

Rethinking Lung Cancer Screening: AI Nodule Detection and Diagnosis Outperforms Radiologists, Leading Models, and Standards Beyond Size and Growth

Sylvain Bodard^{1,2,3,4,+}, Pierre Baudot^{5,*,+}, Benjamin Renoust⁵, Charles Voyton⁵, Gwendoline De Bie⁵, Ezequiel Geremia⁵, Van-Khoa Le⁵, Danny Francis⁵, Pierre-Henri Siot⁵, Yousra Haddou⁵, Vincent Bobin⁵, Jean-Christophe Brisset⁵, Carey C. Thomson^{6,7}, Valérie Bourdès⁵, and Benoît Huet⁵

¹Université de Paris Cité, AP-HP, Hôpital Universitaire Necker Enfants Malades, Service d'Imagerie Adulte, F-75015, Paris, France.

²Memorial Sloan Kettering Cancer Center, Department of Radiology, 1275 York Avenue, New York, NY 10065, USA.

³Massachusetts General Hospital, Center for Transplantation Sciences, Harvard Medical School, Boston, USA.

⁴Sorbonne Université, CNRS UMR 7371, INSERM U 1146, Laboratoire d'Imagerie Biomédicale (LIB), F-75006, Paris, France.

⁵Median Technologies, Eyronis, Valbonne, 06560, France.

⁶Mount Auburn Hospital/Beth Israel Lahey Health, Cambridge MA, USA.

⁷Harvard Medical School, Boston MA, USA.

*Corresponding Author: pierre.baudot@mediantechnologies.com

+These authors contributed equally to this work.

ABSTRACT

Early detection of malignant lung nodules is critical, but its dependence on size and growth in screening inherently delays diagnosis. We present an AI system that redefines lung cancer screening by performing both detection and malignancy diagnosis directly at the nodule level on low-dose CT scans. To address limitations in dataset scale and explainability, we designed an ensemble of shallow deep learning and feature-based specialized models. Trained and evaluated on 25,709 scans with 69,449 annotated nodules, the system outperforms radiologists, Lung-RADS®, and leading AI models (Sybil, Brock, Google, Kaggle). It achieves an area under the receiver operating characteristic curve (AUC) of 0.98 internally and 0.945 on an independent cohort. With 0.5 false positives per scan at 99.3% sensitivity, it addresses key barriers to AI adoption. Critically, it outperforms radiologists across all nodule sizes and stages—excelling in stage 1 cancers—and all growth-based metrics, including the least accurate: Volume-Doubling Time. It also surpasses radiologists by up to one year in diagnosing indeterminate and slow-growing nodules.

Introduction

Lung cancer is the leading cause of cancer deaths worldwide^{1,2}. Early detection through low-dose computed tomography (LDCT) dramatically improves prognosis, enabling curative treatments³⁻⁵, reducing mortality by over 20% and tripling early-stage diagnoses, as demonstrated in pioneer National Lung Screening Trials such as NLST (US), NELSON (Netherlands-Belgium), and PanCan (Canada)⁶⁻⁸. As Lung Cancer Screening (LCS) programs are expanding globally⁹⁻¹², new alarming trends emerged from Asian studies, with lung cancer comprising 60% of new cancer cases¹³.

Lung-RADS® guidelines and standard care rely on nodule size as the primary malignancy predictor, with the PanCan model confirming its dominance among predictive features^{14,15}. Temporal size evolution also plays a crucial role, with Volume-Doubling Time (VDT) considered the “method of choice” for malignancy assessment¹⁶⁻¹⁸ and currently recommended in Europe^{19,20}. However, there is no consensus over VDT measure¹⁴ and evidence for the longitudinal relevance of volumetric over diameter-based measures is still awaited²¹ (see Creamer *et al.*²² for instantaneous measure). Advances in quantitative imaging further reinforce its importance in clinical practice, as highlighted by Larici *et al.*, who stress that “size still matters”^{14,23}.

However, while morphological features contribute to malignancy prediction, their subjective nature and variability remain significant challenges^{24,25}. Clinical trials from the UK and France highlight size and location as key factors in radiologist interpretation^{26,27}.

This continued reliance on size delays diagnosis until they meet predetermined criteria, i.e. size and growth thresholds and

protocol-driven follow up over several months (as per NLST NELSON, LungRADs). This study focuses on the early detection and characterization of solid or part-solid parenchymal nodules (4–30 mm)¹¹, as timely intervention for these lesions is critical to improving survival outcomes²⁹.

AI models—particularly deep neural networks—extract high-dimensional, multi-parametric features, including morphology, that are crucial for detecting early-stage malignancies, when nodules are smallest and treatment most effective. Lung cancer is a central focus of AI-driven diagnostic research owing to the disease’s global prevalence and the availability of open-access imaging data. To date, hundreds of AI models based on CT scans have been developed, with as many as 405 publications reviewed^{30,31}. However, most models rely on a weak ground truth of radiologist assessments and suffer from an under-representation of early-stage cancer cases²¹, as is the case with the LIDC dataset³². Consequently, their effectiveness is inherently limited compared to models trained on histopathology-confirmed diagnoses, such as NLST (as verified by Ma *et al.*³³). Additionally, most computer-aided models are limited in scope, focusing exclusively on either nodule detection (CADE) or diagnosis (CADx), thereby limiting their clinical utility. To address this, we compare our model to Sybil²⁵, Liao *et al.* (Kaggle Data Science Bowl winner; 1,972, teams; \$1M prizes)³⁵, and Ardila *et al.*⁹—all trained on NLST and with proven superior performance in lung screening populations³⁷. We also benchmark against the NLST Brock-Pancan model^{15,38}, recommended by the British Thoracic Society. Unlike our model, however, it is a purely diagnostic tool (CADx), requiring radiologists to detect nodules and assess malignancy-related features such as size, attenuation, and spiculation. Despite AI advancements, clinical adoption of CAD systems remains hindered by high false positive (FP) rates, a major clinical concern³⁹. Recent studies emphasize that “increased FP nodules remain a serious drawback”⁴⁰ and stress the need to improve AI specificity⁴¹. Furthermore, models by Liao *et al.*, Ardila *et al.*, and Sybil provide only patient-level malignancy predictions, failing to provide explicit, directly actionable and explainable CADE/CADx-based nodule-level assessments—still lacking in published lung CAD studies at the exception of Jia *et al.* recent study⁴². This prevents direct comparison of detection performance at nodule level. To address this gap, we compare our model to nnDetection, a self-configuring CAD model based on the nnU-Net design principles for detection tasks. nnDetection has demonstrated superior CADE performance, notably outperforming competitors in the LUNA16 nodule candidate-detection challenge¹⁵.

Results

Model and radiologist performances

We evaluated the ability of our model to detect and characterize malignant lung nodules using two independent test sets: NLST (Test1, $n = 2159$); and an Independent Cohort (IC, $n = 273$). Fig. 1 provides model evaluation through Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), sensitivity, specificity, and accuracy at the Maximum Youden Index (MYI). The model achieved a patient-level AUC of 0.98 on Test1 (Confidence Intervals (CIs) provided in figures). Malignancy likelihood distributions for cancer and non-cancer patients are depicted in Supplementary Fig. 1.

Validation on IC (91 cancer, 182 non-cancer) yielded an AUC of 0.945. Subgroup analysis (Fig. 2) demonstrated stable performance across manufacturers, kernel sharpness, slice thickness, age, and sex. However, the EU subgroup had a lower AUC (0.874) than the US subgroup (0.969), closely matching NLST (US) results. This likely reflects the EU cases originating from a Chronic Obstructive Pulmonary Disease (COPD)-only LCS campaign (AIR¹⁰), a factor known to impair model performance⁴⁵ as corroborated here. The over-representation of GE-manufacturer scans in the EU/COPD subgroup (72%, versus 8% in IC US) further explains weaker performance in this subset.

On Test3 ($n = 481$)—a subset of Test1 with four reads by 20 radiologists of varying levels of seniority, each with at least one year of experience interpreting chest CT scans (cf. Supplementary Material), hereafter “4-radiologists”, assessing nodule detection and malignancy likelihood—the model achieved an AUC of 0.975, significantly surpassing the 4-radiologists’ mean AUC of 0.89, ($p < 0.0001$, see Methods). Individual radiologist performance (Fig. 3, Supplementary Fig. 2) ranged from 0.844 to 0.940 (mean: 0.906 ± 0.028), whereas our model consistently outperformed each of them (mean: 0.974 ± 0.016). Welch’s t-test confirmed non-inferiority to each individual radiologist ($p < 0.0001$). Readers 13 to 20 had insufficient sample sizes for comparison.

On Test4 ($n = 460$)—a subset of Test1 with Lung-RADS® assessments—our model achieved an AUC of 0.976, significantly outperforming radiologists’ mean Lung-RADS® assessment without prior CT scans (AUC=0.885, $p < 0.0001$). Similar superiority was observed on Test5, where prior CT imaging was available (AUC=0.967 vs. 0.893, $p < 0.0001$). As anticipated, Lung-RADS® assessments without prior scans approximated radiologist performance, reflecting the Lung-RADS® standardize the methodology used by radiologists. Tables e and f of Fig. 1 demonstrate that our model outperformed Lung-RADS® scores without and with prior imaging in terms of specificity and sensitivity across Test4 and Test5 ($p < 0.0001$), except for 4B sensitivity score with prior imaging.

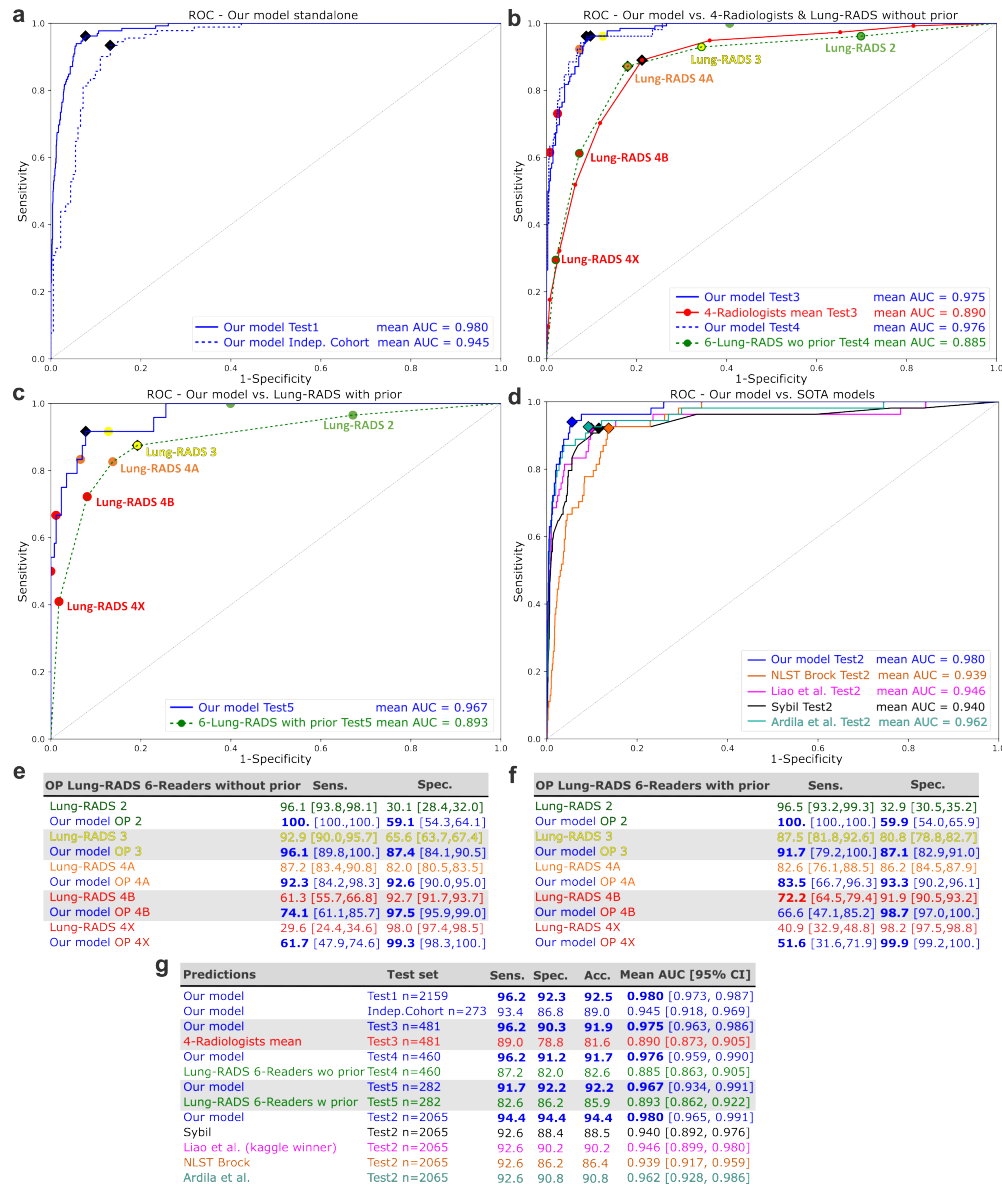


Figure 1. Performance comparison: our model vs. radiologists, lung-RADS®, and SOTA models: **a.**, Patient-level ROC curves for our model’s malignancy prediction on Test1 and Independent Cohort. Our model’s Operating points (OPs) at the Maximum Youden Index (MYI) are depicted with black tilted squares (in all panels of all figures except in **d.** where their colour match the colour of their corresponding curve). **b.**, Patient-level ROC curves for our model’s malignancy predictions on Test3 and Test4, compared with the mean of 4-Radiologists detection and Likelihood Of Malignancy assessment on Test3, as well as of six radiologists Lung-RADS® 2019 score assessments by six radiologists without prior Time Point (TP) evolution on Test4, as provided by Ardila *et al.*⁹. Our model’s OPs equivalent to each Lung-RADS® score are depicted with coloured circles matching the Lung-RADS® scores on the corresponding curves. **c.**, Patient-level ROC curves for malignancy prediction comparing our model and Lung-RADS® 2019 score assessments by six radiologists with prior TP evolution on Test5 as provided by Ardila *et al.*⁹. Same OP representation as in **b.** **d.**, Patient-level ROC curves on Test2 (Sybil and Ardila *et al.* test set), comparing our model, Sybil (five ensembled model), Liao *et al.*, Ardila *et al.* models, and the NLST Brock models using NLST GT for each nodule detected by NLST radiologists. **e.**, the sensitivity and specificity with 95% CI over 5,000 bootstraps for each Lung-RADS® score without prior images, as assessed by six radiologists and for the corresponding accuracy equivalent OP for our model on Test4 (associated to the OPs in **b.**). **f.**, same as **e.**, but for Lung-RADS® with prior on Test5 (associated to the OPs in **c.**). **g.**, Sensitivity, specificity and accuracy at the MYI of each ROC, along with the mean AUC with 95% CI over 5,000 bootstraps.

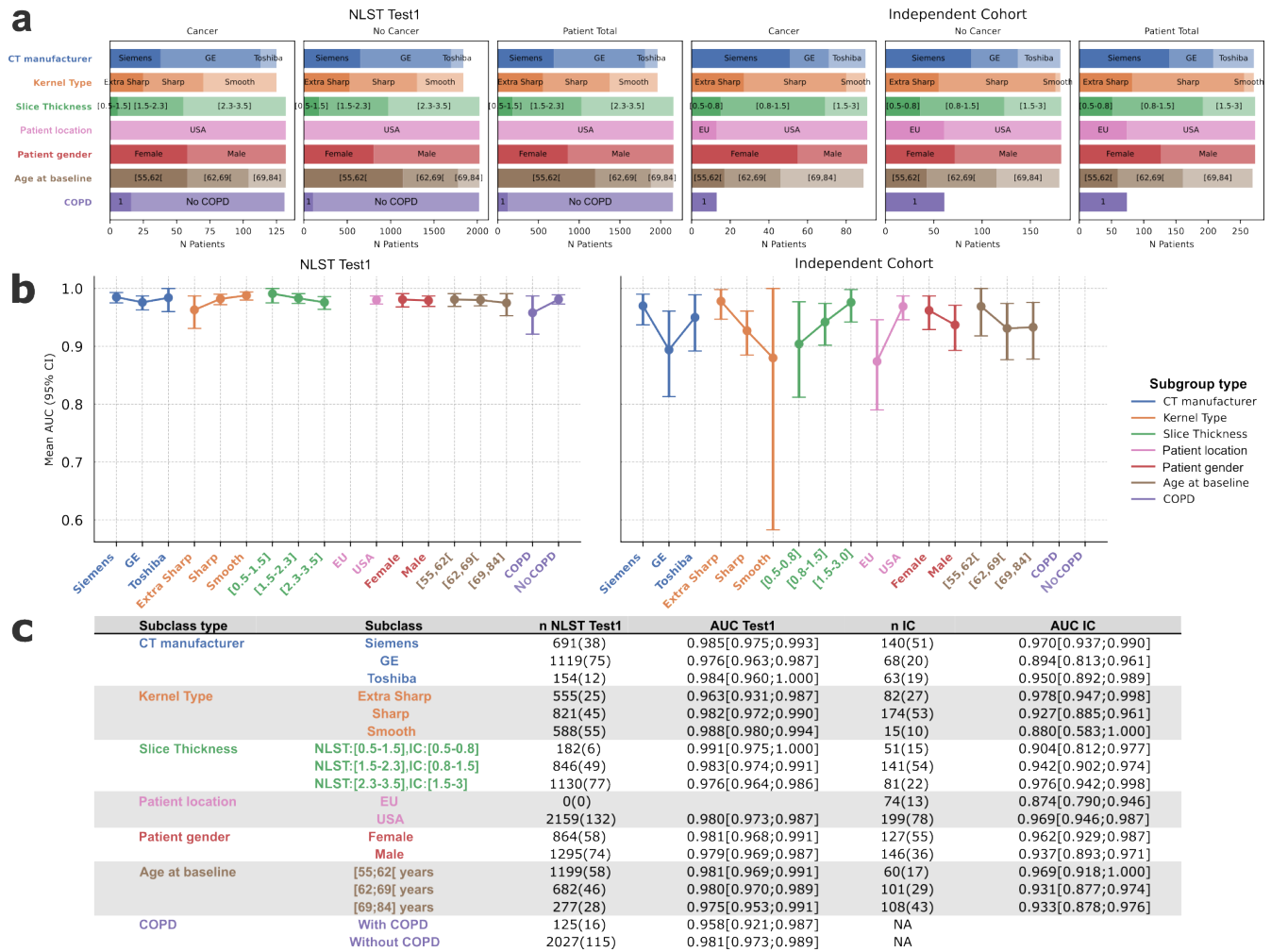


Figure 2. Subgroup demographics and model performance analysis: Subgroup definitions are detailed in Supplementary methods. **a.** The distribution of demographic and scan characteristics in Test1 and IC for cancer, non-cancer and all patients. Sample sizes for Canon and Philips manufacturers are too small to be represented and are therefore replaced by white space. **b.** Mean patient AUC for the various subgroups of our model on Test1 and IC. Vertical bars represent the 95% CI on 5,000 bootstraps. **c.** Table summarizing the values from a. and b. The sample size (n: number of patients) for each subgroup is given with the number of patients with cancer indicated in parenthesis. It provides the mean AUC and 95% CI over 5,000 bootstraps samples. 'NA' stands for 'Not Available': the presence of COPD at baseline was only available for the EU/AIR subset of IC and is 100% ($n = 273(91)$), inclusion criterion).

