training set to compare their performance on AUC metrics.

SSVM3 trained in the space of spliced clinical and imaging features performed optimally in the AUC metrics (Fig. 6), outperforming SSVM2 trained in the space of clinical features alone and SSVM1 trained in the space of imaging features alone, suggesting that splicing the clinical and imaging features helps to improve the predictive performance of the model.

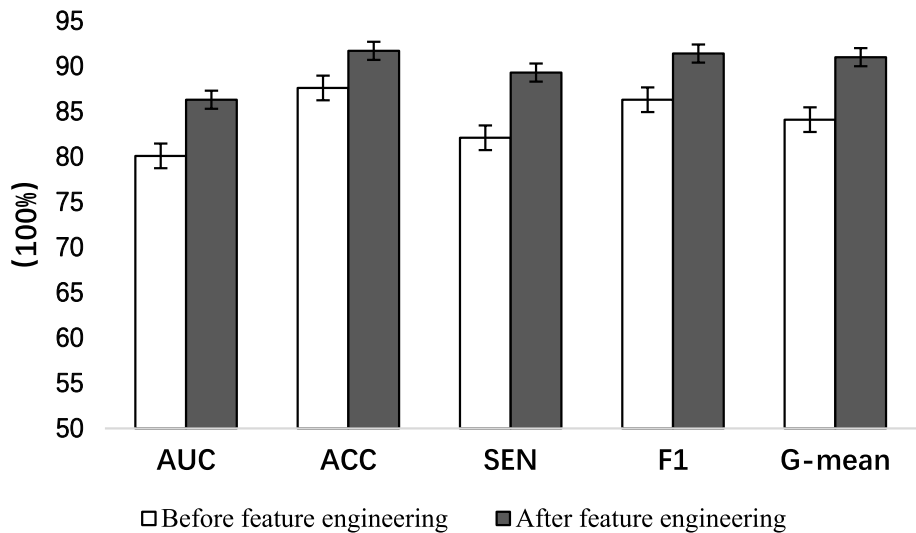


Fig. 4 Comparison of classification performance of SVM before and after feature engineering on Ecoli1 dataset

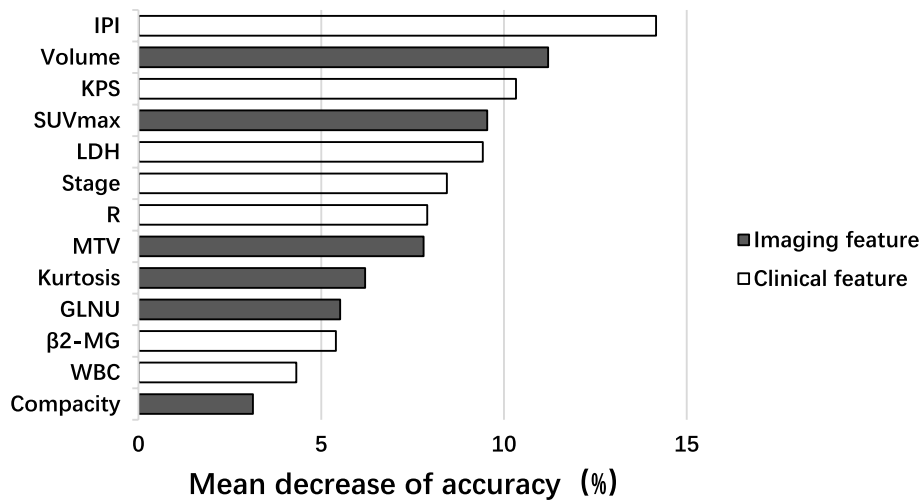


Fig. 5 Feature importance ranking

Table 5 Comparison of classification performance of SVM before and after feature engineering on the DLBCL study dataset

	AUC	ACC	SEN	F1	G
Before feature engineering	0.827 ± 0.024	0.863 ± 0.015	0.778 ± 0.023	0.833 ± 0.02	0.846 ± 0.031
After feature engineering	0.932 ± 0.018	0.961 ± 0.017	0.897 ± 0.015	0.925 ± 0.029	0.924 ± 0.026

In the multi-view model, SVM-2 K outperforms MVMED, ESVM, and CCA with an AUC of 96.2% (95%CI:95.57%~96.83%) for the training set (Fig. 7). SVM-2 K utilizes kernel-typical correlation analysis (KCCA) in conjunction with a support vector machine for optimal integration of the data and learning results.

