Multimodal Deep Learning Model on Interim ^{18}F-FDG PET/CT for Predicting Primary Treatment Failure of Diffuse Large B-cell Lymphoma

Cheng Yuan
Shanghai Jiao Tong University https://orcid.org/0000-0001-8083-3026
Qing Shi
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Xinyun Huang
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Li Wang
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Yang He
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Biao Li
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Weili Zhao
Shanghai Jiao Tong University Medical School Affiliated Ruijin Hospital
Dahong Qian (dahong.qian@sjtu.edu.cn)
Shanghai Jiao Tong University

Research Article

Keywords: Lymphoma, primary treatment failure, prognosis, deep learning, multimodality fusion

Posted Date: February 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1294701/v1

Abstract

Purpose: Prediction of primary treatment failure (PTF) is necessary for patients suffering from diffuse large B-cell lymphoma (DLBCL), since it serves as a prominent means for enhancing front-line outcomes. Utilizing interim ^{18}F-Fluorodeoxyglucose (FDG) positron emission tomography and computed tomography (PET/CT) image data, we aimed to construct multimodal deep learning (MDL) models to predict possible PTF of low-risk DLBCL, which could enable individualized treatment decision-making in clinical practice.

Methods: From June 2016 to November 2020, 205 DLBCL patients undergoing interim ^{18}F-FDG PET-CT scans and the front-line standard-of-care were enrolled. We also collected other 44 patients for the external validation. We built a powerful backbone by redesigning the famous visual recognition network named Conv-LSTM in aspects of network architecture and learning strategy. On top of our improved backbone, multiple MDL models using different feature fusion strategies were developed and compared, including pixel intermixing model, separate channel model, separate branch model, quantitative weighting model, and hybrid learning model. Moreover, we proposed to use a contrastive training objective in the above best model to enhance the modal correlation of semantic embeddings for further improving prediction performance. For visualization, the region of interest was instructed using an activation map.

Results: The MDL model using the hybrid learning strategy provided the best performance in predicting possible PTF with the accuracy of 89.76% (95% confidence interval [CI]: 84.85%–93.20%) in the test cohort. After further optimized by contrastive objective training, the accuracy was improved to 91.22% (95% CI: 86.55%–94.37%). The AUCs of contrastive hybrid learning achieved 0.926 and 0.925 in the test cohort and external validation cohort, respectively.

Conclusion: Our model showed outstanding performance for predicting PTF of low-risk DLBCL and hold promise of improving clinical individualized treatment strategies.

1. Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most frequently observed histologic subtype of lymphoma that is particularly prevalent in Asia [1]. Patients typically present with progressive mass involving lymph nodes and extranodal sites. Currently, over 60% of patients with DLBCL are cured after front-line standard-of-care R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) chemotherapy [2]. Nevertheless, as high as 15% patients undergoing this chemotherapy experience primary treatment failure (PTF) limiting their median survival time to one year at most [3]. As a remedy to this problem, novel emerging therapies, such as chimeric antigen receptor T-cell, were proposed. These therapies show high response rates in relapsed/refractory DLBCL [4] and benefit patients at high risk for PTF towards R-CHOP. Ideally, it is necessary to identify these high-risk patients before ordering them to receive individualized therapies.

Although the revised international prognostic index (R-IPI) and the presence of TP53 mutation are effective in predicting long-term survival among DLBCL patients, they cannot identify those patients likely to experience PTF [5, 6]. With the development of medical imaging, ^{18}F-Fluorodeoxyglucose (FDG) positron emission tomography and computed tomography (PET/CT) emerged as an effective tool that assists in the diagnosis, staging, prognosis, and predicting treatment response in oncology [7–9]. As a prognostic marker in DLBCL, quantitative analysis of initial and interim diagnostic imaging has been recently proposed. Kahle *et al.* [10] conducted a semi-quantitative analysis of ^{18}FDG-PET scans to investigate differences in the presence of necrosis between DLBCL cases with or without a MYC gene rearrangement. Similarly, Schöder *et al.* [11] utilized ^{18}FDG-PET scans at baseline, interim, and end of treatment (EoT) to identify biomarkers of response that are predictive of remission and survival. Recently, Santiago *et al.* [12] built a CT-based radiomics approach that utilizes random forest (RF) machine learning for predicting refractory DLBCL. Senjo *et al.* [13] measured metabolic heterogeneity using ^{18}FDG-PET/CT to predict a worse prognosis. These studies indicated the potential value of texture analysis of both PET and CT scan, which is correlated well with patient survival. Combining the metabolic information of PET scan to the anatomic features of CT scan in lymphoma investigations provides sparse results, albeit these results are possibly of some added value in predicting outcomes [14, 15].

However, diagnostic imaging is routinely used for staging purposes, and few conventional radiological findings have been correlated with PTF in DLBCL [16].

Radiomics has been widely used to correlate feature information from medical images to disease outcome including overall survival and tumor metastasis [17]. Traditional radiomics studies generally involve three steps: (1) the manual delineation of regions of interest (ROIs), (2) the quantitative extraction of hand-crafted radiomic features (i.e., shape, intensity, and texture) from ROIs, and (3) multivariate statistical analysis based on support vector machine (SVM) or RF to determine the correlation [18]. Many studies have presented effective prediction approaches and obtained important conclusions regarding outcomes in oncology. Aerts *et al.* [19] found a multitude of radiomic features with prognostic power in independent datasets concerning patients of lung and head-and-neck cancers. Seidler *et al.* [14] utilized machine learning assisted-texture analysis of dual-energy CT to distinguish metastatic head and neck squamous cell carcinoma lymph nodes from lymphoma, inflammatory, or normal lymph nodes. However, these radiomic-based methods were usually time consuming and labor intensive with a semiautomatic workflow depending on hand-crafted features. Therefore, an automatic approach for learning adequate information from medical image data in a way that exceeds human capabilities is needed.

Despite the prior progress made, there are still two challenges in PET/CT-based multimodal data analysis for predicting PTF in lymphoma. (1) DLBCL lesions are characterized by obvious heterogeneity, and thus designing an appropriate model for complementary feature learning may further enhance prediction accuracy at EoT. To describe high-level semantic features of PET/CT, a multimodal deep learning (MDL) model is needed; such an approach could improve both the investigation of PET and CT complementary characteristics and the interpretation of treatment outcomes. Unfortunately, prior deep learning-based studies on multimodal medical image analysis mostly adopted simple input-level concatenating [20, 21] or output-level averaging [22] as the learning approach of complementary features, and some subtle structural information may be lost in these approaches. (2) On the other hand, image volume has a special 3D structure regarded as a sequence of 2D consecutive slices in many medical analysis [23–25]. Inter-slice context differentiated from intra-slice semantic information is memorized and propagated along the z axis. Opposite to the model with isotropic fully connection after convolution, Donahue *et al.* [26] combined long-range temporal recursion to convolutional layers and enabled a novel end-to-end model named Conv-LSTM for optimized feature mapping and better visual description. This study brought a new inspire in medical image volume processing, but the above limitation is often ignored in constructing a deep learning model specialized in diagnosis, staging, and outcome prediction.

