A Unified Multi-Organ Framework for Universal Lesion Detection, Segmentation and Longitudinal Tracking in CT Imaging

Noëlie Debs[1*], Sophie Loizillon[1*], Omar Ali[1], Rebeca Vétil[1], Alexandre Bône[1], Marc-Michel Rohé[1]

[1] Guerbet Research, Villepinte, France
* These authors contributed equally to this work

noelie.debs@guerbet.com

Abstract. Chest, abdomen, and pelvis (CAP) CT scans are central to cancer management, enabling disease progression assessment through the Response Evaluation Criteria in Solid Tumors (RECIST), a labor-intensive process manually performed by radiologists. AI-based tools have emerged to reduce this burden, yet most existing methods target a single organ and provide single-timepoint analysis, which is a major limitation in oncology where whole-body CT examination at multiple timepoints is required. In this study, we present a unified deep learning framework for CAP CT scans designed to detect, segment, and track lesions across all visible organs in a longitudinal manner. Our framework, trained on over 221,000 manually annotated lesions, achieved a sensitivity of 81.3% on target lesions as defined by RECIST, evaluated on the public Longitudinal-CT database. Regarding tracking performance, we achieved a matching accuracy of 88.7%, demonstrating the framework's capacity to correctly match lesions detected between baseline and follow-up exams. Altogether, these results indicate that automated RECIST assessment is within reach. More broadly, our pipeline, capable of detecting and tracking lesions across CAP CT scans, paves the way toward novel response biomarkers, such as total lesion burden quantification.

Keywords: Longitudinal lesion tracking, Universal lesion detection, Oncology CT imaging

1 Introduction

A chest, abdomen and/or pelvis (CAP) CT scan is a pivotal examination in an oncological context [1]. Patients diagnosed with cancer will undergo multiple CT scans during their treatment planning from staging and response assessment to follow-up and prognostication [2].

For over 25 years, Response Evaluation Criteria in Solid Tumor (RECIST) served as the standard framework for solid tumor measurements [3,4]. RECIST criteria focus on the selection of up to five representative target lesions, whose diameters are measured unidimensionally and summed to quantify tumor burden. In clinical practice,

RECIST is used to assess individual patient response and inform treatment decisions. Interpreting CT scans represents an important workload for radiologists that are already under pressure due to the growing volume of imaging studies to analyze [5,6, 7]. In the context of oncological imaging, assessing RECIST is time consuming, as radiologists need to manually select target lesions and measure them [3]. The longitudinal comparison is labor-intensive, as it involves tracking lesions from previous studies and detecting new ones in order to assess the overall response.

Artificial Intelligence (AI) has the potential to reduce the growing workload in radiology departments by assisting radiologists with image interpretation [8,9]. In recent years, numerous computer-aided-diagnosis systems for radiology have been approved by regulatory authorities and multiple studies have shown that dual reading approaches involving a radiologist and an AI system can reduce diagnostic errors and shorten the interpretation times [10,11,12]. Nevertheless, most of these AI systems focus on a single organ and provide only a single timepoint analysis, such as lung nodule detection [13], liver lesions detection [14], pancreas lesions detection [15], or lymph node detection [16]. This is a major limitation in oncology imaging where treatment response is assessed by examining whole-body CT scans at multiple time points. Recently, a new public database called Longitudinal-CT, containing 600 whole-body CT studies of patients with metastatic malignant melanoma, was released by [17]. Tumor lesions were manually annotated at two timepoints, enabling the development and evaluation of AI tools for lesion detection, segmentation, and longitudinal tracking in an oncological context.

To our knowledge, no framework currently exists that can detect lesions across all organs and track them over time in an oncological context. In this study, we therefore propose a deep learning framework for CAP CT scans designed to assist radiologists in managing cancer patients. Our objectives are to detect, segment, and track lesions on all visible organs of CAP CT scans at different time points. Our key contributions are the following: (1) a unified detection and segmentation pipeline combining nnDetection [18] and nnU-Net [19], coupled with a longitudinal tracking module for lesion progression assessment, trained on over 200,000 manually annotated lesions; (2) a comprehensive evaluation of multi-organ lesion detection and longitudinal tracking on the public Longitudinal-CT database [17], covering a wide range of lesion types and anatomical locations.

