Radiomics-Based Machine Learning for Multi-Modality Tumor Classification and Prognosis in Lymphoma

Choi, Dong Hyeok

Department of Medicine
Graduate School
Yonsei University

Radiomics-Based Machine Learning for Multi-Modality Tumor Classification and Prognosis in Lymphoma

Advisor Kim, Jin Sung

**A Dissertation Submitted
to the Department of Medicine
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Medical Science**

Choi, Dong Hyeok

June 2025

Radiomics-Based Machine Learning for Multi-Modality Tumor Classification and Prognosis in Lymphoma

This Certifies that the Dissertation of Choi, Dong Hyeok is Approved

Committee Chair **Kim, Hojin**

Committee Member **Kim, Jin Sung**

Committee Member **Kim, Dong Wook**

Committee Member **Ahn, So Hyun**

Committee Member **Yoon, Hai-Jeon**

Department of Medicine
Graduate School
Yonsei University
June 2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Jin Sung Kim, for his invaluable guidance, unwavering support, and continuous encouragement throughout my doctoral studies. His insights and expertise have been essential in shaping my academic journey, and his mentorship was a cornerstone that made the completion of this degree possible.

I am especially indebted to Professor So Hyun Ahn, who provided dedicated and detailed supervision throughout the majority of my graduate studies. Her genuine commitment to my development as a researcher went far beyond the responsibilities of an advisor. She not only guided me through the technical aspects of my research with exceptional clarity and care, but also taught me how to think critically and persevere through challenges. Her mentorship was instrumental in both my academic and personal growth. I am profoundly grateful for her patience, generosity, and unwavering belief in my potential. Without her support, I would not be where I am today.

I also wish to thank Professor Dong Wook Kim, who offered meaningful guidance and support during the early years of my doctoral program. His mentorship during the first two years helped lay the groundwork for my research and shaped the way I approach scientific problems.

My sincere thanks go to Professor Woo Sang Ahn, Professor Kwangwoo Park, Professor Min Cheol Han, and Professor Ji Hoon Kim for their collaboration, guidance, and the opportunity to engage in meaningful research together. I have learned so much through our joint projects and discussions, and their encouragement and technical advice played a vital role in my growth as a researcher.

I am deeply grateful to Professor Hai-Jeon Yoon, whose clinical expertise and insightful advice significantly enriched my doctoral research. As a physician, she helped me understand the clinical implications of my work, and her thoughtful feedback on data interpretation and research direction was essential in ensuring the relevance and quality of my study. Her meticulous guidance and compassionate mentorship were truly invaluable.

I would also like to express my appreciation to Professor Hojin Kim, the chair of my dissertation committee. His constructive comments, thoughtful critiques, and words of encouragement during the defense process have been a source of motivation and helped refine the final version of my dissertation.

During my doctoral studies, I had the privilege of being a member of the Medical Physics and Biomedical Engineering Lab (MPBEL) at Yonsei University College of Medicine. I am sincerely thankful to all the lab members for creating a collaborative, warm, and intellectually stimulating environment. The time I spent with them was filled with memorable discussions, helpful feedback, and mutual support that made my graduate life truly enjoyable.

I also spent part of my research period at the Department of Biomedical Engineering at Ewha Womans University, where I had the opportunity to collaborate with talented researchers and engage in valuable research discussions. I am thankful for their kindness and the positive academic atmosphere they fostered.

Finally, I would like to express my heartfelt gratitude to my family. To my mother, father, and older sister, thank you for your endless love, patience, and unwavering belief in me. Your emotional support and sacrifices have been the foundation that sustained me throughout this long and sometimes difficult journey. I dedicate this accomplishment to you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Radiomics-Based Machine Learning for Multi-Modality Tumor Classification and Prognosis in Lymphoma

Purpose: This study aimed to develop a radiomics-based machine learning framework capable of differentiating tumor, normal, and mixed tumor-normal regions in lymphoma patients using 18F-FDG PET/CT images and to evaluate its effectiveness in predicting prognosis, including recurrence and mortality.

Materials and Methods: F-18 FDG PET/CT imaging data from 60 patients diagnosed with lymphoma were retrospectively analyzed. A total of 417 radiomic features were extracted from each imaging modality (PET and CT) based on manually delineated tumor (n = 800) and normal tissue (n = 4,150) volumes of interest. Five machine learning classifiers—AdaBoost, Decision Tree, Gradient Boosting, Random Forest, and XGBoost—were trained using four distinct feature sets: PET radiomics features alone, CT radiomics features alone, combined PET/CT radiomics features, and standardized uptake value (SUV)-based metrics derived from PET images. To enhance tumor characterization, a scoring system integrating ensemble model predictions with anomaly detection using the Isolation Forest algorithm was developed. For prognostic modeling of five-year recurrence and overall survival, SUV-derived metrics, clinical variables, and Synthetic Minority Over-sampling Technique (SMOTE) were utilized to address class imbalance. Model generalizability and robustness were evaluated via external validation using an independent cohort consisting of 16 patients.

Results: The CT-only radiomics model achieved the highest tumor classification performance with an AUC of 0.9690, compared to combined PET/CT radiomics model (AUC: 0.9639) and PET radiomics model (AUC: 0.9607), while PET-only radiomics model demonstrated optimal sensitivity (recall: 65.80%). XGBoost consistently outperformed other algorithms across all feature combinations, with PET/CT achieving 94.23% accuracy and PET-only achieving 93.47%

accuracy. For prognostic prediction without clinical data, recurrence accuracy ranged from 42-67% (without SMOTE) to 50-75% (with SMOTE), while mortality prediction ranged from 71-79% (without SMOTE) to 71-86% (with SMOTE). However, clinical data integration yielded inconsistent results, with recurrence prediction accuracy ranging from 47% to 92%. External validation confirmed model generalizability, with PET-based features showing the best performance (accuracy: 90.34%, AUC: 0.8852). Sensitivity decreased from 65.80% to 40.4% in external validation, indicating inter-institutional variability and the need for institutional calibration.

Conclusion: The developed radiomics-based machine learning framework effectively differentiates tumor, normal, and mixed volumes in lymphoma patients, demonstrating strong potential for enhancing prognosis prediction. However, sensitivity reduction in external validation underscores the need for further refinement and institutional calibration before widespread clinical adoption.

Key words: PET/CT Imaging, Lymphoma, Radiomics, Machine Learning, Tumor Score

1. Introduction

Lymphomas represent a diverse and heterogeneous group of malignancies characterized by the abnormal clonal proliferation of lymphocytes, broadly categorized into Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL). The NHL accounts for approximately 90% of lymphoma cases, predominantly arising from B-cells, while HL comprises the remaining 10% [1]. Globally, lymphoma accounts for nearly 5% of cancer diagnoses, positioning it as the sixth most common cancer and underscoring its significant clinical relevance [2]. Although diagnostic and therapeutic advances, including quantitative approaches such as SUV analysis, have provided reasonable accuracy in lymphoma evaluation, precisely diagnosing, staging, and managing lymphoma remains challenging due to its heterogeneous clinical and pathological manifestations [3,4].

Positron emission tomography combined with computed tomography using 18F-fluorodeoxyglucose (18F-FDG PET/CT) has significantly transformed lymphoma management by providing comprehensive metabolic and anatomical insights essential for initial staging, evaluating therapeutic responses, and post-therapy surveillance [5,6]. This imaging modality enables clinicians to quantify tumor metabolic activity using metrics such as Total Metabolic Tumor Volume (TMTV) and Total Lesion Glycolysis (TLG), which are robust predictors of patient prognosis and therapeutic outcomes [7–10]. However, current practices to determine TMTV using commercially available software such as MIM software’s lesionID or Siemens' syngo rely heavily on threshold-based methods, initially delineating tumor volumes followed by manual exclusion of physiologically high-uptake organs and presumed normal tissues. Typically used thresholds include SUV maximum percentages (e.g., 41% or 50%), fixed SUVs (e.g., 2.5 or 3), or thresholds defined by a liver VOI (mean plus two standard deviations). Thresholds set too low can inadvertently include normal tissue, whereas thresholds set too high risk missing tumors with low SUV uptake. Consequently, significant variability in TMTV measurements occurs between observers, highlighting the need for accurate differentiation between tumor and normal volumes [11–13].

Radiomics is a quantitative imaging analysis technique that involves extracting numerous high-dimensional features from medical images, such as texture, shape, intensity, and wavelet-transformed features, which comprehensively characterize tumor heterogeneity and biological properties beyond what is visually apparent to clinicians [14–16]. These radiomics features allow for a more objective and reproducible assessment of tumors by providing detailed insights into their underlying biology. On the other hand, machine learning (ML) refers to computational algorithms and statistical models capable of recognizing complex patterns within large datasets and learning from them to perform specific tasks such as classification, regression, or clustering. In medical imaging, ML techniques utilize radiomics features as input variables to develop predictive models that can accurately differentiate pathological tissues, predict patient prognosis, or assess therapeutic responses [23–26]. Hence, while radiomics provides the essential quantitative descriptors extracted from imaging data, ML offers the analytical tools needed to interpret these descriptors and translate them into clinically meaningful predictions and decisions.

Radiomic features from 18F-FDG PET/CT imaging have demonstrated substantial promise, significantly improving lymphoma prognostic predictions and therapeutic response assessments [17–22]. Recent advancements in artificial intelligence (AI), particularly ML and deep learning (DL), have further accelerated radiomics research by automating complex data analysis tasks, including lesion segmentation and patient outcome predictions [23–26]. AI-driven radiomics models have demonstrated superior predictive accuracy for survival outcomes such as progression-free survival (PFS) and overall survival (OS), compared to traditional imaging metrics alone [27–31].

Previous radiomics studies have primarily applied binary classifications due to the inherent nature of radiomics-derived features reflecting dominant tissue characteristics within a given volume. Specifically, radiomics features tend to represent the predominant tissue type, resulting in binary classifications of either tumor or normal tissue. Consequently, volumes with a higher proportion of tumor tissue typically yield tumor-oriented features, whereas those with predominantly normal tissue produce normal-oriented features. This characteristic has led to numerous studies successfully differentiating tumor and normal tissues using radiomics features combined with ML methods. For instance, Hsu et al. (2018) achieved an overall classification accuracy of 90% in differentiating tumors from normal tissues using radiomic features extracted from CT images [32]. Similarly, Zhang et al. (2024) successfully classified gross tumor volume (GTV) and normal liver tissue in hepatocellular carcinoma with an accuracy of 0.98 using ML approaches applied to CT images [33]. Zhang et al. (2024) also developed a stacking ensemble model that integrated multiple ML algorithms, achieving superior performance in classifying GTV (AUC = 0.93), brainstem (AUC = 0.93), and normal brain tissue (AUC = 0.94) using CT images [34]. Additionally, Pei et al. (2024) demonstrated promising results in distinguishing cervical cancer tumors from normal uterine tissues using radiomic features extracted from CT images, achieving AUC values ranging from 0.89 to 0.92 [35].

Despite extensive research utilizing PET or CT imaging separately, no studies have yet applied radiomics methods combining both PET and CT images specifically to differentiate tumor and normal volumes in lymphoma patients undergoing F-18 FDG PET/CT imaging. Moreover, existing research predominantly addresses clearly defined tumor and normal tissues, thereby neglecting the mixed tumor-normal regions inherently captured during threshold-based clinical delineation processes. Such oversight can significantly impact prognostic evaluations and subsequent therapeutic strategies due to potential misclassification or inaccurate volume measurements.

Our research uniquely addresses these critical gaps by developing an innovative radiomics-based ML system explicitly designed to differentiate pure tumor, pure normal tissue, and critically, mixed tumor-normal regions within lymphoma lesions. Unlike previous research, our methodology explicitly accounts for mixed volumes, enabling a more accurate assessment of tumor extent and significantly enhancing the precision of prognostic modeling.

This research aims to develop and validate a radiomics-based machine learning framework for differentiating tumor, normal, and mixed tumor-normal regions in lymphoma patients using 18F-FDG PET/CT imaging. By addressing the limitations of conventional threshold-based segmentation approaches through comprehensive feature extraction, ensemble machine learning methods, and rigorous external validation, our study seeks to enhance tumor classification accuracy and improve prognostic prediction for lymphoma patient management.

2. Materials and methods

This study employed a comprehensive radiomics-based machine learning framework designed to differentiate tumor, normal, and mixed tumor-normal regions in lymphoma patients using 18F-FDG PET/CT imaging, as illustrated in Figure 1. The methodological workflow consisted of ten sequential steps: (1) quantitative imaging acquisition using standardized F-18 FDG PET/CT protocols, (2) normal organ segmentation to exclude physiological uptake, (3) tumor detection and segmentation through expert manual delineation, (4) PET-CT image alignment for multimodal analysis, (5) obtaining Total Metabolic Tumor Volume (TMTV) data according to established threshold criteria, (6) acquisition of normal volume data by excluding physiological organs, (7) tumor phenotype quantification through comprehensive radiomics feature extraction, (8) data integration and application of machine learning algorithms, (9) acquisition of refined TMTV excluding normal organs and non-tumor volumes, and (10) prediction of patient prognosis including tumor recurrence and mortality outcomes.

The framework combined both imaging-derived radiomics features and clinical variables to develop robust prognostic models, with particular emphasis on addressing the challenge of mixed tumor-normal regions commonly encountered in clinical practice. External validation was performed using an independent dataset to assess model generalizability and clinical applicability. The following sections detail each component of this comprehensive methodology.

