European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-025-07090-9

ORIGINAL ARTICLE

An explainable transformer model integrating PET and tabular data for histologic grading and prognosis of follicular lymphoma: a multi-institutional digital biopsy study

Chong Jiang[1] · Zekun Jiang[2] · Zitong Zhang[1] · Hexiao Huang[1] · Hang Zhou[3] · Qiuhui Jiang[4] · Yue Teng[5] · Hai Li[6] · Bing Xu[4] · Xin Li[3] · Jingyan Xu[7] · Chongyang Ding[8] · Kang Li[2] · Rong Tian[1]

Received: 14 November 2024 / Accepted: 10 January 2025

Abstract

Background Pathological grade is a critical determinant of clinical outcomes and decision-making of follicular lymphoma (FL). This study aimed to develop a deep learning model as a digital biopsy for the non-invasive identification of FL grade.
Methods This study retrospectively included 513 FL patients from five independent hospital centers, randomly divided into training, internal validation, and external validation cohorts. A multimodal fusion Transformer model was developed integrating 3D PET tumor images with tabular data to predict FL grade. Additionally, the model is equipped with explainable modules, including Gradient-weighted Class Activation Mapping (Grad-CAM) for PET images, SHapley Additive exPlanations analysis for tabular data, and the calculation of predictive contribution ratios for both modalities, to enhance clinical interpretability and reliability. The predictive performance was evaluated using the area under the receiver operating characteristic curve (AUC) and accuracy, and its prognostic value was also assessed.
Results The Transformer model demonstrated high accuracy in grading FL, with AUCs of 0.964–0.985 and accuracies of 90.2-96.7% in the training cohort, and similar performance in the validation cohorts (AUCs: 0.936–0.971, accuracies: 86.4-97.0%). Ablation studies confirmed that the fusion model outperformed single-modality models (AUCs: 0.974 − 0.956, accuracies: 89.8%-85.8%). Interpretability analysis revealed that PET images contributed 81-89% of the predictive value.

Chong Jiang and Zekun Jiang are co-first authors and contributed equally to the work.

✉ Jingyan Xu
xjy1967@sina.com

✉ Chongyang Ding
chongyangding@163.com

✉ Kang Li
likang@wchscu.cn

✉ Rong Tian
rongtiannuclear@126.com

[1] Department of Nuclear Medicine, West China Hospital, Sichuan University, No.37, Guoxue Alley, Chengdu City, Sichuan Province 610041, China

[2] West China Biomedical Big Data Center, West China Hospital, Sichuan University, No.37, Guoxue Alley, Chengdu City, Sichuan Province 610041, China

[3] Department of Nuclear Medicine, Qilu Hospital of Shandong University, Jinan, Shandong, China

[4] Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, Fujian, China

[5] Department of Nuclear Medicine, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu, China

[6] Department of Pathology, The First Affiliated Hospital of Nanjing Medical University, Jiangsu Province Hospital, Nanjing, Jiangsu, China

[7] Department of Hematology, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu, China

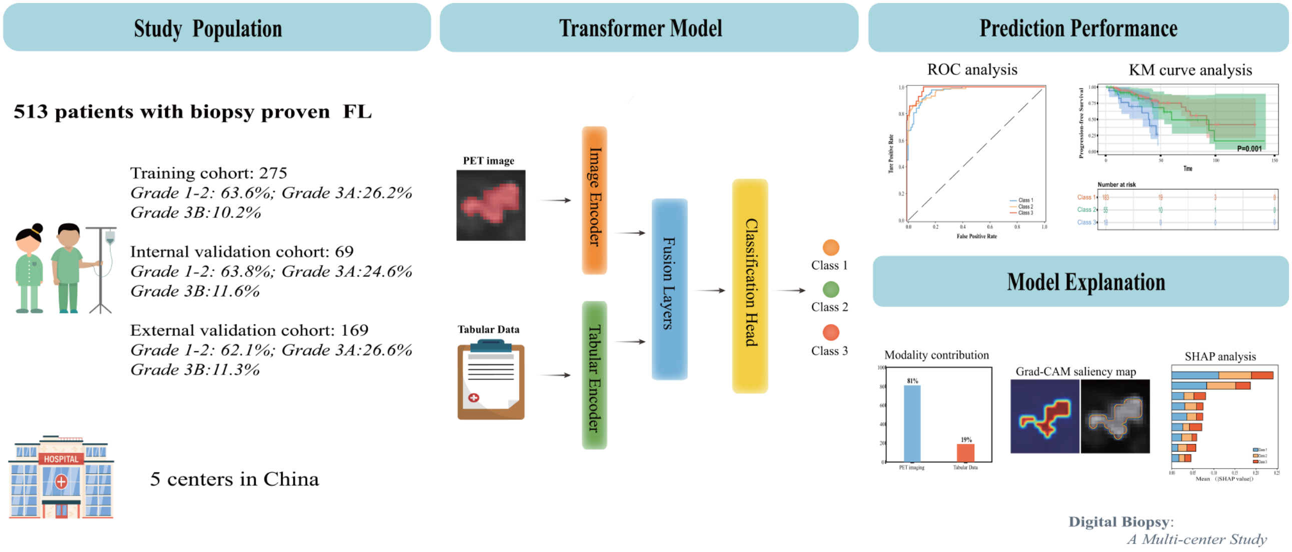
[8] Department of Nuclear Medicine, The First Affiliated Hospital of Nanjing Medical University, Jiangsu Province Hospital, No.321, Zhongshan Road, Nanjing City, Jiangsu Province 210008, China

Published online: 30 January 2025

Grad-CAM highlighted the tumor and peri-tumor regions. The model also effectively stratified patients by survival risk ($P < 0.05$), highlighting its prognostic value.

Conclusions Our study developed an explainable multimodal fusion Transformer model for accurate grading and prognosis of FL, with the potential to aid clinical decision-making.

Graphical Abstract



Keywords Follicular lymphoma · PET/CT · Histologic grade · Prognosis · Transformer model · Digital biopsy

Introduction

Follicular lymphoma (FL) is the most common indolent lymphoma and accounts for approximately 30% of all newly diagnosed non-Hodgkin lymphomas in Western countries [1]. FL is classified into three distinct pathological grades (grade 1–3) based on the number of centroblasts per high-power field [2]. The current WHO classification further subdivides grade 3 into grade 3 A and grade 3B. Patients with grade 1–2 usually exhibit an indolent clinical course, and asymptomatic patients with low tumor burdens are often managed with a watchful-waiting strategy [3]. However, grade 3B follows a more aggressive clinical course with poorer outcomes and is treated similarly to diffuse large B-cell lymphoma (DLBCL) using rituximab-containing poly-chemotherapy regimens, such as R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone) [4]. Moreover, studies have identified differences in survival between FL patients with grade 3 A and those with grade 1–2, as well as between grade 3 A and grade 3B [5, 6]. Therefore, it is crucial for clinicians to distinguish between grade 1–2, grade 3 A, and grade 3B in order to accurately predict the prognosis of FL patients and make informed treatment decisions.

