Multimedia Tools and Applications
https://doi.org/10.1007/s11042-025-20737-x

A machine learning approach for determining solitary pulmonary nodule malignancy in patients undergoing PET/CT examination

Ioannis D. Apostolopoulos[1] · Nikolaos D. Papathanasiou[2] · Dimitris J. Apostolopoulos[2] · Elpiniki I. Papageorgiou[1] · Nikolaos Papandrianos[1]

Received: 16 April 2024 / Revised: 14 February 2025 / Accepted: 5 March 2025

Abstract
Non-small cell lung cancer (NSCLC) constitutes about 85% of all lung cancers and is a leading cause of cancer-related deaths globally. Within the spectrum of lung cancer, Solitary Pulmonary Nodules (SPNs) have become a focal point of research due to their significant implications for mortality. Estimating the malignancy of SPNs is usually performed by medicine experts considering multiple screening methods (Computerised Tomography and Positron Emission Tomography). Machine Learning may simplify this time-consuming procedure and highlight potential human errors. The study presents an efficient methodology for the classification of Solitary Pulmonary Nodule malignancy, addressing critical limitations in existing SPN classification approaches by emphasizing on the synergistic use of PET and CT image features, clinical data, and validation against biopsy-confirmed data. Patient data recorded from a hybrid PET/CT scanner at the University Hospital of Patras, Greece were examined. Human readers annotated 360 SPNs, which were used to train and internally validate the proposed model. SPNs (96) with confirmed histopathological results were used as an external test set. The classification methodology relied on an XGBoost model, which uses manually-extracted SPN features from both imaging modalities. Feature selection was performed to reduce the dimensions of the data and identify the most important predictors. The proposed method exhibited an agreement of 97% with the human readers on the training and validation set. On the external set, the accuracy was 86% (81% sensitivity, 100% specificity). The SUVmax predictor exhibited lower scores (92% agreement on the training and validation set, 85% accuracy on the external set). The model was superior to the advised SUVmax threshold of 2.5 (85% accuracy, 80% sensitivity, 100% specificity).

Keywords Machine Learning · Lung Cancer · Solitary Pulmonary Nodule · Extreme Gradient Boosting

✉ Ioannis D. Apostolopoulos
ece7216@upnet.gr

[1] Department of Energy Systems, University of Thessaly, Gaiopolis Campus, Larisa 41500, Greece

[2] Department of Nuclear Medicine, University Hospital of Patras, Rio 26504, Greece

1 Introduction

Lung cancer diagnosis and treatment pose a significant global health challenge, given the high mortality rates associated with it. Non-small cell Lung Cancer (NSCLC) accounts for approximately 85% of all lung cancers, representing a significant cause of cancer-related deaths worldwide [1]. Among the diverse manifestations of lung cancer, Solitary Pulmonary Nodules (SPNs) have garnered considerable attention due to their impact on mortality [2]. The urgency in comprehending and effectively managing these lesions arises from their potential malignancy, with the prognosis of lung cancer closely tied to the stage of detection [3].

Identifying and classifying SPNs are critical, as delayed detection often results in advanced disease stages and compromises treatment efficacy. Early recognition of SPNs is, therefore, pivotal in improving patient outcomes. Imaging modalities, notably Computerised Tomography (CT) and Positron Emission Tomography (PET) play essential roles in SPN diagnosis [4, 5]. CT scans provide detailed anatomical information, facilitating the characterisation of nodule morphology, size, shape, and density [6]. On the other hand, ^{18}F-fluro-deoxy-glucose (FDG) PET offers metabolic insights, aiding in distinguishing between benign and malignant nodules based on glucose metabolism [7]. In hybrid PET/CT systems, the complementary use of CT and PET enhances diagnostic accuracy by offering a comprehensive understanding of both the anatomical and functional aspects of SPNs.

Despite advancements in imaging technologies, human error in interpreting SPNs remains a challenge, contributing to diagnostic uncertainty [5]. Additionally, the conventional diagnostic process is often time-consuming, leading to delays in initiating treatment. The intricate nature of SPNs, coupled with the necessity for meticulous imaging and clinical data analysis, poses challenges for timely and accurate decision-making.

Machine learning (ML) is a promising approach to augment SPN classification [8]. ML models can process extensive imaging and clinical data, identify patterns, and derive predictive insights [9]. The latter advancement holds the potential to overcome the limitations of traditional diagnostic methods, offering a more efficient and accurate means of classifying SPNs and expediting the initiation of appropriate treatments [8, 10–12].

This study proposes the ML classifier titled XGBoost [13] for predicting the malignancy of SPNs using manually-extracted SPN image features from CT and PET modalities. Feature selection was performed using the XGBoost's inherent feature importance grading operation. Two datasets were used in this particular research. The first contained SPNs labelled by human experts. The second includes SPNs with surgically-confirmed diagnoses as to their malignancy. The algorithm was trained and initially evaluated using the first dataset. Hence, the agreement with the human readers was assessed. The trained algorithm was then used to predict the class of the second dataset, thereby evaluating its diagnostic efficiency.

The study contributes to the field in the following ways:

- The study presents an efficient methodology to handle multi-modal SPN characteristics for automatic classification of SPN malignancy
- Feature selection and feature importance analysis are performed to identify the most important predictors for SPN malignancy identification
- The proposed methodology outperforms the SUVmax predictor, which is the commonly advised threshold for FDG uptake (SUVmax), and exhibits great agreement with the human readers, as well as adequate diagnostic efficiency

The rest of the paper is organised as follows: Sect. 2 presents the related work in the field of ML-based SPN classification, along with existing limitations. The study's methodology including patient data, data preprocessing, and the ML algorithm are presented in Sect. 3. Section 4 presents the results of the study and discussions on the results are held in Sect. 5. Concluding remarks are given in Sect. 6.

2 Related work

In their 2015 study, Chen et al. [14] utilised the k-nearest neighbours classifier as a baseline comparison algorithm for benign–malignant classification in CT images, along with deep belief network and deep Convolutional Neural Network (CNN). The k-nearest neighbours classifier employed geometric descriptors, precisely scale-invariant feature transform (SIFT) and local binary pattern (LBP) to profile quantitative features in the two-dimensional region of interest. Following this, dimension reduction schemes were implemented to reduce the feature size while preserving discriminative power. Subsequently, the reduced feature set was inputted into the k-nearest neighbours classifier for the classification task. The evaluation, conducted on a subset of the LIDC-IDRI [15] dataset featuring nodules with diameters larger than 3 mm (i.e., 2545 nodules from 1010 cases), indicated that the k-nearest neighbours classifier achieved a sensitivity of 75.6% and a specificity of 66.8%.

Examining 48 patients, another research [16] evaluated the efficacy of combining radiomic features from 18F-FDG PET/CT with ML to differentiate between benign and malignant SPNs. The models demonstrated satisfactory performance, with Area Under the Curve (AUC) values approximately equal to 0.81. Notably, the deep learning model exhibited enhanced sensitivity (88% compared to 83%). The classical learning model showcased heightened specificity (86% compared to 79%) in contrast to conventional evaluation. In summary, integrating ML with 18F-FDG PET/CT texture features complements traditional assessments, leading to improved differentiation of lung nodules.

