Stacking Ensemble Learning–Based [^{18}F]FDG PET Radiomics for Outcome Prediction in Diffuse Large B-Cell Lymphoma

Shuilin Zhao*[1–4], Jing Wang*[1–3], Chentao Jin*[1–3], Xiang Zhang[1–3], Chenxi Xue[1–3], Rui Zhou[1–3], Yan Zhong[1–3], Yuwei Liu[1–3], Xuexin He[5], Youyou Zhou[1–3], Caiyun Xu[6], Lixia Zhang[6], Wenbin Qian[7], Hong Zhang[1–3,8,9], Xiaohui Zhang[1–3], and Mei Tian[1–3,10]

[1]*Department of Nuclear Medicine and PET Center, Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China;* [2]*Institute of Nuclear Medicine and Molecular Imaging of Zhejiang University, Hangzhou, China;* [3]*Key Laboratory of Medical Molecular Imaging of Zhejiang Province, Hangzhou, China;* [4]*Cancer Center, Department of Radiology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, China;* [5]*Department of Medical Oncology, Huashan Hospital of Fudan University, Shanghai, China;* [6]*Department of Nuclear Medicine, First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Traditional Chinese Medicine), Hangzhou, China;* [7]*Department of Hematology, Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China;* [8]*College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China;* [9]*Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou, China; and* [10]*Human Phenome Institute, Fudan University, Shanghai, China*

This study aimed to develop an analytic approach based on [^{18}F]FDG PET radiomics using stacking ensemble learning to improve the outcome prediction in diffuse large B-cell lymphoma (DLBCL). **Methods:** In total, 240 DLBCL patients from 2 medical centers were divided into the training set (n = 141), internal testing set (n = 61), and external testing set (n = 38). Radiomics features were extracted from pretreatment [^{18}F]FDG PET scans at the patient level using 4 semiautomatic segmentation methods (SUV threshold of 2.5, SUV threshold of 4.0 [SUV4.0], 41% of SUV_{max}, and SUV threshold of mean liver uptake [PERCIST]). All extracted features were harmonized with the ComBat method. The intraclass correlation coefficient was used to evaluate the reliability of radiomics features extracted by different segmentation methods. Features from the most reliable segmentation method were selected by Pearson correlation coefficient analysis and the LASSO (least absolute shrinkage and selection operator) algorithm. A stacking ensemble learning approach was applied to build radiomics-only and combined clinical–radiomics models for prediction of 2-y progression-free survival and overall survival based on 4 machine learning classifiers (support vector machine, random forests, gradient boosting decision tree, and adaptive boosting). Confusion matrix, receiver-operating-characteristic curve analysis, and survival analysis were used to evaluate the model performance. **Results:** Among 4 semiautomatic segmentation methods, SUV4.0 segmentation yielded the highest interobserver reliability, with 830 (66.7%) selected radiomics features. The combined model constructed by the stacking method achieved the best discrimination performance. For progression-free survival prediction in the external testing set, the areas under the receiver-operating-characteristic curve and accuracy of the stacking-based combined model were 0.771 and 0.789, respectively. For overall survival prediction, the stacking-based combined model achieved an area under the curve of 0.725 and an accuracy of 0.763 in the external testing set. The combined model also demonstrated a more distinct risk stratification than the International Prognostic Index in all sets (log-rank test, all $P < 0.05$). **Conclusion:** The combined model that incorporates [^{18}F]FDG PET radiomics and clinical characteristics based on stacking ensemble learning could enable improved risk stratification in DLBCL.

Key Words: PET; diffuse large B-cell lymphoma; prognosis; machine learning; radiomics

J Nucl Med 2023; 64:1603–1609
DOI: 10.2967/jnumed.122.265244

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of aggressive non-Hodgkin lymphoma. Rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone represents the current first-line treatment, which is effective in approximately 60%–70% of patients (*1*). Patients with refractory disease or relapse after initial treatment have a low probability of cure and dismal outcomes due to the modest response rates for salvage regimens (*2*). Therefore, early identification of those high-risk patients is essential for designing individualized therapeutic intervention. Current prognostic scoring systems, such as the International Prognostic Index (IPI) and the National Comprehensive Cancer Network–IPI, have been the basis for determining prognosis in DLBCL (*3,4*). However, those models are inaccurate in predicting refractory disease, possibly because of their lack of intratumoral metabolic and functional information.