Biopsy combined with immunohistochemical analysis remains the gold standard for confirming the grade of FL. However, this approach is subject to limitations such as sampling bias and the difficulty of obtaining specimens from lesions located in hard-to-reach areas. 18 F-FDG PET/CT is widely recommended for FL staging [7]. Previous research has highlighted the potential of parameters like SUVmax, total metabolic tumor volume (TMTV), and total lesion glycolysis (TLG) in distinguishing between grades 1–2 and grade 3 A FL [8, 9]. However, these studies were limited by small sample sizes and were conducted at single institutions, which raises concerns regarding the generalizability of their findings. Furthermore, these metabolic metrics offer limited insight into the tumor's heterogeneity, which is a critical feature in understanding the biological variability of FL.

In recent years, the large-scale application of artificial intelligence (AI) technology provides an opportunity for in-depth analysis of medical images, particularly in extracting detailed pathological information from tumors [10]. Mu et al. developed a deep learning model using residual convolutional networks to analyze 18 F-FDG PET/CT images of non-small cell lung cancer (NSCLC) patients, achieving high performance in predicting PD-L1 expression (AUC≥0.82) [11]. Similarly, Hu et al. proposed a non-invasive deep learning-based FDG-PET/CT classifier to predict Ki-67 expression in NSCLC, with an accuracy of 0.822 [12]. However, most AI models rely heavily on radiomic features derived from medical images or directly apply deep learning methods to images. Despite some success, these methods often lack effective integration with clinical data, which can provide crucial patient-specific insights into factors affecting both disease diagnosis and progression. Additionally, current AI models typically suffer from a lack of interpretability, limiting their applicability and adoption in clinical settings [13]. For instance, many black-box deep learning models provide only prediction probabilities without revealing the underlying features driving these predictions, making it challenging for clinicians to assess the reliability of these predictions [14, 15]. Finally, much of the research in this field is conducted on limited, single-center datasets [16, 17], raising concerns about the robustness and generalizability of these models.

To address these challenges, our study developed an explainable multimodal fusion transformer model designed to efficiently integrate PET tumor images with patient's tabular information. The model incorporates interpretability mechanisms, such as feature attribution and class activation mapping (CAM) visualization, with the aim of improving the accuracy of FL grading predictions and providing transparent insights into the decision-making process, thereby advancing precision medicine.

Method

Patient cohort

This multicenter research involved a cohort of 513 patients recruited from five distinct medical institutions: West China Hospital, Sichuan University ($n = 195$), Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School ($n = 68$), Qilu Hospital of Shandong University ($n = 81$), Jiangsu Province Hospital, the First Affiliated Hospital of Nanjing Medical University ($n = 115$), and the First Affiliated Hospital of Xiamen University ($n = 54$). The inclusion criteria were as follows: (1) a confirmed histopathological diagnosis of primary follicular lymphoma (grades 1–2, grade 3 A, and grade 3B) based on the World Health Organization classification [18]; (2) no history of other malignancies; and (3) availability of complete medical records. Exclusion criteria included: (1) a prior diagnosis of other cancers; (2) histological transformation of FL; and (3) incomplete medical records. The workflow of patient selection is shown in Fig. 1. Clinical features [gender, age, LDH level, B symptoms, Eastern Cooperative Oncology Group Performance Status (ECOG PS), Ann arbor staging, Hemoglobin level, and β2-microglobulin level] were collected from the medical records. To develop and validate a reliable deep learning model, we carefully structured study cohorts using data from five centers across different regions of China. Specifically, data from West China Hospital, Nanjing Drum Tower Hospital, and Qilu Hospital (from western and eastern China, respectively) were pooled and then randomly divided into training and internal validation cohorts in a 4:1 ratio. Data from Jiangsu Province Hospital

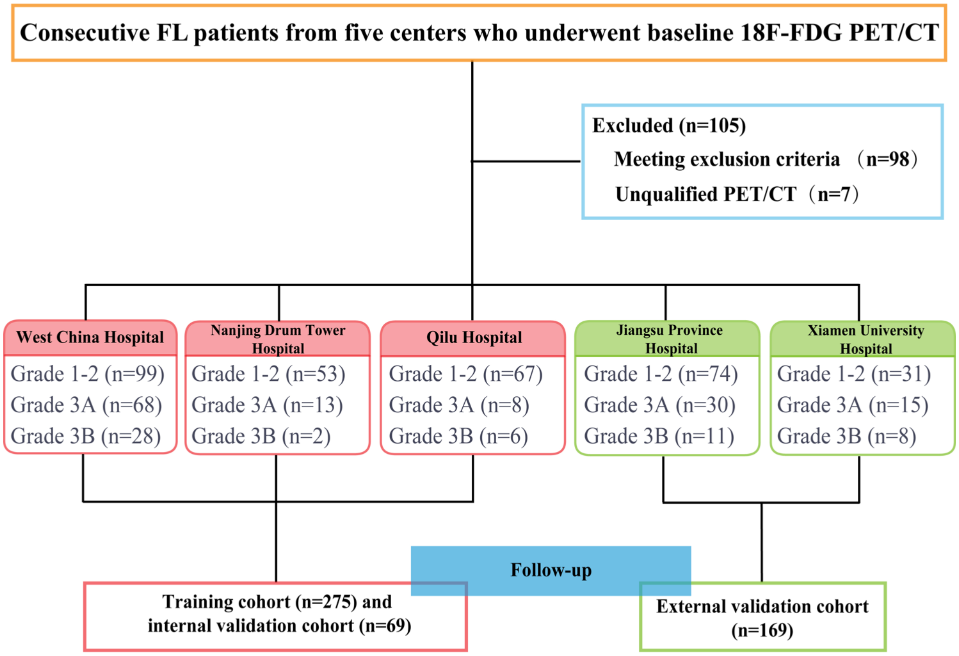


Fig. 1 Flow chart of participant selection

and Xiamen University Hospital were combined to form an independent external validation cohort. Approval for this study was granted by the institutional review board of West China Hospital, Sichuan University, and informed consent was waived due to its retrospective design.

Definition of FL pathological grading

The WHO Classification utilizes a 1–3 grading system based on increased numbers of centroblasts counted per high power field (hpf). Grade 1 FL has 0–5 centroblasts/hpf (follicular small cleaved), Grade 2 FL has 6–15 centroblasts/hpf (follicular mixed), and Grade 3 FL has more than 15 centroblasts/hpf (follicular large cell). Grade 3 has been subdivided into grade 3 A, in which centrocytes are present, and grade 3B, in which there are sheets of centroblasts [18].

PET/CT scanning protocol

All patients underwent PET/CT scans with one of the following systems: Gemini GXL, UM780PET/CT, Biograph 16 PET/CT, GE discovery PET/CT clarity 710, GE Discovery MI, GE Discovery STE. Patients fasted for at least 6 h before scans, resulting in blood glucose levels under 8.7 mmol/L. Then, 185–370 MBq of [18 F] FDG (5.18 MBq/kg) was administered intravenously. The PET/CT scans (from the base of the skull to the upper thigh) were performed 60 min after the radiopharmaceutical injection. Emission data were acquired for 1–2 min in each bed position. Supplementary Table 1 shows the technical parameters for different PET/CT models.

