BMC Cancer

RESEARCH Open Access

Machine learning-based differentiation of benign and malignant adrenal lesions using 18F-FDG PET/CT: a two-stage classification and SHAP interpretation study

Yun Wang[1†], Yuqi Su[1,2†], Jing Li[3], Deying Xie[3], Zhuolin Liu[5], Yuhuang Cai[1,2], Chengyang Sun[1,2], Jingjing Zhang[1,2], Jaesik Jeong[5*], Heqing Yi[1,6*] and Ye Yuan[4,5*]

Abstract

Background Accurately distinguishing benign from malignant adrenal lesions remains a clinical challenge, especially in oncology patients with indeterminate imaging findings. This study aimed to develop and interpret machine learning (ML) models for classifying adrenal lesions based on 18 F-FDG PET/CT imaging and clinical parameters.

Methods A retrospective cohort of 255 patients undergoing 18 F-FDG PET/CT was analyzed. Imaging features—including adrenal SUVmax, SUVpeak, tumor diameter, CT attenuation, and tumor-to-liver SUVmax ratio (T/L SUVmax)—along with clinical variables were extracted. Two classification tasks were constructed: (1) differentiation of benign and malignant adrenal lesions; and (2) subtyping of malignant lesions into lung cancer metastases or lymphoma. Seven ML models were trained and evaluated using 10-fold cross-validation. SHAP (SHapley Additive exPlanations) analysis was applied to elucidate feature contributions.

Results For the benign/malignant classification, ensemble models (Random Forest, Bagging, XGBoost) achieved outstanding performance (AUC > 0.99), with Bagging yielding 100% recall. T/L SUVmax, adrenal SUVmax, and CT attenuation emerged as top predictors. In malignancy subtyping, the artificial neural network (ANN) attained the highest AUC (0.887) and F1-score (0.851). SHAP analysis highlighted distinct metabolic patterns, with lymphoma showing higher SUVmax and T/L ratios, and lung metastases associated with higher CT values.

Conclusion Machine learning models based on PET/CT-derived features enable highly accurate and interpretable classification of adrenal lesions. Integrating metabolic and anatomical parameters improves diagnostic precision,

†Yun Wang and Yuqi Su contributed equally to this work and share the first authorship.

*Correspondence:
Jaesik Jeong
jjs3098@gmail.com
Heqing Yi
yihq@zjcc.org.cn
Ye Yuan
yuanye017@126.com

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

while SHAP analysis offers clinical transparency, supporting personalized decision-making in adrenal lesion management.

Keywords Adrenal lesion, 18F-FDG PET/CT, Machine learning, SHAP, Lung metastasis, Lymphoma, Diagnostic imaging

Introduction

Adrenal lesions are frequently encountered in oncology imaging, particularly in patients with known malignancies. Approximately 50% of adrenal masses identified in individuals with extra-adrenal primary malignancies represent metastases, while benign adrenal lesions may coexist in patients with malignancy, making accurate characterization essential for clinical management [1, 2]. Differentiating benign from malignant adrenal lesions is critical for disease staging, prognosis evaluation, and guiding treatment strategies.

Conventional imaging modalities such as abdominal ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) offer valuable anatomical information but have inherent limitations in functional assessment and lesion characterization. In particular, CT and MRI often struggle to distinguish lipid-poor adenomas from metastases or primary adrenal malignancies, leading to diagnostic uncertainty in a significant proportion of cases [3, 4].

Positron emission tomography/computed tomography (PET/CT) using 18 F-fluorodeoxyglucose (18 F-FDG) provides complementary metabolic information and has shown promise in adrenal lesion evaluation by visualizing glucose metabolism, which is typically upregulated in malignant cells due to the Warburg effect [5, 6]. PET/CT can detect hypermetabolic foci that are not morphologically conspicuous on conventional imaging, thereby improving sensitivity and enabling earlier detection of metastatic disease.

However, visual interpretation of 18 F-FDG PET/CT remains subjective and may vary among observers, particularly regarding the metabolic activity cut-off value used to define malignancy. Previous studies have proposed that the tumor-to-liver SUVmax ratio (T/L SUVmax) may offer superior diagnostic accuracy compared to absolute SUVmax or SUVmean values alone [7, 8]. Nevertheless, the optimal threshold for T/L SUVmax remains uncertain due to limited large-sample validation, and its diagnostic performance in combination with tumor size and unenhanced CT attenuation has not been thoroughly investigated.

Given these gaps, this study aimed to evaluate the diagnostic utility of metabolic parameters—particularly T/L SUVmax and SUVmax—combined with CT-derived morphological features such as tumor diameter and CT value, in distinguishing benign from malignant adrenal lesions using machine learning approaches. Recent systematic reviews have highlighted both the promise and current limitations of artificial intelligence and radiomics in adrenal imaging, emphasizing the need for interpretable and clinically deployable models [9]. By leveraging multi-parametric PET/CT features and advanced analytical techniques, we sought to improve diagnostic precision and address the limitations of traditional visual interpretation.

Methods

Study population and data collection

This retrospective study enrolled 255 patients with adrenal lesions of undetermined origin who underwent PET/CT examinations at Zhejiang Cancer Hospital between December 2016 and December 2023. Inclusion criteria were: (1) histologically or clinically confirmed diagnosis of either benign or malignant adrenal tumors; (2) availability of high-quality PET/CT images and complete clinical records; (3) all patients were treatment-naïve and without secondary primary malignancies; and (4) the cohort included both patients with isolated adrenal lesions and those with multi-organ involvement.

Importantly, adrenal lesions with unenhanced CT attenuation ≤ 10 HU were not automatically included as benign cases. Such lesions were only included if PET/CT was clinically indicated due to oncologic context, ambiguous imaging features, or uncertainty in prior imaging. Lesions with unequivocal benign characteristics that were confidently diagnosed without PET/CT were excluded from the dataset. This approach ensured that all included cases represented diagnostically relevant scenarios in which PET/CT was clinically justified.

Patients with incomplete data or poor imaging quality were excluded. The cohort included 168 males and 87 females, with an average age of 63.0 ± 10.6 years (range, 21–89 years). There were 308 adrenal lesions, including 97 benign lesions and 211 malignant lesions (unilateral in 202 cases and bilateral in 53 cases). Among the 159 patients with malignant adrenal tumors, 32 cases were pathologically confirmed as malignant, while 127 cases were diagnosed through clinical follow-up. Of the 96 patients with benign adrenal tumors, 6 cases received histopathological confirmation of benignancy, with the remaining 91 cases being confirmed by follow-up evaluation. Among the 32 patients with histopathologically confirmed malignant adrenal tumors, diagnosis was established via adrenalectomy in 2 cases (6.3%) and adrenal biopsy in 30 cases (93.7%). Of the 6 patients with

histopathologically confirmed benign adrenal tumors, 3 cases (50%) were diagnosed by adrenalectomy and 3 cases (50%) by adrenal biopsy. Malignant Adrenal Tumors ($n = 159$), Metastases ($n = 106$): All confirmed by histopathology of the primary tumor: Lung cancer (94), Renal cell carcinoma (3), Malignant melanoma (2), Others (1 each): Neuroendocrine tumor, salivary duct carcinoma, uterine malignant mixed Müllerian tumor, rectal cancer, prostate cancer, ovarian cancer, hepatocellular carcinoma. Primary adrenal malignancies ($n = 5$): Adrenocortical carcinoma (3), Malignant pheochromocytoma (2), Adrenal lymphomas ($n = 48$): 14 cases confirmed by adrenal biopsy 34 cases diagnosed via imaging criteria + clinical follow-up, all with extra-adrenal biopsy-confirmed malignancy (sites: lymph nodes [16], left lower gingiva [1], left nasal cavity [1], tonsils [2], left lung [1], chest wall [1], stomach [3], liver [2], pancreas [1], kidney [1], abdominal wall [1], right lumbar region [1], sigmoid colon [1], testis [1], left calf [1]). Benign Adrenal Tumors ($n = 96$), 6 cases confirmed by histopathology, 91 cases diagnosed via imaging criteria + clinical follow-up (Fig. 1). The lymphoma prevalence (30% of malignant lesions) reflects our institution's unique position as a National Cancer Regional Medical Center specializing in hematologic malignancies. This study was approved by the Medical Ethics Committee of Zhejiang Cancer Hospital (IRB-2023-1092), and the requirement of informed patient consent was waived.

PET/CT imaging protocol and feature extraction

Before tracer injection, all patients were instructed to fast for 4–6 h to ensure venous blood glucose (VBG) levels remained below 200 mg/dL. The Discovery 710 PET/CT system (GE Healthcare, Milwaukee, WI, USA) was used to scan all 255 patients. FDG PET/CT image acquisition was performed approximately 1 h after intravenous administration of 18 F-FDG at a dose of approximately 3.7 MBq/kg. Scanning extended from the parietal skull to the upper femur, with each bed position scanned for 2–3 min. Each patient required 7–8 bed positions, resulting in an overall examination time of approximately 20 min. CT data were used for attenuation correction, and image reconstruction was achieved using the ordered-subsets expectation maximization (OSEM) iterative method. Whole-body PET, CT, and PET/CT fusion images were subsequently obtained.

Quantitative imaging features were manually extracted by two experienced nuclear medicine physicians blinded to clinical outcomes. The following variables were recorded: maximum standardized uptake value of the adrenal lesion (SUVmax), peak standardized uptake value (SUVpeak), max liver SUV (liver SUVmax), mean liver SUV (liver SUVmean), and the tumor-to-liver SUVmax ratio (T/L SUVmax). Tumor size (diameter in mm) and unenhanced CT attenuation (in Hounsfield Units) were measured using axial fusion images and region-of-interest (ROI) analysis.

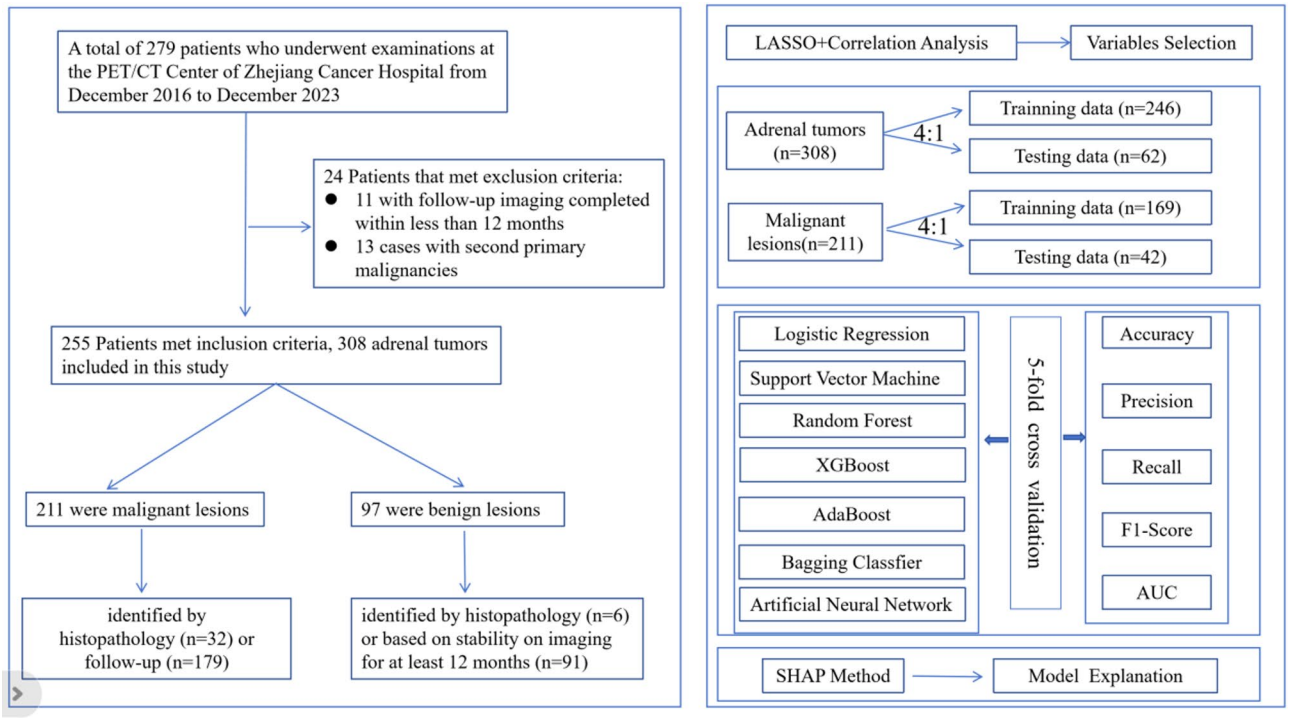

Fig. 1 The research flow chart

Machine learning workflow

Two classification tasks were designed: (1) binary classification to differentiate benign from malignant adrenal lesions; and (2) multiclass classification to distinguish between lung cancer metastases and lymphoma among malignant cases. Prior to modeling, continuous variables were standardized, and missing data (if any) were imputed using median values.

Feature selection was conducted using the least absolute shrinkage and selection operator (LASSO) to reduce dimensionality and identify the most relevant predictors. Seven machine learning models were evaluated: logistic regression, support vector machine (SVM), random forest, XGBoost, AdaBoost, bagging classifier, and artificial neural network (ANN). Model performance was assessed using 5-fold cross-validation.

