DE GRUYTER Journal of Intelligent Systems 2026; 35(1): 20250064

Research Article

Agorastos-Dimitrios Samaras, Serafeim Moustakidis, Ioannis D. Apostolopoulos, Elpiniki Papageorgiou, Nikolaos D. Papathanasiou, Dimitris J. Apostolopoulos and Nikolaos Papandrianos*

Explainable YOLOv8 model for solitary pulmonary nodules classification using positron emission tomography and computed tomography scans

https://doi.org/10.1515/jisys-2025-0064
Received March 23, 2025; accepted August 27, 2025; published online February 17, 2026

Abstract: This study addresses the diagnostic challenges associated with Non-Small Cell Lung Cancer (NSCLC), the most prevalent form of lung cancer often diagnosed at advanced stages. It aims to develop a computer-aided classification model exclusively utilizing medical images from Positron Emission Tomography (PET) and Computed Tomography (CT) scans. The model identifies benign/malignant Solitary Pulmonary Nodules (SPN) related to NSCLC. A dataset comprising of 456 patients, in total, was curated, featuring 48.68 % benign cases. To achieve its objective, four well-established Deep Learning (DL) algorithms were employed. The dataset was split into three different groups of images, each used for a particular task; training, testing and validation of the model. Notably, the study extends beyond predictive accuracy by delving into the prediction process of the best-performing model, thereby enhancing the explainability of the typically opaque Artificial Intelligence (AI) models. This explainability aspect aims to foster trust and confidence in the model's outcomes, allowing users to comprehend the decision-making process. The results indicate that the YOLOv8 algorithm emerged as the most accurate classification model, achieving a maximum accuracy of 91.3 % and a maximum True Positive Rate (TPR) of 93.62 %. This study's importance lies in underscoring the potential of DL approaches in improving NSCLC diagnosis while providing a transparent and understandable classification mechanism. It offers a novel way of explaining classification results from YOLOv8 model and it demonstrates both the effectiveness of DL-assisted predictions in characterizing SPNs and the added value of interpretability, thereby offering a holistic perspective on the model's capabilities.

Keywords: deep learning; non-small cell lung cancer; computer-aided diagnosis; explainable artificial intelligence

***Corresponding author: Nikolaos Papandrianos**, Department of Energy Systems, University of Thessaly, Gaiopolis Campus, 41334, Larisa, Greece, E-mail: npapandrianos@uth.gr
Agorastos-Dimitrios Samaras and Elpiniki Papageorgiou, Department of Energy Systems, University of Thessaly, Gaiopolis Campus, 41334, Larisa, Greece, E-mail: agsamaras@uth.gr (A.-D. Samaras), elpinikipapageorgiou@uth.gr (E. Papageorgiou). https://orcid.org/0000-0003-2498-9661 (E. Papageorgiou)
Serafeim Moustakidis, Department of Energy Systems, University of Thessaly, Gaiopolis Campus, 41334, Larisa, Greece; and AIDEAS OÜ, 10116, Tallinn, Estonia, E-mail: s.moustakidis@aideas.eu
Ioannis D. Apostolopoulos, Department of Medical Physics, School of Medicine, University of Patras, 26504, Patras, Greece, E-mail: ece7216@ac.upatras.gr
Nikolaos D. Papathanasiou and Dimitris J. Apostolopoulos, Department of Nuclear Medicine, University Hospital of Patras, 26504, Patras, Greece, E-mail: nikopapathan@upatras.gr (N.D. Papathanasiou), dimap@upatras.gr (D.J. Apostolopoulos)

1 Introduction

Non-small cell lung cancer (NSCLC) is a prevailing health concern, constituting approximately 85 % of lung cancer cases [1]. The multifaceted diagnosis and treatment of NSCLC relies on factors such as disease stage, patient health, and specific molecular biomarkers [2]. Standard protocols include surgical resection for early-stage NSCLC and a combination of chemotherapy, radiation therapy, and targeted approaches for advanced stages [3]. The advent of immunotherapy, exemplified by drugs like pembrolizumab and nivolumab, has ushered in a paradigm shift in NSCLC treatment, although response variability necessitates ongoing research into predictive biomarkers [4], [5]. The imperative for a personalized approach underscores the management of NSCLC, with continuous research endeavors aimed at refining patient outcomes.

In parallel, recent strides in AI herald transformative prospects for healthcare, promising better precision in diagnoses, disease prognostication, and overall improvements in patient outcomes. A noteworthy application of AI integration lies in customized Medical Decision Support Systems (MDSS), leveraging sophisticated Machine Learning (ML) algorithms to analyze patient data, thereby furnishing personalized diagnoses and tailored treatment recommendations [6]. This computational prowess extends beyond diagnostics, permeating into drug development domains where AI predicts molecular behaviors and identifies potential therapeutic targets. As AI's penetration deepens into healthcare systems, the potential for heightened efficiency, effectiveness, and patient experiences becomes increasingly palpable.

Given the ubiquity of NSCLC, ML and DL algorithms have gained considerable traction in diagnostic, prognostic, and treatment prediction domains [7], [8]. Notable studies, such as that by Guberina et al. [9], have employed multilayer perceptron neural networks and Random Forest (RF) models on image data to predict malignant lymph nodes, achieving commendable accuracy of 87.3 % and sensitivity of 85.7 %. Another study by Yu et al. [10] utilized RF on CT images to predict pathologic stages in NSCLC patients with an accuracy of 86 %. A more recent study by Janik et al. [11] employed RF to predict recurrence probability in early-stage NSCLC patients, yielding an accuracy of 76 %. Additionally, Yu et al. [12] proposed an Extreme Gradient Boosting (XGBoost) classifier for lung cancer prediction, further diversifying the ML methodologies. Furthermore, in Ref. [13], researchers employed a Decision Tree (DT) prediction model to calculate the probability of mediastinal lymph node metastasis in lung cancer, achieving sensitivity and specificity of 40 and 96.23 %, respectively. Lastly, Kureshi et al. [14] employed four distinct classifiers to predict treatment outcomes from EGFR tyrosine kinase inhibitors, achieving a classification accuracy of 76 %. Moreover, in Ref. [15], 11 ML and DL models are evaluated and a maximum accuracy of 84.1 % is reached. Collectively, these studies underscore the burgeoning potential of ML algorithms in refining NSCLC diagnostics and treatment, promising enhanced accuracy and patient outcomes. Table 1 depicts a summary of recent work specifically on NSCLC classification. Further discussion and comparison with the findings of the current work is provided in the Discussion Section.

It is immediately apparent that there is not such a plethora of studies utilizing DL methods to classify NSCLC cases. Importantly, existing studies predominantly leverage datasets rich in image data, primarily PET, SPECT, and CT scans. Indeed, these modalities, especially PET/CT, have proven to be highly sensitive for NSCLC detection. Nonetheless, PET image sets are less represented as there is an oversaturation of CT image sets and clinical data in the body of related research. Additionally, there is a but complete lack of transparency and explainability regarding the proposed models. The current work touches base on both these shortcomings. More specifically, the study achieves peak accuracy by adopting contemporary ML/DL models such as ResNet [24], MobileNet [25], and YOLOv8 [26], underscoring its commitment to cutting-edge methodologies. The explainability aspect of the current study, particularly for YOLOv8, the best-performing model, remains unique, as related work exploring its black-box mechanics is seemingly scarce or none, whatsoever. This incorporation of advanced models enhances the study's predictive capabilities and aligns with the forefront of technological advancements in machine learning and deep learning. The study introduces an essential layer of transparency to the often opaque mechanics of the most accurate classification model. By offering insights into the decision-making process, this transparency contributes to a deeper understanding of the model's functionality. This aspect is particularly significant in addressing the common black-box nature of AI models, enhancing the interpretability and

Table 1: Related work on NSCLC classification.