Delineation of target lesions

The lymph node was selected for delineation based on the patient's biopsy records. Semi-automatic delineation was performed using LIFEx-7.3.0 software (https://www.lifexsoft.org/) [19]. The tumor volume of interest (VOI) was outlined on PET images using a 41% SUVmax threshold, which enabled the calculation of the tumor's SUVmax.

Experimental setup

The study aims to develop a multimodal fusion transformer model to achieve precise grading and prognosis prediction for FL by integrating PET tumor images with patient's tabular data (including SUVmax value and clinical features). The model encodes features from PET tumor images and tabular data through an image encoder network and a tabular encoder network, respectively, transforming them into feature embeddings that are fused within the transformer fusion network layers. The model then outputs a three-class prediction through a classification head. To ensure clinical reliability, interpretability mechanisms were added to enhance decision transparency. The model was developed using the training cohort and tested on both the internal and external validation cohorts. To assess its clinical value, we further conducted a prognosis analysis based on the FL grading results predicted by the model.

Additionally, to validate the necessity of data fusion in our Transformer algorithm, we conducted ablation experiments. We developed single-modality models by separately removing the tabular information input and tabular encoder network, as well as by removing the image information input and image encoder network, following the same modeling strategy for classification. These models were then tested on the validation cohorts. The Delong test was used to compare the predictive performance differences among the three models.

The workflow of the multimodal fusion Transformer model we used is shown in Supplementary Fig. 1. The specific details of data processing, model development, and clinical validation are provided in the subsequent sections.

Data preprocessing

To ensure consistent input quality and improve model performance, preprocessing steps are applied to both the PET images and tabular data (including SUVmax value and eight clinical features). The PET tumor region is extracted and isolated as a 3D VOI around the tumor. The VOI undergoes intensity normalization to account for variations in PET scan intensity across different patients. Following normalization, the VOI is resized to a consistent scale, ensuring uniform input dimensions for the image encoder network. The nine tabular features are also normalized to reduce the impact of scale differences among clinical variables, enabling the tabular encoder network to learn effectively from the tabular data. Additionally, to enhance data diversity and model generalization, we applied augmentation strategies to the image data in the training cohort, such as rotation and scaling. To explore the impact of these preprocessing techniques, we performed comparative analysis on different VOI scales, tabular data normalization methods, and the use of data augmentation strategies.

Multimodal fusion transformer model development

The Transformer network consists of four parts: an image encoder network, a tabular encoder network, a fusion network, and a classification head. The image encoder network processes the preprocessed 3D PET tumor VOI using a 3D Swin Transformer Encoder, which is modified by the Swin_UNTER work [20, 21]. The tabular encoder network processes the standardized tabular data using a Multi-Layer Perceptron (MLP) network. The 512-dimensional image and tabular feature embeddings obtained from the encoding process are input into the multimodal fusion network. The fusion network includes two cross-attention layers to capture inter-modal relationships, followed by three self-attention layers to refine the fused representation. Finally, the fusion network outputs a 512-dimensional fusion feature embedding, serving as a robust fusion representation of both PET tumor image and tabular data features. The final feature embedding is fed into the classification head, which is a fully connected network comprising three fully connected layers and two nonlinear layers, to compute the three-class prediction results. The details of the Transformer network are provided in Supplementary Method 1.

To optimize the classification model, the cross-entropy loss is used. The model is trained end-to-end using the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 0.01. A cosine annealing scheduler is applied for dynamic learning rate adjustment, and dropout is used in the MLP and Transformer layers to prevent overfitting.

Interpretability mechanism

To enhance the model's clinical interpretability, three interpretability mechanisms are added, including Gradient-weighted Class Activation Mapping (Grad-CAM) [22], SHapley Additive exPlanations (SHAP) [23], and cross-modal attention contributions. Grad-CAM generates heatmaps from the image encoder's output, highlighting the most relevant regions in the PET image for each prediction. SHAP is applied to the tabular encoder network to provide feature importance scores, clarifying the influence of each tabular feature on the predictions. Attention weights in the cross-attention layers are analyzed to assess the contributions of PET image and tabular data features to the fused representation. This provides insights into how each modality impacts the model's decision-making process.

Clinical evaluation of the explainable transformer model

The explainable transformer model developed on the training cohort was tested and clinically evaluated on both the internal and external validation cohorts. The model's three-class prediction performance was assessed using receiver operating characteristic (ROC) curves and confusion matrices. We calculated accuracy, precision, recall, and F1 scores for each class, along with the overall macro average. To visually represent the distribution and flow of predictions across the different classes, a Sankey diagram was employed. Additionally, survival analysis was conducted for patient risk stratification.

Statistical analysis

Data analysis was conducted using IBM SPSS 25 and R software (version 4.2.2, www.R-project.org). The model's discrimination ability was assessed by estimating the area under the receiver operating characteristic ROC curve (AUC). Progression-free survival (PFS) was used as the endpoint to evaluate patient prognosis. Survival analysis was performed using the Kaplan-Meier method, with comparisons made via the log-rank test. A p-value of less than 0.05 was considered statistically significant for all analyses.

Results

Patient characteristics

The baseline characteristics of the included patients in the training, internal validation, and external validation cohorts are summarized in Table 1. In the training cohort, 175 patients (63.6%) were classified as histologic grade 1–2, 72 patients (26.2%) as grade 3 A, and 28 patients (10.2%) as grade 3B. In the internal validation cohort, 44 patients (63.8%) were classified as grade 1–2, 17 patients (24.6%) as grade 3 A, and 8 patients (11.6%) as grade 3B. In the external validation cohort, 105 patients (62.1%) were grade 1–2, 45 patients (26.6%) were grade 3 A, and 19 patients (11.2%) were grade 3B. The median follow-up times for the training, internal validation, and external validation cohorts were 27.0, 27.5 and 30.0 months, respectively. A total of 26 patients were lost to follow-up.

The transformer model for histologic grade classification

As shown in Table 2; Fig. 2, in the training cohort, the AUCs were 0.964 (accuracy: 90.2%) for grade 1–2, 0.969 (accuracy: 92.4%) for grade 3 A, and 0.985 (accuracy: 96.4%) for grade 3B. The model correctly classified 171 grade 1–2, 57 grade 3 A, and 19 grade 3B cases. In the internal validation cohort, the AUCs were 0.936 (accuracy: 89.9%) for grade 1–2, 0.971 (accuracy: 94.2%) for grade 3 A, and 0.951 (accuracy: 92.8%) for grade 3B, with 43 grade 1–2, 15 grade 3 A, and 3 grade 3B cases accurately predicted. In the external validation cohort, the AUCs were 0.936 (accuracy: 86.4%) for grade 1–2, 0.927 (accuracy: 88.2%) for grade 3 A, and 0.994 (accuracy: 97.0%) for grade 3B, with correct classification of 100 grade 1–2, 30 grade 3 A, and 15 grade 3B cases.