Model interpretation tools

To enhance model interpretability, SHapley Additive exPlanations (SHAP) analysis was used to calculate the marginal contribution of each feature to individual predictions. Both global feature importance (mean absolute SHAP value) and local interactions (SHAP dependence plots) were visualized for the top-ranked predictors in both classification tasks. These insights were used to explore non-linear effects and to support clinical interpretability of model decisions.

Statistical analysis

Statistical analyses were performed via Python version 3.8 and R version 4.0. Descriptive statistics were used to summarize clinical and imaging variables, and continuous variables were compared using the Mann–Whitney U test or Student's t-test, as appropriate. Categorical variables were compared using the chi-square test or Fisher's exact test. Feature selection was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression to identify the most predictive variables while reducing model overfitting. For supervised classification tasks, multiple machine learning models—including logistic regression, support vector machine (SVM), random forest, XGBoost, AdaBoost, bagging classifier, and artificial neural network (ANN)—were trained and evaluated. Model performance was assessed through 10-fold cross-validation and quantified using area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1-score. SHAP (SHapley Additive exPlanations) analysis was applied to interpret feature contributions and assess both global importance and interaction effects. All statistical tests were two-sided, and a p-value < 0.05 was considered statistically significant. The research flow chart was showed in Fig. 1.

Result

A total of 255 patients were available in this study, the age ranges from 21 to 89 years old. The total number of male patients is 168, accounting for 65.9% of the total number of patients. The total number of female patients was 87, accounting for 34.1% of the total number of patients (Table 1).

Table 1 Baseline characteristics of the study population

Groups	*n*	age	Sex(male/female)	lesion laterality(unilateral/bilateral)	Tumour diameter(cm)	histology
Malignant	159	63.2 ± 11.4	117/42	107/52	3.7 ± 2.4	adrenal metastatic tumor 106 adrenal lymphoma 48 adrenocortical carcinoma 3 adrenal malignant pheochromocytoma 2
Benign	96	63.8 ± 9.1	51/45	95/1	2.0 ± 1.2	adrenal adenoma 67 adrenal myelolipoma 24 adrenal tuberculosis 3 adrenal pheochromocytoma 1 adrenal cyst 1
t	-	0.397	-	-	−6.665	-
P	-	0.692	0.001	0.000	0.000	-

Feature space visualization and classification feasibility for adrenal lesion diagnosis

To investigate the intrinsic structure of the dataset and assess the feasibility of distinguishing benign from malignant adrenal lesions, this study first performed a t-distributed Stochastic Neighbor Embedding (t-SNE) projection on the selected features. As shown in Fig. 2, the t-SNE visualization reveals two well-separated clusters, with each cluster predominantly corresponding to one of the binary classification labels (yellow = benign, purple = malignant).

The projection was based on a subset of clinically relevant features, including adrenal gland lesions SUVmax, tumor plain scan CT value, T/L SUVmax ratio, tumor diameter, liver SUVmean, age, and gender. This clear spatial separation in the two-dimensional embedded space suggests that these input variables contain informative patterns that are strongly associated with lesion type. The distinct clustering indicates low intra-class variance and high inter-class separability, supporting the hypothesis that the malignancy status of adrenal lesions can be effectively inferred from the extracted features.

Elastic net feature coefficient analysis

Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression was applied to identify a parsimonious subset of features most predictive of adrenal lesion malignancy. This method performs both regularization and variable selection by penalizing the absolute size of coefficients, thus shrinking less informative predictors toward zero.

As shown in Fig. 3-A, the coefficient paths demonstrate how variable weights evolve as the regularization parameter (lambda) increases. With stronger penalization (moving rightward along the x-axis), coefficients progressively approach zero. Figure 3-B shows the result of 10-fold cross-validation, where the optimal lambda was determined based on the minimum binomial deviance. This value balances model complexity and predictive performance.

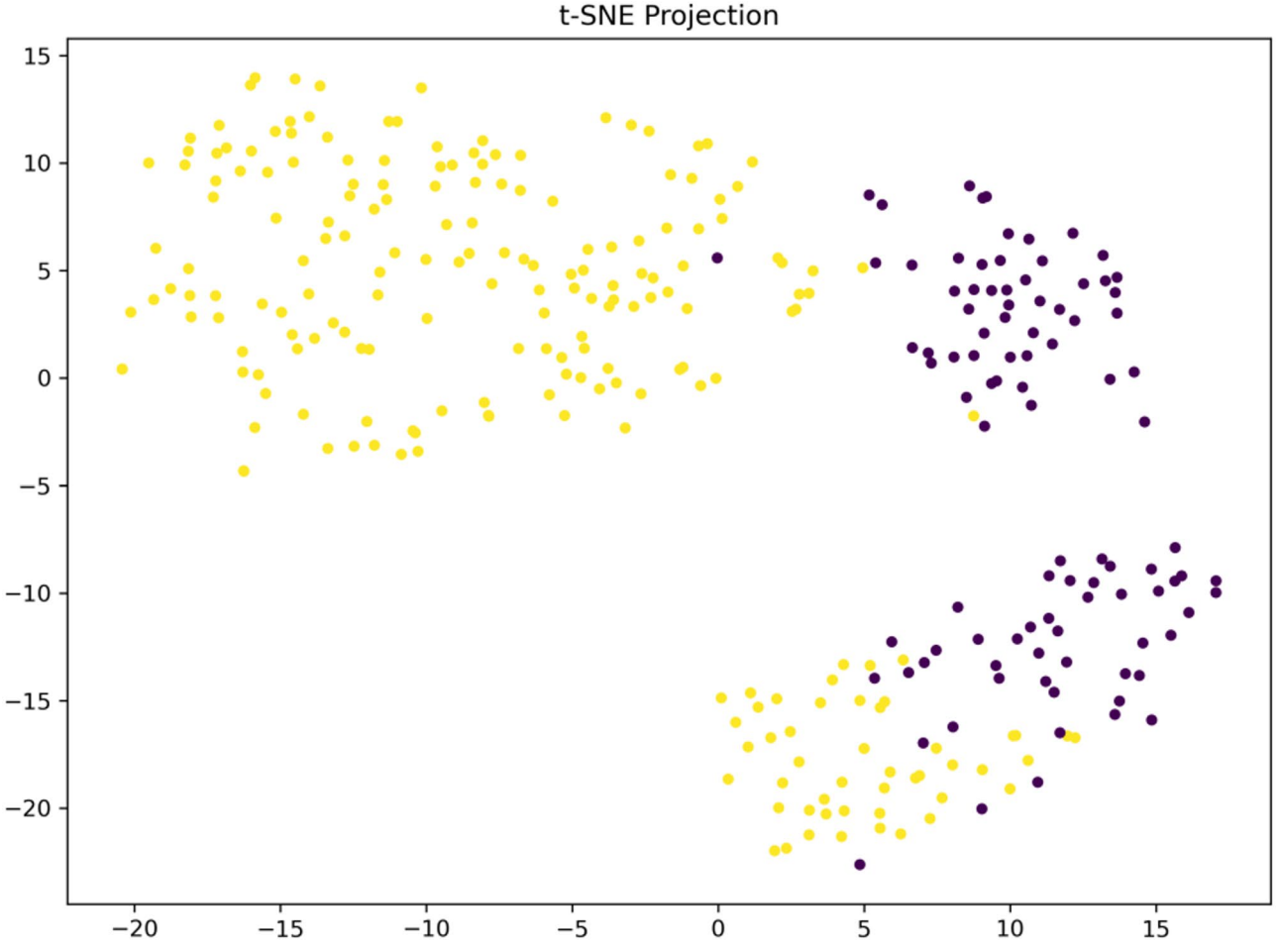

Fig. 2 t-SNE projection of the dataset based on selected features. Each point represents an individual sample, with colors indicating the binary classification labels (yellow = negative, purple = positive). The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm was used to reduce the high-dimensional feature space to two dimensions for visualization

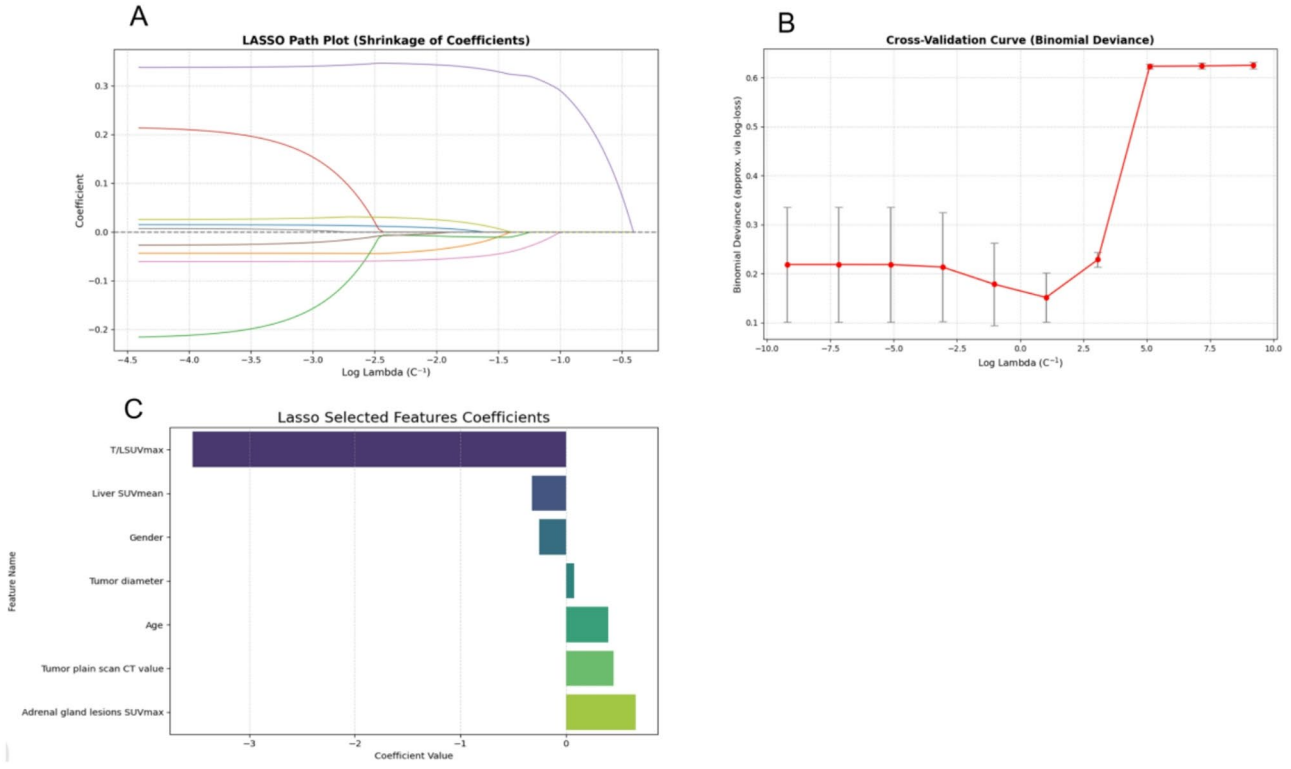

Fig. 3 LASSO feature selection results. **A** LASSO path plot showing the coefficient shrinkage process as a function of log-transformed regularization strength (lambda). **B** 10-fold cross-validation curve indicating the optimal lambda value that minimizes binomial deviance. **C** Final coefficients of selected features based on the optimal lambda. T/L SUVmax was assigned the largest negative weight, while adrenal gland lesions SUVmax and tumor plain scan CT value had positive contributions to the model

The final model (Fig. 3-C) retained seven variables, including: T/L SUVmax, which exhibited the largest negative coefficient, indicating that lower values strongly predict malignant lesions. Adrenal gland lesions SUVmax and tumor plain scan CT value, both showing positive coefficients, suggesting that higher values are associated with increased malignancy probability. Additional features such as age, tumor diameter, gender, and liver SUVmean were retained with smaller but non-zero coefficients, implying secondary predictive roles. Overall, the LASSO-selected feature set reinforces the key role of metabolic and anatomical imaging parameters in malignancy classification, while also providing a compact and interpretable model input set for subsequent supervised learning.

Exploratory analysis of feature distributions across benign and malignant adrenal lesions

In order to evaluate the discriminative capacity of individual clinical and imaging features with respect to the binary classification of adrenal lesions (0 = Benign, 1 = Malignant), a comparative analysis was conducted using boxplots for seven key variables (Fig. 4). These include: Adrenal gland lesions SUVmax, Age, Liver SUVmean, T/L SUVmax ratio, Tumor diameter, Tumor plain scan CT value, Gender.

This exploratory visualization provides insights into the potential of each feature in distinguishing between benign and malignant cases, serving as a basis for feature selection in downstream modeling.

Figure 4A shows the distribution of adrenal gland lesion SUVmax. A clear and significant separation is observed between the two groups: benign cases (Target = 0) exhibit substantially higher SUVmax values, while malignant cases (Target = 1) present consistently low SUVmax values with little variation. This marked difference indicates that adrenal SUVmax is a highly informative feature for malignancy prediction.

Figure 4B depicts the distribution of age. The boxplots reveal overlapping distributions and comparable medians between benign and malignant groups, indicating that age does not significantly differ across classes. Thus, while it may contribute modestly within a multivariate model, it lacks standalone discriminative power.

