

An [18F]-FDG PET/CT Image Dataset for Artificial Intelligence-Based Classification of Lung Cancer

Mohammad Karimpour (MSc)¹, Alireza Mehdizadeh (MD, PhD)², Mehrosadat Alavi (MD)³, Tahereh Mahmoudi (PhD)^{1*}

¹Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran

²Research Center for Neuromodulation and Pain, Shiraz University of Medical Sciences, Shiraz, Iran

³Department of Nuclear Medicine, Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

Lung cancer is a leading cause of cancer-related mortality worldwide. [18F]-fluorodeoxyglucose ([18F]-FDG) Positron Emission Tomography/Computed Tomography (PET/CT) imaging is an advanced medical imaging modality used for lung cancer diagnosis. However, despite the availability of such an imaging modality, current methods of diagnosing cancer and interpreting images are prone to error. Therefore, Artificial Intelligence (AI)-based methods for medical image analysis are introduced and widely applied to enhance diagnostic accuracy. These methods, particularly deep learning techniques, require large datasets. Collecting image datasets, especially PET/CT images, is time-consuming, labor-intensive, and in some instances impossible due to the high costs and low availability of imaging centers. To facilitate further research in AI for lung cancer diagnosis, we present a publicly available 2D [18F]-FDG PET/CT lung image dataset comprising 3,061 image slices obtained from 94 patients. The dataset provides pathology-confirmed ground truth annotations together with patient-level identifiers, making it suitable for developing, validating, benchmarking, and comparing AI algorithms for lung cancer classification, localization, explainable AI, and other PET/CT-based computer-aided diagnostic applications.

Citation: Karimpour M, Mehdizadeh AR, Alavi M, Mahmoudi T. An [18F]-FDG PET/CT Image Dataset for Artificial Intelligence-Based Classification of Lung Cancer. *J Biomed Phys Eng.* 2026;16(5):425-430. doi: 10.31661/jbpe.v0i0.2607-2122.

Keywords

Medical Image Dataset; Artificial Intelligence; Lung Cancer; Small Cell Lung Carcinoma; Non-Small Cell Lung Carcinoma; Positron Emission Tomography Computed Tomography; Computer-Aided Diagnosis

Introduction

Lung cancer is a leading cause of cancer-related mortality worldwide and is one of the most common types of cancer, posing a major global public health burden [1]. Early and accurate diagnosis is essential for improving patient outcomes because survival rates decline substantially with disease progression [2]. Lung cancer is broadly classified into two major histological subtypes: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC), with NSCLC accounting for the majority of cases [3].

Medical imaging modalities, including Computed Tomography (CT) scan and Positron Emission Tomography (PET), have a significant role in diagnosing lung cancer. Despite their widespread clinical use, each modality has inherent limitations. CT scans provide anatomical information but lack functional information, making low-contrast

*Corresponding author: Tahereh Mahmoudi
Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
E-mail: t.mahmoudi94@gmail.com

Received: 10 July 2026
Accepted: 27 August 2026

abnormalities difficult to detect, whereas PET primarily provides functional information with limited anatomical detail [4]. Multimodality imaging techniques, such as PET/CT, overcome many of these limitations by simultaneously providing anatomical and functional information [5]. [18F]-fluorodeoxyglucose ([18F]-FDG) positron emission tomography/computed tomography (PET/CT) is widely used for assessing malignancies; however, even with advanced techniques like PET/CT, current diagnostic methods remain susceptible to errors. Despite its clinical value, PET/CT interpretation remains susceptible to false-positive findings caused by inflammatory or infectious conditions and false-negative findings in low-metabolic or small lesions [6]. Moreover, interpretation of large volumes of PET/CT images is time-consuming and subject to inter- and intra-observer variability.