[^{18}F]FDG PET/CT, a type of molecular imaging and a means to "transpathology" (*5*), has been recommended for staging and response assessment in DLBCL (*6,7*). Quantitative parameters on PET/CT, particularly total metabolic tumor volume (TMTV) and total lesion glycolysis, are considered to have prognostic significance in DLBCL (*8,9*). These parameters may allow for the assessment of whole-body tumor burden but remain limited in their ability to characterize phenotypical profiles such as shape, morphology, spatial distribution, and heterogeneity across individual lesions. For PET/CT image analysis, radiomics has recently been proposed as a novel high-throughput, noninvasive approach that could quantify tumor phenotype at a microscale level via extracting thousands of imaging-derived features (*10*). With the

Received Nov. 23, 2022; revision accepted May 31, 2023.
For correspondence or reprints, contact Mei Tian (tianmei@fudan.edu.cn) or Xiaohui Zhang (zhanghui4127@zju.edu.cn).
*Contributed equally to this work.
Published online Jul. 27, 2023.

assistance of artificial intelligence, such as machine learning, radiomics offers a promising tool for diagnosis, therapeutic response assessment, and outcome prediction in various tumor types (*11*), including DLBCL (*12–16*). Preliminary studies have suggested that the application of machine learning algorithms, such as LASSO (least absolute shrinkage and selection operator) regression (*16*), ridge regression (*13*), and random forest (*17*), may contribute to the improved radiomics feature selection and prognostic modeling in DLBCL. However, most of those studies focused on evaluating a single machine learning approach, whereas only a minority used cross combination of different machine learning algorithms (*14*) or adopted ensemble machine learning (*15*). Stacking, an ensemble approach that combines different base classifiers into 1 metaclassifier, has been suggested to provide optimized performance and simplicity (*18*). In the present study, we aimed to develop an analytic approach based on [^{18}F]FDG PET radiomics using stacking ensemble learning to improve the outcome prediction in DLBCL.

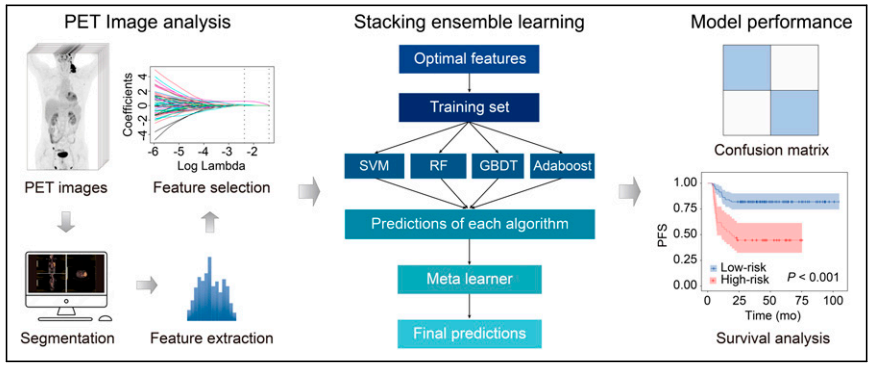

FIGURE 1. Radiomics workflow.

MATERIALS AND METHODS

Study Population

We retrospectively enrolled 240 consecutive patients with newly diagnosed DLBCL at 2 medical centers, including 202 patients at center 1 (the Second Affiliated Hospital of Zhejiang University School of Medicine) and 38 patients at center 2 (the First Affiliated Hospital of Zhejiang Chinese Medical University). Detailed information about the study population is shown in the supplemental materials (available at http://jnm.snmjournals.org) (*19,20*). The flowchart of patient enrollment is shown in Supplemental Figure 1. This study was approved by the Institutional Review Board at each institution, and the requirement to obtain written informed consent was waived.

PET/CT Imaging Protocol

Image acquisition and reconstruction were in accordance with the guidelines of European Association of Nuclear Medicine, version 2.0 (*21*). Patients fasted for at least 6 h and had a blood glucose level below 200 mg/dL before PET/CT examination. They were scanned at about 60 min after intravenous injection of [^{18}F]FDG (3.70 MBq/kg). All PET images were corrected for attenuation using acquired low-dose CT data. Acquisitions differed between the 2 institutions in terms of PET/CT scanners, acquisition protocols, and reconstruction settings (Supplemental Table 1).

PET Image Segmentation and Feature Extraction

PET/CT images were reviewed by 2 independent nuclear medicine physicians, who were masked to patients' clinical outcome. The volumes of interest were semiautomatically delineated using LIFEx software (version 6.30, https://www.lifexsoft.org/index.php) (*22*). Four different segmentation methods were applied to delineate lesions, including an SUV threshold of 2.5, an SUV threshold of 4.0 (SUV4.0), 41% of SUV_{max}, and $SUV_{PERCIST}$ (1.5 × liver SUV_{mean} + 2 SDs) (*21,23*). SUV was calculated as (tissue radioactivity concentration [Bq/mL]) × (body weight [g])/(injected radioactivity [Bq]). According to the European Association of Nuclear Medicine guidelines, the liver SUV_{mean} should be between 1.3 and 3.0 (*21*). Conventional PET parameters including SUV_{max}, SUV_{peak}, TMTV, and total lesion glycolysis of each patient were recorded. The distance between the largest lesion and the lesion farthest from that bulk was also recorded (*16*).