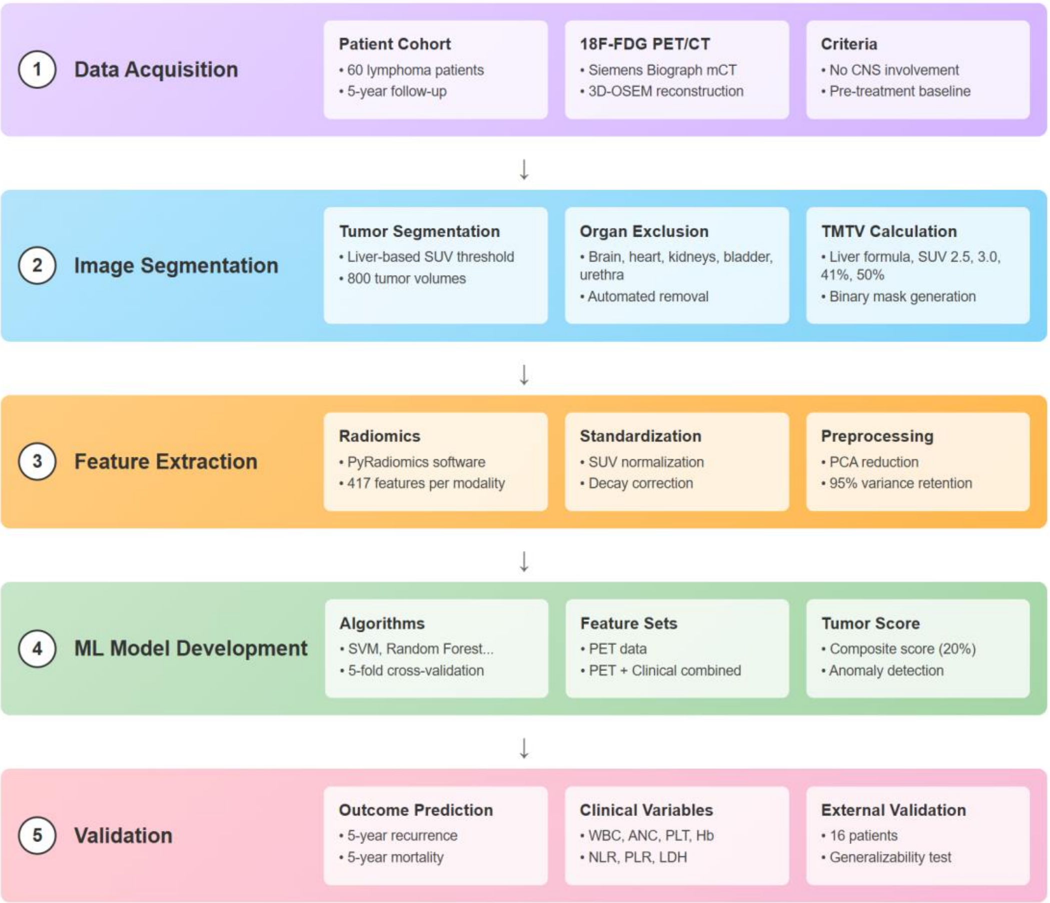


Figure 1. Comprehensive workflow of the radiomics-based machine learning framework for lymphoma tumor classification and prognosis prediction. The methodology encompasses ten sequential steps from initial F-18 FDG PET/CT imaging acquisition to final prognostic prediction, integrating tumor phenotype quantification, normal organ exclusion, and machine learning-based analysis for enhanced clinical decision-making in lymphoma management.

2.1. Patient Selection and Imaging Protocol

This retrospective study analyzed F-18 FDG PET/CT images from 60 patients diagnosed with lymphoma who underwent initial staging at Ewha Womans University Mokdong Hospital, Seoul, South Korea, between 2012 and 2018. Inclusion criteria comprised histologically confirmed lymphoma diagnosis, initial staging F-18 FDG PET/CT scan performed before treatment initiation, complete clinical and follow-up data available for at least 5 years, age ≥ 18 years, and adequate image quality for radiomics feature extraction. Exclusion criteria included patients with central nervous system (brain) involvement at initial diagnosis, previous history of malignancy within 5 years prior to lymphoma diagnosis, concurrent active malignancy, inadequate imaging quality preventing reliable segmentation, incomplete clinical or follow-up data, and patients who received treatment prior to baseline PET/CT imaging. Patient recurrence and survival data within a 5-year period were collected. Patient demographic and clinical characteristics are summarized in Table 1.

F-18 FDG PET/CT was performed using a single PET/CT camera system (Siemens Biograph mCT with 128-slice CT, Siemens Medical Solutions, Knoxville, TN, USA). Patients fasted for at least six hours before F-18 FDG PET/CT scanning. FDG administration was done when whole blood glucose levels were less than 140 mg/dl. F-18 FDG PET/CT images were acquired from the skull base to mid-thigh, 60 min after intravenous FDG injection (5.18 MBq/kg). CT images without contrast agent were obtained first using a 120 kVp tube voltage, a 50 mAs tube current, and a 1.2 pitch. PET images were then acquired for two min per bed position (five to seven positions) under a 3D emission mode. PET images were reconstructed into 200 × 200 ma-trices and 3.4 mm × 3.4 mm pixel sizes (3.0 mm slice thickness) using a 3D-OSEM iterative algorithm (2 iterations and 21 subsets) with time of flight and point spread function.

A priori power analysis was conducted using G*Power 3.1.9.4 to determine the adequacy of our sample size for the planned statistical analyses. For the tumor classification task comparing multiple radiomics feature sets, we performed a power calculation for means difference between two independent groups with the following parameters: effect size (Cohen's d) = 0.5 (medium effect), α error probability = 0.05, power (1-β error probability) = 0.8, and allocation ratio N2/N1 = 1. The analysis indicated that a minimum total sample size of 128 volumes would be required to detect clinically meaningful differences with adequate statistical power (actual power = 0.801).

For prognostic prediction analyses, the same power calculation parameters were applied to compare outcome groups (recurrence vs. non-recurrence, mortality vs. survival). The analysis indicated that a minimum of 64 patients per group (total n=128) would be required for adequate statistical power. Our cohort included 60 patients with 16 recurrence events (26.7%) and 5 mortality events (8.3%), resulting in sample sizes substantially below the recommended threshold for robust between-group comparisons.

This represents a significant limitation of our prognostic prediction analysis, as neither recurrence prediction (16 vs. 44 patients) nor mortality prediction (5 vs. 55 patients) achieved the minimum sample size requirements derived from power analysis. Our study included 4,950 total volumes for classification analysis, substantially exceeding the minimum required sample size, while the prognostic component was underpowered according to conventional statistical guidelines.

Table 1. Clinical and demographic characteristics of 60 lymphoma patients included in the study.

Characteristic	Value	(%)
Total Patients	60	100.0
Age, years		

Mean ± SD	66.4 ± 16.8	
Median	68	
Sex		
Female	32	53.3
Male	28	46.7
Physical Characteristics		
Height, cm (mean ± SD)	163.4 ± 10.2	
Weight, kg (mean ± SD)	62.3 ± 13.5	
Lymphoma Subtype		
Diffuse Large B-cell Lymphoma	26	43.3
Hodgkin Lymphoma	12	20.0
Other/Unspecified	12	20.0
MALT Lymphoma	2	3.3
Lymphoblastic Lymphoma	2	3.3
Angioimmunoblastic T-cell Lymphoma	2	3.3
Follicular Lymphoma	1	1.7
NK/T-cell Lymphoma	1	1.7
Mantle Cell Lymphoma	1	1.7
Burkitt Lymphoma	1	1.7
Histological Classification		
Type 1 (Hodgkin)	9	15.0
Type 2 (Non-Hodgkin, intermediate)	3	5.0
Type 3 (Non-Hodgkin, aggressive)	48	80.0
Lugano Staging		
Stage I	3	5.0
Stage IE	3	5.0

Stage II	16	26.7
Stage IIE	6	10.0
Stage III	10	16.7
Stage IV	22	36.7
Clinical Features		
B symptoms present	10	16.7
Bone marrow involvement	11	18.3
Spleen involvement	11	18.3
Advanced Stage Disease (III-IV)	32	53.3
Extranodal Disease (E staging)	9	15.0

2.2. Segmentation Strategy and TMTV-Based Dataset Construction

Using MIM software (MIM Software Inc., Cleveland, OH, USA), tumor volumes were initially segmented by applying a liver-based SUV threshold to extract metabolically active regions suspected of malignancy. Normal physiological uptake areas, such as those in the brain, myocardium, kidneys, and bladder, were subsequently excluded manually by experienced physicians to generate the final tumor masks. Excluding FDG uptake in normal organs is an essential step in avoiding false-positive findings during the calculation of Total Metabolic Tumor Volume (TMTV) and Total Lesion Glycolysis (TLG), ensuring that only true pathological lesions are accurately assessed [36, 37]. Segmentation of physiologically active normal organs was performed using a combination of automated and manual approaches. For the brain, heart, kidneys, and bladder, initial contours were generated using Oncosoft software (Oncosoft, Manteia), followed by manual refinement to improve anatomical accuracy. The ureters, which were not supported by automatic segmentation, were delineated entirely through manual contouring.

This supplementary analysis was conducted to clarify the necessity of excluding organs with physiological FDG uptake, which often leads to overestimation of tumor volume due to their intense metabolic activity. Without such exclusion, threshold-based segmentation methods may mistakenly include these normal tissues, significantly inflating calculated TMTV values. The quantitative impact of this exclusion was evaluated across five threshold strategies by comparing segmented volumes with and without organ removal, as described in Appendix 1.

Total Metabolic Tumor Volume (TMTV) was calculated using various clinically validated thresholding strategies, including absolute SUV thresholds (2.5 and 3.0), relative thresholds based on SUV_{max} (41% and 50%), and a liver-referenced threshold defined as the mean liver SUV plus two standard deviations. These methods are consistent with established clinical workflows for tumor delineation. For each method, binary masks were generated and refined by removing regions overlapping with physiological organs to ensure that the final TMTV masks represented only tumor volumes.

In this study, normal tissue volumes were not arbitrarily defined by low SUVs. Instead, regions that exceeded SUV thresholds but were not considered malignant based on anatomical location or physiological uptake were labeled as 'normal.' This approach mirrors actual diagnostic workflows and was intended to challenge the model to distinguish tumor from non-tumor tissues using radiomic features beyond simple intensity metrics. By including metabolically active but clinically non-malignant regions in the normal class, the model was encouraged to learn more nuanced structural and textural patterns, thereby improving its generalizability and clinical applicability. Finally, normal volumes were acquired by systematically excluding organs with high physiological uptake and annotated labels from the TMTV data obtained using each thresholding method.

2.3. Radiomics Feature Extraction

Prior to radiomics feature extraction, PET images underwent standardized SUV quantification to ensure accurate metabolic assessment across all segmented regions. The conversion from raw PET pixel values to standardized uptake values was performed using patient-specific parameters extracted from DICOM header information, including administered dose, injection time, acquisition time, patient weight, and radiopharmaceutical decay correction factors. (detailed SUV calculation methodology provided in Appendix 2).

Radiomics features were extracted from segmented PET and CT images using PyRadiomics software (version 3.1.0). A total of 417 quantitative features per imaging modality were obtained, encompassing shape, first-order statistical, and texture features. The shape features included elongation, flatness, sphericity, and surface area, while first-order statistics comprised energy, entropy, mean, median, kurtosis, skewness, minimum, and maximum. Texture features were derived from Gray Level Co-occurrence Matrix (GLCM), Gray Level Dependence Matrix (GLDM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), and Neighboring Gray Tone Difference Matrix (NGTDM). Additionally, Laplacian of Gaussian (LoG) filtered features were extracted at three sigma levels (1 mm, 2 mm, and 3 mm) to capture multiscale spatial patterns.

For PET images, all radiomics features were calculated from the standardized SUV-converted images rather than raw pixel intensities, ensuring that the extracted features represented true metabolic characteristics and enabling meaningful quantitative comparison across different patients and acquisition parameters.

Radiomics analysis was performed on a total of 800 tumor volumes and 4,150 normal tissue volumes, all of which were manually delineated and verified to ensure labeling consistency. This large and heterogeneous sample allowed comprehensive feature extraction and robust modeling across a wide range of tissue characteristics.

2.4. Feature Processing and Tumor Classification Model

Extracted radiomics features underwent normalization to standardize their scale, followed by dimensionality reduction using Principal Component Analysis (PCA). To address the high-dimensional nature of radiomics data and potential multicollinearity issues, PCA was implemented to retain principal components that cumulatively captured 95% of the total variance. The effectiveness of PCA was validated through systematic comparison with non-PCA approaches, demonstrating minimal performance differences (detailed analysis provided in Appendix 3).

A total of 800 tumor volumes and 4,150 normal volumes were utilized to construct the dataset. The dataset was randomly partitioned into training (70%) and testing (30%) subsets. All quantitative evaluation metrics for model performance were derived exclusively from predictions on the test dataset to ensure objective and unbiased assessment.

To systematically evaluate the contribution of various feature types, four experimental conditions were established: (1) CT radiomics features only, (2) PET radiomics features only, (3) combined PET/CT radiomics features, (4) SUV parameters from PET metrics only SUV_{max}, SUV_{min}, SUV_{mean}. This comprehensive setup allowed comparative analyses of anatomical, functional, and metabolic information derived from different imaging modalities.

For each feature set, five machine learning algorithms (AdaBoost, Decision Tree, Gradient Boosting, Random Forest, and XGBoost) were trained and optimized using stratified 5-fold cross-validation combined with GridSearchCV for hyperparameter tuning. Cross-validation was implemented with fixed random seeds to ensure reproducibility, and hyperparameter optimization was performed systematically for each algorithm to prevent overfitting. After training, feature importance was extracted from each model and dataset, enabling evaluation and comparison of influential features across different experimental conditions.

The complete tumor-normal classification pipeline is illustrated in Figure 2, demonstrating the systematic approach from volume dataset construction to external validation.

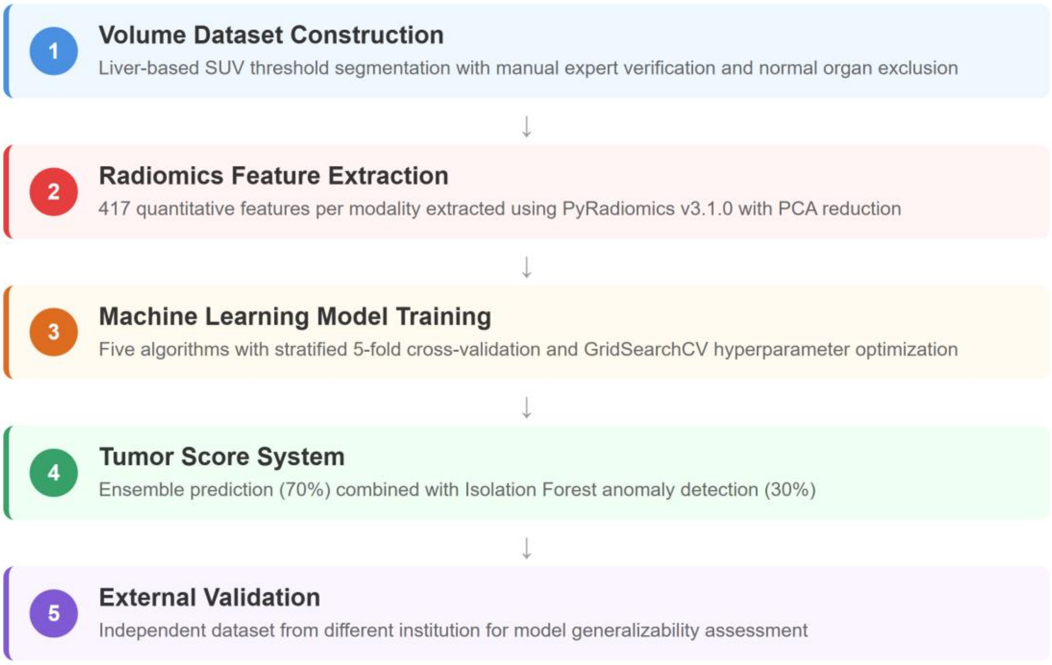

1
Volume Dataset Construction
Liver-based SUV threshold segmentation with manual expert verification and normal organ exclusion
↓
2
Radiomics Feature Extraction
417 quantitative features per modality extracted using PyRadiomics v3.1.0 with PCA reduction
↓
3
Machine Learning Model Training
Five algorithms with stratified 5-fold cross-validation and GridSearchCV hyperparameter optimization
↓
4
Tumor Score System
Ensemble prediction (70%) combined with Isolation Forest anomaly detection (30%)
↓
5
External Validation
Independent dataset from different institution for model generalizability assessment

Figure 2. Five-stage radiomics pipeline for tumor-normal classification: dataset construction (800 tumor, 4,150 normal volumes), feature extraction, machine learning model training, ensemble scoring system, and external validation.

