BMC Medical Informatics and Decision Making

RESEARCH

Open Access

An interpretable machine learning model for predicting bone marrow invasion in patients with lymphoma via ^{18}F-FDG PET/CT: a multicenter study

Xinyu Zhu[1], Denglu Lu[2], Yang Wu[1], Yanqi Lu[1], Liang He[1], Yanyun Deng[2], Xingyu Mu[1*] and Wei Fu[1*]

Abstract

Purpose Accurate identification of bone marrow invasion (BMI) is critical for determining the prognosis of and treatment strategies for lymphoma. Although bone marrow biopsy (BMB) is the current gold standard, its invasive nature and sampling errors highlight the necessity for noninvasive alternatives. We aimed to develop and validate an interpretable machine learning model that integrates clinical data, ^{18}F-fluorodeoxyglucose positron emission tomography/computed tomography (^{18}F-FDG PET/CT) parameters, radiomic features, and deep learning features to predict BMI in lymphoma patients.

Methods We included 159 newly diagnosed lymphoma patients (118 from Center I and 41 from Center II), excluding those with prior treatments, incomplete data, or under 18 years of age. Data from Center I were randomly allocated to training ($n=94$) and internal test ($n=24$) sets; Center II served as an external validation set ($n=41$). Clinical parameters, PET/CT features, radiomic characteristics, and deep learning features were comprehensively analyzed and integrated into machine learning models. Model interpretability was elucidated via Shapley Additive exPlanations (SHAPs). Additionally, a comparative diagnostic study evaluated reader performance with and without model assistance.

Results BMI was confirmed in 70 (44%) patients. The key clinical predictors included B symptoms and platelet count. Among the tested models, the ExtraTrees classifier achieved the best performance. For external validation, the combined model (clinical + PET/CT + radiomics + deep learning) achieved an area under the receiver operating characteristic curve (AUC) of 0.886, outperforming models that use only clinical (AUC 0.798), radiomic (AUC 0.708), or deep learning features (AUC 0.662). SHAP analysis revealed that PET radiomic features (especially PET_lbp_3D_m1_glcm_DependenceEntropy), platelet count, and B symptoms were significant predictors of BMI. Model assistance significantly enhanced junior reader performance (AUC improved from 0.663 to 0.818, $p=0.03$) and improved senior reader accuracy, although not significantly (AUC 0.768 to 0.867, $p=0.10$).

*Correspondence:
Xingyu Mu
muxingyu1992@hotmail.com
Wei Fu
fuwei19700513@163.com

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

© The Author(s) 2025. **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/.

Conclusion Our interpretable machine learning model, which integrates clinical, imaging, radiomic, and deep learning features, demonstrated robust BMI prediction performance and notably enhanced physician diagnostic accuracy. These findings underscore the clinical potential of interpretable AI to complement medical expertise and potentially reduce the reliance on invasive BMB for lymphoma staging.

Keywords Lymphoma, Bone marrow involvement, ^{18}F-FDG PET/CT, Machine learning, Radiomics

Introduction

Lymphoma encompasses a diverse spectrum of hematologic malignancies that often involve the bone marrow, significantly influencing patient prognosis and clinical management [1, 2]. Bone marrow invasion (BMI) critically affects lymphoma staging and treatment decisions, necessitating precise detection [3]. Traditionally, bone marrow biopsy (BMB) serves as the gold standard for detecting BMI; however, this invasive procedure is associated with patient discomfort, sampling error, and potential complications [4].

^{18}F-Fluorodeoxyglucose positron emission tomography/computed tomography (^{18}F-FDG PET/CT) has emerged as a valuable noninvasive imaging modality for lymphoma staging, offering both functional and anatomical information [5, 6]. The ability of ^{18}F-FDG PET/CT to detect BMI has been extensively studied, with varying sensitivity and specificity reported across different lymphoma subtypes [7–11]. However, visual interpretation of PET/CT can be subjective and experience dependent, particularly with subtle or heterogeneous marrow uptake [12].

Recent advances in artificial intelligence and machine learning have created new opportunities to enhance the diagnostic capability of medical imaging [13]. Radiomics, which involves the extraction of high-throughput quantitative features from medical images, has demonstrated promise in capturing tumor heterogeneity beyond what is visually apparent [14, 15]. Similarly, deep learning approaches have achieved remarkable performance in various diagnostic tasks by automatically learning hierarchical feature representations from imaging data [16]. The integration of these computational methods with clinical parameters offers the potential to improve BMI prediction in lymphoma patients.

Although several studies have employed radiomics and machine learning techniques to assist in bone marrow infiltration detection or lymphoma-related diagnostic tasks, the majority have primarily focused on PET-based visual pattern analysis or the construction of simple predictive models using conventional radiomic features [17–20]. However, a significant limitation in current machine learning applications for BMI prediction is the "black box" nature of many algorithms, which hinders their clinical adoption due to their limited interpretability. Clinicians are understandably reluctant to incorporate prediction models into their practice without a clear understanding of the underlying decision-making process [21]. Interpretable machine learning models that provide transparency into feature importance and decision pathways could bridge this gap and facilitate clinical implementation.

This multicenter study aimed to develop an interpretable machine learning model that combines clinical data, PET/CT parameters, radiomics, and deep learning features to predict BMI accurately. We hypothesized enhanced performance with this integrative approach and improved clinical utility through interpretability (SHAP values). Additionally, we evaluated the model's impact on diagnostic accuracy among physicians with varying experience levels.

Materials and methods

The Institutional Review Boards of both participating tertiary medical centers—Liuzhou Worker's Hospital (Center I) and The First Affiliated Hospital of Guilin Medical University (Center II)—approved this multicenter retrospective study (approval numbers: 2023QTLL-16 and XJS2018094, respectively). The ethics committees waived the requirement for informed consent due to the retrospective nature of the study. This research strictly adhered to the Declaration of Helsinki and complied with the Standards for Reporting Diagnostic Accuracy Studies (STARD). The study design is illustrated in Fig. 1.

Study patients and datasets

Patients who were diagnosed with lymphoma and who underwent unilateral iliac bone marrow biopsy (BMB) between January 2022 and August 2024 at the two participating medical centers were retrospectively reviewed for eligibility. The inclusion criteria were as follows: (a) pathologically confirmed lymphoma; (b) comprehensive availability of clinical, laboratory, and pathological data; and (c) completion of an ^{18}F-FDG PET/CT scan within 7 days prior to BMB. Patients meeting any of the following criteria were excluded: (a) prior tumor-related therapy before PET/CT examination; (b) incomplete clinical, pathological, or imaging data; (c) unavailable or poor-quality imaging studies; and (d) age younger than 18 years (Fig. 2).

Patients from Center I were randomly allocated into a training set and an internal test set at a ratio of 8:2. Patients from Center II composed the external validation set.



Fig. 1 Overall workflow of this study

PET standardized acquisition protocol

Patients refrained from strenuous activity for at least 24 h and fasted for a minimum of 6 h before radiotracer administration, maintaining blood glucose levels below 11 mmol/L (200 mg/dL). After intravenous injection of ^{18}F-FDG (3.7 MBq/kg; pH 5–7; purity > 95%), patients rested quietly in the supine position for approximately 60 min. Immediately before imaging, patients were instructed to void and then consume water to achieve gastric distention.

CT scans were acquired via the following parameters: 120 kV, automatic current modulation, 780-mm transaxial field of view, 16 × 1.2 mm collimation, pitch of 0.8, rotation time of 0.5 s, slice thickness of 3 mm, and a 512 × 512 matrix. Subsequently, PET images were acquired in three-dimensional(3D) mode with an acquisition time of 1.5 min per bed position (zoom factor of 1.0, matrix size of 200 × 200). PET image reconstruction utilized the ordered subset expectation maximization algorithm (2 iterations, 21 subsets) with a Gaussian post-reconstruction filter of 5-mm full width at half maximum.

