Journal of Imaging Informatics in Medicine (2025) 38:1552–1562
https://doi.org/10.1007/s10278-024-01304-6

ORIGINAL PAPER

Foundational Segmentation Models and Clinical Data Mining Enable Accurate Computer Vision for Lung Cancer

Nathaniel C. Swinburne[1] · Christopher B. Jackson[2] · Andrew M. Pagano[1] · Joseph N. Stember[1] · Javin Schefflein[1] · Brett Marinelli[1] · Prashanth Kumar Panyam[1] · Arthur Autz[1] · Mohapar S. Chopra[1] · Andrei I. Holodny[1] · Michelle S. Ginsberg[1]

Received: 27 May 2024 / Revised: 10 October 2024 / Accepted: 11 October 2024 / Published online: 22 October 2024

Abstract

This study aims to assess the effectiveness of integrating Segment Anything Model (SAM) and its variant MedSAM into the automated mining, object detection, and segmentation (MODS) methodology for developing robust lung cancer detection and segmentation models without post hoc labeling of training images. In a retrospective analysis, 10,000 chest computed tomography scans from patients with lung cancer were mined. Line measurement annotations were converted to bounding boxes, excluding boxes < 1 cm or > 7 cm. The You Only Look Once object detection architecture was used for teacher-student learning to label unannotated lesions on the training images. Subsequently, a final tumor detection model was trained and employed with SAM and MedSAM for tumor segmentation. Model performance was assessed on a manually annotated test dataset, with additional evaluations conducted on an external lung cancer dataset before and after detection model fine-tuning. Bootstrap resampling was used to calculate 95% confidence intervals. Data mining yielded 10,789 line annotations, resulting in 5403 training boxes. The baseline detection model achieved an internal F1 score of 0.847, improving to 0.860 after self-labeling. Tumor segmentation using the final detection model attained internal Dice similarity coefficients (DSCs) of 0.842 (SAM) and 0.822 (MedSAM). After fine-tuning, external validation showed an F1 of 0.832 and DSCs of 0.802 (SAM) and 0.804 (MedSAM). Integrating foundational segmentation models into the MODS framework results in high-performing lung cancer detection and segmentation models using only mined clinical data. Both SAM and MedSAM hold promise as foundational segmentation models for radiology images.

Keywords Lung neoplasms · Computed tomography · Artificial intelligence · Data mining

Abbreviations

CT	Computed tomography
GPT	Generative pre-trained transformer
GSPS	Grayscale softcopy presentation state
LLM	Large language model
MedSAM	Segment Anything Model for Medical Imaging
MODS	Mining, object detection, and segmentation
MR	Magnetic resonance
PACS	Picture archiving and communication system
SAM	Segment Anything Model
SSL	Semi-supervised learning
YOLO	You Only Look Once

✉ Nathaniel C. Swinburne
swinburn@mskcc.org

[1] Department of Radiology, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, USA

[2] Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, USA

Introduction

Foundational artificial intelligence (AI) models, such as large language models (LLMs) like OpenAI's Generative Pre-trained Transformer (GPT), achieve exceptional performance across a wide array of applications [1], including improving patient care [2] and performing healthcare big data analysis [3]. Modern LLMs have achieved a remarkable capability known as "zero shot" generalization, enabling them to perform a task despite "zero" training data specific to that task. This feat is accomplished through self-supervised learning techniques applied to vast amounts of text data, much of which is publicly available [4–6].

However, unlike text data, medical imaging lacks comparably extensive public datasets, primarily due to concerns surrounding patient privacy, the relative scarcity of medical images in comparison to text repositories, and the need for specialized medical expertise for image annotation. Consequently, there remains a critical need for foundational computer vision models in medical imaging, including radiology models performing computer-aided detection and lesion segmentation models for radiomic analysis and treatment response assessment. These models should be capable of generalizing across diagnostic modalities and resilient to variations in acquisition protocols and radiology scanner parameters.

Previous research has shown that highly effective brain tumor detection models can be developed using "weak" image annotations generated by radiologists within routine clinical workflows [7, 8]. This approach, which relies on data mining instead of expensive de novo image labeling, is scalable and labor efficient. Furthermore, integrating simple unsupervised segmentation techniques, such as Otsu thresholding, with these tumor detection models enables accurate segmentation of enhancing brain tumors, reducing manual annotation time by over 95% compared to fully supervised learning [9]. The success of this data mining, object detection, and segmentation (MODS) methodology for brain magnetic resonance imaging hints at a pathway towards establishing "turnkey" radiology computer vision capabilities. However, to ensure success across diverse diagnostic modalities, including computed tomography (CT) scans, various anatomic regions, and the wide spectrum of tumor appearances—with variations in density, degree of enhancement, and solid versus cystic composition—a generalizable segmentation method must be integrated into the MODS framework.

The Segment Anything Model (SAM), which was trained using the largest known image segmentation dataset in existence (11 million natural scene images and over 1 billion masks), has emerged as a foundational model for general image segmentation [10]. Although SAM was neither trained with nor designed for medical imaging segmentation, it has demonstrated promising results for pathology and radiology images when employed with image region prompts like bounding boxes [11–14]. This success has led to the development of MedSAM [15], a variant model specifically fine-tuned for medical images. SAM-based segmentation thus presents an intriguing potential solution for generalized radiology image segmentation, particularly if integrated with lesion detection models derived from mined clinical radiology data using the MODS methodology.

The purpose of this investigation is twofold. Firstly, it aims to assess the effectiveness of the fully automated MODS methodology in developing computer vision capabilities for chest CT using only mined clinical data. Secondly, it seeks to evaluate the feasibility and performance of SAM-based tumor segmentation within this framework.

Materials and Methods

Approval for this retrospective study was obtained from the local Institutional Review Board, and the requirement for written informed consent was waived. All data storage and handling procedures were conducted in accordance with Health Insurance Portability and Accountability Act regulations.