Table 1 Characteristics of the study population

Characteristic	Training cohort ($n=275$)	Internal validation cohort ($n=69$)	External validation cohort ($n=169$)	*P* value
Gender				
Male	123	33	75	0.880
Female	152	36	94	
Age (years)				
Median#	51	50	52	0.202
Q1–Q3	43–61	44–58	41–60	
Elevate LDH				
No	217	55	141	0.498
Yes	58	14	28	
ECOG PS				
0–1	239	63	142	0.318
≥2	36	6	27	
Ann arbor staging				
I	19	3	15	0.883
II	43	12	19	
III	80	25	54	
IV	133	29	81	
B symptoms				
No	217	223	61	0.019
Yes	58	77	19	
Hemoglobin < 120 g/L				
No	205	52	127	0.985
Yes	70	17	42	
Elevate Serum β2-MG				
No	126	29	132	<0.001
Yes	149	40	37	
Histologic Grade				
1–2	175	44	105	0.937
3 A	72	17	45	
3B	28	8	19	

Differences were assessed by Kruskal-Wallis test or Chi-squared test

Abbreviations: LDH, lactate dehydrogenase; ECOG PS, Eastern Cooperative Oncology Group performance status; LDH, lactate dehydrogenase; β2-MG,β2-microglobulin

#median (range)

Table 2 Prediction performance of the multi-modal transformer model across different classification

Cohort	Classification	Accuracy	Precision	Recall	F1 score	support*
Training cohort	Class 1	0.902	0.881	0.977	0.927	175
	Class 2	0.927	0.919	0.791	0.850	72
	Class 3	0.967	1.000	0.679	0.809	28
Internal validation cohort	Class 1	0.899	0.878	0.977	0.925	44
	Class 2	0.942	0.882	0.882	0.882	17
	Class 3	0.928	1.000	0.375	0.545	8
External validation cohort	Class 1	0.864	0.847	0.952	0.897	105
	Class 2	0.882	0.857	0.667	0.750	45
	Class 3	0.970	0.938	0.789	0.857	19

*The support represents the sample size for each category

To demonstrate the effectiveness of our model in clinical practice, we selected three successfully predicted cases from the independent validation cohort to showcase model predictions, as shown in Fig. 3. By inputting the patient's PET tumor images and relevant tabular features, the Transformer model performs inference to predict the FL grade, with the predictions matching the results confirmed by pathological examination.

Results of the ablation study for the transformer model

The results of the ablation experiments are presented in Table 3. In the training, internal validation, and external validation cohorts, the predictive performance of the fusion model (AUC: 0.974, 0.947, and 0.956; accuracy: 0.898, 0.884, and 0.858) was significantly higher than that of the

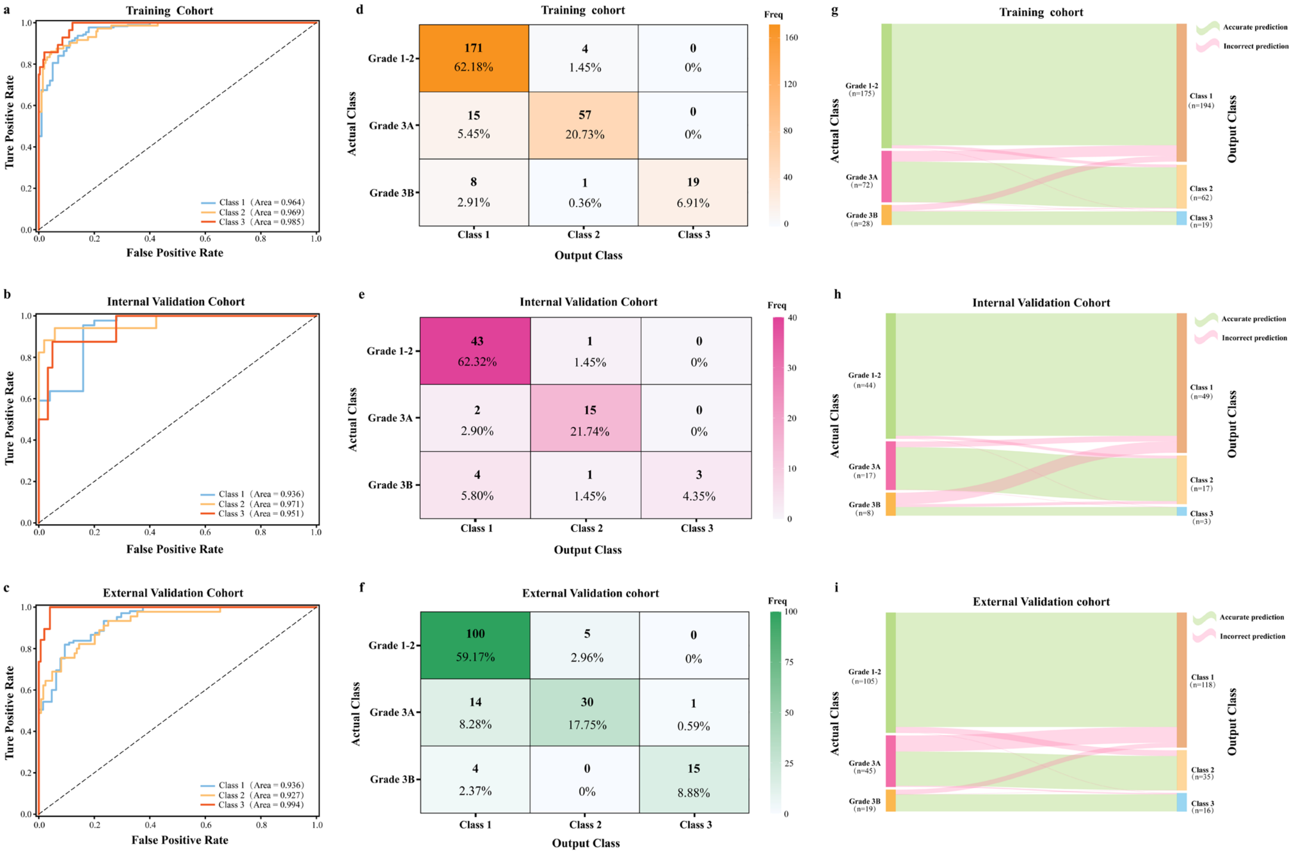


Fig. 2 Performance of the fusion Transformer Model in Different Cohorts: ROC Curves and Confusion Matrices. (**a-c**) ROC curves for the training, internal validation, and external validation cohorts, showing model performance in predicting Grade I/II, Grade 3a, and Grade 3b. (**d-f**) Confusion matrices for the training, internal validation, and external validation cohorts, comparing actual and predicted class distributions across different grades. (**g-i**) The actual and predicted class distributions