Clinical data collection

The following clinical data were systematically collected:

a. General clinical data: patient age, sex, Eastern Cooperative Oncology Group performance status (ECOG PS), presence of B symptoms, International Prognostic Index (IPI) score, Ann Arbor staging, and histopathological diagnosis.
b. Laboratory parameters: hemoglobin (Hb), white blood cell (WBC) count, lymphocyte count, platelet (PLT) count, lactate dehydrogenase (LDH), alkaline phosphatase (ALP), β2-microglobulin (β2-MG), and albumin levels.
c. PET/CT-derived morphological and metabolic parameters: maximum tumor diameter, number of involved lymph node regions, extranodal involvement, maximum and mean standardized uptake values (SUV_{max} and SUV_{mean}), total lesion glycolysis (TLG), and metabolic tumor volume (MTV).
d. Reference standard: Unilateral iliac bone marrow biopsy results served as the definitive criterion for bone marrow invasion classification.

Image preprocessing and ROI segmentation

PET and CT images underwent standardized preprocessing to optimize feature extraction accuracy. The CT images were resampled to isotropic voxel sizes of

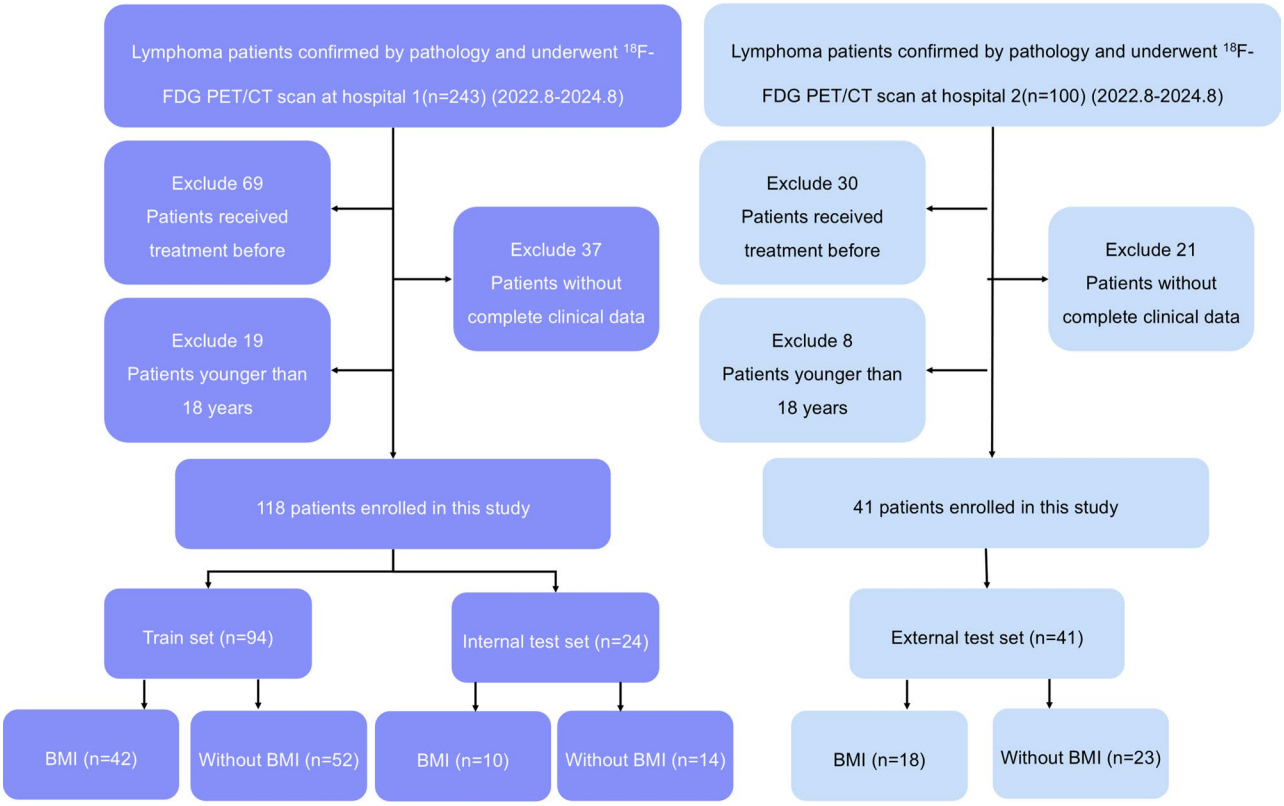


Fig. 2 Patient recruitment workflow

1 × 1 × 1 mm, and the PET images were resampled to 3 × 3 × 3 mm via nearest neighbor interpolation. Standardized window width (1200 HU) and window level (-400 HU) settings were applied to all the CT images. Wavelet transformations were employed to enhance image quality, facilitating smoothing and differentiation.

Regions of interest (ROI) on CT images were automatically segmented via the MONAI Auto3DSeg plugin within 3D Slicer software (version 5.6.2), leveraging a pretrained "Hips and Spine" segmentation model (286 epochs, metric accuracy = 0.902) [22]. We segment the 'left hip', 'right hip', and 'sacrum', all of which were saved as 'pelvic' ROI. Manual adjustments were made as required to ensure segmentation accuracy. PET-derived ROI were delineated by aligning PET images with CT-based segmentation masks, ensuring consistent and robust ROI definition for subsequent radiomic analysis [23].

Feature extraction and selection

Clinical variables and PET/CT-derived features were initially screened through univariate logistic regression analyses. Variables demonstrating statistical significance were subsequently evaluated via multivariate logistic regression analysis to identify independent predictors of BMI.

Radiomic features were extracted from the PET/CT images via the PyRadiomics package (version 3.0.1). For each patient, a total of 3,850 radiomic features—comprising first-order statistics, shape-based features, and texture parameters—were extracted separately from the PET (2,016 features) and CT (1,834 features) images, resulting in a combined radiomic feature set. The detailed extraction parameters are provided in the Supplementary Materials.

Deep learning features were derived via a modified 3D ResNet50 architecture applied to PET/CT images. Prior to deep learning processing, image intensity distributions were standardized via Z score normalization. To address the limited sample size and enhance model robustness, online data augmentation strategies—including random cropping and image flipping—were utilized during training. The Adam optimizer was used with an initial learning rate of 0.001, which was gradually decreased to 0 over 200 iterations, and the softmax cross-entropy loss function was applied for model training. For evaluation (internal and external test sets), only Z score normalization was applied to ensure consistency. The 3D region containing each patient's ROI was extracted and resampled to dimensions of 64 × 64 × 48 voxels, facilitating compatibility with the adapted ResNet50 network architecture. Deep learning features (n = 4,096 per patient)

were subsequently extracted from the penultimate layer average pooling [24].

To mitigate overfitting arising from the high-dimensional feature space, a systematic five-step dimensionality reduction procedure was implemented: (1) Z score normalization: All extracted features were standardized across the entire dataset. (2) Univariate significance testing: Features were evaluated via the Mann–Whitney U test or independent-samples t tests on the basis of normality assumptions, and only those with P values < 0.05 were retained. (3) Correlation filtering: Spearman correlation analysis was performed to remove redundant features, randomly retaining one feature from any highly correlated pair (|correlation coefficient| >0.90). (4) Dimensionality reduction with t-SNE: t-distributed stochastic neighbor embedding (t-SNE) was applied for visualization and dimensionality reduction. (5) Minimum-redundancy maximum-relevance (mRMR): This step further refines feature selection by prioritizing features exhibiting high relevance and minimal redundancy. (6) Least absolute shrinkage and selection operator (LASSO): Predictive features were finalized via LASSO regression with 10-fold cross-validation to identify the most clinically and radiologically meaningful predictors.

For the comprehensive combined model, selected clinical and PET/CT-derived parameters, radiomic features, and deep learning features were integrated into a single unified feature vector. This combined feature vector subsequently underwent the aforementioned rigorous dimensionality reduction and selection process to identify the optimal subset of predictive features for model construction.