Performance of each model on DLBCL test set

The purpose of this section is to evaluate the prediction performance of different single-view and multi-view models on the DLBCL test set, especially SVM-2 K versus other models. We tested the performance of single-view and multi-view models separately. The single-view models included SSVM1, SSVM2, and SSVM3 trained in different feature view spaces (clinical and imaging),

in addition to comparisons with common logistic regression models, C5.0, and MLP. Multi-view models include ESVM, MVMED, CCA, and SVM-2 K, with a focus on comparing the AUC, accuracy (ACC), sensitivity (SEN), and F1 value vs. G-mean of each model.

Single view model performance The performance of SVM models in different feature view spaces in the DLBCL test set is shown in Table 6. SSVM2 trained in clinical feature space outperforms SSVM1 trained in image feature space, while SSVM3 trained in the spliced feature space has the best performance. The specific performance is shown as AUC reaches 86.3% (95%CI:85.93%~86.67%), ACC 91.6% (95%CI:90.71%~92.49%), SEN 83.2% (95%CI:82.32%~84.08%), and F1 value and G-mean 85.7% (95%CI:85.09%~86.31%) and 86.1% (95%

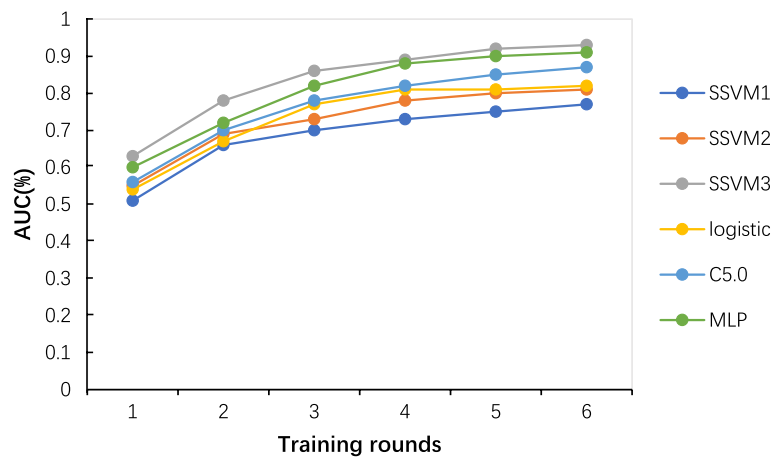


Fig. 6 Comparison of classification performance of single-view learning models on the DLBCL training set

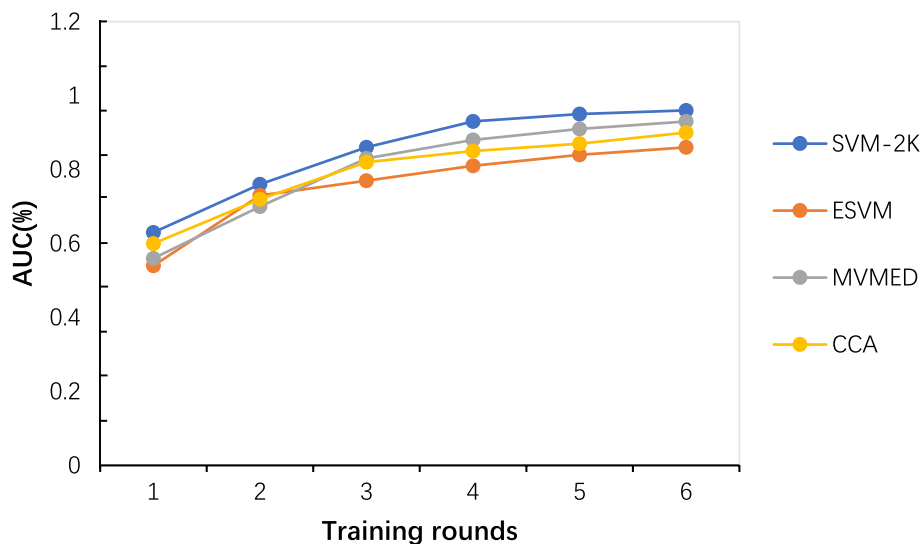


Fig. 7 Comparison of classification performance of MVL models on the DLBCL training set

CI:85.34%~86.86%), respectively. SSVM3 makes full use of the multi-view data by integrating the clinical and imaging features, which makes it superior to the single-view model in terms of performance.