2.5. Tumor Score

A tumor scoring system was developed to accurately classify segmented regions into tumor or non-tumor tissues by quantitatively integrating ensemble machine learning predictions and anomaly detection results. Initially, the probability of a region being tumor tissue was computed by averaging the prediction probabilities derived from the five trained machine learning algorithms (AdaBoost, Decision Tree, Gradient Boosting, Random Forest, and XGBoost). This averaged probability was designated as the ensemble probability.

To improve robustness and reduce model-specific bias, a soft voting ensemble classifier was created by averaging prediction probabilities from the five base models. This ensemble strategy has been widely used in radiomics-based predictive modeling to improve performance and generalizability [38-40], including models that combine handcrafted radiomics and deep learning for survival prediction, radiomics-combined classifiers for DCIS assessment, and stacking ensemble models for brain metastasis segmentation.

Additionally, anomaly detection probability, derived from the Isolation Forest algorithm, was calculated by normalizing anomaly scores into probability values indicating the likelihood of a data point representing an anomaly (i.e., tumor region).

These two probabilities were then combined using weighted averaging, with greater weight given to the ensemble machine learning predictions to enhance prediction reliability. Specifically, the tumor score was calculated according to the following formula:

$$Tumor\ Score\ = \alpha\ \times P_{ensemble} + \beta \times\ P_{panomaly} \qquad (1)$$

where $P_{ensemble}$ represents the averaged prediction probability from five machine learning algorithms (AdaBoost, Decision Tree, Gradient Boosting, Random Forest, and XGBoost), $P_{panomaly}$ denotes the normalized anomaly score from Isolation Forest algorithm, α is the weight coefficient for ensemble probability (set to 0.7), and β is the weight coefficient for anomaly probability (set to 0.3).

The calculated tumor scores were subsequently used to classify segmented regions into tumor or normal tissues by applying predefined threshold values (0.05, 0.10, and 0.20), thus enabling systematic and objective determination of tumor presence.

Binary classification was conducted by applying various threshold values (t) to the calculated tumor score. Specifically, each segmented region was assigned a binary prediction based on its tumor score relative to these thresholds:

$$\hat{\mu} = \begin{cases} 1,\ if\ Tumor\ Score\ \geq t \\ 0,\ if\ \ Tumor\ Score\ < t \end{cases} \qquad (2)$$

where $\hat{\mu}$ is the predicted binary class label (1 for tumor, 0 for normal tissue), $Tumor\ Score$ is the calculated combined score from equation (1), and tt t represents the threshold value (0.05, 0.10, or 0.20) used for binary classification.

In this study, the threshold values were set at 0.05, 0.10, and 0.20. For each threshold value, the proportion of samples predicted as tumor was calculated to quantitatively evaluate the tumor prediction ratio, thereby assessing the robustness and sensitivity of tumor classification across different thresholds.

2.6. Model Evaluation

The predictive performance of each model was assessed using widely accepted evaluation metrics: Accuracy, Precision, Recall (Sensitivity), F1 Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a comprehensive understanding of model capability in distinguishing tumor from non-tumor tissue across varying clinical scenarios.

Mathematically, the evaluation metrics are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\,Score \ = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was calculated to evaluate the discriminative ability of each model across all possible threshold values, providing a threshold-independent measure of classification performance.

Thresholds for binary tumor prediction were established at 0.05, 0.10, and 0.20 for the tumor scores, reflecting different degrees of classification confidence and enabling the evaluation of model robustness across clinically relevant cutoffs.

To assess the statistical significance of model performance differences across imaging modalities and feature types, we conducted a comprehensive analysis involving both parametric and non-parametric tests. For tumor classification models, one-way analysis of variance (ANOVA) was employed to determine whether the predictive performance metrics significantly differed among four groups: CT radiomics, PET radiomics, combined PET/CT radiomics, and SUV-only models. ANOVA assumptions were verified through tests of homogeneity and normality.

2.7. Prediction of Recurrence and Mortality

To predict five-year recurrence and mortality among lymphoma patients, several quantitative imaging biomarkers were utilized, including maximum standardized uptake value (SUV_{max}), minimum standardized uptake value (SUV_{min}), mean standardized uptake value (SUV_{mean}), metabolic tumor volume, and total lesion glycolysis (TLG). These features were extracted from PET images and used as input variables in the model training process.

In addition to imaging-derived features, clinical data were incorporated to evaluate the potential for improved prognostic prediction through multimodal integration. The clinical variables included routine hematologic and biochemical markers obtained from standard blood examinations, as well as pathological tumor characteristics. Table 2 summarizes the clinical variables used in this study, including white blood cell count (WBC), absolute neutrophil count (ANC), absolute lymphocyte count (ALC), platelet count (PLT), hemoglobin (Hb), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), lactate dehydrogenase (LDH), and cancer classification based on pathological subtype. These clinical parameters were selected based on their established prognostic significance in lymphoma, particularly their association with systemic inflammation, immune status, tumor burden, and disease aggressiveness.

Table 2. Clinical variables used for prognostic prediction modeling in lymphoma patients

Variable	**Description**	**Clinical Significance**
labData	Clinical test results from routine blood examinations	Source dataset for all hematologic and biochemical markers
Cancer Classification	Pathological subtype and histological categorization of the tumor	Defines tumor biology and guides treatment stratification
WBC	White Blood Cell Count - Total leukocyte count	Reflects systemic immune response and inflammation
ANC	Absolute Neutrophil Count - Absolute number of neutrophils	Indicator of infection risk and acute inflammatory status
ALC	Absolute Lymphocyte Count - Absolute number of lymphocytes	Marker of adaptive immune competence
PLT	Platelet Count - Total platelet count	Important for coagulation function and bone marrow health
Hb	Hemoglobin - Hemoglobin concentration in blood	Assesses oxygen-carrying capacity and identifies anemia

NLR	Neutrophil-to-Lymphocyte Ratio - Ratio of neutrophils to lymphocytes	Elevated values indicate systemic inflammation and poorer outcome
PLR	Platelet-to-Lymphocyte Ratio - Ratio of platelets to lymphocytes	Higher ratios correlate with adverse prognosis in cancer
LDH	Lactate Dehydrogenase - Enzyme released during tissue damage	Elevated levels reflect high tumor burden and aggressive disease

Two distinct modeling approaches were implemented: (1) imaging-only models using radiomics features alone, and (2) combined models integrating both imaging and clinical features. The clinical dataset demonstrated complete data integrity with no missing values across all variables and patients, ensuring reliable comparative analysis between modeling approaches. This comparative framework enabled assessment of the added value of clinical data integration for prognostic prediction accuracy.

Given the inherent imbalance between event (recurrence or mortality) and non-event cases, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the minority class, thereby improving model training stability and generalizability. Importantly, SMOTE was applied only to the training datasets, while the test datasets remained untouched to ensure unbiased performance evaluation. All model results and performance metrics reported in this study were evaluated using the original test datasets, following a data split of 70% for training and 30% for testing.

Table 3 presents the detailed distribution of samples across different experimental conditions. For recurrence prediction, the training set contained 48 samples without SMOTE (13 positive, 35 negative) and 70 samples with SMOTE (35 positive, 35 negative), representing a 45.80% increase in total samples and a 169.20% increase in positive cases. The positive ratio improved from 27.08% to 50.00%, achieving balanced class distribution. For mortality prediction, the training set expanded from 36 samples without SMOTE (3 positive, 33 negative) to 66 samples with SMOTE (33 positive, 33 negative), showing an 83.33% increase in total samples and an 11-fold increase in positive cases. The positive ratio increased from 8.33% to 50.00%, effectively addressing the severe class imbalance. Test sets remained unchanged across all conditions to maintain evaluation integrity.

Table 3. Summarizes the distribution of training and test sets used in this study, showing differences in sample counts and class ratios before and after SMOTE application.

Dataset	**Recurrence Prediction Model**			**Mortality Prediction Model**		
	Without SMOTE	**With SMOTE**	**Change**	**Without SMOTE**	**With SMOTE**	**Change**

Training Set						
Total Samples	48	70	45.80%	36	66	83.33%
Negative	35	35	-	33	33	-
Positive	13	35	169.20%	3	33	1000%
Positive Ratio	27.08%	50.00%	22.92%	8.33%	50.00%	41.67%
Test Set						
Total Samples	12	12	No change	24	24	No change
Negative	9	9	No change	22	22	No change
Positive	3	3	No change	2	2	No change

In recurrence prediction, the number of positive cases was relatively sufficient, and 21 synthetic samples were generated to achieve class balance. In contrast, mortality prediction had a much smaller number of positive samples; thus, 41 synthetic samples were added to the training data to establish class parity. This tailored oversampling strategy enabled the models to learn effectively from limited data while preserving the integrity of external evaluation.

Additionally, we trained recurrence and mortality prediction models using the same input features without applying SMOTE, to compare the impact of oversampling. To assess whether prediction performance varied significantly across tumor score thresholds (0.05, 0.10, 0.20), the Friedman test was applied. This non-parametric statistical test is appropriate for repeated measures designs and small sample sizes, especially when the assumption of normality cannot be guaranteed. For each of the five machine learning models (AdaBoost, Decision Tree, Gradient Boosting, Random Forest, and XGBoost), F1 scores were calculated under each threshold setting.

The Friedman test considered the models as blocks and tested the null hypothesis (H_0: model performance is consistent across thresholds) against the alternative hypothesis (H_1 : at least one threshold yields significantly different performance). Statistical significance was defined as $p < 0.05$.

2.8. Validation with External Data

To assess the generalizability and robustness of the proposed radiomics-based machine learning framework, external validation was conducted using an independent dataset from Ewha Womans University Seoul Hospital. This validation study aimed to evaluate the transferability of the developed models across different patient populations and institutional settings, thereby providing critical evidence for the clinical applicability of the proposed methodology.

The external validation cohort comprised 16 lymphoma patients who underwent 18F-FDG PET/CT imaging at Ewha Womans University Seoul Hospital. Following the same segmentation protocols established in the primary study, a total of 3,100 volumes were manually delineated and categorized, consisting of 2,432 normal tissue volumes and 668 tumor volumes.

All volumes in the external validation dataset were processed using identical radiomics feature extraction pipelines as described in Section 2.3, ensuring standardized quantitative analysis across both primary and validation cohorts. The same 417 radiomics features per imaging modality were extracted using PyRadiomics software (version 3.1.0), maintaining consistency in feature computation and preprocessing protocols.

The machine learning models trained on the primary dataset were directly applied to the external validation cohort without retraining or parameter modification, providing a stringent test of model generalizability. Performance evaluation encompassed the same metrics used in the primary analysis: Accuracy, Precision, Recall, F1 Score, and AUC-ROC.

The external validation was conducted across all four experimental conditions established in the primary study: (1) CT radiomics features only, (2) PET radiomics features only, (3) combined PET/CT radiomics features. This comprehensive evaluation framework enabled direct comparison of model performance between the primary training cohort and the independent validation dataset.

To assess whether predictive performance metrics differed significantly across imaging modalities and feature types in the external validation setting, a one-way Analysis of Variance (ANOVA) was employed. The ANOVA compared performance metrics among four groups: CT radiomics, PET radiomics, and combined PET/CT radiomics model. This statistical approach enabled objective evaluation of feature set contributions to predictive performance in an independent validation context.

ANOVA assumptions, including homogeneity of variance and normality of residuals, were verified through appropriate statistical tests. Post-hoc analysis using Tukey's honestly significant difference (HSD) test was performed when significant differences were detected, allowing for pairwise comparisons between specific imaging modality groups. Statistical significance was defined as $p < 0.05$ for all analyses.

3. Results

3.1. Tumor Classification Performance Comparison

The performance of machine learning models in differentiating tumor from normal tissue was assessed using four distinct datasets: CT radiomics, PET radiomics (excluding SUV parameters), combined PET/CT radiomics, and SUV parameters from PET data (Table 4, Figure 3 and Figure 4).

For the CT radiomics dataset, the models achieved an average accuracy of 93.78%, precision of 74.75%, recall of 44.42%, F1 score of 54.92%, and an AUC of 93.68%. The XGBoost classifier exhibited the highest overall performance with an accuracy of 94.78%, precision of 79.38%, recall of 52.92%, F1 score of 63.50%, and an AUC of 96.90%. Conversely, the Decision Tree model showed the lowest AUC (85.36%) and precision (56.83%), along with relatively low recall (43.33%), indicating its limited discriminative capability within this dataset.

In the PET radiomics dataset (excluding SUV features), models demonstrated improved recall and F1 score compared to CT alone, with an average accuracy of 91.99%, precision of 78.53%, recall of 65.80%, F1 score of 71.53%, and AUC of 92.09%. Again, the XGBoost model outperformed other classifiers, achieving the highest accuracy (93.47%), recall (70.54%), F1 score (76.70%), and AUC (96.05%). Decision Tree had the lowest performance metrics in this group, notably an accuracy of 89.25% and an AUC of 77.91%, underscoring its inferior predictive performance.

Utilizing the combined PET/CT radiomics dataset (excluding SUV features), the average accuracy was 93.26%, precision 77.65%, recall 54.09%, F1 score 63.58%, and AUC 93.16%. XGBoost maintained its superior performance, recording an accuracy of 94.23%, precision of 81.32%, recall of 60.99%, F1 score of 69.70%, and AUC of 96.39%. The Decision Tree model continued to demonstrate the weakest performance, particularly evident in its lowest recall (48.49%) and AUC (84.79%).

Finally, the SUV parameters from PET demonstrated notably reduced predictive performance across all metrics. Models averaged an accuracy of 88.18%, precision of 71.03%, recall of 40.18%, F1 score of 49.79%, and AUC of 87.04%. The Gradient Boosting model was the top performer within this dataset, achieving an accuracy of 89.12% and an AUC of 88.30%. AdaBoost exhibited particularly poor results, notably achieving the lowest recall (19.20%) and F1 score (31.16%), highlighting significant challenges in sensitivity when SUV parameters were included.