Before feature extraction, all PET images were resampled to a voxel size of 3 × 3 × 3 mm using bilinear interpolation (*24*) and were discretized with a fixed bin size of 0.25 SUV (*25*). In total, 1,245 radiomics features were extracted from the entire segmented disease (patient level) via the open-source toolbox PyRadiomics (version 3.0.1) (*16,26*), consistent with the Image Biomarker Standardization Initiative (*27*). Detailed descriptions of the extracted features are presented in Supplemental Table 2. The radiomics workflow is shown in Figure 1.

Feature Selection

The interobserver repeatability of radiomics features was evaluated using the intraclass correlation coefficient (ICC) in 100 randomly selected patients from center 1. Features with an ICC above 0.80 were considered robust and retained for subsequent analysis. The segmentation method with the maximum number of selected features was considered to be the most reliable method.

The ComBat harmonization method was applied to pool all conventional PET parameters and radiomics features derived from images acquired on the 2 different PET/CT scanners (*28*). Pearson correlation coefficient analysis followed by the LASSO algorithm were applied to select features. Details on feature selection are presented in the supplemental materials.

Stacking Ensemble Learning–Based Model Construction

Stacking ensemble learning is a complex machine learning algorithm that combines the result of several base learners to generate predictions into the metalearner to improve predictive accuracy (*18*). In this study, random forest, support vector machine, gradient boosting decision tree, and adaptive boosting were set as the base learners (first level), whereas random forest served as the metalearner (second level). The methodologic details are presented in the supplemental materials. Logistic regression was also applied to generate predictions. Confusion matrix analytics (including accuracy, F1 score, recall, and precision) were used to compare the performance of different machine learning algorithms. The detailed parameters of these algorithms are presented in Supplemental Table 3.

We evaluated the predictive value of 5 different models, including the radiomics model, the combined clinical–radiomics model, IPI, the model based on TMTV, the distance between the largest lesion and the lesion farthest from that bulk, and SUV_{peak} (*17*), as well as the International Metabolic Prognostic Index (*29*). Receiver-operating-characteristic (ROC) curve analysis was used to compare the predictive performance of different models.

TABLE 1
Patient Characteristics

Characteristic	Training set (n = 141)	Internal testing set (n = 61)	External testing set (n = 38)	P
Sex				0.225
Female	67	30	24	
Male	74	31	14	
Mean age ± SD (y)	57.6 ± 15.1	60.6 ± 13.4	64.3 ± 13.6	0.093
Age (y)				0.269
≤60	70	25	14	
>60	71	36	24	
Ann Arbor stage				0.381
I–II	51	21	18	
III–IV	90	40	20	
B symptoms				0.231
Yes	39	19	16	
No	102	42	22	
Performance status				0.324
<2	102	45	32	
≥2	39	16	6	
Extranodal sites				0.432
<2	88	39	28	
≥2	53	22	10	
LDH				0.217
Normal	61	34	20	
Elevated	80	27	18	
β2-microglobulin				0.745
Normal	95	38	24	
Elevated	46	23	14	
IPI				0.900
≤2	77	35	22	
>2	64	26	16	
Cell of origin				0.182
GCB	59	21	10	
Non-GCB	82	40	28	
Therapy regimens				0.560
R-CHOP	126	54	36	
R-EPOCH	15	7	2	
Endpoints				
2-y PFS (%)	69.5	72.1	71.1	0.855
2-y OS (%)	76.6	80.3	73.7	0.569

LDH = lactate dehydrogenase; GCB = germinal center B-cell–like; R-CHOP = rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone; R-EPOCH = rituximab plus etoposide, prednisone, vincristine, cyclophosphamide, and doxorubicin.

Data are n unless otherwise indicated. P values were calculated by 1-way ANOVA for continuous variables, χ^2 test for categoric variables, and log-rank test for survival rates.

Statistical Analysis

All statistical analysis was performed using SPSS (version 26.0), R (version 4.0.5, http://www.R-project.org), and Python (version 3.10). Progression-free survival (PFS) was defined as the time from diagnosis until lymphoma progression or death from any cause. Overall survival (OS) was defined as the time from diagnosis to death from any cause or to the last follow-up. Patients still alive were censored at the date of last contact. The differences in clinical characteristics were