A reliable model to predict PTF in low-risk DLBCL would guide treatment optimization, thereby improving efficacy and long-term survival. Therefore, we built a powerful backbone by redesigning the Conv-LSTM and further developed and validated multiple MDL models using different feature fusion strategies based on ^{18}F-FDG PET/CT data, including the pixel intermixing model, separate channel model, separate branch model, quantitative weighting model, and hybrid learning model. The model with the best performance was ultimately trained by a contrastive training objective so as to attain the best accuracy. According to what we know, our current investigation of PTF prediction in DLBCL patients is the first such investigation that uses the PET/CT-based MDL approaches. Our results indicate that prediction of the best accuracy to date has been achieved by the present hybrid learning model that employs contrastive objective training. Apparently, our work is capable of securing a noninvasive and accurate method that indicates possible PTF before EoT and promotes DLBCL individualized treatment strategies.

2. Materials And Methods

Patients and dataset

All patients were collected from Ruijin Hospital (Shanghai, China) and were part of a consecutively observational DLBCL cohort from June 2016 to November 2020, in accordance with the declaration of Helsinki. For this retrospective study, we first analyzed ^{18}F-FDG PET/CT data of 205 patients with de novo histologically confirmed DLBCL according to the World Health Organization 2016 classification, of no more than one risk factor according to IPI. A complete flow of data collection is shown in Fig. 1. We excluded patients who (a) underwent surgical resection of all tumor lesions before immunochemotherapy and included all patients in this frame who (b) had available interim ^{18}F-FDG PET/CT examination images after (c) receiving R-CHOP regimen, (d) with definite treatment outcome at EoT. Prior to analysis, the patients were divided into two groups for comparison: PTF and

non-PTF DLBCL. A total of 20 refractory patients, assigned to the PTF group, were defined by progression of disease during R-CHOP, or failure to achieve a complete response (CR) after at least 4 cycles. In the non-PTF group, 185 patients achieved complete metabolic response at EoT without relapse within 6 months of therapy. Treatment response was evaluated according to standardized criteria for non-Hodgkin lymphoma [27]. For model development, the patients were randomly divided into three subsets for training, validation, and test at a ratio of 3:1:1. Following the inclusion criteria of this study, we also supplemented 44 patients from January 2021 to July 2021 for the external validation. In addition, detailed clinical characteristics of all patients were collected, including age (median with interquartile rang [IQR]), its range (≤ 60 years versus > 60 years), gender, IPI (0 versus 1), stage (I–II versus III–IV), eastern cooperative oncology group (ECOG) performance status (0 versus 1), serum lactate dehydrogenase (LDH) level (normal versus elevated), extra-lymphatic involvement, and B-symptoms.

Image acquisition and preprocessing

Image data were acquired from a PET/CT scanner (GE Healthcare, Waukesha, Wisconsin, USA) with the reconstruction method of ordered subset expectation maximization. Each sample contained one CT volume with the resolution of 512 × 512 pixels at 0.98 mm × 0.98 mm and one PET volume with the resolution of 128 × 128 pixels at 5.47 mm × 5.47 mm. Both volumes were reconstructed with the same number of slices, and the inter-distance was 3.27 mm. A standard routine in the first step of dual-modal image preprocessing was a rigid-body registration to eliminate the misalignment in coordinate spaces between PET and CT volumes [28, 29]. Next, the aligned image data were rescaled to the same resolution of 64 × 64 × 32 pixels using bicubic interpolation to reduce computational burden and facilitate model training. Furthermore, PET data were normalized by a transformation to the standard uptake value (SUV); this process was based on the radionuclide total dose of FDG and the weight of each patient [30].

MDL model developing

An overview of our framework for PTF-DLBCL prediction is shown in Fig. 2. The starting point of our model is Conv-LSTM [26], a classic deep learning architecture for natural image recognition and description, that has been recently applied to medical image analysis such as emphysema pattern classification in CT scans and achieved superior performance over traditional radiomics approaches [25]. The powerful network backbone for our model (Fig. 2a) was built from a redesigned Conv-LSTM in aspects of network architecture by constructing two identical encoders for PET and CT data, respectively. To extract hidden image features of input data, four blocks of convolution and pooling operations were conducted. Then, with the introduction of recursive learning framework, it had a structure called “long-short term memory” (LSTM) [31, 32], which performed simple learned gating functions to allow learning parameters to be updated or reset. Above extracted features were concatenated into a sequence, which was then transformed by the LSTM into a composite feature vector for the sample. Thanks to it, complex and heterogeneous information of input data were derived to high-level semantic features reflecting intra-slice spatial structures and inter-slice contextual correlations. The output of the model was a set of two continuous variables representing the prediction probability (on the scale of 0.0 to 1.0) for each category and was treated as a discrete probability distribution. The final prediction was calculated as the probability-weighted average of the categories rounded to the nearest integer.

On top of above improved backbone, multiple MDL models using different feature fusion strategies were developed and compared (Fig. 2b), including the pixel intermixing model (I), separate channel model (II), separate branch model (III), quantitative weighting model (IV), and hybrid learning model (V). The first model (I) is the only input-level kind distinguished from other feature-level fusion approaches. Here, a PET slice and its corresponding CT slice were integrated as one input image via pixel intermixing for single-branch encoding [33]. Second (II), PET and CT data were read into one encoder by separate channels and were simply concatenated after the first group of convolution and pooling operations [34]. As for the third model (III), the output feature maps from PET and CT separate encoding branches were concatenated before fed into the following LSTM predictor [35]. Forth (IV), the model learned the spatial contribution of feature maps from PET and CT encoders by a quantitative weighting strategy which calculated the convolutional result as a weighted matrix [36]. In the last model (V), PET and CT features extracted from two identical encoders were combined by the hybrid learning approach, a modal fusion method we published before [37], which generated spatial fusion maps and quantified the contribution of complementary information.

These fusion maps were then concatenated with specific-modality (i.e. PET and CT) feature maps to obtain a representation of the final-fused feature maps in different scales.

To achieve a better performance, we further aimed to promote the intra-class cohesion and inter-class separation of the semantic embeddings of PTF and non-PTF cases. Thus, we adopted the contrastive learning [38] in the hybrid learning model to achieve that goal. Specifically, a cross-entropy of the prediction and ground truth (Fig. 2c) was integrated with a contrastive training objective (Fig. 2d) derived from the similarity between a pair of samples to generate the overall loss function. Trained in this way, the contrastive hybrid learning model (VI) will be enhanced because the same class lay close to each other regardless of the modal heterogeneity of data source domain, and away from those in different classes. Details of constructing the overall training objective were described in Sec. 1 of the supplementary materials.

Model implementation and visualization

We implemented MDL models using TensorFlow 1.14 [39] on a machine running Windows 10 with CUDA 10.0 and cuDNN 7.6 [40]. Model training was performed on an 11 GB NVIDIA GeForce RTX2080 Ti. Our python code can be found at https://github.com/cyuan-sjtu/MDL-model. With an eye of reducing the overfitting impact, we employed the data enhancement strategy in training the data of each-fold cross validation experiment, wherein we included random horizontal and vertical flipping of the input images. To establish appropriate training parameters, we employed the values of 0.1, 0.01, and 4, respectively for the parameters of the regularization factor, the learning rate, and the batch size. We used Sec. 2 of the supplementary materials to present detailed architecture parameters built in the MDL models.