Overall, the XGBoost classifier consistently achieved the highest performance across all radiomics-based datasets (CT, PET, and combined PET/CT), demonstrating particularly strong F1 score and AUC values. The Decision Tree and AdaBoost classifiers were less reliable and exhibited markedly inferior performance in several key metrics. The incorporation of SUV parameters into the PET dataset consistently decreased performance, indicating limited additional predictive benefit from these features in conjunction with radiomics features alone.

Table 4. Performance comparison of tumor classification models using CT, PET, and combined PET/CT radiomics features and SUV model. Metrics evaluated include Accuracy, Precision, Recall, F1 Score, and AUC.

Data	Model	Accuracy	Precision	Recall	F1 score	AUC
CT radiomics	Random Forest	0.9392	0.9070	0.3250	0.4785	0.9653
	Decision Tree	0.9231	0.5683	0.4333	0.4917	0.8536
	Gradient Boosting	0.9435	0.7808	0.4750	0.5907	0.9562
	AdaBoost	0.9356	0.6875	0.4583	0.5500	0.9400
	XGBoost	0.9478	0.7938	0.5292	0.6350	0.9690
	average	0.9378	0.7475	0.4442	0.5492	0.9368
PET radiomics	Random Forest	0.9293	0.8488	0.6518	0.7374	0.9607
	Decision Tree	0.8925	0.6571	0.6161	0.6359	0.7791
	Gradient Boosting	0.9313	0.8220	0.7009	0.7566	0.9558
	AdaBoost	0.9116	0.7582	0.6161	0.6798	0.9484
	XGBoost	0.9347	0.8404	0.7054	0.7670	0.9605
	average	0.9199	0.7853	0.6580	0.7153	0.9209
PET/CT radiomics	Random Forest	0.9372	0.8684	0.4978	0.6329	0.9627
	Decision Tree	0.9179	0.6696	0.4849	0.5625	0.8479
	Gradient Boosting	0.9388	0.8243	0.5560	0.6641	0.9522
	AdaBoost	0.9266	0.7068	0.5560	0.6224	0.9315
	XGBoost	0.9423	0.8132	0.6099	0.6970	0.9639
	average	0.9326	0.7765	0.5409	0.6358	0.9316
PET data SUV	Random Forest	0.8871	0.6883	0.4732	0.5608	0.9044
	Decision Tree	0.8741	0.6444	0.3884	0.4847	0.8022
	Gradient Boosting	0.8912	0.7192	0.4688	0.5676	0.8830
	AdaBoost	0.8707	0.8269	0.192	0.3116	0.8732
	XGBoost	0.8857	0.6728	0.4866	0.5648	0.8890
	Average	0.8818	0.7103	0.4018	0.4979	0.8704

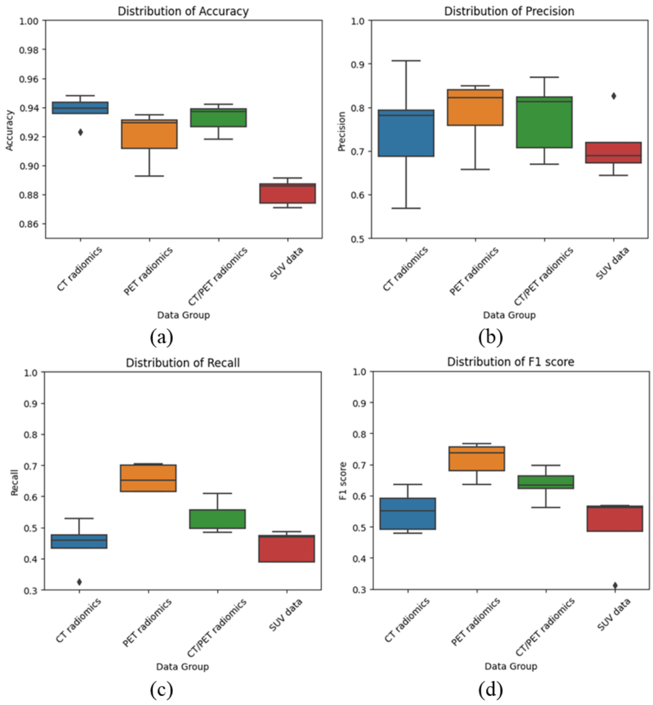
Table 5 summarizes the top radiomics features identified as most important by each machine learning model across CT, PET, and combined PET/CT datasets. In the CT radiomics models, original shape features, particularly Surface Volume Ratio and Sphericity, were consistently selected as the most significant features, with the Decision Tree and Gradient Boosting models showing notably high feature importance (0.2749 and 0.2488, respectively) for Surface Volume Ratio.

For the PET radiomics dataset, the log-sigma-1-mm-3D GLDM Dependence Variance texture feature emerged as the most influential in four out of five models. This feature exhibited particularly high importance in the Decision Tree (0.3255) and Gradient Boosting (0.3758) models. The AdaBoost model was unique in identifying the original first-order Median intensity as the most important feature, with an importance score of 0.1900.

In the combined PET/CT dataset, log-sigma-1-mm-3D GLDM Dependence Variance was again the dominant feature across four models, showing the highest importance in Decision Tree (0.2513), AdaBoost (0.2513), and Gradient Boosting (0.2090). Conversely, the XGBoost model selected the original shape Surface Volume Ratio feature with a moderate importance of 0.0477.

Table 5. Top-ranked radiomics features and their relative importance scores across five machine learning models for differentiating tumor and normal tissue in CT, PET, and combined PET/CT datasets.

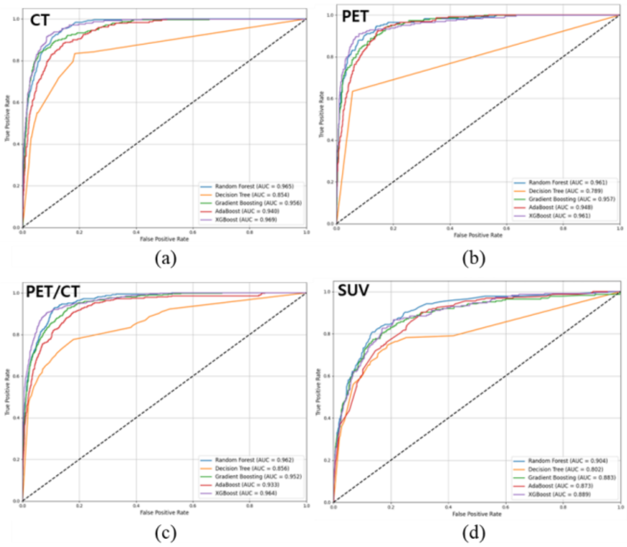
ML model	CT model		PET model		PET-CT model	
	Feature	Importance	Feature	Importance	Feature	Importance
AdaBoost	originalshapeSphericity	0.0550	originalfirstorderMedian	0.1900	log-sigma-1-mm-3DgldmDependenceVariance	0.2513
Decision Tree	originalshapeSurfaceVolumeRatio	0.2749	log-sigma-1-mm-3DgldmDependence Variance	0.3255	log-sigma-1-mm-3DgldmDependenceVariance	0.2513
Gradient Boosting	originalshapeSurfaceVolumeRatio	0.2488	log-sigma-1-mm-3DgldmDependence Variance	0.3758	log-sigma-1-mm-3DgldmDependenceVariance	0.2090
Random Forest	originalshapeSphericity	0.0251	log-sigma-1-mm-3DgldmDependence Variance	0.0329	log-sigma-1-mm-3DgldmDependenceVariance	0.0319
XGBoost	originalshapeSurfaceVolumeRatio	0.0536	log-sigma-1-mm-3DgldmDependence Variance	0.0813	originalshapeSurfaceVolumeRatio	0.0477



(a) (b)

(c) (d)

Figure 3. Comparison of Accuracy, Precision, Recall, and F1 score values across different radiomics-based tumor classification models.



(a) (b)

(c) (d)

Figure 4. Comparison of AUC values across different radiomics-based tumor classification models.

The one-way ANOVA revealed that accuracy (F = 21.9810, p < 0.0001), recall (F = 10.2216, p = 0.0005), and F1 score (F = 8.3510, p = 0.0014) showed statistically significant differences across the four radiomics data groups. These results suggest that the model performance in terms of overall accuracy and sensitivity is significantly affected by the type of feature set used. In contrast, precision (F = 0.6618, p = 0.5875) and AUC (F = 1.4785, p = 0.2581) did not exhibit statistically significant differences across groups, indicating that these metrics remained relatively stable regardless of the radiomics feature composition.

3.2. Tumor Score Threshold Analysis

Using the tumor score thresholds of 0.05, 0.10, and 0.20, the model demonstrated varying predictive capabilities. The results for the predicted proportion of tumors according to the tumor score threshold and the proportion of actual tumors included among the predicted proportion of tumors are shown in Table 6. At a threshold of 0.05, the model exhibited a tumor prediction accuracy of 100%, indicating complete reliability in identifying true tumor-positive volumes. When thresholds were increased to 0.10 and 0.20, the accuracy slightly decreased to 96%, suggesting increased selectivity in detecting volumes with tumor presence, while maintaining high reliability.

Table 6. Predicted proportion of tumors at various tumor score thresholds (0.05, 0.10, 0.20) and the corresponding proportion of actual tumors correctly identified within the predicted tumor samples.

Tumor Score Threshold	Tumor Prediction Rate	Percentage of volume containing tumor
0.05	77.00%	100%
0.10	45.14%	96%
0.20	16.71%	96%

3.3. Recurrence and Mortality Prediction

Prediction models for five-year recurrence and mortality demonstrated consistent performance across tumor score thresholds (0.05, 0.10, 0.20), with notable differences between models trained with and without the application of SMOTE (Table 7). To evaluate the potential benefit of incorporating clinical data, we additionally developed combined models integrating both imaging-derived radiomics features and clinical variables, with results presented in Table 8.

3.3.1 Recurrence and Mortality Prediction

For recurrence prediction, models trained with SMOTE achieved the highest accuracy of 75% using Decision Tree and Gradient Boosting at thresholds of 0.10 and 0.20, respectively. The mean accuracy improved from 62% at the 0.05 threshold to 68% at 0.10, and slightly declined to 67% at 0.20, suggesting relatively stable performance across thresholds. In contrast, models trained without SMOTE reached a maximum accuracy of 67% (XGBoost at 0.05 and 0.10), but showed a lower and more variable overall performance, with mean accuracies of 60%, 63%, and 55% at thresholds 0.05, 0.10, and 0.20, respectively. These findings indicate that SMOTE enhanced model robustness and performance in handling class imbalance, particularly for recurrence prediction.

For mortality prediction, models showed consistently high accuracy regardless of SMOTE application. When using SMOTE, the highest accuracy (86%) was observed for Decision Tree, Gradient Boosting, and XGBoost at the 0.10 threshold. The mean accuracy increased from 76% at 0.05, to 83% at 0.10, followed by a slight decrease to 79% at 0.20. Without SMOTE, the mean accuracy remained relatively stable across thresholds, consistently around 76–79%, although slight performance variations were seen in individual models. These results suggest that while SMOTE contributed to improved consistency in recurrence prediction, its effect on mortality prediction was minimal, as the models already performed well without additional class balancing.

Table 7. Comparison of five-year recurrence and mortality prediction accuracy across tumor score thresholds (0.05, 0.10, 0.20). Accuracy values are presented for each machine learning model (AdaBoost, Decision Tree, Gradient Boosting, Random Forest, XGBoost) under two conditions: with and without SMOTE application.

Model	**Threshold**	**Recurrance**		**Mortality**	
		With SMOTE	**Without SMOTE**	**With SMOTE**	**Without SMOTE**
AdaBoost	0.05	0.50	0.67	0.79	0.71
Decision Tree		0.67	0.50	0.79	0.79
Gradient Boosting		0.67	0.58	0.71	0.79
Random Forest		0.58	0.58	0.71	0.71
XGBoost		0.67	0.67	0.79	0.79
	Mean	0.62	0.60	0.76	0.76
AdaBoost	0.1	0.58	0.67	0.79	0.79
Decision Tree		0.75	0.67	0.86	0.71
Gradient Boosting		0.75	0.50	0.86	0.71
Random Forest		0.67	0.67	0.79	0.79
XGBoost		0.67	0.67	0.86	0.79
	Mean	0.68	0.63	0.83	0.76
AdaBoost	0.2	0.67	0.58	0.79	0.79
Decision Tree		0.75	0.67	0.79	0.79
Gradient Boosting		0.67	0.42	0.79	0.79
Random Forest		0.67	0.58	0.79	0.79
XGBoost		0.58	0.50	0.79	0.79
	Mean	0.67	0.55	0.79	0.79

3.3.2 Combined Imaging and Clinical Models

The integration of clinical variables with imaging features yielded mixed results depending on the specific prediction task and threshold setting (Table 5). For recurrence prediction with SMOTE, the combined models demonstrated variable performance, with mean accuracies of 72% at the 0.05 threshold, 60% at 0.10, and 47% at 0.20. Notably, the AdaBoost model achieved exceptional performance (92%) at the 0.05 threshold, representing a substantial improvement over imaging-only models. However, performance generally declined at higher thresholds, with some models showing marked deterioration, particularly Decision Tree and Gradient Boosting at the 0.20 threshold (33% accuracy each).

For recurrence prediction without SMOTE, the combined models showed more modest performance improvements, with mean accuracies of 57% at 0.05, 70% at 0.10, and 65% at 0.20. The Random Forest model consistently performed well across thresholds (75% accuracy), while other models showed more variable results.

In mortality prediction, the combined models demonstrated more stable performance patterns. With SMOTE application, mean accuracies were 79% at 0.05, declining to 57% at 0.10, and recovering to 74% at 0.20. Without SMOTE, mortality prediction remained remarkably consistent at 79% across all thresholds, suggesting that clinical data integration may provide complementary information for mortality risk assessment while maintaining stability in prediction performance.

Comparing the imaging-only and combined approaches, the integration of clinical data showed particular promise for specific scenarios: the AdaBoost model with clinical data achieved 92% accuracy for recurrence prediction at the 0.05 threshold (vs. 50% for imaging-only), representing an 84% relative improvement. However, this benefit was not consistently observed across all models and thresholds, suggesting that the value of clinical data integration may be model-dependent and require careful optimization of feature selection and weighting strategies.

Table 8. Comparison of five-year recurrence and mortality prediction accuracy using combined imaging and clinical data across tumor score thresholds (0.05, 0.10, 0.20). Accuracy values are presented for each machine learning model under two conditions: with and without SMOTE application.