TABLE 2
AUCs of Different Models

Model	Training set		Internal testing set		External testing set	
	PFS	OS	PFS	OS	PFS	OS
Combined	0.791 (0.725–0.857)	0.843 (0.786–0.899)	0.762 (0.618–0.906)	0.741 (0.572–0.911)	0.771 (0.594–0.948)	0.725 (0.534–0.916)
Radiomics	0.765 (0.697–0.834)	0.787 (0.724–0.850)	0.715 (0.559–0.870)	0.637 (0.447–0.827)	0.707 (0.515–0.899)	0.661 (0.450–0.871)
IPI	0.715 (0.624–0.807)	0.729 (0.734–0.823)	0.717 (0.569–0.864)	0.670 (0.497–0.843)	0.715 (0.531–0.900)	0.689 (0.495–0.884)
TMTV + $Dmax_{bulk}$ + SUV_{peak}	0.696 (0.604–0.789)	0.720 (0.623–0.817)	0.623 (0.457–0.788)	0.722 (0.551–0.893)	0.652 (0.452–0.851)	0.640 (0.432–0.848)
IMPI	0.765 (0.681–0.849)	0.765 (0.676–0.854)	0.699 (0.546–0.851)	0.659 (0.479–0.839)	0.660 (0.470–0.850)	0.689 (0.495–0.884)

$Dmax_{bulk}$ = distance between largest lesion and lesion farthest from that bulk; IMPI = International Metabolic Prognostic Index.
Data in parentheses are 95% CIs.

assessed using the χ^2 test and 1-way ANOVA, when appropriate. Patients were stratified into high- and low-risk groups using ROC curve analysis and maximizing the Youden index (*30*). Survival curves were estimated by the Kaplan–Meier analysis, and survival distributions were compared using the log-rank test. A *P* value of less than 0.05 was considered statistically significant.

RESULTS

Patient Characteristics and Outcome

Patients' clinical characteristics are summarized in Table 1. No clinical characteristic had statistically significant differences among different datasets (all $P > 0.05$). The median follow-up intervals for the training, internal testing, and external testing sets were 41 mo (range, 4–105 mo), 44 mo (range, 6–104 mo), and 39 mo (range, 4–69 mo), respectively. By the end of follow-up, relapse and progression occurred in 56, 21, and 14 patients in the training, internal testing and external testing sets, respectively, whereas 45, 16, and 10 patients, respectively, had died.

Feature Selection

Among 4 segmentations, SUV4.0 segmentation showed the highest reliability, with 830 features (66.7%) retained in the context of an ICC of more than 0.8 (Supplemental Table 4). After the Pearson correlation coefficient test, 88 radiomics features were selected for SUV4.0 segmentation. The optimal features were obtained by the LASSO algorithm for construction of different stacking models (Supplemental Table 5).

Model Performance Evaluation

The model performance for 2-y PFS prediction based on different machine learning algorithms is shown in Supplemental Table 6. For the radiomics model, the stacking classifier showed better performance than the other 4 base classifiers and logistic regression,

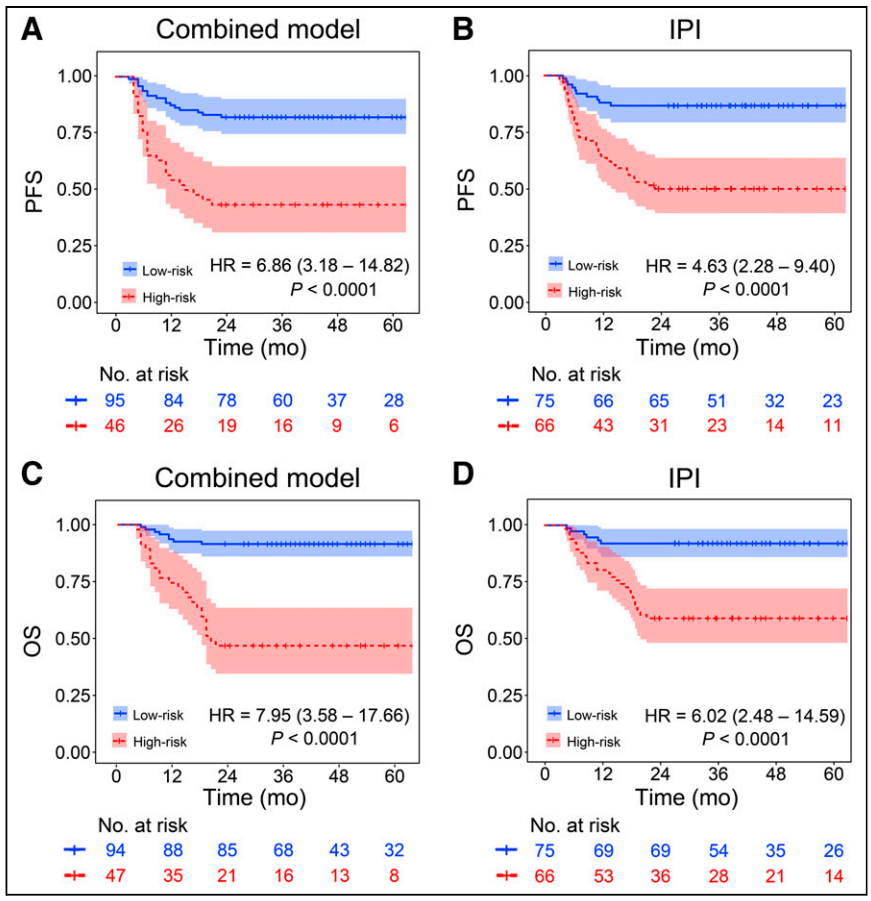

FIGURE 2. Kaplan–Meier curves for PFS of combined model (A), PFS of IPI (B), OS of combined model (C), and OS of IPI (D) in training set. Hazard ratio with 95% CI and log-rank *P* value are reported. HR = hazard ratio.