Comparison with other single-view learning models Table 7 demonstrates the results of SSVM3 compared with common single-view learning models (Logistic regression, C5.0 and MLP). SSVM3 outperforms the other models across the board. The specific performance is shown as AUC reaches 86.3% (95%CI:85.93%~86.67%), ACC 91.6%(95%CI:90.71%~92.49%), SEN 83.2% (95%CI:82.32%~84.08%), and F1 value and G-mean 85.7% (95%CI:85.09%~86.31%) and 86.1% (95%CI:85.34%~86.86%), respectively. This result indicates that SSVM3 is able to better capture key information when combining clinical and imaging features, significantly improving the predictive ability of the model.

Multi-view model performance The test results of the multi-view model are shown in Table 8. SVM-2 K performed the best on all the metrics, with an AUC of 92.1% (95%CI:91.69%~92.51%), an ACC of 96.9% (95%CI:96.22%~97.16%), and F1 and G-mean values of 92.8% (95%CI:92.39%~93.21%) and 91.4% (95%CI:90.78%~92.02%), respectively. Although its sensitivity (SEN=90.9% (95%CI:90.28%~91.52%)) was slightly lower than that of MVMED (SEN=91.1% (95%CI:90.44%~91.76%)), SVM-2 K had the best overall performance. Through the combination of kernel typical correlation analysis (KCCA) and support vector machine, SVM-2 K is able to effectively integrate the complementary information of different feature views, which significantly improves the overall performance.

Model AUC comparison Figure 8 shows the comparison of the AUC values of the models on the DLBCL test set. It can be seen that the AUCs of the multi-view models are all better than the single-view model, and SVM-2 K has the best performance with an AUC of 92.1% (95%CI:91.49%~92.71%). This further validates that SVM-2 K can effectively improve the prediction performance in multi-view learning.

Table 6 Comparison of SVM performance on different feature views of the test set (%)

Performance indicators	SSVM1	SSVM2	SSVM3
AUC	72.4±2.43	81.1±1.13	86.3±1.31
ACC	81.3±3.41	87.1±2.41	91.6±3.14
SEN	70.2±4.65	78.9±1.91	83.2±3.11
F1 score	75.0±2.33	77.3±1.84	85.7±2.16
G-mean	77.6±1.44	80.1±2.01	86.1±2.66

Table 7 Comparison of classification performance of single-view models on the test set (%)

Performance indicators	MLP	C5.0	logistic	SSVM3
AUC	84.1±2.13	82.1±2.41	79.6±1.13	86.3±1.31
ACC	90.3±1.98	85.2±3.33	80.5±2.15	91.6±3.14
SEN	82.6±2.89	80.4±1.21	76.1±2.47	83.2±3.11
F1 score	85.2±3.15	81.8±2.16	79.2±1.66	85.7±2.16
G-mean	83.2±2.83	82.4±1.95	78.7±1.77	86.1±2.66

Discussion

DLBCL has a high relapse rate. Therefore, for each DLBCL patient it is necessary to make a comprehensive assessment of the efficacy of treatment in advance and prepare a personalized treatment plan accordingly. The IPI scoring system evaluates the patient's prognosis only with reference to the pre-treatment clinical characterization and ignores the role of the patient's prognostic factors with respect to imaging PET/CT, including semiquantitative features (such as SUV) and quantitative features (such as radiological features). This is because the clinical data and imaging data originate from different sources and have different structures; in addition, their nature is different, making it difficult to use them in an integrated manner. Therefore, the aim of this study was to construct a prognostic model for DLBCL patients with high predictive performance by using a combination of KCCA and SVM to make full use of patients' imaging data and clinical data.

After feature screening, as shown in Fig. 5, most of the clinical features selected have been shown to be associated with the prognosis of DLBCL. IPI [25, 26] is a recognized prognostic indicator for DLBCL, and studies have demonstrated that high IPI scores are significantly associated with poor patient prognosis; advanced disease stage is a risk factor for DLBCL prognosis, and is associated with low patient survival and short survival time [26]. Patients with primary gastrointestinal DLBCL with elevated LDH [27] have a corresponding reduction in overall survival and

Table 8 Comparison of classification performance of multi-view models on the test set (%)