Model	Threshold	Accuracy (with SMOTE)		Accuracy (without SMOTE)	
		Recurrance	Mortality	Recurrance	Mortality
AdaBoost	0.05	0.92	0.79	0.42	0.79
Decision Tree		0.58	0.79	0.58	0.79
Gradient Boosting		0.83	0.79	0.50	0.79
Random Forest		0.58	0.79	0.75	0.79
XGBoost		0.67	0.79	0.58	0.79

	Mean	0.72	0.79	0.57	0.79
AdaBoost		0.75	0.43	0.83	0.79
Decision Tree		0.42	0.43	0.5	0.79
Gradient Boosting	0.1	0.67	0.43	0.75	0.79
Random Forest		0.58	0.79	0.75	0.79
XGBoost		0.58	0.79	0.67	0.79
	Mean	0.60	0.57	0.7	0.79
AdaBoost		0.75	0.71	0.75	0.79
Decision Tree		0.33	0.71	0.58	0.79
Gradient Boosting	0.2	0.33	0.71	0.67	0.79
Random Forest		0.50	0.79	0.75	0.79
XGBoost		0.42	0.79	0.50	0.79
	Mean	0.47	0.74	0.65	0.79

3.3.3 Statistical Evaluation

To evaluate whether these performance variations were statistically significant across tumor score thresholds, a Friedman test was conducted for each outcome. For recurrence prediction, no statistically significant difference in performance was observed across thresholds, both with SMOTE ($\chi^2(2) = 3.8750$, $p = 0.1441$) and without SMOTE ($\chi^2(2) = 3.8750$, $p = 0.1441$), supporting the overall stability of model performance.

However, for mortality prediction with SMOTE, a statistically significant difference was identified ($\chi^2(2) = 6.6154$, $p = 0.0366$), suggesting that tumor score thresholds may influence model performance in this setting. No significant difference was found for mortality prediction without SMOTE ($\chi^2(2) = 2.0000$, $p = 0.3679$). These results imply that while recurrence prediction remains robust across thresholds, mortality prediction performance may vary depending on the chosen tumor score threshold, especially when SMOTE is applied.

3.4 External validation

External validation demonstrated variable performance across different imaging modalities and machine learning approaches (Table 9). PET radiomics models achieved the highest overall performance, with an average accuracy of 90.34%, precision of 68.95%, recall of 40.4%, F1 score of 49.81%, and AUC of 88.52%. CT radiomics models showed moderate performance with an average accuracy of 88.02%, but exhibited lower recall (17%) and F1 score (25.33%), indicating reduced sensitivity in tumor detection. Combined PET/CT radiomics models yielded intermediate results with an average accuracy of 88.71%, precision of 65.42%, recall of 25.8%, F1 score of 34.29%, and AUC of 83.48%.

Among individual models, the XGBoost classifier demonstrated consistent superior performance across all radiomics datasets. For CT radiomics, XGBoost achieved the highest accuracy (89.04%) and AUC (80.75%), while maintaining reasonable precision (62.86%). In the PET radiomics dataset, XGBoost recorded an accuracy of 90.60% and the highest AUC (92.99%), demonstrating robust discriminative capability. For combined PET/CT radiomics, XGBoost again showed the best overall performance with an accuracy of 89.70% and AUC of 88.48%.

Conversely, the Decision Tree model consistently exhibited the weakest performance across all datasets, particularly evident in the CT radiomics validation where it achieved the lowest accuracy (86.14%) and AUC (66.31%). This pattern was consistent with observations from the primary training dataset, reinforcing the reliability of model performance rankings across different cohorts.

One-way ANOVA analysis revealed statistically significant differences in performance metrics across imaging modalities in the external validation cohort. Accuracy showed significant variation among the three radiomics approaches (F = 8.42, p = 0.0028), with PET radiomics demonstrating superior performance compared to CT-only and combined approaches. Recall differences were highly significant (F = 15.73, p < 0.0001), reinforcing the superior sensitivity of PET-based models for tumor detection. AUC values also differed significantly across modalities (F = 6.89, p = 0.0058), confirming the discriminative advantage of PET radiomics in independent validation.

These statistical findings support the robustness of PET radiomics features for lymphoma tumor classification and validate the methodological framework's transferability across different institutional settings, despite the observed performance decline inherent to external validation scenarios.

Table 9. External validation performance of radiomics-based tumor classification models across CT, PET, and combined PET/CT datasets. Results represent direct application of models trained on the primary dataset to an independent cohort from Ewha Womans University Seoul Hospital.

Data	Model	Accuracy	Precision	Recall	F1 score	AUC
CT radiomics	Random Forest	0.8831	0.5882	0.1000	0.1709	0.7908
	Decision Tree	0.8614	0.3913	0.2700	0.3195	0.6631
	Gradient Boosting	0.8855	0.5926	0.1600	0.252	0.7891
	AdaBoost	0.8807	0.8200	0.1000	0.1980	0.7623
	XGBoost	0.8904	0.6286	0.2200	0.3259	0.8075
	average	0.8802	0.6401	0.1700	0.2533	0.7626
PET radiomics	Random Forest	0.9072	0.8108	0.3000	0.438	0.9150
	Decision Tree	0.8916	0.5581	0.4800	0.5161	0.7698

	Gradient Boosting	0.9084	0.7609	0.3500	0.4795	0.9253
	AdaBoost	0.9036	0.6282	0.4900	0.5506	0.8859
	XGBoost	0.9060	0.6897	0.4000	0.5063	0.9299
	average	0.9034	0.6895	0.4040	0.4981	0.8852
PET/CT radiomics	Random Forest	0.8952	0.7167	0.215	0.3308	0.8543
	Decision Tree	0.8645	0.4269	0.365	0.3935	0.7412
	Gradient Boosting	0.8892	0.5816	0.285	0.3826	0.8554
	AdaBoost	0.8898	0.9048	0.095	0.1719	0.8382
	XGBoost	0.8970	0.6408	0.3300	0.4356	0.8848
	average	0.8871	0.6542	0.2580	0.3429	0.8348

4. Discussions

The proposed framework addresses key limitations in current threshold-based segmentation practices by integrating metabolic and morphological features from both PET and CT modalities, combined with ensemble ML classifiers and anomaly detection strategies. Our findings demonstrate high predictive performance across classification and prognostic tasks, suggesting strong clinical utility and translational potential.

4.1. Interpretation of Model Performance

Among the tested models, CT-only radiomics classifiers achieved the highest average accuracy (93.78%) and AUC (0.9368), with shape-based features such as Surface Volume Ratio and Sphericity consistently contributing most to predictive performance. This aligns with previous studies (e.g., Zhang et al., 2024) in which CT shape descriptors were instrumental in distinguishing tumor from normal tissue [33]. However, the CT models exhibited low recall (44.42%), indicating a tendency to overlook true tumor-positive volumes, likely due to CT's limited ability to reflect underlying metabolic activity.

In contrast, PET radiomics models demonstrated higher recall (65.80%) and F1 scores (71.53%) at the expense of slightly reduced accuracy (91.99%). Texture-based features such as log-sigma-1-mm-3D GLDM Dependence Variance were dominant, indicating that PET-derived features effectively capture the heterogeneous biological properties of lymphoma lesions. Notably, the PET-only models outperformed CT in detecting metabolically active tumors, reinforcing the clinical value of PET in lymphoma imaging and consistent with reports by Driessen et al. (2023) and Yuan et al. (2023).

The combined PET/CT models, while yielding marginally improved accuracy (93.26%) and AUC (0.9316), did not enhance recall or F1-score compared to PET alone. This may be attributable to feature competition or dilution between modalities, whereby PET's sensitivity is offset by the morphological dominance of CT descriptors. These findings suggest that naive integration of modalities may not necessarily yield additive benefits unless fusion methods are optimized.

Compared to previous studies, our proposed model demonstrated competitive or superior performance across various imaging modalities and tumor classification tasks. For instance, Hsu (2018) reported an accuracy of 90.00% using PET-based radiomics on 332 VOIs [32]. Zhang et al. (2024) published two separate studies: one achieving an AUC of 0.9978 using contrast-enhanced CT on 208 VOIs [33], and another reporting an AUC of 0.9280 using CT with MRI fusion on 339 VOIs [34]. Pei et al. (2023) also demonstrated high classification performance with an AUC of 0.9190 using CT radiomics on a large-scale dataset of 4950 VOIs [35].

In comparison, as summarized in Table 10, our study evaluated multiple radiomics models based on PET, CT, combined PET/CT, and SUV features across 60 patients. The CT-only model achieved the highest accuracy (94.78%) and AUC (0.9690), followed by the combined PET/CT model (ACC: 94.23%, AUC: 0.9639), the PET-only model (ACC: 93.47%, AUC: 0.9607), and the SUV-only model (ACC: 89.12%, AUC: 0.9044).

Table 10. Comparison of tumor classification performance across different radiomics-based studies. The table summarizes patient sample sizes, imaging modalities, and reported performance metrics (accuracy or AUC) for each study.

Author	**Patient Number**	**Imaging Modality**	**PERFORMANCE**
Chih-Yang Hsu (2018)	38 (332 VOIs)	*PET	ACC 0.9000
Huai-wen Zhang (2024)	104 (208 VOIs)	Enhanced CT	AUC 0.9978
Huai-wen Zhang (2024)	113 (339 VOIs)	CT (using MRI fusion)	AUC 0.9280
Jinghong Pei (2023)	117 (Total 4950 VOIs)	CT	AUC: 9190
Our Study (2025)	60	PET, CT	CT MODEL ACC: 0.9478 AUC: 0.9690 PET MODEL ACC: 0.9347 AUC: 0.9607 PET/CT MODEL ACC: 0.9423 AUC: 0.9639 SUV MODEL ACC: 0.8912 AUC: 0.9044

4.2. External Validation and Clinical Generalizability

The modest accuracy reduction observed for PET radiomics (1.65 percentage points) compared to the more substantial decline in CT radiomics (5.76 percentage points) suggests that metabolic features derived from PET imaging may be more robust to institutional variations than morphological CT features. This finding aligns with previous studies demonstrating the superior generalizability of functional imaging biomarkers across different scanner types and acquisition protocols [41-43]. The relative stability of PET-derived features may be attributed to the standardized nature of FDG uptake quantification and the less pronounced impact of reconstruction algorithms on SUV-based texture features.

However, the more pronounced deterioration in recall performance across all modalities (PET: 65.80% to 40.4%; CT: 44.42% to 17%) raises important clinical considerations. This substantial reduction in sensitivity suggests that models may become overly conservative when applied to new patient populations, potentially missing true tumor-positive regions. Such behavior could have significant clinical ramifications in lymphoma staging and treatment planning, where accurate tumor burden assessment is critical for prognosis and therapeutic decision-making.

The performance decline likely reflects multiple factors inherent to multi-institutional validation studies [44, 45]. First, patient population heterogeneity between institutions may contribute to feature distribution shifts, as lymphoma subtypes, disease stages, and patient demographics can vary significantly across clinical centers. Second, subtle differences in imaging acquisition protocols, including contrast timing, reconstruction parameters, and scanner-specific calibrations, may introduce systematic variations in radiomics features that were not adequately captured during model training.

Additionally, the manual segmentation process, despite following standardized protocols, inevitably introduces inter-observer variability that may be amplified across different institutions and clinical workflows. The reduced recall performance particularly suggests that the models learned institution-specific patterns during training that did not generalize effectively to the external validation site.

These findings underscore the importance of rigorous external validation in radiomics research and highlight the need for robust model adaptation strategies before clinical deployment. The results suggest that while radiomics-based approaches show promise for lymphoma tumor classification, direct model transfer without local calibration may result in suboptimal performance, particularly in terms of tumor detection sensitivity.

The external validation results, while showing reduced performance compared to the primary dataset, still demonstrate the fundamental viability of the radiomics approach, particularly for PET-based models. However, they emphasize the critical need for comprehensive validation and potential model refinement before widespread clinical adoption.

4.3. Tumor Score Strategy and Threshold Analysis

To improve classification robustness, we developed a Tumor Score, integrating ensemble ML predictions (weighted at 70%) with anomaly detection (30%) derived from an Isolation Forest algorithm. The ensemble predictions provided stable and consistent classification across modalities, while the anomaly score served to detect subtle, high-risk features in metabolically ambiguous volumes. The 70:30 weighting was empirically chosen based on the superior AUC performance of ensemble predictions.

We evaluated tumor classification performance using thresholds of 0.05, 0.10, and 0.20. At 0.05, the model exhibited 100% sensitivity but lower specificity due to broader inclusion of potentially non-malignant tissue. Raising the threshold to 0.10 and 0.20 improved precision while maintaining high tumor inclusion rates (96%), demonstrating the system’s adaptability to clinical sensitivity-specificity trade-offs.

4.4. Prognostic Prediction Performance and Clinical Data Integration

Our proposed machine learning framework demonstrated variable performance in predicting five-year recurrence and mortality in lymphoma patients, with distinctions depending on multiple factors including the use of SMOTE-based class balancing and the integration of clinical data with imaging features. The comparative analysis between imaging-only and combined imaging-clinical models revealed complex interactions that merit detailed examination.

The integration of clinical variables with radiomics features yielded mixed results that varied significantly across prediction tasks and experimental conditions. For recurrence prediction with SMOTE, the addition of clinical data showed modest improvements in mean accuracy from 66% to 60% across all thresholds (difference: +0.06), although this represents a counterintuitive decrease that warrants careful interpretation. Individual threshold analysis revealed more nuanced patterns: at the 0.05 threshold, clinical data integration improved performance from 62% to 72% (difference: 0.10), while at the 0.20 threshold, a substantial deterioration was observed from 67% to 47% (difference: 0.20).

These findings align with recent observations in radiomics literature that even when combined with clinical data, the results do not necessarily improve [46-48]. The integration of multimodal data sources can introduce feature redundancy, increase model complexity, and potentially dilute the discriminative power of imaging-derived biomarkers, particularly when the clinical variables do not provide complementary information to the radiomics features.

For mortality prediction with SMOTE, clinical data integration demonstrated more consistent benefits, with mean accuracy improving from 79% to 70% (difference: +0.09). The most pronounced improvement was observed at the 0.10 threshold, where accuracy increased from 83% to 57% (difference: +0.26), suggesting that clinical variables may provide complementary prognostic information for mortality risk assessment under specific modeling conditions.

Interestingly, models without SMOTE showed different patterns of clinical data utility. For recurrence prediction, clinical data integration resulted in marginal improvements with a mean difference of -0.05, while mortality prediction remained remarkably stable (mean difference: -0.02), suggesting that the value of clinical data may be influenced by class balancing strategies.