Model construction

Four distinct predictive models—clinical-only, radiomic-only, deep learning-only, and combined (clinical, PET/CT, radiomic, and deep learning) models—were developed to predict BMI in lymphoma patients. The performance of three machine learning classifiers—ExtraTrees, support vector machine (SVM), and random forest—was systematically evaluated, and the optimal performing algorithm was subsequently selected on the basis of accuracy metrics. The hyperparameter optimization of machine learning classifiers are presented in supplementary materials. Each classifier was trained on the labeled dataset, providing probabilities of BMI (ranging from 0 to 1) for the test datasets.

Interpretation of the combined model

To facilitate clinical interpretability and transparency, the combined predictive model was analyzed via SHapley Additive exPlanations (SHAP) [25]. SHAP values enabled visualization of the importance of global features for BMI prediction, particularly within the external test set. SHAP summary plots illustrating and ranking each feature's relative contribution across the patient cohort, depicting patient-level risk with a color gradient (red indicates high-risk cases, and blue indicates low-risk cases). Additionally, waterfall plots were employed to offer detailed interpretability at the individual patient level. Generalized additive models (GAMs) were further applied to elucidate nonlinear relationships between predictive features and their SHAP values, thereby strengthening the clinical trustworthiness and interpretability of the predictive model.

PET readers' visual interpretation and combined Model-assisted diagnosis

To evaluate the clinical applicability and diagnostic performance, two nuclear medicine physicians—one junior reader (with less than 1 year of PET interpretation experience) and one senior reader (with more than 5 years of PET interpretation experience)—performed blinded visual assessments of all the external test set PET/CT scans. During initial interpretation, focal abnormal ^{18}F-FDG uptake within unilateral iliac bones was qualitatively identified as indicative of BMI. Readers were permitted to qualitatively interpret PET and CT images and extract relevant parameters such as SUV or Hounsfield units (HUs).

After a one-month interval, both readers performed a second, separate assessment of BMI status. During this round, they were additionally provided with the combined model's predicted BMI probability and patient-specific SHAP values, thereby evaluating the impact of model-assisted interpretation on diagnostic accuracy.

Statistical analysis

Categorical variables were compared via chi-square or Fisher's exact tests, as appropriate. Calibration curves alongside the Hosmer–Lemeshow test were employed to evaluate model calibration. Continuous variables were assessed via either the Mann–Whitney U test or independent-samples t tests on the basis of normality assumptions. Predictive model performance was evaluated by generating receiver operating characteristic (ROC) curves and calculating the corresponding area under the curve (AUC). Additionally, radar plots were constructed to visually summarize the accuracy, specificity, sensitivity, precision, and F1 scores across the models. Differences in AUC values between models were statistically compared via the DeLong test.

Clinical utility was quantified through decision curve analysis (DCA). Differences in PET reader diagnostic performance (with and without model assistance) were evaluated via McNemar's test, while inter-reader agreement in the initial visual assessment was quantified using

Cohen's kappa coefficient. A two-sided p value of < 0.05 was considered statistically significant.

All the statistical analyses were conducted via R software (version 4.3.1) and Python (version 3.7.12). The machine learning models were implemented via the scikit-learn package (version 1.0.2), whereas the deep learning analyses utilized MONAI (version 0.8.1) and PyTorch (version 1.8.1) on an NVIDIA 4090 GPU workstation.

Results

Clinical and PET/CT characteristics

A total of 184 patients were excluded from the two participating centers: 125 from Center I (69 due to prior tumor-related therapy before PET/CT, 37 due to incomplete data, and 19 under 18 years of age) and 59 from Center II (30 due to prior treatment, 21 due to incomplete data, and 8 under 18 years of age). Ultimately, 159 lymphoma patients met the inclusion criteria (Fig. 2). Data from 118 patients at Center I were randomly allocated into training ($n = 94$) and internal test ($n = 24$) sets, whereas data from 41 patients at Center II formed the external validation set.

The baseline demographic and clinical characteristics of the patients in each dataset are presented in Table 1. The mean patient ages were similar across groups: 57.0 years (training), 56.0 years (internal test), and 59.0 years (external test). The proportion of female patients was 46.8% in the training set, 21.8% in the internal test set, and 43.9% in the external test set. Overall, among the 159 patients, 89 (56.0%) had no bone marrow involvement (non-BMI), whereas 70 (44.0%) had confirmed BMI. The detailed characteristics of the participants stratified by BMI status are summarized in Supplementary Table S1.

The semiquantitative PET/CT parameters extracted from the segmented bone marrow regions and their ability to discriminate between the BMI and non-BMI groups are detailed in Table 2. The MTV was not significantly different between the groups. In contrast, the SUV_{max}, SUV_{mean} and TLG were significantly elevated in the BMI-positive group (all $p < 0.01$).

Feature selection and model construction

We comprehensively analyzed clinical variables, PET/CT parameters, radiomic features, and deep learning features derived from ^{18}F-FDG PET/CT to identify robust predictors of BMI. Logistic regression analyses identified B symptoms and the PLT as significant clinical predictors (Table S2). From the radiomic feature set, 11 informative features strongly associated with BMI were selected after dimensionality reduction (Supplementary Fig. 1). For the deep learning features, the LASSO algorithm identified 8 predictive features (Supplementary Fig. 1). The combined model integrated 12 features that demonstrated the strongest predictive associations (Supplementary Fig. 1). Further details and weights of selected features across modalities are provided in Tables S3-5.

Three machine learning algorithms were evaluated for the radiomics-only, deep learning-only, and combined models within the training and internal test sets (Tables S3–S5). After careful assessment, the ExtraTrees classifier demonstrated the highest performance and was selected for subsequent analyses. Figure 3 compares the diagnostic performance of the clinical, radiomic, deep learning, and combined models across datasets.

In the internal set, the combined model achieved the highest area under the curve (AUC = 0.900; 95% CI: 0.689-1.000), which surpassed those of the clinical (AUC = 0.739; 95% CI: 0.454–0.958), radiomic (AUC = 0.770; 95% CI: 0.500-1.000), and deep learning models (AUC = 0.563; 95% CI: 0.293-0.800). The combined model also consistently demonstrated superior performance in the external test sets (AUC = 0.886; 95% CI: 0.769–0.975) to other models(Table 3). Although DeLong tests indicated statistically significant superiority of the combined model over the clinical ($p = 0.005$) and deep learning ($p = 0.028$) models in the training set, no statistically significant differences emerged in the internal or external test sets (Supplementary Fig. 2).

Radar plots were generated to illustrate the diagnostic performance metrics of ExtraTrees classifier across datasets (Fig. 4). Compared with single-modality models, the combined model consistently achieved superior accuracy, specificity, sensitivity, precision, and F1 scores (Table S3). Calibration assessed by the Hosmer–Lemeshow test indicated statistically significant calibration only within the training set ($p = 0.013$), whereas no significant calibration differences appeared in the internal ($p = 0.247$) and external validation sets ($p = 0.171$) (Fig. 3). DCA demonstrated that the combined model consistently yielded greater net clinical benefit than did the individual models across various thresholds (Fig. 3).