For each sample, model attention can be visualized for physician comprehension and validation. Here, focused regions of hidden-layer feature maps were instructed in the form of rough location heat map [41], which highlighted the entry area of prediction targets and interpreted the explanatory nature of such models about what kinds of features contributed to outputs. Through this way, complex features that passed deep convolutional and pooling layers were projected on the original input image.

Statistical analysis

To reveal the difference between clinical characteristics of the PTF group and the non-PTF one, we utilized the statistical package SPSS version 22.0 for various aspects of univariate analysis, including Mann-Whitney U test for numerical variables and Pearson's chi-squared test and, if necessary, Fisher's exact tests for categorical variables. Here, p-values of < 0.05 were considered as usual to be statistically significant. Prediction results were drawn from a 5-fold cross-validation, where samples of dataset 1 were divided into three cohorts for training, validating, and testing of models at ratios of 3:1:1. Samples of dataset 2 were only tested for external validation. The main metrics used to evaluate the model performance were accuracy as well as sensitivity (True Positive Rate [TPR]), specificity (True Negative Rate [TNR]), and the positive and negative predictive values (PPV and NPV). Sec. 3 of the supplementary materials lists the specific formulas for these evaluation metrics. Corresponding 95% confidence intervals (CIs) of the various variables were calculated based on the Wilson score interval. In addition, we plotted the confusion matrices and calculated the area under the receiver operating characteristic curves (AUCs) to assess the discrimination ability and predictive accuracy of multiple MDL models. Moreover, we compared the AUCs of various MDL models utilizing the Delong test.

3. Results

Baseline clinical characteristics

From June 2016 to November 2020, 205 low-risk DLBCL patients (median age: 55.00 years, 95 females, 110 males) were collected for model development and assigned to the primary dataset. Besides them, the data of 44 patients (median age: 54.50 years, 23 females, 21 males) were included for prospective validation and assigned to the external dataset. Table 1 displayed the baseline characteristics of patients in the primary and external dataset. The PTF rates were 9.76% (20/205) and 9.10% (4/44), respectively, showing a similarity between the two datasets. The percentages of PTF patients and non-PTF patients who

possessed one IPI risk factor each were 95.00% and 54.05% (100/185), respectively, which demonstrates the existence of a significant difference between the two cohorts in the primary dataset ($p < 0.001$).

Performance comparison of MDL models

The prediction performances of MDL models in the test cohort and external validation cohort were listed in Table 2. All four MDL models using feature-level fusion strategies (separate channel model, separate branch model, quantitative weighting model, and hybrid learning model) provided better accuracy than the pixel intermixing model in the test and external validation cohorts, the only one using input-level fusion strategy. Due to the class imbalance between PTF and non-PTF patients, PPVs of all MDL models were relatively low, which we still considered as a meaningful result. Overall speaking, the hybrid learning model achieved the best performance among all evaluation metrics (sensitivity = 65.00% [95% CI: 43.29–81.88%], specificity=92.43% [95% CI: 87.70–95.44%], accuracy=89.76% [95% CI: 84.85–93.20%], PPV=48.15% [95% CI: 30.74–66.01%], and NPV=96.07% [95% CI: 92.11–98.08%]) in the test cohort for predicting PTF. This indicated that the hybrid learning feature fusion strategy increased both the ratio of true positives to false positives and that of true negatives to false negatives. Fig. 3 shows that the AUC was 0.837 in the test cohort and 0.869 in the external validation cohort. Although the quantitative weighting model achieved a better AUC of 0.844 in the test cohort, we have taken the external validation cohort for the most important indicator, and, therefore, we integrated the contrastive training objective with the hybrid learning model and established the contrastive hybrid learning model.

The contrastive hybrid learning model achieved AUCs of 0.926 and 0.925 in the test cohort and the external validation cohort, respectively. In addition, the Delong test demonstrated that the AUC of this model was significantly better than that of each of the pixel intermixing model ($p < 0.001$), separate channel model ($p < 0.05$), separate branch model ($p < 0.05$), and quantitative weighting model ($p < 0.05$) in the test cohort. Quantitative comparisons of other five evaluation metrics are exhibited in Table 2. Compared with the quantitative weighting method, which has shown good performances in computer aided diagnosis [36], its predictive accuracy, sensitivity, specificity, and predictive values were all improved. In addition, the normalized confusion matrices of all MDL models in distinguishing PTF from non-PTF in the test cohort were shown in Fig. 4. Notably, the contrastive hybrid learning model achieved continuous improvements in the overall sensitivity for PTF groups, as indicated from subfigure 4a to 4f.

Interpretability of MDL models

For each sample, model attention can be visualized for clinical comprehension and validation. Here, we aimed to understand which areas of input images and what kinds of features contributed to the prediction. Fig. 5 shows input PET/CT images and the heat maps of corresponding locations for three patients randomly chosen from the test cohort, which demonstrates the existence of a common pattern that is consistently shared among all samples. The contrastive hybrid learning model paid great attention (highlighted in red) to the structure of lesion from physiological uptake interference of the heart and bones. This suggested that this model actively sought tumor lesion distribution areas to classify PTF and non-PTF. Notably, anatomic features derived from CT-modality data were meaningful, even though radiologists mostly referred PET information in clinical diagnosis. In addition, the compared models viewed different lesion adjacent areas for the same patient, a discrepancy which explains why these models differed in the prediction performance attained by each of them. Specifically, the heat map of the contrastive hybrid learning model contained more gradient attention on specific regions related to tumor-self, except necrosis and peripheral inflammation [42].

4. Discussion

The prediction of PTF for DLBCL patients has been a prominent challenge facing clinicians for a long time. In this work, we developed and validated a deep learning model that learned complementary high-level semantic features from interim ^{18}F-FDG PET/CT images and achieved individualized and noninvasive prediction of PTF in patients with low-risk DLBCL at EoT. The major findings of our experiments covered the following issues: (1) As far as we know, our work is a seminal pioneering one, apparently the first that applies the MDL approach on interim PET/CT images acquired from DLBCL patients for predicting PTF. (2) The prediction performance of the present MDL model, based on both the incorporation of hybrid learning feature fusion

strategy and the enhancement of contrastive training objective, significantly outperformed (almost all AUCs, $p < 0.05$) other models using different feature fusion strategies in ablation comparisons. (3) Our work provides solid evidence that the contrastive hybrid learning approach on PET/CT images secured an effective method for the prediction of PTF and the stratification of risk for patients suffering from DLBCL.

From the clinical point of view, the tumor involvement changes calculated from pathological FDG uptake in PET imaging can indicate possible PTF, but its use remains limited because of its dependence on data both prior to and after treatment [43]. Moreover, the identification of CT parameters in the tumor region has been demonstrated to be an effective discrimination technique for predicting the PTF of different carcinomas [14, 15]. However, present models on extracting and combining invisible imaging features of PET and CT are not still adequate for predicting PTF of DLBCL. In the current work, we proposed several MDL models based on ^{18}F-FDG PET/CT to predict PTF, and we investigated ways of fully utilizing the PET/CT data so as to achieve the best performance. Compared with using the input-level fusion strategy for identifying PTF of DLBCL, other MDL models with various feature-level fusion strategies actively indicated more accurate tumor involvement areas in PET/CT images. The hybrid learning model demonstrated particularly outstanding predictive performance for PTF, as it effectively integrated PET metabolic features with corresponding CT anatomic features.