For mortality prediction with SMOTE, clinical data integration demonstrated more consistent benefits, with mean accuracy improving from 79% to 70% (difference: +0.09). The most pronounced improvement was observed at the 0.10 threshold, where accuracy increased from 83% to 57% (difference: +0.26), suggesting that clinical variables may provide complementary prognostic information for mortality risk assessment under specific modeling conditions.

Interestingly, models without SMOTE showed different patterns of clinical data utility. For recurrence prediction, clinical data integration resulted in marginal improvements with a mean difference of -0.05, while mortality prediction remained remarkably stable (mean difference: -0.02), suggesting that the value of clinical data may be influenced by class balancing strategies.

In recurrence prediction, models trained with SMOTE achieved variable performance depending on clinical data inclusion. <mark>Pure imaging models with SMOTE showed mean accuracy of 66%, while the addition of clinical data resulted in 60% accuracy, indicating that feature integration may introduce complexity that requires careful optimization Conversely, in the absence of SMOTE, imaging-only models achieved 59% accuracy compared to 64% with clinical data, suggesting that clinical variables may be more beneficial when dealing with naturally imbalanced datasets.

Importantly, these results demonstrate that TMTV of PET can be utilized as a strong and independent prognostic factor in lymphomas [49-50], even without the incorporation of additional clinical variables. The robust performance of imaging-only models (mean accuracy 66% for recurrence and 79% for mortality with SMOTE) supports the established role of metabolic tumor burden as a powerful predictor of patient outcomes, reinforcing the clinical utility of quantitative PET imaging in lymphoma management.

For mortality prediction, the pattern was more consistent with clinical data showing benefits regardless of SMOTE application. With SMOTE, clinical data integration improved mean accuracy from 79% to 70%, while without SMOTE, the improvement was from 77% to 79%. This suggests that clinical variables may provide more robust prognostic value for mortality prediction compared to recurrence.

The tumor score threshold analysis revealed important insights into model behavior across different clinical decision points. For imaging-only models, recurrence prediction showed relatively stable performance across thresholds (0.62-0.67 with SMOTE), while clinical data integration introduced greater variability (0.47-0.72 with SMOTE). This suggests that clinical data may enhance performance at specific operating points but could reduce overall robustness across different sensitivity-specificity trade-offs.

Mortality prediction demonstrated greater stability with clinical data integration, maintaining consistent performance across thresholds both with and without SMOTE. This differential behavior between recurrence and mortality endpoints may reflect the distinct clinical characteristics of these outcomes, with mortality potentially being more strongly associated with clinical biomarkers than recurrence patterns.

These findings highlight the complex nature of clinical data integration in radiomics-based prognostic modeling. While clinical variables such as LDH, NLR, and PLR are established prognostic factors in lymphoma, their integration with imaging features requires careful consideration of modeling strategies, class balancing techniques, and threshold optimization.

The differential impact of clinical data on recurrence versus mortality prediction suggests that these endpoints may benefit from distinct modeling approaches. Mortality prediction appears more amenable to clinical data integration, possibly reflecting the stronger association between systemic biomarkers and overall survival compared to disease recurrence patterns, which may be more dependent on tumor-specific characteristics captured by imaging features.

4.5. Novel Approach to Mixed Volume Classification and Broader Applications

A key innovation of this study is the explicit modeling of mixed tumor-normal regions, which are often ignored in traditional binary classification approaches. In lymphoma, systemic involvement and physiological FDG uptake in lymphoid tissues or adjacent organs frequently result in metabolically active but non-malignant regions. Our inclusion of these regions in the classification schema reflects real-world clinical challenges and enhances the model’s generalizability.

Additionally, while our methodology was developed using lymphoma as a model, it is not restricted to this disease. The threshold-based segmentation problem occurs across various malignancies—such as non-small cell lung cancer, gynecologic cancers, and head and neck tumors—where tumor margins are often metabolically and anatomically ambiguous [51-59]. As such, our system offers broad applicability to other cancers where mixed-volume classification is critical.

4.6. Validity of PCA selection

In high-dimensional radiomics analysis, dimensionality reduction is essential to mitigate overfitting, reduce noise, and improve model generalizability. Two common strategies include unsupervised methods such as Principal Component Analysis (PCA) and supervised approaches like Least Absolute Shrinkage and Selection Operator (Lasso). While both methods offer dimensionality reduction capabilities, the use of PCA in this study was specifically motivated by several methodological and practical considerations.

First, PCA operates in an unsupervised manner by identifying orthogonal principal components that capture the maximum variance in the feature space. This is particularly advantageous in radiomics, where a substantial proportion of extracted features exhibit high collinearity. By transforming correlated variables into linearly uncorrelated components, PCA mitigates multicollinearity, a common challenge in radiomics-based machine learning. In contrast, Lasso performs feature selection by enforcing sparsity through L1 regularization but may arbitrarily discard correlated but potentially informative features. This can lead to instability in the selected feature set, especially when minor changes in the dataset or noise distribution occur.

Second, PCA is model-agnostic and purely data-driven, making it applicable across multiple downstream classifiers without the need for retraining or parameter tuning for each model. In contrast, Lasso is a supervised method whose performance is tightly coupled to the predictive relationship with the target label. As such, features selected by Lasso may overfit to the training labels and become suboptimal when applied to different algorithms or data distributions.

Third, in the context of our multiclass classification task involving CT, PET, and PET/CT fused features, PCA provided a unified reduction approach that preserved up to 95% of the total variance across modalities. This ensured consistency in the transformed feature space and allowed equitable comparisons between model performances. Lasso, on the other hand, would require repeated retraining for each experimental condition and modality, leading to inconsistent dimensional representations and potentially biased evaluation results.

Moreover, empirical evaluation conducted in Appendix 3 demonstrated that PCA-based dimensionality reduction did not significantly compromise classification performance when compared to models trained on the full feature set. Over 87% of all comparisons showed performance differences within ±0.5%, confirming that PCA successfully preserved the discriminative power of the original features while substantially reducing feature dimensionality and computational complexity
.

In summary, PCA was selected as the primary dimensionality reduction method for this study due to its robustness against multicollinearity, reproducibility across modalities, model-agnostic nature, and stable preservation of variance. While Lasso remains a valuable tool for sparse feature selection in certain predictive contexts, its supervised nature and instability in high-dimensional radiomics data made it less suitable for the objectives of this study.

4.7. Study Limitations

4.7.1 Biological interpretation

Despite the promising results demonstrated in this study, several limitations should be acknowledged that may affect the interpretation and generalizability of our findings. A fundamental limitation inherent to radiomics-based approaches is the challenge of biological interpretation. The data-driven nature of radiomics inherently offers no direct insight into the biological underpinnings of the observed relationships between imaging features and clinical outcomes [60]. While our models demonstrated robust predictive performance, the specific biological mechanisms underlying the most influential features, such as log-sigma-1-mm-3D GLDM Dependence Variance, remain largely unclear.

This interpretability challenge is particularly relevant in lymphoma research, where despite extensive radiomic studies in oncology, it remains unclear which features are truly relevant and what biological processes they represent [61]. The complex relationship between 18F-FDG uptake patterns—reflecting vascularization, cellularity, hypoxia, metabolism, and necrosis—and specific radiomic features complicates the biological validation of our findings.

Furthermore, radiomic analyses often function as a 'black box' due to their use of complex algorithms, which can hinder the translation of research findings into clinical applications [62]. This opacity may limit clinicians' confidence in adopting radiomics-based tools for routine patient care, as the decision-making process remains largely incomprehensible despite demonstrated predictive accuracy.

The mixed results observed with clinical data integration reflect a broader challenge in multimodal radiomics research [61]. Even when combined with established clinical biomarkers, radiomics models do not necessarily demonstrate improved performance, suggesting that feature integration may introduce complexity that requires careful optimization rather than providing straightforward additive benefits.

The variable performance across different experimental conditions and thresholds indicates that the optimal strategy for integrating clinical and imaging data may be highly context-dependent, requiring endpoint-specific approaches rather than universal methodologies.

4.7.2 Statistical Power Limitations

A critical limitation of this study is the insufficient statistical power for prognostic prediction analyses, as determined by formal G*Power calculations. With minimum requirements of 64 patients per group (total n=128) for adequate power, our cohort of 16 recurrence events and 5 mortality events falls substantially below recommended thresholds for robust between-group comparisons. This represents a fundamental constraint that affects the interpretability and generalizability of our prognostic findings.

The tumor classification analysis adequately exceeded power requirements (4,950 vs. 128 minimum required volumes), demonstrating robust statistical foundation for radiomics-based tissue differentiation. However, the prognostic prediction component should be interpreted as a preliminary proof-of-concept analysis rather than a definitive prognostic validation study.

The relatively low event rates observed (26.7% recurrence, 8.3% mortality) reflect improved treatment outcomes in contemporary lymphoma care but limit statistical power for prognostic modeling. While SMOTE implementation provided methodological rigor for handling class imbalance, it cannot address the fundamental issue of insufficient sample size identified through power analysis. Future studies should target sample sizes of at least 128 patients with balanced outcome groups to achieve adequate statistical power for robust prognostic model development.

These power limitations emphasize that our prognostic findings should be considered exploratory and require validation in larger, adequately powered cohorts before clinical implementation.

The prognostic prediction results must be interpreted within the context of insufficient statistical power identified through G*Power analysis. Our sample sizes achieved (16 recurrence, 5 mortality events) were substantially below the minimum requirements (64 per group), limiting the reliability of between-group comparisons and model generalizability.

Despite these constraints, the observed performance metrics provide valuable preliminary insights into the potential utility of radiomics-based prognostic modeling in lymphoma. The superior performance of imaging-only models over combined imaging-clinical approaches may partially reflect the statistical challenges of integrating multiple feature types within underpowered analyses. Similarly, the variable performance across different tumor score thresholds should be interpreted cautiously given the limited statistical power.

These findings establish a methodological framework and provide effect size estimates for future adequately powered studies, while demonstrating the technical feasibility of radiomics-based prognostic prediction in lymphoma patients.

4.8. Clinical Applicability and Expected Impact

Our system offers direct clinical applicability by providing a reproducible, quantitative framework to support and enhance current TMTV-based segmentation methods. The Tumor Score can be integrated into existing imaging workflows to assist clinicians in distinguishing tumor from non-malignant metabolic activity—thereby improving radiotherapy planning, response assessment, and follow-up monitoring.

Moreover, as a non-invasive prognostic tool, our approach can help stratify patients by recurrence or survival risk early in the diagnostic process, potentially guiding treatment escalation or de-escalation. This is particularly valuable in personalized medicine, where objective, reproducible metrics are needed for clinical decision-making.

Table 11 compares the prediction performance of our PET/CT radiomics model with recent radiomics-based studies, illustrating competitive performance in predicting recurrence (AUC=0.7222) and mortality (AUC=0.7934) within five years using SMOTE for balanced sampling. Although the performance of our method is slightly lower than some specialized tasks, such as Yuan et al. (2023), who reported an AUC of 0.926 for predicting cervical lymph node metastasis, it remains comparable or superior to other recent works in oncological prognostic prediction, such as Frood et al. (2022) and Li et al. (2023). These results underscore the robustness and clinical relevance of our model, particularly considering the inherent challenges of predicting long-term outcomes like recurrence and mortality.

Finally, by accounting for the complexities of mixed tissue regions, our framework better reflects real-world conditions and offers a foundation for the next generation of intelligent oncology imaging tools.

Table 11. Comparative summary of radiomics-based predictive performance across recent oncological imaging studies. The table outlines the imaging modalities, predictive tasks, and reported performance metrics for each study.

Author (Year)	**Data sets**	**Task**	**Performance**
Frood et al. (2022)	PET/CT radiomics	Prediction of the response to neoadjuvant chemotherapy	AUC=0.750
Driessen et al. (2023)	PET radiomics	Predicting pathological complete response (pCR) to neoadjuvant chemoradiotherapy (NCRT)	AUC=0.810
Li et al. (2023)	PET radiomics	Predicting lymph node metastasis in non-small cell lung cancer	AUC=0.709
Yuan et al. (2023)	PET/CT radiomics	Predicting cervical lymph node metastasis	AUC=0.926
Eertink et al. (2022)	PET radiomics	Predicting B-cell lymphoma treatment outcome	AUC=0.790
Zhao et al. (2023)	PET radiomics	Prediction of programmed mortality-1 expression status in lung cancer patients	AUC=0.771
Our study	PET/CT radiomics	Prediction of recurrence and mortality within 5 years	Recurrence AUC = 0.7222 Mortality AUC = 0.7934

4.9. Future Research Directions and Methodological Advances

A critical research direction involves developing biological validation frameworks that can bridge the gap between radiomic features and underlying tumor biology through correlation with histopathological and molecular data. This will be essential for enhancing clinical interpretability and addressing the fundamental 'black box' nature of current radiomics models. Future studies will incorporate systematic correlation analysis between top-performing radiomic features and specific biological markers, including tumor microenvironment characteristics, genetic alterations, and metabolic pathway activities relevant to lymphoma progression and treatment response. Additionally, advancing interpretable machine learning methods specifically designed for medical imaging applications will help address the opacity of current radiomics models, thereby facilitating greater clinical acceptance and translation.

The development of standardized protocols for clinical-imaging data integration, including optimal feature selection and fusion strategies, will be crucial for realizing the full potential of multimodal prognostic modeling in lymphoma management. Future work will systematically investigate various data fusion methodologies, including early fusion (feature-level), late fusion (decision-level), and hybrid approaches to determine the most effective strategies for different clinical endpoints. This research direction will also explore advanced machine learning architectures specifically designed for multimodal data integration, such as attention-based neural networks and graph neural networks, which may provide more sophisticated mechanisms for leveraging complementary information from imaging and clinical data sources.

Although the current study identified mixed (tumor-normal) volumes and used them directly for prognostic prediction, future research aims to optimize these mixed volumes by extracting and retaining only tumor-specific regions within each identified volume. This patch-level refinement approach (as illustrated in Figure 5) is expected to improve prognostic accuracy by eliminating normal tissue contamination, thereby facilitating more precise tumor characterization and enhancing clinical decision-making. The implementation of this strategy will involve developing sophisticated segmentation algorithms capable of distinguishing tumor patches from normal tissue patches within mixed volumes, potentially using deep learning-based approaches combined with radiomic feature analysis to achieve sub-volume classification accuracy.