On Test2 ($n = 2065$)—a subset of Test1 that overlaps with Sybil's and Ardila et al's test sets—our model achieved an AUC of 0.980, significantly outperforming Sybil's AUC of 0.94 ($p < 0.0001$)²⁵. Our model also significantly outperformed Liao *et al.* (AUC=0.946, $p < 0.0001$)³⁵. The hitherto best state-of-the-art (SOTA) model by Ardila *et al.*⁹, achieved an AUC of 0.962 on Test2, significantly below ours ($p < 0.0001$). Moreover, Ardila *et al.*'s model outperformed Sybil and Liao *et al.*

Most published models focus on characterization (CADx) rather than detection (CAdE), often relying on the LIDC dataset³² (16 references in Ma *et al.*³³), where malignancy assessments are radiologist-based rather than biopsy-confirmed. Given our model's superior performance over radiologist assessments (Fig. 1), it inherently outperforms models trained on weak radiological Ground Truth (GT).

We further compared our model to the widely used NLST Brock model^{15,38}. This CADx models require radiologists to detect nodules and assess up to 11 nodule and patient-based features. Our model requires no radiologist feature assessments and significantly outperformed Brock (AUC=0.939, $p < 0.0001$).

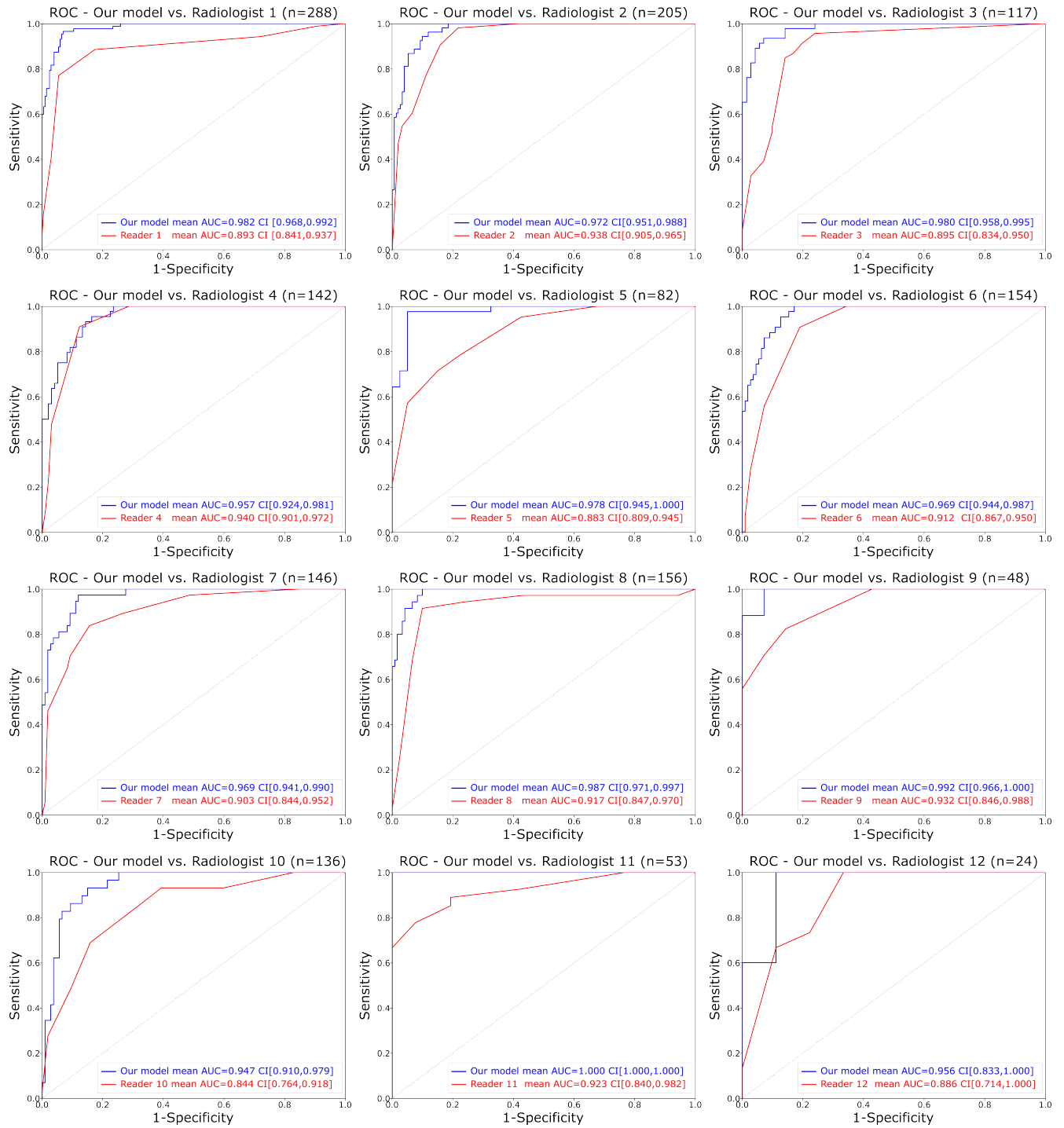


Figure 3. Performance comparison: our model vs. individual radiologists: patient-level ROC curves for each of the 12 radiologist who annotated at least 15 patients with cancer and for our model on the same annotated sample of Test1, as detailed in Supplementary methods (the sample size is provided in the title of each ROC). Their mean AUC and CI over 5,000 bootstraps are indicated in the labels, and reported in detail in Supplementary Fig.2.

The performances of Sybil, Ardila *et al.* and NLST Brock were slightly higher than the originally reported values (0.92²⁵, 0.959⁹ and 0.912³⁸, respectively).

Model and radiologist detection performances

This section details the detection performance of our integrated CAdE/CAdx model and demonstrates how it effectively mitigates the high FP rate challenge outlined in the Introduction.

At the nodule level, we compared our model's performance to radiologists in detecting malignant nodules using the Free-response Receiver Operating Characteristic (FROC) curve, which measures sensitivity as a function of the mean number of FPs per scan and for Test1 and IC. Fig. 4a,b,d illustrates the FROCs and sensitivity values closest to 0.5 FP/scan and 1 FP/scan.

On Test1, the model achieves 99.3% sensitivity at 0.5 FP/scan and 1 FP/scan. Performance drops slightly on the IC, with 90.2% sensitivity at 0.5 FP/scan and 92.6% at 1 FP/scan. For comparison, we include the FROC of the nnDetection model¹⁵, re-trained on Train1 to detect malignant and benign nodules as separate targets (CAdE/CAdx). As illustrated in Fig. 4a, the nnDetection FROC remains below ours due to CAdx module's effectiveness in reducing FPs. On Test1, nnDetection achieved mean sensitivities of 84.9% at 1 FP/scan, and 82.0% at 0.5 FP/scan—both significantly lower than our model ($p < 0.0001$).

Traditional CAdE systems focus on nodule detection rather than malignancy classification, leading to significantly higher FP rates than our integrated CAdE/CAdx model. This disparity arises because benign nodules are considerably more prone to detection errors than malignant ones, as illustrated in Fig. 4c. Direct evaluations of our model's CAdE component, alongside of nnDetection—retrained on NLST for separate benign and malignant detection and nnDetection trained by Baumgartner *et al.*¹⁵ on LUNA16—for single-class nodule detection, further substantiate this observation (Fig. 4a). Furthermore, the nnDetection model trained on LUNA16 by Baumgartner *et al.* (single-class nodule detection) outperforms the retrained nnDetection on NLST for malignant and benign nodule as separate class after joining the classes, likely reflecting the added complexity in distinguishing benign from malignant nodules within a finite model size.

Our model's CAdE component continues to outperform both nnDetection-based CAdE models as a result of its large ensembling strategy while concurrently exhibiting much higher FP rates than its CAdE/CAdx counterparts. As expected, the performance of nnDetection on NLST is markedly lower than on LUNA16¹⁵, since LUNA16 official evaluation includes numerous suspicious findings that are not penalized as FP if detected, thus artificially inflating performance. Accordingly, Fig. 4d shows both our CAdE and LUNA16-trained nnDetection perform strongly on the LUNA16 task, matching previous reports of “outperforming all previous methods on the nodule-candidate-detection task”¹⁵. On this set, nnDetection showed a slight performance advantage over our CAdE model, likely because the former was specifically trained on LUNA16, whereas ours was trained on NLST.

Our CAdE/CAdx also surpassed radiologists on the same task (Fig. 4b). On Test3, at 0.5 and 1 FP/scan, the mean sensitivity of 4-Radiologists was 82.4% and 88.4%, respectively, significantly below our model's 98.5% and 99.3% ($p < 0.0001$). Supplementary Fig. 3&4 further demonstrate that our model's FROC outperforms each radiologist on their annotated subsets. Radiologists failed to detect 8.2% of malignant nodules (mean diameter 7.3 mm $\pm$ 4.1 s.d), all detected by our model.

Beyond size with AI predictions

The correlation between our model's predictions and nodule size is moderate (Pearson $\rho = 0.499$), as evidenced by the joint-distribution of diameter and malignancy prediction for detections in Supplementary Fig. 5. Marginal distributions of model-estimated diameter, alongside malignancy predictions for malignant and benign detections, highlight that size alone is a weak discriminator between malignancy and benignity, with complete overlap in size but only partial overlap in malignancy likelihood. Using diameter alone yields an AUC of 0.832 (CI[0.803,0.859]), significantly lower than our model's 0.996 (CI[0.995,0.998]), corroborating the recent results of Creamer *et al.* who reported an AUC of 0.822 (CI[0.787,0.857]) for diameter on the SUMMIT Study²².

At the patient level, Fig. 5a presents ROC curves for cancer cases with largest malignant nodule in the [4, 10]mm range ($n = 11$), and for non-cancer patients with largest nodule in the same range on Test3 ($n = 225$). For these patients with small nodules, our model achieves an AUC of 0.960, significantly outperforming 4-Radiologists (AUC=0.773, $p < 0.0001$). Fig. 5c extends this comparison across GT diameter ranges ([4, 10]mm, [10, 20]mm, and [20, 30]mm), with the model significantly outperforming radiologists across all subgroups ($p < 0.0001$).

Fig. 5d expands this analysis at nodule level on Test3, comparing malignancy detection and classification performance by the 4-Radiologists and our model. Diameters were measured from their respective segmentations. Across all diameter ranges, the model's AUC remains stable, maintaining strong discriminative power even for small malignant nodules and significantly outperforming radiologists across all size ranges ($p < 0.0001$).

As recommended by Lancaster *et al.*²¹—and supported by the cancer stage distribution in Supplementary Fig.9—NLST mainly includes early-stage cancers. To evaluate early-stage cancers performance, Fig.5e presents AUC for stage 1A ($n = 81$, see also Fig.5a), stage 1 (1A+1B, $n = 95$) and later stages in Test1. The model achieved AUCs of 0.972 for stage 1A, 0.975 for stage 1, and 0.975 for later stages in Test3 (0.979, 0.981, 0.978 in Test1), significantly outperforming the mean of 4-Radiologists ($p < 0.0001$).

Further analyses examined benign nodules in the [4, 10[mm range that underwent unnecessary invasive procedures (e.g., biopsy-confirmed benign cases), constituting radiologist-induced FPs, with considerable patient risks and healthcare burden. In Test1, nine benign nodules meet this criterion. Using the MYI threshold, the model would have correctly classified six of these (67%), thus holding promise for preventing unnecessary interventions.

Fig.5f displays CT patches of all 19 malignant nodules within the [4, 10[mm range from Test1, with their automatic segmentations generated by our model. Four nodules fall below the MYI threshold, and are thus misclassified as False Negatives at this OP. Notably, three are Solid Pleura-Attached Nodules (SPANS), a diagnostically challenging class often considered as benign^{19,46}. Jiang *et al.*⁴⁷ describe SPANS as rarely detected by radiologists, particularly when small. Our model successfully detects these nodules, but underestimates malignancy likelihood for some of them.

Among the 15 correctly classified malignant nodules, all were solid or part-solid solitary nodules, with one exhibiting spiculation. Of these, five were SPANS, all correctly identified by the model, indicating only partial misclassifications (3/9, 33%).

Longitudinal predictions: beyond size growth

Nodule size growth is the key diagnostic feature, with Volume-Doubling Time (VDT) widely used to quantify changes^{17,18,48,49}, notably in the NELSON screening protocol³, which classifies nodules using 400- and 600-day VDT thresholds. Jennings *et al.*⁵¹ introduced Reciprocal Doubling Time ($RDT=365/VDT$), strictly equivalent to Volume Growth (%) in AUC performance and increasing with suspicion of malignancy. Other standard measures include diameter change ($\Delta\text{diameter} (D_{-1} - D_{-2})$), used by Lung-RADS^{®14}, and volume change ($\Delta\text{volume} (V_{-1} - V_{-2})$).

The results above summarize AI predictions at a single time point (TP): either the scan closest to cancer diagnosis within 12 months (T_{-1}) or the earliest available scan for non-cancer patients. To further assess predictive performance at earlier TP T_{-2} (12–24 months before cancer diagnosis, when available), following Sybil and Ardila *et al.*^{9,25} or the earliest pairs of consecutive available scan for non-cancer patients, we analyzed its predictive evolution alongside standard growth measures.

Fig.6a compares the ROC curves of our model's "instantaneous" prediction at T_{-1} , the evolution of its prediction (using D'Arcy Thompson's growth function^{55,56} p.160), and three standard growth measures (Δvolume , $\Delta\text{diameter}$, RDT) for all nodules longitudinally paired by radiologists in GT on Test6.

The best-performing function for prediction evolution follows D'Arcy Thompson's autocalytic model (logistic) of biological growth: $P_{-1}/(P_{-1} - P_{-2})_n$, both outperforming simple ratios or Δ -based functions. Surprisingly, this function remains marginally inferior to the instantaneous prediction at T_{-1} . Our model's instantaneous and evolution predictions significantly outperform all standard size growth measures ($p < 0.0001$), with AUCs of 0.992 and 0.989, respectively. Supplementary Fig.6 illustrates that the growth of model prediction yields less overlap between benign and malignant nodule distributions than traditional metrics. RDT performs worst with an AUC of 0.924, whereas Δvolume , as expected, slightly outperforms $\Delta\text{diameter}$ (AUC=0.969), confirming that accurately measured size growth is the primary indicator of malignancy¹⁸. Fig.6b shows VDT's non-linearity requires either median-centering correction⁵³, or pre-selecting nodules with $\geq 25\%$ growth, as in NELSON⁵². Even after median correction, VDT performed weakest (AUC=0.901); dropping to 0.441 uncorrected. Limiting to nodules with $\geq 25\%$ growth boosts AUC to 0.795, exceeding the 0.67 reported by⁵². Fig.6b reveals that normalizing VDT or Δvolume to daily precision exerts negligible non-significant effects on AUC compared to year-precision normalization used for Δvolume or $\Delta\text{diameter}$ ($p > 0.1$). However, Fig.6c reveals ranking differences of up to 40% between Δ and volume ratios, e.g., a nodule ranked first over 1,000 by volume ratio may rank 400th by Δvolume (-40%). This ranking discrepancy explains the performance gap between RDT/VDT and Δvolume .

To assess clinical impact, we computed the same ROC curves (Fig.6d) following the NELSON protocol, which prioritizes the largest nodule at the initial TP for size growth analysis and our model's patient-level predictions. Patient-level results mirror with nodule-level findings—except RDT slightly outperforming $\Delta\text{diameter}$, reflecting size selection errors—while the gap between AI predictions and standard measures widens (AUC=0.969 for AI vs. 0.81 for RDT).

Our RDT estimates for malignant nodules strongly correlate with Jennings *et al.*⁵¹ (Pearson $\rho = 0.908$, $p < 0.0001$;

Supplementary Fig.7), confirming consistency with prior work. Fig.6e quantifies NELSON's protocol⁵⁴ on NLST, selecting indeterminate (50-500 mm³) and growing ($\geq 25\%$ Volume Growth) NODCAT III nodules at initial TP (T_{-2}) and evaluating VDT over short follow-up. In this subgroup, our model achieved an AUC of 0.954, improving sensitivity by 11.9% and specificity by 7.4% over NELSON's VDT criteria at their respective MYI. These findings also validate NELSON's 400-day VDT and 1.5 mm Δ diameter screening thresholds (Fig.6ad) aligning closely with their MYI on ROC curves.

Using T_{-2} predictions, we compared early AI diagnosis to that of 4-radiologists on the follow-up scan acquired one year later at T_{-1} . As anticipated, for NELSON's fastest-growing nodules (GROWCAT C, VDT<400 days), radiologists outperformed AI at T_{-2} but were subsequently surpassed by it at T_{-1} (Supplementary Fig.8). In contrast, Fig.6f demonstrates that for slow-growing and indeterminate nodules (GROWCAT A+B, VDT $\geq$ 400 days), our model at T_{-2} significantly outperforms radiologists at T_{-1} ($p < 0.0001$, AUC=0.908 for AI vs. 0.861 for 4-radiologists).