By optimizing the hybrid learning model with the contrastive training objective enhancement, the proposed contrastive hybrid learning model significantly outperformed all the other MDL prediction models in AUC (almost all, $p < 0.05$). By contrast, the quantitative weighting model implemented the convolutional result from a fusion unit as a weighted matrix that was then multiplied by the PET and CT feature maps [36]. Element multiplication was used to encompass the level of importance given to information from each modality, although it considerably weakened the natural characteristics of each modality. The separate branch strategy was widely used in medical data analysis especially with different-class information [35]. As shown in Table 2, the separate branch model was second to the quantitative weighting model in terms of sensitivity, specificity, and AUC. It included a layer-level fusion strategy based on simple concatenation and then applied prediction layers. Thus, some useful information associated with complementary features may be lost. In addition, the separate channel model combined the PET and CT images resulting after the first convolutional layer to derive fused feature maps [34], which led to lopsided attention to a modality with dominant pixel strength. The pixel intermixing strategy was used to construct a type of early fusion model [33]. Here, the PET and CT images were initially fused via pixel intermixing and the intermixed images were used as model inputs. This approach shared a similar weakness with the separate channel model, and such a limitation reduced the prediction accuracy.

To interpret the MDL model, we visualized the ROIs of network by generating rough location heat maps. The activated areas in the heat map were found to be primarily located in the tumor lesion and its surrounding areas according to several MDL models. All these areas were consistent with the predictive region observed by experienced radiologists. Notably, the contrastive hybrid learning model paid more precise attention meanwhile avoiding interference of physiological uptake. These common patterns served as a clue for the working principle of MDL models for analyzing PET/CT data.

Although our results are definitely promising, our present work has a few limitations, thereby leaving room for several future improvements. First, this work is a retrospective study that is based on a relatively small sample size, especially on positive samples. Although large numbers of PET-CT scans are difficult to obtain, the addition of in-house data would definitely be of paramount importance to the current work. We are actively working on this task. Notably, after calibration and targeted optimization, deep learning models would have the natural advantages of stability, repeatability, and ease of migration. As more PET/CT images from other institutions are supplemented for model developing, our MDL model is quite likely to be applied to data from different institutions and would achieve better generalizing capabilities. Second, to indicate accurate PTF for planning personalized treatments, the sensitivity and accuracy should be very high. Thus, the MDL model provided a reference result but not a direct decision for clinical practice, given the need for a sufficiently high PPV and prospective validation. Finally, we did not consider other possible prognostic factors. We suggest that integrating biologic markers including blood biomarkers and pathological and genetic features, may improve the accuracy and robustness of our model.

5. Conclusions

Our work developed several MDL models, and the best model integrated hybrid learning strategy with contrastive training objective enhancement that exhibited satisfactory performance in predicting PTF for patients suffering from DLBCL. It enabled complementary information generation and feature adaptation in multimodal learning. Adequate performance in ablation experiments proved the effectiveness and superiority of the contrastive hybrid learning model. In addition, it is end-to-end trainable and avoids the need for radiomics experience and time-consuming manual delineations. Therefore, it provides a proof-of-concept for multimodal data analysis and further helps clinicians for individualized decision-making in DLBCL clinical practice.

Declarations

Acknowledgements

This study was supported, in part, by research funding from the National Natural Science Foundation of China (81974276, 81830007, 81520108003, 81670176, and 82070204), Chang Jiang Scholars Program, Shanghai Municipal Education Commission Gaofeng Clinical Medicine Grant Support (20152206, and 20152208), Clinical Research Plan of Shanghai Hospital Development Center (SHDC2020CR1032B), Multicenter Clinical Research Project by Shanghai Jiao Tong University School of Medicine (DLY201601), Collaborative Innovation Center of Systems Biomedicine, and the Samuel Waxman Cancer Research Foundation.

Author contribution

Conception and design: Cheng Yuan, Qing Shi, Li Wang, Weili Zhao, Dahong Qian.

Code: Cheng Yuan.

Data analysis: Cheng Yuan, Qing Shi, Xinyun Huang.

Data acquisition: Qing Shi, Xinyun Huang, Li Wang, Yang He, Biao Li, Weili Zhao.

Manuscript draft: Cheng Yuan, Qing Shi.

Manuscript revision: Cheng Yuan, Qing Shi, Xinyun Huang, Li Wang, Yang He, Biao Li, Weili Zhao, Dahong Qian.

Code availability

The code of our study is publicly accessible at https://github.com/cyuan-sjtu/MDL-model.

Ethics approval

This retrospective study was approved by the Review Board of Shanghai Ruijin Hospital with informed consent obtained from all patients following the principles of the Declaration of Helsinki.

Conflict of interest

The authors declare no competing interests.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49. https://doi.org/10.3322/caac.21660
2. Feugier P, Van Hoof A, Sebban C, Solal-Celigny P, Bouabdallah R, Christian FC, et al. Long-term results of the R-CHOP study in the treatment of elderly patients with diffuse large B-cell lymphoma: a study by the Groupe d'Etude des Lymphomes de l'Adulte. J Clin Oncol. 2005;23:4117–26. https://doi.org/10.1200/JCO.2005.09.131