Performance indicators	ESVM	MVMED	CCA	SVM-2K
AUC	88.1±2.22	90.1±1.76	88.7±2.55	92.1±1.45
ACC	92.8±1.32	92.8±1.12	91.3±1.87	96.9±1.65
SEN	85.4±3.11	91.1±2.31	82.9±1.34	90.9±2.17
F1 score	87.4±2.31	90.3±2.46	87.1±2.23	92.8±1.45
G-mean	86.8±1.14	90.7±2.77	82.2±2.32	91.4±2.17

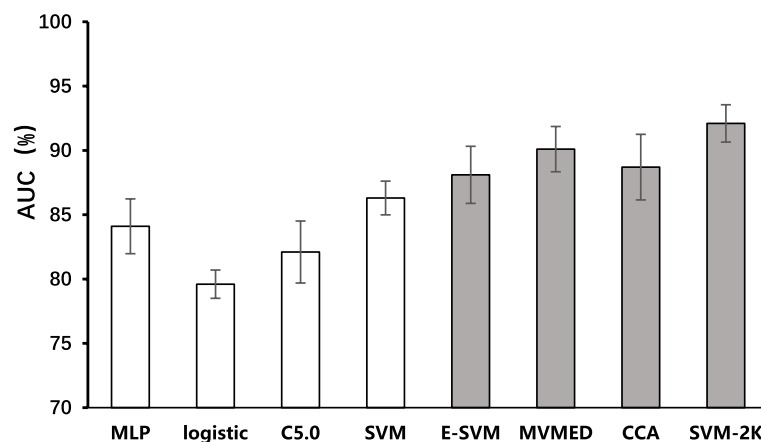


Fig. 8 Comparison of the area under the ROC curve achieved by each model on the test set

progression-free survival. According to a study by Kanemasa et al. [27], β_2 -M is an important prognostic factor in patients with DLBCL, and studies have also shown that the use of immunotherapy containing rituximab (R) is an effective chemotherapy for patients with limited aggressive non-Hodgkin's lymphoma with a poor prognosis or high disease grade [28]. Of the imaging features, SUVmax and MTV are the most common semiquantitative radiological features: they have been extensively reported to correlate with disease-free survival in DLBCL [29]. Volume and compacity reflect the response to tumor shape and volume, which is a factor that influences disease severity and treatment difficulty. Kurtosis is a statistic of voxel points in the histogram analysis of the patient's PET image, and is quantitatively descriptive of the tumor in radiology. GLNU is a higher-order statistical feature extracted from the voxel points of the image by matrix computation using the GLRLM statistical model [29, 30], and it also shows the prognostic impact of higher-order radiologic features on DLBCL.

To solve the problem that the small sample size and unbalanced class distribution of DLBCL patient data may affect the performance of prediction models, this study used feature engineering including PCA and instance weighting. Table 4 shows that feature engineering is helpful for improving classification performance: the AUC of the SVM learner on the DLBCL study dataset was increased by 10.5% after PCA and instance weighting. Many studies in recent years have shown that the problem of class imbalance in data seriously affects the classification accuracy of prediction models. Because there are far more instances in the majority class than the minority class in the original data, and the minority class instances are less informative, the classification model will be largely

biased toward the majority class in the training process because this achieves higher accuracy. Therefore, the model may not be sufficiently sensitive to the minority class instances, resulting in low sensitivity. In addition, instances in different locations have different importance in the classification task. Boundary instances have a significant effect on classification, whereas internal instances contribute little to the classification task. Therefore, it is unreasonable to treat all instances in the training set equally. Appropriately assigning high weights to boundary instances can provide advantages for the classification task. In the proposed method, location weights are added to the category weights, both to optimize the final decision surface of the classifier and to enhance the classification performance for the minority class, thereby improving the sensitivity and thus the AUC.

To solve the problem of how to adequately and rationally use the imaging data of DLBCL patients to provide reference values for patient prognosis, this paper presents a MVL method that combines KCCA with an SVM to construct the SVM-2 K model for patient prognosis, and compares its results with those of a single-view learning model.

Table 6 shows that, of the single-view learning models, SVM achieved the best performance. MLP and C5.0 both realize linear to nonlinear conversion by adding intermediate layers. However, because of the problem of overfitting and the poor robustness of decision trees, single decision trees are often not as good as they should be in practical applications. SVM makes effective use of kernel methods for nonlinear transformations and has lower computational complexity than MLP and decision trees, which makes it a good choice for solving nonlinear problems on small datasets. As shown in Fig. 6 and Table 5, the classification performance of SSVM3 (AUC = 86.3%),

trained in the feature space that combines imaging features with clinical features, was better than that of SSVM1 (AUC=72.4%) and SSVM2 (AUC=81.1%). This proves that combining the clinical features with imaging features for training is helpful for improving the performance of classification. It also proves indirectly that the semiquantitative features and quantitative radiological features of PET/CT images are valuable for the prognosis of patients with DLBCL. The performance of SSVM2 was better than that of SSVM1, which indicates that the clinical features extracted in this study have higher prognostic value than the imaging features that were extracted.