Visualizing and interpreting the model

SHAP analysis of the combined model, which was based on the external test dataset, elucidated feature contributions to BMI prediction. The SHAP summary plot (Fig. 5A) identified PET_lbp_3D_m1_glcm_DependenceEntropy as the most impactful feature, followed closely by platelet count and B symptoms. Additional influential radiomic features derived from PET and CT images were identified, highlighting their critical predictive roles. The mean absolute SHAP values (Fig. 5B) confirmed PET_lbp_3D_m1_glcm_DependenceEntropy as the top feature, supported by additional predictive features such as platelet count, B symptoms, PET_wavelet_LLL_firstorder_10Percentile, and PET_lbp_3D_m1_firstorder_InterquartileRange. Figure 6 visually

Table 1 Characteristics of patients with lymphoma in the training and test set

Characteristics	All patients	Training set	Internal test set	External test set	*P* value*
	(*n* = 159)	(*n* = 94)	(*n* = 24)	(*n* = 41)	
Age (years)	57.23 ± 13.39	56.89 ± 14.15	55.62 ± 11.98	58.95 ± 12.47	0.585
Gender					0.02
Male	87	50	19	23	
Female	72	44	5	18	
Pathological type					0.268
HL	16	10	1	5	
Aggressive NHL	108	64	14	30	
Indolent NHL	35	20	9	6	
Maximum tumor diameter (cm)	5.75 ± 4.30	5.85 ± 4.65	6.48 ± 4.81	5.08 ± 2.91	0.419
Number of involved lymph node regions	8.12 ± 5.36	7.96 ± 5.36	8.12 ± 5.66	8.49 ± 5.29	0.871
The number of extranodal involvements	1.22 ± 1.48	1.11 ± 1.23	1.29 ± 1.60	1.44 ± 1.90	0.474
WBC (×10^9/L)	8.43 ± 13.37	9.45 ± 16.77	5.75 ± 2.72	7.65 ± 6.40	0.441
Lymphocytes (×10^9/L)	3.41 ± 12.57	4.35 ± 15.92	1.61 ± 0.81	2.31 ± 5.45	0.516
PLT (×10^9/L)	222.81 ± 121.49	218.37 ± 123.23	239.04 ± 159.49	223.46 ± 90.38	0.759
Hb (g/L)	108.44 ± 30.41	108.26 ± 32.13	115.04 ± 29.44	104.96 ± 26.75	0.436
LDH (U/L)	461.82 ± 734.24	407.07 ± 618.07	487.04 ± 557.07	572.57 ± 1021.21	0.479
ALP (U/L)	105.33 ± 85.28	106.00 ± 78.82	101.33 ± 101.15	106.11 ± 91.61	0.969
β2-MG (mg/L)	4.12 ± 2.66	4.32 ± 2.89	4.38 ± 2.56	3.51 ± 2.04	0.231
Albumin (g/L)	36.93 ± 6.64	36.70 ± 6.86	39.29 ± 7.18	36.06 ± 5.52	0.146
ECOG PS					0.875
0–1	142	84	21	37	
2–4	17	10	3	4	
B symptoms					0.036
Yes	53	28	5	20	
No	106	66	19	21	
Ann Arbor stage					0.762
I	4	3	1	0	
II	21	14	2	5	
III	36	19	5	12	
IV	98	58	16	24	
IPI score					0.149
0–2	71	45	11	15	
3	40	26	6	8	
4–5	48	23	7	18	

Data are shown as n (%) or mean ± standard deviation (SD)

*Statistical comparisons are one-way ANOVA for continuous variables and chi-square test for categorical variables

Hodgkin lymphoma (HL); non-Hodgkin lymphoma (NHL); Eastern cooperative oncology group performance status (ECOG PS), International prognostic index score (IPI score), Hemoglobin (Hb), White blood cell count (WBC), Platelet count (PLT), lactate dehydrogenase (LDH), Alkaline phosphatase (ALP), β2-microglobulin(β2-MG)

demonstrates these predictive contributions within three representative patient cases.

Comparison with visual interpretation and the interpreted combined Model-assisted diagnosis

The diagnostic performance of the combined model when assisting in PET interpretation was evaluated across two nuclear medicine physicians of differing experience levels (Fig. 7). Cohen' s kappa coefficient was used to assess the interobserver agreement of the initial visual assessments prior to model assistance in the external test set, yielding a κ value of 0.784. For the junior reader, model assistance significantly improved diagnostic accuracy, increasing the AUC from 0.663 (visual interpretation alone) to 0.818 ($p = 0.03$). Although the senior reader's diagnostic performance improved from an AUC of 0.768 to 0.867, this difference did not reach statistical significance ($p = 0.10$). Notably, model assistance substantially reduced false-positive and false-negative rates for both readers, with pronounced improvements evident for the junior reader. Detailed comparative diagnostic indices for both readers across evaluation rounds are presented in Table 4.

Table 2 Analysis of the ^{18}F-FDG PET/CT semiquantitative parameters extracted from the maximum tumor for predicting BMI

Parameters	Total (n = 159)	Without BMI (n = 89)	BMI (n = 70)	P
SUV_{max}	3.20 (2.27, 5.19)	2.55 (2.10, 3.66)	4.10 (3.07, 7.61)	<0.01
SUV_{mean}	1.22 (0.97, 1.60)	1.08 (0.86, 1.25)	1.57 (1.27, 2.37)	<0.01
TLG	537.41 (410.45, 730.85)	460.04 (332.19, 589.69)	708.08 (518.28, 939.08)	<0.01
MTV	438.81 (375.65, 500.12)	438.81 (373.91, 503.61)	439.46 (383.99, 494.42)	0.98

Maximum standardized uptake value (SUV_{max}); Mean standardized uptake value (SUV_{mean}); Total lesion glycolysis (TLG); Metabolic tumor volume (MTV)

Discussion

Bone marrow biopsy, although the gold standard for assessing bone marrow invasion in lymphoma, has inherent risks, including patient discomfort, bleeding, infection, and sampling errors, especially due to focal marrow infiltration patterns. An accurate, noninvasive predictive tool could significantly mitigate these clinical risks by potentially sparing patients from unnecessary invasive procedures, particularly in cases with distinctly low or high pretest probabilities. Our study successfully developed and externally validated an interpretable machine learning model to predict BMI in lymphoma patients, integrating clinical variables, PET/CT parameters, radiomic features, and deep learning features. The combined model demonstrated superior predictive