Figure 4C presents liver SUVmean, showing a slightly higher median in the benign group. However, the wide overlap in distributions diminishes its utility as an independent predictor. In Fig. 4D, the T/L SUVmax ratio demonstrates a higher median and broader range in

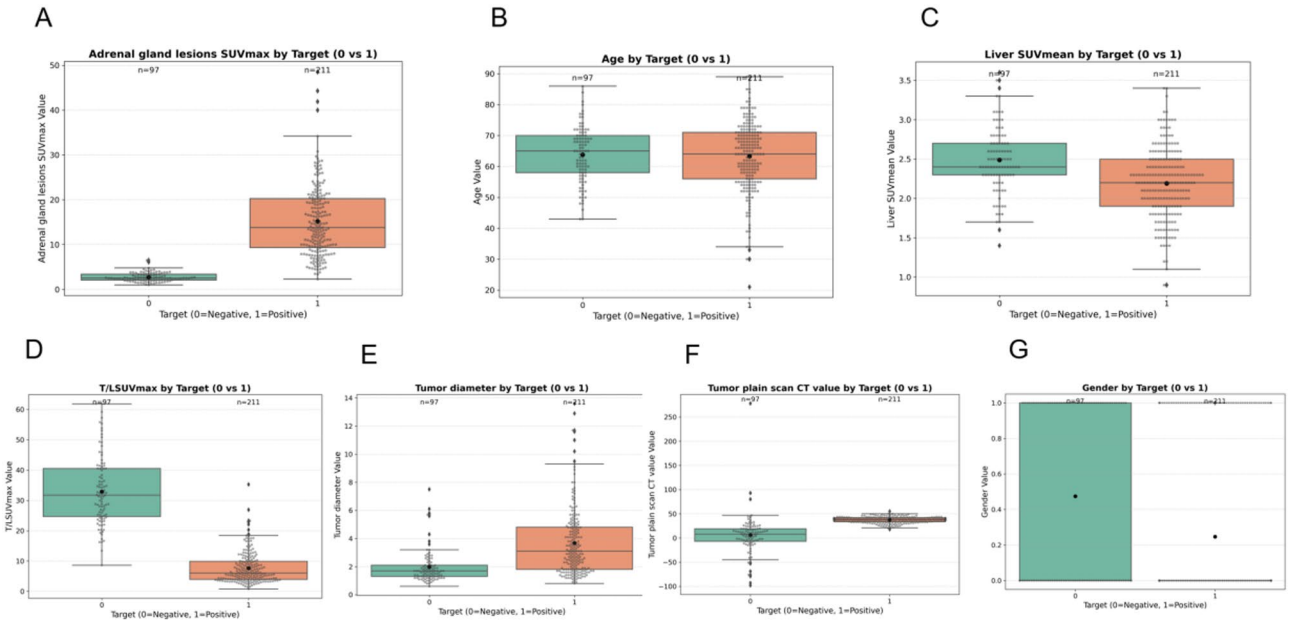

Fig. 4 Distribution of clinical and imaging features by binary target outcome (0 = Negative, 1 = Positive). Panels A–G show comparisons of each variable using boxplots: **A** Adrenal gland lesions SUVmax, **B** Age, **C** Liver SUVmean, **D** T/L SUVmax ratio, **E** Tumor diameter, **F** Tumor plain scan CT value, and **G** Gender

benign cases, suggesting that this ratio may serve as a valuable imaging biomarker for differentiating disease types. Fig. 4E highlights tumor diameter, which tends to be larger in benign cases and shows a wider interquartile range, reinforcing the hypothesis that tumor size correlates with lesion type. Fig. 4F illustrates tumor plain scan CT values. Benign cases consistently present lower CT values than malignant cases, suggesting that radiodensity may be inversely associated with malignancy, and further supports the value of CT in classification. Lastly, Fig. 4G reveals a distinct gender imbalance between the two groups. The majority of benign cases are concentrated in one gender, while malignant cases predominate in the opposite gender, underscoring gender's potential role as a relevant demographic factor in classification.

The most discriminative features include adrenal gland SUVmax, T/L SUVmax, tumor diameter, tumor plain scan CT value, and gender, all showing substantial separability between benign and malignant lesions. Age and liver SUVmean, while less distinct on their own, may still provide supplementary value in multivariate modeling. This analysis lays the groundwork for informed feature selection in subsequent supervised learning workflows.

Feature correlation analysis

Figure 5 presents a correlation heatmap illustrating pairwise Pearson correlation coefficients among clinical, imaging, and demographic features, including the binary classification target. This analysis provides insight into potential feature redundancy and the direct associations between variables.

Among all variables, adrenal gland lesions SUVmax demonstrated the strongest positive correlation with the target variable, indicating that higher SUVmax values in adrenal lesions are strongly associated with the positive class. Conversely, T/L SUVmax exhibited a strong negative correlation with the target, suggesting that lower tumor-to-liver ratios are predictive of the positive outcome—a finding consistent with model-based feature importance analysis.

Tumor plain scan CT value and tumor diameter also showed moderate positive correlations with the target, while gender, age, and liver SUVmean exhibited weak or near-zero correlations, implying limited direct linear association with the outcome. Notably, the correlation coefficients between input features remained moderate ($|r| < 0.7$), suggesting minimal multicollinearity and supporting the inclusion of all features in multivariate models without significant risk of redundancy or instability.

Model performance evaluation and comparison

Continue to comprehensively assess the effectiveness of various classification algorithms in predicting the malignancy of adrenal gland lesions, seven commonly used models: Logistic Regression, Random Forest, AdaBoost, SVM, XGBoost, Bagging, and Artificial Neural Network (ANN) were evaluated. Model performance was compared using multiple metrics, including Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver

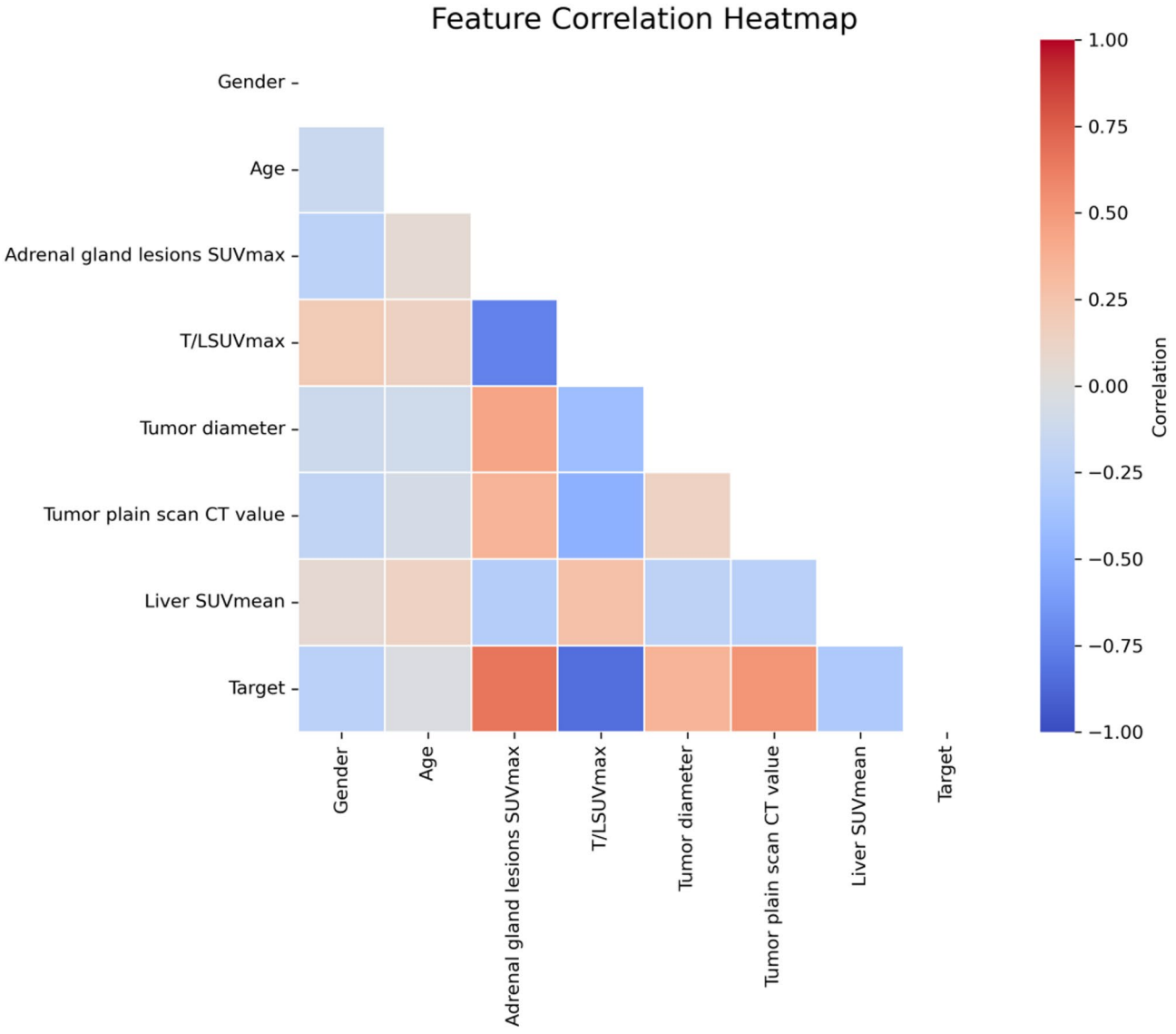

Fig. 5 Feature correlation heatmap. The plot illustrates pairwise Pearson correlation coefficients among all selected variables and the binary target. Color gradients range from − 1.0 (strong negative correlation, blue) to + 1.0 (strong positive correlation, red), with neutral values shown in white

Operating Characteristic Curve (AUC), as presented in Fig. 6 and summarized in Table 2.

Figure 6(A) displays the ROC curves for all models, illustrating high true positive rates across classifiers. Most models achieved an AUC greater than 0.98, with Random Forest, XGBoost, and Bagging reaching perfect or near-perfect scores (AUC = 1.00, 0.998, and 0.999, respectively), indicating excellent discriminative ability. Logistic Regression and SVM also performed robustly (AUC = 0.994 and 0.986), while AdaBoost showed slightly lower performance (AUC = 0.985), though still in the high range.

Figure 6(B) provides a visual comparison of model performance across Accuracy, F1-Score, and AUC. The Bagging model demonstrated the highest overall performance, achieving the highest Recall (1.000), strong Precision (0.955), and the best F1-Score (0.977), indicating its ability to balance sensitivity and precision effectively. Random Forest and XGBoost also showed consistent superiority across metrics, confirming their reliability in this binary classification task.

Table 2 further supports these findings with precise numerical comparisons. Logistic Regression, Random Forest, and Bagging all reached identical Accuracy levels (0.968), while Logistic Regression yielded perfect Precision (1.000) but had slightly lower Recall (0.952), indicating a tendency toward conservative prediction. In contrast, Bagging achieved perfect Recall (1.000), which may be advantageous for clinical applications where minimizing false negatives is critical.

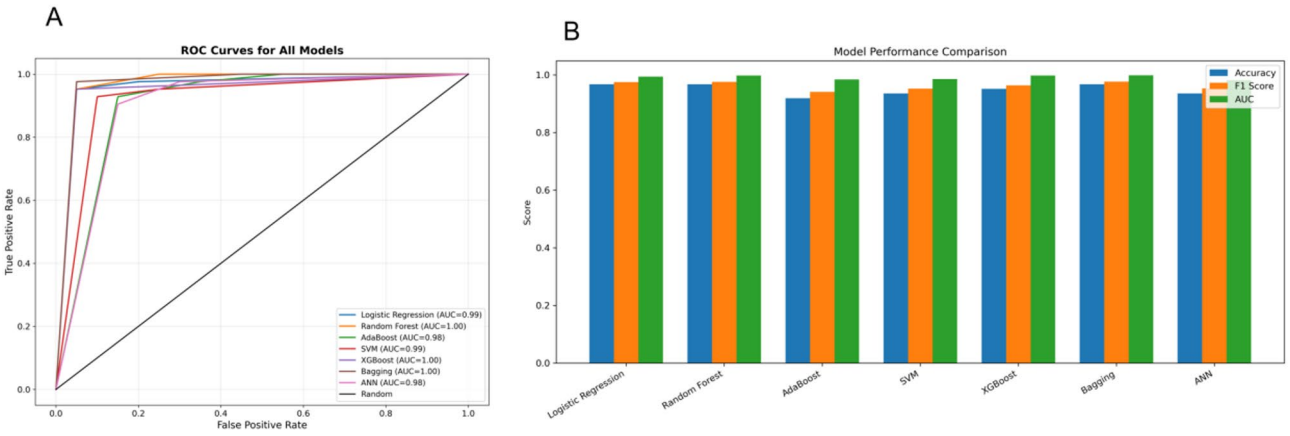

Fig. 6 Model performance comparison using ROC and evaluation metrics. **A** Receiver Operating Characteristic (ROC) curves for seven classification models: Logistic Regression, Random Forest, AdaBoost, Support Vector Machine (SVM), XGBoost, Bagging, and Artificial Neural Network (ANN). All models demonstrate high predictive ability, with AUC values ranging from 0.98 to 1.00. **B** Bar plot comparison of Accuracy, F1 Score, and AUC across the seven models, showing that ensemble methods (Random Forest, XGBoost, Bagging) consistently achieve top performance across all metrics

Table 2 Comparison of machine learning models for predicting adrenal lesions on benign or malignant

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.968	1	0.952	0.976	0.994
Random Forest	0.968	0.976	0.976	0.976	0.998
AdaBoost	0.919	0.930	0.952	0.941	0.985
SVM	0.935	0.952	0.952	0.952	0.986
XGBoost	0.952	0.976	0.952	0.964	0.998
Bagging	0.968	0.955	1	0.977	0.999
ANN	0.935	0.932	0.976	0.953	0.982

Among the ensemble methods, Bagging and XGBoost consistently outperformed AdaBoost, which had the lowest Accuracy (0.919) and F1-Score (0.941). ANN also demonstrated competitive results, particularly in Recall (0.976), suggesting potential value when incorporated into hybrid decision systems. The ensemble-based models—particularly Bagging and Random Forest—exhibited the most robust and stable predictive performance across all evaluation metrics. These results underscore the feasibility and reliability of machine learning algorithms in assisting with the binary classification of adrenal lesion malignancy, providing a solid foundation for clinical decision support tools.