Figure 8 compares the performance of the MVL models ESVM (AUC=88.1%), MVMED (AUC=90.1%), CCA (AUC=88.7%), and SVM-2 K (AUC=92.1%). All results are better than the results of the best-performing single-view learning model SSVM3. The results show that the performance of MVL models on the dataset used in this study is better than that of single-view learning models. Most previous DLBCL prognostic studies were conducted using either clinical data alone or imaging data alone. Wang et al. [31] assessed the predictive accuracy of tumor recurrence in DLBCL patients by using ROC curves and obtained a sensitivity of 83.3%, specificity of 78.3%, and AUC of 0.864. Gong et al. [32] analyzed the risk factors associated with recurrence in DLBCL patients, and the 1-year, 2-year, 3-year, and 4-year AUC values were 0.812, 0.850, 0.837, and 0.801, respectively; this performance is not as good as that of the MVL model designed in our study. If it uses clinical data alone, a model cannot make use of the important imaging data generated during patient treatment; if it uses imaging data alone, it omits key clinical information. In addition, clinical data and imaging data are not distinguished by algorithms that merge all features to adapt to the learning environment when dealing with multi-view feature data. Such direct merging ignores the attributes of the same object distributed in different feature spaces with different data-specific statistical properties and different physical significance. This reduces the amount of information that the model learns from the data, resulting in worse performance.

The comparison of classification performance on the test set by each multi-view model, in Table 7, shows that the proposed model SVM-2 K achieved the best overall performance (AUC=92.1%, accuracy=96.9%, sensitivity=90.9%, F1=92.8%, and G-mean=91.4%). The CCA and MVMED models both consider the correlation between view features but ignore the differences between different views (i.e., complementarity). The proposed model SVM-2 K uses an insensitive L1 parametrization when considering the constraints of correlation and uses slack variables to measure the number of them

that do not conform to the similarity; to some extent, this can make better use of the complementary information between different features in the data and thereby improve the performance of the model. In a sense, MVL can also be referred to as integrated-view learning. The ESVM evaluated in this study uses the method of training an SVM classifier for the clinical view and imaging view separately, and integrating the classification results obtained under each view by the voting method for late integration. SVM-2 K combines KCCA and SVM and constructs their respective kernel matrices using kernel functions for different feature types to represent the local characteristics, evaluates and fuses each kernel matrix to reflect the global characteristics, and finally classifies them through the fused kernel matrix. Farquhar et al. [33] presented both experimental and theoretical analyses of SVM-2 K, showing superior results.

The innovations of this study include the following. First, in this study, not only conventional semi-quantitative metrics were used, but also more complex quantitative features of higher-order imaging were extracted by a radiomics approach, which provided more detailed and potentially more valuable information about the prognosis of patients with DLBCL, and improved the predictive performance of the model and the value of its clinical application. Second, because clinical data and imaging data have different patient categories and are of different nature, the modeling scheme of MVL is used to rationalize their use and improve the performance of classification. Third, our model handles small samples of class-imbalanced data using PCA and instance weighting to achieve superior performance.

The shortcomings of this study are as follows. First, although the genetic data of DLBCL patients have been shown to have prognostic value in a large number of studies, because of the difficulties of collecting and testing genetic data of patients, such data have not been used for modeling in this study. Second, this research was based on the data provided by a specific hospital. Therefore, external validation is necessary to evaluate the generalizability of the model and reproducibility of the results.

Conclusions

In this paper, we mainly propose the use of KCCA and SVM methods to model the prognosis of DLBCL patients. After feature engineering, such as research object screening, feature extraction, feature screening, dimensionality reduction, and instance weighting, we divide the data into training and test sets, and finally train the proposed model and evaluate its performance. The results lead to the following conclusions.