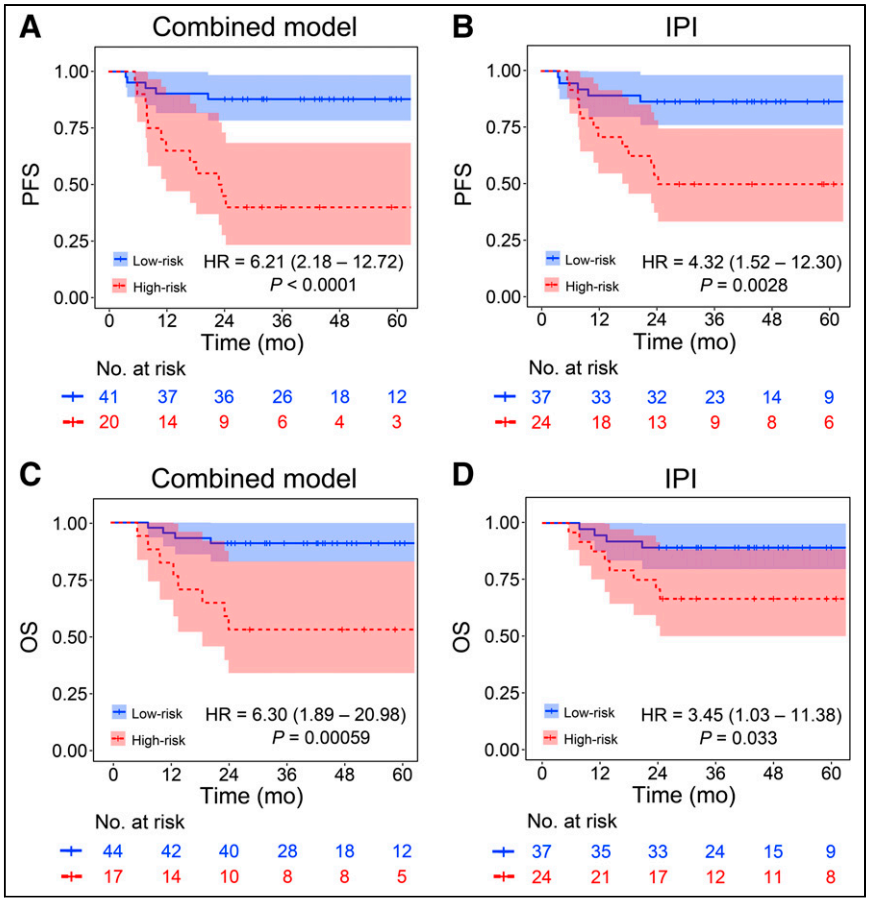

FIGURE 3. Kaplan–Meier curves for PFS of combined model (A), PFS of IPI (B), OS of combined model (C), and OS of IPI (D) in internal testing set. Hazard ratio with 95% CI and log-rank *P* value are reported. HR = hazard ratio.

except for recall in the training set. For the combined model, the stacking classifier also demonstrated better performance than the other classifiers in the training set, internal testing set, and external testing set. Furthermore, the stacking-based combined model had higher predictive power than the radiomics model and IPI across nearly all evaluation metrics.

The model performance for 2-y OS prediction is shown in Supplemental Table 7. For the radiomics model, the stacking classifier demonstrated superior performance to the other base classifiers and logistic regression, except for precision in the internal testing set and accuracy and recall in the external testing set. For the combined model, the stacking classifier had relatively balanced performance in the training set but outperformed the other base classifiers in the internal testing set and the external testing set. Moreover, the stacking-based combined model performed better than the radiomics model and IPI.

We compared the performance of the stacking-based combined models by various combinations of base classifiers. As shown in Supplemental Tables 8 and 9, the combination of 4 base classifiers had a more balanced performance for PFS and OS prediction than did the other combinations. We also evaluated the performance of the radiomics and combined models trained on PFS prediction for predicting OS and vice versa; the results are shown in Supplemental Tables 10 and 11.

The results of ROC analysis are shown in Table 2. The combined model outperformed the other models for PFS prediction, with the area under the ROC curve (AUC) being 0.791, 0.762, and 0.771 in the training set, internal testing set, and external testing set, respectively. A similar trend was observed for OS prediction (the AUCs of the combined model were 0.843, 0.741, and 0.725 for the training set, internal testing set, and external testing set, respectively).