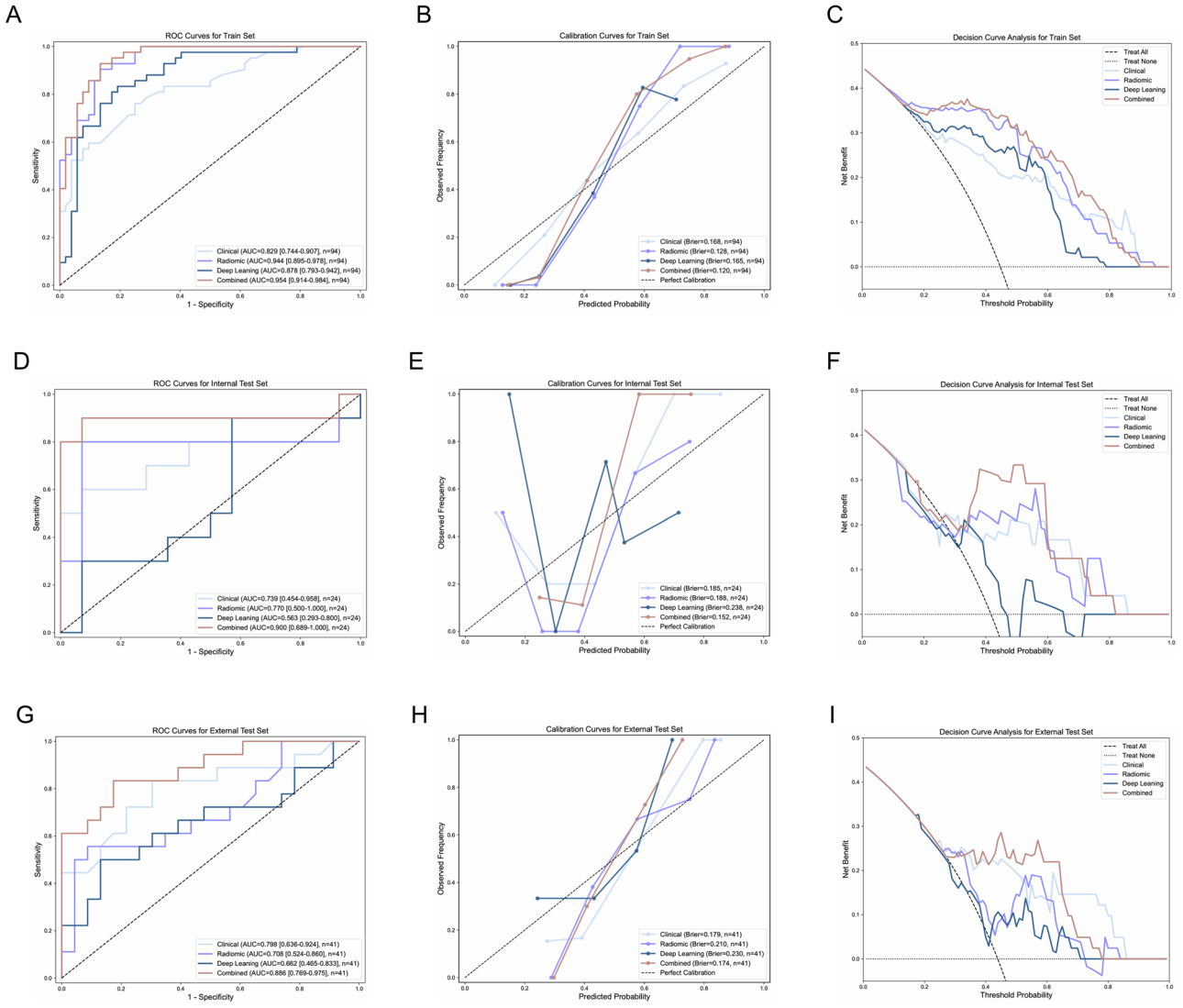


Fig. 3 Diagnostic performance of the four models. (**A**) ROC curve of the training set; (**B**) calibration curve of the training set; (**C**) DCA curve of the training set; (**D**) ROC curve of the internal test set; (**E**) calibration curve of the internal test set; (**F**) DCA curve of the internal test set; (**G**) ROC curve of the external test set; (**H**) calibration curve of the external test set; (**I**) DCA curve of the external test set

Table 3 The performance of different model in internal and external dateset using extra trees classier

Models	AUC (95%CI)	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 score	Group
Clinical Model	0.739(0.454-0.958)	0.750	0.500	0.929	0.833	0.500	0.625	Internal test
Radiomic Model	0.770(0.500-1.000)	0.833	0.700	0.929	0.875	0.700	0.778	Internal test
DL Radiomic Model	0.563(0.293-0.800)	0.583	0.800	0.429	0.500	0.800	0.615	Internal test
Combined Model	0.900(0.689-1.000)	0.875	0.800	0.929	0.667	0.778	0.718	Internal test
Clinical Model	0.798(0.636-0.924)	0.732	0.778	0.696	0.667	0.778	0.718	External test
Radiomic Model	0.708(0.524-0.860)	0.732	0.500	0.913	0.818	0.500	0.621	External test
DL Radiomic Model	0.662(0.465-0.833)	0.683	0.444	0.870	0.727	0.444	0.552	External test
Combined Model	0.886(0.769-0.975)	0.805	0.778	0.826	0.778	0.778	0.778	External test

AUC the area under the receiver operating characteristics curve, PPV the positive predictive value, NPV the negative predictive value

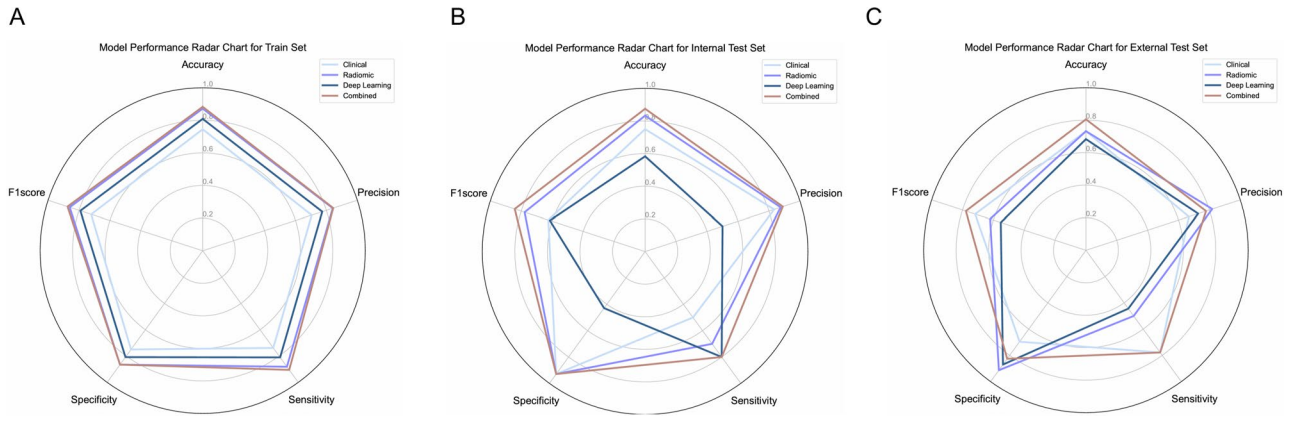


Fig. 4 Radar plots illustrate the diagnostic metrics of each model based on ExtraTrees classifier across datasets. (**A**) Radar plots of the training set. (**B**) Radar plots of the internal test set. (**C**) Radar plots of the external test set

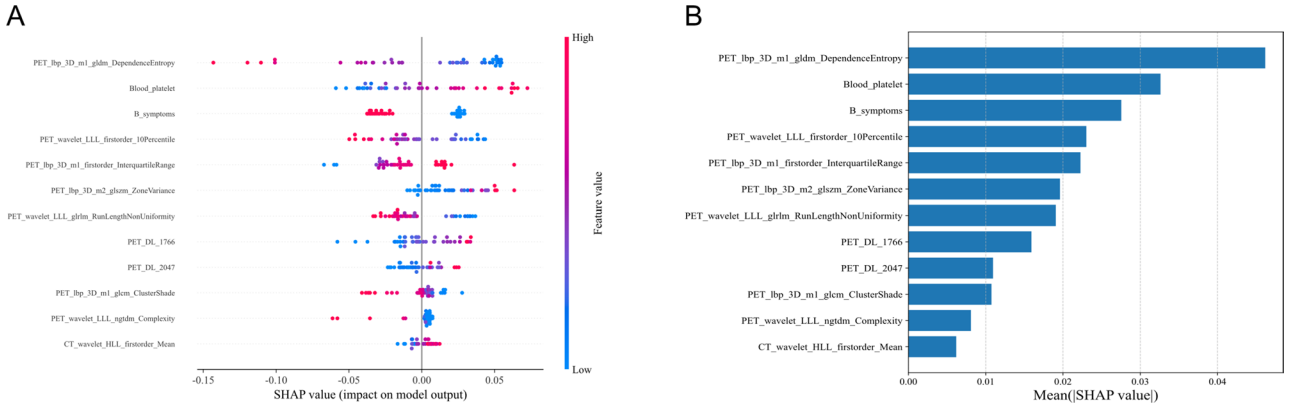


Fig. 5 The model's interpretation is based on an external test set. (**A**) Importance rankings of all risk factors with stability and interpretation via the optimal model. (**B**) The importance ranking of all variables according to the mean (|SHAP value|); the higher the SHAP value of a feature is, the higher the risk. of death the patient would have. The red part of the feature value represents a higher value

performance (AUC = 0.886 in the external validation), significantly outperforming single-modality approaches. SHAP analysis further identified key predictive features, notably PET_lbp_3D_m1_glcm_DependenceEntropy, platelet count, and B symptoms, highlighting the clinical relevance and interpretability of the model. Notably, the combined model significantly improved the diagnostic accuracy for junior readers (AUC: 0.663–0.818; $p = 0.03$) and substantially improved the diagnostic accuracy for senior readers (AUC: 0.768–0.867; $p = 0.10$), underscoring its potential utility in clinical practice.