1. Feature engineering, including dimensionality reduction and instance weighting, largely solves the problem of decreased sensitivity when the model is trained on small samples with imbalanced class distribution. 2. The semiquantitative features of PET/CT images and the quantitative radiological features have some prognostic value for patients with DLBCL. 3. The performance of the MVL model is better than that of the single-view learning model; the proposed model SVM-2 K has the best overall performance and can accurately predict the prognosis of patients, which is important for the treatment of DLBCL patients and assisting doctors' decision-making.

Acknowledgements

We thank Liwen Bianji (Edanz) (www.liwenbianji.cn) for editing the English text of a draft of this manuscript.

Authors' contributions

YanHong Luo, ZhenHuan Yang, and YongAo Li designed the study, and Jie Zhou provided helpful advice. YongAo Li, ZhenHuan Yang, Kai Yu, YuJiao Guo, XueMan Wang, Na Yang, Yan Zhang, and TingTing Zheng collected and organized the dataset. YanBo Zhang, HongMei Yu, and ZhiQiang Zhao provided useful guidance. ZhenHuan Yang designed the code. ZhenHuan Yang and YongAo Li wrote the paper, and drafted and revised the work with the help of YanHong Luo. Both Jie Zhou and YanHong Luo are corresponding authors.

Funding

This work was supported by the National Natural Science Foundation of China [Grant Nos. 81502897, 82273742 and 82173631], Applied Basic Research Program of Shanxi Province [Grant No.202103021224245], Shanxi Province Graduate Education Innovation Program [Grant No.2024JG088], and Teaching Reform and Innovation Program for Higher Education Institutions in Shanxi Province (General Program) [Grant No.J20240531].

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

This study did not involve any human trials. We have obtained ethics approval from the Shanxi Cancer Hospital ethics committee, with reference number 201,835. The data did not contain personal and health information that could be connected back to the original identifiers. The data used in this study was anonymized before its use. The requirement to obtain informed consent was waived because of the secondary nature of the de-identified data in the retrospective study design.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan 030001, China. ²Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment, Taiyuan 030001, China. ³Department of Hematology, Shanxi Province Cancer Hospital/Shanxi Hospital Affiliated to Cancer Hospital, Chinese Academy of Medical Sciences/Cancer Hospital Affiliated to Shanxi Medical University, Taiyuan 030013, China. ⁴Department of Nuclear Medicine, Shanxi Province Cancer Hospital/Shanxi Hospital Affiliated to Cancer Hospital, Chinese Academy of Medical Sciences/Cancer Hospital Affiliated to Shanxi Medical University, Taiyuan 030013, China.