Survival Prediction

Kaplan–Meier survival estimates of the combined model and IPI in the training set, internal testing set, and external testing set are shown in Figures 2, 3, and 4, respectively. The Kaplan–Meier survival estimates of the radiomics model are shown in Supplemental Figure 2. The differences in survival rates between low- and high-risk groups were significant except for OS in the radiomics model in the external testing set ($P = 0.053$). Moreover, the combined model demonstrated a more distinct risk stratification than the radiomics model and IPI, with larger differences between subgroups for both PFS and OS prediction (all $P < 0.05$).

DISCUSSION

In this study, we developed an analytic approach based on [^{18}F]FDG PET radiomics using stacking ensemble learning for outcome prediction in DLBCL. Radiomics and combined clinical–radiomics models constructed by the stacking method outperformed those built on other single machine learning classifiers. Furthermore, the combined models integrating radiomics features and clinical information exhibited predictive performance superior to that of radiomics-only models and IPI.

To the best of our knowledge, this was the first study to evaluate the prognostic effect of [^{18}F]FDG PET radiomics through a stacking ensemble learning approach in patients with DLBCL. Several previous studies have found that machine learning–based PET radiomics could be of prognostic importance in DLBCL (*12–14*). A multicenter study with 317 DLBCL patients suggested that the radiomics model based on LASSO logistic regression was predictive of 2-y time to progression, with an AUC of 0.76 (*16*). Another study using a LASSO-Cox algorithm reported an AUC of 0.748 for the radiomics model in the test set for PFS prediction (*12*). In a recent study, Jiang et al. used cross combination of 7 different machine learning algorithms for feature selection and found that the radiomics signature obtained by the support vector machine–support vector machine was highly predictive of PFS (AUC, 0.757) (*14*). Despite these encouraging findings, a recently developed ensemble learning approach has revealed diagnostic and prognostic advantages over a single machine learning method by aggregating multiple algorithms to achieve higher prediction accuracy (*31,32*). In our current study, the radiomics model built on a stacking ensemble learning approach outperformed those developed by the other 4 base classifiers and logistic regression, with AUCs of 0.715 and 0.707 for PFS prediction in the internal and external testing sets, respectively. This finding is consistent with

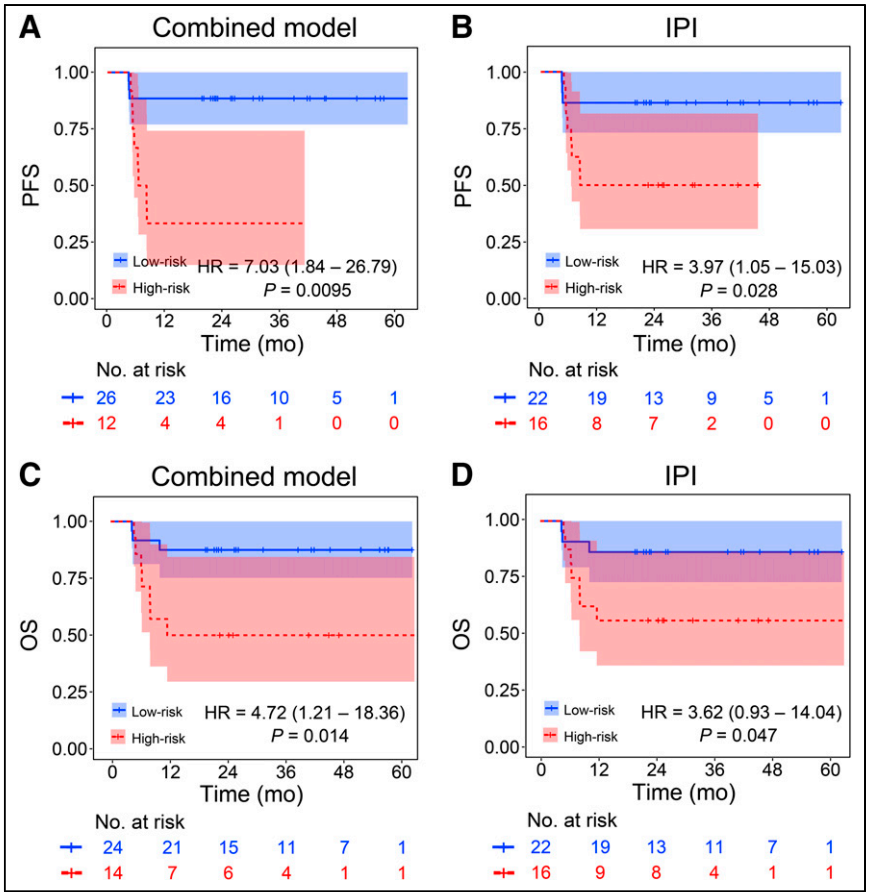

FIGURE 4. Kaplan–Meier curves for PFS of combined model (A), PFS of IPI (B), OS of combined model (C), and OS of IPI (D) in external testing set. Hazard ratio with 95% CI and log-rank *P* value are reported. HR = hazard ratio.

the results from a recent radiomics study on DLBCL, in which a soft voting ensemble–based model showed higher accuracy than those based on single machine learning classifiers for 2-y event-free survival prediction (*15*). Notably, voting considers only linear relationships among classifiers whereas stacking is able to learn complex associations when individual base classifiers are heterogeneous (*33*). In our study, the combined model developed by 4 classifiers showed a more balanced performance than the other combinations, supporting the potential of stacking ensemble learning for radiomics analysis in DLBCL.