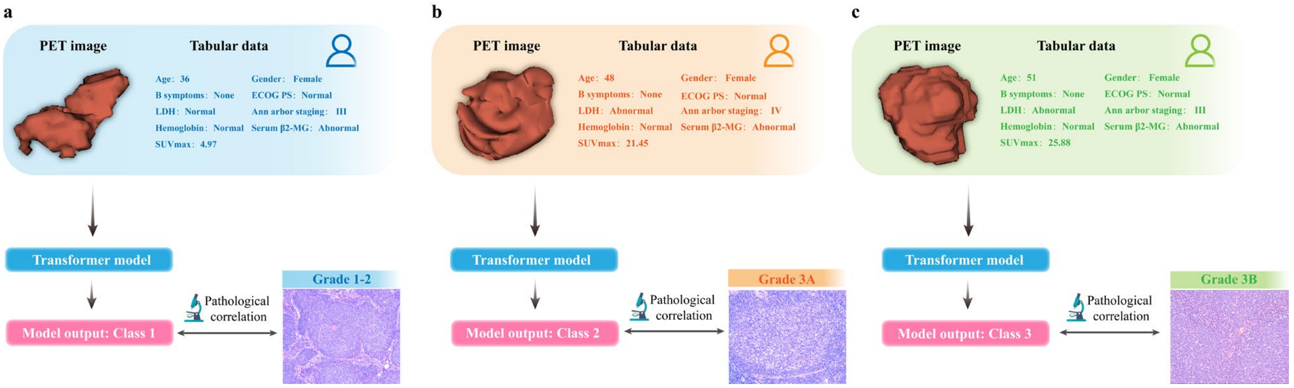


Fig. 3 Illustration of fusion model predictions in clinical practice for patients with varying pathology grades: Grade 1–2 (**a**), Grade 3 A (**b**) and Grade 3B (**c**)

Table 3 Ablation study of the multi-modal transformer algorithm

Cohort	Model	Accuracy	Precision	Recall	F1 score	AUC
Training cohort	Fusion	0.898	0.933	0.816	0.862	0.974
	PET	0.735	0.819	0.532	0.579	0.859
	Tabular	0.633	0.504	0.499	0.481	0.695
Internal validation cohort	Fusion	0.884	0.92	0.745	0.784	0.947
	PET	0.812	0.901	0.591	0.62	0.801
	Tabular	0.580	0.395	0.373	0.369	0.576
External validation cohort	Fusion	0.858	0.881	0.803	0.835	0.956
	PET	0.751	0.767	0.604	0.645	0.835
	Tabular	0.580	0.381	0.406	0.363	0.571

PET single-modality model (AUC: 0.859, 0.801, and 0.835; accuracy: 0.735, 0.812, and 0.751; $P<0.05$) and the tabular single-modality model (AUC: 0.695, 0.576, and 0.571; accuracy: 0.633, 0.580, and 0.580; $P<0.01$). Additionally, the predictive performance results of the PET and tabular single-modality models are provided in Supplementary Figs. 2 and 3, respectively.

Interpretability analysis of the transformer model

While the Transformer model achieved accurate predictions, three types of interpretability analyses were calculated and output, as shown in Fig. 4. First, cross-attention scores revealed that PET images contributed the majority of predictive value (81-89%) during both training and validation, while tabular information also provided predictive value (11-19%), as shown in Fig. 4a and c. Second, SHAP analysis of the MLP visualized the importance ranking of different tabular features in the prediction process, with rankings remaining largely consistent across the three cohorts; notably, age and SUVmax ranked in the top two across all cohorts, as shown in Fig. 4d and f. Lastly, the Grad-CAM maps generated by the Swin Transformer visualized the model's focus areas within the PET tumor region during prediction, showing that the attention was primarily concentrated within the tumor area and partially on specific peri-tumor regions (see Fig. 4g and i).

Prognostic value of model classification outcomes

Based on the Kaplan-Meier (KM) curves (Fig. 5), the Transformer model demonstrated significant prognostic stratification ability for PFS across different cohorts. In the training cohort, the Transformer model effectively divided patients into Class 1–3 groups, showing distinct PFS outcomes ($P<0.05$). The results from the internal and external validation cohorts further confirmed the model's effectiveness. Table 4 shows that the Transformer model achieved survival rates of 85.8%, 78.2%, and 50.0% for Class 1, Class 2, and Class 3 in the training cohort; 78.3%, 64.7%, and 0% in the internal validation cohort; and 83.5%, 70.6%, and 56.3% in the external validation cohort, respectively. These results demonstrate the Transformer model's effective risk stratification across cohorts, providing predictive performance that is comparable to that of the pathologist.

Discussion

This study developed an explainable multimodal fusion Transformer model that takes both PET tumor images and tabular features as input, enabling the accurately prediction of FL grade.

Based on the results from the Transformer model across the training, internal validation, and external validation cohorts, the model demonstrates high predictive performance and strong generalization ability, with AUC values exceeding 0.9 in all cohorts (see Fig. 2). The confusion matrix indicates that predictions for grade 1–2 generally exhibit higher recognition performance, whereas performance for Grades 3 A and 3B is somewhat lower. This is partly due to biological differences between the various grades of follicular lymphoma. Studies have confirmed that Grade 1–2 patients show significant clinical and gene expression differences compared to Grade 3 A, which exhibits more aggressive behavior (e.g., higher frequency of translocations of BCL2, BCL6, and MYC, elevated LDH levels, higher Ki-67 index, and poorer prognosis), whereas Grade 3 A and 3B share more similar biological characteristics [5, 24].

The multimodal fusion Transformer network in this study adopts a dual-tower encoder network structure, encoding the features of image and tabular information separately before alignment and fusion. This approach ensures the integrity of each modality's features and allows for flexible similarity calculations in a shared space, making it well-suited for large-scale data and one of the most commonly used methods in current foundation model research [25, 26]. The multimodal fusion network effectively combines image and tabular feature embeddings through cross-attention and self-attention layers, capturing inter-modal and intra-modal dependencies, respectively [27]. The cross-attention