References	Dataset size	Data types	Best results of study
Cai et al. [16]	LADC: 452 SQCLC: 359	Clinical data	Accuracy: 86.54 %
Rabbani et al. [17]	227 patients	Clinical, demographical and image data	AUC: 88.2 %
Prelaj et al. [18]	200 patients	Clinical and laboratorial data	Accuracy: 75.6 % F1: 72.2 % AUC: 82 %
Han et al. [15]	ADC: 867 SCC: 552	Image data	Accuracy: 84.1 % AUC: 90.3 %
Apostolopoulos et al. [19]	112 SPNs	Image data	Accuracy: 94 % Sensitivity: 92.7 % Specificity: 95.2 %
Apostolopoulos et al. [20]	LIDC-IDRI: 1113 SPNs PET/CT: 172 SPNs	Image data	LIDC-IDRI accuracy: 92.07 % LIDC-IDRI sensitivity: 89.35 % LIDC-IDRI specificity: 94.8 % PET/CT accuracy: 84.46 % PET/CT sensitivity: 91.12 % PET/CT specificity: 77.75 %
Zhang et al. [21]	55 patients	Image data	Sensitivity: 85.2 % Specificity: 82.1 %
Laros et al. [22]	148 patients	Image and laboratorial data	Accuracy: 92 % AUC: 87 %
Bashir et al. [23]	106 patients	Image data	AUC: 82 %

reliability of the study's outcomes. The study's contributions encompass three main facets. Firstly, it achieves highly accurate results in SPN malignancy classification, without the need for an excessively large image set. Secondly, it conducts a comprehensive evaluation of four well-established DL algorithms for lung cancer prediction, demonstrating a commitment to thorough analysis and benchmarking against leading models. Lastly, the study delves into the complexities of a black-box AI prediction model, providing valuable insights for its seamless integration into a computer-aided decision-making system for NSCLC diagnosis and classification.

The subsequent sections of the paper are organized as follows. Section 2 provides analytical details on the patient dataset, the ML models harnessed for classification, and the evaluation process, inclusive of a proposed explainability analysis. Section 3 presents the experimental results, elucidating performance metrics and highlighting noteworthy findings. Section 4 conducts a comprehensive discussion, delineating both the strengths and weaknesses of the proposed model, and touches on the mechanics of the best-performing algorithm. The study concludes in Section 5, summarizing key findings and proposing potential avenues for future research.

2 Materials and methods

The methodology of this research was executed in a Linux environment, specifically on a system equipped with a 14-core i5 CPU, 32GB DDR4 RAM, RTX-3060 GPU and running Ubuntu 20.04LTS. Python served as the primary programming language for the project, complemented by various machine learning-specific libraries such as sklearn, shap, and genetic_selection. The MDSS application was created using Streamlit, a web-like Graphical User Interface (GUI) library for back-end code running on Python.

2.1 Dataset – patient population

The current study comprised 456 participants, among whom 234 received a diagnosis of malignant SPN from medical experts, constituting 51.32 % of the cohort. The remaining participants were diagnosed with benign cases, resulting in a highly balanced dataset. In terms of demographics, a majority of participants (70.37 %) were male, spanning ages from 46 to 89 years, with an average age of 67 years. Body Mass Index (BMI) ranged from 14.36 to 40.88, covering categories from underweight to obesity class III, based on World Health Organization (WHO) BMI classifications [27].

This investigation employed a hybrid PET/CT scanner, specifically the "GE Healthcare IQ3" model. The scanner features three detector rings with a 15 cm field of view, reconstructing 35 axial images at 4.25 mm intervals. Three-dimensional (3D) volumes representing the entire body were acquired through various bed positions while the patient assumed a supine position. For CT image acquisition, 16 detectors of 3.75 mm size, 1.5 pitch, and 5 mm collimation were employed, with exposure factors set at 120–140 kVp and 80 mAs. The CT images were acquired in free-breathing mode. SPN malignancy characterization was conducted by physicians or radiologists through a cross-examination of the produced PET scan, CT1.25 scan, and derived markers such as the SUVmax score. Scans lacking Solitary Pulmonary Nodules were excluded, and studies featuring lung nodules exceeding a diameter of 3 cm were disregarded.

Patient data were systematically recorded at the Clinical Sector of the Department of Nuclear Medicine, University Hospital of Patras, spanning from February 16, 2019, to February 28, 2022. Ethical approval for data collection was granted by the Ethical and Research Committee of the University General Hospital of Patras (protocol number 108/10-3-2022). Given the retrospective nature of the study, the need for informed consent from participants was waived. Rigorous measures were implemented to ensure anonymity throughout all data-related processes. The execution of all procedures in this study adhered strictly to the principles outlined in the Declaration of Helsinki.

2.2 Medical images

PET scans involve injecting a small amount of a radioactive tracer, typically fluorodeoxyglucose (FDG) in the case of NSCLC, into the patient's bloodstream. As cells absorb the tracer, emitted positrons interact with electrons, producing gamma rays. Detectors capture these gamma rays, enabling the creation of three-dimensional images that highlight areas of increased metabolic activity. PET scans are valuable in diagnosing and staging various diseases, including cancer [28].

CT scans employ X-rays to create detailed cross-sectional images of the body. The patient lies on a table that moves through the CT scanner, which rotates to capture X-ray images from various angles. A computer processes these images to generate detailed cross-sectional slices, providing a comprehensive view of internal structures. CT scans are crucial for diagnosing conditions such as tumors, fractures, and internal injuries, offering high-resolution anatomical information [29].

The current work relies solely on the medical scans from the hybrid PET/CT scanner. A recent study by our research team [30] proposed an explainable classification model, utilizing solely clinical data for the same NSCLC cases. Therefore, the current work expands the particular effort by implementing an explainable classification model for NSCLC medical scans.

More specifically, two main sets of images were created; PET and CT. Each particular dataset was then divided into three separate subsets/subdirectories, test, train and valid. The analogy by which this split was applied was:
- train set: 70 %
- validation set: 20 %
- test set: 10 %

As per common practice in the research community on image classification, the train subset (the largest of the three) was used in tandem with the test set for the training and recalibration of each model considered in the scope of the current work. Lastly, the validation set was used in order to validate the trained classification models and extract the evaluation metrics.