Previous studies [10, 12, 26, 27] have proposed the use of ^{18}F-FDG PET/CT as a possible alternative to BMB in specific lymphoma subtypes, highlighting both diagnostic and prognostic potential. Although multiple studies have applied radiomics and machine learning techniques for disease diagnosis, research employing explainable radiomics-deep learning integration specifically for

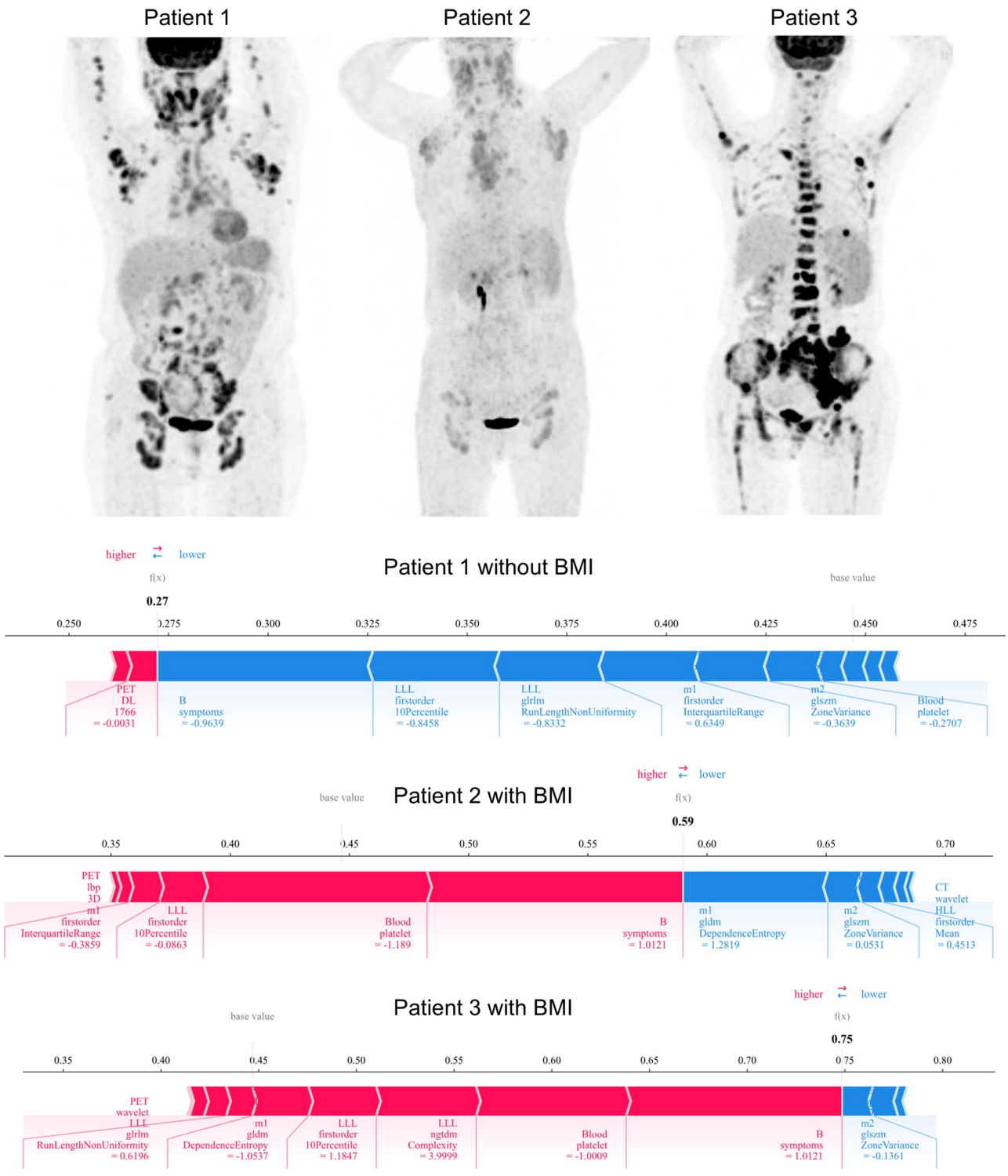


Fig. 6 Interpretation of lymphoma patients with or without BMI assisted by the interpreted combined model

lymphoma bone marrow infiltration remains limited [17, 19]. In contrast, our combined model, which uses a multimodal approach, significantly improved diagnostic performance beyond that of clinical, radiomic, or deep learning models alone. While our clinical-only model demonstrated respectable performance (AUC = 0.798) and the radiomic-only model also showed reasonable predictive ability (AUC = 0.708) during external validation, integration with deep learning features considerably enhanced performance. This aligns with recent evidence

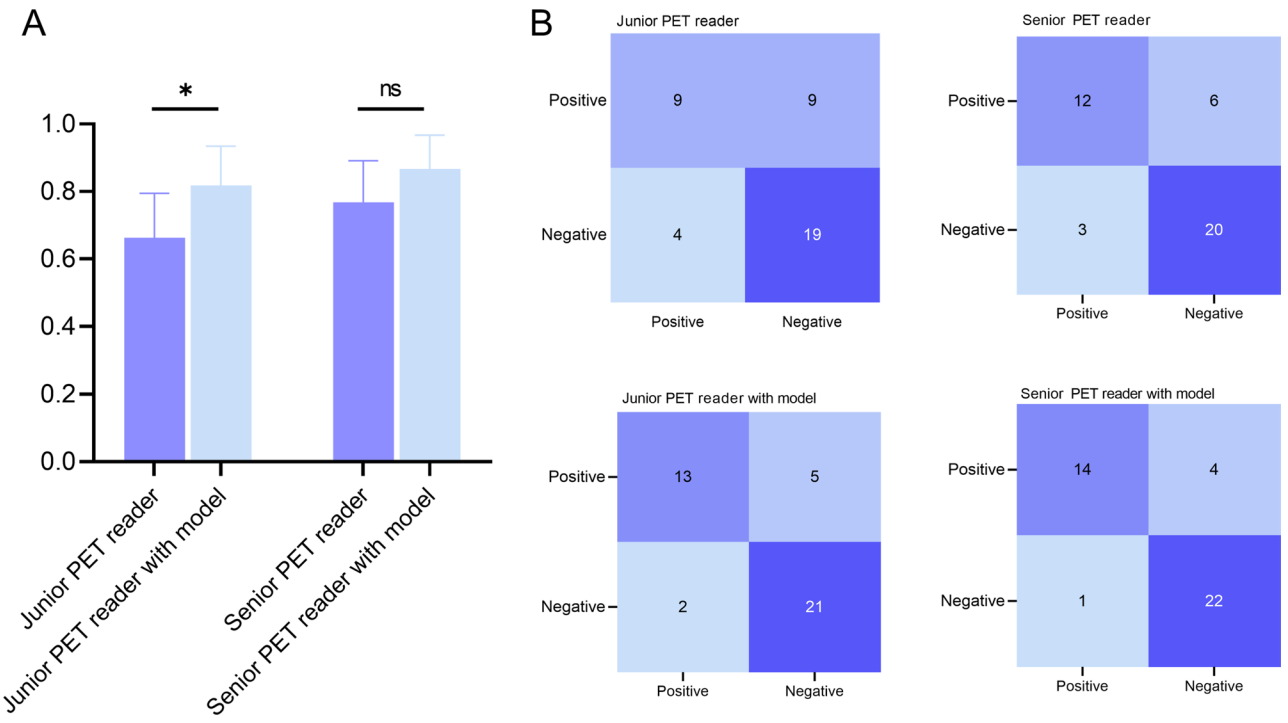


Fig. 7 The diagnostic result of each PET reader in two rounds

Table 4 Performance comparison with visual interpretation and combined radiomic model assisted diagnosis in test set

	AUC (95%CI)	Accuracy	Sensitivity	Specificity	Precision	F1-Score
Combined model	0.886 (0.769-0.975)	0.805	0.778	0.826	0.778	0.778
Visual interpretation						
Junior physician PET Reader	0.663(0.523-0.795)	0.683	0.500	0.826	0.692	0.581
Senior physician PET Reader	0.768(0.629-0.895)	0.781	0.667	0.870	0.800	0.727
Visual with model						
Junior physician PET Reader	0.818(0.687-0.935)	0.829	0.722	0.913	0.867	0.788
Senior physician PET Reader	0.867(0.751-0.967)	0.829	0.778	0.957	0.933	0.849

AUC the area under the receiver operating characteristics curve. The cutoff for the Combined model is based on the Youden index, the cutoff for visual interpretation by the PET reader is set at 0.5

highlighting the benefits of ensemble or integrative models in complex diagnostic tasks [28–31]. Aide et al. [32], for example, previously reported improved lymphoma characterization via combined clinical-radiomics approaches, although their reported AUC values were lower than ours. Our study extends these previous findings by incorporating a comprehensive combination of radiomics, deep learning, and clinical parameters, representing a more robust approach to lymphoma staging.