3. Crump M, Neelapu SS, Farooq U, Van Den Neste E, Kuruvilla J, Westin J, et al. Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. Blood. 2017;130(16):1800-8. https://doi.org/10.1182/blood-2017-03-769620
4. Gisselbrecht C, Van Den Neste E. How I manage patients with relapsed/refractory diffuse large B cell lymphoma. Br J Haematol. 2018;182:633-43. https://doi. org/10.1111/bjh.15412
5. Sehn LH, Chhanabhai M, Fitzgerald C, Gill K, Hoskins P, Klasa R, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. Blood. 2007;190:1857–61. https://doi.org/ 10.1182/blood-2006-08-038257
6. Xu-Monette ZY, Wu L, Visco C, Tai YC, Tzankov A, Liu W, et al. Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an international DLBCL rituximab-CHOP consortium program study. Blood. 2012;120(19):3986–96. https://doi.org/10.1182/blood-2012-05-433334
7. Fuglø HM, Jørgensen SM, Loft A, Hovgaard D, Petersen MM. The diagnostic and prognostic value of ^{18}F-FDG PET/CT in the initial assessment of high-grade bone and soft tissue sarcoma. a retrospective study of 89 patients. Eur J Nucl Med Mol Imaging. 2012;39(9):1416-24.
8. Guo Y, Wang Q, Guo Y, Zhang Y, Zhang H. Preoperative prediction of perineural invasion with multi-modality radiomics in rectal cancer. Sci Rep. 2021;11.
9. Liu Y, Fan L, Zhang C, Zhou T, Shen D. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med Image Anal. 2021;69:101953.
10. Kahle XU, Hovingh M, Noordzij W, Seitz A, Diepstra A, Visser L, et al. Tumour necrosis as assessed with ^{18}F-FDG PET is a potential prognostic marker in diffuse large B cell lymphoma independent of MYC rearrangements. Eur Radiol. 2019;29:6018-28. https://doi.org/10.1007/s00330-019-06178-9
11. Schöder H, Polley MC, Knopp MV, Hall N, Kostakoglu L, Zhang J, et al. Prognostic value of interim FDG-PET in diffuse large cell lymphoma: results from the CALGB 50303 clinical trial. Blood. 2020;135(25):2224-34. https://doi. org/10.1182/blood.2019003277
12. Santiago R, Jimenez JO, Forghani R, Muthukrishnan N, Corpo OD, Karthigesu S, et al. CT-based radiomics model with machine learning for predicting primary treatment failure in diffuse large B-cell Lymphoma. Translational Oncology. 2021;14(10). https://doi.org/10.1016/j.tranon.2021.101188.
13. Senjo H, Hirata K, Izumiyama K, Minauchi K, Tsukamoto E, Itoh K, et al. High metabolic heterogeneity on baseline ^{18}FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma, Blood Adv. 2020;4(10):2286–96. https://doi.org/10.1182/bloodadvances.2020001816
14. Seidler M, Forghani B, Reinhold C, Pérez-Lara A, Romero-Sanchez G, Muthukrishnan N, et al. Dual-energy CT texture analysis with machine learning for the evaluation and characterization of cervical lymphadenopathy. Computational and Structural Biotechnology Journal. 2019;17:1009-15. https://doi.org/10.1016/j.csbj.2019.07.004
15. Ganeshan B, Miles KA, Babikir S, Shortman R, Afaq A, Ardeshna KM, et al. CT-based texture analysis potentially provides prognostic information complementary to interim FDG-PET for patients with Hodgkin's and aggressive non-Hodgkin's lymphomas. Eur Radiol. 2017;27:1012–20. https:// doi.org/10.1007/s00330-016-4470-8
16. Adams HJA, Klerk JMH, Fijnheer R, Dubois SV, Nievelstein RAJ, Kwee TC. Prognostic value of tumor necrosis at CT in diffuse large B-cell lymphoma. Eur J Radiol. 2015;84(3):372–7. https://doi.org/10.1016/j.ejrad.2014.12.009
17. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Aerts JHWL, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;43(4):441-6. https:// doi.org/10.1016/j.ejca.2011.11.036
18. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563-577. https:// doi.org/10.1148/radiol.2015151169
19. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. https://doi.org/10.1038/ncomms5006

20. Bi L, Kim J, Kumar A, Wen L, Feng D, Fulham M. Automatic detection and classification of regions of FDG uptake in whole-body PET-CT lymphoma studies. Comput Med Imaging Graph. 2017;60:3-10.
21. Jin C, Yu H, Ke J, Ding P, Yi Y, Jiang X, et al. Predicting treatment response from longitudinal images using multi-task deep learning. Nat Commun. 2021;12:1851. https://doi.org/10.1038/s41467-021-22188-y
22. Hu H, Shen L, Zhou T, Decazes P, Su R. Lymphoma segmentation in PET images based on multi-view and Conv3D fusion strategy. In Proc. International Symposium on Biomedical Imaging (ISBI). Apr 2020.
23. Cai J, Lu L, Xie Y, Xing F, Yang L. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. In Proc. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Sep 2017.
24. Zhang L, Lu L, Wang X, Zhu RM, Bagheri M, Summers RM. Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data. IEEE Transactions on Medical Imaging. 2019;39(4):1114-26.
25. Humphries SM, Notary AM, Centeno JP, Strand MJ, Crapo JD, Silverman EK, et al. Deep learning enables automatic classification of emphysema pattern at CT. Radiology. 2019;294(2). https://doi.org/10.1148/radiol.2019191022
26. J. Donahue, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(4):677-91. https://doi.org/10.1109/TPAMI.2016.2599174
27. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, 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):3059-3068. https://doi.org/10.1200/JCO.2013.54.8800
28. Zhong Z, Kim Y, Plichta Y, Bryan GA, Zhou L, Buatti JM, et al. Simultaneous co-segmentation of tumors in PET-CT images using deep fully convolutional networks. Medical Physics. 2019;46(2):619-33. https://doi.org/10.1002/mp.13331
29. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Physics in Medicine Biology. 2018;64(1). https://doi.org/10.1088/1361-6560/aaf44b
30. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. J Nucl Med. 2004;45(9). https://doi.org/10.1016/j.nuclcard.2004.07.002
31. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-80. https://doi.org/10.1162/neco.1997.9.8.1735
32. Lu N, Wu Y, Feng L, Song J. Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data. IEEE Journal of Biomedical and Health Informatics. 2019;23(1):314-23.
33. Zhong Z, Kim Y, Zhou L, Plichta K, Allen B, Buatti J, et al. 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In Proc. International Symposium on Biomedical Imaging (ISBI). 2018. https://doi.org/10.1109/ISBI.2018.8363561.
34. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage. 2015;108:214-24. https://doi.org/10.1016/j.neuroimage.2014.12.061.
35. Peng Y, Bi L, Guo Y, Feng D, Kim J. Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies. In Proc. Engineering in Medicine and Biology Society (EMBC). Oct 2019.
36. Kumar A, Fulham MJ, Feng D, Kim J. Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Transactions on Medical Imaging. 2020;39(1):204-17.
37. Yuan C, Zhang M, Huang X, Xie W, Lin X, Zhao W, et al. Diffuse large B-cell lymphoma segmentation in PET-CT images via hybrid learning for feature fusion. Medical Physics. 2021;48(7):3665-78. https://doi.org/10.1002/mp.14847
38. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. 2020. arXiv:2002.05709.
39. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In Proc. USENIX conference on Operating Systems Design and Implementation. Nov 2016.

40. Chetlur A, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, et al. cuDNN: efficient primitives for deep learning. 2014. arXiv: arXiv:1410.0759.
41. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In Proc. Conference on Computer Vision and Pattern Recognition (CVPR). Jun 2016.
42. Du D, Feng H, Lv W, Ashrafinia S, Yuan Q, Wang Q, et al. Machine learning methods for optimal radiomics-based differentiation between recurrence and inflammation: application to nasopharyngeal carcinoma post-therapy PET/CT images. Mol Imaging Biol. 2020;22:730–8. https://doi.org/10.1007/s11307-019-01411-9