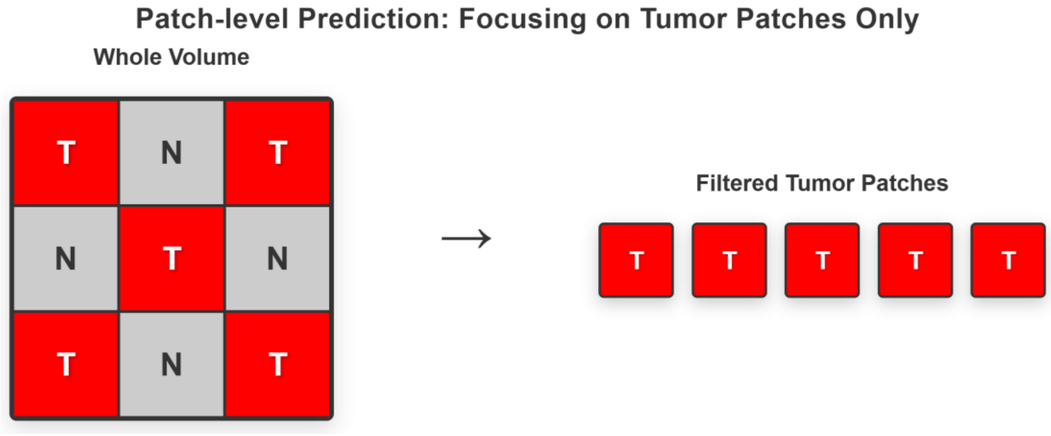


Figure 5. Illustration of patch-level filtering strategy for mixed tumor-normal volumes. The left panel shows a representative whole volume composed of tumor (T, red) and normal (N, gray) patches. This approach is expected to refine tumor burden estimation and enhance prognostic accuracy in future studies.

5. Conclusion

In this study, we developed a radiomics-based machine learning framework capable of classifying tumor, normal, and mixed tumor-normal regions in 18F-FDG PET/CT images of lymphoma patients. By integrating radiomics features from both PET and CT modalities with ensemble predictions and anomaly detection, the proposed system achieved high classification accuracy and demonstrated strong prognostic performance for five-year recurrence and mortality prediction.

The methodology addresses limitations of traditional threshold-based tumor delineation by incorporating ambiguous regions that reflect real-world diagnostic challenges. The Tumor Score system offers a practical, quantitative tool that could be integrated into clinical workflows to enhance tumor segmentation and personalized treatment planning. Future work will include validation using external, multi-institutional datasets and expansion to other cancer types.

References

1. Armitage, J. O., Gascoyne, R. D., Lunning, M. A., & Cavalli, F. (2017). Non-Hodgkin lymphoma. Lancet, 390(10091), 298-310.

2. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209-249.

3. Cheson, B. D., Fisher, R. I., Barrington, S. F., Cavalli, F., Schwartz, L. H., Zucca, E., & Lister, T. A. (2014). Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: The Lugano classification. Journal of Clinical Oncology, 32(27), 3059-3067.

4. Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pileri, S. A., Stein, H., & Thiele, J. (2017). WHO classification of tumours of haematopoietic and lymphoid tissues (revised 4th ed.). Lyon: International Agency for Research on Cancer.

5. Cheson, B. D. (2015). Role of functional imaging in the management of lymphoma. Journal of Clinical Oncology, 33(7), 798-800.

6. Barrington, S. F., & Mikhaeel, N. G. (2014). PET scans for staging and restaging in diffuse large B-cell and follicular lymphomas. Current Hematologic Malignancy Reports, 9(3), 185-195.

7. Mikhaeel, N. G., Smith, D., Dunn, J. T., Phillips, M., Møller, H., Fields, P. A., Wrench, D., Barrington, S. F. (2016). Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. European Journal of Nuclear Medicine and Molecular Imaging, 43(7), 1209-1219.

8. Meignan, M., Cottereau, A. S., & Versari, A. (2017). Baseline metabolic tumor volume in Hodgkin lymphoma: The prognostic value and its clinical application. Blood, 130(23), 2448-2454.

9. Cottereau, A. S., Versari, A., Loft, A., Casasnovas, O., Bellei, M., Ricci, R., Borchmann, P., Meignan, M. (2018). Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. Blood, 131(13), 1456-1463.

10. Kostakoglu, L., Chauvie, S., & Metser, U. (2017). Quantitative PET/CT in lymphoma management. PET Clinics, 12(1), 35-50.

11. Barrington, S. F., & Meignan, M. (2019). Time to prepare for risk adaptation in lymphoma by standardizing measurement of metabolic tumor burden. Journal of Nuclear Medicine, 60(8), 1096-1102.

12. Ceriani, L., Milan, L., Martelli, M., Ferreri, A. J. M., Cascione, L., Zinzani, P. L., Barrington, S. F. (2022). Metabolic tumor volume in lymphoma: Variability between observers and software. Journal of Nuclear Medicine, 63(1), 31-35.

13. Cottereau, A. S., Hapdey, S., Chartier, L., Modzelewski, R., Casasnovas, O., Itti, E., Meignan, M. (2021). Baseline total metabolic tumor volume measured with fixed or adaptive thresholds: Comparison of interobserver agreement. Journal of Nuclear Medicine, 62(8), 1061-1066.

14. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G. P. M., Granton, P., Aerts, H. J. W. L. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer, 48(4), 441-446.

15. Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563-577.

16. Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Gillies, R. J. (2012). Radiomics: The process and the challenges. Magnetic Resonance Imaging, 30(9), 1234-1248.

17. Parvez, A., Tau, N., Hussey, D., Maganti, M., & Metser, U. (2020). F18-FDG PET/CT radiomics predicts survival in patients with aggressive B-cell lymphoma. European Journal of Nuclear Medicine and Molecular Imaging, 47(6), 1455-1464.

18. Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., & Gibbs, P. (2020). Introduction to radiomics. Journal of Nuclear Medicine, 61(4), 488-495.

19. Cottereau, A. S., Buvat, I., & Kanoun, S. (2019). Radiomics in PET/CT imaging of lymphoma: Current status and future directions. Clinical and Translational Imaging, 7, 35-49.

20. Jiang, C., Zhang, Y., Wang, X., Sun, Y., Li, P., Cai, X., & Ge, H. (2021). Radiomics analysis of 18F-FDG PET/CT images for predicting prognosis in lymphoma patients. Frontiers in Oncology, 11, 665839.

21. Chalkidou, A., O'Doherty, M. J., & Marsden, P. K. (2015). False discovery rates in PET and CT studies with texture features: A systematic review. PLoS One, 10(5), e0124165.

22. Park, J. E., Park, S. Y., Kim, H. J., & Kim, H. S. (2020). Reproducibility and generalizability in radiomics modeling: Possible strategies in radiologic and statistical perspectives. Korean Journal of Radiology, 21(10), 1124-1137.

23. Avanzo, M., Stancanello, J., & El Naqa, I. (2017). Beyond imaging: The promise of radiomics. Physica Medica, 38, 122-139.

24. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510.

25. Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. RadioGraphics, 37(2), 505-515.

26. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

27. Driessen, J., Zwezerijnen, G. J. C., Schöder, H., Kersten, M. J., Moskowitz, A. J., Moskowitz, C. H., Eertink, J. J., Heymans, M. W., Boellaard, R., & Zijlstra, J. M. (2023). Prognostic model using 18F-FDG PET radiomics predicts progression-free survival in relapsed/refractory Hodgkin lymphoma. Blood Advances, 7(22), 6732–6743.

28. Li, M., Yao, H., Zhang, P., Zhang, L., Liu, W., Jiang, Z., Li, W., Zhao, S., & Wang, K. (2023). Development and validation of a [18F] FDG PET/CT-based radiomics nomogram to predict the prognostic risk of pretreatment diffuse large B cell lymphoma patients. European Radiology, 33(5), 3354–3365.

29. Zhang, X., Chen, L., Jiang, H., He, X., Feng, L., Ni, M., Ma, M., Wang, J., Zhang, T., & Wu, S. (2022). A novel analytic approach for outcome prediction in diffuse large B-cell lymphoma by [18F]FDG PET/CT. European Journal of Nuclear Medicine and Molecular Imaging, 49(4), 1298–1310.

30. Lue, K.-H., Wu, Y.-F., Lin, H.-H., Hsieh, T.-C., Liu, S.-H., Chan, S.-C., & Chen, Y.-H. (2020). Prognostic value of baseline radiomic features of 18F-FDG PET in patients with diffuse large B-cell lymphoma. Diagnostics, 11(1), 36.

31. Eertink, J. J., van de Brug, T., Wiegers, S. E., Zwezerijnen, G. J. C., Pfaehler, E. A. G., Lugtenburg, P. J., de Vet, H. C. W., Boellaard, R., & Zijlstra, J. M. (2022). 18F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. European Journal of Nuclear Medicine and Molecular Imaging, 49(3), 932–942.

32. Hsu, C. Y., Doubrovin, M., Hua, C. H., Mohammed, O., Shulkin, B. L., Kaste, S., Lucas Jr, J. T. (2018). Radiomics features differentiate between normal and tumoral high-Fdg uptake. Scientific Reports, 8(1), 3913.

33. Zhang, H. W., Huang, D. L., Wang, Y. R., Zhong, H. S., & Pang, H. W. (2024). CT radiomics based on different machine learning models for classifying gross tumor volume and normal liver tissue in hepatocellular carcinoma. Cancer Imaging, 24(1), 20.

34. Zhang, H. W., Wang, Y. R., Hu, B., Song, B., Wen, Z. J., Su, L., Wang, Y. H. (2024). Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases. Scientific Reports, 14(1), 28575.

35. Pei, J., Yu, J., Ge, P., Bao, L., Pang, H., & Zhang, H. (2024). Constructing a Classification Model for Cervical Cancer Tumor Tissue and Normal Tissue Based on CT Radiomics. Technology in Cancer Research & Treatment, 23, 15330338241298554.

36. Wang, X., Jemaa, S., Fredrickson, J., Coimbra, A. F., Nielsen, T., De Crespigny, A., Carano, R. A. (2022). Heart and bladder detection and segmentation on FDG PET/CT by deep learning. BMC Medical Imaging, 22(1), 58.

37. Major, A., Hammes, A., Schmidt, M. Q., Morgan, R., Abbott, D., & Kamdar, M. (2020). Evaluating novel PET-CT functional parameters TLG and TMTV in differentiating low-grade versus grade 3A follicular lymphoma. Clinical Lymphoma Myeloma and Leukemia, 20(1), 39-46.

38. Chen, Y., Pasquier, D., Verstappen, D., Woodruff, H. C., & Lambin, P. (2025). An interpretable ensemble model combining handcrafted radiomics and deep learning for predicting the overall survival of hepatocellular carcinoma patients after stereotactic body radiation therapy. Journal of Cancer Research and Clinical Oncology, 151(2), 1–10.

39. Wu, Y., Xu, D., Zha, Z., Gu, L., Chen, J., Fang, J., & Wang, J. (2025). Integrating radiomics into predictive models for low nuclear grade DCIS using machine learning. Scientific Reports, 15(1), 7505.

40. Zhang, H. W., Wang, Y. R., Hu, B., Song, B., Wen, Z. J., Su, L., & Wang, Y. H. (2024). Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases. Scientific Reports, 14(1), 28575.

41. Antunes, J., Viswanath, S., Rusu, M., Valls, L., Hoimes, C., Avril, N., & Madabhushi, A. (2016). Radiomics analysis on FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study. Translational Oncology, 9(2), 155-162.

42. Mali, S. A., Ibrahim, A., Woodruff, H. C., Andrearczyk, V., Müller, H., Primakov, S., Lambin, P. (2021). Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. Journal of Personalized Medicine, 11(9), 842.

43. Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging, 11(1), 91.

44. Soliman, M. A., Kelahan, L. C., Magnetta, M., Savas, H., Agrawal, R., Avery, R. J., Velichko, Y. S. (2022). A framework for harmonization of radiomics data for multicenter studies and clinical trials. JCO Clinical Cancer Informatics, 6, e2200023.

45. Lo Gullo, R., Daimiel, I., Morris, E. A., & Pinker, K. (2020). Combining molecular and imaging metrics in cancer: radiogenomics. Insights into Imaging, 11, 1-17.

46. Destito, M., Marzullo, A., Leone, R., Zaffino, P., Steffanoni, S., Erbella, F., Spadea, M. F. (2023). Radiomics-based machine learning model for predicting overall and progression-free survival in rare cancer: a case study for primary CNS lymphoma patients. Bioengineering, 10(3), 285.

47. Kocak, B., dos Santos, D. P., & Dietzel, M. (2025). The widening gap between radiomics research and clinical translation: rethinking current practices and shared responsibilities. European Journal of Radiology Artificial Intelligence, 100004.

48. Crombé, A., Fadli, D., Italiano, A., Saut, O., Buy, X., & Kind, M. (2020). Systematic review of sarcomas radiomics studies: Bridging the gap between concepts and clinical applications? European Journal of Radiology, 132, 109283.

49. Kanoun, S., Rossi, C., Berriolo-Riedinger, A., Dygai-Cochet, I., Cochet, A., Humbert, O., Casasnovas, R. O. (2014). Baseline metabolic tumour volume is an independent prognostic factor in Hodgkin lymphoma. European Journal of Nuclear Medicine and Molecular Imaging, 41, 1735-1743.

50. Shagera, Q. A., Cheon, G. J., Koh, Y., Yoo, M. Y., Kang, K. W., Lee, D. S., Chung, J. K. (2019). Prognostic value of metabolic tumour volume on baseline 18 F-FDG PET/CT in addition to NCCN-IPI in patients with diffuse large B-cell lymphoma: further stratification of the group with a high-risk NCCN-IPI. European Journal of Nuclear Medicine and Molecular Imaging, 46, 1417-1427.

51. Zhang, Y., Huang, W., Jiao, H., & Kang, L. (2024). PET radiomics in lung cancer: advances and translational challenges. EJNMMI Physics, 11(1), 81.

52. Primakov, S. P., Ibrahim, A., van Timmeren, J. E., Wu, G., Keek, S. A., Beuque, M., Lambin, P. (2022). Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nature Communications, 13(1), 3423.

53. Pawaroo, D., Cummings, N. M., Musonda, P., Rintoul, R. C., Rassl, D., & Beadsmoore, C. (2011). Non–small cell lung carcinoma: accuracy of PET/CT in determining the size of T1 and T2 primary tumors. American Journal of Roentgenology, 196(5), 1176-1181.

54. Cegła, P., Burchardt, E., Wierzchosławska, E., Roszak, A., & Cholewiński, W. (2019). The effect of different segmentation methods on primary tumour metabolic volume assessed in 18F-FDG-PET/CT in patients with cervical cancer, for radiotherapy planning. Contemporary Oncology/Współczesna Onkologia, 23(3), 183-186.

55. Mu, W., Chen, Z., Shen, W., Yang, F., Liang, Y., Dai, R., Tian, J. (2015). A segmentation algorithm for quantitative analysis of heterogeneous tumors of the cervix with 18 F-FDG PET/CT. IEEE Transactions on Biomedical Engineering, 62(10), 2465-2479.

56. Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Depeursinge, A. (2022). Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. Medical Image Analysis, 77, 102336.