Clinical utility assessment via decision curve analysis

Furthermore evaluate the clinical applicability of each predictive model for distinguishing benign from malignant adrenal gland lesions, we conducted Decision Curve Analysis (DCA), as illustrated in Fig. 7. DCA estimates the net benefit of a model across a continuum of threshold probabilities, thereby quantifying its value in clinical decision-making beyond traditional accuracy-based metrics.

Across a wide range of clinically relevant threshold probabilities (10%–90%), ensemble-based models—particularly Random Forest, XGBoost, and Bagging—demonstrated superior net benefits. Their performance consistently exceeded both the "Treat All" and "Treat None" strategies, indicating that these models offer more favorable risk-benefit trade-offs when guiding decisions about lesion management. This supports their utility in clinical scenarios where accurately identifying malignancy is essential to optimizing patient outcomes while minimizing overtreatment.

Logistic Regression, SVM, and Ada-Boost also exhibited favorable and stable net benefit profiles, especially within the lower-to-intermediate threshold range (10%–60%). These models may therefore be appropriate in more conservative or risk-averse clinical settings, where the cost of false positives is acceptable in exchange for minimizing false negatives.

In contrast, the Artificial Neural Network (ANN) showed a marked decline in net benefit beyond the 70% threshold, accompanied by instability in curve behavior. This suggests potential limitations in its use for high-confidence clinical decisions, particularly in cases requiring stringent thresholding to justify intervention.

Notably, the XGBoost and Bagging models maintained the highest net benefit across nearly the entire decision threshold spectrum. Their consistent out-performance underscores their robustness and potential for integration into real-world diagnostic workflows.

In summary, the DCA results confirm that Random Forest, Bagging, and XGBoost not only yield high predictive performance but also deliver the greatest clinical value in distinguishing benign from malignant adrenal lesions. These models are thus well-suited to serve as decision-support tools in clinical practice, where optimizing diagnostic precision is paramount.

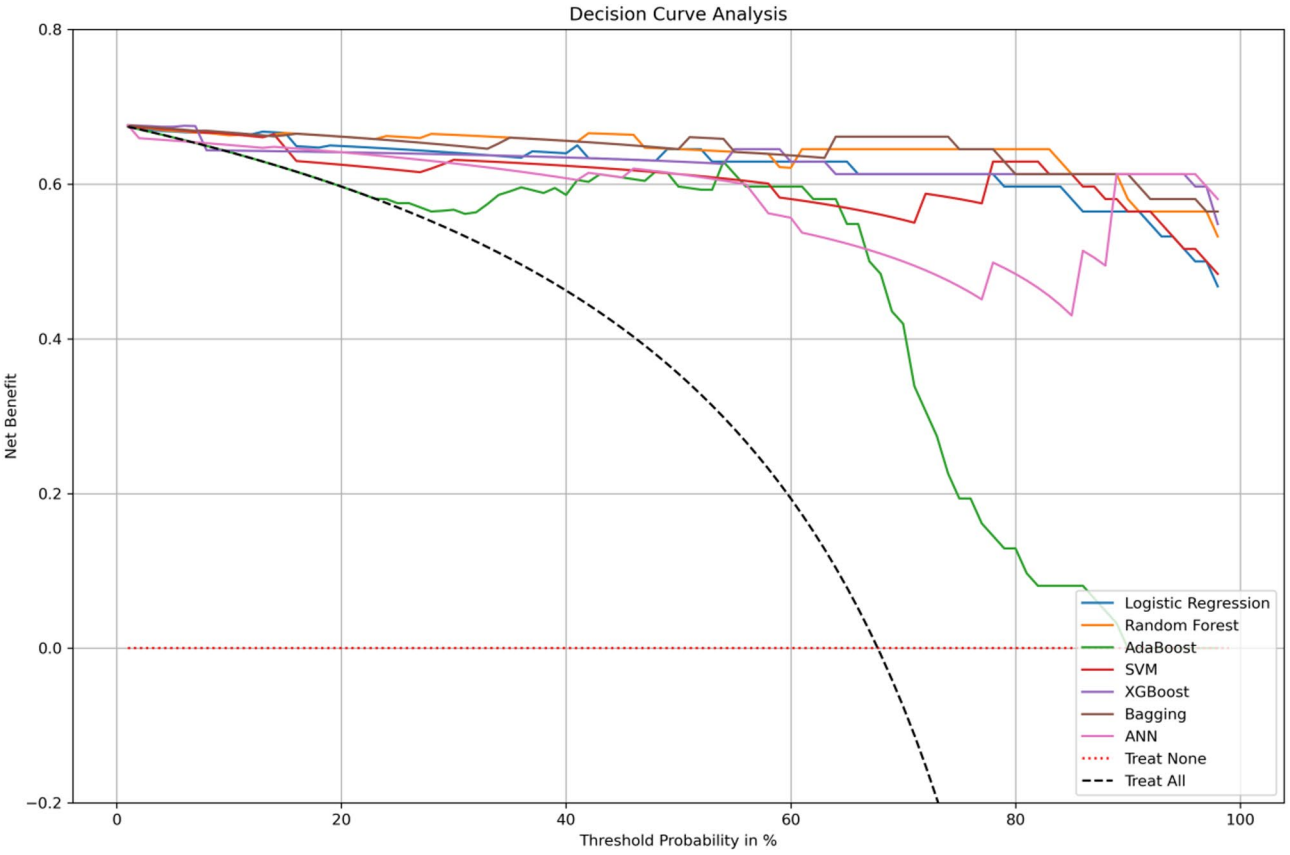

Fig. 7 Decision curve analysis (DCA) comparing the net clinical benefit of different classification models across a range of threshold probabilities. The y-axis represents the net benefit, and the x-axis represents the threshold probability at which a patient would opt for treatment. Models shown include Logistic Regression, Random Forest, AdaBoost, Support Vector Machine (SVM), XGBoost, Bagging, and Artificial Neural Network (ANN). The "Treat All" and "Treat None" strategies are included as reference baselines (dashed lines)

Model explanation using SHAP analysis

This study also gain insight into the interpretability and internal decision logic of the machine learning model, SHAP (SHapley Additive exPlanations) analysis was performed. Figure 8 provides a comprehensive evaluation of both the global and instance-level feature contributions to model output.

Figure 8-A presents the mean absolute SHAP values for each predictor, reflecting their average impact on the model's decision. The feature Adrenal gland lesions SUVmax emerges as the most influential variable, with a mean SHAP value of + 2.31, indicating its dominant role in distinguishing between benign and malignant adrenal lesions. This is followed by Tumor plain scan CT value (+ 1.54) and T/L SUVmax ratio (+ 1.13), both of which also show substantial contributions. Features such as Gender, Tumor diameter, and Liver SUVmean exhibit moderate influence, while Age has the lowest importance, suggesting minimal contribution to the predictive outcome.

Figure 8-B offers a summary SHAP plot, which captures not only the magnitude but also the directional effect of each feature on model predictions. The x-axis represents SHAP values (i.e., impact on the model output), while the color gradient encodes feature values from low (blue) to high (red). This dual encoding allows for interpretation of how specific feature ranges affect classification. Key observations include: Adrenal gland lesions SUVmax: Low values (blue) strongly increase SHAP scores toward the positive class, suggesting a strong association with malignant outcomes. High values (red) are associated with lower SHAP scores, aligning with benign classifications. Tumor plain scan CT value: Higher CT values (red) correspond to positive SHAP values, indicating association with malignancy. T/L SUVmax ratio: Low ratios (blue) decrease the SHAP value (toward benign), while high ratios increase it, supporting its role as a malignancy marker. Liver SUVmean and Tumor diameter: Exhibit more dispersed SHAP distributions, with mixed impact depending on value range. Gender: Shows a modest effect but with notable separation, potentially reflecting imbalanced class associations. Age: Demonstrates limited impact across all SHAP values, reinforcing its weak discriminative power.

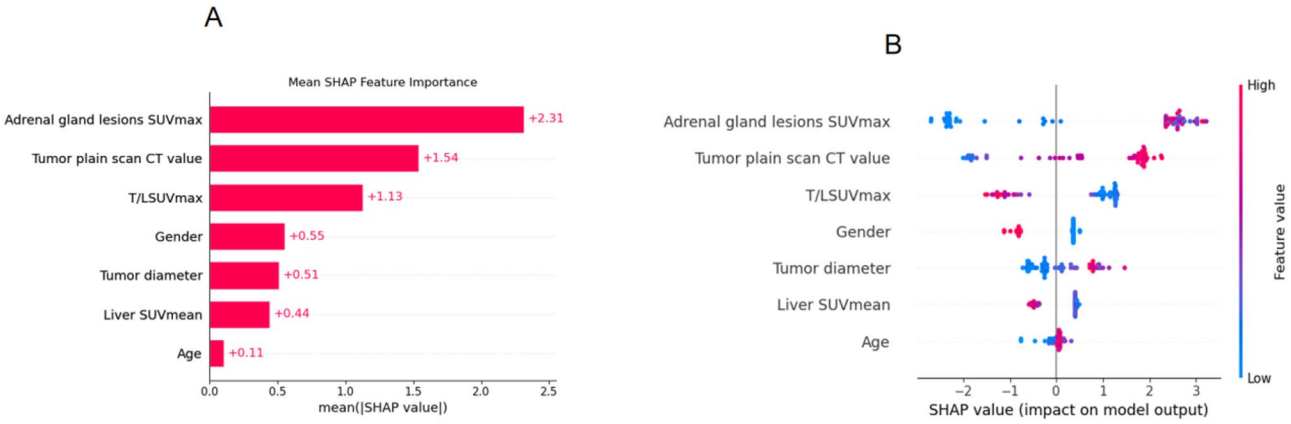

Fig. 8 **A** The bar chart shows the mean absolute SHAP values for each feature, representing their average contribution to the model's prediction of adrenal lesion malignancy. Features with higher SHAP values are more influential. **B** The SHAP summary plot visualizes the distribution and direction of each feature's impact. Each point represents an individual patient sample. The x-axis indicates whether the feature pushes the prediction toward benign (left) or malignant (right). The color scale reflects the actual feature value: red for high values and blue for low values. For example, lower adrenal SUVmax (blue) tends to be associated with malignancy in our dataset, while higher CT values (red) also contribute to malignancy prediction. This plot helps visualize nonlinear effects and variable interactions on the model output.

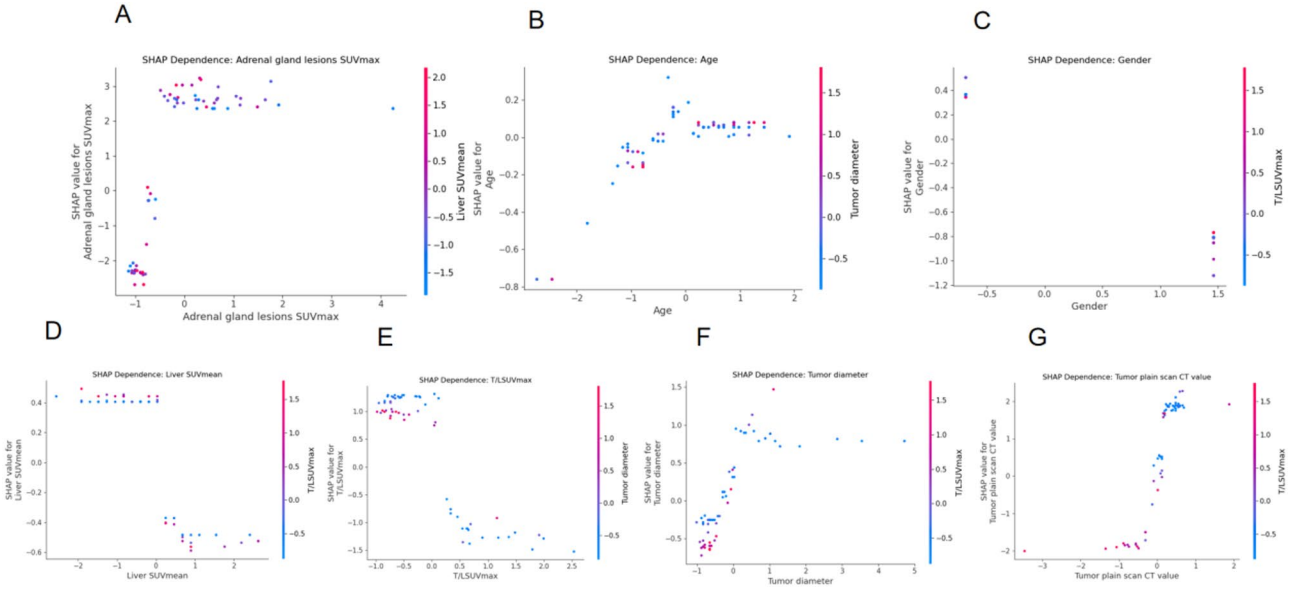

Fig. 9 SHAP dependence plots showing nonlinear and interaction effects among key predictors.Each subplot (**A**–**G**) illustrates the relationship between a specific feature's value (x-axis) and its SHAP value (y-axis), which indicates the feature's contribution to predicting malignancy. A higher SHAP value means the feature is pushing the model toward predicting malignancy.The color of each point represents the value of a second interacting feature, helping to reveal interaction effects. For example, in (A), lower adrenal SUVmax is associated with higher SHAP values (malignancy), and the liver SUVmean (color gradient) modulates this relationship. These plots help explain the nonlinear behavior of features and their combined impact on prediction, making the model more interpretable to clinicians

The SHAP results provide robust evidence of feature importance and directionality, complementing the model's statistical performance with clear clinical interpretability. These findings underscore the significance of SUV-based imaging markers and CT attenuation values as key predictors in differentiating adrenal lesion malignancy.