Recent advances in artificial intelligence (AI), particularly deep learning, have significantly improved computer-aided diagnosis (CAD) in medical imaging by enhancing lesion detection, image classification, and radiologists' diagnostic performance [7-12]. However, the clinical translation, reproducibility, and objective comparison of these approaches remain constrained by the limited availability of large, high-quality, and publicly accessible PET/CT imaging datasets. Because data acquisition and expert annotation are expensive, time-consuming, and subject to ethical and privacy constraints, most published studies rely on private institutional datasets that are unavailable to the research community, limiting independent validation, reproducibility, and standardized benchmarking of AI algorithms. Publicly available, high-quality annotated datasets are therefore essential for transparent method comparison, external validation, and reproducible research, ultimately accelerating the translation of AI technologies into clinical practice.

To address these challenges, we present a publicly available, expert-annotated 2D

[18F]-FDG PET/CT lung image dataset comprising three clinically relevant categories: normal, NSCLC, and SCLC. The dataset incorporates pathology-confirmed subtype annotations, expert-reviewed image labels, and patient-level identifiers that facilitate data partitioning without leakage for AI studies. This dataset can support several research applications. It provides a benchmark for designing and assessing deep learning architectures for lung cancer classification from PET/CT image data, addressing the field's reliance on private, non-shareable institutional datasets [13]. Additionally, the dataset is suited to explainable AI methods, such as class activation mapping, which localize discriminative image regions without requiring manual segmentation [14]. The dataset can also facilitate research on generalizability across patient groups and acquisition techniques by acting as an external validation cohort for models trained elsewhere. Together, these characteristics provide reliable ground truth and make the dataset well suited for developing, validating, benchmarking, and comparing AI algorithms for lung cancer classification, explainable AI, and other computer-aided diagnostic applications. We anticipate that this dataset will serve as a valuable resource for future PET/CT-based lung cancer research and promote the development of robust, reproducible, and clinically translatable AI models. In the following sections, we describe the patient cohort and imaging acquisition protocol, the expert annotation process, and the structural organization of the dataset, and we discuss its potential applications in AI-based lung cancer studies.

Methods

The data were retrospectively collected from patients who underwent standard medical treatment at Kowsar Hospital in Shiraz, Iran. All collected data were anonymized to protect patient confidentiality. Participants' names were removed, and their original identifiers were replaced with anonymized codes to

ensure the protection of identifiable information throughout the study.

Patient selection and image acquisition

This retrospective study included patients who underwent [18F]-FDG PET/CT imaging at Kowsar Hospital between June 2019 and March 2025. A total of 94 patients (3,061 2D images) were included in the study. Out of 3,061 total images, 1,913 images were classified as normal with no signs of malignancy, while 612 images were diagnosed as NSCLC, and 536 as SCLC. The cohort comprised 57 males and 37 females with a mean age of 54.13 ± 17.03 years. Figure 1 depicts details of data collection and number of cases in each class. No fixed patient-to-image ratio was imposed during case selection. Instead, all thoracic images demonstrating malignancy confirmed by pathology were included regardless of the number of images per patient. This inclusive approach was adopted because PET/CT imaging data, and lung cancer subtype imaging in particular, remain scarce and costly to acquire and annotate, and datasets with

image-level ground truth labels are especially rare [15,16]. Consequently, every image containing confirmed pathological evidence was considered valuable for inclusion rather than being subsampled.

After fasting for 6 hours and waiting approximately 60–80 minutes following intravenous FDG injection, whole-body PET/CT images were acquired using a Philips Ingenuity TF 64 PET/CT scanner. Although the entire body was scanned, only the thoracic images relevant to the study objectives were selected. The scanner provided 2D PET/CT image slices with dimensions of $363 \times 657 \times 3$. To remove unnecessary information, these images were cropped to a final size of $245 \times 457 \times 3$.