PACS Data Mining and Curation

From the research database at our institution, we randomly selected 10,000 patients with an ICD-9 diagnosis of primary lung cancer who underwent a non-contrast chest CT from 2012 to 2022. The institution's picture archiving and communication system (PACS; Universal Viewer, version 7; GE Healthcare) was mined to extract the chest CT Digital Imaging and Communications in Medicine (DICOM) images and their accompanying grayscale softcopy presentation state (GSPS) data. All DICOM data were deidentified and stored in our XNAT research server (https://www.xnat.org), which was mounted to our research computing platform for data processing. An in-house software utility was employed to extract all line measurement annotations from the DICOM GSPS files [7].

An automated data curation pipeline was constructed to convert line annotations to bounding boxes by geometrically squaring each line annotation. Overlapping boxes were merged, and boxes less than 1 cm or greater than 7 cm in length were excluded. Since bounding box merging tends to introduce padding around lesions, overestimating their true longest diameter, boxes smaller than 1 cm in length were considered likely too small for reliable detection by a two-dimensional (2D) model due to confounding from partial volume averaging. Boxes larger than 7 cm in length were excluded to remove non-lesion annotations, such as biopsy needle trajectory markup from procedural planning. To mitigate class imbalance, the annotated image set was augmented with randomly selected unannotated images (not containing mined annotations) from the same CT scans as the annotated images. To avoid data leakage, the dataset was randomly split at the subject level (i.e., all scans for a given patient were grouped together) into a training set (90%) and a test partition (10%).

Establishing the Internal Test Dataset

To establish the reference standard internal test dataset, 125 patients were randomly selected from the test partition. Each

image was manually annotated by coauthor NS (10 years of radiology experience) using the VGG Image Annotator application [16] to label all tumors with bounding boxes. For the segmentation task, NS manually segmented all tumors in the test dataset using ITK-SNAP [17]. These images and respective annotations comprised the final internal test dataset used to score the detection and segmentation models.

Visual Inspection of the Mined Annotation Subset

To estimate the effectiveness of the PACS data mining process, 500 annotated images from the training set were randomly selected and reviewed by authors NS and AP (8 years of radiology experience). Each mined annotation was classified as a pulmonary nodule/mass, pleural mass, lymph node, other abnormality, or spurious (false positive) annotation.

Detection Model Training Approach

All CT images were resized to 512×512 pixels and scaled from − 1350 to 150 Hounsfield units to optimize lung tissue display before being converted from DICOM to portable network graphics format. Detection models were trained using the You Only Look Once (YOLO) v8 framework (github.com/ultralytics). YOLO, in contrast to region proposal architectures that require multiple passes through the model to process different sections of a single image, features a unified model that performs the detection and classification of an entire image in a single pass, increasing speed while maintaining accuracy.

To ensure a fair comparison, all detection models were initiated using identical YOLO pretrained weights and trained using a batch size of 200 for a maximum of 500 epochs. Data augmentation methods utilized were random $\pm 10°$ rotations and zooming from 100 to 110%. Initial learning rates and optimizer were determined automatically at train time by the YOLO framework.

Lung Tumor Detection Using Teacher-Student Learning

To address the challenge of incomplete labeling inherent in clinically generated image annotations, where often only the largest cross-section of several index tumors is measured by the radiologist, teacher-student semi-supervised learning was employed. First, a baseline "teacher" model was trained using the mined training data. Subsequently, the teacher model was used to self-label the complete image stacks represented by the mined training images, adding these images and "pseudo-labels" to the original mined training set. This augmented training set was then used to train a final "student" model.

To estimate the performance of the self-labeling, a random sample of 500 self-labeled images underwent review by NS and AP to classify each bounding box generated by the teacher model.

Finally, both the teacher and student tumor detection models underwent evaluation using the held-out test set. For all detection tasks, a true positive was identified if there was a minimum intersection over union of 0.5 with respect to the ground truth box.

Assessing Model Generalizability to External Data

To assess generalizability, the tumor detection models underwent further evaluation using the public Lung Image Database Consortium image collection (LIDC-IDRI) dataset from The Cancer Imaging Archive [18–20]. This dataset comprises 1018 chest CT scans obtained from seven different medical centers for lung cancer screening or diagnosis, accompanied by consensus volumetric tumor segmentations conducted by four radiologists.

The primary objective in utilizing the LIDC-IDRI dataset was to evaluate the performance of the internal detection and SAM-based segmentation models on external data. Since model performance often degrades when applied to out-of-distribution data, we additionally simulated the acquisition of external training data via PACS mining, specifically images and bounding boxes delineating the tumors' maximum cross-section, for fine-tuning the internal detection models. Notably, the LIDC-IDRI ground truth masks were only used for testing, and not training, the segmentation models.

Upon downloading the LIDC-IDRI data, preprocessing was conducted to enable performance comparison with the internal dataset. First, the images were filtered to isolate only the largest cross-sectional image slice through each tumor, thereby mimicking data from a typical clinical workflow obtained through data mining, and tumors less than 1 cm in maximum dimension were excluded. Bounding boxes were automatically generated from the ground truth segmentation masks for use in tumor detection re-training and testing. The tumor-containing images were balanced with randomly selected tumor-free images, and the patients were split randomly into train (90%) and test (10%) subsets.

The teacher and student detection models were scored using the LIDC-IDRI test set. Following this, both detection models underwent fine-tuning using the LIDC-IDRI training set and were subsequently retested.