2 Material

2.1 Training Dataset

We built a retrospective multicenter database comprising 10,887 cancer patients and 12,554 CT scans, with a mean patient age of 61.6 ± 19.4 years (4,659 male, 3,972 female, 2,256 with missing demographic information). Images were acquired on 95 different scanner models from four manufacturers, GE Healthcare (63%), Philips (15%),

Siemens Healthineers (14%), and Toshiba (6%), ensuring broad scanner diversity. The dataset covers a wide range of acquisition protocols: body part coverage was predominantly abdomen-pelvis (50%), followed by thorax (22%) and thorax-abdomen-pelvis (15%); contrast phase was mostly portal-venous (54%), with unenhanced (16%), arterial (10%), and delayed (4%) phases also represented; and reconstruction kernel was smooth in the majority of cases (73%), followed by intermediate (13%) and sharp (13%).

All CT scans were reviewed and annotated by a pool of trained radiologists using an internally developed 3D Slicer module, enabling slice-by-slice 3D segmentation of all visible lesions, both benign and malignant, with access to the original radiology reports. In total, 221,346 lesions were manually annotated, with a median of 14 lesions per study (range: 1–310) and a mean volume of 5.5 ± 68.4 mL. Lesions span 13 anatomical regions, with lymph nodes (27%), lung (20%), liver (18%), bone (17%), and kidney (11%) being the most prevalent.

2.2 Testing Dataset

We evaluate our method on Longitudinal-CT [17], a publicly available dataset of whole-body CT scans from 300 melanoma patients, each comprising a baseline and a follow-up acquisition performed on Siemens scanners with portal-venous phase and smooth kernel reconstruction. In total, 7,182 tumor lesions were manually annotated by experts, with associated anatomical location, volume, and longitudinal correspondence information, making it well-suited for joint evaluation of lesion detection and tracking. Annotation was exhaustive, with every lesion judged malignant by expert readers segmented on all CT volumes regardless of size. Unlike our own protocol, benign lesions were not segmented in Longitudinal-CT.

3 Methods

3.1 Overall Pipeline

Our framework for universal lesion detection, segmentation, and longitudinal tracking in CAP CT oncology imaging is illustrated in Figure 1. The pipeline consists of two modules: (1) a lesion detection and segmentation module operating across multiple organs; and (2) a longitudinal tracking module that establishes lesion correspondences across multiple timepoints when serial acquisitions are available. Each module is described in detail in the following sections.

Detection and Segmentation module at the lesion level. The detection and segmentation module performs detection and segmentation of lesions located in the thorax, abdomen, and pelvis. Detection is performed with nnDetection [18], a deep learning framework specifically designed for medical object detection, while segmentation is performed with nnU-Net, which is also specifically designed for medical images and

automatically adapts its architecture, preprocessing, training and postprocessing pipelines to a given dataset [19]. The outputs of both models are combined using the ensembling method of [20] to reduce false positives: nnU-Net's voxel-based predictions and nnDetection's box-based predictions are fused into a final detection map in which each lesion is represented as a connected component associated with a single probability value. This fusion involves two steps: (i) spatial matching between lesions segmented by nnUnet and those detected by the nnDetection, and (ii) each matched lesion is assigned the detection probability score from nnDetection.