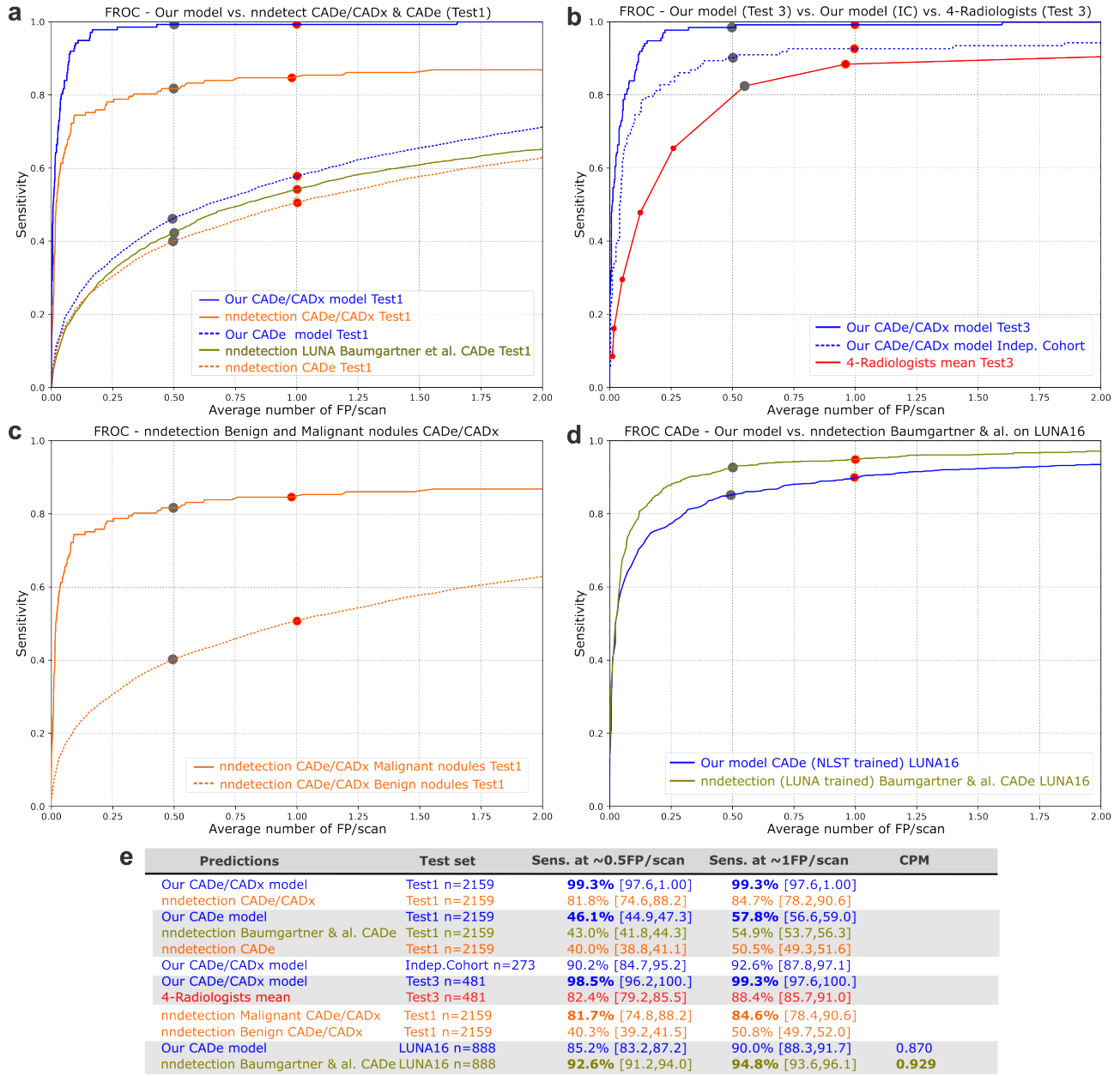


Figure 4. Comparison of detection performances: our model vs. radiologists, CADe and SOTA model, benign vs. malignant nodules (nnDetection), & performance comparison of our model vs. original nnDetection on LUNA16 challenge: **a.** The Free-response Receiver Operating Curves (FROC) on Test1 comparing our CADe/CADx model, and the nnDetection CADe/CADx model retrained on Train1 (malignant class detection only). Also shown: FROCs of our CADe-only module for the nodule detection task (malignant and benign class detection), of the nnDetection model retrained on NLST (both for malignant and benign class detection), and of the nnDetection model trained by Baumgartner *et al.*¹⁵ on LUNA16 (both for malignant and benign class detection). **b.** FROCs of our CADe/CADx model on Test3 and IC, compared to the mean performance of 4-Radiologists assessments. **c.** The FROC curves of the nnDetection CADe/CADx model retrained on NLST (Train3) for Benign and Malignant nodules detection tasks as separate classes on Test1 (the Malignant nodule detection curve is the same as in **a**). **d.** FROC curve of our CADe model trained on NLST alongside the nnDetection model trained by Baumgartner *et al.*¹⁵ on LUNA16, on the LUNA16 nodule detection challenge dataset. The Competition Performance Metric (CPM) of 0.929 achieved in this study for the Baumgartner *et al.* model reproduces their published CPM of 0.930, thereby "outperforming all previous methods on the nodule-candidate-detection task"¹⁵. **e.** Table of the mean sensitivity over 5,000 bootstraps samples (in percent) at FROC OPs closest to 1 and 0.5 mean FP/scan for the various predictions in **a,b,c,d**, with 95% CI over 5,000 bootstraps samples. The CPM for LUNA16 challenge task are also included.

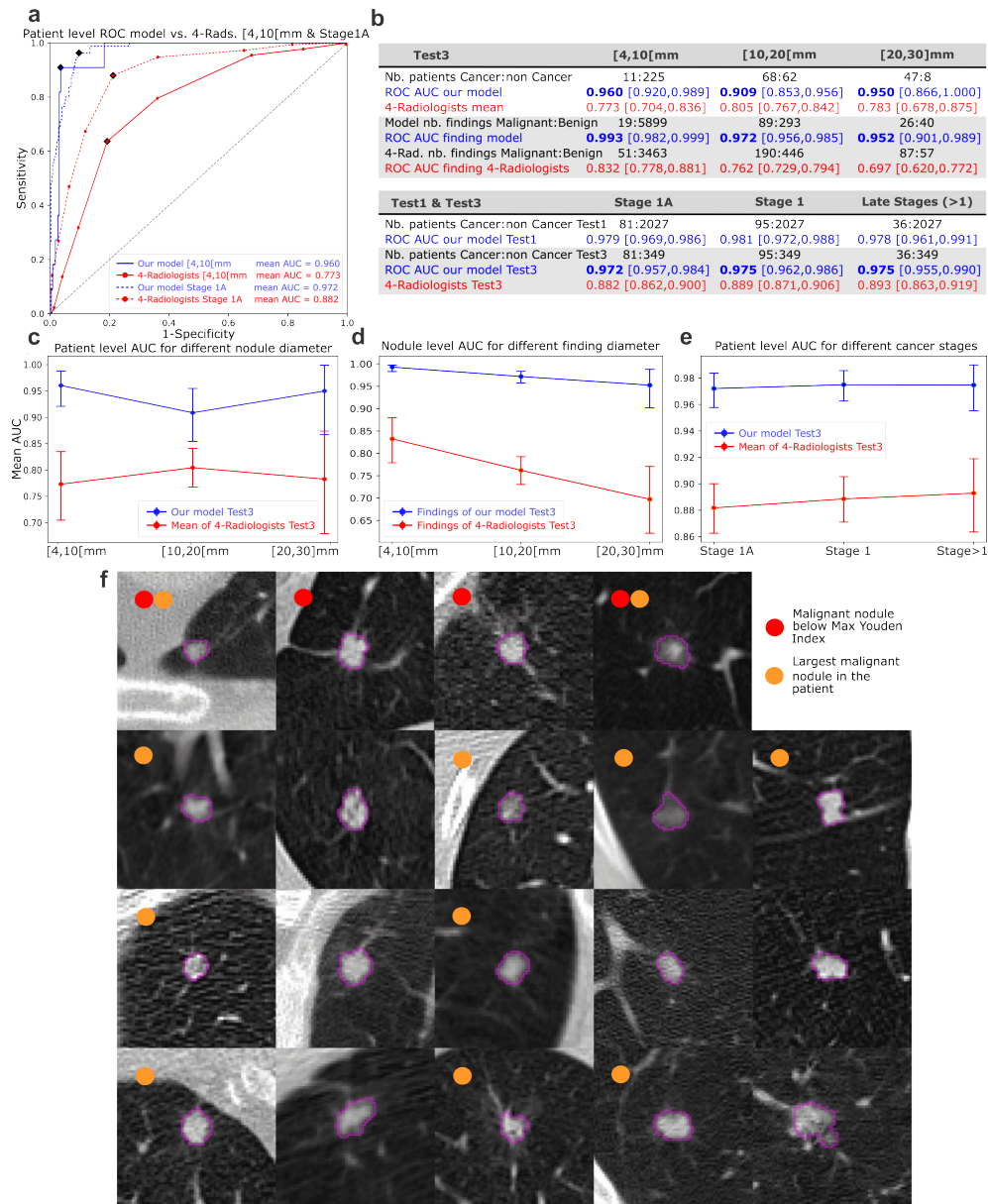


Figure 5. Performance across size-based subgroups: **a.** Patient-level ROC curves on Test3 comparing our model's malignancy prediction and the mean of 4-Radiologists' assessments for patients whose largest nodule has a diameter GT within [4,10]mm range, and for stage 1A (vs. non-cancer) patients. **b.** Table summarizing panels **a,c,d** and **e** of patient- and nodule-level AUCs for our model and for the 4-Radiologists across nodule size and cancer stage subpopulations with the corresponding malignant and benign sample size. 95% are provided over 5,000 bootstraps. The discrepancy between the 19 malignant nodules (nodule level) and the 11 cancers (patient-level) reflects patients with multiple malignant nodules and differences in diameter between GT and our model estimates. **c.** Mean AUC for subgroups of patients whose largest nodule has a diameter GT within [4,10]mm, [10,20]mm and [20,30]mm ranges on Test3. Horizontal bars represent the 95% CI over 5,000 bootstraps (as for other panels). **d.** Nodule-level mean AUC of the mean of four Radiologists' findings compared to our model's findings with computed diameter in the [4,10]mm, [10,20]mm and [20,30]mm ranges on Test3. **e.** Mean AUC for patients with stage 1A, stage 1 (stage 1A and stage 1B), and late stages (stage >1), compared to non-cancer patient. This stratification is imposed by the small prevalence of stage 1B and late stages in NLST (see Supplementary Information). **f.** Our model's outputs for the 19 malignant nodules in the [4–10]mm range in GT. Pink contours represent the automatic segmentation of the nodule. Each CT patch is 40*40mm. Nodules are ordered by ascending malignancy prediction in reading order. Red dots mark the four nodules misclassified as malignant (False Negative at Maximum Youden Index). Orange dots denote the 11 nodules identified as the patient's largest nodule of the patient (as in **b**), in all other cases, the patient presents some larger malignant nodule.

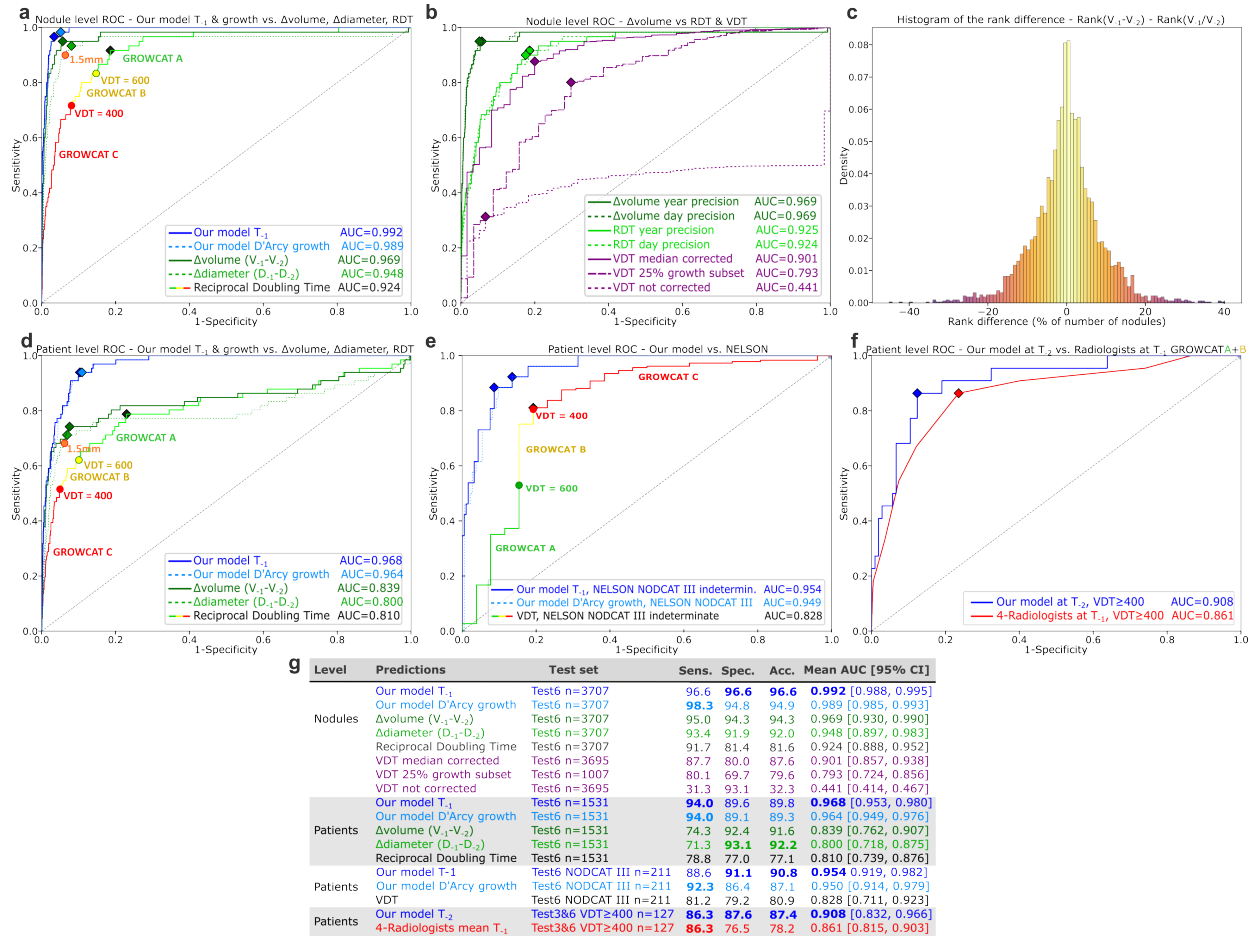


Figure 6. AI prediction and temporal evolution vs. standard growth metrics & comparison of radiologists' predictions vs. AI predictions made one year earlier (in slow-growing or indeterminate nodules). On each ROC the Maximum Youden Index is represented by a square matching the curve color. **a.** Nodule-level ROC of model predictions at T_1 , of the D'Arcy Thompson growth of model predictions, for Δ volume, Δ diameter, and RDT (or, equivalently, the percent of Volume Growth) for all nodules longitudinally paired by radiologists in GT on Test6. For RDT, the OP of 600 and 400 days thresholds—defining the GROWCAT A, B and C categories of nodule growth in the NELSON protocol³—are depicted by yellow and red circles, while for Δ diameter the threshold of 1.5 mm of Lung-RADS[®]¹⁴ is depicted by an orange circle. **b.** Nodule-level ROC curves comparing Δ volume with precision in years (same as in **a**), or precision in days ($365 * (V_1 - V_2) / (T_1 - T_2)$, T in days) with RDT with precision in years, or with precision in days (same as in **a**) and VDT in three variants: without correction; with $\geq 25\%$ growth selection as in NELSON⁵²; and VDT with median centering correction. All VDT measures use daily precision; 12 nodules were removed with respect to **a** due to infinite VDT values (stable volume). ROC curves are computed with permuted benign/malignant GT. The median centering correction is $VDT_c = (VDT - median_{VDT})^2$ following the results and conclusion of Ho⁵³. The VDT with $\geq 25\%$ growth selection yielded an AUC of 0.795, exceeding but comparable to the 0.67 reported by⁵² in different population. **c.** Distribution of rank difference—expressed as a percentage of the total number of nodules—between the Δ volume and the volume ratio (as for RDT) over the nodules longitudinally paired in GT of Test6. **d.** Patient-level ROC of the same variables as in **a** with the same representations, obtained by selecting the largest nodule at first TP for size growth analysis, per the NELSON protocol, using our model's patient-level predictions for the curves and its temporal evolution. The sample of 1,531 patients includes all Test6 patients with at least one nodule longitudinally paired by radiologists in GT. **e.** Performance comparison of the NELSON protocol using NELSON criteria^{3,54} and our model on NLST: Patient-level ROC curves for our model and VDT on the subgroup of patients with nodule with volume at T_2 in the $[50 - 500]mm^3$ range and showing at least a 25% volume increase (NODCAT III indeterminate category). **f.** Patient-level ROC comparing our model at T_2 vs. the mean of 4-Radiologist at T_1 for the subgroup of patients of Test6 with nodules with VDT $\geq$ 400 days, i.e., corresponding to the negative and indeterminate growth category in the NELSON protocol³ (GROWCAT A+B). **g.** Mean sensitivity, specificity and accuracy at the MYI, sample size and mean AUC with 95% CI over 5,000 bootstraps for each ROC presented in **a,b,c,d,e,f**.

Discussion

Previous studies have demonstrated AI models can equal or surpass radiologists in diagnostic accuracy across various medical imaging tasks^{57–60}. In LCS, Ardila *et al.* reported superior AI performance⁹, and the Brock model improved triage over Lung-RADS®2019⁶¹. Our study confirms these observations, and presents an architecture that significantly outperforms existing models, including those by Ardila *et al.*⁹, NLST Brock³⁸, Liao *et al.*³⁵ (winner of the Kaggle Data Science Bowl), and Sybil²⁵. However, Liao and Brock performances may be inflated in our comparison due to possible training–Test1 overlap and resulting overfitting.

In contrast, our model outperforms the mean performance of 4-radiologists and each individual radiologist, as well as Lung-RADS® assessment with or without prior imaging—whereas previous studies only matched Lung-RADS® when prior scans were available⁹. Additionally, our model consistently surpasses radiologist performance across all nodule sizes and cancer stage subgroups. However, by design, our model is limited to solid and part-solid parenchymal nodules [4,30] mm, and is therefore not intended to detect hilar, mediastinal, Ground Glass Opacity (GGO) nodules, or incidental findings, which still rely entirely on the reading of radiologists. Interestingly, for cancers with slow or indeterminate growth ($VDT \geq 400$ days), our model applied to CT scans acquired [24, 12] months before diagnosis outperforms radiologists reviewing diagnostic scans acquired one year later (AUC=0.908 vs. 0.861, $p < 0.0001$). This suggests the model’s potential for detecting malignancy well before overt metastatic progression, enabling earlier intervention and improved patient outcomes.