The authors of [17] aimed to develop an automatic analysis method using ML for diagnosing early-stage lung cancer based on PET/CT data. In a retrospective cohort study involving 187 cases of NSCLC and 190 benign pulmonary nodules, a diagnosis model was trained using twelve PET and CT features. The model identified standardised uptake value (SUVmax) as a crucial biochemical factor for early-stage lung cancer. The PET/CT diagnosis model exhibited a sensitivity and AUC of 86.5% and 0.89, respectively. Validation in cohorts of 462 and 229 cases demonstrated consistently high sensitivity and AUC. The proposed model significantly enhanced clinical discrimination for SPNs at the early stage. The study suggests that machine learning-based analysis of PET/CT scans can automate and improve the accuracy and positive predictive value (PPV) of early-stage SPNs, with the potential for optimisation into a less biased PET/CT automatic diagnosis system.

In [18], the authors investigated the efficacy of a ML radiomics model in detecting small-cell lung cancer (SCLC) on CT scans, specifically among nodules at least 1 cm in size. The research analysed CT scans from a single institution using four classification models, dividing them into SCLC and other categories. The ML models, including support vector classifier, random forest, XGBoost, and logistic regression, generated radiomic models with promising performance. The AUC ranged from 0.77 to 0.88 for different models and scan groups. Specific radiomic features were identified as predictive of SCLC in non-contrast and contrast-enhanced scans. Six radiomic features were found to overlap between the two groups.

Koo et al. [19] assessed the performance of multiparametric MRI radiomic features, either alone or combined with standard-of-care methods, in classifying pulmonary nodules. The impact of segmentation variability on feature reproducibility and reliability was also examined. Radiomic features from 74 pulmonary nodules were compared with histopathology and conventional quantitative imaging values. Eleven diffusion-weighted features effectively distinguished malignant from benign nodules, outperforming traditional imaging parameters. Multiparametric features are associated with the CT Hounsfield unit and SUVmax. ML models, particularly XGBoost, demonstrated strong performance in nodule classification. The study identified MRI radiomic features that could enhance the non-invasive and radiation-free assessment of pulmonary nodules, potentially establishing thoracic MRI as an alternative to standard practices.

Lin et al. [20] developed a comprehensive model integrating deep learning (DL), radiomics, and clinical data for classifying lung nodules into benign or malignant categories, as well as different pathological subtypes and Lung Imaging Reporting and Data System (Lung-RADS) scores. The model utilised a stacked ensemble approach with an AutoGluon-Tabular classifier and was trained on three datasets, including a public dataset (LUNA16 [21]) and two private datasets (LNOP and LNHE). The proposed model achieved high accuracy in classifying lung nodules into benign or malignant categories (92.8%), different pathological subtypes (F1-score of 75.5%), and Lung-RADS scores (F1-score of 80.4%).

While significant progress has been made, several limitations persist in the current body of related work. One major limitation in current research is the non-synergistic use of PET and CT images for SPN classification. PET provides functional information about metabolic activity, while CT offers anatomical details. Many studies have focused on either PET or CT images in isolation, neglecting the potential synergy between the two modalities. The failure to integrate these complementary sources of information may result in suboptimal classification performance and limit the comprehensive understanding of SPN characteristics.

Current approaches often focus on image-related features without incorporating relevant clinical data, such as patient demographics, smoking history, or symptoms. The lack of a comprehensive combination of image-related and clinical features hinders the development of holistic models for SPN classification. Integrating clinical information with imaging data has the potential to enhance the accuracy and reliability of classification models, providing a more robust foundation for clinical decision-making.

Interpretability is critical to developing trust in ML models, especially in medical applications [22–24]. Many existing models lack transparency, making understanding the reasoning behind classification decisions challenging. The absence of interpretability and feature explanation poses a significant barrier to the widespread clinical adoption of SPN classification models. Clinicians require insights into the features contributing to a classification decision to make informed and confident decisions.

While various studies have proposed SPN classification models, a significant limitation is the scarcity of evaluations based on biopsy or follow-up data. Validation using ground truth information from biopsy results or long-term follow-up is essential for assessing these models' real-world applicability and generalizability. Without robust validation of clinical outcomes, the reliability of SPN classification models remains uncertain, limiting their potential impact on actual patient care.

3 Methods

The overall study's approach is showcased in Fig. 1. The PET and CT images acquired by a PET/CT scanner (DISCOVERY iQ3 sl16, GE Healthcare) are employed for each patient case. Physicians analysed the resulting scans and collected auxiliary patient information. For each SPN case, additional features were extracted and integrated into the dataset, such as the SPN's diameter, type, margins, and SUVmax uptake. The latter is derived from the PET scan. An XGBoost classifier model is developed to classify SPNs into benign and malignant classes using the extracted features after a feature selection process, which is performed using the inherent feature importance functions of the XGBoost classifier. Due to the absence of follow-up or surgery verification of the findings, the model is trained on a partition of the dataset containing labels assigned by human experts. Hence, the model learns to exhibit agreement with the human readers. The trained model is then deployed to predict the malignancy of follow-up and biopsy-verified SPNs (Fig. 2).

3.1 Patient data

Over a span of five years (2018–2022), more than 800 PET/CT scans with confirmed lung nodules were meticulously reviewed to identify potential participants. Data were extracted from the University Hospital of Patras. Scans lacking SPNs or featuring nodules greater than 3 cm or less than 0.6 cm in diameter were excluded, resulting in 456 qualifying PET/CT scans. The medical staff analysed the resulting scans and extracted the corresponding SPN features (diameter, type, margins, and SUVmax uptake). The human readers also characterised SPN malignancy through (a) biopsy results, (b) patient follow-up, or (c) a verdict based on a combination of image and clinical factors (age, tobacco smoking or