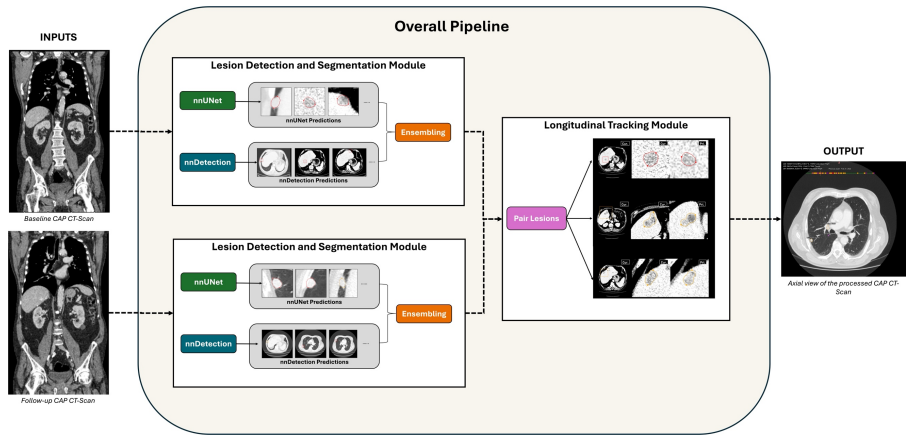

Fig. 1. Multi-organ framework for universal lesion detection, segmentation, and longitudinal tracking in CAP CT scans. Longitudinal processing pipeline, in which baseline and follow-up CT scans are independently processed through the same lesion-level detection and segmentation module, and their respective outputs are subsequently fed into a longitudinal tracking module to establish lesion-to-lesion correspondences across timepoints.

Longitudinal tracking module. This module is responsible for pairing lesions across multiple studies in the context of longitudinal processing. The 3D position of each lesion from the previous study is reconstructed in the current study coordinate space to best reproduce its relative position with respect to neighboring organs, as segmented by TotalSegmentator. For every reference organ, its centroid, boundaries and volume are stored. Each lesion is then encoded by storing its distances to the centroids and boundaries of the reference organs, together with an anatomical label. To estimate the location of a lesion from the previous study in the current study, the reconstruction is performed in two stages. First, the displacement vectors of the reference organ centroids between the two studies are used to interpolate a continuous displacement field using a 3D ordinary Kriging interpolation. This resulting field is then applied to the location of the previous lesion to estimate its approximate position in the current study. Second, this position is refined by a local grid search that identifies the location that best preserves the lesion's original anatomical relationships with the surrounding organs. A distance matrix is then computed between the estimated current positions of lesions from the previous study and the actual lesion centroids detected in the current study. Candidate pairs with excessive spatial distance or incompatible anatomical locations are discarded. The remaining lesions are matched by solving a linear assignment

problem that minimizes the total matching distance while allowing lesions to remain unmatched, accounting for newly apearing or disappearing lesions..

3.2 Experiments

Training of the detection and segmentation module at the lesion level. Both nnU-Net and nnDetection were trained on our training dataset described in Section 2.1 comprising 12,554 CAP CT scans with 221,346 manually annotated lesions, split into 12,157 training images and 397 validation images. For the segmentation and detection tasks, both models were trained using an anisotropic resampling strategy that fixes the in-plane resolution while adapting the z-spacing. The segmentation model (nnU-Net ResEnc, ~850M parameters) used a patch size of [96, 128, 128], a learning rate of 1e-3, and a batch size of 2 over 80,000 epochs. The detection model (nnDetection, ~70M parameters) used a patch size of [64, 96, 96], a learning rate of 5e-3, and a batch size of 16 over 2000 epochs.

3.3 Evaluation

Detection Metrics. Lesion detection performance was evaluated using Free-Response Receiver Operating Characteristic (FROC) analysis, first globally and then stratified by anatomical region (adrenal gland, bone, kidney, liver, lung, lymph node, other abdomen, other thorax, skin, and unknown), RECIST category (target, measurable, and non-measurable lesions, with measurable lesions defined per RECIST as >10 mm in longest axis or >15 mm in short axis for lymph nodes), lesion size (defined by longest-axis diameter as tiny, <5 mm; small, 5–10 mm; medium, 10–30 mm; and large, >30 mm), and timepoint (baseline or follow-up). FROC curves were complemented by the operating point at a detection threshold of 0.5, reporting the corresponding sensitivity and false-positive rate per scan. False-positive rates were reported only for stratifications allowing assignment of predicted lesions to the corresponding subgroup. This was directly available for lesion size and measurable status, whereas anatomical region assignment relied on TotalSegmentator predictions [21] and was therefore limited to the classes supported by the model. False-positive rates could not be reported for target and non-target categories because these labels were available only for ground-truth lesions.