Our model also achieves a detection sensitivity of 99.3% at 0.5 FP/scan, a substantial improvement over SOTA detection models, far lower than traditional CADe systems, which typically focus on detecting all nodules without assessing malignancy. For example, the DL-LND⁴⁰ reported a sensitivity of 93.0% at 7 FP/scan, and Murchison *et al.* achieved 95.9% sensitivity at 10.9 FP/scan⁶². Moreover, many of our model’s FP are benign nodules that conventional CADe models would classify as true positives, highlighting the advantage of an integrated CADe/CADx approach. Since benign nodules are considerably more challenging to detect (cf. Fig. 4)—due to ambiguity in GT assessments and inter-reader variability⁶³—traditional CADe systems generally exhibit poor specificity. In contrast, malignant nodule detection benefits from more reliable GT, enabling deep learning models to achieve higher accuracy. Clinically, this distinction is critical: CADx-supported malignancy ranking improves screening efficiency by prioritizing the most relevant nodules.

This gain is primarily driven by the effectiveness of the CADx module in filtering out FPs, powered by a massively parallel ensemble of highly diverse classifiers—including 2D and 3D CNNs and tree-based models—that collectively bolster its performance and generalization. Unlike nnDetection, Sybil, or Ardila—which rely on end-to-end backpropagation—our model adopts a factorized learning strategy, that separates detection (‘where’) from characterization (‘what’) within a frozen learning framework. This design partially mirrors the dual-stream architecture of human visual cortical processing, enhancing explainability, by automatically linking malignancy suspicion to precise nodule localization and detailed features like diameter, volume, and radiomics. This strategy facilitates more efficient learning, particularly with medium-sized datasets. Three factors underlie our model’s superior performance. First, its architecture—comparable in size to ViT-Huge—adopts a fundamentally different approach using a modular, factorized, and large-scale ensembling of specialized shallow modules, optimized for mid-sized medical datasets. Second, it leverages a highly accurate GT, from biopsy-confirmed diagnoses and follow-up data, complemented by detailed internal annotations. Third, the training dataset includes over 23,000 CT scans from LIDC and NLST, making it one of the largest publicly available lung cancer AI datasets, second only to Chen *et al.*⁶⁴. Generic large multi-modal transformers remain hard to scale for 3D imaging due to their high data volume and computational demands. Even Med-Gemini^{65,66} or MedMPT (0.824 AUC on NLST⁶⁷) still far underperform specialized models like ours. Scarcity of curated, high-quality training data—not methodology—is now the main bottleneck. Under current constraints, specialized shallow models ensemble remains the most effective solution.

Radiologists rely heavily on size for malignancy triage, but our AI model is more accurate and robust across all nodule sizes, suggesting a shift in diagnostic logic in which “size does [not] matter [that much]” with AI. A single AI malignancy prediction at T_{-1} outperforms all size-based growth metrics, supporting the prioritization of AI-based malignancy prediction over traditional measures. Without AI, Δ Volume remains the strongest size-derived predictor. The integration of such AI capabilities into clinical workflows holds substantial promise for transforming LCS. Our model could reduce unnecessary follow-up imaging and invasive procedures—both major contributors to patient distress and overall healthcare expenditure. Furthermore⁶⁸, early identification of malignancies among small nodules currently managed conservatively may drive changes in clinical decision-making, promoting earlier treatment and improved outcomes. These practical implications call for a re-evaluation of existing workflows in light of AI-enabled tools.

In conclusion, our findings set a new benchmark for AI-assisted LCS. Our model significantly outperforms radiologists in both malignant nodule detection and cancer diagnostic—despite using a single time point, unlike radiologists who rely on longitudinal scans—challenging the limitations of size-based assessments and redefining what is possible in screening

workflows. Future work should prioritize the real-world integration of AI tools, to enhance early diagnosis, reduce unnecessary interventions, and ease the healthcare burden. Importantly, the broader economic benefits of AI-enabled earlier detection and treatment deserves further investigation to inform policy and reimbursement strategies.

Methods

Data

The dataset consists of three cohorts: LIDC-IDRI, NLST, and an IC.

Datasets

LIDC-IDRI: A publicly available US LCS database with 1,018 CT scans from 1,010 patients³². It provides high-quality nodule detection, segmentation, and malignancy assessments from up to eight annotators. However, histopathological confirmation is available only for 77 malignant and 36 benign cases. Further details are provided in Supplementary methods.

NLST: The National Lung Screening Trial is a large multi-centric U.S. screening programme with longitudinal follow-up extending up to seven years⁶⁹. It provides biopsy-confirmed cancer diagnoses but lacks precise nodule segmentation and localization. To address this, we enriched its GT via two re-annotation campaigns:

- Complementing GT (Test1, Train1) with full nodule detection, longitudinal pairing, segmentation, and characterization.
- Assessing radiologist performance using a protocol similar to LIDC, with four annotations per scan from 20 radiologists (Test3).

For further details, see Supplementary methods.

Independent Cohort: A multi-centric dataset from three providers:

- Segmed (US, $n = 255$, 49%)
- Gradient (US, $n = 131$, 25.2%)
- AIR Trial¹⁰ (France, $n = 134$, 25.8%), comprising exclusively patients with COPD.

Further details are in Supplementary methods. Demographic and scan parameter distributions for NLST and the IC are presented in Fig.2.

Data inclusion/exclusion

A complete overview of the dataset inclusion/exclusion flow is illustrated in Fig.7. Fig.7 a. depicts the different stages of the inclusion/exclusion process of the study, alongside the corresponding sample sizes. All steps of inclusion/exclusion process and corresponding criteria are described in the Supplementary methods. For NLST, we adhered as closely as possible to the inclusion/exclusion workflow and criteria of Ardila *et al.*⁹, considering our specific nodule level malignancy detection aims. This enables direct comparison by ensuring our patient and volume datasets closely approximate those used in Ardila *et al.* (and Sybil²⁵). Our deviations from their protocol are that we exclude cancer patients presenting only malignant (pure) GGO nodules—focusing on both solid and part-solid—owing to the higher malignancy risk associated with part-solid nodules⁷⁰ and the comparatively lower likelihood of malignancy for pure GGO (see⁷¹ and references therein), and that, implementing the definition of nodule of the Fleischner society¹¹, we focus only on malignant nodule in the [4,30]mm diameter range while malignant nodules above 30 mm—considered as masses—are excluded (notably mediastinal and hilar masses). GGO display atypical slow growth and excellent survival outcomes, may unlikely progress to malignancy only after becoming part-solid^{72,73}, and have diameter and volume inherently fuzzily defined, inducing considerable measurement variability⁷⁴, rendering them unsuitable for the purpose of this study. As a result of this selection, 2,065 patients (Test2) overlap both the 2,357 and 2,203 patients of the respective test sets of Ardila *et al.*²⁵ and of Sybil²⁵, and 481 (Test3) overlap the 501 annotated test set of Ardila *et al.*

For NLST and IC, the selection consists of three main steps with inclusion/exclusion criteria specified (see Supplementary methods for details):

1. Availability: scans available without download error;
2. Quality control (QC): Exclusion of kernels, patients and scans lacking complete information (e.g. three TPs for NLST) or unreadable images,

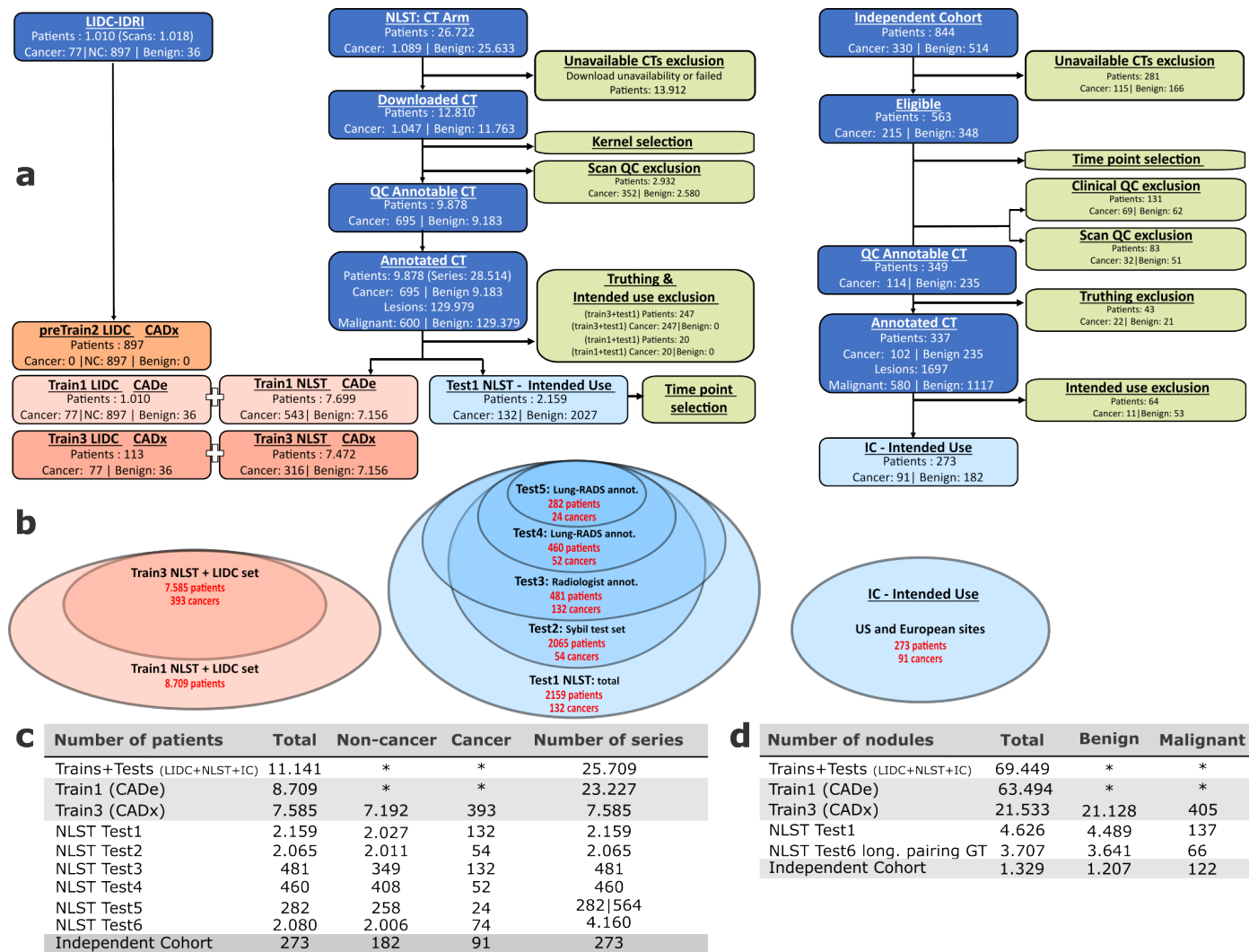


Figure 7. Datasets inclusion/exclusion: a. Study inclusion–exclusion flowchart with sample sizes. For LIDC train sets, ‘Cancer’ refers to cases confirmed by histopathology study following biopsy or resection; ‘Non-cancer’ refers to cases confirmed negative for cancer by at least one year of follow-up; ‘not confirmed’ includes all other cases (for which a malignancy visual assessment based solely on CT is given by multiple annotators) **b.** Venn diagrams representing inclusions of our train and test sets. **c,d.** Patient, series and nodule counts for each of our train and test sets. * denotes values that are not applicable due to lack of definitive cancer GT in full LIDC datasets.

3. Eligibility: exclusion based on study purpose and truthing exclusion. This step excludes all non-parenchymal cancers (hilar, mediastinal), patients with exclusively ground-glass malignant nodules, and patients whose malignant nodules fall outside the diameter ranges of [3,40]mm and [4,30]mm for tests and training sets, respectively. Additionally, scans from cancer patients that could not be disambiguated—where our radiologists were unable to identify the malignant nodule on the available CT scan—are also excluded.
4. Time Point selection: for Test1–5, the scan closest to cancer diagnosis within 12 months (T_{-1}) or, for non-cancer patients, the earliest available scan. For Test6 (longitudinal), for cancer patients the same TP scan as test1–5 for T_{-1} and the scan closest to cancer diagnosis within [12,24] months for T_{-2} ; for non-cancer patients, the earliest pair of consecutive TPs available.

For LIDC, which serves solely as a training dataset and is already curated, we only included cases confirmed by biopsy or follow-up (see Supplementary methods for details).

Train and test sets

Following the inclusion/exclusion process, the NLST-LIDC dataset was split at the patient level into two subsets—a training set and a test set—by “leave-out” such that: (1) they are fully independent; (2) the test set ($n = 2,159$) has maximal overlap with the test set in Ardila *et al.*⁹ obtained from random splitting, and hence supposed to preserve covariate distributions of intended use population; and (3). the test comprises 132 cancer patients, i.e., the minimum sample size determined by a preliminary statistical power analysis to ensure robust overall and subgroup analyses. As the test set of Ardila *et al.*⁹ contained only 66 cancer cases after exclusions, we supplemented it with 66 additional cancer patients randomly selected from NLST to enrich our test set (Test1).

Training Sets:

- **Train1:** used for CADe training, includes all TPs, excludes pure GGO cancers, and retains only nodules of interest (see Supplementary methods);
- **Pretrain2:** used for CADx pretraining, includes LIDC-IDRI scans excluded from Train3 (i.e., without confirmed cancer or non-cancer status);
- **Train3:** used for CADx training, includes one single TP per patient, excludes pure GGO cancers, and retains nodules in the [3,40]mm range.

Train1 and Train3 were subsequently split into training and tuning subsets for CADe and CADx training, respectively, as detailed in Supplementary methods. Fig.7c. reports the sample sizes as well as the cancer-non-cancer breakdown and class imbalance across the training and test sets.

Test Sets:

- **Test1** ($n = 2,159$; 132 cancers): the largest test set, reflecting real-world LCS conditions, derived from the extensive multi-centre NLST campaign and enriched to 6.1% cancer prevalence to support more robust subgroup analyses;
- **Test2** ($n = 2,065$; 54 cancers): Subset of Test1 used for benchmarking against SOTA models. It maximally overlaps with the test sets of Sybil²⁵ and Ardila *et al.*⁹, thereby mitigating overfitting bias. This set exhibits an imbalance (2.6% cancer prevalence) that better approximates the expected LCS population.
- **Test3** ($n = 481$; 132 cancers): Subset of Test1, enriched in cancers (28%) for robust subgroup analysis. Annotated by 4-Radiologists for nodule detection and malignancy scoring, it maximizes overlap with the 507-patient test set from Ardila *et al.*⁹.
- **Test4** ($n = 460$; 52 cancers): Subset of Test1 originating from Ardila *et al.*⁹, featuring Lung-RADS® assessments by six radiologists without access to prior CT scan information;
- **Test5** ($n = 282$; 24 cancers): Subset of Test1 derived from Ardila *et al.*⁹, featuring Lung-RADS® assessments by six radiologists, with access to prior CT scan information;
- **Test6** ($n = 2080$; 74 cancers): the earliest pair of consecutive TP scans available for patients from Test1. Two subsets of Test6 were used for longitudinal analysis: one with all (patients with) nodules longitudinally paired by radiologists in GT ($n = 1531$, 66 cancers, $n = 3707$ nodules, 66 malignant); the second intersecting with Test3 ($n = 419$, 74 cancers).

Fig.7c,d summarizes training and test set composition.

Model

Strength in Specialization: A Factorized Ensemble of Small Models to Address Data and Explainability Constraints

Medical datasets are typically smaller than those used for training Large Language Models and involve handling high-dimensional 3D medical images. This makes very deep Convolutional Neural Networks (CNNs) and transformers inefficient due to their large number of parameters and high GPU memory requirements⁷⁵. Transformers generalize CNNs by replacing spatially limited convolutional attention with exhaustive attention, thereby increasing learnable parameters, particularly in 3D⁷⁶, hence requiring far larger datasets for effective training. In theory, with unlimited data and computational capacity, an optimal approach might involve a single-voxel self-attention classifier taking the entire CT scan as input or 3D CNNs in more limited settings, with detection and segmentation naturally emerging as sub-tasks. However, in data-constrained settings, ‘weak learners’ (e.g., gradient boosted trees), shallow networks, and ensemble methods often outperform deep models^{24,77}. Studies have shown that smaller networks like EfficientNet-b0 (45K parameters) can outperform larger counterparts in CT classification⁷⁹.

To address the challenges of training on a mid-sized dataset (10,000 cases), we designed a factorized ensembling strategy that combines diverse and complementary CNNs in a massively parallel architecture to enhance robustness and performance. Drawing inspiration from the ventral and dorsal visual pathways in the human cortex⁸⁰, we factor learning into separate detection (localization) and characterization tasks—an approach successfully adopted by top-performing Kaggle teams. Our model employs a hybrid ensemble architecture that integrates a detection module—built from 3D-SegResnet ensembles and 3D-retina UNet for lung and nodule segmentation—alongside a characterization module that combines 15 2D-CNNs, 15 3D-CNNs, 15 radiomics-based gradient boosted trees (leveraging 122 radiomics and 3D-morphomics features²²), and a full-CT scan 3D classifier adapted from a Sybil-inspired attention model²⁵. This diverse ensemble effectively balances local and global feature extraction, for strong performance on mid-sized medical datasets.

Each component of this 'factorized' model uses a SOTA approach: a Sybil-inspired²⁵ full-CT scan model captures not only local nodule features, but also contextual features (e.g., emphysema, nodule density, and location); a gradient boosted trees classifier leverages standard radiomics²¹ and 3D geometrical curvature features²², as in recent studies^{83–85}; and a relatively shallow 39-layer 3D DenseNet builds on prior success of similar CNNs in medical imaging studies⁸⁶. Additionally, this factorization strategy into subtasks also aligns with the expected standards for explainability in the medical field^{87,88}: the model not only predicts cancer diagnosis but also, through the intermediate module outputs, provides detailed information on each detected finding that contributed to the final prediction. This includes precise localization of findings ranked by malignancy suspicion, their segmentation, and standard characteristics such as size, volume, and other radiomic features. However, explainability remains partial, and the performance gains offered by AI over feature-based models^{8,38} is distributed throughout the entire 3D scan voxels and millions of parameters.