Evaluating SAM-Based Segmentation of Lung Tumors

To evaluate segmentation performance, the best-performing tumor detection models as assessed on the internal and

external datasets were paired with SAM (SAM large model; github.com/ultralytics) and MedSAM (github.com/bowang-lab/MedSAM). The common architecture used by SAM and MedSAM combines an image encoder with a prompt encoder. The image encoder processes the input image to generate a high-dimensional feature map. The prompt encoder interprets the user-provided prompt and creates an embedding that reflects the prompt's content. These two components are then fused, allowing the model to use the combined information to predict the segmentation mask. In the current implementation, the bounding boxes generated by the tumor detection models served as prompts for the SAM-based segmentation models. Subsequently, the segmentation predictions were compared to the ground truth masks to score the model's performance.

The complete MODS pipeline is illustrated in Fig. 1.

Statistical Analysis

For each model metric, 95% confidence intervals were calculated by bootstrap resampling 2000 times. During each resampling iteration, a sample of images equivalent in size to the full test set was randomly selected with replacement, and each model was scored using this sample. The 2.5th and 97.5th percentile values from the aggregated results represented the 95% confidence interval for the metric.

Results

PACS Data Mining and Curation

The PACS data query yielded a total of 10,789 line annotations extracted from 4536 unique CT scans, resulting in 5891 bounding boxes that met the inclusion criteria (Fig. 2). The combined training and testing datasets comprised 17,591 image slices from 4177 patients, with a mean age of 68.6 ± 12.8 years, including 2301 women and 1876 men. Various lung cancer subtype diagnoses were represented among the patients, with adenocarcinoma being the most prevalent. Detailed patient characteristics are presented in Table 1. Manual review of the randomly selected mined image subset found that a large majority of bounding boxes annotated pulmonary nodules or masses (over 80% combined), while the remaining annotations were primarily distributed between pleural masses and lymph nodes (Table 2).

Lung Tumor Detection Using Teacher-Student Learning

The teacher detection model attained an F1 score of 0.847 (95% CI, 0.812–0.880) on the internal test set. Self-labeling of the full CT scan image stacks (637,170 image

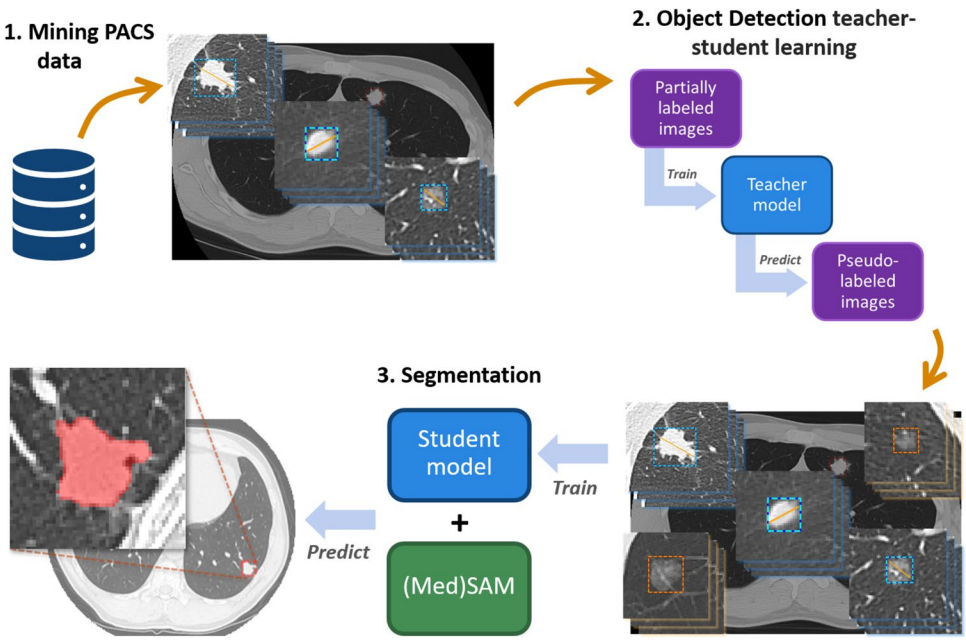

Fig. 1 Schematic of the Mining, Object Detection, and Segmentation (MODS) pipeline. In Step 1, PACS data mining extracts clinical line measurement annotations, which are converted to bounding boxes (blue dashed boxes). Step 2 utilizes teacher-student learning for object detection: a teacher model is first trained using the baseline annotations. The teacher model generates pseudo-labels (orange dashed boxes) for unannotated lesions, which are then combined with the original annotations to train the final student detection model. In Step 3, the trained student detection model is integrated with foundational segmentation methods for tumor segmentation. MedSAM, Segment Anything Model for Medical Imaging. SAM, Segment Anything Model