Data Labelling

All image labeling was performed by a highly experienced nuclear medicine physician with over 30 years of clinical experience. Furthermore, subtype classification was determined by histopathological diagnosis rather than imaging interpretation alone, which minimizes the impact of single-reader subjectivity on the final ground-truth labels. The

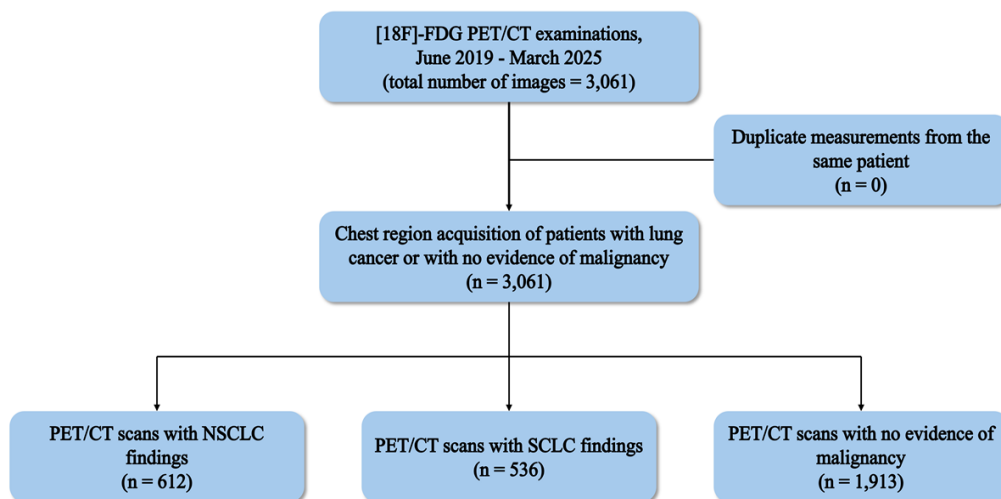


Figure 1: The workflow of patient selection, image acquisition, annotation, and dataset organization. The final dataset comprised three classes: normal cases, non-small cell lung cancer (NSCLC), and small cell lung cancer (SCLC).

([18F]-FDG: [18F]-fluorodeoxyglucose, PET/CT: Positron Emission Tomography/Computed Tomography)

labelling process consisted of two sequential steps. Each [18F]-FDG PET/CT image slice was first reviewed and labeled as either lung cancer or normal (no evidence of malignancy) based on the imaging findings. Image slices identified as lung cancer were subsequently classified into SCLC or NSCLC according to the corresponding histopathological reports (Figure 2).

Histopathological diagnosis was considered the reference standard for subtype annotation. Since SCLC and NSCLC often exhibit overlapping metabolic and anatomical characteristics on PET/CT images, reliable differentiation based solely on imaging findings is challenging. Using diagnoses verified by pathology substantially improves the reliability of the dataset labels and provides high-quality ground truth suitable for the development, validation, and benchmarking of supervised artificial intelligence algorithms. Whenever uncertainty existed regarding image interpretation or subtype assignment, the nuclear medicine physician reviewed the patients' complete clinical records together with the official pathological reports before assigning the final label. Patients with incomplete pathological confirmation or insufficient clinical documentation were excluded to minimize the risk of incorrect labelling.

Data Records

The lung cancer PET/CT dataset can be accessed via a public zenodo repository [17]. In this repository, three separate folders exist

containing the PET/CT images of the three classes. A folder named "Normal Cases" contains cases with no signs of lung cancer. Other two folders named "Non-Small Cell Lung Cancer (NSCLC)" and "Small Cell Lung Cancer (SCLC)" respectively contain images of patients diagnosed with non-small cell lung cancer and small cell lung cancer. The images are in TIFF format, and scans belonging to the same patient are assigned a unique numerical identifier to facilitate proper data splitting and prevent data leakage. The numerical identifier preceding the parentheses in each image filename represents the patient case ID, while the number within the parentheses indicates the image slice number associated with that patient case. Therefore, images sharing the same numerical identifier before the parentheses belong to the same patient case. The dataset does not contain any direct identifiers, and the assigned identifiers have been modified and do not correspond to the original clinical IDs to ensure the anonymization of potentially identifiable information.

This PET/CT dataset can support a range of artificial intelligence applications beyond image classification [18], including explainable AI approaches (e.g., class activation mapping), transfer learning, and methods for improving model generalizability across different imaging centers and scanners. We anticipate that this dataset will support the development, validation, and objective comparison of future PET/CT-based artificial intelligence algorithms for lung cancer.