Ensembling these diverse and complementary models with targeted pretraining modules enhances both performance and generalizability. Importantly, even simpler classifiers contribute meaningfully to final predictions, as confirmed by a preliminary ablation study. Although our model has a high total parameter count as a result of this parallel structure (965 million parameters, see Fig. 8), its design fundamentally differs from traditional large CNN and transformer models—in both structure and training strategy—notably by introducing systematic sub-network freezing and integrating models trained with substantially different and specialized loss, reflecting a heavily expert- and experience-driven supervised 'Frankenstein modeling' paradigm. For comparison, classic CNNs such as 2D VGG-16, ResNet-50, DenseNet169, and Inception-v4 have 138M, 23M, 7M, and 41M parameters, respectively⁷⁹, while the original ViT-Huge model (2D) has 632M parameters and SOTA large-generic medical model MedGemma-only “approaching the performance of task-specific models”—has 4B or 27B parameters⁸⁹. Our model is designed for a commercial fully automated assistant for LCS radiologists.

Our model

The model was developed incrementally, by integrating the ensemble of 3D-CNN, the ensemble of 2D-CNN, an XGBoost classifier based on radiomics and morphomics, and full-CT scan model—each addition validated to significantly improve classification performance, as measured by AUC on tuning set. The modules were trained sequentially in feed-forward order, using outputs from preceding modules as input, with a basic version of sub-network freezing method⁹⁰ that enables the use of different losses.

The sequential modules that make up the CADe/CADx system are detailed below, and illustrated in Fig. 8a (from left to right):

- **Lung segmentation:** ensemble of 3D-SegResNet models¹⁷.
- **Nodule detection (CADe):** ensemble of Five Retina U-net¹⁶ models with eight test-time augmentation.
- **Patch extraction and filter in lung (CADe):** filters out any detection outside the lung parenchyma—with Bounding boxes having Intersection over Union (IoU)>0 with the lung mask.
- **Nodule segmentation (CADe):** ensemble of 3D-U-Net¹⁷.
- **Diameter measurement (CADe):** extracts minimum, maximum, and mean diameters from the convex hull of the nodule segmentation. It filters out any detection outside the [4,40]mm diameter range.
- **Nodule characterization (CADx):** ensemble of a 15 shallow 3D-CNN densenet39¹⁹ module, a 15 shallow 2D CNN module, and a 15 XGBoost²³ classifiers module based on 111 extracted radiomic and 11 3D-morphomic features²².
- **Nodule Predictions ensembling and calibration (CADx):** averages outputs from the 15 3D, 2D and XGBoost models, then computes the optimal weights ensembling (convex optimal stacking²⁴) to combine them into a single prediction per finding. It then applies optimal calibration using the NetCal logistic model by Küppers et al¹.

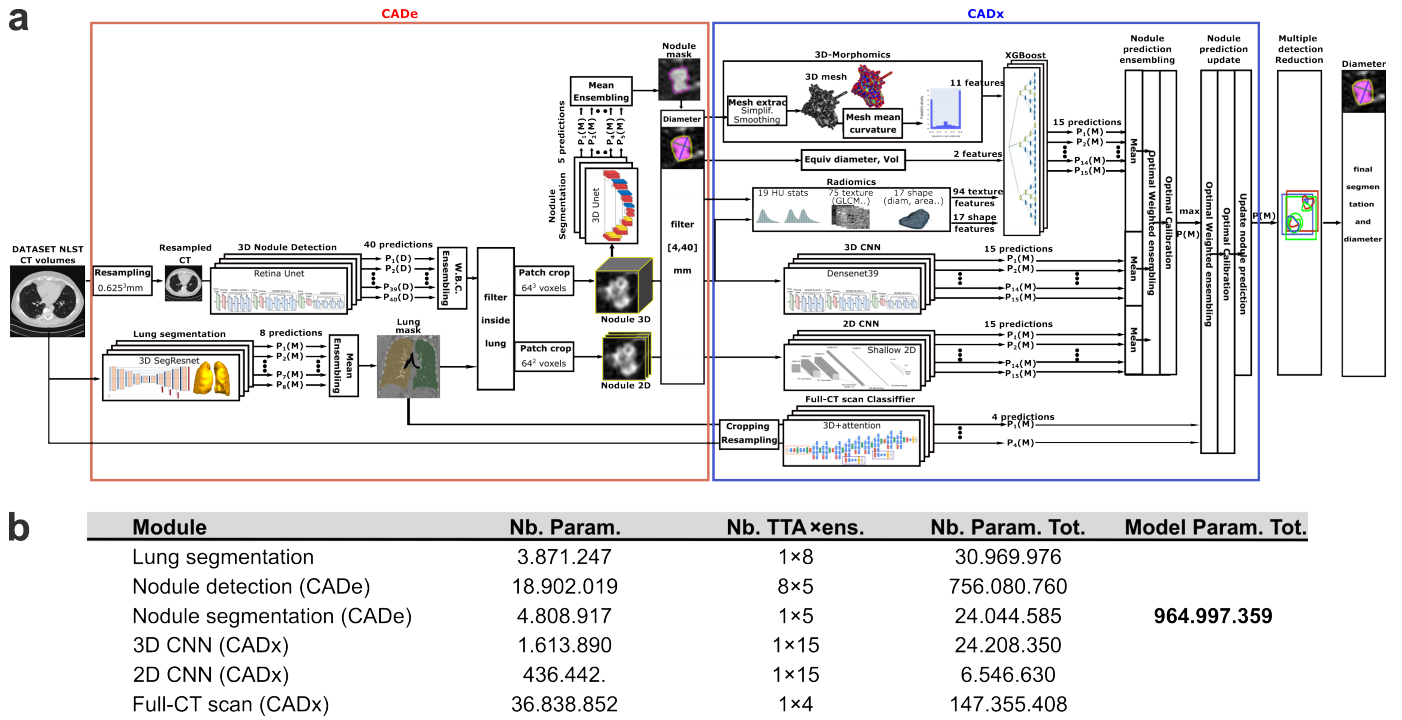


Figure 8. Modular architecture of the CADe/CADx system: **a.** The model comprises two sequential components: a CADe module for nodule detection (left, red box); and a CADx module for malignancy characterization (right, blue box). The CADe module includes: an ensemble of 3D CNNs for lung segmentation; another for nodule detection with FP reduction; and a third for nodule segmentation. The CADx module combines: an ensemble of gradient-boosted decision tree classifiers using radiomic and morphomic features; ensembles of local patch-based 2D and 3D CNNs; and a full-CT scan 3D CNN with an attention mechanism. Predictions at the nodule- and patient-levels are ensembled and integrated via calibrated stacking. A final post-processing step removes spurious and duplicate detections. **b.** Number of parameters per Deep Neural Network submodules: Number of learnable parameters of each individual model, number of Test Time Augmentation (TTA) and number of ensembled models (ens.), and the total number of parameters for each Deep Neural Network modules of our model. The last column gives the total number of parameters of the whole model.

- **Full-CT scan characterization (CADx):** ensemble of 4 3D-CNN module with trained attention mechanism adapted from Sybil²⁵ and retrained on Train3, taking the full CT scan in input.
- **Nodule prediction update and calibration (CADx):** computes the optimal weights ensembling (convex stacking²⁴) of the four full-CT scan predictions alongside the maximum nodule predictions; and computes the optimally calibrated transformation of this patient prediction¹. Nodule predictions are then updated by adding a correction factor proportional to the difference (in percentage) between the final patient prediction and the maximum of detected nodule predictions. As a result of this last step, the patient and maximum nodule predictions are equal, calibrated (prediction value optimally reflects the real observed accuracy), and bounded within the [0, 1] interval.
- **Multiple detections reduction module and diameter measurement (CADx):** first, the central connected component of each finding is selected. Overlapping central components from different findings are merged by union, with the merged finding inheriting the highest nodule prediction. Diameters and volume are then extracted from the merged findings. For this step, the previous diameter estimation method is slightly adapted to better reflect radiological practice.

A detailed description of each of those modules, of their training and parameters is provided in the Supplementary methods.

Other models

Sybil: The original Sybil model²⁵ (<https://github.com/reginabarzilaygroup/Sybil>), using the predictions of an ensemble of five models, is run in test mode on our Test2 using the default configuration, one year before diagnosis (which provided the best performance for Sybil and corresponds to our model's settings).

Liao *et al.*: The original model of Liao *et al.*³⁵ (<https://github.com/lfz/DSB2017>) is run in test mode on our Test2 using the default configuration.

Ardila *et al.*: We use the original predictions of Ardila *et al.*⁹ for patients on Test2 provided in their supplementary files and using the selection method described in Supplementary Information.

NLST Brock model: The NLST Brock model, adapted from the PanCan-Brock model and recalibrated for NLST dataset complemented with a pack-year feature (thereby addressing low PPV and overestimation issues in the original model), was re-implemented following³⁸ and applied to Test2 using nodule features provided by NLST radiologists GT (available at <https://cdas.cancer.gov/datasets/nlst/>).

MAYO model: Following the same method as for the NLST Brock model, we re-implemented and ran the MAYO model as described in⁹⁶.

nnDetection model: The nnDetection model¹⁵ (<https://github.com/MIC-DKFZ/nnDetection>) was retrained on Train3 for CADe/CADx and Train1 for CADe to separately detect malignant nodules (CADe/CADx) or malignant and benign nodules (CADe) using the default configuration. We also tested the "original" CADe model trained on LUNA16 provided by Baumgartner¹⁵.

Data analysis

ROCs and FROCs

Malignancy likelihood predictions for each patient is given by the highest malignancy prediction value among all detected findings for models operating at nodule level (such as ours, the Brock model and the radiologists' readings), and from the output of Sybil, Ardila *et al.* and Liao *et al.* (that only provides a global patient prediction). We report the MYI, which indicates the optimal cutoff threshold for a deterministic binary decision between predicted malignant and benign outcomes (cf.⁹⁷ p.106): $J = \text{Max}_c(\text{Sensitivity}_c + \text{Specificity}_c - 1)$, where c is the optimal cutoff⁹⁷ p.107. ROC curves are computed using standard Scikit learn function, with AUCs estimated via trapezoidal (linear interpolation) integration rule and averaged over 5,000 bootstrap samples. Note that Lung-RADS[®] discretizes malignancy scores into only six classes, which introduces an uncertainty on the ROC and AUC estimations due to interpolation.

Radiologists ROC: the mean ROC curve and its corresponding AUC across the four sets of annotations per scan (from a total of 20 radiologists) is obtained assuming equivalence of readers (permutation invariance). In practice, this simply involves pooling the whole set of radiologist predictions against GT, essentially generalizing the non-parametric ROC method described in^{98,99} while allowing for varying sample sizes per annotator. To evaluate individual radiologist performance relative to our model and assess inter-radiologist variability, we select the 12 radiologists who annotated at least 15 cancer cases. We compute the ROC curves for each of these as well as our model using the same set of patients and scans for comparison. The resulting 12 AUCs and corresponding CIs are derived from 5,000 bootstraps from both individual readers and the model, using identical samples. Results are summarized in Fig.3.

ROC curves for subgroups: For patient-level subgroups focusing specifically on patient characteristics, we simply consider the associated patient subsets and compute their ROC. For diameter-based patient subgroups, we define three ranges—[4, 10[, [10, 20[, [20, 30]mm—based on the largest nodule per patient. For nodule-level diameter subgroups analysis, findings from either radiologists or our model are selected based on their respective diameter estimations.

Statistical tests

For AUC, accuracy, sensitivity and specificity, the estimations of the mean, standard deviation, 95% CI and distributions are obtained by computing 5,000 bootstrap samples with replacement¹⁰⁰. For these measures, performance comparisons between our model and others were conducted over the same bootstrap samples by applying a superiority t-test with unequal variance using scipy.stats (one-sided Welch t-test, our prediction vs. other). A superiority test assuming equal variance (t-test) produced consistent results at equivalent p levels across all cases.

Lung-RADS[®] scores equivalent OPs of the model

The OPs of our model, designed to correspond to Lung-RADS[®] scores, follow the methodology of Ardila *et al.*⁹, with one key modification: equivalence is based on accuracy rather than precision, because our model's predictions are calibrated to reflect empirical accuracy¹. The accuracies of Lung-RADS[®] accuracy scores are derived from NLST estimates reported by

Pinsky (See Table 2)¹⁰¹ by uniformly distributing the ambiguous categories— i.e., classes 3 or 4A or 4B—across included classes ("Lung-RADS® 1": 0.010499583, "Lung-RADS® 2": 0.565082253, "Lung-RADS® 3": 0.870543047, "Lung-RADS® 4A": 0.934665051, "Lung-RADS® 4B": 0.972519143, "Lung-RADS® 4X": 0.983397773]) OPs in our training set with accuracy values closest to these Lung-RADS® accuracies are then selected as the equivalent OPs.

Size and size evolution measures

Diameter: In line with Fleischner Society guidelines for radiological best practices for measuring pulmonary nodules at¹⁰² and Lung-RADS® 2022¹⁴, the diameter of a finding is defined as the mean of the long- and short-axis diameter of the segmentation mask, either annotated by radiologists or automatically generated by our model using the algorithm detailed in Supplementary Information.

Volume: Nodule volume (in mm^3) computed from the segmentation mask provided by either the radiologist or by the model.

Cancer stages: AJCC6¹⁰³ stages provided by NLST.

VDT, RDT and Volume Growth: Calculated following established methods², using the Schwartz formula¹⁶ $VDT = ((t_{-1} - t_{-2}) \ln 2) / (\ln(V_{-1}/V_{-2}))$ and $RDT = 365/VDT$, where t_{-1} and V_{-1} are time (in days) and volume measured at the final CT; and t_{-2} and V_{-2} are measured at initial CT. Volume Growth is given by³ $100 \times (V_{-1} - V_{-2})/V_{-1}$.

D'Arcy Thompson growth of model prediction: given by the formula $P_{-1}/(P_{-1} - P_{-2})_n$. The logarithm is omitted here, as its monotonicity leaves the ROC invariant. The term $(P_{-1} - P_{-2})$ is normalized to the $[0, 1]$ interval to prevent issues with negative values.

Nodule and detection pairing

Pairing between GT and detected nodules (be it by models or radiologists) is addressed using an IoU criteria of $IoU > 0.1$ between GT and 3D Bounding Boxes, as recommended for 3D contexts¹⁶.

Acknowledgements

The authors gratefully acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for establishing and maintaining open access to the NLST and LIDC/IDRI Databases, which proved essential to this study. We also thank the DKFZ team, especially M. Baumgartner for sharing their code and LUNA16 trained nnDetection model.

Author contributions

S.B., P.B., C.V., B.R., B.H. conceived the experiments, P.B., G.D.B, E.G., V.K.L, D.F., B.H. conceived the model, E.G., B.R., C.V., V.B., Y.H., P.B., V.K.L, G.D.B, D.F., carried out data acquisition and preprocessing, P.B., G.D.B, E.G., V.K.L, D.F., P.H.S., Y.H., C.V., B.R., B.H. analysed the results, S.B., P.B., B.H. wrote the manuscript, J-C.B., V.B., B.R., C.C.T., provided analysis suggestion and manuscript corrections. All authors reviewed the manuscript.

Competing interests

The authors declare the following competing interests: Median Technologies has filed for patent protections (application numbers: US11810299B1, EU22315156.4, CN 202380053824.2, and US12272063B2, EU22315034.3, CN202380028642X, JP024-548634) on behalf of B.H., D.F., P.B. for the work related to 2D and 3D CNN ensembling and on behalf of B.H., P.B., E.M., E.G., J-C.B., V.G. for the work related to 3D-morphomics. P.B., B.R., C.H., G.D.B, E.G., V.K.L, D.F., P.H.S., Y.H., J-C.B., V.B., B.H. are employees of Median Technologies and own Median Technologies stock as part of the standard compensation package.

Data availability

LIDC-IDRI and NLST are open-source dataset available at <https://www.cancerimagingarchive.net/collection/lidc-idri/>³², and <https://www.cancerimagingarchive.net/collection/nlst/>¹⁰⁵. The IC is subject to restrictions and cannot be publicly released by the providers. The Multi-Reader annotation dataset, with a total of 2,004 scan annotations performed by radiologists—comprising 23,099 detected and annotated lesions, among which the 6,505 most suspicious nodules are segmented with malignancy assessment, for 501 patients (including those from Test3) of the public NLST dataset— is shared as open source, thereby

supplementing the public LIDC annotations on NLST scans (submitted to The Cancer Imaging Archive (TCIA under the short name LIDC-annot-NLST501)).

Code availability

To enable automatic and full reproducibility of all figures, statistical analyses, and results presented in the article, to support further benchmarking of the reported performances, and to generically evaluate arbitrary CAdE/CAdx systems, we publicly release the evaluation chain at <https://github.com/EYONIS-AIDS-DS/CAdE-CAdx-evaluation>, together with the corresponding model predictions, nodule volumes, and diameters (...). The code used for the implementation of the model has dependencies on internal tooling and infrastructure, and as such, is protected by patent applications filed under the following numbers: US11810299B1, EU22315156.4, CN 202380053824.2, and US12272063B2, EU22315034.3, CN202380028642X, JP024-548634), preventing public release. In line with reproducibility and transparency standards in AI emphasized by Haibe-Kains *et al.*¹³, all experiments and model implementations are described thoroughly in the Methods and Supplementary Information (Methods) sections to facilitate replication with non-proprietary libraries. All relevant links to the necessary library repositories are provided. To promote responsible innovation, we are actively seeking collaboration with research partners, regulators, and healthcare providers to evaluate and research secure applications and support further model advancement and improvement.