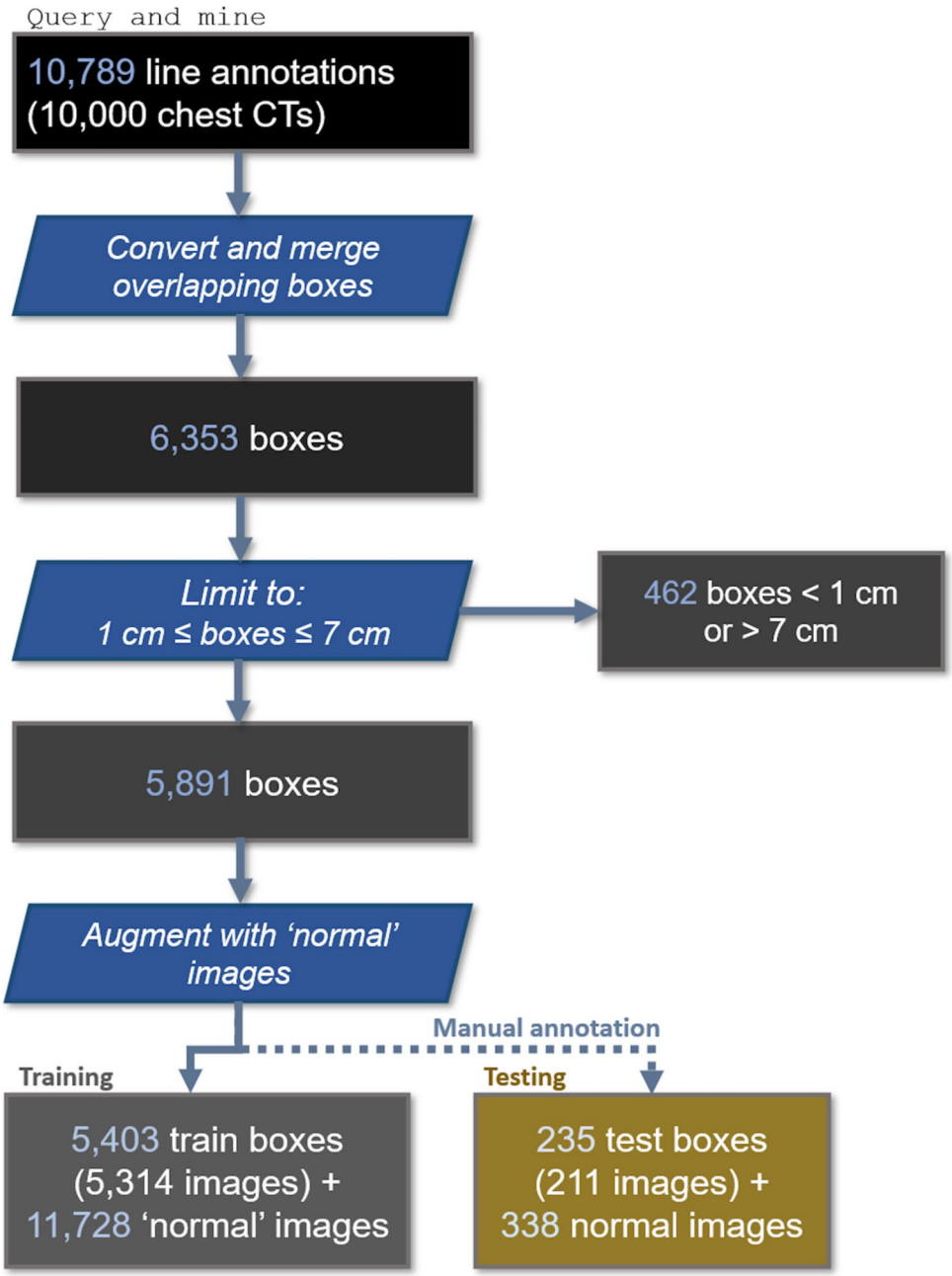

Fig. 2 Flow diagram shows data inclusion and exclusion criteria for the baseline training and test datasets. Additional image annotations were subsequently added to the training dataset via the self-labeling method. The test dataset was manually annotated with bounding boxes and segmentation masks

slices) from the train dataset added 66,993 new bounding boxes for student model training. Manual review of a subset of these self-labeled boxes revealed that most annotated pulmonary nodules or masses (> 70%), with a false positive rate of less than 10% (Table 2). The student detection model slightly outperformed the teacher, attaining an F1 score of 0.860 (95% CI, 0.825–0.893). Examples of false positive and false negative detections for the student model are shown in Fig. 3.

External Validation of Lung Tumor Detection

After pre-processing, the external LIDC-IDRI dataset included 979 image slices (with 541 bounding boxes) for fine-tuning and 115 image slices (with 65 bounding boxes) for testing. The total number of slices (1094) exceeded the number of CT scans in the dataset (1018) because some scans contained more than one tumor. The teacher detection model achieved an initial F1 score of 0.723 (95% CI, 0.649–0.800), which improved to 0.790 (95% CI,

Table 1 Patient characteristics

	Training group	Testing group	Total
Patient demographics			
• No. of patients	4052	125	4177
• Mean age (y)	68.6 ± 12.7	68.6 ± 13.8	68.6 ± 12.8
• Sex	2233 women, 1819 men	68 women, 57 men	2301 women, 1876 men
Lung cancer subtype			
• Adenocarcinoma	2506	70	2576
• Squamous cell carcinoma	454	15	469
• Neuroendocrine	148	6	154
• Small cell carcinoma	126	3	129
• Other, multiple, or unavailable	818	31	849
No. of unique scans	4400	136	4536
Image slice thickness (mm)			
• 5.0	13,378	381	13,759
• 1.25	3640	168	3808
• Other	24	0	24
Mean box length (cm)	2.08 ± 1.09	2.15 ± 1.07	2.08 ± 1.09

Unless otherwise indicated, data are numbers of patients.

Table 2 Results of manual inspection of mined and self-labeled image annotations

Total boxes	Pulmonary nodules/masses	Pleural masses	Lymph nodes	Other	False positives/spurious
Mined image subset (500 images)					
505	Solid: 283 Sub-solid: 130	40	46	6	0
Self-labeled image subset (500 images)					
500	Solid: 253 Sub-solid: 111	42	32	14	48

Results of manual inspection conducted on randomly selected subsets of both mined and self-labeled image annotations. In the mined image subset, other abnormalities consisted of bullae, esophageal masses, and the tracheal lumen diameter. In the self-labeled subset, other abnormalities consisted of bullae, loculated pleural fluid, apical pleural scarring, and regions of fibrosis. False positives observed in this subset were predominantly vessels imaged in cross-section and respiratory motion artifacts.

0.723–0.852) after fine-tuning. As observed with the internal test set, the student model outperformed the teacher, with an initial F1 score of 0.765 (95% CI, 0.679–0.841) that increased to 0.832 (95% CI, 0.764–0.895) with fine-tuning (Table 3). Precision and recall metrics are included in supplemental Tables S-1 and S-2.