Received: 14 March 2024 Accepted: 27 November 2024

Published online: 05 December 2024

References

- Brooks E, Fang P, Pinnix C. Salvage radiotherapy for primary refractory and relapsed diffuse large B-Cell lymphoma. *Br J Radiol.* 2021;94(1127):20210360.
- Stefano A. Challenges and limitations in applying radiomics to PET imaging: possible opportunities and avenues for research. *Comput Biol Med.* 2024;179:108827.
- Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol.* 2014;32(27):3059.
- Shagera QA, Cheon GJ, Koh Y, et al. Prognostic value of metabolic tumour volume on baseline 18F-FDG PET/CT in addition to NCCN-IPI in patients with diffuse large B-cell lymphoma: further stratification of the group with a high-risk NCCN-IPI. *Eur J Nucl Med Mol Imaging.* 2019;46(7):1417–27.
- Tang J, Tian Y. A survey on multi-view learning (in Chinese). *Mathematical modeling and its applications.* 2017;6(3):16.
- Wang Z, Chen L, Zhang J, et al. Multi-view ensemble learning with empirical kernel for heart failure mortality prediction. *Commun Numer Methods Eng.* 2020;36(1):e3273.1–e.17.
- Ali H, Salleh MNM, Saedudin R, et al. Imbalance class problems in data mining: a review. *Indonesian J Electric Eng Comp Sci.* 2019;14(3):1560–71.
- Wang Y, Yao Q, Kwok JT, et al. Generalizing from a few examples: a survey on few-shot learning. *ACM Comp Surveys (CSUR).* 2020;53(3):1–34.
- Xu C, Tao D, Xu C. A survey on multi-view learning[J]. *arXiv preprint arXiv:1304.5634.* 2013.
- Zhu Z, Wang Z, Li D, et al. Multiple partial empirical kernel learning with instance weighting and boundary fitting. *Neural Netw.* 2020;123:26–37.
- Wang Y, Yao Q. Few-shot learning: a survey. 2019.
- Zhao W. Research on the deep learning of the small sample data based on transfer learning[C]. *AIP conference proceedings.* AIP Publishing. 2017;1864(1).
- Koh M, Hayakawa Y, Akai T, et al. Novel biomarker, phosphorylated T-LAK cell-originated protein kinase (p-TOPK) can predict outcome in primary central nervous system lymphoma. *Neuropathology.* 2018;38(3):228–36.
- Song MK, Yang DH, Lee GW, et al. High total metabolic tumor volume in PET/CT predicts worse prognosis in diffuse large B cell lymphoma patients with bone marrowinvolvement in rituximab era. *Leukemia Res.* 2016;42:1–6.
- Rahim MK, Kim SE, So H, et al. Recent trends in PET image interpretations using volumetric and texture-based quantification methods in nuclear oncology. *Nucl Med Mol Imaging.* 2014;48(1):1–15.
- Forgacs A, Pall H, Dahlbom M, et al. A study on the basic criteria for selecting heterogeneity parameters of F18-FDG PET images. *PloS One.* 2016;11(10):e0164113.
- Wierda WG, Byrd JC, Abramson JS, et al. NCCN guidelines insights: chronic lymphocytic leukemia/small lymphocytic lymphoma, version 2.2019. *J Nat Comprehensive Cancer Netw.* 2019;17(1):12–20.
- Louw N, Steel S. Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. *Comput Stat Data Anal.* 2006;51(3):2043–55.
- Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev: Comput Stat.* 2010;2(4):433–59.
- Alickovic E, Subasi A. Ensemble SVM method for automatic sleep stage classification. *IEEE Trans Instrum Meas.* 2018;67(6):1258–65.
- Sun S, Chao G. Multi-view maximum entropy discrimination[C]. *Twenty-third international joint conference on artificial intelligence.* 2013.
- Myles AJ, Feudale RN, Liu Y, et al. An introduction to decision tree modeling. *J Chemometrics: A J Chemometrics Soc.* 2004;18(6):275–85.
- Zou X, Hu Y, Tian Z, et al. Logistic regression model optimization and case analysis[C]. *2019 IEEE 7th international conference on computer science and network technology (ICCSNT).* IEEE; 2019. p. 135–9.
- Riedmiller M, Lerner A. Multi layer perceptron. *Machine learning lab special lecture: University of Freiburg;* 2014. p. 7–24.

25. Coiffier B, Lepage E, Brière J, et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *New England J Med*. 2002;346(4):235–42.
26. Zhang A, Ohshima K, Sato K, et al. Prognostic clinicopathologic factors, including immunologic expression in diffuse large B-cell lymphomas. *Pathol Int*. 1999;49(12):1043–52.
27. Mao L, Wang X, Wang C, et al. Evaluation of different staging systems and prognostic analysis of 110 primary gastrointestinal diffuse large B cell lymphoma. *Zhonghua Yi Xue Za Zhi*. 2019;99(24):1853–8.
28. Kanemasa Y, Shimoyama T, Sasaki Y, et al. Beta-2 microglobulin as a significant prognostic factor and a new risk model for patients with diffuse large B-cell lymphoma. *Hematol Oncol*. 2017;35(4):440–6.
29. Xiao G, Meng J, Zhang J, et al. Clinical application of detecting 21-gene recurrence score in predicating prognosis and therapy response of patients with breast cancer from two medical centers. *Cancer Invest*. 2017;35(10):639–46.
30. Zhou Y, Ma XL, Pu LT, et al. Prediction of overall survival and progression-free survival by the 18FFDG PET/CT radiomic features in patients with primary gastric diffuse large B-cell lymphoma[J]. *Contrast Media Mol Imaging*. 2019;2019(1):5963607.
31. Wang C, Zhou X, Liu GY, et al. Analysis of different protein expression levels in peripheral blood circulating tumor cells from patients with diffuse large B-cell lymphoma and their predictive efficiency for recurrence. *Zhonghua Yi Xue Za Zhi*. 2023;103(17):1328–33.
32. Gong Y, Yan H, Yang Y, et al. Construction and validation of a novel nomogram for predicting the recurrence of diffuse large B cell lymphoma treated with R-CHOP. *Pharmgenomics Pers Med*. 2023;16:291–301.
33. Farquhar JDR, Hardoon DR, Meng H, et al. Two view learning: SVM-2K, theory and practice[C]. *Proceedings of the 18th International Conference on Neural Information Processing Systems*. 2005. p. 355–62.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.