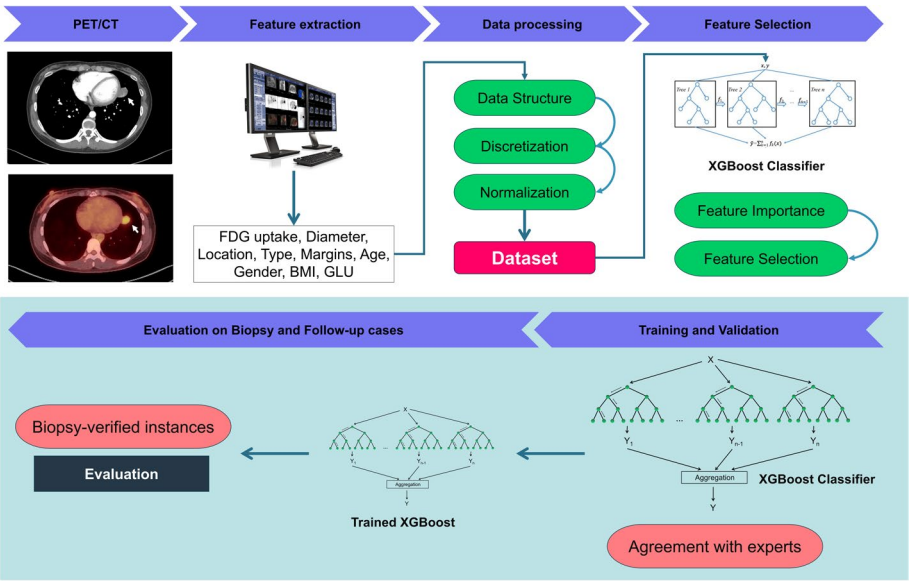

Fig. 1 Experiment overview

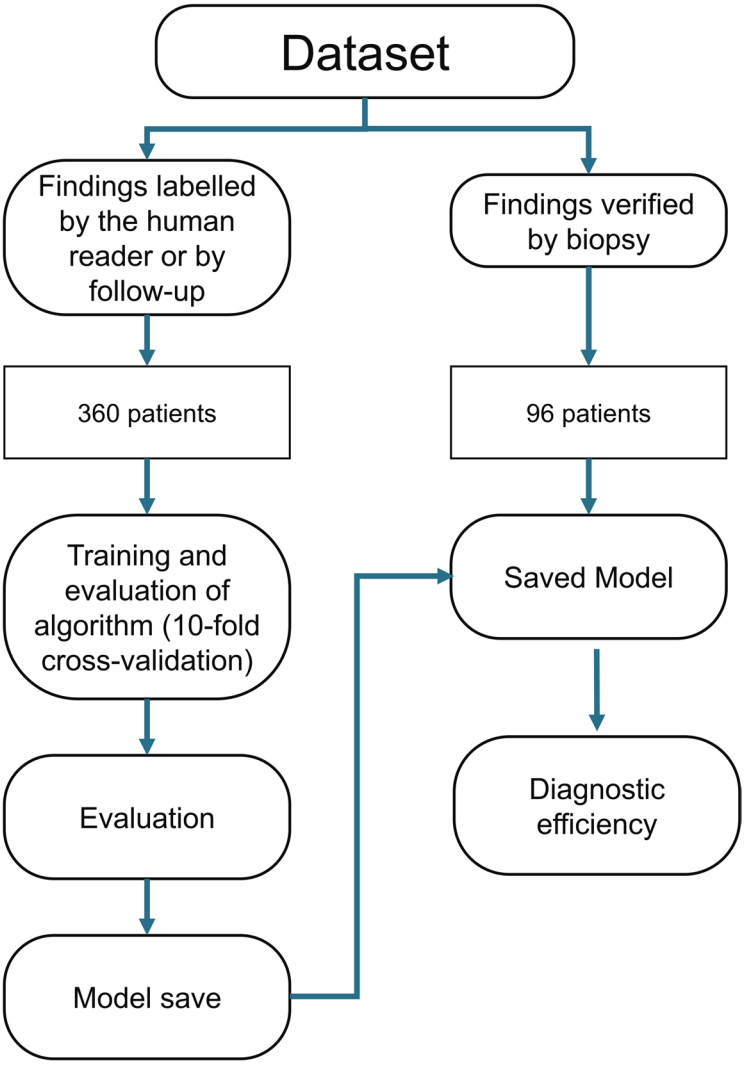

Fig. 2 Dataset split for model validation and evaluation

exposure to secondhand smoke, occupational hazards, previous chronic lung diseases). The nature of the survey waived the necessity for obtaining informed consent from patients.

The participants exhibited a mean age of 66 years, with 69% being male and 31% female. Benign SPNs constituted 49% of identifications, while malignant SPNs accounted for 51%. Fine-needle biopsy results or histopathological confirmation post-surgery were available in 13% of cases. Data from this subgroup was utilised for ablation studies and control groups. Approximately 9% of SPN labels were assigned based on patient follow-up and re-examination. Notably, 78% of labels were assigned by experts relying on their extensive 20 + years of experience. Table 1 showcases the dataset's characteristics.

Stringent measures were implemented for data collection, encompassing PET, CT, and clinical information. Sensitive DICOM data were promptly removed post-image collection,

Table 1 Clinical characteristics of the NSCLC dataset

Clinical characteristics	Frequency
No of participants	456
Age (mean ± sd)	66 ± 8
Sex (male/female)	69% male / 31% female
Benign SPNs (histopathological)	27
Benign SPNs (Human Expert or Follow-up)	195
Total Benign SPNs	222 (48.6%)
Malignant SPNs (histopathological)	69
Malignant SPNs (Human Expert or Follow-up)	165
Total Malignant SPNs	234 (51.4%)
Total SPNs with histopathological confirmation	96
Total SPNs characterised by the human expert or by follow–up	360

and clinical data utilised anonymous identification numbers instead of patient names. This ethical approach aligns with the principles of the Declaration of Helsinki, ensuring participant privacy and upholding the highest standards in medical research.

The dataset consists of 8 features, as explained in Table 2. The features are numeric or categorical and cover a wide range of image features the medical staff considers for the diagnosis.

3.2 Data processing

We split the data into two sub-datasets. The first contained SPNs without histopathological confirmation of their malignancy (the labels have been assigned by the human reader by

Table 2 Details of the NSCLC dataset's features

Feature Name	Type of feature	Type of values	Potential values
Gender	Demographic	Male / Female	M, F
Age	Demographic	Numeric	
SUVmax	FDG Uptake	Numeric	
Diameter	SPN Feature	Numeric	0.6 to 3 cm
Location	SPN Feature	Categorical	Left Lower Lobe (LLL) Lingula Middle Right Upper Lobe (RUL) Right Lower Lobe (RLR)
Type	SPN Feature	Categorical	Solid Semi-solid Ground-class
Margins	SPN Feature	Categorical	Well-defined Lobulated Spiculated Ill-defined

performing patient follow-up). The second dataset contained SPNs with a histopathological confirmation of their malignancy. The distribution of the SPNs is detailed in Table 3.

A comprehensive set of preprocessing measures was instituted in the preparatory phase of tabular data for the classification task to augment the dataset's appropriateness for subsequent ML analyses. The reader shall note that no images were used for model development. Instead, SPN features (e.g. diameter) were manually extracted from the corresponding PET and CT scans by the medical experts.

Primarily, the discretisation of potential feature values was executed to mitigate inconsistencies inherent in their representation. Doctors assigned the potential values for each categorical attribute (Table 2).

Furthermore, entries were methodically partitioned based on discerning labels and demarcating categories encompassing biopsy outcomes, patient follow-up, and human reader verdicts. Strategic preparation for feature selection transpired through extracting novel features from those bearing categorical values; a salient instance is the "type" feature with its tripartite categorisation, which was differentially delineated into discrete features. Notably, the presence of missing values was deliberately not rectified during the preprocessing stage, as modern ML algorithms inherently possess adaptive mechanisms to autonomously manage such occurrences.

In addition to preprocessing steps we performed data normalization, which ensures that all features have a consistent scale. Specifically, we used the min–max scaling, which transforms each feature x_i to a x_i' within a specified range, typically [0, 1]. This transformation is mathematically expressed as:

$$x_i\prime = \frac{x_i - \min(x)}{\max(x) - mix(x)}$$

This normalization technique preserves the relative relationships between data points while ensuring that each feature contributes equally to the analysis, regardless of its original scale.

3.3 Machine learning-based feature selection

In the pursuit of optimal model performance and interpretability in the context of SPN classification, feature selection through ML was investigated. The entire dataset was provided for analysis, and an XGBoost classifier [13] was deployed to discern the intrinsic importance of each feature. The latter classifier is analytically explained in the following sections. A comprehensive feature importance ranking was generated after the classifier's evaluation, illuminating the relative significance of individual attributes in influencing

Table 3 Datasets of the study

Dataset name	Description	No. of SPNs	Benign SPNs	Malignant SPNs
Train dataset	Contains SPNs labelled by the human readers or by patient follow-up	360	195	165
Biopsy dataset	Contains SPNs with histopathological confirmation of their malignancy	96	27	69

the classification outcomes. These results served as a compass for selecting features that yielded substantial discriminatory power. Table 4 presents the importance of each feature.