Evaluating SAM-Based Segmentation of Lung Tumors

For the tumor segmentation task, the top-performing detection models—the student model for the internal test set and the fine-tuned student model for the external test set—were selected to pair with the SAM-based segmentation models. On the internal test dataset, SAM exhibited a slight advantage over MedSAM, with Dice score coefficients (DSCs) of 0.842 (95% CI, 0.805–0.873) versus 0.822 (95% CI, 0.785–0.851), respectively. On the external LIDC-IDRI test dataset, SAM was marginally outperformed by MedSAM, with DSCs of 0.802 (95% CI, 0.714–0.870) and 0.804 (95% CI, 0.730–0.858), respectively (Table 4).

Representative examples of the highest and lowest accuracy segmentations from the SAM-based models are provided in Fig. 4.

Discussion

In this investigation, we leveraged clinical image annotations from PACS to develop chest CT tumor detection and segmentation models, achieving excellent internal performance without the need for post hoc labeling of training images.

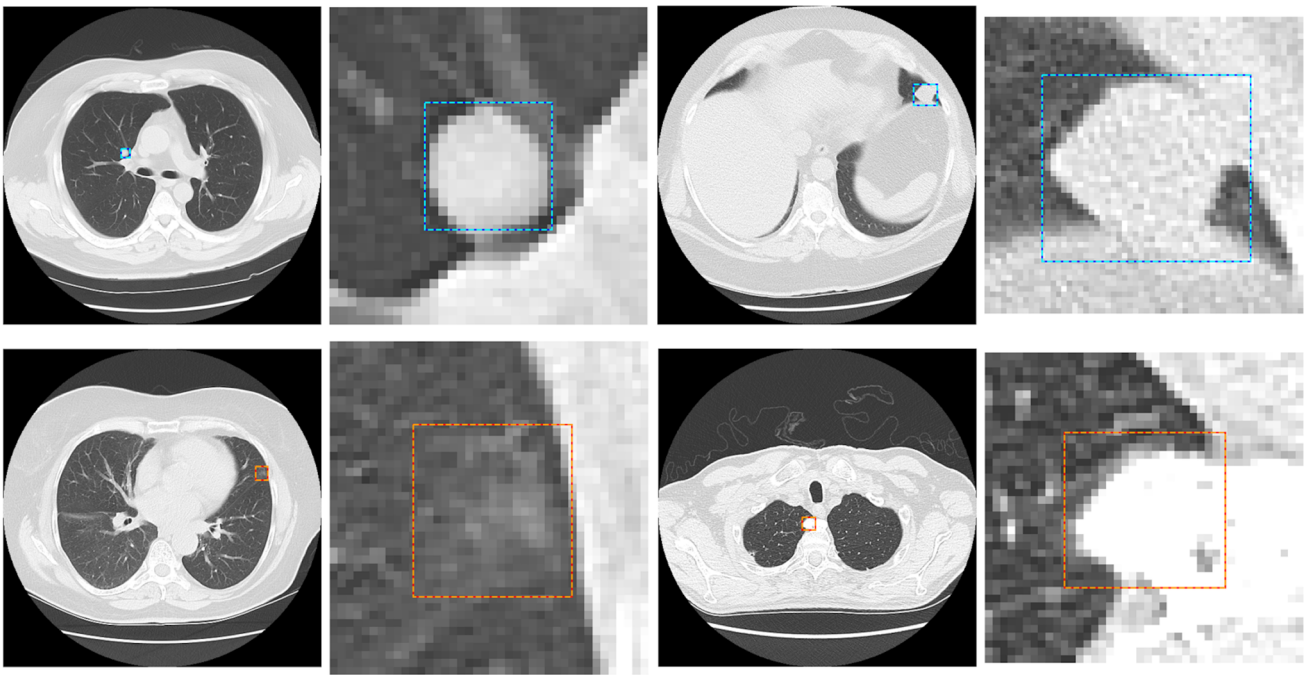

Fig. 3 Representative false negative (cyan/blue boxes) and false positive (orange/red boxes) predictions made by the best-performing (student) tumor detection model, as evaluated on the internal test set. False negatives include a perihilar nodule (top left) and lung base nodule (top right). False positives include a region of focal scarring (bottom left) and vertebral body osteophyte (bottom right)

Table 3 Tumor detection model performance

Model	Internal test set	External LIDC-IDRI test set	
	F1 score (95% CI)	F1 score before fine-tuning (95% CI)	F1 score after fine-tuning (95% CI)
Teacher	0.847 (0.812–0.880)	0.723 (0.649–0.800)	0.790 (0.723–0.852)
Student	**0.860** (0.825–0.893)	0.765 (0.679–0.841)	**0.832** (0.764–0.895)

Detection F1 scores obtained from the teacher and student tumor detection models using both the internal and external test sets. For assessment on the external dataset, the detection models were scored before and after fine-tuning. Values in bold represent the highest performance achieved for each test set.

CI, confidence interval; *LIDC-IDRI*, Lung Image Database Consortium image collection

Table 4 Integrated tumor detection and segmentation model performance

	Internal test set DSC (95% CI)	External LIDC-IDRI test set DSC (95% CI)
Student detection model + SAM	**0.842** (0.805–0.873)	0.802 (0.714–0.870)
Student detection model + MedSAM	0.822 (0.785–0.851)	**0.804** (0.730–0.858)

Segmentation Dice similarity coefficients (DSCs) obtained from the integrated tumor detection and segmentation models using both the internal and external test sets. The top-performing detection models—student model for the internal test set and fine-tuned student model for the external test set—were paired with SAM and MedSAM for segmentation. Bold values indicate the highest performance for each test set.