The original images were preprocessed for DL. As a first step, an extra layer was added to each image, pinpointing its specific class/label (Benign/Malignant). Then the images were resized accordingly as per the needs of each particular application (training, validation or explanation). The final step to produce the best model for each DL algorithm was then to fine-tune the parameters of every algorithm based on every run's results so as to

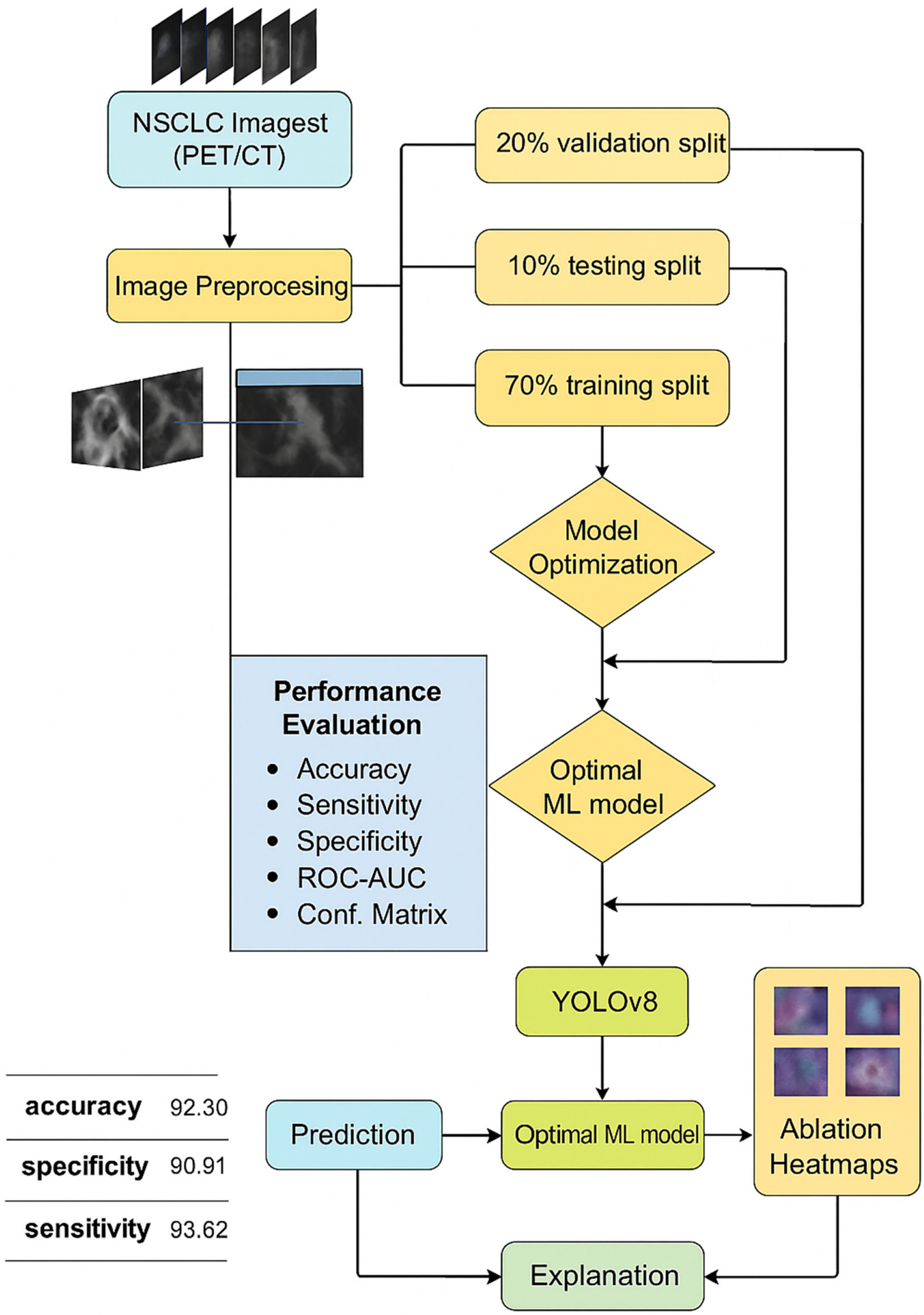


Figure 1: Study's high-level design. (*Source*: created by the author).

achieve maximum classification/prediction accuracy. The high-level design of the present study is also depicted on Figure 1.

2.3 DL models

In the scope of the current study, four different well-known DL architectures were tested; YOLOv8 (You Only Look Once v8), VGG-16 (Visual Geometry Group 16), ResNet (Residual Network) and MobileNet. This selection of DL models has been actively employed on image classification works with proven efficiency/accuracy [31]–[38].

YOLO [26] is a real-time object detection model, that tackles computer vision with its singular pass approach, predicting bounding boxes and class probabilities simultaneously. Developed by Joseph Redmon and Santosh Divvala, YOLO emerged in 2016, introducing an innovative single-pass architecture for real-time object detection. YOLO's architecture underwent enhancements in subsequent versions. Renowned for speed, YOLO's unified architecture excels in diverse applications, including medical image analysis. In healthcare, its efficiency proves invaluable for rapid identification and localization of abnormalities in diagnostic imaging. Its real-time capabilities make it particularly apt for scenarios requiring immediate decision support. YOLO's versatility and speed make it a prominent choice in the evolving landscape of medical image interpretation, offering swift and accurate detection of anomalies for improved patient care.

VGG-16 [39] has at its base a deep convolutional neural network. It originated from the Visual Geometry Group at the University of Oxford, led by Andrew Zisserman and Karen Simonyan. Introduced in 2014, VGG-16 gained prominence for its simplicity and effectiveness. With its 16-layer architecture (hence its naming), VGG-16 captures intricate features, setting a benchmark in computer vision tasks. In medical imaging, its depth proves advantageous for discerning subtle nuances in pathology. The network's ability to hierarchically process visual information aligns with the complex nature of medical images, offering detailed insights for precise diagnostics. VGG-16's applications in medical image analysis showcase robust performance, providing a foundation for accurate classification and interpretation.

ResNet [24], based on residual learning, was proposed by Kaiming He et al. in 2015. It addresses challenges in training deep neural networks. With skip connections, ResNet enables the successful training of exceptionally deep networks, vital for extracting complex hierarchical features. As far as medical imaging application is concerned, ResNet's depth allows the model to pinpoint intricate patterns, a feat that is particularly crucial for tasks such as tumor detection. Its contribution to mitigating vanishing gradient issues enhances the network's ability to learn and adapt, positioning ResNet as a valid option for medical image analysis applications.

The last DL model to be evaluated was MobileNet [25]. Developed by Google researchers Andrew G. Howard et al., MobileNet debuted in 2017. It is designed for computational efficiency, and can play a pivotal role in resource-constrained environments. Its architecture, featuring depth-wise separable convolutions, balances high accuracy with low computational demands. In medical applications, MobileNet's lightweight design can prove advantageous for real-time image processing on mobile devices. Its adaptability to diverse settings, without compromising performance, positions MobileNet as a practical choice in scenarios where computational resources are limited. MobileNet's efficiency aligns seamlessly with the evolving landscape of portable medical imaging devices, contributing to accessible and swift diagnostic solutions.

2.4 Results evaluation

The study conducted a thorough evaluation of each model's outcomes, employing five widely adopted metric scores. These metrics included accuracy, sensitivity, specificity, Receiver Operating Characteristics (ROC) curve, and the confusion matrix. The analysis involved examining the correlation among True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) instances through these metrics, providing a comprehensive understanding of each model's performance. This systematic evaluation facilitated a nuanced assessment of the models' strengths and weaknesses, ultimately guiding the selection of the optimal model.

Traditionally, accuracy stands as the predominant metric for AI model assessment, quantifying the proportion of correctly classified instances among the total instances. Its calculation follows the mathematical formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Essential for evaluating diagnostic tests or decision-making tools, sensitivity and specificity serve as pivotal measures. Sensitivity gauges the precision of true positive identification, while specificity assesses the precision of true negative identification. Particularly pertinent in binary classification scenarios, such as the benign/malignant classification in this study, these metrics illuminate a model's capability to accurately identify individuals within the first class (benign SPN) and those outside it (specificity). The mathematical expressions for calculating these metrics are articulated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

In appraising the effectiveness of a binary classification model, the ROC curve provides a valuable visual representation, depicting the true positive rate (sensitivity) against the false positive rate (1-specificity) across diverse threshold values. This graphical tool aids in selecting an optimal threshold, striking a balance between sensitivity and specificity. Additionally, the area under the ROC curve (AUC) serves as a widely utilized metric, quantifying the overall performance of the classification model. A perfect model attains an AUC of 1, while randomness yields an AUC of 0.5.