3.4 Machine learning classification model based on extreme Gradient Boosting

XGBoost, which stands for eXtreme Gradient Boosting, is a powerful and efficient ML algorithm that has gained immense popularity for its performance in various data science competitions. Developed by Tianqi Chen [13], XGBoost implements gradient-boosted decision trees designed for speed and performance. It excels in handling structured data, making it a popular choice for classification and regression tasks. XGBoost belongs to the family of ensemble learning methods [25], where multiple weak learners are combined to form a robust and accurate predictive model. The key idea behind XGBoost is to build a series of decision trees sequentially, with each tree correcting the errors of the previous one. The ensemble of trees improves predictive accuracy and generalisation.

The foundational element in XGBoost is the decision tree, specifically a shallow tree known as a weak learner [26]. These trees are constructed iteratively by recursively splitting the dataset based on features, forming a binary tree structure. In XGBoost,

Table 4 The importance of each feature to the classification outcome based on the XGBoost classifier. Based on the feature importance, the selected attributes are 14 (from 25)

No	Feature	Importance (%)	Selected
1	AGE	2.4%	Yes
2	BMI	1.2%	Yes
3	GLU	0.0%	No
4	SUV	34.3%	Yes
5	DIAMETER	6.1%	Yes
5	Gender_F	1.2%	No
7	Gender_M	0.0%	No
8	LOCATION_Left_Lower_Lobe	3.1%	Yes
9	LOCATION_Left_Upper_Lobe	0.0%	No
10	LOCATION_Lingula	7.2%	Yes
11	LOCATION_MIDDLE	0.0%	No
12	LOCATION_Middle	0.0%	No
13	LOCATION_Right_Lower_Lobe	3.8%	Yes
14	LOCATION_Right_Upper_Lobe	2.4%	Yes
15	LOCATION_middle	0.0%	No
16	TYPE_Consolidated	2.7%	Yes
17	TYPE_Ground-class	0.0%	No
18	TYPE_Semi-solid	4.5%	Yes
19	TYPE_Solid	3.7%	Yes
20	TYPE_Speckled	0.0%	No
21	TYPE_calcified	0.0%	No
22	MARGINS_ill-defined	0.0%	No
23	MARGINS_lobulated	8.1%	Yes
24	MARGINS_spiculated	10.5%	Yes
25	MARGINS_well defined	4.9%	Yes

shallow trees are emphasised to prevent overfitting and enhance the model's generalisation capability.

The features considered, such as gender, age, BMI, SUVmax, diameter, and various anatomical and compositional indicators, serve as the cornerstone of the algorithm's analytical process. For example, the diameter feature, representing the size of a pulmonary nodule, holds significant importance and can heavily influence the classification outcome. XGBoost meticulously identifies patterns in the dataset, recognising, for instance, the crucial roles played by SUVmax (a metabolic indicator) and nodule type in differentiating between benign and malignant cases.

XGBoost employs a regularised objective function comprising a loss function, measuring model error, and a regularisation term, penalising complexity [27]. This objective function guides the training process by assessing model performance and discouraging overly complex models. The algorithm constructs trees sequentially, with each subsequent tree correcting errors made by preceding ones. This training process involves minimising the objective function by iteratively adding trees. The gradient of the objective function concerning the model's prediction is calculated, and a new tree is fitted to the negative gradient of the loss function.

To mitigate overfitting, XGBoost integrates regularisation terms into the objective function, encompassing both L1 (Lasso) and L2 (Ridge) regularisation [27]. These terms control the complexity of individual trees and the overall model.

The XGBoost model and the corresponding experiments were performed using Python 3.9 with the aid of the Scikit Learn library on a 64 GB-RAM machine with an Intel Core i7 processor.

4 Results

4.1 Agreement with the human reader

We present the results of the tenfold cross-validation performed on the training dataset. The particular dataset is labelled following the human reader's verdict. Hence, the accuracy metric reflects the agreement between the XGBoost model and the experts.

XGBoost shows a 97.22% agreement with the human reader, corresponding to a sensitivity of 96.36%, a specificity of 97.95%, and an AUC score of 0.9961.

If the FDG uptake (SUVmax) was used as the main predictor for the diagnosis using a threshold of 2.96, it would achieve an accuracy of 91.67% (lower than XGBoost). The threshold of 2.96 was the optimal value based on the particular dataset. Similarly, a threshold of 2.5, which is generally advisable by the guidelines ([4]), would yield an agreement of 90.56%. The results are also presented in Table 5.

Table 5 Evaluation metrics of the XGBoost classifier with reference to the human reader-assigned labels

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
XGBoost	97.22	96.36	97.95	0.9961
SUVmax = 2.5	90.56	89.09	91.79	0.9044
SUVmax = 2.96	91.67	88.48	94.36	0.9142

In addition to the notable accuracy demonstrated by the XGBoost model when compared to FDG uptake thresholds, further examination reveals significant performance indicators. The sensitivity, standing at 96.36%, underscores the model's proficiency in accurately identifying true positive cases, affirming its ability to recognise instances with genuine positive outcomes. Simultaneously, the specificity of 97.95% emphasises the model's effectiveness in correctly pinpointing true negative cases, reinforcing its capability to distinguish instances with authentic negative outcomes.

The AUC score of 0.9961 serves as a comprehensive gauge of the model's discriminative ability, highlighting its effectiveness in distinguishing between positive and negative instances across different threshold values. This substantial AUC score underscores the robustness of the XGBoost model in capturing the true positive rate in contrast to the false positive rate.

Figure 3 illustrates the ROC Curve and the Learning curve of the XGBoost classifier. The curves were extracted following the tenfold cross-validation on the study's training dataset. The ROC curve verifies the effectiveness of XGBoost and underlines the exceptional agreement with the human reader. The Learning curve is used to validate the importance of the training data amount during training. It is observed that an improvement follows the increase of training data in cross-validation scores. The latter ensures that the algorithm is not under-fitting our data.

In contrast, relying solely on FDG uptake (SUVmax), even with an optimised threshold of 2.96, yields a marginally lower accuracy of 91.67% compared to XGBoost. This difference underscores the enhanced predictive capability of the XGBoost algorithm over depending solely on FDG uptake. It suggests that integrating multiple features and their interactions, as captured by the XGBoost model, significantly contributes to superior diagnostic accuracy.

Moreover, the comparison of threshold values, particularly the model-derived 2.96, against the guideline-recommended 2.5, highlights the subtle nature of decision-making in diagnostic settings. While the model's optimised threshold outperforms the guideline-recommended value, the latter still demonstrates a commendable agreement of 90.56%. This observation underscores the importance of aligning model outputs with established clinical guidelines, providing a practical balance between algorithmic precision and medical interpretability.