Dependence and interaction effects

Regarding to elucidate the model's nonlinear behavior and inter-feature dependencies, SHAP dependence plots were generated for the seven most influential variables (Fig. 9). These plots provide a nuanced view of how each feature individually—and in interaction with others—affects the prediction of adrenal lesion malignancy. Adrenal gland lesions SUVmax: A clear threshold effect is observed. While SUVmax above ~1.5 contributes

minimally, values below this threshold sharply increase SHAP values, strongly influencing malignancy predictions. This effect is further modulated by liver SUVmean, highlighting the relevance of background metabolic activity. Tumor plain scan CT value: Demonstrates a strong and consistent positive SHAP response to increasing CT values, especially under higher T/L SUVmax conditions, reinforcing the interaction between anatomical density and metabolic intensity. T/L SUVmax ratio: Shows a monotonic relationship with SHAP value—higher ratios are consistently predictive of malignancy. Interaction with tumor diameter suggests that metabolic significance is partially size-dependent. Gender: Although a categorical variable, gender displays distinct SHAP value separation, with metabolic ratio (T/L SUVmax) further differentiating the effect, hinting at sex-specific biological patterns. Age: Exhibits a mild positive relationship with malignancy risk, especially in larger tumors, again suggesting size-age synergy. Liver SUVmean and Tumor diameter: Both reveal discrete transitions in SHAP values—rather than gradual trends—indicating the presence of threshold behavior in decision logic.

Overall, the dependence plots reinforce the dominance of Adrenal SUVmax, CT value, and T/L SUVmax as primary drivers of prediction. Moreover, the color-coded interaction overlays uncover clinically relevant combinatorial effects, thereby enhancing model transparency and interpretability. These findings support the feasibility of incorporating such models into diagnostic workflows for adrenal lesion risk stratification.

Subgroup analysis: discriminating lung cancer from lymphoma in malignant adrenal lesions

Following the successful development of models for distinguishing benign from malignant adrenal lesions, we further refined our analysis by focusing solely on the malignant cases. Specifically, we aimed to differentiate between two common types of adrenal malignancies: lung cancer metastases and lymphoma. Accurate discrimination between these sub-types holds significant clinical relevance, as therapeutic strategies and prognoses differ substantially. To address this, a secondary dataset was constructed comprising only malignant cases, with the target variable (Primary stove) indicating either lung cancer (coded as 3) or lymphoma (coded as 4). A new binary classification task was thus established, and the complete machine learning pipeline was reapplied to evaluate model performance, interpretability, and clinical utility.

Feature selection via LASSO regularization

Identify the most relevant features for subclass classification is very important, this study applied LASSO logistic regression. The regularization path plot showed progressive coefficient shrinkage as the penalty increased, while 5-fold cross-validation identified the optimal lambda minimizing binomial deviance. Seven features were retained in the final model: Adrenal gland lesions SUVmax, Adrenal gland lesions SUVpeak, T/LSUVmax, Tumor diameter, Tumor plain scan CT value, Liver SUVmean, Age and Gender. These features reflect a combination of metabolic, anatomical, and demographic characteristics, suggesting that both localized uptake and systemic parameters contribute to subtype differentiation, as shown in Fig. 10.

Univariate distribution analysis

Figure 11 presents comparative box-plots illustrating the univariate distribution of key imaging and clinical features between patients diagnosed with lung cancer and those with lymphoma. The analysis yields several statistically and clinically relevant observations:

Adrenal gland lesions SUVmax and SUVpeak: Both parameters, indicative of metabolic activity within adrenal lesions, were significantly higher in lymphoma patients. This aligns with the biological behavior of lymphomas, which are generally characterized by higher FDG avidity due to intense cellular proliferation and glycolytic activity. These findings suggest that hypermetabolic adrenal lesions are more likely associated with lymphoma in patients presenting with malignancy.

T/LSUVmax ratio: This ratio quantifies lesion-to-liver contrast and was markedly elevated in the lymphoma subgroup, indicating that the background-normalized metabolic contrast is more pronounced in lymphomatous involvement. This supports the clinical utility of the T/L ratio in improving the discrimination of highly metabolic lesions.

Tumor plain CT value: Lung cancer patients tended to exhibit higher non-contrast CT values, suggesting that metastatic lesions in these cases may present with greater tissue density, possibly due to fibrosis, calcification, or hemorrhagic components. This morphological difference provides an anatomic correlate to the metabolic heterogeneity between subtypes.

Tumor diameter: Although differences in median tumor size between groups were modest, lymphomas showed a wider interquartile range, reflecting variable presentations from bulky disease to small-volume involvement. This heterogeneity is consistent with the known variable burden of lymphoma and may impact classification robustness.

Liver SUVmax and SUVmean: Both parameters were slightly elevated in the lung cancer group, which may reflect systemic inflammatory or paraneoplastic metabolic activation, especially in advanced-stage disease. This trend, though modest, adds context to the differential metabolic background.

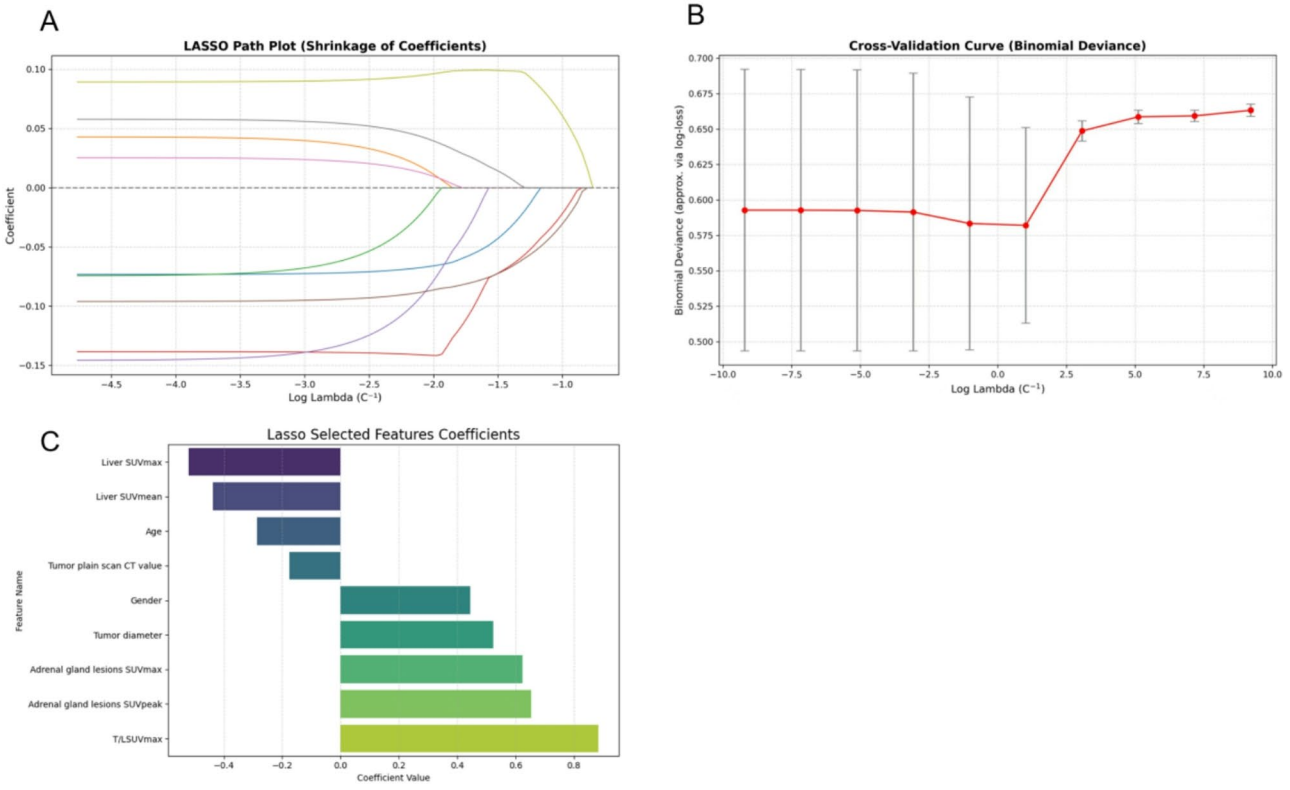

Fig. 10 LASSO-based feature selection for subclassification of malignant adrenal lesions (lung cancer vs. lymphoma). **A** LASSO coefficient path illustrating the shrinkage of coefficients as the regularization parameter (log λ) increases. **B** Cross-validation curve showing binomial deviance across different λ values. The optimal λ is selected based on the minimum deviance with the one-standard-error rule. **C** Coefficients of the features retained by LASSO at the optimal λ. Eight variables were selected, including T/LSUVmax, Adrenal gland SUVmax, SUVpeak, Tumor diameter, Gender, Tumor CT value, Age, and Liver SUV metrics. Notably, T/LSUVmax and Adrenal SUV metrics showed the strongest positive associations with lymphoma classification

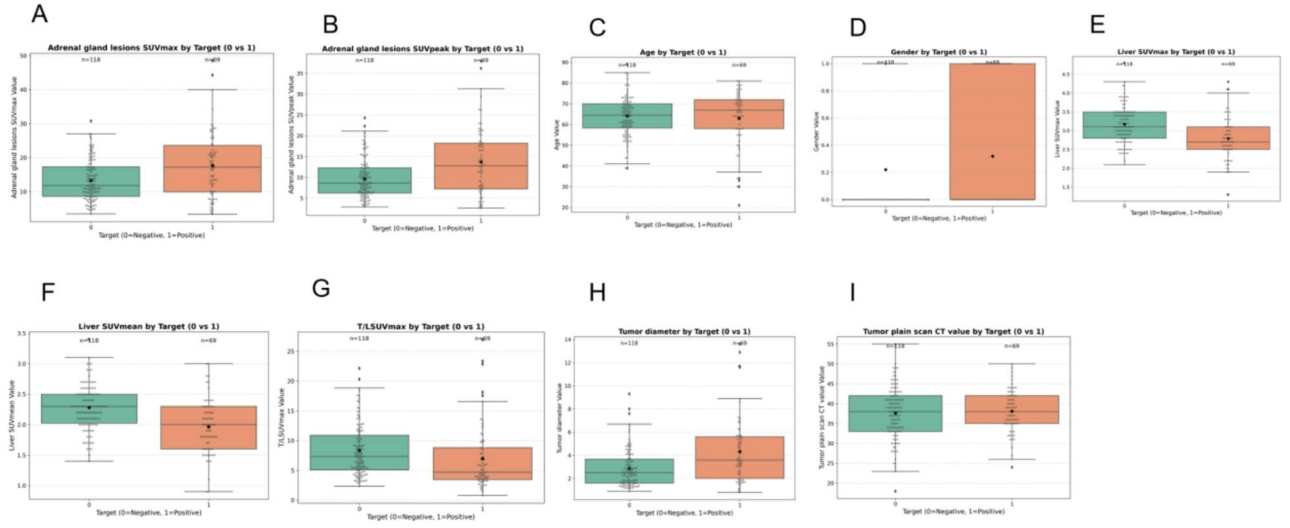

Fig. 11 Boxplot Comparison of Feature Distributions Between Lung Cancer and Lymphoma Subgroups. Panels A–I display the distribution of nine key imaging and demographic variables across two diagnostic groups: lung cancer (Target = 0) and lymphoma (Target = 1). **A**–**B** Adrenal gland lesions SUVmax and SUVpeak. **C**–**D** Age and Gender. **E**–**F** Liver SUVmax and SUVmean (**G**) T/LSUVmax ratio. **H** Tumor diameter. **I** Tumor plain scan CT value