Concluding the evaluation toolkit, the confusion matrix proves instrumental. This tabular representation succinctly summarizes the model's predicted and actual outcomes. In binary classification scenarios, the matrix assumes a 2×2 format, delineating TP, TN, FP and FN predictions. Scrutinizing these matrix values yields insights into the model's accuracy, precision, and recall.

2.5 Results' interpretation

Within the methodology of this study, a pivotal concluding step involves interpreting the best performing prediction model. This holds paramount significance for all black-box prediction models, particularly in medical applications. The imperative for explainability stems from the potential ramifications of opacity, as it can erode trust among both patients and medical personnel [40]–[42]. In the medical domain, adherence to regulations mandating explanations for automated decision-making systems is imperative [43]. Hence, for AI to be a valuable tool in healthcare, it necessitates a transparent prediction process. Furnishing insights into the decision-making mechanisms of AI systems is indispensable for bolstering trustworthiness and fostering acceptance within the medical practice.

Within the contemporary body of scientific work, there is variety of tools employed to explain ML prediction patterns. Afore mentioned tools include analysis systems, such as SHapley Additive exPlanations (SHAP) [44] and Local Interpretable Model-Agnostic Explanations (LIME) [45], but also, especially when explaining images, Gradient-weighted Class Activation Mapping (Grad-CAM) [46]. These different forms of explainability focus on the explanation of predictions with unique formulae. Notwithstanding, as the optimal model of the current study is based on DL algorithm YOLOv8, this introduces some extra complications. Due to the complexity of the underlying classification model, the more standard approaches listed above were not feasible. Namely, they could either not be applied on the current prediction model due their unique analytical approach, or they needed tampering of the internal neural network powering the optimal model. Ergo, we opted for ablation analysis, in order to shed light unto the prediction process of the model.

This innovative approach is introduced for enhancing explainability in a black-box ML classification model through an ablation heat-map. This methodology dissects the model's decision-making process by systematically deactivating/obscuring parts of the input features/images and observing the resultant impact on predictions. The generated heat-map visually highlights the pivotal features influencing the model's classifications, offering interpretable insights into its decision rationale. By overlaying the image on top of the original image, we create a user-friendly means of showcasing the most crucial points of interest of the original image. This ablation heat-map approach aims to not only elucidates the black-box model's inner workings but also contributes to increased transparency, fostering trust and understanding crucial for the effective integration of AI systems in scientific and medical applications. This is a radical approach utilizing an established analysis pattern to offer explainability for a state-of-the-art DL classification model (YOLOv8). Also and most importantly this approach is case-agnostic, meaning that it can be applied for the specific model on any image classification scenario, not exclusively for PET/CT image sets. Hopefully, this works will help to further exploit the specific model and expand the explainable AI (XAI) research on related fields.

As this novel ablation heat-map approach required some form of validation, comparative analyses against established XAI techniques (SHAP and LIME) were conducted, using the same PET/CT images. While ablation analysis was uniquely compatible with the primary YOLOv8 architecture, SHAP and LIME explanations were implemented on the ResNet18 model (secondary classifier with 85 % accuracy) to enable direct methodological comparison.

3 Results

YOLOv8, VGG-16, ResNet and MobileNet are four established DL models and networks that were tested in the context of the current work. All algorithms were individually trained on both image sets; PET and CT. Furthermore, some fine-tuning with hyperparameter selection was conducted on each experiment/run, so as to maximize the efficiency of every model. Eventually, the optimal model was found to be the one based on YOLOv8 architecture and this model's prediction performance was explained with the use of ablation heat-maps, in order to produce a XAI system. The above mentioned results are showcased in the following sections.

3.1 Results using the PET image set

Table 2 illustrates the accuracy, sensitivity, and specificity outcomes derived from applying the models on the validation image set of PET medical scans. The YOLOv8-derived model demonstrated superior performance, achieving the highest scores for accuracy and specificity. On the other hand, the highest sensitivity was reached by the resNet-based classification model. This underscores the efficacy of the YOLOv8 model in accurately discerning between benign and malignant SPNs, establishing it as the optimal selection for the classification task at hand. Additionally, Figure 2 depicts the confusion matrices of all tested architectures.

3.2 Results using the CT image set

Similarly, Table 3 illustrates the accuracy, sensitivity, and specificity outcomes derived from applying the models on the validation image set of CT medical scans. Once again, the YOLOv8-derived model demonstrated superior performance, achieving the highest scores for accuracy and specificity. This time, however, it also achieved the highest sensitivity score across all architectures, making it the optimal prediction architecture for CT image sets, as well. Additionally, Figure 3 depicts the confusion matrices of all tested architectures.

3.3 Interpretation of best performing model

As depicted on Tables 2 and 3, the optimal prediction system for both PET and CT images was the one utilizing the YOLOv8 algorithm. As described in the Methods section, through the process of ablation, explainable heat-maps were generated for each case. The more magenta colored sections indicate a point of interest/focus. These informative heat-maps highlight the points of interest and focuses of the classification mechanism. Ergo, they

Table 2: Metrics scores achieved on the PET image set.

	YOLOv8	VGG-16	ResNet	MobileNet
Accuracy	**89.00 %**	82.41 %	84.62 %	82.42 %
Specificity	**97.73 %**	79.55 %	79.55 %	79.55 %
Sensitivity	80.85 %	85.11 %	**89.36 %**	85.11 %

Bold values are the higher results for each algorithm.