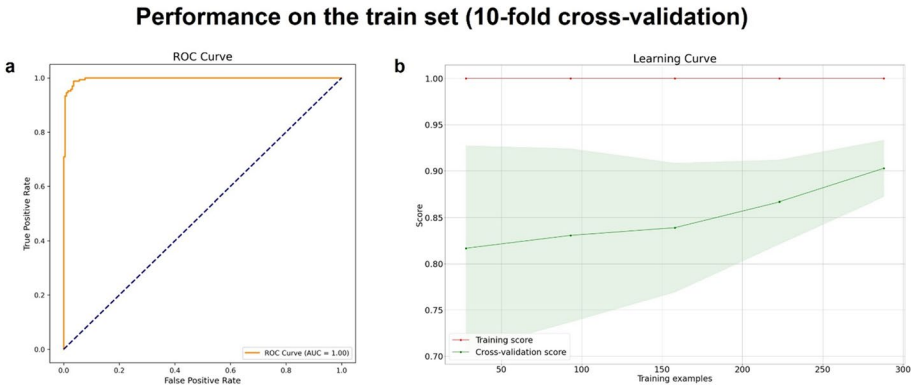

Fig. 3 ROC curve (**a**) and Learning Curve (**b**) of the XGBoost classifier extracted under a tenfold cross-validation procedure on the train set

Table 6 Evaluation metrics of the XGBoost classifier with reference biopsy-confirmed cases

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
XGBoost	86.46	81.16	100	0.8921
SUVmax = 2.5	85.42	79.71	100	0.8986
SUVmax = 2.96	78.12	69.57	100	0.8478

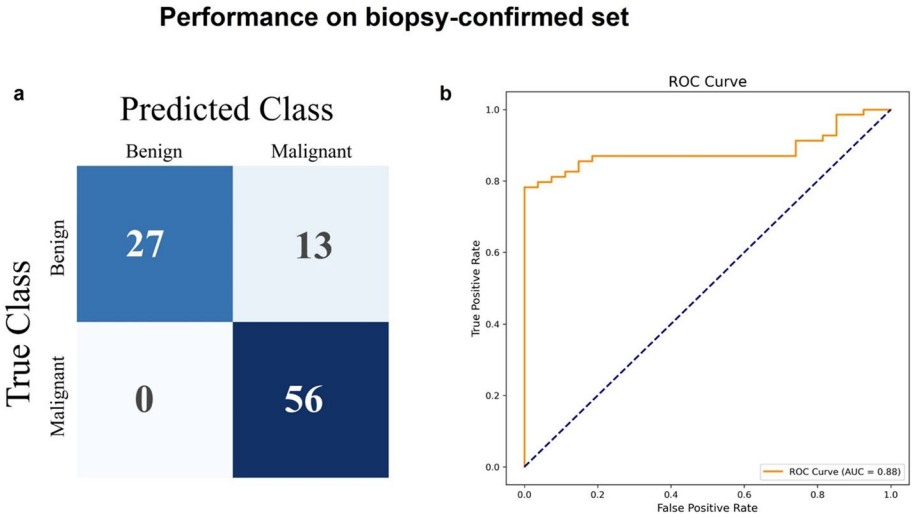

Fig. 4 Confusion Matrix (**a**) and ROC Curve (**b**) of the XGBoost classifier when evaluated on the Biopsy dataset

4.2 Performance on external surgically-verified patient cases

This section presents the results of the test phase. The Biopsy dataset was used to evaluate the XGBoost's performance. The model was trained using the entire training dataset, and its parameters were retained afterwards. Next, the Biopsy dataset was used to test the diagnostic efficiency of the algorithm. The overall results are presented in Table 6. The confusion matrix and the ROC curve are illustrated in Fig. 4. A performance summary is illustrated in Fig. 5.

The robust performance of XGBoost in predicting the malignancy of biopsy-confirmed SPNs is further emphasised by its ability to achieve an accuracy of 86.46%. This accuracy surpasses alternative methods, such as the SUVmax-based classification at the commonly advised threshold of 2.5, which yielded a slightly lower accuracy of 85.42%. The comparison underscores the efficacy of the XGBoost algorithm in distinguishing between malignant and non-malignant cases.

Moreover, XGBoost exhibits a commendable sensitivity of 81.16%, indicating its capability to correctly identify a significant portion of true positive cases among the biopsy-confirmed SPNs. In contrast, the SUVmax-based classification at the threshold of 2.5 lags behind in sensitivity, achieving a rate of 79.71%.

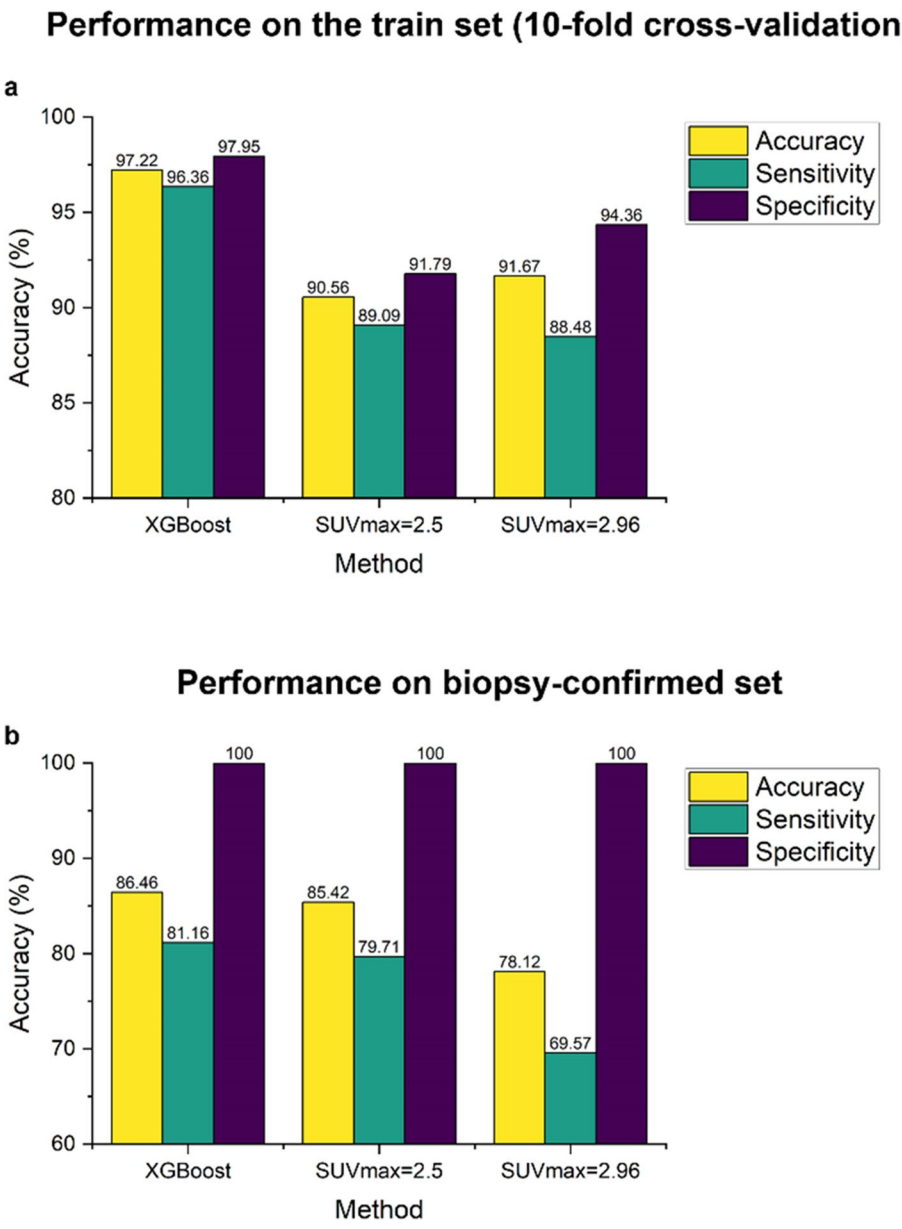

Fig. 5 Performance comparison in terms of accuracy, sensitivity, and specificity between the XGBoost classifier and the SUVmax-based classification on the Train and Biopsy datasets

The specificity of both XGBoost and the SUVmax-based classification at 2.5 is noteworthy, with both methods achieving 100%. The AUC score further underscores the superiority of XGBoost. With an AUC score of 0.8921, XGBoost demonstrates a robust ability to differentiate between malignant and benign cases, outperforming the SUVmax-based classification at 2.5, which achieved an AUC score of 0.8686. Fig. 3 presents the performance

comparison in terms of accuracy, sensitivity, and specificity between the XGBoost classifier and the SUVmax-based classification on the Train and Biopsy datasets.