References

1. Barta, J. A., Powell, C. A. & Wisnivesky, J. P. Global Epidemiology of Lung Cancer. *Annals Glob. Heal.* **85**, 8, DOI: [10.5334/aogh.2419](https://doi.org/10.5334/aogh.2419) (2019).
2. Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H. & Jemal, A. Cancer statistics, 2025. *CA: A Cancer J. for Clin.* **75**, 10–45, DOI: [10.3322/caac.21871](https://doi.org/10.3322/caac.21871) (2025). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21871>.
3. Alberg, A. J., Brock, M. V., Ford, J. G., Samet, J. M. & Spivack, S. D. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **143**, e1S–e29S, DOI: [10.1378/chest.12-2345](https://doi.org/10.1378/chest.12-2345) (2013).
4. Henschke, C. I. *et al.* A 20-year Follow-up of the International Early Lung Cancer Action Program (I-ELCAP). *Radiology* **309**, e231988, DOI: [10.1148/radiol.231988](https://doi.org/10.1148/radiol.231988) (2023).
5. Mathur, P. N., Edell, E., Sutedja, T., Vergnon, J.-M. & of Chest Physicians, A. C. Treatment of early stage non-small cell lung cancer. *Chest* **123**, 176S–180S, DOI: [10.1378/chest.123.1_suppl.176s](https://doi.org/10.1378/chest.123.1_suppl.176s) (2003).
6. National Lung Screening Trial Research Team *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New Engl. J. Medicine* **365**, 395–409, DOI: [10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873) (2011).
7. de Koning, H. J. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *The New Engl. J. Medicine* **382**, 503–513, DOI: [10.1056/NEJMoa1911793](https://doi.org/10.1056/NEJMoa1911793) (2020).
8. Tammemagi, M. C. *et al.* Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. *The Lancet Oncol.* **18**, 1523–1531, DOI: [10.1016/S1470-2045\(17\)30597-1](https://doi.org/10.1016/S1470-2045(17)30597-1) (2017).
9. Wait, S. *et al.* Implementing Lung Cancer Screening in Europe: Taking a Systems Approach. *JTO Clin. Res. Reports* **3**, 100329, DOI: [10.1016/j.jtocrr.2022.100329](https://doi.org/10.1016/j.jtocrr.2022.100329) (2022).
10. Parker, K. *et al.* Invitation methods for Indigenous New Zealand Māori in lung cancer screening: Protocol for a pragmatic cluster randomized controlled trial. *PLOS ONE* **18**, e0281420, DOI: [10.1371/journal.pone.0281420](https://doi.org/10.1371/journal.pone.0281420) (2023).
11. Wu, J. *et al.* Application of artificial intelligence in lung cancer screening: A real-world study in a Chinese physical examination population. *Thorac. Cancer* **15**, 2061–2072, DOI: [10.1111/1759-7714.15428](https://doi.org/10.1111/1759-7714.15428) (2024).
12. Bhamani, A. *et al.* Low-dose CT for lung cancer screening in a high-risk population (SUMMIT): a prospective, longitudinal cohort study. *The Lancet Oncol.* **0**, DOI: [10.1016/S1470-2045\(25\)00082-8](https://doi.org/10.1016/S1470-2045(25)00082-8) (2025). Publisher: Elsevier.
13. Lam, D. C.-L. *et al.* Lung Cancer Screening in Asia: An Expert Consensus Report. *J. Thorac. Oncol.* **18**, 1303–1322, DOI: [10.1016/j.jtho.2023.06.014](https://doi.org/10.1016/j.jtho.2023.06.014) (2023). Publisher: Elsevier.
14. Christensen, J. *et al.* ACR Lung-RADS v2022: Assessment Categories and Management Recommendations. *J. Am. Coll. Radiol.* **21**, 473–488, DOI: [10.1016/j.jacr.2023.09.009](https://doi.org/10.1016/j.jacr.2023.09.009) (2024). Publisher: Elsevier.
15. McWilliams, A. *et al.* Probability of cancer in pulmonary nodules detected on first screening CT. *The New Engl. J. Medicine* **369**, 910–919, DOI: [10.1056/NEJMoa1214726](https://doi.org/10.1056/NEJMoa1214726) (2013).

16. Schwartz, M. A biomathematical approach to clinical tumor growth. *Cancer* **14**, 1272–1294, DOI: [10.1002/1097-0142\(196111/12\)14:6<1272::AID-CNCR2820140618>3.0.CO;2-H](https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142(196111/12)14:6<1272::AID-CNCR2820140618>3.0.CO;2-H) (1961). [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142%28196111/12%2914%3A6%3C1272%3A%3AAID-CNCR2820140618%3E3.0.CO%3B2-H](https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142%28196111/12%2914%3A6%3C1272%3A%3AAID-CNCR2820140618%3E3.0.CO%3B2-H).
17. Steel, G. G. & Lamerton, L. F. The growth rate of human tumours. *Br. J. Cancer* **20**, 74, DOI: [10.1038/bjc.1966.9](https://doi.org/10.1038/bjc.1966.9) (1966).
18. Yankelevitz, D. F., Yip, R., Jirapatnakul, A. & Henschke, C. I. The Winner and still champion: Nodule volume doubling times. *Eur. J. Cancer* **216**, DOI: [10.1016/j.ejca.2024.115184](https://doi.org/10.1016/j.ejca.2024.115184) (2025). Publisher: Elsevier.
19. Snoeckx, A. *et al.* Lung cancer screening with low-dose CT: definition of positive, indeterminate, and negative screen results. A nodule management recommendation from the European Society of Thoracic Imaging. *Eur. Radiol.* DOI: [10.1007/s00330-025-11648-4](https://doi.org/10.1007/s00330-025-11648-4) (2025).
20. Prokop, M. *et al.* Aggressiveness-guided nodule management for lung cancer screening in Europe—justification for follow-up intervals and definition of growth. *Eur. Radiol.* DOI: [10.1007/s00330-025-11647-5](https://doi.org/10.1007/s00330-025-11647-5) (2025).
21. Lancaster, H. L. *et al.* Action plan for an international imaging framework for implementation of global low-dose CT screening for lung cancer. *Eur. J. Cancer* **220**, 115323, DOI: [10.1016/j.ejca.2025.115323](https://doi.org/10.1016/j.ejca.2025.115323) (2025).
22. Creamer, A. W. *et al.* Performance of volume and diameter thresholds in malignancy prediction of solid nodules in lung cancer screening. *Thorax* DOI: [10.1136/thorax-2024-222086](https://doi.org/10.1136/thorax-2024-222086) (2025). Publisher: BMJ Publishing Group Ltd Section: Lung cancer.
23. Larici, A. R. *et al.* Lung nodules: size still matters. *Eur. Respir. Rev. An Off. J. Eur. Respir. Soc.* **26**, 170025, DOI: [10.1183/16000617.0025-2017](https://doi.org/10.1183/16000617.0025-2017) (2017).
24. Nair, A. *et al.* Variable radiological lung nodule evaluation leads to divergent management recommendations. *The Eur. Respir. J.* **52**, 1801359, DOI: [10.1183/13993003.01359-2018](https://doi.org/10.1183/13993003.01359-2018) (2018).
25. Han, D. *et al.* Influence of lung nodule margin on volume- and diameter-based reader variability in CT lung cancer screening. *The Br. J. Radiol.* **91**, 20170405, DOI: [10.1259/bjr.20170405](https://doi.org/10.1259/bjr.20170405) (2018).
26. Asmara, O. D. *et al.* Location and volume of screen-detected lung nodules: post hoc analysis from the UK Lung Cancer Screening (UKLS) trial. *Eur. Respir. J.* **64**, DOI: [10.1183/13993003.congress-2024.PA3337](https://doi.org/10.1183/13993003.congress-2024.PA3337) (2024). Conference Name: Location and volume of screen-detected lung nodules: post hoc analysis from the UK Lung Cancer Screening (UKLS) trial Publisher: European Respiratory Society Section: Lung cancer.
27. Storme, N. *et al.* Characteristics of lung nodules detected by low-dose CT scan: results from the French lung cancer screening study DEP KP80. *Eur. Respir. J.* **64**, DOI: [10.1183/13993003.congress-2024.OA4678](https://doi.org/10.1183/13993003.congress-2024.OA4678) (2024). Conference Name: Characteristics of lung nodules detected by low-dose CT scan: results from the French lung cancer screening study DEP KP80 Publisher: European Respiratory Society Section: Lung cancer.
28. Bankier, A. A. *et al.* Fleischner Society: Glossary of Terms for Thoracic Imaging. *Radiology* **310**, e232558, DOI: [10.1148/radiol.232558](https://doi.org/10.1148/radiol.232558) (2024). Publisher: Radiological Society of North America.
29. Gierada, D. S. & Pinsky, P. F. Survival Following Detection of Stage I Lung Cancer by Screening in the National Lung Screening Trial. *Chest* **159**, 862–869, DOI: [10.1016/j.chest.2020.08.2048](https://doi.org/10.1016/j.chest.2020.08.2048) (2021).
30. Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A. & Abdani, S. R. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics* **13**, 2617, DOI: [10.3390/diagnostics13162617](https://doi.org/10.3390/diagnostics13162617) (2023).
31. Javed, R. *et al.* Deep learning for lungs cancer detection: a review. *Artif. Intell. Rev.* **57**, 197, DOI: [10.1007/s10462-024-10807-1](https://doi.org/10.1007/s10462-024-10807-1) (2024).
32. Armato, S. G. *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.* **38**, 915–931, DOI: [10.1118/1.3528204](https://doi.org/10.1118/1.3528204) (2011).
33. Ma, L., Wan, C., Hao, K., Cai, A. & Liu, L. A novel fusion algorithm for benign-malignant lung nodule classification on CT images. *BMC Pulm. Medicine* **23**, 474, DOI: [10.1186/s12890-023-02708-w](https://doi.org/10.1186/s12890-023-02708-w) (2023).
34. Mikhael, P. G. *et al.* Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **41**, 2191–2200, DOI: [10.1200/JCO.22.01345](https://doi.org/10.1200/JCO.22.01345) (2023).

35. Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. *IEEE transactions on neural networks learning systems* **30**, 3484–3495, DOI: [10.1109/TNNLS.2019.2892409](https://doi.org/10.1109/TNNLS.2019.2892409) (2019).
36. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Medicine* **25**, 954–961, DOI: [10.1038/s41591-019-0447-x](https://doi.org/10.1038/s41591-019-0447-x) (2019).
37. Li, T. Z. *et al.* No winners: Performance of lung cancer prediction models depends on screening-detected, incidental, and biopsied pulmonary nodule use cases, DOI: [10.48550/arXiv.2405.10993](https://doi.org/10.48550/arXiv.2405.10993) (2024). ArXiv:2405.10993 [q-bio].
38. Winter, A., Aberle, D. R. & Hsu, W. External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. *Thorax* **74**, 551–563, DOI: [10.1136/thoraxjnl-2018-212413](https://doi.org/10.1136/thoraxjnl-2018-212413) (2019).
39. Mohammad, B. A., Brennan, P. C. & Mello-Thoms, C. A review of lung cancer screening and the role of computer-aided detection. *Clin. Radiol.* **72**, 433–442, DOI: [10.1016/j.crad.2017.01.002](https://doi.org/10.1016/j.crad.2017.01.002) (2017). Publisher: Elsevier.
40. Fukumoto, W. *et al.* External validation of the performance of commercially available deep-learning-based lung nodule detection on low-dose CT images for lung cancer screening in Japan. *Jpn. J. Radiol.* DOI: [10.1007/s11604-024-01704-2](https://doi.org/10.1007/s11604-024-01704-2) (2024).
41. Geppert, J. *et al.* Software using artificial intelligence for nodule and cancer detection in CT lung cancer screening: systematic review of test accuracy studies. *Thorax* **79**, 1040–1049, DOI: [10.1136/thorax-2024-221662](https://doi.org/10.1136/thorax-2024-221662) (2024).
42. Jia, R., Liu, B. & Ali, M. Establishing an AI-based diagnostic framework for pulmonary nodules in computed tomography. *BMC Pulm. Medicine* **25**, 339, DOI: [10.1186/s12890-025-03806-7](https://doi.org/10.1186/s12890-025-03806-7) (2025).
43. Baumgartner, M., Jäger, P. F., Isensee, F. & Maier-Hein, K. H. nnDetection: A Self-configuring Method for Medical Object Detection. In de Bruijne, M. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 530–539, DOI: [10.1007/978-3-030-87240-3_51](https://doi.org/10.1007/978-3-030-87240-3_51) (Springer International Publishing, Cham, 2021).
44. Leroy, S. *et al.* Circulating tumour cells as a potential screening tool for lung cancer (the AIR study): protocol of a prospective multicentre cohort study in France. *BMJ Open* **7**, e018884, DOI: [10.1136/bmjopen-2017-018884](https://doi.org/10.1136/bmjopen-2017-018884) (2017).
45. Piskorski, L. *et al.* Malignancy risk stratification for pulmonary nodules: comparing a deep learning approach to multiparametric statistical models in different disease groups. *Eur. Radiol.* DOI: [10.1007/s00330-024-11256-8](https://doi.org/10.1007/s00330-024-11256-8) (2025).
46. Zhu, Y., Yankelevitz, D. F. & Henschke, C. I. How I Do It: Management of Pleural-attached Pulmonary Nodules in Low-Dose CT Screening for Lung Cancer. *Radiology* **314**, e240091, DOI: [10.1148/radiol.240091](https://doi.org/10.1148/radiol.240091) (2025).
47. Jiang, J. *et al.* Differentiation of pulmonary solid nodules attached to the pleura detected by thin-section CT. *Insights into Imaging* **14**, 146, DOI: [10.1186/s13244-023-01504-8](https://doi.org/10.1186/s13244-023-01504-8) (2023).
48. Arai, T. *et al.* Tumor Doubling Time and Prognosis in Lung Cancer Patients: Evaluation from Chest Films and Clinical Follow-up Study. *Jpn. J. Clin. Oncol.* **24**, 199–204, DOI: [10.1093/oxfordjournals.jjco.a039706](https://doi.org/10.1093/oxfordjournals.jjco.a039706) (1994).
49. Nakahashi, K., Nakatsuka, M., Endo, M. & Shiono, S. Tumor volume doubling time as a potential predictor of prognosis in clinical stage I lung squamous cell carcinoma. *J. Thorac. Dis.* **15**, 3849–3859, DOI: [10.21037/jtd-23-292](https://doi.org/10.21037/jtd-23-292) (2023).
50. Xu, D. M. *et al.* Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer (Amsterdam, Netherlands)* **54**, 177–184, DOI: [10.1016/j.lungcan.2006.08.006](https://doi.org/10.1016/j.lungcan.2006.08.006) (2006).
51. Jennings, S. G., Winer-Muram, H. T., Tann, M., Ying, J. & Dowdeswell, I. Distribution of Stage I Lung Cancer Growth Rates Determined with Serial Volumetric CT Measurements1. *Radiology* DOI: [10.1148/radiol.2412051185](https://doi.org/10.1148/radiol.2412051185) (2006). Publisher: Radiological Society of North America.
52. Zhang, R., Tian, P., Qiu, Z., Liang, Y. & Li, W. The growth feature and its diagnostic value for benign and malignant pulmonary nodules met in routine clinical practice. *J. Thorac. Dis.* **12**, 2019–2030, DOI: [10.21037/jtd-19-3591](https://doi.org/10.21037/jtd-19-3591) (2020).
53. Ho, K. M. Effect of non-linearity of a predictor on the shape and magnitude of its receiver-operating-characteristic curve in predicting a binary outcome. *Sci. Reports* **7**, 10155, DOI: [10.1038/s41598-017-10408-9](https://doi.org/10.1038/s41598-017-10408-9) (2017). Publisher: Nature Publishing Group.
54. Klaveren, R. J. v. *et al.* Management of Lung Nodules Detected by Volume CT Scanning. *New Engl. J. Medicine* **361**, 2221–2229, DOI: [10.1056/NEJMoa0906085](https://doi.org/10.1056/NEJMoa0906085) (2009). Publisher: Massachusetts Medical Society _eprint: <https://www.nejm.org/doi/pdf/10.1056/NEJMoa0906085>.
55. Thompson, D. W. *On Growth and Form*. Canto (Cambridge University Press, Cambridge, 1945), 1992 edn.
56. Reed, H. S. & Holland, R. H. The Growth Rate of an Annual Plant *Helianthus*. *Proc. Natl. Acad. Sci. United States Am.* **5**, 135–144, DOI: [10.1073/pnas.5.4.135](https://doi.org/10.1073/pnas.5.4.135) (1919).

57. Rodriguez-Ruiz, A. *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J. Natl. Cancer Inst.* **111**, 916–922, DOI: [10.1093/jnci/djy222](https://doi.org/10.1093/jnci/djy222) (2019).
58. Wu, N. *et al.* Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE transactions on medical imaging* **39**, 1184–1194, DOI: [10.1109/TMI.2019.2945514](https://doi.org/10.1109/TMI.2019.2945514) (2020).
59. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94, DOI: [10.1038/s41586-019-1799-6](https://doi.org/10.1038/s41586-019-1799-6) (2020). Publisher: Nature Publishing Group.
60. Pacilè, S. *et al.* Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. *Radiol. Artif. intelligence* **2**, e190208, DOI: [10.1148/ryai.2020190208](https://doi.org/10.1148/ryai.2020190208) (2020).
61. McWilliams, A. M. *et al.* PL02.14 Triaging ILST Screening Participants at Program Entry: Comparative Performance of PanCan versus LungRADSV1.1 Protocol. *J. Thorac. Oncol.* **19**, S3–S4, DOI: [10.1016/j.jtho.2024.09.016](https://doi.org/10.1016/j.jtho.2024.09.016) (2024).
62. Murchison, J. T. *et al.* Validation of a deep learning computer aided system for CT based lung nodule detection, classification, and growth rate estimation in a routine clinical population. *PLoS ONE* **17**, e0266799, DOI: [10.1371/journal.pone.0266799](https://doi.org/10.1371/journal.pone.0266799) (2022).
63. Armato, S. G. *et al.* Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence on the Definition of “Truth”. *Acad. Radiol.* **16**, 28–38, DOI: [10.1016/j.acra.2008.05.022](https://doi.org/10.1016/j.acra.2008.05.022) (2009).
64. Chen, J. *et al.* CancerUniT: Towards a Single Unified Model for Effective Detection, Segmentation, and Diagnosis of Eight Major Cancers Using a Large Collection of CT Scans. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 21270–21281, DOI: [10.1109/ICCV51070.2023.01950](https://doi.org/10.1109/ICCV51070.2023.01950) (IEEE, Paris, France, 2023).
65. Yang, L. *et al.* Advancing Multimodal Medical Capabilities of Gemini, DOI: [10.48550/arXiv.2405.03162](https://doi.org/10.48550/arXiv.2405.03162) (2024). ArXiv:2405.03162.
66. Saab, K. *et al.* Capabilities of Gemini Models in Medicine, DOI: [10.48550/arXiv.2404.18416](https://doi.org/10.48550/arXiv.2404.18416) (2024). ArXiv:2404.18416.
67. Ma, L. *et al.* A vision–language pretrained transformer for versatile clinical respiratory disease applications. *Nat. Biomed. Eng.* 1–19, DOI: [10.1038/s41551-025-01544-z](https://doi.org/10.1038/s41551-025-01544-z) (2025). Publisher: Nature Publishing Group.
68. Loh, S. E. K. *et al.* CT-guided thoracic biopsy: evaluating diagnostic yield and complications. *Annals Acad. Medicine, Singapore* **42**, 285–290 (2013).
69. Jemal, A. & Fedewa, S. A. Lung Cancer Screening With Low-Dose Computed Tomography in the United States—2010 to 2015. *JAMA Oncol.* **3**, 1278–1281, DOI: [10.1001/jamaoncol.2016.6416](https://doi.org/10.1001/jamaoncol.2016.6416) (2017).
70. Naidich, D. P. *et al.* Recommendations for the Management of Subsolid Pulmonary Nodules Detected at CT: A Statement from the Fleischner Society. *Radiology* **266**, 304–317, DOI: [10.1148/radiol.12120628](https://doi.org/10.1148/radiol.12120628) (2013). Publisher: Radiological Society of North America.
71. Li, Q., Xiao, T., Li, J., Niu, Y. & Zhang, G. The diagnosis and management of multiple ground-glass nodules in the lung. *Eur. J. Med. Res.* **29**, 305, DOI: [10.1186/s40001-024-01904-6](https://doi.org/10.1186/s40001-024-01904-6) (2024).
72. Korb, M. L. & Burt, B. M. The elusive ground glass opacity, revealed. *J. Thorac. Dis.* **10**, S3828–S3830, DOI: [10.21037/jtd.2018.09.41](https://doi.org/10.21037/jtd.2018.09.41) (2018).
73. Herskovitz, E., Solomides, C., Barta, J., Evans, N. & Kane, G. Detection of lung carcinoma arising from ground glass opacities (GGO) after 5 years - A retrospective review. *Respir. Medicine* **196**, 106803, DOI: [10.1016/j.rmed.2022.106803](https://doi.org/10.1016/j.rmed.2022.106803) (2022).
74. Goodman, L. R., Gulsun, M., Washington, L., Nagy, P. G. & Piacsek, K. L. Inherent Variability of CT Lung Nodule Measurements In Vivo Using Semiautomated Volumetric Measurements. *Am. J. Roentgenol.* **186**, 989–994, DOI: [10.2214/AJR.04.1821](https://doi.org/10.2214/AJR.04.1821) (2006). Publisher: American Roentgen Ray Society.
75. Pu, Q., Xi, Z., Yin, S., Zhao, Z. & Zhao, L. Advantages of transformer and its application for medical image segmentation: a survey. *BioMedical Eng. OnLine* **23**, 14, DOI: [10.1186/s12938-024-01212-4](https://doi.org/10.1186/s12938-024-01212-4) (2024).
76. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020).
77. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **35**, 507–520, DOI: [10.48550/arXiv.2207.08815](https://doi.org/10.48550/arXiv.2207.08815) (2022).
78. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ. - Comput. Inf. Sci.* **35**, 757–774, DOI: [10.1016/j.jksuci.2023.01.014](https://doi.org/10.1016/j.jksuci.2023.01.014) (2023).