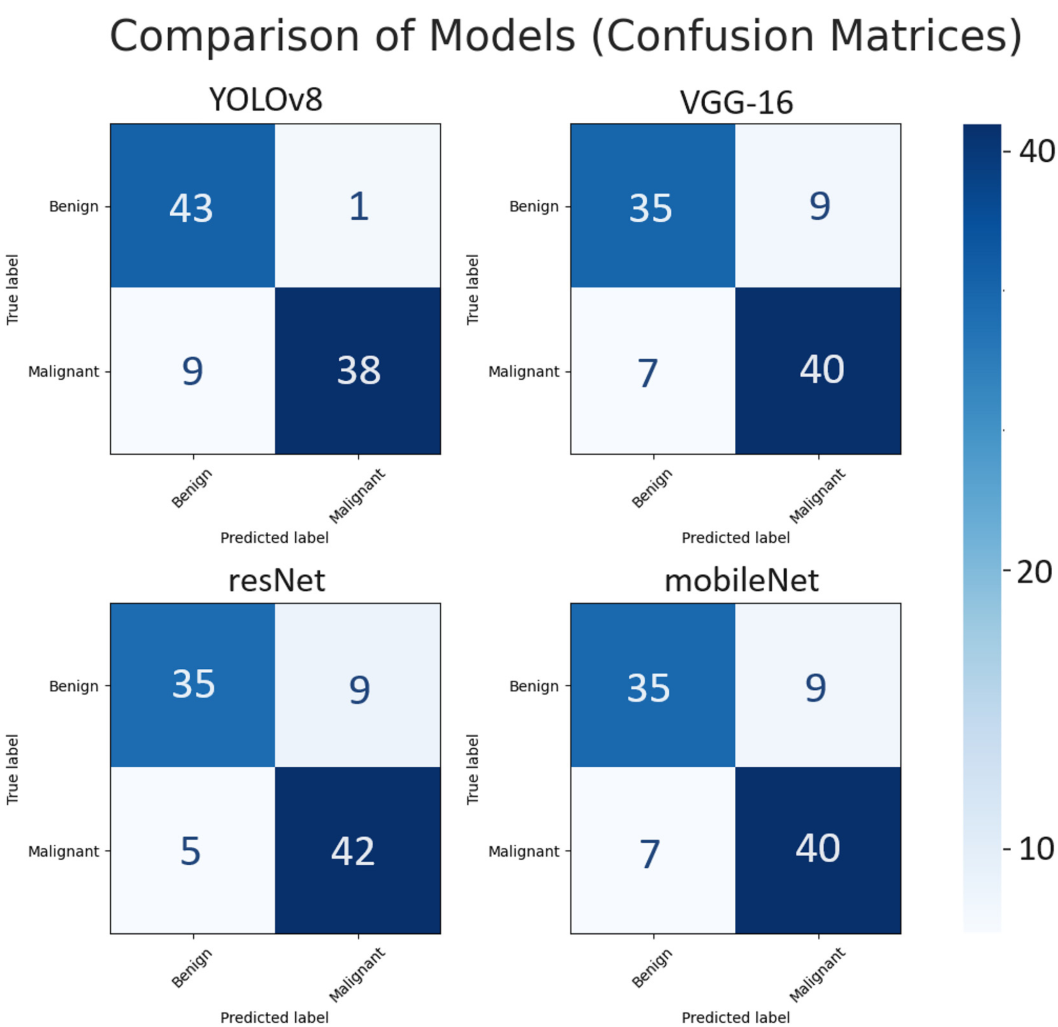


Figure 2: Comparison of models' performance; indicative PET confusion matrices. (*Source*: created by the author).

convey the exact points of the medical scan's images that were the most crucial in deciding whether the case is a malignant one or a benign. Following, some characteristic heat-map overlays are presented for each test case.

3.4 Comparative explainability analysis

The current study introduces among others a new XAI technique based on ablation heat-maps and as such existing, established XAI techniques were employed to validate the results. Namely, explanation through SHAP

Table 3: Metrics scores achieved on the PET image set.

	YOLOv8	VGG-16	ResNet	MobileNet
Accuracy	**92.30 %**	87.91 %	85.71 %	83.52 %
Specificity	**90.91 %**	86.36 %	86.36 %	81.82 %
Sensitivity	**93.62 %**	89.36 %	85.11 %	85.11 %

Bold values are the higher results for each algorithm.

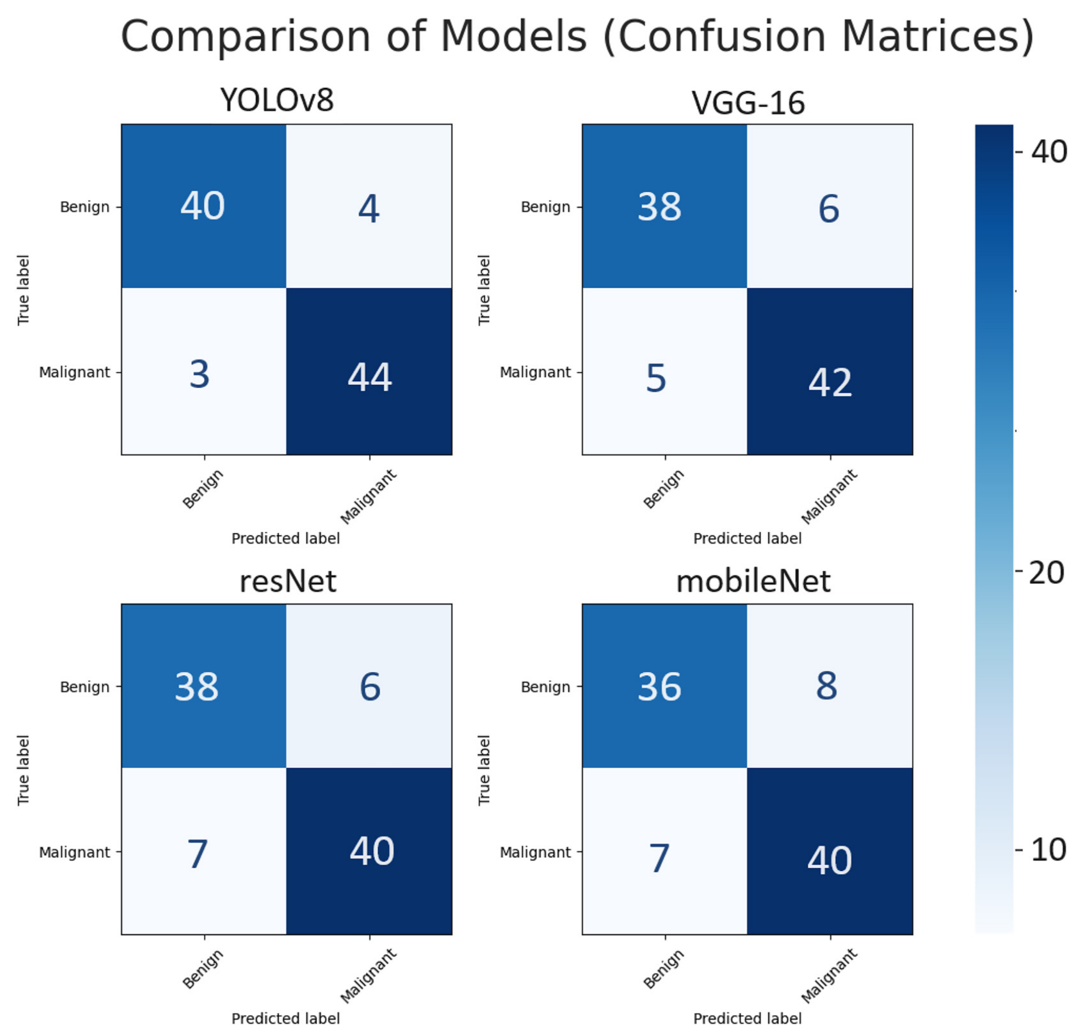


Figure 3: Comparison of models' performance; indicative CT confusion matrices. (*Source*: created by the author).

and LIME was applied to the same images and was compared with the relative ablation heat-map based XAI image. Since ablation analysis was uniquely suited for the primary YOLOv8 architecture (which does not offer an explanation alternative), the second best performing model in the current study was selected (i.e. ResNet18) classifier on the same PET/CT dataset to enable SHAP and LIME comparisons. In the following images characteristic malignant cases are showcased.

4 Discussion

The current study has a multifaceted focus. First, it set to produce an explainable approach for NSCLC classification, by developing an accurate computer-aided decision-making system that can in equal efficiency discern between benign and malignant NSCLC cases. Furthermore, another objective was to offer valid insight to the decision process of the system. This is what actually makes this approach explainable, as it reveals which points/factors played major role in the eventual prediction.

Indeed, as Tables 2 and 3 showcase, the proposed model (i.e. the one based on the YOLOv8 architecture) achieved top of the class accuracies. More specifically, as far as PET images are concerned, it classified between malignant and benign cases with an accuracy of 89 %. Also and most importantly it exhibited an extremely low False Positive Rate (FPR) of 2.27 %. On the other hand, for CT clinical scans the results were even more accurate, with a top accuracy of 92.3 % and FPR of 9.1 %. These metrics mean that not only the model is accurate in NSCLC classification, but also (and perhaps most crucially) it has a very low FPR. This actually means that very few false malignant cases were predicted, a factor that can have a great implicit impact. It can save both the patient and the National Healthcare System (NHS) extreme psychological and economical toll. On a side note, when trained on CT image sets, apparently all tested models performed slightly better than with PET images.