The comparison with an alternative threshold for SUVmax (2.96), extracted from the Train dataset, reveals a significant performance drop. The accuracy diminishes to 78.12%, sensitivity decreases to 69.57%, and specificity remains 100%. The AUC score also experiences a decline, reaching 0.8478. This emphasises the importance of careful threshold selection and highlights the robustness of the XGBoost model in maintaining superior performance across different evaluation metrics.

The confusion matrix in Fig. 4 presents the total number of True Positives (56), True Negatives (27), False Positives (13), and False Negatives (0). The zero False Negative cases is an auspicious outcome in the case of SPN classification because it decrease the possibility of a malignant finding being classified as benign. The latter is accompanied by a perfect True Positive Rate (TPR) of 100%. However, the True Negative Rate (TNR) is concerning (67.5%).

4.3 Comparison with alternative algorithms

We compared XGBoost with alternative algorithms in terms of their performance using the Biopsy dataset as the reference point. The results are showcased in Table 7 and Fig. 6.

Notably, XGBoost emerges as the most robust model (Table 7), attaining an accuracy of 86.46%, sensitivity of 81.16%, perfect specificity (100%), and a substantial AUC value of 90.58. These results suggest that XGBoost excels in achieving an optimal balance between true positive and true negative rates, positioning it as a promising choice for the considered classification problem. Conversely, Naïve Bayes exhibits a distinct pattern, yielding perfect sensitivity but limited accuracy (42.86%) due to its challenges in correctly identifying true negatives. The study underscores the importance of carefully considering both sensitivity and specificity metrics and overall accuracy and AUC when selecting a suitable ML model for classification tasks.

In comparison to other models, Light Gradient Boosting Machine (LightGBM) and Support Vector Machine (SVM) also demonstrate commendable performances with accuracies of 75.93% and 74.07%, respectively. These findings contribute valuable insights into diverse machine learning algorithms' comparative strengths and weaknesses, facilitating

Table 7 Evaluation metrics with reference to the biopsy-confirmed cases (Biopsy dataset) for alternative ML classifiers

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
CatBoost	78.13	69.57	100	0.8980
Light Gradient Boosting Machine	85.42	79.71	100	0.9
Naïve Bayes	75.00	65.22	100	0.9791
Random Forest	82.29	75.36	100	0.8921
K-Nearest Neighbours	79.22	62.32	100	0.8722
AdaBoost	77.08	68.12	100	0.8798
Linear Discriminant Analysis	78.13	69.57	100	0.9533
Support Vector Machine	80.21	72.46	100	0.9946
XGBoost	86.46	81.16	100	0.8921

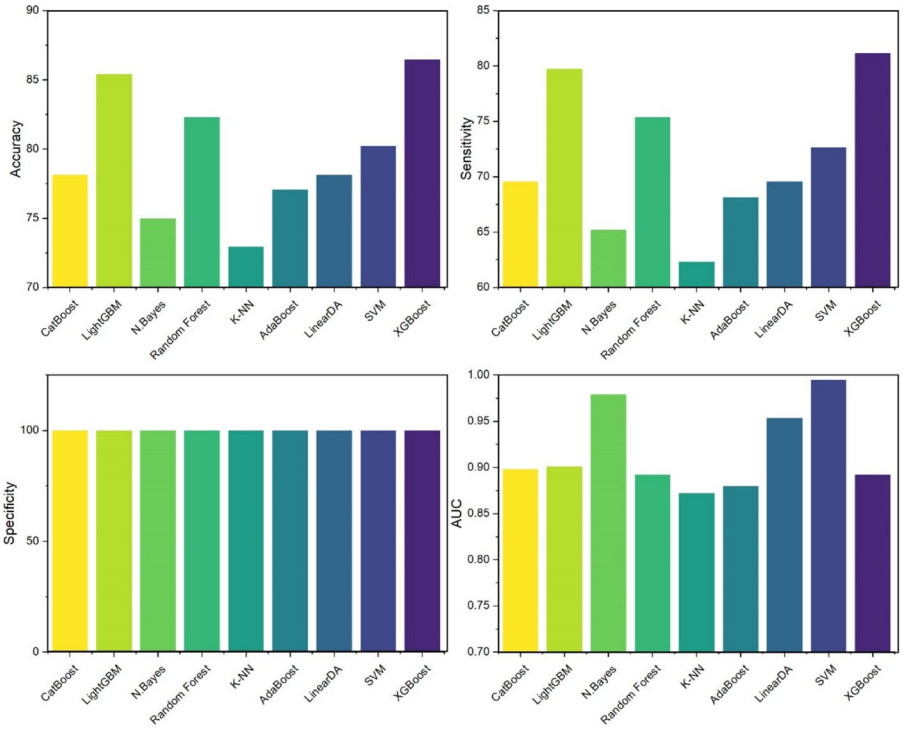

Fig. 6 Performance comparison between the XGBoost classifier and alternative methods based on the Biopsy dataset

informed decision-making in selecting the most apt model for specific classification scenarios.

5 Discussion

The current body of research on SPN classification exhibits notable advancements, yet several critical limitations persist. A significant drawback lies in the insufficient synergistic utilisation of PET and CT images. While PET reveals metabolic activity, and CT offers anatomical details, many studies focus on either modality in isolation, neglecting their potential synergy. This oversight may compromise classification accuracy and hinder a comprehensive understanding of SPN characteristics. Additionally, prevalent approaches often concentrate solely on image-related features, disregarding essential clinical data like patient demographics and symptoms. Integrating both sets of information holds promise for enhancing classification model accuracy and reliability. Moreover, interpretability remains a crucial concern, as many existing models lack transparency, impeding trust in medical applications. Clinicians require insights into classification decision rationale for informed decision-making. Another significant limitation is the scarcity of evaluations based on biopsy or follow-up data in proposed SPN classification models. The absence

of robust validation using ground truth information undermines these models' real-world applicability and generalizability, limiting their potential impact on actual patient care.

The study results indicate that the XGBoost model, developed for the classification of SPNs, exhibits a high level of accuracy and outperforms alternative methods, particularly in predicting the malignancy of SPNs with biopsy confirmation.

The selection of features in the developed XGBoost model for SPN classification reveals a thoughtful consideration of both clinical and imaging variables. Among the features, SUVmax, with a substantial importance of 34.3%, emerges as the most influential predictor. This aligns with the understanding that metabolic activity, as captured by PET imaging through SUVmax, plays a crucial role in distinguishing between benign and malignant SPNs. The high importance of SUVmax underscores its significance in accurately characterising nodules, reflecting the model's reliance on metabolic information for classification.