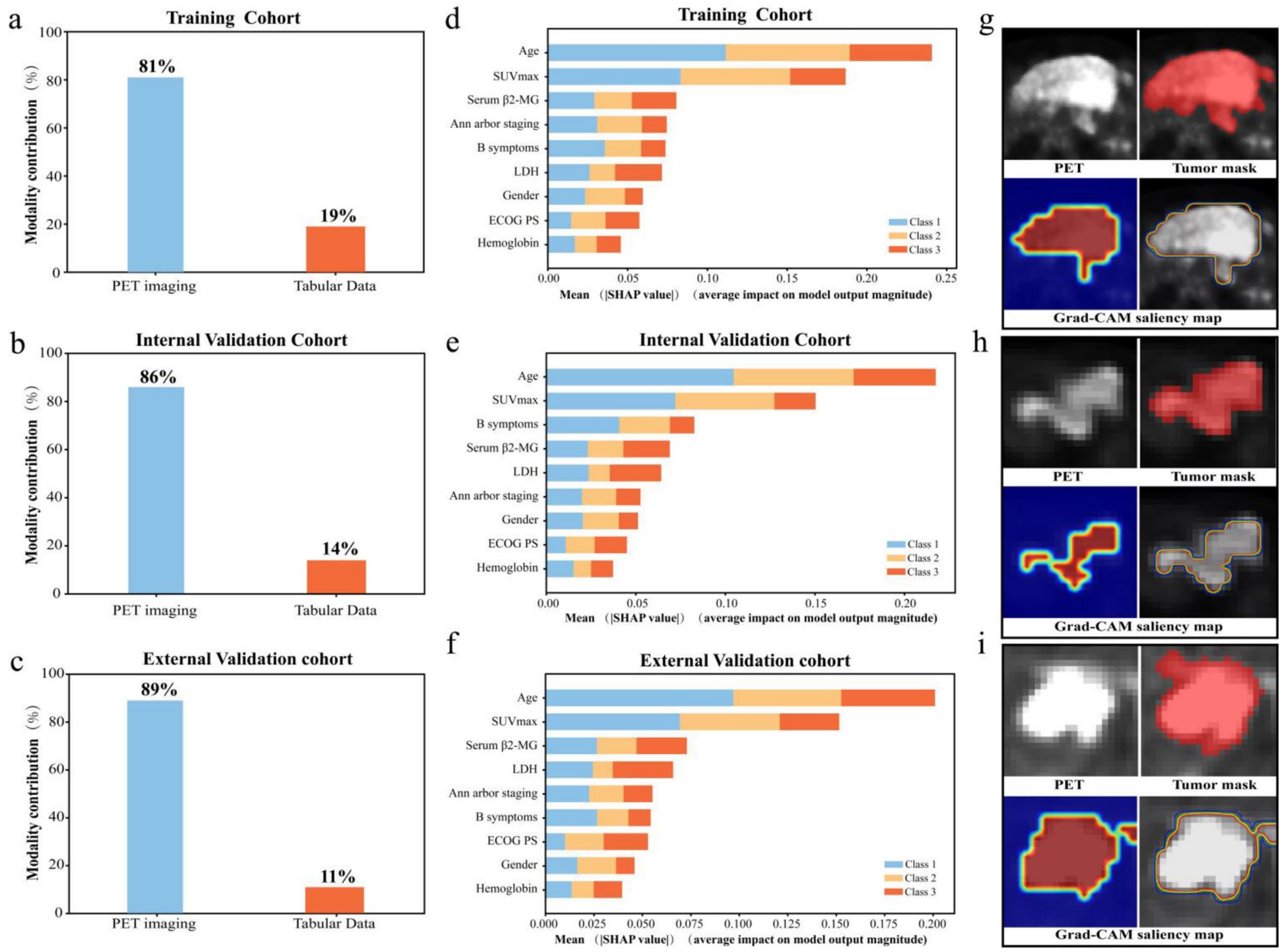


Fig. 4 Interpretability analysis results of the Transformer model. (**a-c**) The cross-modal attention scores for the three cohorts, illustrating the contribution ratios of PET images and tabular information; (**d-f**) The SHAP plots for tabular features across the three cohorts, demonstrating the contribution levels of various tabular features to the decision; (**g-i**) The Grad-CAM visualizations for PET images in the three cohorts, highlighting the areas the model focuses on during decision-making

mechanism allows PET image features to attend to tabular information and vice versa, facilitating inter-modal interaction and enhancing data integration. Additionally, in the image encoder network, we adopted the Swin Transformer architecture, a type of vision Transformer [20]. With recent advances in deep learning, vision Transformer algorithms have achieved diagnostic performance surpassing convolutional networks in large-scale medical image analysis, particularly achieving great success in pathology prediction [28]. Overall, the Transformer network we designed is a flexible and effective multimodal learning algorithm capable of efficiently integrating various modalities in a shared space.

Another major contribution of our algorithm is the comprehensive interpretability analysis. Cross-modal attention scores, SHAP, and Grad-CAM are used to explore decision interpretability between and within the image and tabular data modalities, respectively. First, focusing on interpretability between modalities, the results of the cross-modal attention scores (Fig. 4a and c) indicate that PET images and tabular information contribute differently to FL grading predictions in the fusion model. PET images provide the primary predictive contribution, while tabular information holds relatively less predictive value. This highlights the critical importance of the PET imaging modality in the FL grading model. Of course, the contribution of tabular factors should not be overlooked. The ablation experiment (Table 3) further shows that while the PET model achieves higher predictive performance than the tabular model alone, the fusion model that integrates both PET and tabular information demonstrates a significant improvement in predictive performance ($P < 0.05$).

For interpretability analysis within each modality, we utilized SHAP to analyze the contribution of tabular data to the prediction of FL pathology grade in the model. The results showed that various indicators contributed differently,