As far as misses (False Negative and False Positive instances) are concerned, on the case of CT medical scans, the misses were too few (4 False Positive and 3 False Negative occurrences) to reach a meaningful conclusion or error pattern. The same applies to the False Positive instances for the PET case. However, the PET False Negative occurrences were enough (9 instances) for a meaningful error analysis. Based on the affected scans, it was noticed that the PET False Negative ablation maps involved patients with minimal tracer uptake. Hence, their corresponding PET images were of inherently low diagnostic value. Nevertheless, these patients were ultimately confirmed positive (a phenomenon that can occasionally occur), which indirectly vindicates the selected YOLOv8 model, as the ground-truth labels were established by biopsy or clinical follow-up rather than by imaging alone.

The second main objective of this study, the matter of explainability, proved to be a challenge. The optimal model was based on YOLOv8 architecture, which is a quite modern/new algorithm. Even though previous versions of YOLO architecture have been successfully analyzed with traditional explainability analyses such as SHAP, this version had no relative library which could be utilized for that purpose. Thus, a new approach was to be found in order to shed light unto this black-box system. This was accomplished by employing ablation on the original images. By consistently hiding parts of the image and measuring their impact on the prediction result, ablation heat-maps were created. These heat-maps contained the information regarding which parts of the images were the most crucial for the eventual prediction. By overlaying the produced heat-maps on top of the image, we produced a clear method to explain the classification results by highlighting the most important parts of medical image, which were evidently the SPNs themselves.

Hence, regarding the PET image set (Figures 4 and 5), the explainability heat-maps clearly identify the malignant SPNs. For instance, in Figure 4a the prediction model accurately identified the existence of two SPNs (the characteristically fluorescent points). Moreover, Figure 4a–c. showcase that the existence of the SPNs was pinpointed and it was what led to the eventual prediction. On the contrary, as Figure 5a and b depict, for benign cases, the focal points of the heat-maps were somewhat all over the image. This can explain the choice of the benign class, as no specific malignant finding was spotted.

Similarly, CT image sets showcased the same behavior. Due to the way CT procedure functions, the malignant SPNs on the original images may not be equally discernible to the untrained human eye as in their PET counterparts. Nevertheless, the heat-maps can be utilized to identify the malignant instances. In particular, Figure 6a–c depict malignant cases and which points of the image led to the final classification. On the other hand, the "heat" in Figure 7a and b is somewhat less focused and more spread over the image, showcasing a lack of a main point interest, i.e. a malignant node.

Moreover, Figures 8, 9, 10, and 11 present the validation results of ablation heat-maps against SHAP and LIME explanations. This dual-analysis paradigm revealed strong spatial agreement between techniques, with ablation heat-maps identifying the majority of the high-importance regions highlighted by SHAP and LIME, while

PET Image set

a. b. c.

Figure 4: Characteristic ablation heat-maps overlays for PET scans; **malignant** cases. (*Source*: created by the author).

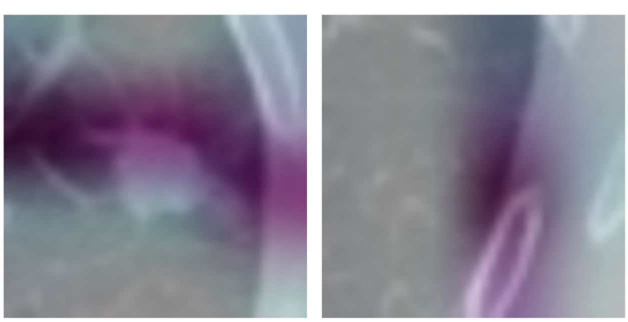
a. b.

Figure 5: Characteristic ablation heat-maps overlays for PET scans; **benign** cases. (*Source*: created by the author).

CT Image set

a. b. c.

Figure 6: Characteristic ablation heat-maps overlays for CT scans; **malignant** cases. (*Source*: created by the author).



a. b.

Figure 7: Characteristic ablation heat-maps overlays for CT scans; **benign** cases. (*Source*: created by the author).

maintaining YOLOv8's real-time processing capabilities. In some cases, ablation heat-maps provided even more accurate focus on the important regions than the other two techniques. The concordance across methods substantiates that the ablation technique could provide clinically relevant explanations without compromising the architectural advantages of YOLOv8.

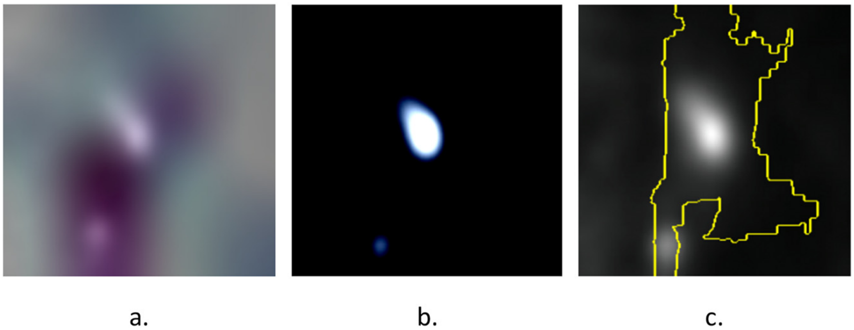


Figure 8: XAI images for malignant PET cases; a. Ablation heat-map (YOLOv8), b. SHAP (ResNet), c. LIME (ResNet). (*Source*: created by the author).

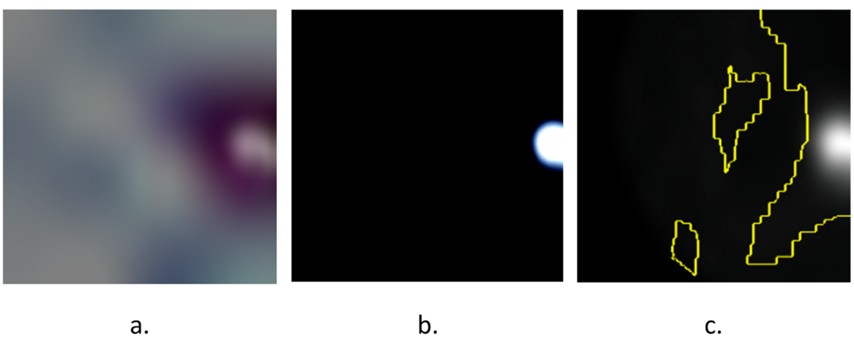


Figure 9: XAI images for malignant PET cases; a. Ablation heat-map (YOLOv8), b. SHAP (ResNet), c. LIME (ResNet). (*Source*: created by the author).

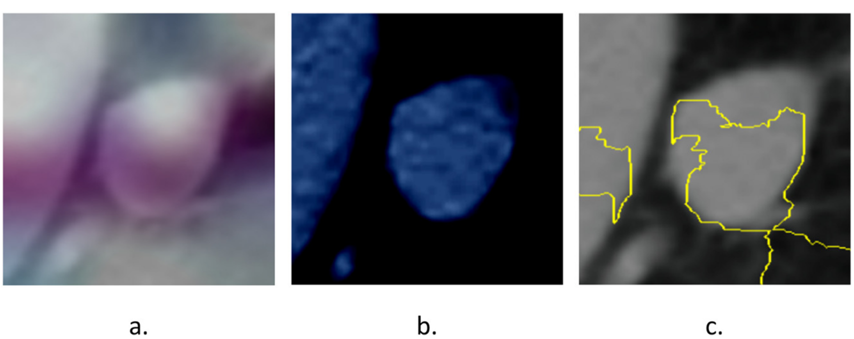


Figure 10: XAI images for malignant CT cases; a. Ablation heat-map (YOLOv8), b. SHAP (ResNet), c. LIME (ResNet). (*Source*: created by the author).