Additionally, the diameter of the SPN is assigned a notable importance of 6.1%. As indicated by diameter, SPN size is a well-established factor in assessing pulmonary nodules, with larger nodules often associated with a higher likelihood of malignancy. This inclusion reinforces the model's consideration of anatomical characteristics to enhance classification accuracy. Incorporating location features, such as the specific lobe and lung region, further refines the model's predictive capabilities. Notably, features like ill-defined margins and lobulated and spiculated margins, each with varying degrees of importance, highlight the model's sensitivity to the morphological characteristics of SPNs.

Beyond the imaging features, the model incorporates demographic information, such as age and gender, though with relatively lower importance. With an importance of 2.4%, age suggests a modest but discernible impact on the classification process. The consideration of gender-related variables, even with lower importance, acknowledges the potential influence of gender in SPN characterisation.

The overall performance metrics of the XGBoost model are noteworthy. With a remarkable agreement of 97.22% with human readers, the model demonstrates a sensitivity of 96.36%, a specificity of 97.95%, and an impressive area under the curve (AUC) score of 0.9961. These metrics underscore the model's ability to effectively distinguish between benign and malignant SPNs and mimic how human experts perform the diagnostic procedure.

Comparative analysis with a commonly advised threshold for FDG uptake (SUVmax) further emphasises the robustness of the XGBoost model.

In this study, the XGBoost algorithm's performance in predicting the malignancy of SPNs is thoroughly evaluated using a set of biopsy-confirmed SPNs. The model achieves an impressive overall accuracy of 86.46%, outperforming alternative methods like SUVmax-based classification at the commonly advised threshold of 2.5, which achieved a slightly lower accuracy of 85.42%. The XGBoost algorithm demonstrates robust sensitivity (81.16%) and perfect specificity (100%), highlighting its ability to correctly identify true positive cases and distinguish between malignant and non-malignant SPNs.

The confusion matrix further illustrates the model's promising outcomes, with zero false-negative cases, reducing the risk of classifying malignant findings as benign. However, attention is drawn to the TNR of 67.5%, indicating a potential area for improvement in the classification of non-malignant cases. Overall, these findings underscore the efficacy of the XGBoost algorithm in enhancing diagnostic efficiency for SPNs.

This study performed well when compared to other studies in the existing literature (Table 8), particularly in terms of the AUC value. The AUC was found to be 0.8921 in our research using the XGBoost method. This places our work firmly towards the top-end of the studies mentioned. It is also worth noting the evolution and enhancement of outcomes

Table 8 Comparison with existing literature

Study	Year	Method	Data size	ACC (%)	SEN (%)	SPE (%)	AUC
Hawkins et. al [28]	2014	DT, RB, SVM, NB	219	77.5	x	x	0.71
Chen et. al [14]	2015	DBN	2545	x	73.4	82.2	x
Kumar et. al [29]	2015	DT	157	75.01	83.35	x	x
Orozco et. al [30]	2015	SVM	61	82	90.9	73.91	0.805
Sergeeva et. al [31]	2016	NN, RF, KNN, NB, SVM	79	81.3	x	x	0.89
Gonçalves et. al [32]	2017	KNN, SVM	177	x	x	x	0.96
Sahu et. al [33]	2020	SVM	114	89	88	89	0.92
Shah et. al [18]	2021	SVM, RF, XGBoost, LR	103	x	x	X	0.88
Salihoğlu et. al [16]	2022	XGBoost, DNN	48	x	88	86	0.81
Wang et. al [17]	2023	LR	462	x	85.7	x	0.87
This study	2024	XGBoost	96	86.46	81.16	100	0.8921

DBN: Deep Belief Network; *DT*: Decision Tree; *RB*: Rule-Basd; *NB*: Naïve Bayes; *SVM*: Support Vector Machine; *NN*: Neural Network; *KNN*: K-nearest Neighbours; *RF*: Random Forest; *LR*: Logistic Regression

over the years. The earliest study, done by Hawkins et. al in 2014 [28] using multiple methodologies including Decision Tree (DT), Rule-Based (RB), Support Vector Machine (SVM), and Naive Bayes (NB), documented an AUC value of 0.71. In comparison, our study has shown significant improvement, indicating the development and refinement of algorithms and techniques over the decade. Therefore, it is evident that our study is an important contribution to the field and demonstrates the advances made in predictive modelling using XGBoost.

Moreover, it is evident that the AUC value of our study is consistent with the emerging trends in the research literature, confirming that the current approach aligns with the expected performance range. It is interesting to note that while our study consisted of a smaller data size (96) compared to some other studies it managed to maintain a strong performance in AUC. However, it is also important to consider that larger datasets may provide a more comprehensive and potentially more accurate understanding due to the demographic and individual variances they could cover. Therefore, in future contributions, increasing the data size could be considered to enhance the estimation efficiency and generalize the results more effectively.

A significant limitation of this study is the reliance on expert-labeled data for training the model, rather than biopsy-confirmed diagnoses. While the expertise of radiologists is valuable, human labeling introduces potential biases due to subjective interpretations, which can vary between experts and institutions. This variability could affect the consistency and accuracy of the labels used to train the model, ultimately influencing its predictions. In contrast, biopsy-confirmed cases provide a definitive diagnosis, offering a more reliable ground truth for model development. However, only a small subset of the test data in this study consisted of biopsy-confirmed cases, which limits the strength of the validation process. This discrepancy between training and testing with definitive clinical outcomes raises concerns about the model's real-world applicability, particularly in high-stakes clinical decision-making, where biopsy data should ideally be used as a benchmark.

Furthermore, the scarcity of data in this study, particularly when compared to other studies in the literature, presents another challenge. Some studies have been able to leverage large datasets with thousands of SPN cases, providing a more robust foundation for

model training and evaluation. In contrast, our dataset, though carefully curated, is relatively small, which limits the diversity of nodules and patient profiles that the model has been exposed to. This may hinder the model's ability to generalize across different patient populations and imaging conditions. Larger, more diverse datasets would not only improve the model's learning process but also reduce the potential for overfitting to a specific set of characteristics, ultimately increasing its generalizability and reliability in broader clinical settings. Therefore, addressing both the labeling limitations and the dataset size in future studies is crucial for improving the model's performance and clinical applicability.

6 Conclusions

This study presents an efficient methodology for the classification of SPN malignancy, leveraging the synergistic integration of PET- and CT-based SPN features coupled with clinical data. The application of feature selection and feature importance analysis successfully identified the most critical predictors for SPN malignancy, marking a significant stride forward in diagnostic studies of NSCLC. The XGBoost method demonstrated impressive concordance with human expert labels, achieving an agreement of 97% on the training and internal validation set, and 86% accuracy on an external set. Notably, our model exhibited superior performance to the commonly used SUVmax predictor, underscoring the vast potential of Machine Learning in cancer diagnosis and facilitating early and accurate interventions.

However, several limitations of our study must be acknowledged. The potential biases introduced by the reliance on labels assigned by human experts for model training are notable. Additionally, the lack of follow-up or surgical verification of the results leaves room for improvement in evaluating the model's real-world performance. Future work should seek to address these limitations by incorporating larger and more diverse datasets, along with rigorous validation against biopsy and follow-up data, to ensure the model's reliability and wider applicability in varied clinical settings.