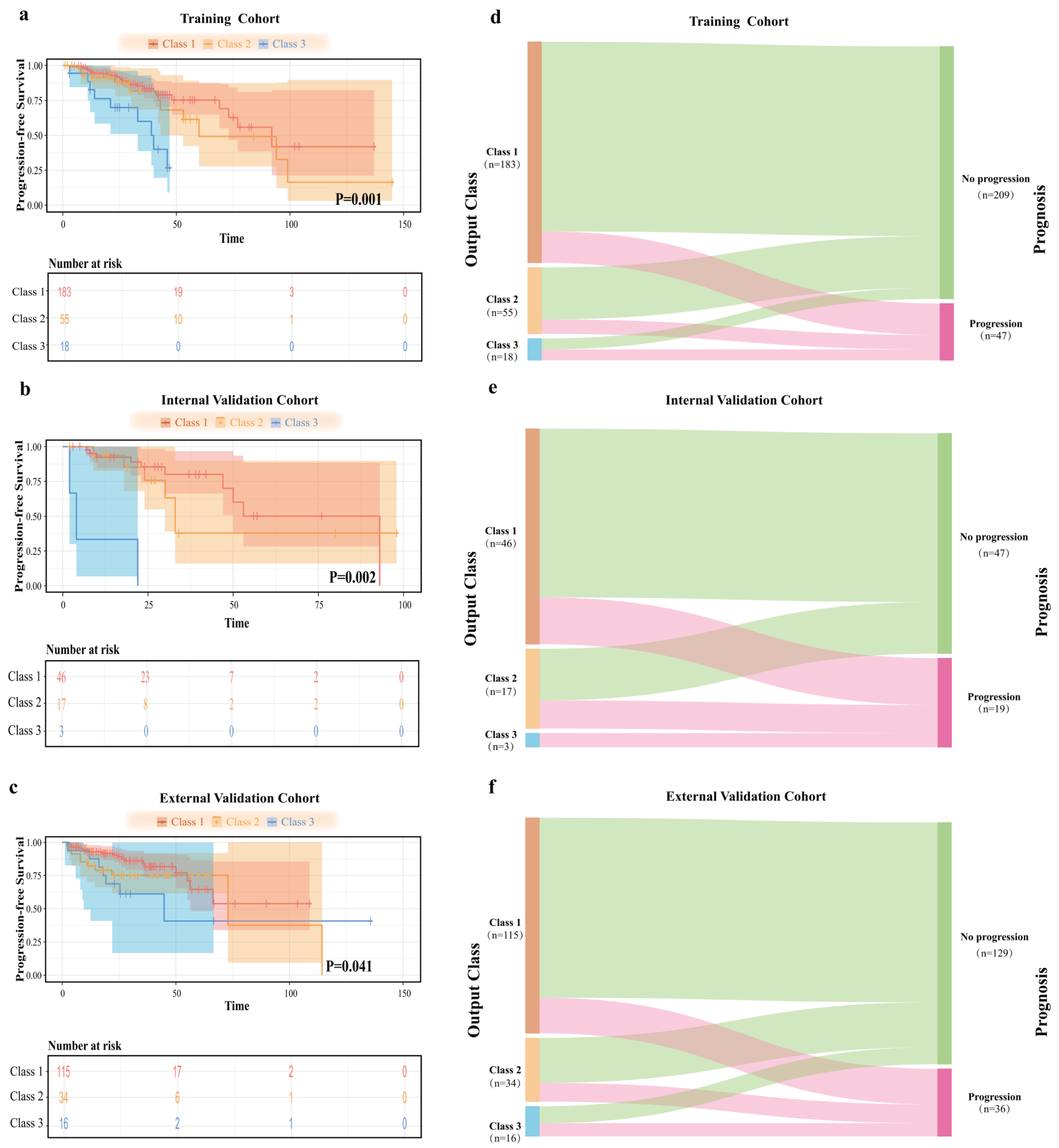


Fig. 5 (**a-c**) KM survival curves depicting the differences in PFS among different classes as determined by the fusion Transformer Model in the training, internal validation, and external validation cohorts. (**d-f**) Sankey diagram of showing the correspondence between prognosis and predicted class distributions

especially age and SUVmax, which ranked among the top two in different groups. Alterations in tumor-host biology are observed in older patients, which may indicate that lymphomas in this age group are biologically more complex and aggressive [29]. It has been reported that FL Grades 3a and 3b are more common in the older population [30]. Additionally, there is a clear trend toward higher 18 F-FDG uptake in more aggressive histologic subtypes of lymphoma [31, 32]. Previous studies have demonstrated the utility of SUVmax in differentiating between Grade 1–2 and Grade 3 A of FL [8, 9]. Our SHAP analysis aligns well with the findings from the aforementioned study, which highlighted

Table 4 Prognostic performance of classification results using transformer model and pathologists

Cohort	Progression-free survival					
	Transformer Model			Pathologist		
	Classification	Survival (%)	χ^2	Grade	Survival (%)	χ^2
Training cohort	Class 1	85.8	11.898	Grade 1–2	86.7	6.557
	Class 2	78.2		Grade 3 A	78.5	
	Class 3	50.0		Grade 3B	57.7	
Internal validation cohort	Class 1	78.3	9.358	Grade 1–2	80.5	15.803
	Class 2	64.7		Grade 3 A	64.7	
	Class 3	0.0		Grade 3B	37.5	
External validation cohort	Class 1	83.5	4.159	Grade 1–2	83.3	2.064
	Class 2	70.6		Grade 3 A	70.5	
	Class 3	56.3		Grade 3B	68.4	

the significant roles of age and SUVmax in the pathological grading of FL. SHAP explains the model's decision-making process by quantifying the contribution of each tabular feature to the model's predictions and calculating their marginal impact on the output, thus providing a clear ranking of feature importance [33]. Meanwhile, we implemented Grad-CAM to visualize the primary focus areas of the Transformer model on PET tumor images. CAM is a technique for visualizing the regions that deep learning models focus on in image classification and recognition tasks, but it requires a convolutional network architecture [34]. The improved Grad-CAM, however, generates heatmaps by calculating the gradient of the target class with respect to the model's feature maps, thereby revealing the regions the model attends to during prediction [22]. This allows it to be more flexibly applied to different neural network structures, such as Transformers. Our Grad-CAM visualizations revealed that the Transformer model primarily focuses on the internal regions of the PET tumor, with some attention to surrounding areas, which is consistent with clinical knowledge.

By integrating these interpretability techniques, our Transformer model not only enables accurate predictions but also provides interpretable analyses of the model's decision-making process. On one hand, these interpretability results (Fig. 4) can help doctors understand how the Transformer model works. On the other hand, the insights gained from tumor images and tabular features can also provide reference and evidence for clinical decision-making. With the development of explainable AI (XAI) techniques, how to further enhance the transparency and interpretability of medical AI models will remain an important research question.

In this study, we further explored the value of the grading prediction results in the prognostic risk stratification of FL patients. The results showed significant prognostic differences among patients in different classes. This finding aligns with clinical studies, which indicate that prognosis differs across Grades of FL patients, with Grade 1–2 having the best prognosis, followed by Grade 3 A, and Grade 3B having the poorest prognosis [5, 6]. Interestingly, while the pathological grading results did not show statistically significant differences in the KM curves of the external validation cohort ($P=0.151$) (seen in Supplementary Fig. 4), the model-based classification was able to stratify patient risk levels ($P=0.041$). We hypothesize that this discrepancy suggests that our classification model not only captures the pathological grading features but also incorporates prognostic heterogeneity of the tumor.

Another issue worth discussing is that, in clinical practice, multimodal data analysis often involves data heterogeneity and modality missing. Regarding data heterogeneity, it is usually caused by factors such as different regions, centers, and imaging equipment, as seen in our study, which incorporated data from five centers across different regions of China. To address the problem, we adopted a mixed grouping strategy and standardized data preprocessing steps. Additionally, through extensive experimentation, we determined the final preprocessing methods (Supplementary Tables 2–4). For example, regarding the scale of the input PET VOIs, we found that smaller sizes might lose more image information, while larger sizes increase the computational load and difficulty during model training, both of which lead to reduced model performance. Appropriate preprocessing and normalization are prerequisites for efficient model training, but there is a lack of in-depth exploration in this area [35]. As discussed in the previous study [35], the specific preprocessing steps required for PET/CT deep learning models are not standardized and depend on the research itself and the algorithms used. On the other hand, regarding modality data missing, our study employed a rigorous data inclusion and exclusion process, ensuring that both PET and tabular data were complete and of high quality, with no missing data. However, our algorithmic framework is not limited to PET and tabular data; theoretically, it can also incorporate additional modalities, such as CT and pathology images. When incorporating additional modalities, a significant challenge will be how to train efficient multimodal AI models despite missing modality data. We are currently exploring this aspect. Of course, this does not imply that adding more modalities will always enhance model performance. Incorporating additional modalities, such as CT images, may also introduce additional computational burden, making it more challenging to develop the fusion model successfully.

Our study also has several limitations. Firstly, due to the retrospective nature of the study, there may be selection bias. Secondly, as mentioned earlier, there was heterogeneity in PET/CT image acquisition among the patient cohorts from 5 independent medical centers, which may have influenced the extracted features and subsequently the performance of the model. To address this, we conducted thorough data preprocessing, including normalization and resizing to a consistent scale. Additionally, to better develop and validate a robust model, we carefully structured study cohorts using data from five centers across different regions of China. Specifically, data from West China Hospital, Nanjing Drum Tower Hospital, and Qilu Hospital (from western and eastern China, respectively) were pooled and then randomly divided into training and internal validation cohorts. Data from Jiangsu Province Hospital and Xiamen University Hospital were combined to form an independent external validation cohort. This grouping ensures that our trained model can perform pathology grading on a broader population, supporting its effectiveness in clinical practice.

We developed an explainable multimodal fusion Transformer model that enables the efficient integration of PET images and tabular information for accurate FL grading predictions. Additionally, the model's decision-making interpretability is enhanced through three interpretability mechanisms, promoting its application in clinical practice and advancing precision medicine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-025-07090-9.