Tables

Table 1. Baseline clinical characteristics of the primary and external dataset

Variable	Primary dataset (n=205)		*p*-value	External dataset (n=44)		*p*-value
	non-PTF (n=185)	PTF (n=20)		non-PTF (n=40)	PTF (n=4)	
Age (y), median (IQR)	55.00 (45.50–63.00)	57.50 (42.00–62.75)	0.671[a]	54.50 (44.25–62.25)	49.50 (40.75-59.00)	0.653[a]
Age range			0.162[b]			0.559[b]
≤ 60 years	130 (70.27%)	11 (55.00%)		30 (75.00%)	4 (100.00%)	
> 60 years	55 (29.73%)	9 (45.00%)		10 (25.00%)	0 (0.00%)	
Gender			0.284[b]			0.335[b]
Female	88 (47.57%)	7 (35.00%)		22 (55.00%)	1 (25.00%)	
Male	97 (52.43%)	13 (65.00%)		18 (45.00%)	3 (75.00%)	
IPI			**<0.001**[b]			1.000[b]
0	85 (45.95%)	1 (5.00%)		19 (47.50%)	2 (50.00%)	
1	100 (54.05%)	19 (95.00%)		21 (52.50%)	2 (50.00%)	
Stage			1.000[b]			1.000[b]
I–II	174 (94.05%)	19 (95.00%)		33 (82.50%)	2 (50.00%)	
III–IV	11 (5.95%)	1 (5.00%)		7 (17.50%)	2 (50.00%)	
ECOG performance status			0.293[b]			1.000[b]
0	163 (88.11%)	16 (80.00%)		35 (87.50%)	4 (100.00%)	
1	22 (11.89%)	4 (20.00%)		5 (12.50%)	0 (0.00%)	
LDH level			0.543[b]			1.000[b]
Normal	152 (82.16%)	15 (75.00%)		36 (90.00%)	4 (100.00%)	
Elevated	33 (17.84%)	5 (25.00%)		4 (10.00%)	0 (0.00%)	
Extra-lymphatic involvement	76 (41.08%)	4 (20.00%)	0.066[b]	24 (60.00%)	2 (50.00%)	1.000[b]
B-symptoms	23 (12.43%)	3 (15.00%)	0.725[b]	2 (5.00%)	0 (0.00%)	1.000[b]

[a] Mann-Whitney U test, [b] Pearson's chi-squared test, if necessary, Fisher's exact test.

Note that data are number of patients; data in parentheses are percentage.

Table 2. The performance comparison of multiple MDL models

Models	Cohorts	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Model I Pixel intermixing	Test	40.00 (21.88–61.34)	86.49 (80.81–90.68)	81.95 (76.11–86.61)	24.24 (12.83–41.03)	93.02 (88.20–95.96)	0.774
	External validation	50.00 (15.00–85.00)	77.50 (62.50–87.68)	75.00 (60.56–85.43)	18.18 (5.14–47.70)	94.00 (80.39–98.32)	0.688
Model II Separate channel	Test	45.00 (25.82–65.79)	87.03 (81.42–91.13)	82.93 (77.18–87.46)	27.27 (15.07–44.22)	93.60 (88.91–96.39)	0.794
	External validation	50.00 (15.00–85.00)	85.00 (70.93–92.94)	81.82 (68.04–90.49)	25.00 (7.15–59.07)	94.44 (81.86–98.46)	0.706
Model III Separate branch	Test	50.00 (29.93–70.07)	90.27 (85.15–93.76)	86.34 (80.97–90.38)	35.71 (20.71–54.17)	94.35 (89.91–96.90)	0.810
	External validation	50.00 (15.00–85.00)	85.00 (70.93–92.94)	81.82 (68.04–90.49)	25.00 (7.15–59.07)	94.44 (81.86–98.46)	0.725
Model IV Quantitative weighting	Test	45.00 (25.82–65.79)	91.35 (86.41–94.61)	86.83 (81.52–90.79)	36.00 (20.25–55.48)	93.89 (89.39–96.55)	0.844
	External validation	50.00 (15.00–85.00)	87.50 (73.89–94.54)	84.09 (70.63–92.07)	28.57 (8.22–64.11)	94.59 (82.30–98.50)	0.766
Model V Hybrid learning	Test	65.00 (43.29–81.88)	92.43 (87.70–95.44)	89.76 (84.85–93.20)	48.15 (30.74–66.01)	96.07 (92.11–98.08)	0.837
	External validation	75.00 (30.06–95.44)	87.50 (73.89–94.54)	86.36 (73.29–93.60)	37.50 (13.68–69.43)	97.22 (85.83–99.51)	0.869
Model VI Contrastive hybrid learning	Test	**70.00 (48.10–85.45)**	**93.51 (89.01–96.25)**	**91.22 (86.55–94.37)**	**53.85 (35.46–71.24)**	**96.65 (92.88–98.45)**	**0.926**
	External validation	**75.00 (30.06–95.44)**	**90.00 (76.95–96.04)**	**88.64 (76.02–95.05)**	**42.86 (15.82–74.95)**	**97.30 (86.18–99.52)**	**0.925**
Comparison between the contrastive hybrid learning model and other models using Delong test							
Compared method			Model I	Model II	Model III	Model IV	Model V
Test cohort			< 0.001	0.006	0.020	0.028	0.051
External validation cohort			0.030	0.021	0.054	0.018	0.088

The best metrics are shown in the bold numbers. Numbers in parentheses are 95% confidence intervals.

Figures

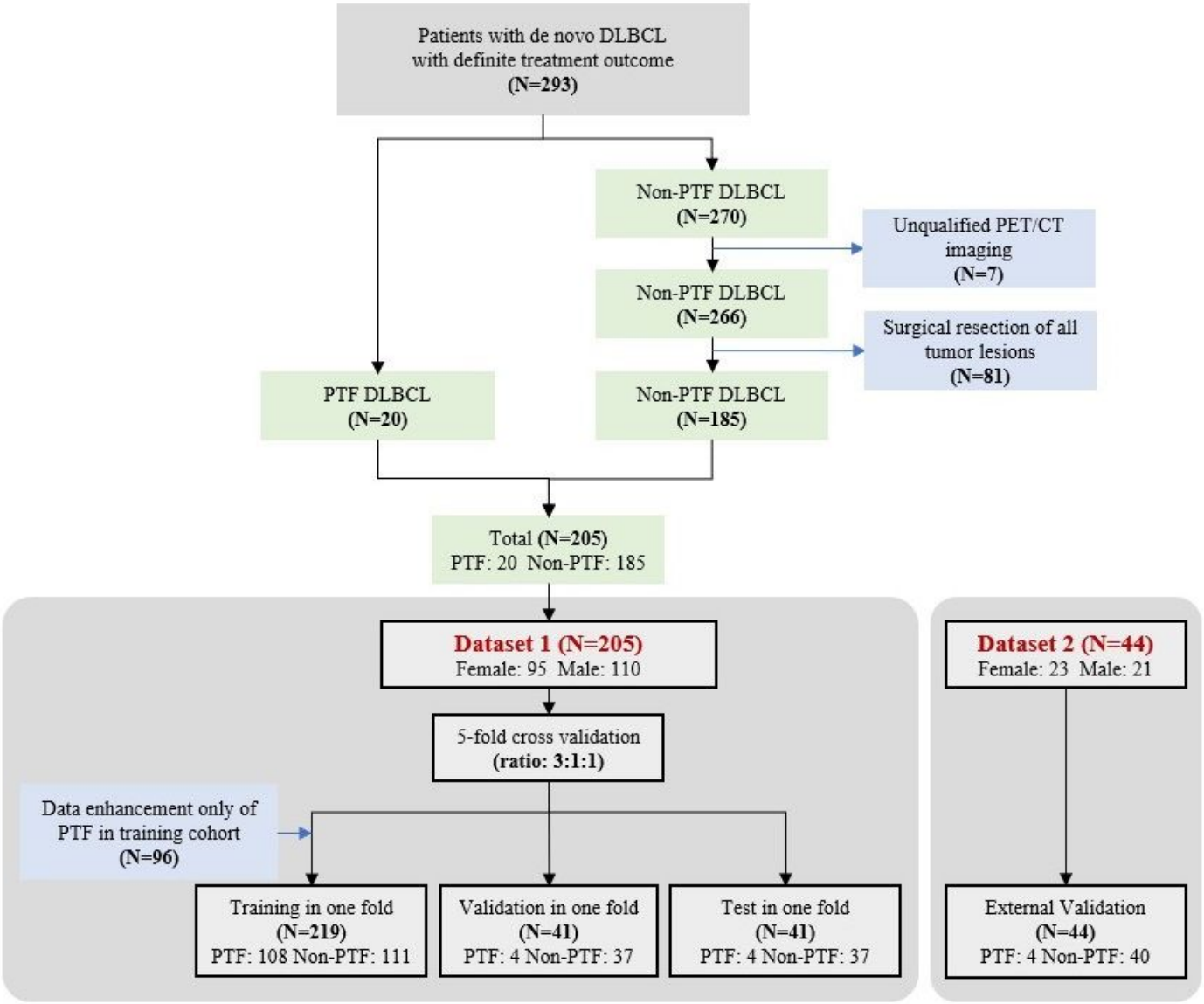

Figure 1

The complete flow of data collection.