Author contributions Conceptualization: I.D.A. and N.P.; Methodology: N.D.P., D.J.A.; Software: I.D.A., E.I.P.; Validation: D.J.A., N.P.; Writing – Original Draft: I.D.A., N.D.P.; Writing – Review & Editing: E.I.P., D.J.A.; Supervision: N.P.

Funding Open access funding provided by HEAL-Link Greece. The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers" (Project Number: 3656).

Data availability The datasets of the study are public can be communicated by the research team on a reasonable request.

Declarations

Consent statement All procedures in this study were in accordance with the Declaration of Helsinki. This study was approved on the 3rd of March 2022 by the ethical committee of the University General Hospital of Patras (Ethical & Research Committee of University Hospital of Pa-tras—protocol number 108/10–3–2022). The requirement to obtain informed consent was waived by the director of the diagnostic center due to its retrospective nature.

Competing interests No potential conflicts of interest were disclosed.

References

1. Torre LA, Siegel RL, Jemal A (2016) Lung cancer statistics. In: Ahmad A, Gadgeel S (eds) Lung Cancer and Personalized Medicine. Springer International Publishing, Cham, pp 1–19
2. Ko JP, Bagga B, Gozansky E, Moore WH (2022) Solitary pulmonary nodule evaluation: pearls and pitfalls. Semin Ultrasound CT MRI 43:230–245. https://doi.org/10.1053/j.sult.2022.01.006
3. Bade BC, Cruz CSD (2020) Lung cancer 2020: epidemiology, etiology, and prevention. Clin Chest Med 41:1–24
4. Cruickshank A, Stieler G, Ameer F (2019) Evaluation of the solitary pulmonary nodule. Intern Med J 49:306–315. https://doi.org/10.1111/imj.14219
5. Mazzone PJ, Lam L (2022) Evaluating the patient with a pulmonary nodule: a review. JAMA 327:264. https://doi.org/10.1001/jama.2021.24287
6. Greenspan BS (2017) Role of PET/CT for precision medicine in lung cancer: perspective of the society of nuclear medicine and molecular imaging. Transl Lung Cancer Res 6:617–620. https://doi.org/10.21037/tlcr.2017.09.01
7. Alongi P, Zanoni L, Incerti E et al (2015) 18F-FDG PET/CT for early postradiotherapy assessment in solitary bone plasmacytoma. Clin Nucl Med 40:e399–e404. https://doi.org/10.1097/RLU.0000000000000819
8. Apostolopoulos ID, Pintelas EG, Livieris IE et al (2021) Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques. Med Biol Eng Comput 59:1299–1310. https://doi.org/10.1007/s11517-021-02378-y
9. De Bruijne M (2016) Machine learning approaches in medical image analysis: From detection to diagnosis. Elsevier
10. Causey JL, Zhang J, Ma S et al (2018) Highly accurate model for prediction of lung nodule malignancy with CT scans. Sci Rep 8:9286
11. Apostolopoulos ID, Papathanasiou ND (2021) Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning. Biocybern Biomed Eng 41:1243–1257. https://doi.org/10.1016/j.bbe.2021.08.006
12. Apostolopoulos ID, Apostolopoulos DJ, Panayiotakis GS (2022) Solitary Pulmonary Nodule malignancy classification utilising 3D features and semi-supervised Deep Learning. In: 2022 13th international conference on information, intelligence, systems & applications (IISA). IEEE, Corfu, Greece, pp 1–6
13. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp 785–794
14. Hua K-L, Hsu C-H, Chen Y-J et al (2015) Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Ther 8:2015–2022. https://doi.org/10.2147/OTT.S80733
15. Armato SG, McLennan G, Bidaut L et al (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT Scans: the LIDC/IDRI thoracic CT database of lung nodules. Med Phys 38:915–931. https://doi.org/10.1118/1.3528204
16. Salihoğlu YS, Uslu Erdemir R, Aydur Püren B et al (2022) Diagnostic performance of machine learning models based on 18F-FDG PET/CT radiomic features in the classification of solitary pulmonary nodules. Mol Imaging Radionucl Ther 31:82–88. https://doi.org/10.4274/mirt.galenos.2021.43760
17. Wang H, Li Y, Han J et al (2023) A machine learning-based PET/CT model for automatic diagnosis of early-stage lung cancer. Front Oncol 13:1192908. https://doi.org/10.3389/fonc.2023.1192908

18. Shah RP, Selby HM, Mukherjee P et al (2021) Machine learning radiomics model for early identification of small-cell lung cancer on computed tomography scans. JCO Clin Cancer Inform 746–757:746. https://doi.org/10.1200/CCI.21.00021
19. Koo CW, Kline TL, Yoon JH et al (2022) Magnetic resonance radiomic feature performance in pulmonary nodule classification and impact of segmentation variability on radiomics. Br J Radiol 95:20220230. https://doi.org/10.1259/bjr.20220230
20. Lin C-Y, Guo S-M, Lien J-JJ et al (2023) Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. Radiol Med (Torino). https://doi.org/10.1007/s11547-023-01730-6
21. Setio AAA, Traverso A, De Bel T et al (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 42:1–13. https://doi.org/10.1016/j.media.2017.06.015
22. Dwivedi R, Dave D, Naik H et al (2023) Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv 55:1–33. https://doi.org/10.1145/3561048
23. Lipton ZC (2018) The mythos of model interpretability. Queue 16:31–57
24. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp 1135–1144
25. Dietterich TG (2000) Ensemble methods in machine learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1
26. Kotsiantis SB (2013) Decision trees: a recent overview. Artif Intell Rev 39:261–283. https://doi.org/10.1007/s10462-011-9272-4
27. Cortes C, Mohri M, Rostamizadeh A (2012) L2 regularization for learning kernels. ArXiv Prepr ArXiv12052653
28. Hawkins SH, Korecki JN, Balagurunathan Y et al (2014) Predicting outcomes of nonsmall cell lung cancer using CT image features. IEEE Access 2:1418–1426. https://doi.org/10.1109/ACCESS.2014.2373335
29. Kumar D, Wong A, Clausi DA (2015) Lung nodule classification using deep features in CT images. In: 2015 12th conference on computer and robot vision. IEEE, Halifax, NS, Canada, pp 133–138
30. Madero Orozco H, Vergara Villegas OO, Cruz Sánchez VG et al (2015) Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng OnLine 14:9. https://doi.org/10.1186/s12938-015-0003-y
31. Sergeeva M, Ryabchikov I, Glaznev M, Gusarova N (2016) Classification of pulmonary nodules on computed tomography scans. Evaluation of the effectiveness of application of textural features extracted using wavelet transform of image. In: 2016 18th conference of open innovations association and seminar on information security and protection of information technology (FRUCT-ISPIT). IEEE, St-Petersburg, Russia, pp 291–299
32. Gonçalves L, Novo J, Cunha A, Campilho A (2017) Evaluation of the degree of malignancy of lung nodules in computed tomography images. In: International Conference on Computer Vision Theory and Applications, February 27, 2017, Porto, Portugal, vol 7. SCITEPRESS, pp 74–80. https://www.scitepress.org/Papers/2017/61162/
33. Sahu SP, Londhe ND, Verma S et al (2021) Improved pulmonary lung nodules risk stratification in computed tomography images by fusing shape and texture features in a machine-learning paradigm. Int J Imaging Syst Technol 31:1503–1518. https://doi.org/10.1002/ima.22539