Author contributions Chong Jiang, Zekun Jiang, Zitong Zhang, Hexiao Huang, Hang Zhou, Qiuhui Jiang, Yue Teng, Hai Li, Bing Xu, Xin Li collected data and analyzed data; Jingyan Xu, Chongyang Ding, Kang Li, and Rong Tian participated in the research design and and supervised the study; Chong Jiang and Zekun Jiang equally contributed to the writing of the manuscript and discussing data; Jingyan Xu,, Chongyang Ding, Kang Li, and Rong Tian equally contributed to conceptualization, supervision, and project administration. All authors performed data analysis and interpretation and read and approved the final version of the manuscript.

Funding This study was supported by the National Natural Science Foundation of China (grant 81971653) and the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (grant ZYAI24014).

Data availability The datasets generated and analysed during the current study are available in West China Hospital, Sichuan University, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Jiangsu Province Hospital, the First Affiliated Hospital of Nanjing Medical University, Qilu Hospital of Shandong University and the First Affiliated Hospital of Xiamen University.

Declarations

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This study was approved by the institutional review board of West China Hospital, Sichuan University.

Consent to participate For the nature of retrospective study, written informed consent was waived.

Competing interests The authors declare that they have no conflict of interest.

References

1. Carbone A, Roulland S, Gloghini A, et al. Follicular lymphoma. Nat Rev Dis Primers. 2019;5(1):83.
2. Jacobsen E. Follicular lymphoma: 2023 update on diagnosis and management. Am J Hematol. 2022;97(12):1638–51.
3. Fowler N. Frontline strategy for follicular lymphoma: are we ready to abandon chemotherapy? Hematol Am Soc Hematol Educ Program. 2016;2016(1):277–83.
4. Barraclough A, Bishton M, Cheah CY, et al. The diagnostic and therapeutic challenges of Grade 3B follicular lymphoma. Br J Haematol. 2021;195(1):15–24.
5. Zha J, Chen Q, Ye J, et al. Differences in clinical characteristics and outcomes between patients with grade 3a and grades 1–2 follicular lymphoma: a real-world multicenter study. Biomark Res. 2023;11(1):16.
6. Xue T, Yu BH, Yan WH, et al. Prognostic significance of histologic grade and Ki-67 proliferation index in follicular lymphoma. Hematol Oncol. 2020;38(5):665–72.
7. Metser U, Hussey D, Murphy G. Impact of (18)F-FDG PET/CT on the staging and management of follicular lymphoma. Br J Radiol. 2014;87(1042):20140360.
8. Li H, Wang M, Zhang Y, et al. Prediction of prognosis and pathologic grade in follicular lymphoma using 18F-FDG PET/CT. Front Oncol. 2022;12:943151.
9. Major A, Hammes A, Schmidt MQ, et al. Evaluating Novel PET-CT functional parameters TLG and TMTV in differentiating low-grade Versus Grade 3A Follicular Lymphoma. Clin Lymphoma Myeloma Leuk. 2020;20(1):39–46.
10. Zhang C, Gu J, Zhu Y, et al. AI in spotting high-risk characteristics of medical imaging and molecular pathology. Precis Clin Med. 2021;4(4):271–86.
11. Mu W, Jiang L, Shi Y, et al. Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images. J Immunother Cancer. 2021;9(6):e002118.
12. Hu D, Li X, Lin C, et al. Deep learning to predict the cell proliferation and prognosis of Non-small Cell Lung Cancer based on FDG-PET/CT images. Diagnostics (Basel). 2023;13(19):3107.
13. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Mak. 2020;20(1):310.

14. Goriparthi RG. Interpretable Machine Learning Models for Healthcare Diagnostics: addressing the Black-Box Problem[J]. Revista De Inteligencia Artif en Med. 2022;13(1):508–34.
15. Nasarian E, Alizadehsani R, Acharya UR et al. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework[J]. Inform Fusion, 2024: 102412.
16. de Jesus FM, Yin Y, Mantzorou-Kyriaki E, et al. Machine learning in the differentiation of follicular lymphoma from diffuse large B-cell lymphoma with radiomic [18F]FDG PET/CT features. Eur J Nucl Med Mol Imaging. 2022;49(5):1535–43.
17. Faudemer J, Aide N, Gac AC, et al. Diagnostic value of baseline 18FDG PET/CT skeletal textural features in follicular lymphoma. Sci Rep. 2021;11(1):23812.
18. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127(20):2375–90.
19. 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(16):4786–9.
20. Tang Y, Yang D, Li W et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20730–20740).
21. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. 2022. Swin UNETR: Swin transformers for Semantic Segmentation of Brain tumors in MRI images. arXiv preprint arXiv:2201.01266.
22. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. Int J Comput Vision. 2020;128:336–59.
23. Nohara Y, Matsumoto K, Soejima H, et al. Explanation of machine learning models using shapley additive explanation and application for real data in hospital[J]. Volume 214. Computer Methods and Programs in Biomedicine; 2022. p. 106584.
24. Horn H, Kohler C, Witzig R, et al. Gene expression profiling reveals a close relationship between follicular lymphoma grade 3A and 3B, but distinct profiles of follicular lymphoma grade 1 and 2. Haematologica. 2018;103(7):1182–90.
25. Kim C, Gadgil SU, DeGrave AJ et al. Transparent medical image AI via an image–text foundation model grounded in medical literature[J]. Nat Med, 2024: 1–12.
26. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology[J]. Nat Med. 2024;30(3):863–74.
27. Zhou HY, Yu Y, Wang C, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics[J]. Nat Biomedical Eng. 2023;7(6):743–55.
28. Xu H, Usuyama N, Bagga J et al. A whole-slide foundation model for digital pathology from real-world data[J]. Nature, 2024: 1–8.
29. Castellino A, Santambrogio E, Nicolosi M, et al. Follicular lymphoma: the management of Elderly Patient. Mediterr J Hematol Infect Dis. 2017;9(1):e2017009.
30. Hsi ED, Mirza I, Lozanski G, et al. A clinicopathologic evaluation of follicular lymphoma grade 3A versus grade 3B reveals no survival differences. Arch Pathol Lab Med. 2004;128(8):863–8.
31. Schöder H, Noy A, Gönen M, et al. Intensity of 18fluorodeoxyglucose uptake in positron emission tomography distinguishes between indolent and aggressive non-hodgkin's lymphoma. J Clin Oncol. 2005;23(21):4643–51.
32. Ngeow JYY, Quek RHH, Ng DCE, et al. High SUV uptake on FDG-PET/CT predicts for an aggressive B-cell lymphoma in a prospective study of primary FDG-PET/CT staging in lymphoma. Ann Oncol. 2009;20(9):1543–7.
33. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. 2021;298:103502.
34. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol. 2021;66(4):04TR01.
35. Fallahpoor M, Chakraborty S, Pradhan B, et al. Deep learning techniques in PET/CT imaging: a comprehensive review from sinogram to image space. Comput Methods Programs Biomed. 2024;243:107880.

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