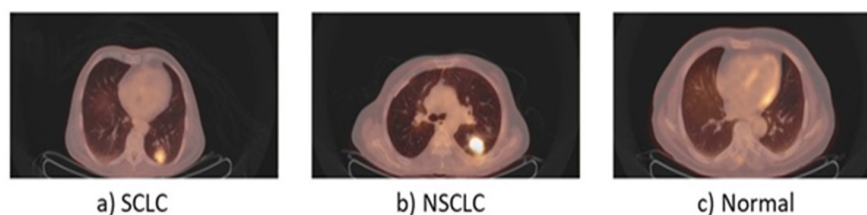


Figure 2: Image slices of positron emission tomography/computed tomography (PET/CT) scans for three categories: (a) small cell lung cancer (SCLC), (b) non-small cell lung cancer (NSCLC), and (c) normal.

Technical Validation

Motion artifacts, image misregistration, and incomplete PET/CT acquisition are well-recognized sources of diagnostic error and reduced image quality in PET/CT imaging [19,20]. Accordingly, images exhibiting these issues, as well as cases with insufficient clinical documentation, were excluded from the dataset. All labels were assigned by an experienced nuclear medicine physician and verified against each patient's pathological and clinical reports. These quality-control procedures were intended to maximize annotation reliability and support the development of high-quality, trustworthy data for AI research.

Usage Notes

The dataset is publicly available through the repository specified in this article. All users are required to appropriately cite this article and the associated dataset publication in any research, publication, or derivative work based on this dataset.

Acknowledgment

The authors gratefully acknowledge the support of Kowsar Hospital, Shiraz, Iran, for providing access to clinical data and imaging resources used in this study.

Authors' Contribution

M. Karimpour was involved in data preprocessing, methodology validation, and writing the paper. T. Mahmoudi and AR. Mehdizadeh were involved in supervision, conceptualization, methodology, and checking the final version of the manuscript. M. Alavi participated in conceptualization, data collection, and checking the labels. All authors read the final version of the manuscript and approved it.

Ethical Approval

All research methods in this study followed the tenets of the Declaration of Helsinki, and the project was found to be in accordance with the ethical considerations and the nation-

al standards of conducting research in Iran. The Research Ethics Committee of Shiraz University of Medical Sciences (SUMS) approved the study (Approval No. IR.SUMS.REC.1401.715).

Conflict of Interest

AR. Mehdizadeh, Editor-in-Chief and Chairperson, was not involved in the peer-review and decision-making processes for this manuscript. The non-author, Editorial Board and reviewers oversaw the peer review process for this paper.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* 2021;**71**(3):209-49. doi: 10.3322/caac.21660. PubMed PMID: 33538338.
2. Nooreldeen R, Bach H. Current and Future Development in Lung Cancer Diagnosis. *Int J Mol Sci.* 2021;**22**(16):8661. doi: 10.3390/ijms22168661. PubMed PMID: 34445366. PubMed PMCID: PMC8395394.
3. Hendriks LEL, Remon J, Faivre-Finn C, Garassino MC, Heymach JV, Kerr KM, et al. Non-small-cell lung cancer. *Nat Rev Dis Primers.* 2024;**10**(1):71. doi: 10.1038/s41572-024-00551-9. PubMed PMID: 39327441.
4. Kwee TC, Kwee RM, Nievelstein RA. Imaging in staging of malignant lymphoma: a systematic review. *Blood.* 2008;**111**(2):504-16. doi: 10.1182/blood-2007-07-101899. PubMed PMID: 17916746.
5. Padmanabhan P, Nedumaran AM, Mishra S, Pandarinathan G, Archunan G, Gulyás B. The Advents of Hybrid Imaging Modalities: A New Era in Neuroimaging Applications. *Adv Biosyst.* 2017;**1**(8):e1700019. doi: 10.1002/adbi.201700019. PubMed PMID: 32646180.
6. Hochegger B, Alves GR, Irion KL, Fritscher CC, Fritscher LG, Concatto NH, Marchiori E. PET/CT imaging in lung cancer: indications and findings. *J Bras Pneumol.* 2015;**41**(3):264-74. doi: 10.1590/S1806-37132015000004479. PubMed PMID: 26176525. PubMed PMCID: PMC4541763.
7. Ambrosini RD, Wang P, O'Dell WG. Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching. *J Magn Reson Imaging.* 2010;**31**(1):85-93. doi:

- 10.1002/jmri.22009. PubMed PMID: 20027576. PubMed PMCID: PMC2799295.
8. Castellino RA. Computer aided detection (CAD): an overview. *Cancer Imaging*. 2005;**5**(1):17-9. doi: 10.1102/1470-7330.2005.0018. PubMed PMID: 16154813. PubMed PMCID: PMC1665219.
9. Liang M, Tang W, Xu DM, Jirapatnakul AC, Reeves AP, Henschke CI, Yankelevitz D. Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers. *Radiology*. 2016;**281**(1):279-88. doi: 10.1148/radiol.2016150063. PubMed PMID: 27019363.
10. Hong TS, Tomé WA, Harari PM. Heterogeneity in head and neck IMRT target design and clinical practice. *Radiother Oncol*. 2012;**103**(1):92-8. doi: 10.1016/j.radonc.2012.02.010. PubMed PMID: 22405806. PubMed PMCID: PMC3694728.
11. Li XA, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG Multi-Institutional and Multiobserver Study. *Int J Radiat Oncol Biol Phys*. 2009;**73**(3):944-51. doi: 10.1016/j.ijrobp.2008.10.034. PubMed PMID: 19215827. PubMed PMCID: PMC2911777.
12. Van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*. 2011;**261**(3):719-32. doi: 10.1148/radiol.11091710. PubMed PMID: 22095995.
13. Abumohsen M, Costa-Montenegro E, García-Méndez S, Owda AY, Owda M. Machine learning and deep learning in lung cancer diagnostics: A systematic review of technical breakthroughs, clinical barriers, and ethical imperatives. *AI*. 2026;**7**(1):23. doi: 10.3390/ai7010023.
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision; Venice, Italy: IEEE; 2017. p. 618-26.
15. Gatidis S, Hepp T, Früh M, La Fougère C, Nikolaou K, Pfannenberger C, et al. A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Sci Data*. 2022;**9**(1):601. doi: 10.1038/s41597-022-01718-3. PubMed PMID: 36195599. PubMed PMCID: PMC9532417.
16. Fallahpoor M, Chakraborty S, Pradhan B, Faust O, Barua PD, Chegeni H, Acharya R. Deep learning techniques in PET/CT imaging: A comprehensive review from sinogram to image space. *Comput Methods Programs Biomed*. 2024;**243**:107880. doi: 10.1016/j.cmpb.2023.107880.
17. Karimpour M, Mehdizadeh AR, Alavi M, Mahmoudi T. An [18F]-FDG PET/CT Image Dataset for Artificial Intelligence-Based Classification of Lung Cancer. Zenodo; 2026. doi: 10.5281/zenodo.22065809.
18. Karimpour M, Taghinezhad N, Mehdizadeh AR, Alavi M, Mahmoudi T. A computer-aided diagnosis (CAD) system based on convolutional neural networks for lung cancer diagnosis from 2D [18F]- PET/CT images. *J Appl Clin Med Phys*. 2025;**26**(10):e70285. doi: 10.1002/acm2.70285. PubMed PMID: 41065340. PubMed PMCID: PMC12509238.
19. Tsai YJ, Liu C. Pitfalls on PET/CT Due to Artifacts and Instrumentation. *Semin Nucl Med*. 2021;**51**(6):646-56. doi: 10.1053/j.semnuclmed.2021.06.015. PubMed PMID: 34243906. PubMed PMCID: PMC8490278.
20. Van Dalen JA, Vogel WV, Corstens FH, Oyen WJ. Multi-modality nuclear medicine imaging: artefacts, pitfalls and recommendations. *Cancer Imaging*. 2007;**7**(1):77-83. doi: 10.1102/1470-7330.2007.0014. PubMed PMID: 17535775. PubMed; PMCID: PMC1931515.