The identification of PET_lbp_3D_m1_glcm_DependenceEntropy as the most influential radiomic feature provides meaningful insights into lymphoma marrow involvement patterns. This specific radiomic parameter quantifies randomness within gray-level dependency matrices, thus reflecting the inherent heterogeneity associated with lymphoma-induced marrow infiltration. Additionally, the high predictive contributions of clinical factors such as platelet count and B symptoms reinforce known clinical associations: marrow infiltration typically displaces hematopoietic elements, causing thrombocytopenia, whereas systemic "B symptoms" (including fever, night sweats, and weight loss) often indicate an advanced disease stage [33, 34]. In addition, the PET deep learning features "DL-1766" and "DL-2047" exhibited a modest impact on model prediction, potentially reflecting more complex radiological characteristics. These findings validate our model biologically and offer clinicians familiar and interpretable parameters that correspond with established lymphoma pathophysiology.

Model interpretability remains a critical barrier to the adoption of machine learning applications in clinical practice. Many predictive models, while accurate, function as "black boxes" that clinicians find difficult to trust and integrate into diagnostic workflows. Our utilization of SHAP analysis addresses this crucial gap by explicitly quantifying each feature's contribution to individual

predictions, thus providing clinicians with transparent decision-making pathways. The implementation of SHAP has been recommended by Iguchi et al. [35] and Jiang et al. [36] For medical imaging applications, few studies have systematically incorporated this interpretability framework within nuclear medicine. Our study's comprehensive visualization of predictive features and their relative impact facilitates clinical understanding and supports informed clinical decision-making, thus advancing the practical integration of artificial intelligence into routine lymphoma staging.

The diagnostic performance evaluation conducted with nuclear medicine readers further emphasized our model's practical value. Model assistance significantly elevated the diagnostic accuracy for less experienced readers (AUC improved from 0.663 to 0.818; $p = 0.03$) and notably improved the accuracy for more experienced readers (AUC improved from 0.768 to 0.867; $p = 0.10$), suggesting particular benefits in scenarios involving less experienced clinicians; however, a smaller sample size may not lead to significant differences among highly experienced clinicians. Similar patterns of benefit for junior readers from AI assistance have been reported previously by Mu et al. [37], who evaluated 99mTc-MDP bone scans for skull-base invasion. Our findings further expand these insights into lymphoma staging, highlighting the potential of AI to standardize diagnostic accuracy and reduce interobserver variability across clinicians with varying experience levels.

Several limitations of our study should be acknowledged. First, although the total sample size of 159 patients from two centers compares favorably with that of similar studies, the internal ($n = 24$) and external ($n = 41$) validation sets are relatively modest, limiting the generalizability of the results. Second, minor differences in PET/CT acquisition protocols and image reconstruction between the two centers might introduce variations in radiomic feature extraction, although the robust external validation performance suggests resilience against such variations. Third, our retrospective cohort could involve selection bias, as we excluded patients who underwent prior treatments or had incomplete data, potentially limiting the applicability of our findings to broader clinical populations. Fourth, while BMI may occur across multiple skeletal sites, our ROI delineation was confined to the pelvic region, where bone marrow biopsy was conducted in our cohort. Thus, the results may not fully capture systemic bone marrow infiltration patterns. Fifth, despite the extraction of numerous radiomic and deep learning features, other relevant imaging characteristics or clinical predictors may not be included in our analysis. Finally, as our findings are derived from retrospective analyses, prospective multicenter studies are needed to validate the model comprehensively before its clinical implementation.

Conclusion

In summary, our study demonstrated that an interpretable machine learning model integrating clinical parameters, PET/CT imaging, radiomics, and deep learning features can accurately predict bone marrow invasion in lymphoma patients, significantly increasing the diagnostic accuracy for clinicians. By potentially reducing reliance on invasive bone marrow biopsies and improving lymphoma staging accuracy, this approach represents an important step toward personalized, precision medicine in lymphoma management.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03110-8.

Supplementary Material 1

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xinyu Zhu, Denglu Lu, Yang Wu, Yanqi Lu and Liang He. The first draft of the manuscript was written by Xinyu Zhu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant numbers [82402332] and [82460356]) and the Natural Science Foundation of Guangxi Zhuang Autonomous Region (Grant numbers [2025GXNSFHA069001]).

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This research adheres to the Helsinki standards for research on human subjects. The Institutional Review Boards of both participating tertiary medical centers—Liuzhou Worker's Hospital (Center I) and The First Affiliated Hospital of Guilin Medical University (Center II)—approved this multicenter retrospective study (approval numbers: 2023QTLL-16 and XJS2018094, respectively). The ethics committees waived the requirement for informed consent due to the retrospective nature of the study.

Consent to publish

The authors affirm that human research participants provided informed consent for publication of the images in Fig. 6.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Nuclear Medicine, The First Affiliated Hospital of Guilin Medical University, Guilin, Guangxi Zhuang Autonomous Region 541001, China
[2]Department of Nuclear Medicine, Liuzhou Workers' Hospital, Liuzhou, Guangxi Zhuang Autonomous Region 545000, China