CI, confidence interval; *MedSAM*, Segment Anything Model for Medical Imaging; *SAM*, Segment Anything Model; *LIDC-IDRI*, Lung Image Database Consortium image collection

Deployed internally for clinical practice, these models could serve a variety of roles such as monitoring the growth kinetics of pulmonary nodules under surveillance, performing radiomic tumor characterization, and assessing treatment response. Although applying these models to external data led to an anticipated performance degradation, a common occurrence in the presence of domain shifts [21], performance notably improved following fine-tuning with a small, weakly annotated external training set of fewer than 1000 images, a quantity readily obtainable through data mining PACS. The resulting segmentation accuracy, with a maximum DSC of 0.804 on the external LIDC-IDRI test set, is comparable to models trained with LIDC-IDRI ground truth segmentation masks [22, 23], which achieve DSCs of 0.777 and 0.822. These findings highlight the effectiveness

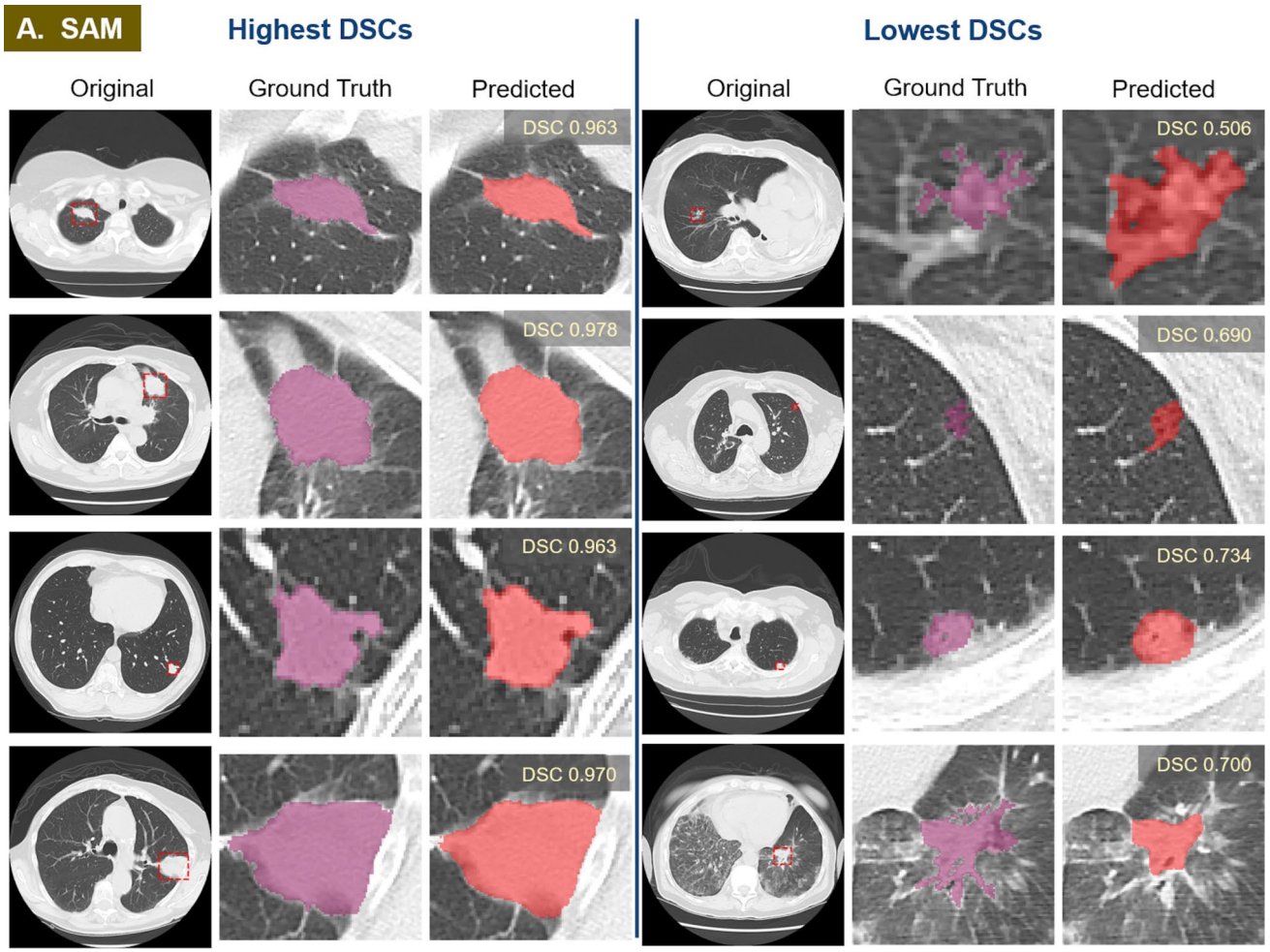

Fig. 4 Representative highest and lowest scoring segmentations generated by SAM (**A**) and MedSAM (**B**) on the internal test set. Red bounding boxes denote the detected tumors and serve as prompts for the segmentation models. In both sets of predictions, the highest scores (left set in both figures) typically encompass solid tumors in the peripheral lungs, with both models adept at distinguishing atelectasis and rib from the lesions. For SAM, the lowest scores (right set of **A**) include, from the top-right downwards, solid nodules with prominent adjacent vasculature, ground glass nodules, pleural lesions, and centrally located lesions with background motion artifact. For MedSAM, the lowest scores (right set of **B**) include solid nodules with prominent adjacent vasculature (first and third from the top), ground glass nodules (second from the top), and centrally located lesions with background motion artifact (bottom). Note is made of the globular morphology of several low-performing segmentation masks (arrows). DSC, Dice similarity coefficient

of the MODS methodology in efficiently developing radiology computer vision models that can adapt to different modalities and anatomical regions, and generalize externally through fine-tuning with mined clinical data from the deploying radiology practice.