79. Yang, Y. *et al.* A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions. *Comput. Biol. Medicine* **139**, 104887, DOI: [10.1016/j.combiomed.2021.104887](https://doi.org/10.1016/j.combiomed.2021.104887) (2021).
80. Milner, A. D. & Goodale, M. A. Two visual systems re-viewed. *Neuropsychologia* **46**, 774–785, DOI: [10.1016/j.neuropsychologia.2007.10.005](https://doi.org/10.1016/j.neuropsychologia.2007.10.005) (2008).
81. Munoz, E. *et al.* 3D-Morphomics, Morphological Features on CT Scans for Lung Nodule Malignancy Diagnosis. In Ali, S. *et al.* (eds.) *MICCAI Cancer Prevention Through Early Detection*, Lecture Notes in Computer Science, 3–13, DOI: [10.1007/978-3-031-17979-2_1](https://doi.org/10.1007/978-3-031-17979-2_1) (Springer Nature Switzerland, Singapore, 2022).
82. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, DOI: [10.1038/ncomms5006](https://doi.org/10.1038/ncomms5006) (2014). 00015.
83. Xie, J. *et al.* Differential diagnosis of benign and lung adenocarcinoma presenting as larger solid nodules and masses based on multiscale CT radiomics. *PLOS ONE* **19**, e0309033, DOI: [10.1371/journal.pone.0309033](https://doi.org/10.1371/journal.pone.0309033) (2024). Publisher: Public Library of Science.
84. Wang, M. *et al.* The Value of Topological Radiomics Analysis in Predicting Malignant Risk of Pulmonary Ground-Glass Nodules: A Multi-Center Study. *Technol. Cancer Res. & Treat.* **23**, 15330338241287089, DOI: [10.1177/15330338241287089](https://doi.org/10.1177/15330338241287089) (2024). Publisher: SAGE Publications Inc.
85. Kim, R. Y. *et al.* Clinical Utility of an Artificial Intelligence Radiomics-based Tool for Risk Stratification of Pulmonary Nodules. *JNCI Cancer Spectr.* pkae086, DOI: [10.1093/jncics/pkae086](https://doi.org/10.1093/jncics/pkae086) (2024).
86. Astaraki, M., Zakko, Y., Dasu, I. T., Smedby, O. & Wang, C. Benign-malignant pulmonary nodule classification in low-dose CT with convolutional features. *Phys. Medica: Eur. J. Med. Phys.* **83**, 146–153, DOI: [10.1016/j.ejmp.2021.03.013](https://doi.org/10.1016/j.ejmp.2021.03.013) (2021). Publisher: Elsevier.
87. Fillon, M. Key issues face AI deployment in cancer care. *CA: A Cancer J. for Clin.* **74**, 320–322, DOI: [10.3322/caac.21860](https://doi.org/10.3322/caac.21860) (2024). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21860>.
88. Hantel, A. *et al.* Perspectives of Oncologists on the Ethical Implications of Using Artificial Intelligence for Cancer Care. *JAMA Netw. Open* **7**, e244077, DOI: [10.1001/jamanetworkopen.2024.4077](https://doi.org/10.1001/jamanetworkopen.2024.4077) (2024).
89. Sellergren, A. *et al.* MedGemma Technical Report, DOI: [10.48550/arXiv.2507.05201](https://doi.org/10.48550/arXiv.2507.05201) (2025). ArXiv:2507.05201 [cs].
90. Brock, A., Lim, T., Ritchie, J. M. & Weston, N. FreezeOut: Accelerate Training by Progressively Freezing Layers, DOI: [10.48550/arXiv.1706.04983](https://doi.org/10.48550/arXiv.1706.04983) (2017). ArXiv:1706.04983 [stat].
91. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241, DOI: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28) (Springer International Publishing, Cham, 2015).
92. Jaeger, P. F. *et al.* Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection, DOI: [10.48550/arXiv.1811.08661](https://doi.org/10.48550/arXiv.1811.08661) (2018). ArXiv:1811.08661 [cs].
93. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, DOI: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243) (2017). ISSN: 1063-6919.
94. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785) (Association for Computing Machinery, New York, NY, USA, 2016).
95. Kupperts, F., Kronenberger, J., Shantia, A. & Haselhoff, A. Multivariate Confidence Calibration for Object Detection. *2020 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. Work. (CVPRW)* 1322–1330, DOI: [10.1109/CVPRW50498.2020.00171](https://doi.org/10.1109/CVPRW50498.2020.00171) (2020). Conference Name: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) ISBN: 9781728193601 Place: Seattle, WA, USA Publisher: IEEE.
96. Swensen, S. J., Silverstein, M. D., Ilstrup, D. M., Schleck, C. D. & Edell, E. S. The Probability of Malignancy in Solitary Pulmonary Nodules: Application to Small Radiologically Indeterminate Nodules. *Arch. Intern. Medicine* **157**, 849–855, DOI: [10.1001/archinte.1997.00440290031002](https://doi.org/10.1001/archinte.1997.00440290031002) (1997).
97. Kallner, A. Formulas. In Kallner, A. (ed.) *Laboratory Statistics (Second Edition)*, 1–140, DOI: [10.1016/B978-0-12-814348-3.00001-0](https://doi.org/10.1016/B978-0-12-814348-3.00001-0) (Elsevier, 2018).
98. Chen, W. & Samuelson, F. W. The average receiver operating characteristic curve in multireader multicase imaging studies. *The Br. J. Radiol.* **87**, 20140016, DOI: [10.1259/bjr.20140016](https://doi.org/10.1259/bjr.20140016) (2014).

99. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874, DOI: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010) (2006).
100. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall/CRC, New York, 1994).
101. Pinsky, P. F. *et al.* Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. *Annals Intern. Medicine* **162**, 485–491, DOI: [10.7326/M14-2086](https://doi.org/10.7326/M14-2086) (2015).
102. Bankier, A. A. *et al.* Recommendations for Measuring Pulmonary Nodules at CT: A Statement from the Fleischner Society. *Radiology* **285**, 584–600, DOI: [10.1148/radiol.2017162894](https://doi.org/10.1148/radiol.2017162894) (2017).
103. Lung. In Greene, F. L. *et al.* (eds.) *AJCC Cancer Staging Manual*, 167–177, DOI: [10.1007/978-1-4757-3656-4_19](https://doi.org/10.1007/978-1-4757-3656-4_19) (Springer, New York, NY, 2002).
104. Jennings, S. G., Winer-Muram, H. T., Tann, M., Ying, J. & Dowdeswell, I. Distribution of Stage I Lung Cancer Growth Rates Determined with Serial Volumetric CT Measurements. *Radiology* **241**, 554–563, DOI: [10.1148/radiol.2412051185](https://doi.org/10.1148/radiol.2412051185) (2006).
105. Team, N. L. S. T. R. Data from the National Lung Screening Trial (NLST) [Data set], DOI: <https://doi.org/10.7937/TCIA.HMQ8-J677> (2013).
106. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16, DOI: [10.1038/s41586-020-2766-y](https://doi.org/10.1038/s41586-020-2766-y) (2020). Publisher: Nature Publishing Group.

Supplementary Information Supplementary Figures

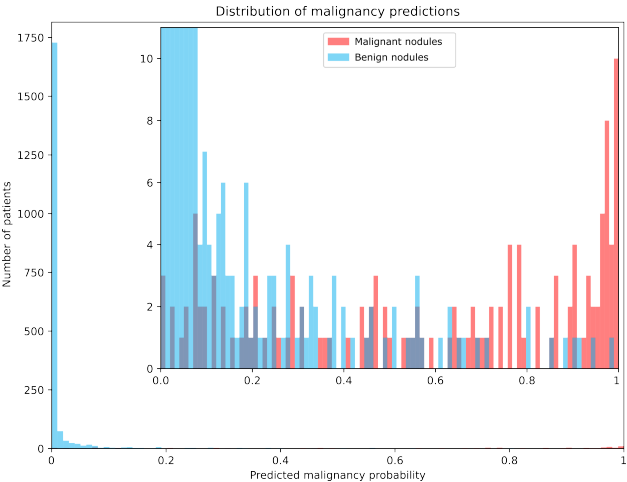


Figure S1. Distribution of model’s patient-level Likelihood Of Malignancy on Test1. Thanks to the calibration¹ of the output prediction on the training set, these output distributions optimally reflect the model’s empirical accuracy.

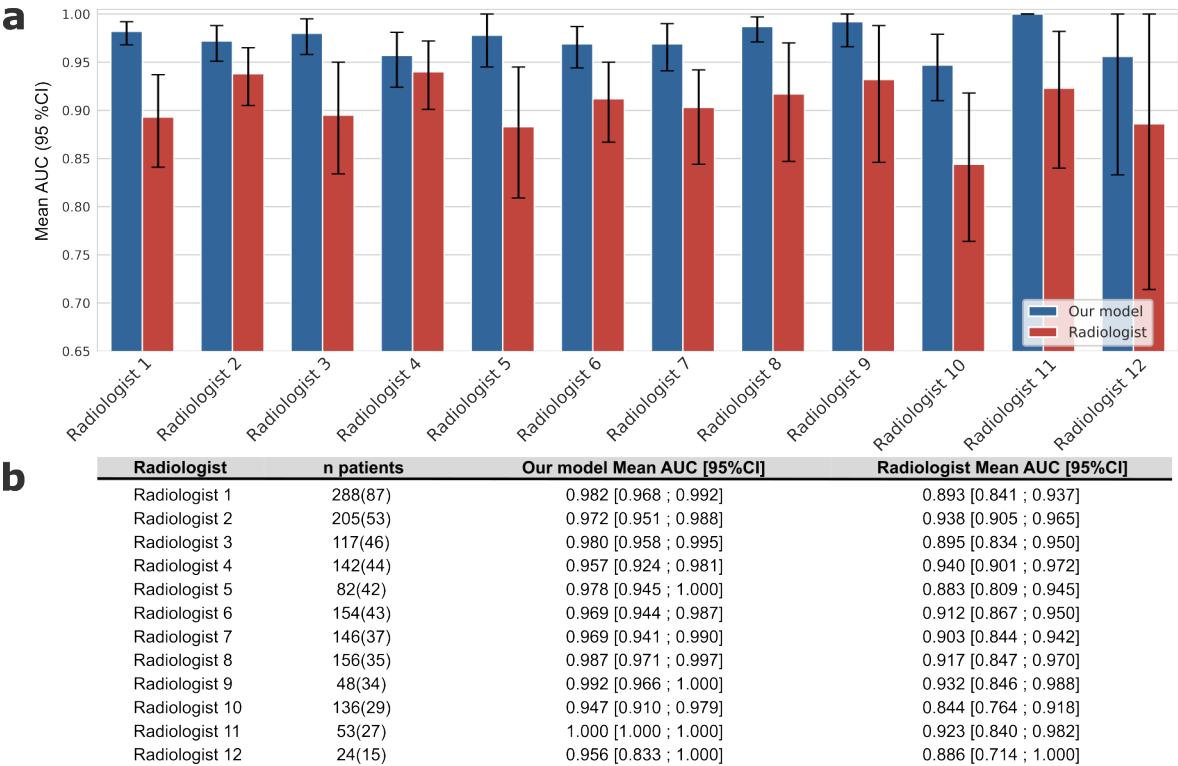


Figure S2. Comparison of mean AUC between our model and individual radiologists: a. Mean AUCs over 5,000 bootstraps of our model and of each of the 12 radiologists on the same scan subsets of Test3, as shown in Fig.3. Vertical bars indicates the 95% CI. **b.** Values of mean AUCs and 95% CI illustrated in **a**. The sample size (n: number of patients) for each annotator with the number of cancer cases indicated in parenthesis. The Welch t-tests on AUC distributions over the 5,000 bootstraps confirm our model’s non-inferiority of our model AUC over radiologists in all cases, with $p < 0.0001$.

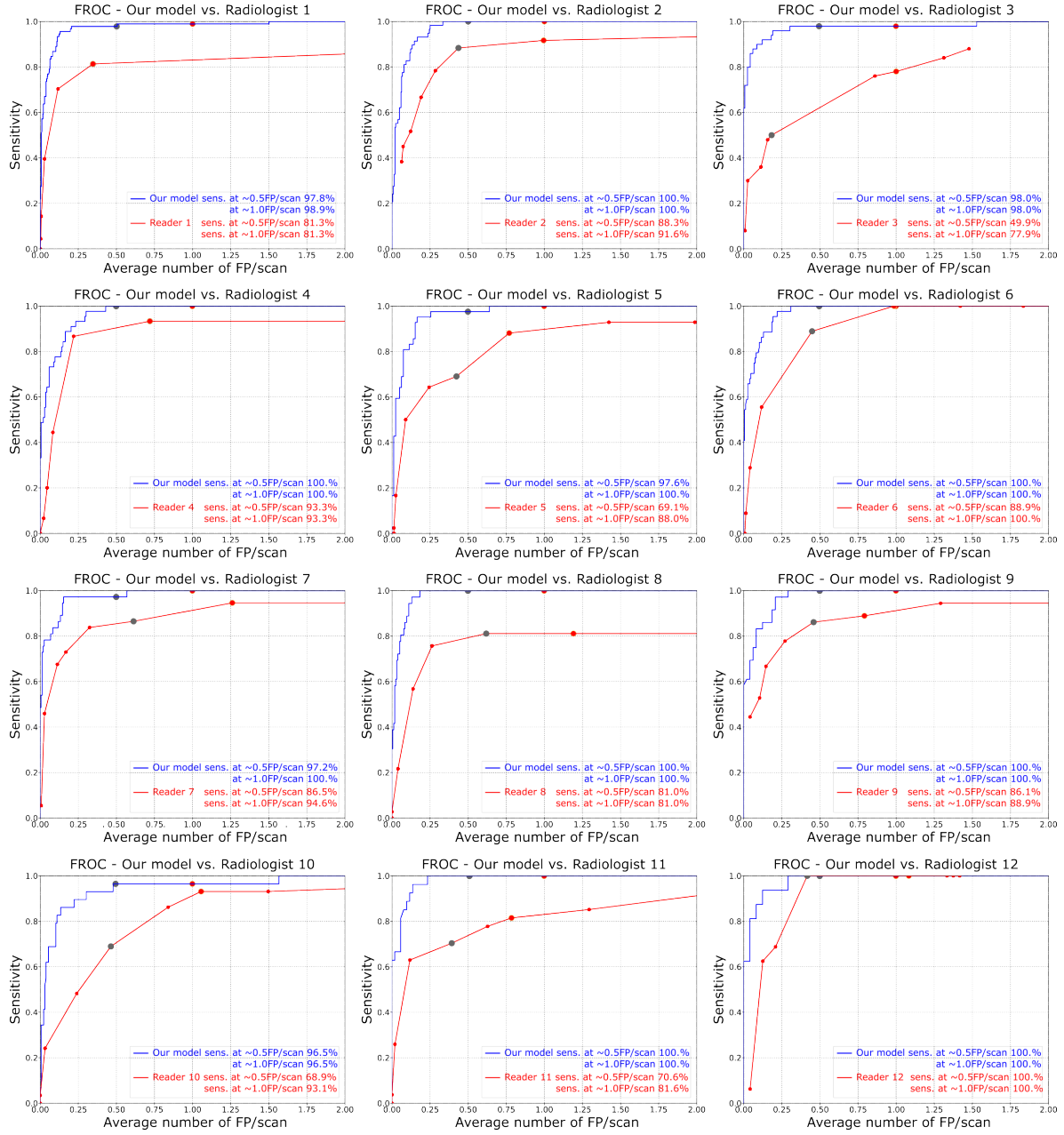


Figure S3. Detection performance: our model vs. individual radiologists The 12 FROCs of our model and of the 12 radiologists who annotated at least 15 cancer patients for the same scans subsets of Test3, as detailed in Supplementary methods. Mean sensitivity over 5,000 bootstraps at the closest OP to 0.5 and 1 FP/scan is provided in labels, marked by a large grey and a large red circle, respectively.