Segmentation Metrics. Segmentation performance was evaluated using the lesion-level Dice Similarity Coefficient (DSC) between each ground-truth lesion and its matched prediction. Two metrics were reported: (i) overall DSC, assigning a DSC of 0 to missed lesions (false negatives) to account for detection failures, and (ii) matched-lesion DSC, computed only for successfully detected lesions to assess segmentation quality independently of detection performance. Both metrics were aggregated at the study level and summarized as mean ± standard deviation. Stratified analyses were performed according to anatomical region (lung, lymph node, liver, skin, bone), lesion size (tiny, small, medium, large), RECIST category (target, measurable, non-measurable), and study cohort.

Tracking Metrics. Longitudinal lesion association was further assessed using matching accuracy, defined as the proportion of baseline lesion pairs (ground truth–prediction) successfully matched to the corresponding lesion at follow-up. A lesion was considered successfully associated if its ground-truth identity was preserved at follow-up through either a one-to-one correspondence or a merge/split relationship, as specified by the dataset annotations.

4 Results

On the Longitudinal-CT database, our framework achieved a mean lesion detection sensitivity per study of 61.3%, with a mean overall DSC of 48.5 ± 23.7 and a matched-lesion DSC of 74.0 ± 12.4. The model produced on average 24.1 ± 13.2 false positives per study (median 22.0, IQR: 16.0-29.0) and 4.57 ± 7.5 false negatives per study (median 2.0, IQR: 0.0-5.8). Results stratified by anatomical region, lesion size, RECIST category and timepoint are detailed in Table 1.

Table. 1. Detection and segmentation performance stratified by anatomical region, lesion size, RECIST category and timepoint. Detection performance is reported as sensitivity (%) at the FROC operating point of maximum number of false positives, along with the corresponding number of false positives per volume (FP/volume). Segmentation performance is reported as overall DSC (assigning a score of 0 to missed lesions) and matched-lesion DSC (computed only on correctly detected lesions), expressed as mean ± standard deviation.

		Detection		Segmentation	
		Sensitivity (max)	FP/volume (max)	Overall DSC (mean ± std)	Matched-lesion DSC (mean ± std)
Global		61.3	24.1	48.5 ± 23.7	74.0 ± 12.4
Anatomical Region	Lung	70.1	7.80	54.3 ± 37.2	77.5 ± 13.2
	Lymph Node	60.9	-	38.9 ± 38.5	63.8 ± 29.0
	Liver	83.4	8.97	55.6 ± 30.8	66.7 ± 19.8
	Skin	35.4	-	27.1 ± 38.9	76.4 ± 22.4
	Bone	50.2	48.4	31.3 ± 36.2	62.3 ± 25.8
Lesion Size	Large	85.5	1.6	59.1 ± 35.3	69.2 ± 27.7
	Medium	72.6	10.4	51.0 ± 37.6	70.3 ± 24.3
	Small	58.5	13.0	43.0 ± 38.6	73.5 ± 17.4
	Tiny	34.9	11.7	25.4 ± 35.9	72.6 ± 15.8
RECIST	Target	81.3	-	63.3 ± 34.1	77.9 ± 17.1
	Measurable	71.7	13.9	48.0 ± 37.7	67.0 ± 26.6
	Non Measurable	48.5	21.6	35.4 ± 38.4	72.9 ± 17.4
Timepoint	Baseline	60.4	22.0	43.1 ± 38.9	71.4 ± 22.1
	Follow-up	62.5	26.2	44.7 ± 38.7	71.5 ± 21.7