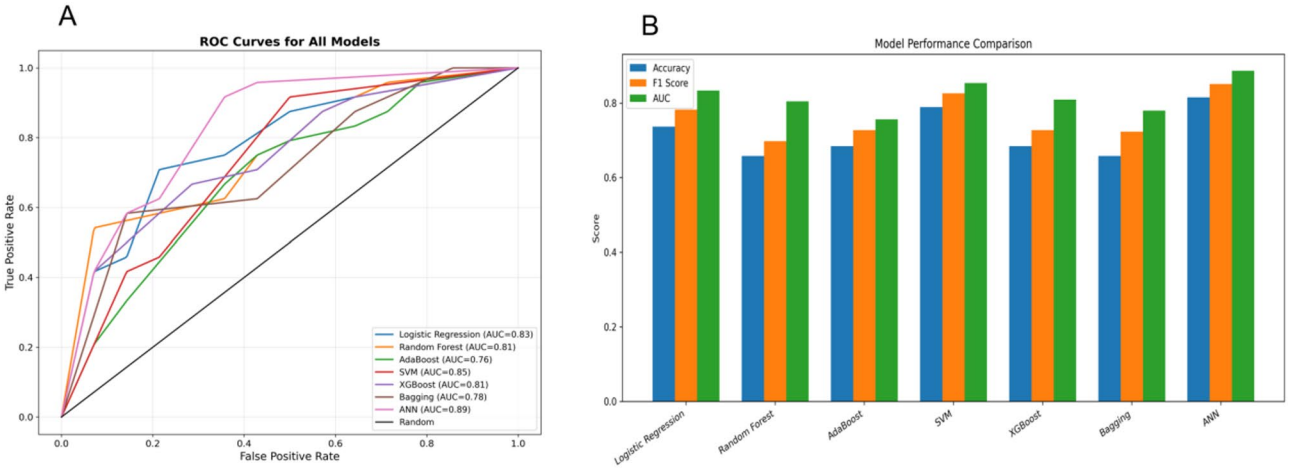

Fig. 12 Performance comparison of machine learning models for classifying lung cancer vs. lymphoma. **A** Receiver operating characteristic (ROC) curves for all eight models evaluated. **B** Bar plots showing Accuracy, F1-score, and AUC for each model

Demographics – Age and Gender: Younger age and male gender were more frequently observed in the lymphoma group. This demographic distribution corresponds well with known epidemiological patterns, where certain subtypes of lymphoma (e.g., Hodgkin) have a higher prevalence in young males. These variables, although not strongly predictive individually, may serve as supporting covariates in multivariate modeling.

Model construction and performance comparison

To differentiate between lung cancer and lymphoma among malignant adrenal lesions, this study constructed and evaluated a series of supervised machine learning classifiers based on imaging-derived features and clinical variables. The input features included quantitative PET/CT parameters such as adrenal SUVmax, SUVpeak, T/LSUVmax, liver SUVmean and SUVmax, tumor diameter, and plain CT value. These were selected using LASSO regression, while age and gender were retained to account for potential demographic confounding.

Seven classifiers were developed and tuned using stratified cross-validation, including logistic regression, random forest, AdaBoost, support vector machine (SVM), XGBoost, bagging, and artificial neural network (ANN). This comprehensive modeling strategy ensured both linear and nonlinear decision boundaries were considered, capturing a wide range of decision-making patterns.

Performance was first evaluated using ROC curves. As shown in Fig. 12A, ANN achieved the highest area under the curve (AUC = 0.887), indicating excellent ability to distinguish between the two malignancy sub-types. SVM followed with an AUC of 0.854, and logistic regression also demonstrated strong discrimination (AUC = 0.833), underscoring the predictive strength of the selected features even within a linear modeling framework. In contrast, Ada-Boost showed the lowest AUC (0.756), possibly due to overfitting or sensitivity to sample imbalance.

Table 3 Comparison of machine learning models for predicting distinguishing benign from malignant adrenal lesions

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.737	0.818	0.75	0.783	0.833
Random Forest	0.658	0.789	0.625	0.698	0.805
AdaBoost	0.684	0.8	0.667	0.727	0.756
SVM	0.789	0.864	0.792	0.826	0.854
XGBoost	0.684	0.8	0.667	0.727	0.810
Bagging	0.658	0.739	0.708	0.723	0.780
ANN	0.816	0.870	0.833	0.851	0.887

A more detailed comparison of model performance is provided in Fig. 12B; Table 3, which report metrics including accuracy, precision, recall, F1-score, and AUC. ANN consistently outperformed all other models, with the highest accuracy (0.816), precision (0.870), recall (0.833), and F1-score (0.851). SVM also exhibited strong and balanced performance, while logistic regression offered a desirable trade-off between accuracy (0.737) and interpretability. Bagging and Ada-Boost, although part of ensemble methods, yielded relatively lower performance across most metrics, suggesting limited benefit in this dataset.

While ANN provides the best predictive performance, its complex structure hinders transparency, which may reduce its acceptability in clinical practice. In contrast, models like logistic regression and random forest, although slightly less accurate, provide intuitive interpretability and decision logic, which are critical for clinical integration. These results suggest that while ANN and SVM are highly effective for sub-type classification, simpler models remain valuable for transparent clinical deployment. The optimal model choice may thus depend

on the specific diagnostic scenario and the balance between accuracy and interpretability required for real-world application.

Clinical utility evaluation by decision curve analysis (DCA)

To assess the clinical relevance of predictive models, decision curve analysis (DCA) was employed, which estimates the net benefit of each model by balancing true positives against false positives across a continuum of clinical threshold probabilities. This approach simulates real-world decision-making conditions and allows comparison of model utility in guiding diagnostic or therapeutic actions. As illustrated in Fig. 13, the artificial neural network (ANN), support vector machine (SVM), and XGBoost consistently yielded higher net benefits across a broad range of thresholds (10% to 80%), indicating their strong clinical applicability in scenarios involving diagnostic uncertainty for adrenal malignancies. Compared to baseline strategies such as "Treat All" or "Treat None," all machine learning models demonstrated superior net benefit, reinforcing the advantage of algorithmic triage. Notably, the ANN model sustained its performance even at higher threshold probabilities, suggesting robustness under conservative decision thresholds. While ANN demonstrated the highest overall utility, SVM and XGBoost also offered favorable performance with potentially lower computational costs and better interpretability, making them attractive alternatives for institutions with limited resources or a higher need for model transparency.

SHAP-based interpretability analysis

SHAP (SHapley Additive exPlanations) analysis was conducted on the ANN classifier to enhance transparency and interpretability of the model's decision-making process. As shown in Fig. 14. This approach quantifies the marginal contribution of each feature to individual predictions, allowing insight into how specific variables drive classification outcomes. The analysis revealed that adrenal SUVmax was the most influential predictor, with elevated values significantly increasing the likelihood of lymphoma. SUVpeak and T/LSUVmax also ranked among the top contributors, further highlighting the importance of metabolic activity in distinguishing lymphoma from lung cancer within adrenal lesions. Conversely, higher liver SUVmax and SUVmean were

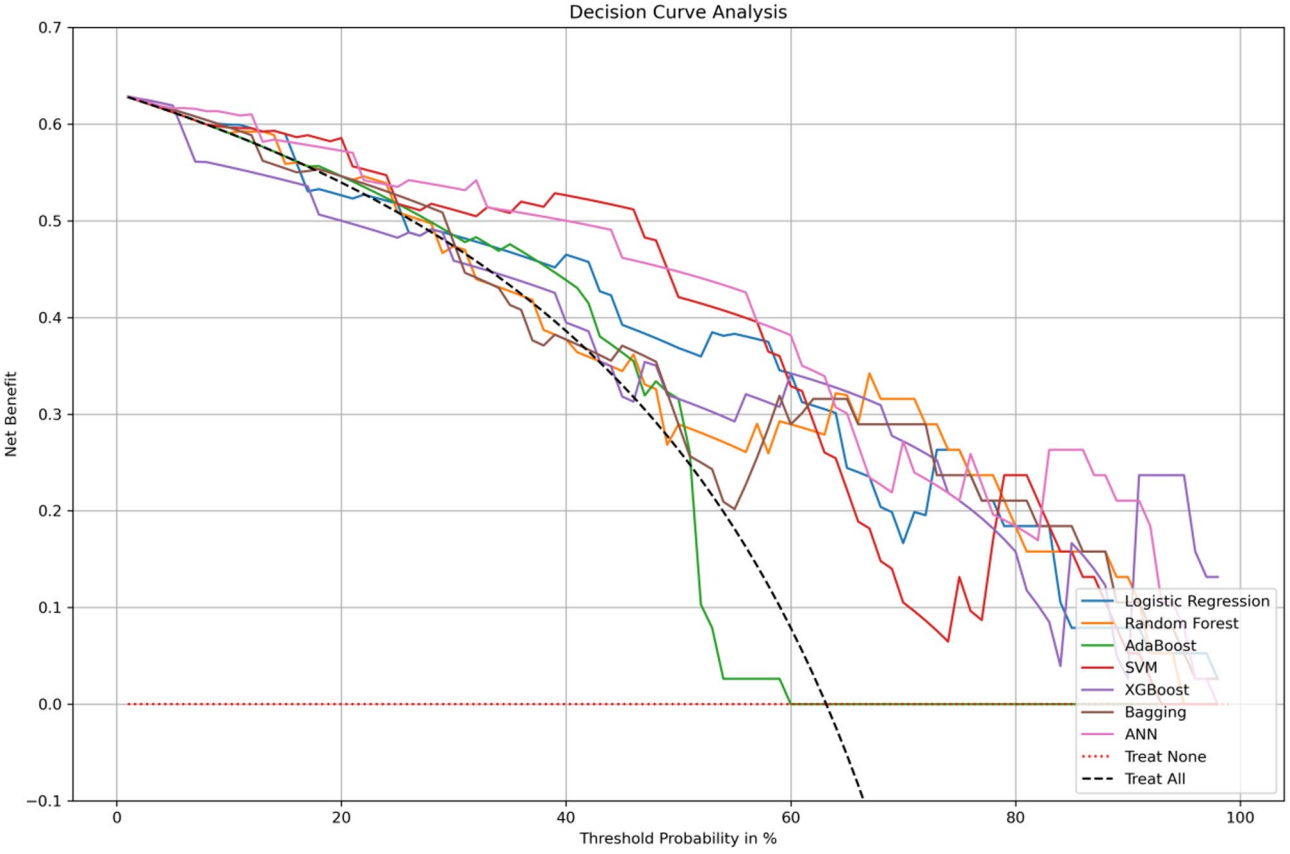

Fig. 13 Decision curve analysis (DCA) for clinical utility evaluation of classification models. This plot illustrates the net clinical benefit of various predictive models across a range of threshold probabilities (0–100%). The y-axis represents the net benefit, and the x-axis indicates the threshold probability at which a clinician would choose to intervene

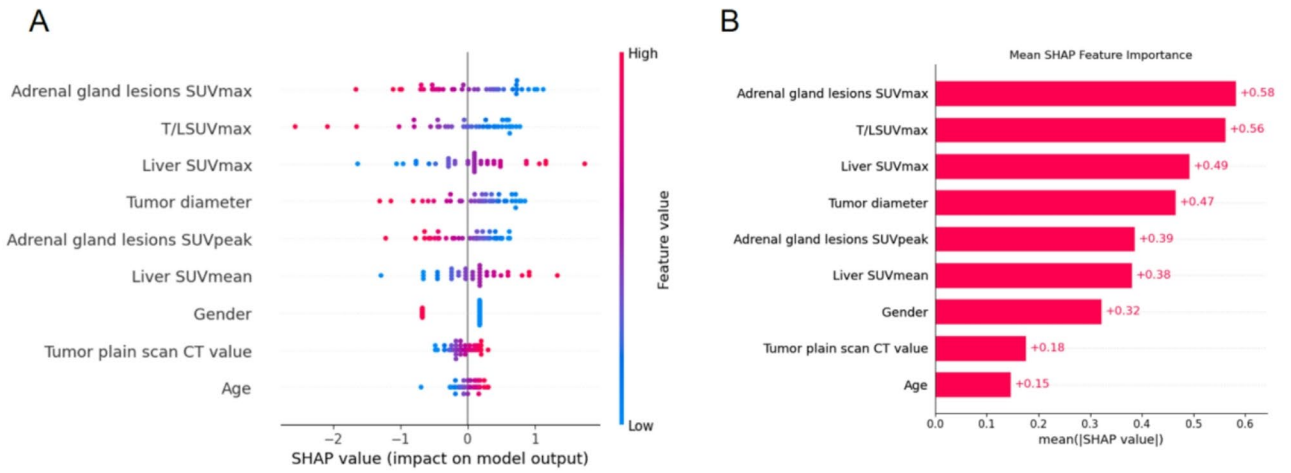

Fig. 14 SHAP Analysis for Interpreting Model Predictions. **A** SHAP summary plot showing the distribution of feature effects on individual predictions. Each dot represents a sample, and color indicates the magnitude of the feature value. **B** Mean SHAP values quantifying the global importance of each feature

more commonly associated with lung cancer, potentially reflecting systemic inflammatory responses or liver uptake characteristics related to pulmonary primary tumors. These inverse associations were clearly evident in the SHAP value distributions. The tumor plain CT value demonstrated a moderate but complex contribution, with its predictive impact varying depending on interactions with adrenal SUV metrics, suggesting it serves a complementary role rather than a primary determinant. Demographic variables, including age and gender, showed weaker yet consistent effects, improving overall model calibration when integrated with imaging-derived features. Finally, the mean SHAP importance rankings confirmed the dominance of adrenal SUVmax, T/LSUVmax, and SUVpeak, reinforcing their central role in model behavior and clinical interpretation.