Our study also demonstrated that the combined models incorporating patient-level PET radiomics and clinical characteristics yielded higher AUCs and more distinct risk stratifications than IPI for outcome prediction in DLBCL, which is in line with previous observations (*12,14,16*). Recent studies suggested that the predictive ability of IPI has been weakened in the rituximab era (*4*). In this context, PET radiomics might add a new perspective on the phenotypic characteristics of DLBCL through profiling the intratumoral metabolic heterogeneity. Therefore, it is likely that considering both clinical and imaging features in analysis may offer a deeper understanding of the complex biologic properties of malignancy and thereby provide a better prognosis estimation.

Radiomics analysis in lymphoma remains challenging because of the lack of a primary site and the complexity of lesion delineation, particularly for disseminated disease. To date, no consensus has been reached on which segmentation method for lesion delineation in DLBCL is preferable. Although the 41%-of-SUV_{max} method has been recommended by the European Association of Nuclear Medicine for TMTV evaluation (*21*), this method is more likely to be influenced by interobserver variability (*34*). Other studies indicated that the SUV4.0 method could give a good approximation of TMTV for prediction of disease progression (*35*). On top of these, the impact of different segmentations on radiomics features for prognosis prediction in DLBCL remains to be explored. In our study, we compared the reliability of radiomics features based on 4 different segmentation methods. The SUV4.0 method yielded the highest interobserver reliability, with 830 features (66.7%) retained in ICC analysis, which is in line with the results from a recent study suggesting that SUV4.0 is the most stable approach (with excellent reliability for 84.8% of all features) among 6 semiautomatic segmentation methods (*36*). By contrast, the interobserver reliability of radiomics features based on 41%-of-SUV_{max} segmentation was the lowest in the current study, with only 46 features (3.7%) having excellent reliability. This discrepancy may correlate with differences in TMTV delineation. Previous studies demonstrated that variations in segmentation methods could have a marked effect on the outer contour of the segmentation, thereby influencing radiomics features, especially morphologic metrics (*36,37*). In our study, the SUV4.0 method exhibited a higher TMTV estimation and more stable radiomics features than the 41%-of-SUV_{max} method, indicating that a higher TMTV may cause the segmentation method to have less of an impact on radiomics features.

Several limitations of our study deserve mention. First, since this was a retrospective study with a relatively small sample size, our results need to be further validated in prospective multicenter studies involving a larger cohort of patients. Second, we applied only patient-level radiomics analysis; further studies are required to compare the impact of different lesion selection methods on radiomics analysis. Third, we applied ICC, Pearson correlation analysis, and LASSO for feature selection; further studies will be required to assess the performance of other strategies, for example, minimum redundancy maximum relevance and ReliefF. Fourth, to facilitate comparison with previous results, we used only PET images for radiomics analysis. A combination of PET and CT images may lead to the discovery of radiomics features that are more predictive. Fifth, Ki-67 expression and MYC/BCL-2 double-hit status are established prognostic factors but were not assessed in this study because of the incompleteness of the available data.

CONCLUSION

In the present study, we proposed an analytic approach using stacking ensemble learning for outcome prediction in DLBCL

based on [^{18}F]FDG PET radiomics. The stacking-based combined model that incorporates radiomics features and clinical characteristics could enable improved risk stratification in DLBCL patients.

DISCLOSURE

This study was partially supported by the National Natural Science Foundation of China (32027802), the National Key R&D Program of China (2021YFE0108300 and 2022YFE0118000), and the Key R&D Program of Zhejiang (2022C03071). No other potential conflict of interest relevant to this article was reported.

KEY POINTS

QUESTION: Can stacking ensemble learning–based [^{18}F]FDG PET radiomics improve outcome prediction in patients with DLBCL?

PATIENT FINDINGS: In a retrospective study of 240 DLBCL patients, a stacking ensemble learning–based model that incorporates radiomics features and clinical characteristics enabled improved risk stratification.

IMPLICATIONS FOR PATIENT CARE: The stacking ensemble learning–based model incorporating PET radiomics and clinical information can be useful for better survival prediction and therapeutic decision making.