Detection Performance. Sensitivity was highest for liver lesions (83.4%), followed by lung lesions (70.1%) and lymph nodes (60.9%), while bone (50.2%) and skin (35.4%) lesions remained challenging. Sensitivity decreased substantially with lesion size, ranging from 85.5% for large lesions to 34.9% for tiny lesions, with medium and small lesions achieving sensitivities of 72.6% and 58.5%, respectively. For RECIST categories, sensitivity reached 81.3% for target lesions, 71.7% for measurable lesions, and 48.5% for non-measurable lesions. Sensitivity was comparable at baseline (60.4%) and follow-up (62.5%). The mean number of false positives per scan remained substantial (24.1), largely driven by bone lesions (FP/volume max: 48.4). The corresponding FROC curves illustrating lesion-level detection performance across anatomical regions and lesion size categories are presented in Fig. 2.

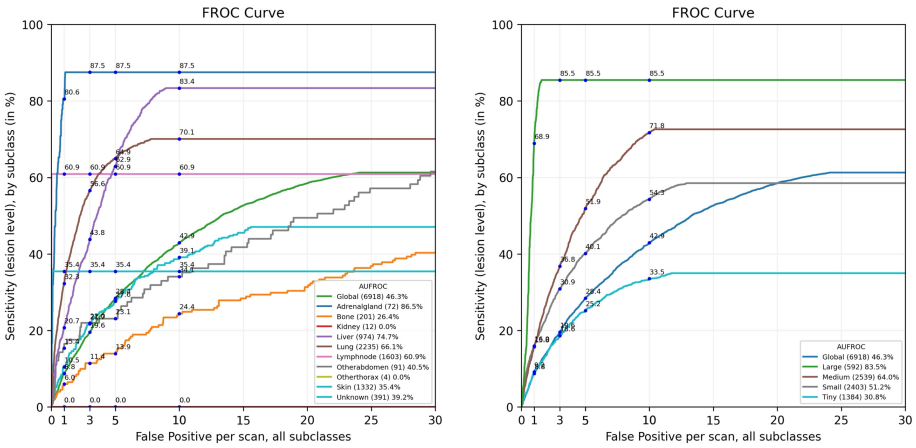

Fig. 2. FROC curves illustrating lesion-level detection sensitivity as a function of the average number of false positives per scan, stratified by anatomical region (left) and lesion size (right).

Segmentation Performance. Stratified analyses showed that overall DSC varied substantially according to lesion characteristics. Performance was highest for liver and lung lesions (55.6% and 54.3%, respectively), large lesions (59.1%), and target lesions (63.3%), whereas lower overall DSC were observed for skin (27.1%), bone (31.3%), lymph node (38.9%), tiny lesions (25.4%), and non-measurable lesions (35.4%). In contrast, matched-lesion DSC remained consistently high across all subgroups, ranging from 62.3% to 77.9%, with the highest values observed for target (77.9%), lung (77.5%), and skin lesions (76.4%), and the lowest for lymph node (63.8%) and bone lesions (62.3%). Segmentation performance was similar between the two timepoints for both matched-lesion DSC (71.4% vs. 71.5%) and overall DSC (43.1% vs. 44.7%). Representative examples of true-positive and false-positive detections are shown in Fig. 3.