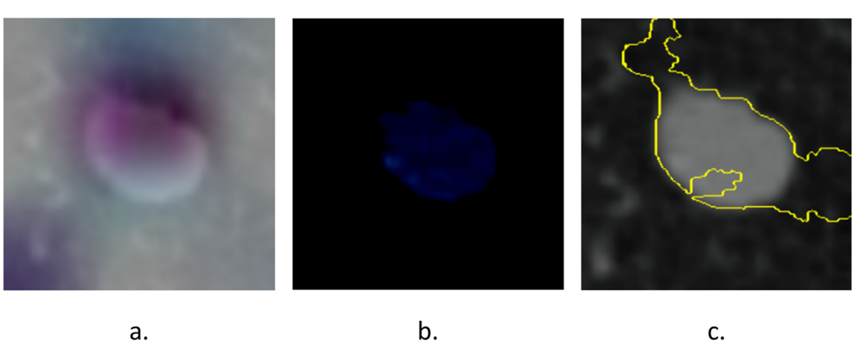


Figure 11: XAI images for malignant CT cases; a. Ablation heat-map (YOLOv8), b. SHAP (ResNet), c. LIME (ResNet). (*Source*: created by the author).

Overall, the integration of explainability is paramount for comprehending the model's decision-making process and the variables influencing its predictions. This transparency is especially vital to gain trust from the medical community, ensuring confidence in the model's predictive accuracy. Last but not least, such insights derived from these techniques can greatly aid in identifying outlier cases and any potential errors in input images.

In conclusion, this study presents actionable proposals tailored for the medical-biomedical community. Firstly, considering both PET and CT modalities is recommended due to their distinct information content, as algorithms can independently glean valuable insights from each. An argument is made for integration of both clinical characteristics and manually extracted image features by medical professionals. Employing explainability-enhancing methods is advocated to provide physicians with a more comprehensive understanding of the neural network's decision-making process, allowing them to assess the model's reliability. Drawing from research results that depict an increase in accuracy when the human professional is included in the process [47], incorporating the "doctor in the loop" concept in future work is also highly advisable. Seeking the input of medical professionals can enhance diagnostic accuracy through collaborative refinement with machine learning models.

Although the diagnostic performance of the proposed model was the primary focus, deployment feasibility remains a crucial aspect for clinical integration. Preliminary assessments indicate that the YOLOv8-based model achieves near real-time inference speeds (~25 ms per image) and maintains a relatively compact size (~45 MB), suggesting suitability for deployment on standard hospital servers. Future work will include comprehensive benchmarking on various hardware platforms, including edge devices, to assess performance trade-offs and optimize for resource-constrained environments.

However, the study bears certain limitations. Specifically, the image set employed was relatively small and represented a limited patient population, which may constrain the generalizability of the findings. Expanding the dataset to include a more diverse and extensive cohort from multiple institutions could enhance the robustness and applicability of the model. In addition, future efforts may benefit from the integration of transfer learning techniques, leveraging pre-trained models from related medical imaging domains to improve performance in data-scarce scenarios. The continued use and refinement of data augmentation strategies – such as geometric transformations and intensity normalization – also present viable avenues to mitigate overfitting and enhance model generalization. Ultimately, creating a multimodal approach for the matter at hand, combining insights from image and clinical data could enhance even further the current approach.

5 Conclusions

The primary aim of the current work was to develop a highly accurate explainable classification system for NSCLC cases. Notably, on the front of accuracy, the optimal model selected achieved an accuracy of 92.3 % for CT image sets and 89 % for PET image sets, demonstrating equal proficiency in predicting both benign and malignant cases with high rates.

On top of that, the introduction of a novelty explainability mechanism into the black-box NSCLC classification system yielded valuable insights. The YOLOv8 being a much contemporary architecture has still not been treated by more traditional explainability analyses. Hence, the current study offers a new method of explaining classifications from the particular DL algorithm, with the use of ablation heat-maps. This methodology affords transparency into the model's decision-making process, enhancing comprehension of contributing factors. This transparency fosters trust, crucial in medical contexts where prediction accuracy profoundly impacts patient outcomes. Highlighting impactful features enables users to scrutinize predictions, promoting error identification and informed decision-making.

Future endeavors encompass developing a holistic MDSS for NSCLC utilizing the results of the current study as well as its accompanying one for clinical data [30]. Thus, a complete solution on this subject will be offered. Moreover, expanding the model's applicability to diverse datasets from various external sources is also part of our forthcoming exploration, ensuring a broader and more comprehensive scope for advancing our understanding of NSCLC.

Author contributions: Design and conduct of the study (A.D.S.); data collection (I.A., N.P., D.A., N.P., E.P.); data processing (A.D.S., I.A., N.P., D.A.); manuscript preparation (A.D.S., S.M., I.A.); manuscript review (N.P., E.P.); medical expertise (N.P., D.A., N.P.). All figures in the current manuscript were created by the authors.

Funding information: 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 sup-port Faculty Members & Researchers” (Project Number 3656).
Competing interests: The authors declare no competing interests.
Data availability: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
Additional information: Correspondence and requests for materials should be addressed to E.P.