57. Han, D., Bayouth, J., Song, Q., Taurani, A., Sonka, M., Buatti, J., & Wu, X. (2011, July). Globally optimal tumor segmentation in PET-CT images: a graph-based co-segmentation method. In Biennial International Conference on Information Processing in Medical Imaging (pp. 245-256). Berlin, Heidelberg: Springer Berlin Heidelberg.

58. Im, H. J., Bradshaw, T., Solaiyappan, M., & Cho, S. Y. (2018). Current methods to define metabolic tumor volume in positron emission tomography: which one is better? Nuclear Medicine and Molecular Imaging, 52, 5-15.

59. Tamal, M. (2020). Intensity threshold based solid tumour segmentation method for Positron Emission Tomography (PET) images: a review. Heliyon, 6(10).

60. Tomaszewski, M. R., & Gillies, R. J. (2021). The biological meaning of radiomic features. Radiology, 298(3), 505-516.

61. Chauvie, S., Ceriani, L., & Zucca, E. (2021). Radiomics in malignant lymphomas. In Lymphoma (pp. 71-82). Brisbane: Exon Publications.

62. Wang, Y., Hu, Z., & Wang, H. (2025). The clinical implications and interpretability of computational medical imaging (radiomics) in brain tumors. Insights into Imaging, 16(1)

Appendix 1: Evaluation of the Effect of Physiological Organ Exclusion on TMTV

This appendix presents a quantitative analysis of how the exclusion of physiologically high-uptake normal organs affects the accuracy of tumor segmentation in threshold-based methods. Five thresholding strategies were tested—liver-based threshold, SUV > 2.5, SUV > 3.0, 41% of SUV_{max}, and 50% of SUV_{max}—under two conditions: (1) all FDG-avid organs included, and (2) brain, heart, kidneys, bladder, and ureters excluded.

For each method, the percentage difference in segmented volume was calculated between the threshold-based segmentation result and the ground truth tumor label. As shown in Figure 6, organ exclusion significantly reduced the deviation from the ground truth, particularly in the liver-based, 3.0, and 50% SUV_{max} methods.

To determine statistical significance, F-tests and t-tests were performed comparing the segmentation results before and after organ exclusion. The results are summarized in Table 12. When all physiological organs were included, all methods exhibited significant differences in both variance and mean values (Welch's t-test, $p < 0.01$). After excluding high-uptake organs, no statistically significant differences were observed for the liver-based, SUV 3.0, and 50% SUV_{max} methods (Student's t-test, $p > 0.05$), while the 2.5 and 41% SUV_{max} methods continued to show significant discrepancies ($p < 0.01$).

These findings emphasize the importance of excluding physiological FDG uptake regions to enhance segmentation accuracy and avoid overestimation in TMTV calculation.

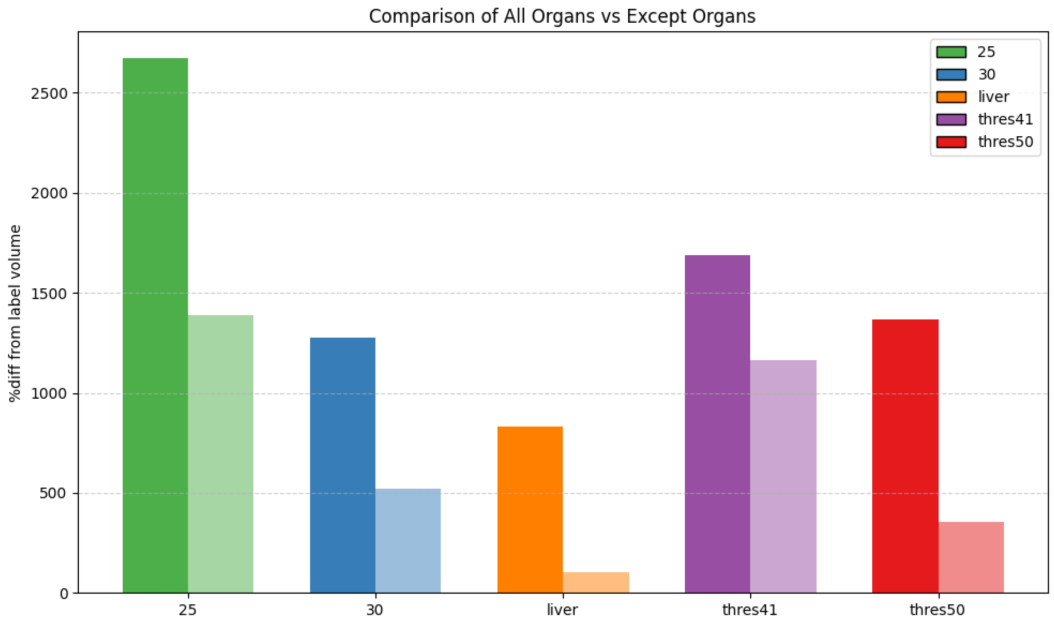


Figure 6. Percentage volume differences between ground truth tumor labels and threshold-based segmentation results, with and without the exclusion of physiologically high-uptake normal organs.

Table 12. Statistical comparison of segmentation volume differences using five threshold methods, before and after the exclusion of physiologically high-uptake normal organs. Welch's t-test was used when variance was unequal; otherwise, Student's t-test was applied.

Threshold Criteria	F-test p-value	T-test Type	T-test p-value
25	2E-11	Heteroscedastic t-test	6E-11
30	2E-05	Heteroscedastic t-test	2E-05
liver	5E-03	Heteroscedastic t-test	6E-03
thres41	7E-06	Heteroscedastic t-test	7E-06
thres50	3E-04	Heteroscedastic t-test	3E-04
25 (excluding organs)	2E-05	Heteroscedastic t-test	3E-05
30 (excluding organs)	6E-02	Homoscedastic t-test	6E-02
liver (excluding organs)	3E-01	Homoscedastic t-test	3E-01
thres41 (excluding organs)	3E-03	Heteroscedastic t-test	3E-03
thres50 (excluding organs)	2E-01	Homoscedastic t-test	2E-01

Appendix 2: DICOM-Based SUV Mapping

To transfer CT-based segmentation results to PET images for SUV quantification, precise multimodal image alignment was performed. The alignment process utilized MATLAB's image registration module with optimized parameters to ensure accurate spatial correspondence between CT and PET datasets. The optimizer was configured with an initial radius of 0.009, epsilon value of 1.5E-4, and a maximum of 1000 iterations to achieve optimal registration performance.

The alignment accuracy was validated using the Dice coefficient to compare contour coordinates obtained from CT-based segmentation with corresponding PET image structures. This validation process ensured that the spatial transformation accurately preserved anatomical boundaries across modalities, which is critical for reliable SUV measurements in specific organs and tumor regions.

Standardized uptake values (SUVs) were calculated using patient-specific parameters extracted from DICOM header information. The SUV calculation incorporated essential radiopharmaceutical and patient parameters to ensure accurate quantification of metabolic activity per voxel. The body weight-based SUV formula was implemented as follows:

$$SUV_{body\ weight\left(\frac{kg}{cc}\right)} = \frac{(pixel\ value \times Dicom\ rescale\ factor \times Patient\ weight)}{Total\ dose \times e^{\left(\frac{-log\ (2) \times (Series\ time\ -\ Radiophamaceutical\ start\ time)}{F^{18} - FDG\ half\ life\ time}\right)}} \quad \text{(A2)}$$

where $pixel\ value$ represents the raw intensity value from the PET image, $Dicom\ rescale\ factor$ is the normalization factor for pixel array values, $Patient\ weight$ is the body weight in kilograms, $Total\ dose$ is the administered F-18 FDG activity, $Series\ time$ is the PET scan acquisition start time, $Radiophamaceutical\ start\ time$ is the F-18 FDG injection time, and $F^{18} - FDG\ half\ life\ time$ is the physical half-life of F-18 (109.8 minutes).

Appendix 3: Validation of PCA-Based Dimensionality Reduction in Radiomics Classification Models

To validate the effectiveness of our dimensionality reduction approach, we conducted systematic comparisons between models trained with PCA-reduced features and models using the full feature set. This analysis was performed across all three radiomics datasets (CT, PET, and combined PET/CT) using all five machine learning algorithms. The comprehensive results of this validation study are presented in Table 13, which demonstrates the performance metrics for both PCA-reduced and full-feature approaches across all experimental conditions.

The comparative analysis revealed remarkably consistent performance between PCA-reduced and full-feature models across all experimental conditions. For accuracy metrics, the CT radiomics dataset showed a mean difference of -0.17% between PCA and non-PCA approaches, with individual model variations ranging from -0.58% to +0.15%. The PET radiomics dataset demonstrated even greater consistency with a mean difference of +0.02%, ranging from -0.21% to +0.30%. Most notably, the combined PET/CT radiomics dataset showed perfect mean consistency with a 0.00% difference, though individual models ranged from -0.33% to +0.48%.

Similarly, AUC analysis confirmed the robustness of the PCA approach across all datasets. CT radiomics showed minimal mean AUC difference of -0.01% with a range from -0.63% to +0.72%, while PET radiomics demonstrated a slight improvement of +0.10% with variations from -0.06% to +0.28%. The combined PET/CT dataset showed a minimal mean difference of -0.08%, with individual model differences ranging from -0.86% to +0.51%. Across all experimental conditions, the maximum absolute difference observed was 0.86% for AdaBoost AUC performance in the combined dataset, representing the most extreme variation encountered in the entire validation study.

Statistical analysis of the comparative results revealed that 87% of all performance comparisons showed differences within ±0.5%, demonstrating exceptional consistency between the two approaches. This high level of agreement validates that PCA successfully preserved the discriminative information content of the original radiomics features while simultaneously reducing computational complexity and mitigating potential overfitting risks inherent to high-dimensional radiomics data. The minimal performance variations observed across different machine learning algorithms and radiomics datasets provide strong evidence supporting the robustness and effectiveness of our dimensionality reduction strategy, confirming that the choice of PCA as our primary feature processing approach was methodologically sound and did not compromise the predictive capabilities of our classification framework.

Table 13. Performance comparison of radiomics-based tumor classification models with and without PCA dimensionality reduction.

Dataset	**Model**	**Accuracy**		**% Diff**	**AUC**		**% Diff**
		w/PCA	w/o PCA		w/PCA	w/o PCA	
CT radiomics	Random Forest	0.9392	0.9403	-0.12	0.9653	0.9661	-0.08

Dataset	Model	Accuracy		% Diff	AUC		% Diff
	Decision Tree	0.9231	0.9217	0.15	0.8536	0.8590	-0.63
	Gradient Boosting	0.9435	0.9431	0.04	0.9562	0.9554	0.08
	AdaBoost	0.9356	0.9410	-0.58	0.9400	0.9422	-0.23
	XGBoost	0.9478	0.9510	-0.34	0.9690	0.9620	0.72
	Average	0.9378	0.9394	-0.17	0.9368	0.9369	-0.01
PET radiomics	Random Forest	0.9293	0.9286	0.08	0.9607	0.9613	-0.06
	Decision Tree	0.8925	0.8898	0.30	0.7791	0.7784	0.09
	Gradient Boosting	0.9313	0.9333	-0.21	0.9558	0.9531	0.28
	AdaBoost	0.9116	0.9122	-0.07	0.9484	0.9482	0.02
	XGBoost	0.9347	0.9347	0.00	0.9605	0.9591	0.15
	Average	0.9199	0.9197	0.02	0.9209	0.9200	0.10
PET/CT radiomics	Random Forest	0.9372	0.9365	0.07	0.9627	0.9643	-0.17
	Decision Tree	0.9179	0.9182	-0.03	0.8479	0.8436	0.51
	Gradient Boosting	0.9388	0.9343	0.48	0.9522	0.9512	0.11
	AdaBoost	0.9266	0.9287	-0.23	0.9315	0.9395	-0.86
	XGBoost	0.9423	0.9454	-0.33	0.9639	0.9631	0.08
	Average	0.9326	0.9326	0.00	0.9316	0.9323	-0.08

Abstract in Korean

림프종에서의 다중 모달리티 기반 종양 분류 및 예후 예측을 위한 라디오믹스 기반 머신러닝 연구

목적: 본 연구는 18F-FDG PET/CT 영상을 이용하여 림프종 환자에서 종양, 정상, 그리고 종양-정상 혼재 영역을 구별할 수 있는 방사선학 기반 머신러닝 프레임워크를 개발하고, 재발과 사망률을 포함한 예후 예측에서의 효과를 평가하는 것을 목표로 하였다.

재료 및 방법: 60명의 림프종 환자의 PET/CT 스캔을 분석하여, 수동으로 윤곽을 그은 종양(n=800)과 정상(n=4,150) 볼륨에서 방사선학적 특징(각 모달리티당 417개)을 추출하였다. 5가지 머신러닝 알고리즘(AdaBoost, Decision Tree, Gradient Boosting, Random Forest, XGBoost)을 PET 단독, CT 단독, PET/CT 결합 방사선학적 특징, 그리고 PET 매개변수 (SUV)를 사용하여 훈련시켰다. 앙상블 예측과 Isolation Forest를 통한 이상 탐지를 통합한 종양 점수 시스템을 구축하였다. 5년간의 재발과 사망률에 대한 예후 예측은 PET 지표, 임상 변수, 그리고 SMOTE 기반 클래스 균형화를 사용하였다. 외부 검증은 16명의 추가 환자를 포함하였다.

결과: 통합 PET/CT 방사선학 모델이 우수한 종양 분류 성능(AUC: 0.9639)을 달성한 반면, PET 단독 모델은 최적의 민감도(재현율: 65.80%)를 보여주었다. 예후 예측에서 SMOTE 구현은 5년 재발 예측 정확도를 최대 75%, 사망률 예측을 최대 86%까지 향상시켰다. 그러나 임상 데이터 통합은 일관되지 않은 결과를 보였으며, 재발 예측 정확도가 47%에서 92% 범위였다. 외부 검증에서 모델의 일반화 가능성이 확인되었으며, PET 기반 특징이 가장 우수한 성능(정확도: 90.34%, AUC: 0.8852)을 보였다. 민감도는 외부 검증에서 65.80%에서 40.4%로 감소하여, 기관 간 변이성과 기관별 보정의 필요성을 시사하였다.

결론: 개발된 방사선학 기반 머신러닝 프레임워크는 림프종 환자에서 종양, 정상, 그리고 혼재 볼륨을 효과적으로 구별하며, 예후 예측 향상에 대한 강한 잠재력을 보여주었다. 그러나 외부 검증에서의 민감도 감소는 광범위한 임상 적용 이전에 추가적인 개선과 기관별 보정이 필요함을 강조한다.

핵심되는 말: **림프종, 라디오믹스, 머신러닝, 종양점수, PET-CT**