The successful integration of SAM-based segmentation into the framework further streamlines the MODS pipeline originally proposed in [9] by eliminating the need for teacher-student learning to train and refine sequential segmentation models. Collectively, this updated methodology offers tremendous potential time savings compared to traditional fully supervised learning methods by avoiding the need for post hoc manual annotation of training images and eliminating the typically lengthy training time for consecutive segmentation models. By incorporating SAM-based segmentation, the challenge of lesion segmentation is effectively reduced to the simpler task of lesion detection, a process facilitated by leveraging image annotations generated from routine clinical workflows. Moreover, as even better foundational segmentation models emerge, they can be seamlessly integrated into this framework.

While MedSAM was specifically trained for medical image segmentation, our findings suggest that it does not necessarily surpass SAM for radiology lesion segmentation. Upon qualitative review of the lowest-scoring test image masks from these segmentation models, SAM appears to perform poorly when normal structures with densities similar to tumors, such as cross-sectioned blood vessels, are

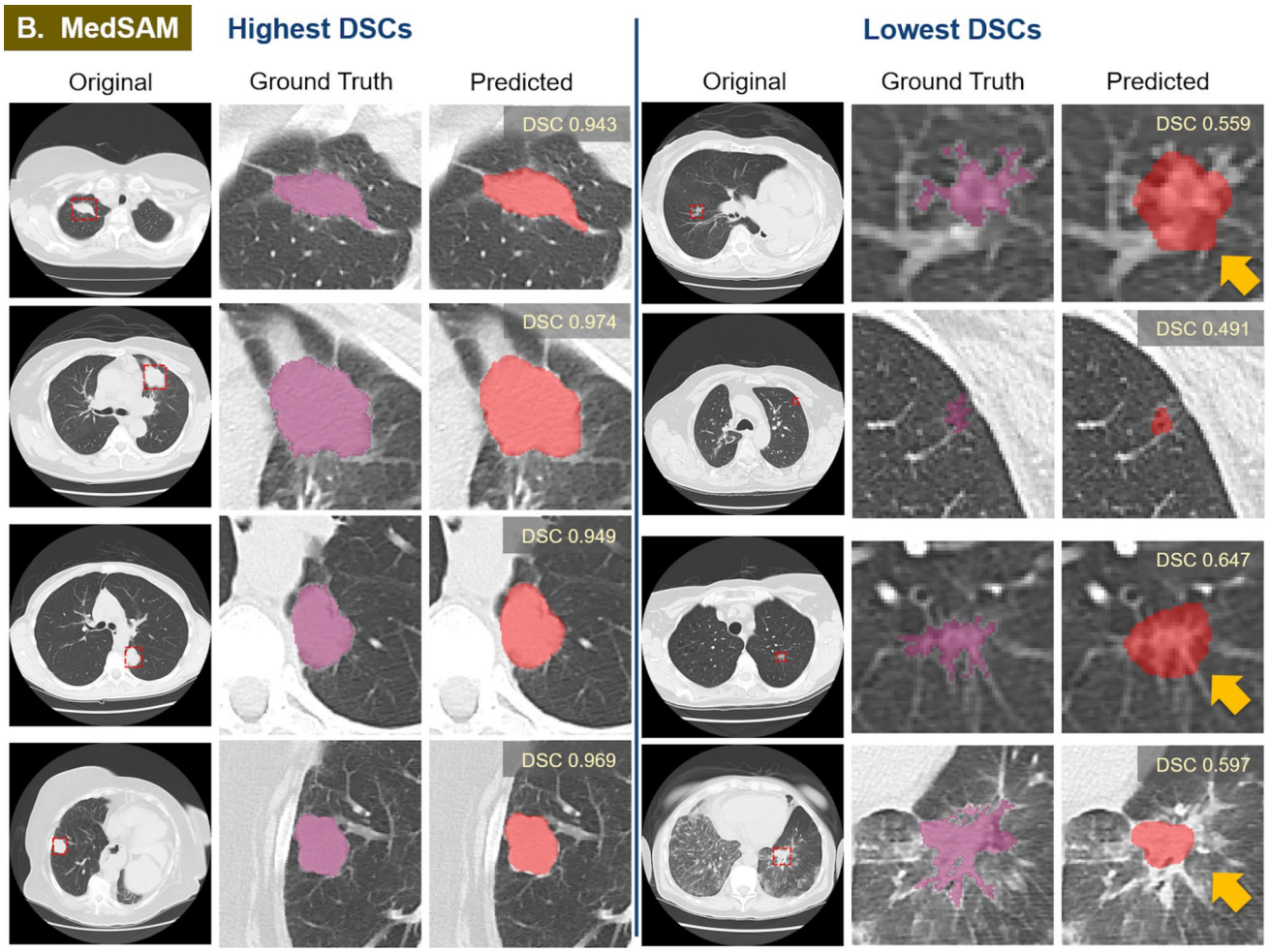

Fig. 4 (continued)

situated adjacent to the tumor. This proximity leads to the erroneous inclusion of the normal structure in the tumor mask. In contrast, the problematic MedSAM segmentations exhibit less intuitive characteristics, featuring masks with globular morphologies centered on the nodule but with seemingly arbitrary boundaries that encompass adjacent aerated lung tissue. Further investigation is warranted to compare the performance of SAM and MedSAM across additional radiology imaging modalities.