Local feature dependence analysis based on SHAP

To further elucidate the non-linear influence of individual variables and their interactions on the model's predictions, SHAP dependence plots were generated for each of the selected features, as shown in Fig. 15. These plots illustrate the relationship between feature values and their corresponding SHAP contributions, with a second variable encoded by color to reveal potential feature interactions. Among all predictors, adrenal lesion metabolic indicators—namely SUVmax and SUVpeak—exhibited the strongest positive SHAP values, with higher metabolic uptake consistently pushing predictions toward the lymphoma category. The observed color gradients further indicate that elevated SUVpeak values amplify the effect of SUVmax, reflecting a synergistic interaction between the two features in characterizing metabolically active lesions. T/LSUVmax, a composite lesion-to-liver contrast metric, showed a sharp non-linear increase in its SHAP contribution once a normalized threshold was exceeded, emphasizing its threshold-like behavior in distinguishing lymphomas.

In contrast, liver SUVmax and SUVmean were negatively associated with lymphoma prediction, indicating that greater background hepatic uptake favors lung cancer classification, potentially reflecting systemic metabolic activation. Age and gender exerted relatively smaller effects, but still showed meaningful patterns: younger male patients tended to be classified as lymphoma, and age displayed a U-shaped SHAP distribution modulated by other variables such as T/LSUVmax. Additionally, tumor diameter and plain CT value contributed modestly to model output. Smaller tumor sizes were linked to negative SHAP values, while higher CT densities shifted predictions toward lung cancer, consistent with denser metastatic morphology. Overall, these dependence analyses underscore that the model not only relies on individual feature magnitudes but also captures their complex interactions. The interpretability afforded by SHAP enhances transparency in decision-making and supports the biological plausibility of the identified predictive patterns.

Discussion

Overview of study objectives and approach

This study aimed to explore the feasibility and clinical utility of applying interpretable machine learning algorithms to classify adrenal lesions using quantitative features derived from 18 F-FDG PET/CT imaging and clinical variables. The research was structured into two major predictive tasks: first, the binary classification of adrenal lesions as benign or malignant; and second, the subclassification of malignant lesions into lung cancer metastases and lymphoma. By integrating

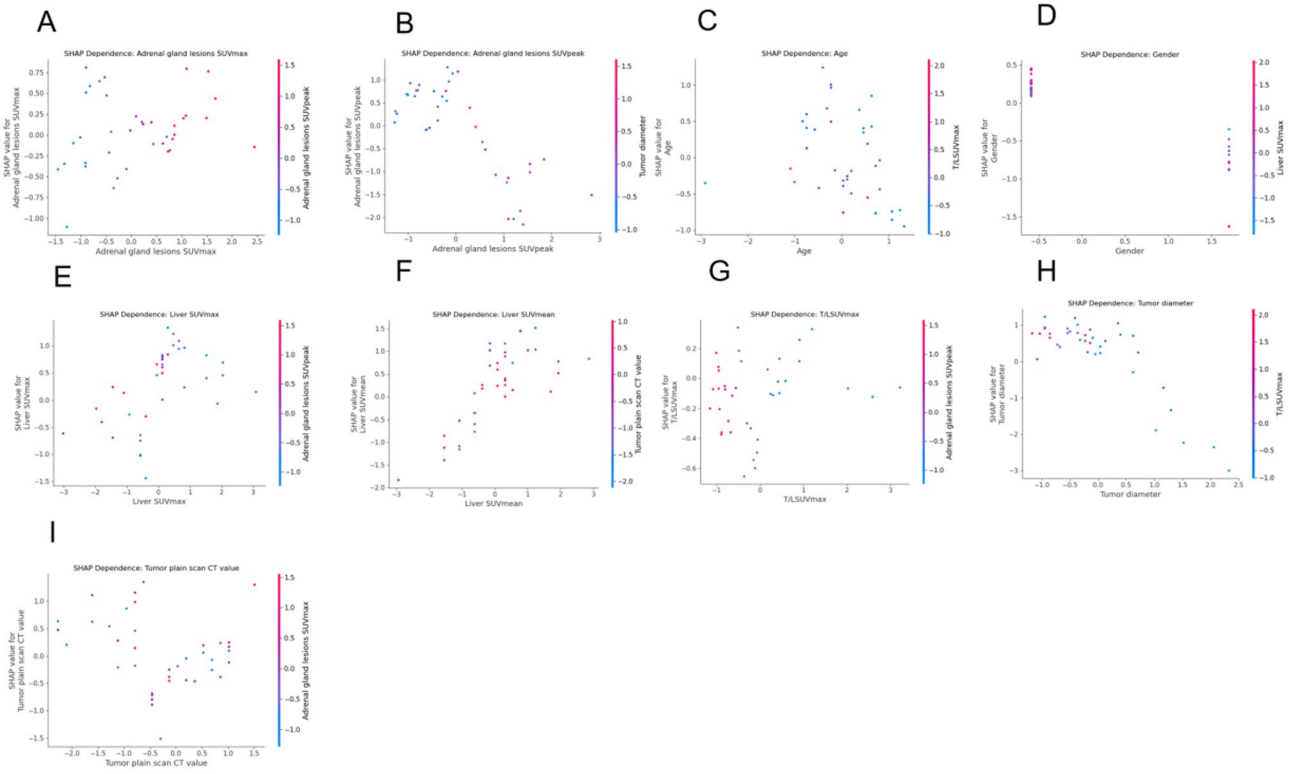

Fig. 15 SHAP dependence plots of selected predictors. Each subplot (**A**–**I**) displays the SHAP dependence plot for one of the selected features in the ANN model. The x-axis represents the standardized feature value, and the y-axis shows the SHAP value (i.e., the feature's contribution to model output). Color gradients represent a second interacting feature to reveal potential feature interactions

both metabolic and anatomical imaging biomarkers with clinical demographics, the models were trained and evaluated to identify patterns that support diagnostic decision-making.

Classification performance in benign vs. malignant lesions

In the primary task, metabolic features such as adrenal SUVmax and T/L SUVmax ratio, as well as anatomical variables like non-contrast CT value and tumor diameter, emerged as highly discriminative indicators. These findings are consistent with known oncologic imaging principles, in which malignant tumors tend to exhibit increased glycolytic activity and altered tissue structure [10, 11]. Among all tested algorithms, ensemble models—particularly Bagging, Random Forest, and XGBoost—achieved near-perfect performance, with Bagging achieving an AUC greater than 0.99 and 100% recall. Such high sensitivity is especially desirable in clinical contexts where failing to detect malignancy may lead to harmful delays in intervention. While logistic regression offered excellent precision, its slightly lower recall suggests a tendency toward under-identification of malignancies, which may limit its utility in screening applications where sensitivity is prioritized.

Classification performance in malignant subtype discrimination

In the secondary task, distinguishing between malignant lesion subtypes, namely lung cancer and lymphoma, the best-performing models were artificial neural networks (ANN) and support vector machines (SVM). The ANN model achieved the highest area under the curve (AUC = 0.887) and F1-score (0.851), demonstrating its capacity to capture complex nonlinear patterns among PET/CT features. Subgroup analysis revealed that lymphoma cases were associated with elevated adrenal SUVmax, SUVpeak, and T/L SUVmax, reflecting the high metabolic intensity and homogeneous cellularity characteristic of lymphoproliferative tumors. In contrast, lung metastases demonstrated higher CT attenuation and elevated liver metabolic background, possibly due to systemic inflammation or necrotic tumor burden.

SHAP-based feature interpretation

To further elucidate the decision logic of the models, SHAP (SHapley Additive exPlanations) analysis was conducted for both classification tasks. Importantly, the most influential predictors varied between the benign–malignant classification and the subtype discrimination, reflecting underlying biological and metabolic distinctions.

SHAP analysis for benign vs. malignant lesions

In the task of differentiating benign from malignant adrenal lesions, SHAP analysis identified adrenal SUVmax, T/L SUVmax ratio, and tumor CT value as the top contributors. Their high impact is consistent with well-established clinical and biological mechanisms:

Adrenal SUVmax measures the maximum intensity of FDG uptake and reflects glycolytic activity, which is typically elevated in malignant tumors due to the Warburg effect—upregulation of glucose transporters (GLUT1) and hexokinase [12, 13]. Interestingly, in our model, lower SUVmax values were associated with malignancy, particularly in adrenal metastases from lung cancer, which often exhibit moderate or diffuse uptake due to necrosis or low cellular density. This seemingly counterintuitive finding can be explained by several biological factors. First, adrenal metastases—particularly those from lung cancer—frequently exhibit central necrosis, hemorrhage, or cystic degeneration, resulting in reduced FDG accumulation and heterogeneous uptake patterns that lower the measured SUVmax. Second, certain benign lesions, such as lipid-poor adenomas or inflammatory myelolipomas, may display unexpectedly high SUVmax due to increased cellular density or macrophage activity, mimicking malignancy on PET/CT. Third, the inclusion of T/L SUVmax as a normalized parameter helped contextualize SUVmax within the patient's metabolic background. Lesions with relatively low T/L ratios despite moderately elevated SUVmax were more likely to be malignant, which aligns with prior studies suggesting that absolute SUVmax alone may not be a reliable discriminator. This complex interplay highlights the value of using SHAP analysis to uncover non-linear and non-obvious relationships that reflect underlying pathophysiological diversity. Meanwhile, some benign adenomas may display relatively high SUVmax because of lipid-poor composition or inflammatory activity, leading to overlapping distributions [5].

T/L SUVmax ratio, which normalizes lesion uptake to hepatic background, accounts for interpatient variability and scanner differences. A low T/L ratio is a robust indicator of malignancy, as it reflects diminished metabolic contrast between aggressive tumors and liver parenchyma. This ratio has shown superior diagnostic performance in previous PET-based adrenal assessments compared to SUVmax alone [14].

While unenhanced CT attenuation remains a widely accepted first-line discriminator for adrenal lesion characterization—particularly the use of ≤10 Hounsfield Units (HU) to identify lipid-rich adenomas, as stated in the ESE/ENSAT Guidelines [15]—this criterion alone is often insufficient for evaluating lipid-poor adenomas or lesions in oncologic patients. Recent studies, such as that by Feliciani et al. [16], have emphasized the limitations of relying solely on density-based metrics for accurate classification.

In our cohort, although benign lesions generally exhibited lower CT values than malignant ones, our machine learning model integrated this anatomical feature with metabolic (e.g., SUVmax, T/L SUVmax) and morphologic parameters (e.g., diameter), enhancing the ability to distinguish indeterminate lesions, including those that fall outside the classic HU threshold.This is consistent with recent findings by Feliciani et al. (2023), who demonstrated that radiomic features extracted from unenhanced CT can improve the characterization of lipid-poor adrenal adenomas, overcoming the limitations of traditional HU-based thresholds alone [17].

The SHAP interpretability analysis further revealed that the contribution of CT attenuation to prediction was context-dependent, interacting with features such as T/L SUVmax and lesion size. This underscores the incremental value of multivariate, data-driven approaches in scenarios where traditional thresholds may not suffice for confident diagnosis.

Similarly, tumor diameter has long been recognized as a relevant factor in adrenal lesion assessment, with larger tumors often considered suspicious for malignancy. However, our SHAP analysis reveals that diameter contributes to prediction in a non-linear and context-dependent manner, rather than functioning as a fixed cutoff (e.g., >4 cm). For example, the malignancy risk associated with a 3.5 cm lesion may vary significantly depending on its metabolic activity and density. By modeling these interactions, our machine learning approach offers incremental diagnostic value beyond individual feature thresholds, especially in ambiguous or borderline lesions where conventional rules may be insufficient.

These findings emphasize that the strength of the model lies not in rediscovering known predictors, but in integrating them holistically to generate interpretable, individualized risk assessments in clinically relevant populations.

SHAP dependence analysis further confirmed: A threshold effect for SUVmax below ~1.5 in predicting malignancy, especially with low liver SUVmean. An amplifying interaction between T/L SUVmax and tumor size. Enhanced CT-based discrimination when combined with low metabolic contrast. Secondary variables like gender, age, tumor diameter, and liver SUVmean played modest roles, contributing to risk calibration rather than direct classification.

SHAP analysis for lung cancer vs. lymphoma in malignant lesions

For malignant cases, the second classification task aimed to distinguish lung cancer metastases from lymphoma, and SHAP revealed a distinct feature importance profile:

Adrenal SUVmax and SUVpeak were strongly associated with lymphoma, which is characterized by dense, homogenous cell populations with high metabolic rates. Unlike carcinomas, lymphomas lack necrosis and calcification, leading to intense, uniform FDG uptake across the lesion [18]. These features made SUVmax and SUVpeak decisive predictors in subtype classification.

T/L SUVmax also favored lymphoma when elevated, reflecting strong lesion-to-liver contrast typical of lymphoproliferative disease. Conversely, lung metastases may appear metabolically less active or heterogeneous, particularly in partially necrotic lesions.