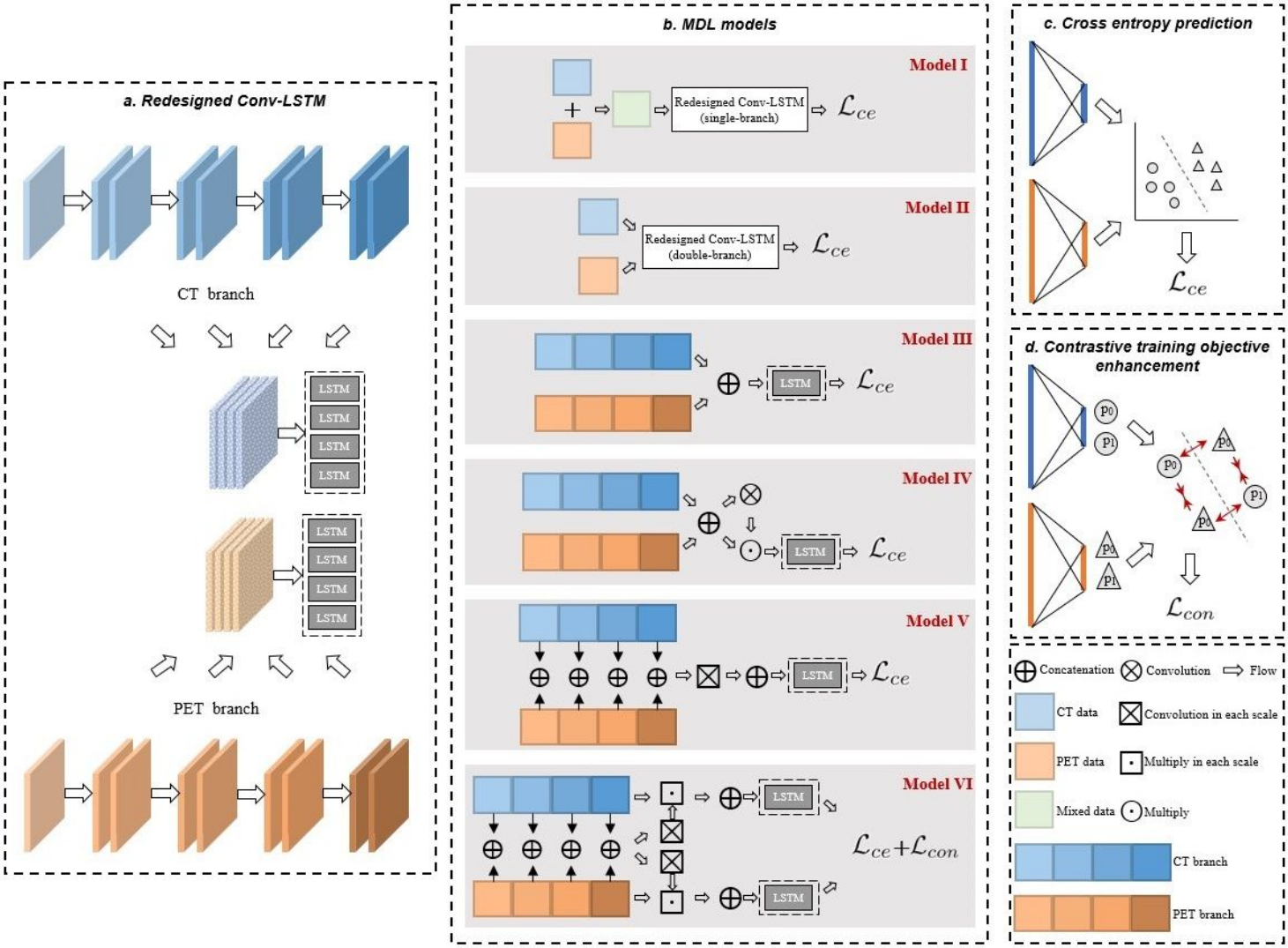

Figure 2

The workflow of MDL models based on PET/CT for predicting primary treatment failure (PTF) of patients suffering from diffuse large B-cell lymphoma (DLBCL). Legend is shown in the bottom right corner.

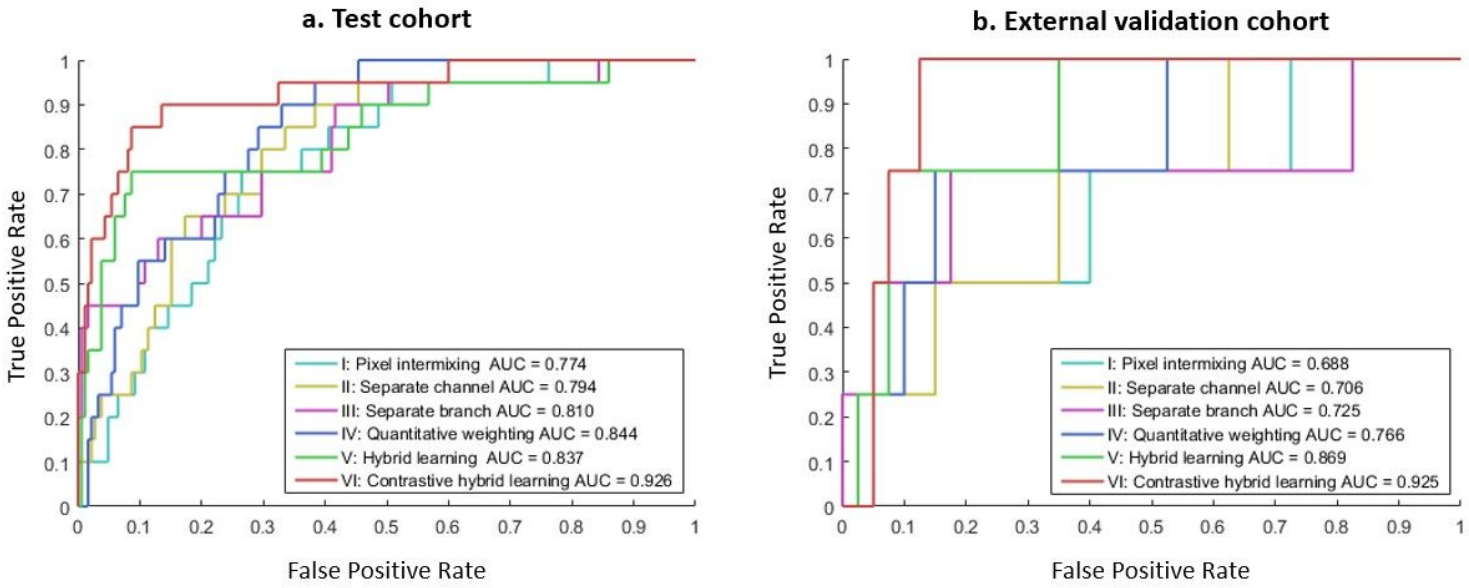

Figure 3

Receiver operating characteristic (ROC) curves of compared MDL models for predicting PTF of low-risk DLBCL. **a** Test cohort. **b** External validation cohort.