There are several potential limitations to this investigation. First, the initial dataset of 10,000 chest CTs produced a relatively small number of bounding boxes for training, which likely constrained the performance of the teacher detection model and subsequently impacted the yield of the self-labeling process. The modest performance improvement of the student model contrasts with previous radiology computer vision studies, where statistically significant gains in student model performance were observed [8, 24]. Ultimately, the pipeline is dependent on the performance of the final detection model, and any false negative or false positive detections will lower the segmentation performance as a result. Nevertheless, the number of annotated images available for training the teacher model could be increased by simply expanding the initial cohort of CT scans or incorporating patients with metastatic lung disease. Importantly, the scalability of this pipeline allows for substantial dataset expansion without requiring de novo manual annotation, offering a path to improved performance through the integration of larger, more diverse datasets. Second, both the YOLO object detection and SAM-based segmentation models utilize 2D architectures, whereas small pulmonary nodules may be difficult to differentiate from normal structures without the benefit of a three-dimensional (3D) context. Future efforts could explore using these 2D models to generate volumetric detection and segmentation training sets for 3D models. Furthermore, the incorporation of additional categories of mined image annotations like arrows, which are frequently used to annotate very small findings, could increase the number of small pulmonary nodule annotations in the training set. Third, the model was trained and tested

exclusively on images from patients with a diagnosis of lung cancer, with the training set balanced between tumor-containing and non-tumor-containing images. Although the background lung parenchyma varied from relatively normal to showing evidence of chronic pulmonary disease, the model's performance on a non-cancer population requires further investigation.

In conclusion, our investigation showcases the development of highly effective pulmonary tumor detection and segmentation models solely from mined clinical data, underscoring the utility of the MODS framework in creating radiology computer vision models across diverse modalities and diseases. Moreover, the integration of foundational SAM-based segmentation streamlines this methodology, expediting radiology model development and enhancing model portability. Looking ahead, future work will explore broader clinical applications, with an emphasis on prospective validation, refinement of the models in diverse clinical settings, and the incorporation of continuous learning to counter data drift and ensure long-term robustness.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-01304-6.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Nathaniel Swinburne, Christopher B. Jackson, Andrew Pagano, and Michelle Ginsberg. The first draft of the manuscript was written by Nathaniel Swinburne, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study has received funding by the National Institutes of Health/National Cancer Institute (Cancer Center Support Grant P30 CA008748).

Data Availability The internal chest CT imaging data used in this study are not publicly available at this time. The external LIDC-IDRI images are available for download from TCIA.

Declarations

Ethics Approval Institutional Review Board approval was obtained.

Consent to Participate Written informed consent was waived by the Institutional Review Board.

Consent for Publication No individual case reports are included.

Competing Interests Nathaniel Swinburne: Research grant from the Innovation in Cancer Informatics (ICI) program; patent issued to Memorial Sloan Kettering Cancer Center related to the use of semisupervised learning for radiology computer vision.

References

1. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. arXiv; 2023. https://doi.org/10.48550/arXiv.2303.08774.
2. Gebrael G, Sahu KK, Chigarira B, et al. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. Cancers. Multidisciplinary Digital Publishing Institute; 2023;15(14):3717. https://doi.org/10.3390/cancers15143717.
3. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature. Nature Publishing Group; 2023;619(7969):357–362. https://doi.org/10.1038/s41586-023-06160-y.
4. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv; 2023. https://doi.org/10.48550/arXiv.1706.03762.
5. Common Crawl - Open Repository of Web Crawl Data. . https://commoncrawl.org/. Accessed February 14, 2024.
6. Gao L, Biderman S, Black S, et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv; 2020. https://doi.org/10.48550/arXiv.2101.00027.
7. Swinburne NC, Mendelson D, Rubin DL. Advancing Semantic Interoperability of Image Annotations: Automated Conversion of Non-standard Image Annotations in a Commercial PACS to the Annotation and Image Markup. J Digit Imaging. 2020;33(1):49–53. https://doi.org/10.1007/s10278-019-00191-6.
8. Swinburne NC, Yadav V, Kim J, et al. Semisupervised Training of a Brain MRI Tumor Detection Model Using Mined Annotations. Radiology. Radiological Society of North America; 2022;210817. https://doi.org/10.1148/radiol.210817.
9. Swinburne NC, Yadav V, Murthy KNK, et al. Fast, light, and scalable: harnessing data-mined line annotations for automated tumor segmentation on brain MRI. Eur Radiol. 2023;33(9):6582–6591. https://doi.org/10.1007/s00330-023-09583-3.
10. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.02643.
11. Deng R, Cui C, Liu Q, et al. Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.04155.
12. Roy S, Wald T, Koehler G, et al. SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.05396.
13. Hu C, Xia T, Ju S, Li X. When SAM Meets Medical Images: An Investigation of Segment Anything Model (SAM) on Multi-phase Liver Tumor Segmentation. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.08506.
14. He S, Bao R, Li J, et al. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.09324.
15. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. Nature Publishing Group; 2024;15(1):654. https://doi.org/10.1038/s41467-024-44824-z.
16. Dutta A, Zisserman A. The VIA Annotation Software for Images, Audio and Video. Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery; 2019. p. 2276–2279. https://doi.org/10.1145/3343031.3350535.
17. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage. 2006;31(3):1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.
18. Armato SG, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38(2):915–931. https://doi.org/10.1118/1.3528204.

19. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26(6):1045–1057. https://doi.org/10.1007/s10278-013-9622-7.
20. Armato III SG, McLennan G, Bidaut L, et al. Data From LIDC-IDRI. The Cancer Imaging Archive; 2015. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.
21. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology: Artificial Intelligence. Radiological Society of North America; 2022;4(3):e210064. https://doi.org/10.1148/ryai.210064.
22. Shuo Wang null, Mu Zhou null, Gevaert O, et al. A multi-view deep convolutional neural networks for lung nodule segmentation. Annu Int Conf IEEE Eng Med Biol Soc. 2017;2017:1752–1755. https://doi.org/10.1109/EMBC.2017.8037182.
23. Wang S, Zhou M, Liu Z, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Med Image Anal. 2017;40:172–183. https://doi.org/10.1016/j.media.2017.06.014.
24. Lin E, Yuh EL. Semi-supervised learning for generalizable intracranial hemorrhage detection and segmentation. Radiology: Artificial Intelligence.