Tumor CT value was higher in lung cancer metastases, likely due to fibrotic or hemorrhagic components. Lymphomas, being soft-tissue dominant and less prone to necrosis, tended to show lower CT attenuation.

Liver SUVmean was slightly elevated in lung cancer cases, possibly due to systemic inflammation or metabolic activation, particularly in advanced disease.

Demographic factors such as age and gender contributed minor improvements in calibration. For instance, younger male patients were more frequently classified as lymphoma, which is consistent with epidemiological patterns in diseases like Hodgkin lymphoma [19].

SHAP interaction plots highlighted: SUVmax × SUVpeak synergy: co-elevation pushed predictions toward lymphoma. T/L SUVmax × CT value: low-density lesions with high contrast further supported lymphoma classification. These observations confirm that the models captured not only individual predictive strength but also biologically plausible interactions among features, enhancing diagnostic transparency.

Clinical implications

The findings of this study suggest that integrating PET/CT-derived quantitative features with interpretable machine learning models can provide valuable support in the clinical management of adrenal lesions. These models can be embedded into structured radiology reports, offering automatic malignancy risk scores and highlighting key predictive features (e.g., low T/L SUVmax, high CT value), thus aiding clinicians in interpretation and reducing inter-reader variability.

The high sensitivity and interpretability of ensemble models—particularly Bagging and Random Forest—make them suitable for use as clinical decision support systems during multidisciplinary tumor boards. For instance, in cases with indeterminate imaging findings, these models could help prioritize patients for biopsy versus follow-up, improving triage efficiency.

Furthermore, the Decision Curve Analysis (DCA) confirms that these models offer net clinical benefit over a broad range of risk thresholds, reflecting real-world uncertainty in adrenal lesion management. By integrating DCA findings into clinical decision-making (e.g., thresholds for when to biopsy or observe), the models can be used to optimize individualized patient care, reduce unnecessary invasive procedures, and support evidence-based discussions at tumor boards. Accurate non-invasive characterization of adrenal lesions is not only diagnostically valuable but may also facilitate appropriate selection for minimally invasive treatments, such as image-guided thermal ablation, as discussed by Ierardi et al. [20].

In the future, such models could be incorporated into PACS or electronic health record systems as interactive modules, where clinicians can input imaging parameters and receive real-time malignancy risk predictions with SHAP-based explanations to support transparency and trust.

Limitations and future directions

Despite the promising results, this study has several limitations that warrant consideration and offer direction for future work. First, the sample size, particularly in the malignant subtype classification task, was relatively modest, which may limit the generalizability of the models and introduce statistical instability in minority class predictions. Second, the study was conducted using retrospective data from a single institution, and the absence of external validation raises concerns regarding potential overfitting and lack of applicability across diverse patient populations and imaging platforms. Furthermore, we acknowledge that differences in PET/CT scanner types, image acquisition protocols, and patient population characteristics across institutions could significantly impact model reproducibility. Variations in SUV measurement due to scanner calibration, reconstruction algorithms, and FDG uptake times may lead to inconsistencies in radiomic feature values. To address these challenges in future work, multicenter collaborations should be pursued. Standardizing imaging protocols and adopting harmonization techniques (e.g., ComBat) can help reduce inter-center variability. In addition, model fine-tuning or domain adaptation methods may enhance transferability to external datasets. Federated learning frameworks could also facilitate cross-institutional model training without requiring raw data sharing, ensuring both privacy and robustness. some clinical variables, such as prior treatment history, hormonal function status, or genomic alterations, were not included, which may have limited the model's ability to capture deeper biological heterogeneity. Although efforts were made to mitigate demographic imbalances, factors such as gender and age distribution could still contribute to hidden confounding. Future research should aim to validate these findings in large-scale, multicenter prospective cohorts and consider incorporating multimodal data—including laboratory

results, histopathology, and molecular features—to further enhance model robustness, interpretability, and clinical utility [21]. Moreover, real-time integration of such models into clinical workflows through PACS or EHR systems will be essential for practical deployment and adoption, ensuring that the benefits of AI-driven decision support are fully realized in everyday medical practice, As noted in recent reviews [9], many AI models for adrenal imaging lack interpretability and generalizability, which our study seeks to address through SHAP-based feature analysis and discussion of multicenter validation strategies.

Conclusion

In this study, we developed and validated interpretable machine learning models based on PET/CT-derived radiomic and CT-derived morphological features to classify adrenal lesions as benign or malignant, and to further distinguish between lung cancer metastases and lymphoma among malignant cases. The models showed promising classification performance in this retrospective cohort, with ensemble methods and neural networks demonstrating favorable discrimination ability. SHAP analysis contributed to model transparency by revealing clinically relevant feature contributions and interactions, highlighting the potential utility of SUV-based imaging markers and CT attenuation values in adrenal lesion assessment.

However, given the single-center and retrospective nature of the study, and the lack of external validation, these findings should be interpreted as preliminary. While our results support the feasibility of explainable machine learning in adrenal imaging, prospective validation in larger, multicenter cohorts is essential before clinical integration can be considered.

Abbreviations

AUC	Area Under the Receiver Operating Characteristic Curve
DCA	Decision Curve Analysis
SHAP	SHapley Addictive exPlanations
SVM	Support Vector Machine
ANN	Artificial Neural Network
ML	Machine Learning
t-SNE	t-distributed Stochastic Neighbor Embedding
LASSO	Least Absolute Shrinkage and Selection Operator

Acknowledgements

We would like to share our sincere gratitude to the staffs who are worked in Zhejiang Cancer Hospital for their contributions in data collection, maintenance, distribution and so on, and thanks to all the individuals who participated in this study.

Authors' contributions

Yun Wang: Data collection, Writing-original draft, Manuscript revision. Yuqi Su: Data collection, Methodology, Validation, Formal analysis, Writing-original draft. Jing Li, Deying Xie, Yuhang Cai, Chengyang Sun, Jingjing Zhang: Data collection, Methodology. Zhuolin Liu: Software. Ye Yuan: Statistical analysis, Writing -original draft. Jaesik Jeong and Heqing Yi: Funding acquisition, Writing-Review and Supervision. All authors reviewed the results and approved the final version of the manuscript.

Funding

This study was supported by the Medical Health Science and Technology Project of the Zhejiang Provincial Health Commission (2023KY068/2022PY043) and the Zhejiang Province Natural Science Foundation of China (Grant number: LTGY24H180013), and Supported by the BK 21 Four (Fostering Outstanding University for Research, NO. 5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

Data availability

The dataset and materials generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Zhejiang Cancer Hospital. The requirement for informed consent was waived due to the retrospective nature of the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Nuclear Medicine, Zhejiang Cancer Hospital, Hangzhou, Zhejiang 310022, China
[2]Postgraduate Training Base Alliance of Wenzhou Medical University, ZheJiang Cancer Hospital, Hangzhou, Zhejiang 310022, China
[3]Department of Nuclear Medicine, Aksu Prefecture First People's Hospital, Aksu, Xinjiang 843000, China
[4]School of Mental Health, Wenzhou Medical University, Wenzhou 325000, China
[5]Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Republic of Korea
[6]Key Laboratory of Head and Neck Cancer Translational Research of Zhejiang Province, Zhejiang Cancer Hospital, Hangzhou, Zhejiang 310022, China

Received: 20 August 2025 / Accepted: 24 October 2025
Published online: 07 November 2025

References

1. Boland GW, Blake MA, Holalkere NS, Hahn PF. PET/CT for the characterization of adrenal masses in patients with cancer: qualitative versus quantitative accuracy in 150 consecutive patients. AJR Am J Roentgenol. 2009;192(4):956–62. https://doi.org/10.2214/AJR.08.1431.
2. Rowe NE, Kumar R, Schieda N, Siddiqi F, McGregor T, McAlpine K, et al. Diagnosis, management, and follow-up of the incidentally discovered adrenal mass: CUA guideline endorsed by the AUA. J Urol. 2023;210(4):590–9. https://doi.org/10.1097/JU.0000000000003644.
3. Tüdös Z, Čtvrtlík F. Possible impact of CT histogram analysis in incidentally discovered adrenal masses. Abdom Radiol (New York). 2020;45(9):2937–8. https://doi.org/10.1007/s00261-020-02596-2.
4. Blake MA, Kalra MK, Maher MM, Sahani DV, Sweeney AT, Mueller PR, Hahn PF, Boland GW. Pheochromocytoma: an imaging chameleon. Radiographics: Rev Publication Radiological Soc North Am Inc. 2004;24(Suppl 1):S87–99. https://doi.org/10.1148/rg.24si045506.
5. Metser U, Miller E, Lerman H, Lievshitz G, Avital S, Even-Sapir E. 18F-FDG PET/CT in the evaluation of adrenal masses. J Nuclear Medicine: Official Publication Soc Nuclear Med. 2006;47(1):32–7.
6. Schöder H, Larson SM. Positron emission tomography for prostate, bladder, and renal cancer. Semin Nucl Med. 2004;34(4):274–92. https://doi.org/10.1053/j.semnuclmed.2004.06.004.
7. Kim SJ, Lee SW, Pak K, Kim IJ, Kim K. Diagnostic accuracy of 18F-FDG PET or PET/CT for the characterization of adrenal masses: a systematic review and

meta-analysis. Br J Radiol. 2018;91(1086):20170520. https://doi.org/10.1259/bjr.20170520.
8. Romanisio M, Daffara T, Pitino R, et al. [18F]FDG-PET/CT in adrenal lesions: diagnostic performance in different clinical settings. Endocrine. 2025;87:325–33. https://doi.org/10.1007/s12020-024-04042-5.
9. Ferro M, Tataru OS, Carrieri G, Busetto GM, Falagario UG, Maggi M, Crocetto F, Barone B, Del Giudice F, Marchioni M, Terracciano D, Lucarelli G, Ditonno P, Gherasim R, Todea-Moga C, Fallara G, Tozzi M, Cioffi A, Bianchi R, Digiacomo A, Rocco B. Artificial intelligence and radiomics applications in adrenal lesions: a systematic review. Ther Adv Urol. 2025;17:17562872251352553. https://doi.org/10.1177/17562872251352553.
10. El-Galaly TC, Villa D, Gormsen LC, Baech J, Lo A, Cheah CY. FDG-PET/CT in the management of lymphomas: current status and future directions. J Intern Med. 2018;284(4):358–76.
11. Song JH, Chaudhry FS, Mayo-Smith WW. The incidental adrenal mass on CT: prevalence of adrenal disease in 1,049 consecutive adrenal masses in patients with no known malignancy. AJR Am J Roentgenol. 2008;190(5):1163–8. https://doi.org/10.2214/AJR.07.2799.
12. Boland GW, Blake MA, Hahn PF, Mayo-Smith WW. Incidental adrenal lesions: principles, techniques, and algorithms for imaging characterization. Radiology. 2008;249(3):756–75. https://doi.org/10.1148/radiol.2493070976.
13. Yun M, Kim W, Alnafisi N, Lacorte L, Jang S, Alavi A. 18F-fdg pet in characterizing adrenal lesions detected on CT or MRI. Journal of nuclear medicine: official publication Soc Nuclear Med. 2001;42(12):1795–9.
14. Blake MA, Kalra MK, Sweeney AT, Lucey BC, Maher MM, Sahani DV, et al. Distinguishing benign from malignant adrenal masses: multi–detector row CT protocol with 10-minute delay. Radiology. 2006;238(2):578–85.
15. Fassnacht M, Arlt W, Barzon L, et al.European Society of Endocrinology (ESE) and European Network for the Study of Adrenal Tumors (ENSAT) guidelines: management of adrenal incidentalomas, update 2023.Lancet Diabetes Endocrinol. 2023;11(6):469–482.https://doi.org/10.1016/S2213-8587(23)00113-6.
16. Feliciani G, Serra F, Menghi E, Ferroni F, Sarnelli A, Feo C, Zatelli MC, Ambrosio MR, Giganti M, Carnevale A. Radiomics in the characterization of lipid-poor adrenal adenomas at unenhanced CT: time to look beyond usual density metrics. Eur Radiol. 2024;34(1):422–432. https://doi.org/10.1007/s00330-023-10090-8.
17. Yuan, H., Kang, B., Sun, K. et al. CT-based radiomics nomogram for differentiation of adrenal hyperplasia from lipid-poor adenoma: an exploratory study. BMC Med Imaging. 2023;23:4. https://doi.org/10.1186/s12880-022-00951-x.
18. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009;324(5930):1029–33. https://doi.org/10.1126/science.1160809.
19. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74. https://doi.org/10.1016/j.cell.2011.02.013.
20. Ierardi AM, Patella F, Petrillo M, Angileri SA, Pellegrino C, Pignatti AR, Carrafiello G. Minimally invasive image-guided treatment of adrenal lesions: current status. Gland Surg. 2020;9(4):1249–57. https://doi.org/10.21037/gs.2020.03.22.
21. Groussin L, Bonardel G, Silvéra S, Tissier F, Coste J, Abiven G, et al. 18F-fluorodeoxyglucose positron emission tomography for the diagnosis of adrenocortical tumors: a prospective study in 77 operated patients. J Clin Endocrinol Metab. 2009;94(5):1713–22.

Publisher's Note