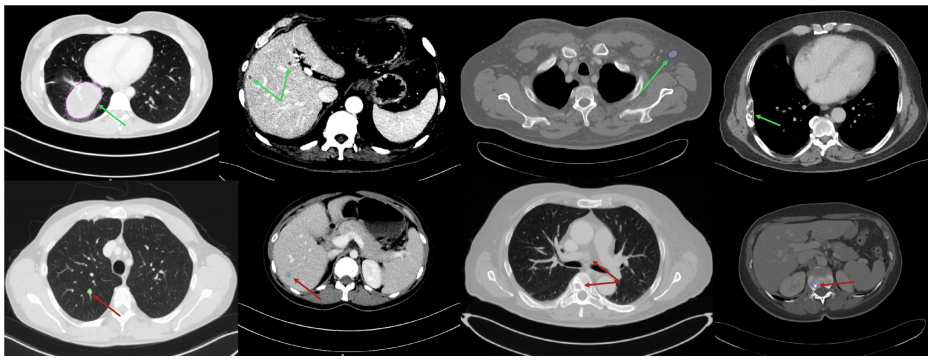

Fig. 3. Example outputs generated by our multi-organ framework for universal lesion detection and segmentation in CAP CT scans. The first row shows to true positive lesions in the lung, liver, skin and bone (left to right). The second row highlights examples of false positive detections.

Tracking Performance. Our pipeline achieved a matching accuracy of 88.7% (839/946): among the 946 ground-truth lesion pairs correctly detected at both baseline and follow-up, 839 were correctly matched.

5 Discussion and Conclusion

We presented a unified framework for lesion detection, segmentation, and tracking in whole-body CT scans, evaluated on the Longitudinal-CT public database. Our results demonstrate promising performance for automated lesion detection and tracking, paving the way toward AI-assisted RECIST assessment and broader oncological response evaluation. Nevertheless, our method exhibits limitations in challenging anatomical regions, particularly for bone and skin lesions, as well as for small lesions (5–10 mm), which remain difficult to detect reliably. We also observe substantial number of false positive detections per scan (>20). This range is consistent with results reported by [17] on Longitudinal-CT (27.8 ± 18.7 false positives per study) confirming that limiting false positive detection remains an open challenge. Part of this false positive may stem from a discrepancy between the training and evaluation definitions of a lesion: our pipeline was trained to detect all visible lesions, including benign findings such as cysts, ground-glass opacities and non-pathological lymph node, whereas the Longitudinal-CT ground truth only contains malignant lesions. This may require additional filtering steps to classify lesions malignancy and represents an important direction for future work.

Beyond RECIST automation, we believe this work opens a broader perspective for oncology response assessment. Current clinical practice restricts RECIST evaluation to a maximum of five target lesions, which poorly captures the spatial and biological heterogeneity of metastatic disease. A fully automated system capable of tracking the entire detectable tumor burden across timepoints would enable more comprehensive tumor response assessments.

Acknowledgments. This work was performed using HPC resources from GENCI–IDRIS (2026-103370).

Disclosure of Interests. All authors are employees of Guerbet.

References

[1] S. Hermena and M. Young, "CT-scan Image Production Procedures," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Oct. 17, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK574548/

[2] R. Paudyal *et al.*, "Artificial Intelligence in CT and MR Imaging for Oncological Applications," *Cancers*, vol. 15, no. 9, p. 2573, Jan. 2023, doi: 10.3390/cancers15092573.

[3] E. A. Eisenhauer *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *Eur. J. Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009, doi: 10.1016/j.ejca.2008.10.026.

[4] P. Therasse *et al.*, "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J. Natl. Cancer Inst.*, vol. 92, no. 3, pp. 205–216, Feb. 2000, doi: 10.1093/jnci/92.3.205.

[5] R. J. M. Bruls and R. M. Kwee, "Workload for radiologists during on-call hours: dramatic increase in the past 15 years," *Insights Imaging*, vol. 11, no. 1, p. 121, Nov. 2020, doi: 10.1186/s13244-020-00925-z.

[6] Ö. Kasalak, H. Alnahwi, R. Toxopeus, J. P. Pennings, D. Yakar, and T. C. Kwee, "Work overload and diagnostic errors in radiology," *Eur. J. Radiol.*, vol. 167, p. 111032, Oct. 2023, doi: 10.1016/j.ejrad.2023.111032.