REFERENCES

1. Tilly H, Gomes da Silva M, Vitolo U, et al. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2015;26(suppl 5):v116–v125.
2. Crump M, Neelapu SS, Farooq U, et al. Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. *Blood.* 2017;130:1800–1808.
3. International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med.* 1993;329:987–994.
4. Zhou Z, Sehn LH, Rademaker AW, et al. An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. *Blood.* 2014;123:837–842.
5. Tian M, He X, Jin C, et al. Transpathology: molecular imaging-based pathology. *Eur J Nucl Med Mol Imaging.* 2021;48:2338–2350.
6. Barrington SF, Kluge R. FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. *Eur J Nucl Med Mol Imaging.* 2017;44(suppl 1):97–110.
7. Zhang X, Jiang H, Wu S, et al. Positron emission tomography molecular imaging for phenotyping and management of lymphoma. *Phenomics.* 2022;2:102–118.
8. Cottereau AS, Lanic H, Mareschal S, et al. Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-cell lymphoma. *Clin Cancer Res.* 2016;22:3801–3809.
9. Toledano MN, Desbordes P, Banjar A, et al. Combination of baseline FDG PET/CT total metabolic tumour volume and gene expression profile have a robust predictive value in patients with diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2018;45:680–688.
10. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14:749–762.
11. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin.* 2019;69:127–157.
12. Zhang X, Chen L, Jiang H, et al. A novel analytic approach for outcome prediction in diffuse large B-cell lymphoma by [^{18}F]FDG PET/CT. *Eur J Nucl Med Mol Imaging.* 2022;49:1298–1310.
13. Frood R, Clark M, Burton C, et al. Discovery of pre-treatment FDG PET/CT-derived radiomics-based models for predicting outcome in diffuse large B-cell lymphoma. *Cancers (Basel).* 2022;14:1711.
14. Jiang C, Li A, Teng Y, et al. Optimal PET-based radiomic signature construction based on the cross-combination method for predicting the survival of patients with diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2022;49:2902–2916.
15. Ritter Z, Papp L, Zámbó K, et al. Two-year event-free survival prediction in DLBCL patients based on in vivo radiomics and clinical parameters. *Front Oncol.* 2022;12:820136.
16. Eertink JJ, van de Brug T, Wiegers SE, et al. ^{18}F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2022;49:932–942.
17. Eertink JJ, Zwezerijnen GJC, Cysouw MCF, et al. Comparing lesion and feature selections to predict progression in newly diagnosed DLBCL patients with FDG PET/CT radiomics features. *Eur J Nucl Med Mol Imaging.* 2022;49:4642–4651.
18. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol.* 2018;33:459–464.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–357.
20. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res.* 2012;13:281–305.
21. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging—version 2.0. *Eur J Nucl Med Mol Imaging.* 2015;42:328–354.
22. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res.* 2018;78:4786–4789.
23. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50(suppl 1):122S–150S.
24. Shiri I, Vafaei Sadr A, Amini M, et al. Decentralized distributed multi-institutional PET image segmentation using a federated deep learning framework. *Clin Nucl Med.* 2022;47:606–617.
25. Pfaehler E, van Sluis J, Merema BBJ, et al. Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-dimensionally printed phantom inserts. *J Nucl Med.* 2020;61:469–476.
26. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 2017;77:e104–e107.
27. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020;295:328–338.
28. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med.* 2018;59:1321–1328.
29. Mikhaeel NG, Heymans MW, Eertink JJ, et al. Proposed new dynamic prognostic index for diffuse large B-cell lymphoma: International Metabolic Prognostic Index. *J Clin Oncol.* 2022;40:2352–2360.
30. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J.* 2008;50:419–430.
31. Chassagnon G, Vakalopoulou M, Battistella E, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal.* 2021;67:101860.
32. Papp L, Spielvogel CP, Grubmuller B, et al. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [^{68}Ga]Ga-PSMA-11 PET/MRI. *Eur J Nucl Med Mol Imaging.* 2021;48:1795–1805.
33. Heisler M, Karst S, Lo J, et al. Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography. *Transl Vis Sci Technol.* 2020;9:20.
34. Ilyas H, Mikhaeel NG, Dunn JT, et al. Defining the optimal method for measuring baseline metabolic tumour volume in diffuse large B cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2018;45:1142–1154.
35. Barrington SF, Zwezerijnen B, de Vet HCW, et al. Automated segmentation of baseline metabolic total tumor burden in diffuse large B-cell lymphoma: which method is most successful? A study on behalf of the PETRA consortium. *J Nucl Med.* 2021;62:332–337.
36. Eertink JJ, Pfaehler EAG, Wiegers SE, et al. Quantitative radiomics features in diffuse large B-cell lymphoma: does segmentation method matter? *J Nucl Med.* 2022;63:389–395.
37. Belli ML, Mori M, Broggi S, et al. Quantifying the robustness of [^{18}F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Phys Med.* 2018;49:105–111.