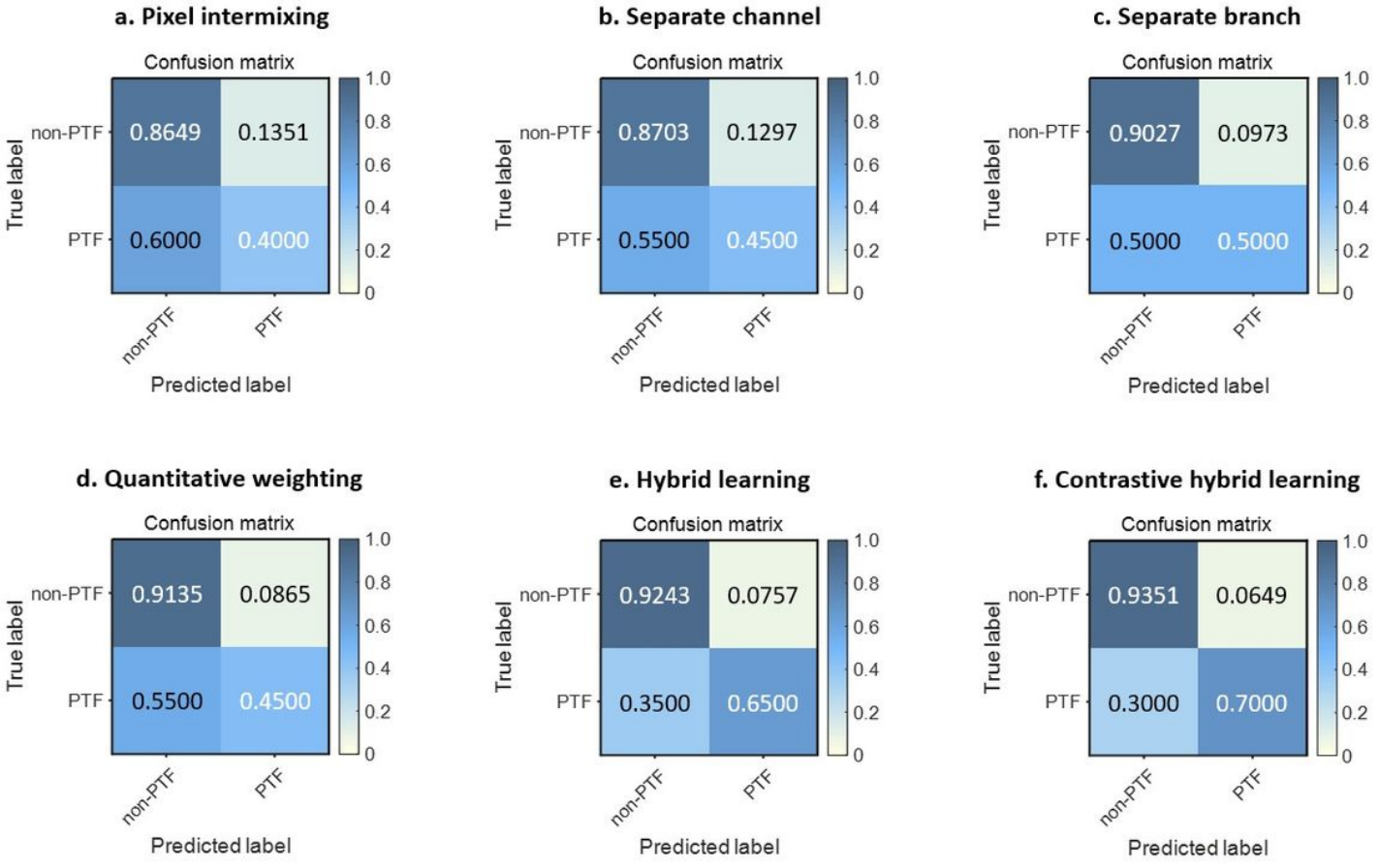

Figure 4

Confusion matrices of compared MDL models for predicting PTF of low-risk DLBCL in the test cohort.

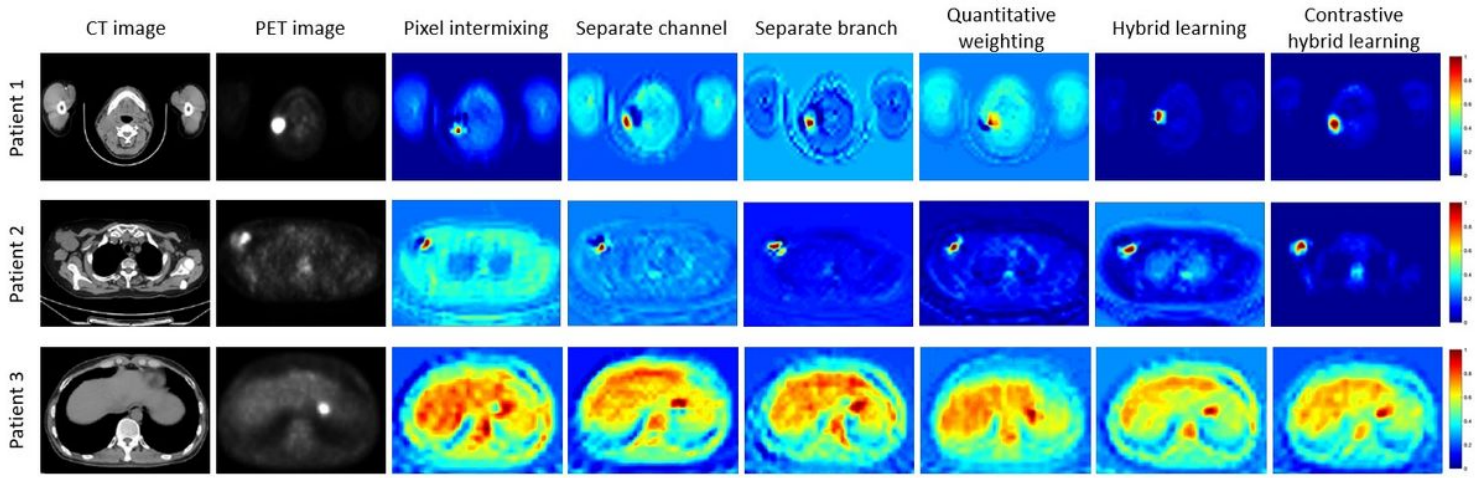

Figure 5

Visualization of three PTF-DLBCL patient examples. The activated regions are presented in red with a larger weight, which can be decoded by the color legend on the right.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterials.docx