[7] Y.-C. Peng, W.-J. Lee, Y.-C. Chang, W. P. Chan, and S.-J. Chen, "Radiologist burnout: Trends in medical imaging utilization under the national health insurance system with the universal code bundling strategy in an academic tertiary medical centre," *Eur. J. Radiol.*, vol. 157, p. 110596, Dec. 2022, doi: 10.1016/j.ejrad.2022.110596.

[8] S. Katal, B. York, and A. Gholamrezanezhad, "AI in radiology: From promise to practice − A guide to effective integration," *Eur. J. Radiol.*, vol. 181, p. 111798, Dec. 2024, doi: 10.1016/j.ejrad.2024.111798.

[9] M. H. Rezazade Mehrizi, P. van Ooijen, and M. Homan, "Applications of artificial intelligence (AI) in diagnostic radiology: a technography study," *Eur. Radiol.*, vol. 31, no. 4, pp. 1805–1811, Apr. 2021, doi: 10.1007/s00330-020-07230-9.

[10] L. Topff *et al.*, "Artificial intelligence-assisted double reading of chest radiographs to detect clinically relevant missed findings: a two-centre evaluation," *Eur. Radiol.*, vol. 34, no. 9, pp. 5876–5885, Sep. 2024, doi: 10.1007/s00330-024-10676-w.

[11] S. L. van Winkel *et al.*, "Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study," *Eur. Radiol.*, vol. 31, no. 11, p. 8682, May 2021, doi: 10.1007/s00330-021-07992-w.

[12] H. M. L. Frazer *et al.*, “Comparison of AI-integrated pathways with human-AI interaction in population mammographic screening for breast cancer,” *Nat. Commun.*, vol. 15, no. 1, p. 7525, Aug. 2024, doi: 10.1038/s41467-024-51725-8.

[13] J. Geppert *et al.*, “Software using artificial intelligence for nodule and cancer detection in CT lung cancer screening: systematic review of test accuracy studies,” *Thorax*, vol. 79, no. 11, pp. 1040–1049, Oct. 2024, doi: 10.1136/thorax-2024-221662.

[14] H. Ying *et al.*, “A multicenter clinical AI system study for detection and diagnosis of focal liver lesions,” *Nat. Commun.*, vol. 15, no. 1, p. 1131, Feb. 2024, doi: 10.1038/s41467-024-45325-9.

[15] C. Abi Nader *et al.*, “Automatic Detection of Pancreatic Lesions and Main Pancreatic Duct Dilatation on Portal Venous CT Scans Using Deep Learning,” *Invest. Radiol.*, vol. 58, no. 11, p. 791, Nov. 2023, doi: 10.1097/RLI.0000000000000992.

[16] A.-I. Iuga *et al.*, “Automated detection and segmentation of thoracic lymph nodes from CT using 3D foveal fully convolutional neural networks,” *BMC Med. Imaging*, vol. 21, no. 1, p. 69, Apr. 2021, doi: 10.1186/s12880-021-00599-z.

[17] S. Gatidis *et al.*, “A longitudinal whole-body CT dataset with manually annotated tumor lesions,” *Sci. Data*, May 2026, doi: 10.1038/s41597-026-07466-y.

[18] M. Baumgartner, P. F. Jaeger, F. Isensee, and K. H. Maier-Hein, “nnDetection: A Self-configuring Method for Medical Object Detection,” vol. 12905, 2021, pp. 530–539. doi: 10.1007/978-3-030-87240-3_51.

[19] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[20] N. Debs, A. Routier, C. Abi-Nader, A. Marcoux, A. Bône, and M.-M. Rohé, “Ensembling Voxel-Based and Box-Based Model Predictions for Robust Lesion Detection,” in *Applications of Medical Artificial Intelligence*, S. Wu, B. Shabestari, and L. Xing, Eds., Cham: Springer Nature Switzerland, 2024, pp. 42–51. doi: 10.1007/978-3-031-47076-9_5.

[21] J. Wasserthal *et al.*, “TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images,” *Radiol. Artif. Intell.*, vol. 5, no. 5, p. e230024, Sep. 2023, doi: 10.1148/ryai.230024.