Received: 11 May 2025 / Accepted: 9 July 2025

Published online: 15 July 2025

References

1. Chigrinova E, Mian M, Scandurra M, Greiner TC, Chan WC, Vose JM, Inghirami G, Chiappella A, Baldini L, Ponzoni M, et al. Diffuse large B-cell lymphoma with concordant bone marrow involvement has peculiar genomic profile and poor clinical outcome. Hematol Oncol. 2011;29(1):38–41.
2. Park MJ, Park SH, Park PW, Seo YH, Kim KH, Seo JY, Jeong JH, Kim MJ, Ahn JY, Hong J. Prognostic impact of concordant and discordant bone marrow involvement and cell-of-origin in Korean patients with diffuse large B-cell lymphoma treated with R-CHOP. J Clin Pathol. 2015;68(9):733–8.
3. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, Lister TA. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol. 2014;32(27):3059–68.
4. Lidén Y, Landgren O, Arnér S, Sjölund KF, Johansson E. Procedure-related pain among adult patients with hematologic malignancies. Acta Anaesthesiol Scand. 2009;53(3):354–63.
5. Khan AB, Barrington SF, Mikhaeel NG, Hunt AA, Cameron L, Morris T, Carr R. PET-CT staging of DLBCL accurately identifies and provides new insight into the clinical significance of bone marrow involvement. Blood. 2013;122(1):61–7.
6. Albano D, Bosio G, Pagani C, Re A, Tucci A, Giubbini R, Bertagna F. Prognostic role of baseline 18F-FDG PET/CT metabolic parameters in Burkitt lymphoma. Eur J Nucl Med Mol Imaging. 2019;46(1):87–96.
7. Chen-Liang TH, Martin-Santos T, Jerez A, Senent L, Orero MT, Remigia MJ, Muiña B, Romera M, Fernandez-Muñoz H, Raya JM, et al. The role of bone marrow biopsy and FDG-PET/CT in identifying bone marrow infiltration in the initial diagnosis of high grade non-Hodgkin B-cell lymphoma and hodgkin lymphoma. Accuracy in a multicenter series of 372 patients. Am J Hematol. 2015;90(8):686–90.
8. Alzahrani M, El-Galaly TC, Hutchings M, Hansen JW, Loft A, Johnsen HE, Iyer V, Wilson D, Sehn LH, Savage KJ, et al. The value of routine bone marrow biopsy in patients with diffuse large B-cell lymphoma staged with PET/CT: a Danish-Canadian study. Ann Oncol. 2016;27(6):1095–9.
9. Ródenas Quiñonero I, Marco-Ayala J, Chen-Liang TH, de la Cruz-Vicente F, Baumann T, Navarro JT, Martín García-Sancho A, Martin-Santos T, López-Jiménez J, Andreu R et al. The Value of Bone Marrow Assessment by FDG PET/CT, Biopsy and Aspirate in the Upfront Evaluation of Mantle Cell Lymphoma: A Nationwide Cohort Study. *Cancers (Basel)* 2024, 16(24).
10. Adams HJ, Nievelstein RA, Kwee TC. Opportunities and limitations of bone marrow biopsy and bone marrow FDG-PET in lymphoma. Blood Rev. 2015;29(6):417–25.
11. Doma A, Studen A, Jezeršek Novaković B. The impact of bone marrow involvement on prognosis in diffuse large B-Cell lymphoma: an 18F-FDG PET/CT volumetric segmentation study. Cancers (Basel) 2024, 16(22).
12. El-Galaly TC, Villa D, Gormsen LC, Baech J, Lo A, Cheah CY. FDG-PET/CT in the management of lymphomas: current status and future directions. J Intern Med. 2018;284(4):358–76.
13. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10.
14. Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, Cook G. Introduction to radiomics. J Nucl Med. 2020;61(4):488–95.
15. Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics. Eur J Nucl Med Mol Imaging. 2019;46(13):2656–72.
16. Häggström I, Leithner D, Alvén J, Campanella G, Abusamra M, Zhang H, Chhabra S, Beer L, Haug A, Salles G, et al. Deep learning for [(18)F]fluorodeoxyglucose-PET-CT classification in patients with lymphoma: a dual-centre retrospective analysis. Lancet Digit Health. 2024;6(2):e114–25.
17. Faudemer J, Aide N, Gac AC, Damaj G, Vilque JP, Lasnon C. Diagnostic value of baseline (18)FDG PET/CT skeletal textural features in follicular lymphoma. Sci Rep. 2021;11(1):23812.
18. Aide N, Talbot M, Fruchart C, Damaj G, Lasnon C. Diagnostic and prognostic value of baseline FDG PET/CT skeletal textural features in diffuse large B cell lymphoma. Eur J Nucl Med Mol Imaging. 2018;45(5):699–711.
19. Milara E, Sarandeses P, Jiménez-Ubieto A, Saviatto A, Seiffert AP, Gárate F, Moreno-Blanco D, Poza M, Gómez EJ, Gómez-Grande A. Machine Learning Models Based on [18 F] FDG PET Radiomics for Bone Marrow Assessment in Non-Hodgkin Lymphoma. Applied Sciences (2076–3417) 2024, 14(22).
20. Almaimani J, Tsoumpas C, Feltbower R, Polycarpou I. FDG PET/CT versus bone marrow biopsy for diagnosis of bone marrow involvement in non-Hodgkin lymphoma: a systematic review. Appl Sci. 2022;12(2):540.
21. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J Med Chem. 2020;63(16):8761–77.
22. Diaz-Pinto A, Alle S, Nath V, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. Med Image Anal. 2024;95:103207.
23. Garcia DA, Jeans EB, Morris LK, Shiraishi S, Laughlin BS, Rong Y, Rwigema JM, Foote RL, Herman MG, Qian J. A Radiomics-Based classifier for the progression of oropharyngeal Cancer treated with definitive radiotherapy. Cancers (Basel) 2023, 15(14).
24. Niu W, Yan J, Hao M, Zhang Y, Li T, Liu C, Li Q, Liu Z, Su Y, Peng B, et al. MRI transformer deep learning and radiomics for predicting IDH wild type TERT promoter mutant gliomas. NPJ Precis Oncol. 2025;9(1):89.
25. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
26. Chen J, Zhao Y. Pre-treatment [(18)F]FDG PET/CT for assessing bone marrow involvement and prognosis in patients with newly diagnosed peripheral T-cell lymphoma. Hematology. 2024;29(1):2325317.
27. Yu M, Chen Z, Wang Z, Fang X, Li X, Ye H, Lin T, Huang H. Diagnostic and prognostic value of pretreatment PET/CT in extranodal natural killer/T-cell lymphoma: a retrospective multicenter study. J Cancer Res Clin Oncol. 2023;149(11):8863–75.
28. Kenawy MA, Khalil MM, Abdelgawad MH, El-Bahnasawy HH. Correlation of texture feature analysis with bone marrow infiltration in initial staging of patients with lymphoma using (18)F-fluorodeoxyglucose positron emission tomography combined with computed tomography. Pol J Radiol. 2020;85:e586–94.
29. Li H, Xu C, Xin B, Zheng C, Zhao Y, Hao K, Wang Q, Wahl RL, Wang X, Zhou Y. (18)F-FDG PET/CT radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia. Theranostics. 2019;9(16):4730–9.
30. Feng L, Yang X, Lu X, Kan Y, Wang C, Sun D, Zhang H, Wang W, Yang J. (18) F-FDG PET/CT-based radiomics nomogram could predict bone marrow involvement in pediatric neuroblastoma. Insights Imaging. 2022;13(1):144.
31. Feng L, Yang X, Lu X, Kan Y, Wang C, Zhang H, Wang W, Yang J. Diagnostic value of (18)F-FDG PET/CT-Based radiomics nomogram in bone marrow involvement of pediatric neuroblastoma. Acad Radiol. 2023;30(5):940–51.
32. Aide N, Fruchart C, Nganoa C, Gac AC, Lasnon C. Baseline (18)F-FDG PET radiomic features as predictors of 2-year event-free survival in diffuse large B cell lymphomas treated with immunochemotherapy. Eur Radiol. 2020;30(8):4623–32.
33. Martini V, Melzi E, Comazzi S, Gelain ME. Peripheral blood abnormalities and bone marrow infiltration in canine large B-cell lymphoma: is there a link? Vet Comp Oncol. 2015;13(2):117–23.
34. Sugihara A, Katsuya H, Ureshino H, Kido S, Kimura S. Alveolar rhabdomyosarcoma with massive bone marrow infiltration and disseminated intravascular coagulation mimicking acute leukemia. EJHaem. 2022;3(3):1062–3.
35. Iguchi T, Kojima K, Hayashi D, Tokunaga T, Okishio K, Yoon H. Preoperative maximum standardized uptake value emphasized in explainable machine learning model for predicting the risk of recurrence in resected Non-Small cell lung Cancer. JCO Clin Cancer Inf. 2025;9:e2400194.
36. Jiang C, Jiang Z, Zhang X, Qu L, Fu K, Teng Y, Lai R, Guo R, Ding C, Li K, et al. Robust and interpretable deep learning system for prognostic stratification of extranodal natural killer/T-cell lymphoma. Eur J Nucl Med Mol Imaging. 2025;52(5):1739–50.
37. Mu X, Ge Z, Lu D, Li T, Liu L, Chen C, Song S, Fu W, Jin G. Deep learning model using planar whole-body bone scintigraphy for diagnosis of skull base invasion in patients with nasopharyngeal carcinoma. J Cancer Res Clin Oncol. 2024;150(10):449.

Publisher's note

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