References

[1] L. A. Torre, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, “Global cancer statistics, 2012,” *CA*, vol. 65, no. 2, pp. 87–108, 2015, https://doi.org/10.3322/caac.21262.
[2] R. S. Herbst, D. Morgensztern, and C. Boshoff, “The biology and management of non-small cell lung cancer,” *Nature*, vol. 553, no. 7689, pp. 446–454, 2018, https://doi.org/10.1038/nature25183.
[3] D. E. Wood, “National comprehensive cancer network (NCCN) clinical practice guidelines for lung cancer screening,” *Thorac. Surg. Clin.*, vol. 25, no. 2, pp. 185–197, 2015, https://doi.org/10.6004/jnccn.2010.0056.
[4] L. Gandhi *et al*., “Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer,” *N. Engl. J. Med.*, vol. 378, no. 22, pp. 2078–2092, 2018, https://doi.org/10.1056/NEJMoa180100.
[5] J.-C. Soria *et al*., “Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer,” *N. Engl. J. Med.*, vol. 378, no. 2, pp. 113–125, 2018, https://doi.org/10.1056/NEJMoa1713137.
[6] S. Spänig *et al*., “The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes,” *Artif. Intell. Med.*, vol. 100, p. 101706, 2019, https://doi.org/10.1016/j.artmed.2019.101706.
[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, https://doi.org/10.1016/j.csbj.2014.11.005.
[8] R. Cuocolo, M. Caruso, T. Perillo, L. Ugga, and M. Petretta, “Machine learning in oncology: A clinical appraisal,” *Cancer Lett.*, vol. 481, pp. 55–62, 2020, https://doi.org/10.1016/j.canlet.2020.03.032.
[9] M. Guberina *et al*., “Prediction of malignant lymph nodes in NSCLC by machine-learning classifiers using EBUS-TBNA and PET/CT,” *Sci. Rep.*, vol. 12, no. 1, p. 17511, 2022, https://doi.org/10.1038/s41598-022-21637-y.
[10] L. Yu *et al*., “Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis,” *BMC Cancer*, vol. 19, no. 1, pp. 1–12, 2019, https://doi.org/10.1186/s12885-019-5646-9.
[11] A. Janik *et al*., “Machine learning-assisted recurrence prediction for early-stage non-small-cell lung cancer patients,” arXiv preprint arXiv: 2211.09856, 2022.
[12] D. Yu *et al*., “Copy number variation in plasma as a tool for lung cancer prediction using extreme gradient boosting (XGBoost) classifier,” *Thorac. Cancer*, vol. 11, no. 1, pp. 95–102, 2020, https://doi.org/10.1111/1759-7714.13204.
[13] K. Pak *et al*., “A decision tree model for predicting mediastinal lymph node metastasis in non-small cell lung cancer with F-18 FDG PET/CT,” *PLoS One*, vol. 13, no. 2, p. e0193403, 2018, https://doi.org/10.1371/journal.pone.0193403.
[14] N. Kureshi, S. S. R. Abidi, and C. Blouin, “A predictive model for personalized therapeutic interventions in non-small cell lung cancer,” *IEEE J. Biomed. Health Inf.*, vol. 20, no. 1, pp. 424–431, 2016, https://doi.org/10.1109/JBHI.2014.2377517.
[15] Y. Han *et al*., “Histologic subtype classification of non-small cell lung cancer using PET/CT images,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 48, pp. 350–360, 2021, https://doi.org/10.1007/s00259-020-04771-5.
[16] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao, “Classification of lung cancer using ensemble-based feature selection and machine learning methods,” *Mol. BioSyst.*, vol. 11, no. 3, pp. 791–800, 2015, https://doi.org/10.1039/c4mb00659c.
[17] M. Rabbani, J. Kanevsky, K. Kafi, F. Chandelier, and F. J. Giles, “Role of artificial intelligence in the care of patients with nonsmall cell lung cancer,” *Eur. J. Clin. Invest.*, vol. 48, no. 4, p. e12901, 2018, https://doi.org/10.1111/eci.12901.
[18] A. Prelaj *et al*., “Machine learning using real-world and translational data to improve treatment selection for NSCLC patients treated with immunotherapy,” *Cancers*, vol. 14, no. 2, p. 435, 2022, https://doi.org/10.3390/cancers14020435.
[19] I. D. Apostolopoulos, I. E. Livieris, D. J. Apostolopoulos, N. D. Papathanasiou, P. E. Pintelas, and G. S. Panayiotakis, “Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques,” *Med. Biol. Eng. Comput.*, vol. 59, no. 6, pp. 1299–1310, 2021, https://doi.org/10.1007/s11517-021-02378-y.
[20] I. D. Apostolopoulos, N. D. Papathanasiou, and G. S. Panayiotakis, “Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning,” *Biocybern. Biomed. Eng.*, vol. 41, no. 4, pp. 1243–1257, 2021, https://doi.org/10.1016/j.bbe.2021.08.006.
[21] M. Zhang *et al*., “Accurate classification of non-small cell lung cancer (NSCLC) pathology and mapping of EGFR mutation spatial distribution by ambient mass spectrometry imaging,” *Front. Oncol.*, vol. 9, p. 804, 2019, https://doi.org/10.3389/fonc.2019.00804.
[22] S. S. Laros *et al*., “Machine learning classification of mediastinal lymph node metastasis in NSCLC: A multicentre study in a Western European patient population,” *EJNMMI Phys.*, vol. 9, no. 1, pp. 1–10, 2022.

[23] U. Bashir *et al*., "Non-invasive classification of non-small cell lung cancer: A comparison between random forest models utilising radiomic and semantic features," *Br. J. Radiol.*, vol. 92, no. 1099, p. 20190159, 2019, https://doi.org/10.1259/bjr.20190159.
[24] B. Koonce, *ResNet 50. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, Berkeley, CA, Springer Nature, 2021, pp. 63–72.
[25] A. G. Howard *et al*., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017, https://doi.org/10.1007/978-1-4842-6168-2.
[26] J. Redmon *et al*., "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, IEEE, 2016.
[27] W. H. Organisation, "A healthy lifestyle - WHO recommendations," 2010. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle—who-recommendations.
[28] E. Tsukamoto and S. Ochi, "PET/CT today: System and its impact on cancer diagnosis," *Ann. Nucl. Med.*, vol. 20, pp. 255–267, 2006, https://doi.org/10.1007/bf02984642.
[29] P. Huang *et al*., "Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study," *Radiology*, vol. 286, no. 1, pp. 286–295, 2018, https://doi.org/10.1148/radiol.2017162725.
[30] I. D. A. Agorastos Dimitrios Samaras, M. Serafeim, N. D. Papathanasiou, D. J. Apostolopoulos, E. Papageorgiou, and N. Papandrianos, "Explainable classification for non-small cell lung cancer based on positron emission tomography features and clinical data," in *14th International Conf. IISA*, Volos, IEEE, 2023.
[31] W. Wang *et al*., "A new image classification approach via improved MobileNet models with local receptive field expansion in shallow layers," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–10, 2020, https://doi.org/10.1155/2020/8817849.
[32] X. Liu *et al*., "Real-time marine animal images classification by embedded system based on mobilenet and transfer learning," in *OCEANS 2019-Marseille*, Marseille, France, IEEE, 2019.
[33] A. Inui *et al*., "Detection of elbow OCD in the ultrasound image by artificial intelligence using YOLOv8," *Appl. Sci.*, vol. 13, no. 13, p. 7623, 2023, https://doi.org/10.3390/app13137623.
[34] Y. Li, Q. Fan, H. Huang, Z. Han, and Q. Gu, "A modified YOLOv8 detection network for UAV aerial image recognition," *Drones*, vol. 7, no. 5, p. 304, 2023, https://doi.org/10.3390/drones7050304.
[35] Z.-P. Jiang, Y. Y. Liu, Z. E. Shao, and K. W. Huang, "An improved VGG16 model for pneumonia image classification," *Appl. Sci.*, vol. 11, no. 23, p. 11185, 2021, https://doi.org/10.3390/app112311185.
[36] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in *2019 International Conf. on Information Technology (ICIT)*, IEEE, 2019.
[37] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer," *Procedia Comput. Sci.*, vol. 179, pp. 423–431, 2021, https://doi.org/10.1016/j.procs.2021.01.025.
[38] A. Mahajan and S. Chaudhary, "Categorical image classification based on representational deep network (RESNET)," in *2019 3rd International Conf. on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, IEEE, 2019.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[40] S. Kundu, "AI in medicine must be explainable," *Nat. Med.*, vol. 27, no. 8, p. 1328, 2021, https://doi.org/10.1038/s41591-021-01461-z.
[41] J. Caspers, "Translation of predictive modeling and AI into clinics: A question of trust," *Eur. Radiol.*, vol. 31, no. 7, pp. 4947–4948, 2021, https://doi.org/10.1007/s00330-021-07977-9.
[42] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "AI-assisted decision-making in healthcare: The application of an ethics framework for big data in health and research," *Asian Bioeth. Rev.*, vol. 11, pp. 299–314, 2019, https://doi.org/10.1007/s41649-019-00096-0.
[43] S. E. Middleton *et al*., "Trust, regulation, and human-in-the-loop AI: Within the European region," *Commun. ACM*, vol. 65, no. 4, pp. 64–68, 2022.
[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, NIPS, 2017.
[45] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, Association for Computational Linguistics, 2016.
[46] R. R. Selvaraju *et al*., "Grad-CAM: Why did you say that?," arXiv preprint arXiv:1611.07450, 2016.
[47] A.-D. Samaras, S. Moustakidis, I. D. Apostolopoulos, N. Papandrianos, and E. Papageorgiou, "Classification models for assessing coronary artery disease instances using clinical and biometric data: An explainable man-in-the-loop approach," *Sci. Rep.*, vol. 13, no. 1, p. 6668, 2023, https://doi.org/10.1038/s41598-023-33500-9.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/jisys-2025-0064).