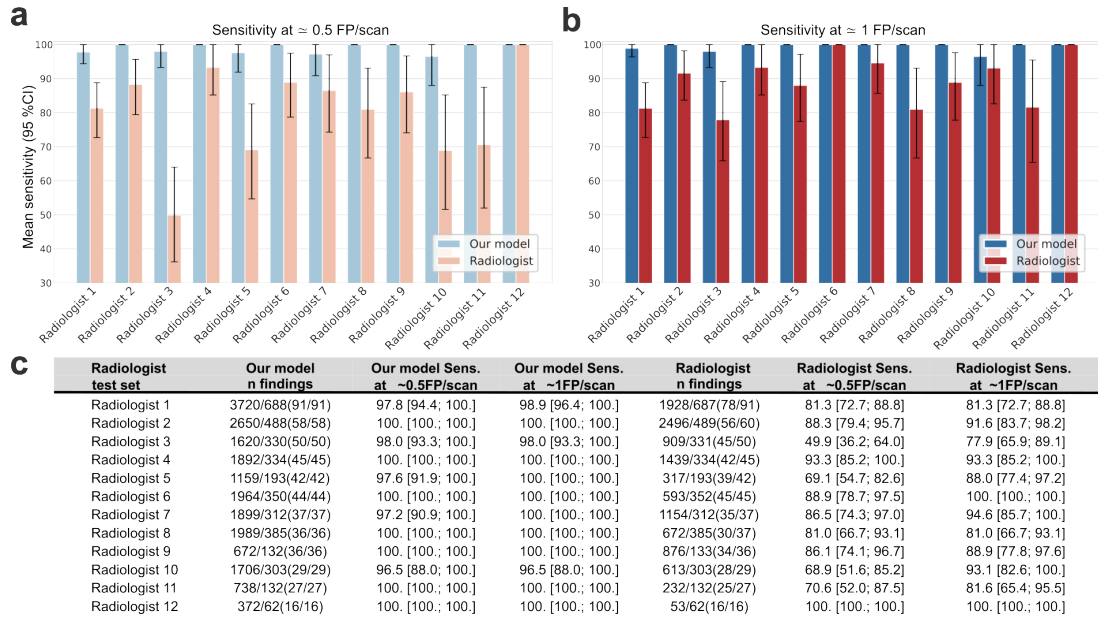


Figure S4. Detection Sensitivity comparison at fixed FP/scan: our model vs. individual radiologists Mean sensitivity over 5,000 bootstraps at the closest OP to ≈ 0.5 in **a.** and at the closest OP to ≈ 1 FP/scan in **b.**, for our model and the 12 radiologists on the same subsets of Test3 corresponding to the FROCs shown in Supplementary Fig.S3. Vertical bars indicate the 95% CI. **c.** Values of mean sensitivity and 95% CI illustrated in **a.** and **b.** The sample size (n: number of findings: TP+FP of CADe) and number of findings in GT (CADe TP+FN of CADe) for each annotator and the model are provided, with malignant finding cases indicated in parenthesis. The Welch t-tests on sensitivity distributions demonstrate the non-inferiority of the sensitivity of our model over each radiologist in all cases, with $p < 0.0001$ both at 0.5 and 1FP/scan—except for Radiologist 12 whose small sample size leads to saturation in this range. Notably, at lower FP/scan our model still outperforms Radiologist 12, as shown in Supplementary Fig.S3. Radiologist's count of finding in GT can slightly exceed that of our model in GT (which matches the number of nodules in GT for malignant nodules), since radiologists may detect multiple findings with different scores for one same nodule in GT. Additionally, the closest OPs to 0.5 and 1 FP/scan on the radiologists' curves can sometimes deviate far from 0.5 and 1 FP/scan value, as shown in Supplementary Fig.S3, due to malignancy scores being discretized on a 1–10 scale, limiting direct OP comparisons. Finally, the number of findings retrieved by our model, as presented in the table, may be excessively large and arbitrary, and should be restricted in practice by setting a threshold to a low FP/scan threshold (e.g., 2 FP/scan or below) without compromising performance, especially in the asymptotic constant regime.

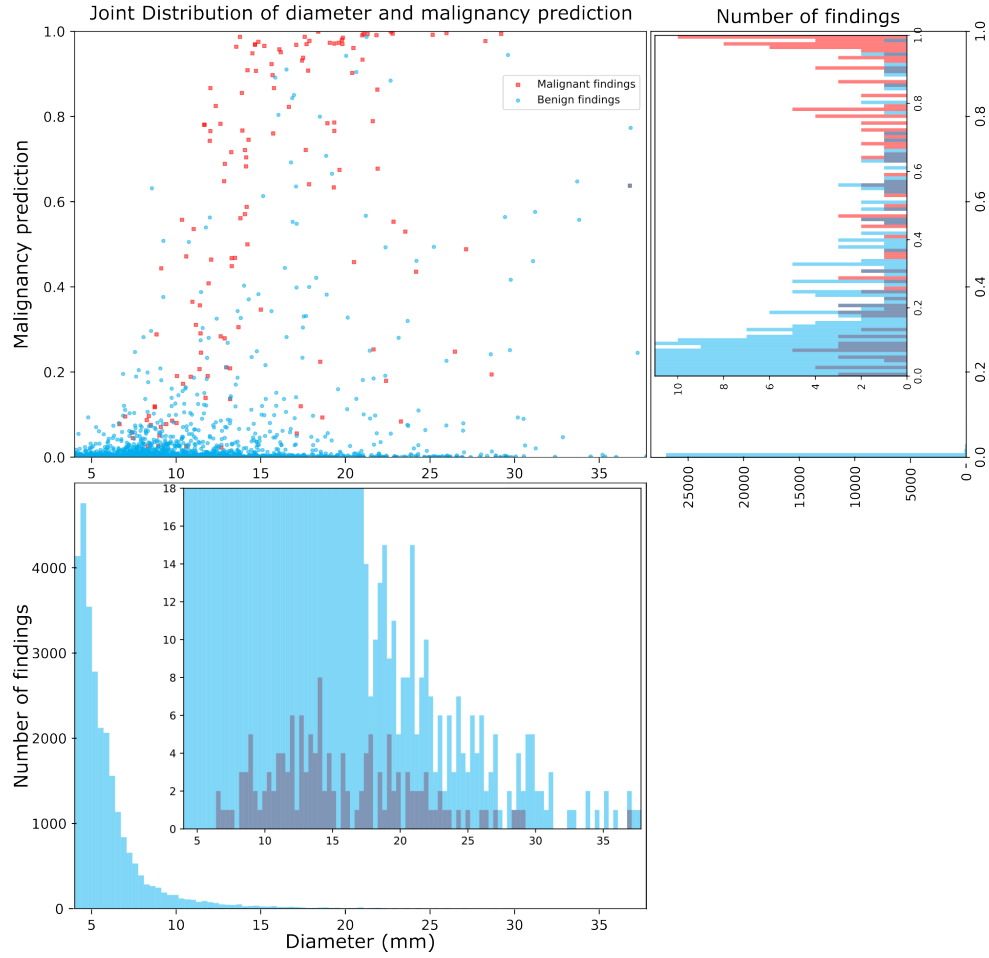


Figure S5. Diameter and malignancy prediction joint-distributions and marginals: The diameter (mm) estimated from our model's automatic segmentation vs. our model's malignancy prediction for all our model detections on Test1 (with diameter greater or equal to 4 mm). The corresponding marginal distribution histograms of diameters and malignancy prediction are provided in the bottom and right panel, respectively. Due to the large imbalance between malignant and benign findings in Test1, causing cancer distributions to be nearly invisible, a zoomed-in cartoon on the distribution of the model's predictions is included in both marginal distributions.

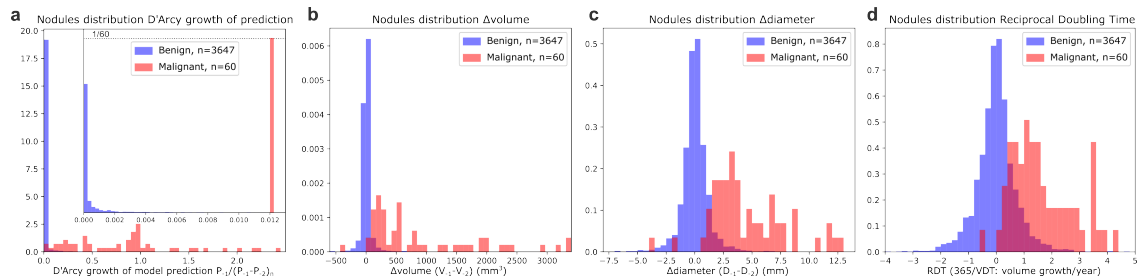


Figure S6. Distributions of Malignant and Benign probability corresponding to the ROCs in Fig.6a: of D'Arcy Thompson growth of our model prediction with a zoom on small values (a), of Δ volume (b), of Δ diameter (c) and of RDT (d). For clarity, Benign and Malignant distributions are independently normalized to an integral of 1, and for Δ diameter distributions, two nodules with outlying large delta were excluded.

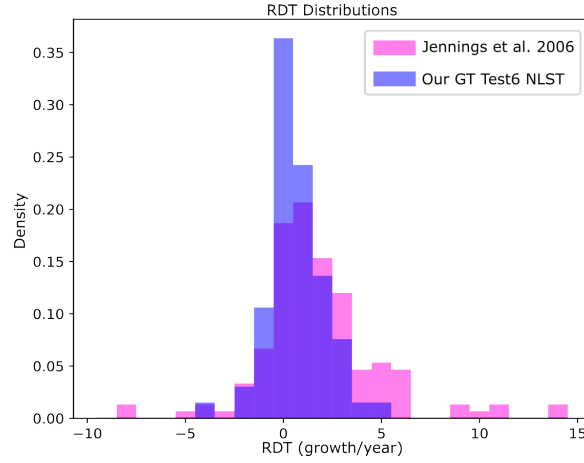


Figure S7. Comparison of the distribution of cancer patient RDT of Jennings *et al.*² with the RDT of our GT over the nodules longitudinally paired in GT of Test6 (redrawn from²).

	NOT GROWING/NOT SUSPICIOUS		GROWING/SUSPICIOUS	
Δ Diameter (Lung-RADS)	D(T₋₁)-D(T₋₂) < 1.5mm		D(T₋₁)-D(T₋₂) ≥ 1.5mm	
Our model T ₋₁ mean AUC	0.969 [0.944,0.987]		0.977 [0.941,1.000]	
Our model T ₋₂ mean AUC	0.864 [0.784,0.929]		0.824 [0.696,0.929]	
4 Radiologist mean T ₋₁ mean AUC	0.887 [0.853,0.918]		0.856 [0.799,0.904]	
Volume (NELSON)	Volume Growth < 25%		Volume Growth ≥ 25%	
Our model T ₋₁ mean AUC	0.960 [0.930,0.983]		0.981 [0.957,0.996]	
Our model T ₋₂ mean AUC	0.829 [0.723,0.913]		0.872 [0.799,0.933]	
4 Radiologist mean T ₋₁ mean AUC	0.861 [0.816,0.900]		0.902 [0.870,0.930]	
NELSON growth category	GROWCAT A: VDT>600		GROWCAT B: 400≤VDT≤600	GROWCAT C: VDT<400
NELSON screening outcome	NEGATIVE		INDETERMINATE	POSITIVE
Our model T ₋₁ mean AUC	0.987 [0.942,1.000]		0.982 [0.912,1.000]	0.980 [0.937,1.000]
Our model T ₋₂ mean AUC	0.916 [0.749,1.000]		0.928 [0.800,1.000]	0.768 [0.587,0.917]
4 Radiologist mean T ₋₁ mean AUC	0.802 [0.706,0.885]		0.934 [0.882,0.976]	0.942 [0.904,0.972]

Figure S8. Mean ROC-AUC and 95% CI over 5,000 bootstraps samples for prediction by our model at T₋₁, at T₋₂ and the mean of 4-Radiologists at T₋₁ on the same intersecting subgroup of Test6 and Test3, for the various subgroups of growing nodules category, Δdiameter below or above 1.5 mm (Lung-RADS®), volume growth below or above 25% (NELSON³) and with VDT above 600, between 400 and 600 or below 400 days, corresponding to the NELSON's GROWCAT A,B and C categories, respectively³.

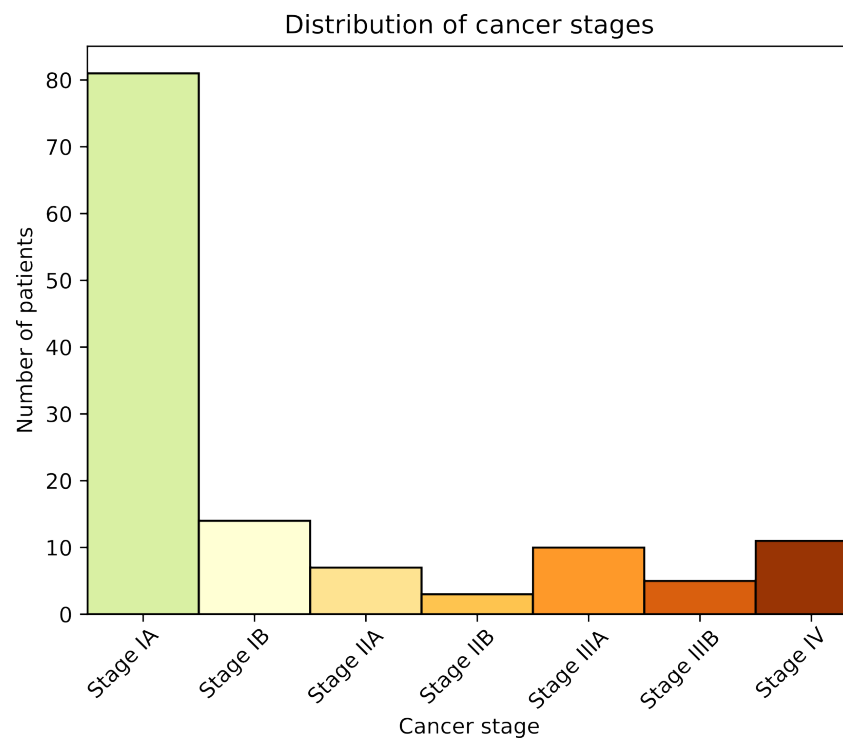


Figure S9. The distribution of cancer stages across patients in Test1.

Supplementary methods

Data

Annotations and GT

NLST GT annotation campaign: The annotation campaign was conducted by Quantified Imaging, a medical consulting and annotation company. The annotations were conducted by one expert medical doctor specialized in chest radiology and one expert radiologist, both of whom followed a predefined annotation protocol (summarized below). The clinician annotations process consisted of two main tasks:

- **Task of nodule localization, segmentation and longitudinal tracking:** This task consists in detecting and semi-manually segmenting the solid component of all segmentable nodules that can be identified in the scan. It also includes linking longitudinally identical nodules across multiple TPs. Radiologists completed this task with LMS software (FDA approved, commonly used for VDT quantification⁴), that provides a semi-automatic "one-click" segmentation of nodule solid components, originally designed for RECIST quantification. When a radiologist disagreed with the semi-automatic segmentation, they were required to adjust the segmentation manually.
- **Disambiguation task:** This task consists in linking each annotated nodule to its corresponding entry in the NLST "Spiral CT Abnormalities dataset". Each nodule not referenced in NLST records was annotated with a new GT label. For cancer patients at the time of diagnosis, malignant nodules were localized according to biopsy-confirmed outcome (size, grade, position) and other characteristics (z position, lobe position) provided by NLST. When multiple highly suspicious nodules were found at the cancer location indicated by NLST, or when no site was specified but a highly suspicious nodule was present elsewhere, the radiologist also classified the nodule as malignant.

As a result of this annotation protocol, an augmented pool of nodules is localized in each scan, associated with a mask, tracked across TPs under a unique ID, and linked to their NLST GT (or assigned a new GT if not previously recorded in NLST). Among the GT entries for each abnormality in NLST (notably in the "Abnormalities" dictionary in NLST), the 'Abnormality description' field (*set ab desc*) specifies lesion types, which can be classified into clinically distinct abnormality categories:

- Parenchymal lung nodular lesions: 51="Non-calcified nodule or mass (opacity $\geq$ 4 mm diameter)"; 52="Non-calcified micronodule(s) (opacity $<$ 4 mm diameter)"; 53="Benign lung nodule(s) (benign calcification)"; 62="6 or more nodules, not suspicious for cancer (opacity $\geq$ 4 mm)";
- Parenchymal lung lesion non-solid or with sometimes pseudo-nodular appearance: 54="Atelectasis, segmental or greater"; 58="Consolidation"; 61="Reticular/reticulonodular opacities, honeycombing, fibrosis, scar";
- Parenchymal Lung lesion without nodular shape: 59="Emphysema"; 55="Pleural thickening or effusion";
- Non parenchymal lesions: 56="Non-calcified hilar/mediastinal adenopathy or mass ($\geq$ 10 mm on short axis)"; 57="Chest wall abnormality (bone destruction, metastasis, etc.)"; 60="Significant cardiovascular abnormality"; 63="Other potentially significant abnormality above the diaphragm"; 64="Other potentially significant abnormality below the diaphragm"; 65="Other minor abnormality noted".

For annotation convenience, all NLST abnormality types 52 and 62, as well as newly-discovered nodular parenchymal lesions were pooled in a single class, labeled 50.

NLST LIDC-like radiologist assessment annotation campaign: In a second step, we conducted a radiologist annotation campaign covering Test3—a subset of Test1—to evaluate performance on a malignant nodule detection task equivalent to a CADe/CADx assessment. Independent readers—practicing radiologists with a valid medical license and at least one year of experience interpreting chest CT scans—were recruited through an independent company, Ingedata, which managed the annotation campaign. To optimize flexibility, cost and time of the annotation campaign, we adapted the k-fold Split-Plot Multi-Reader Multi-Case (MRMC) protocol⁵. The new protocol relaxes the constraint that each reader assess an identical number of cases, thereby accommodating individual availability, while ensuring each case/scan is annotated independently exactly four times, by different readers. Although statistically less robust than classical MRMC design, this protocol is sufficient for our current objectives and claims; a comprehensive MRMC study is ongoing and will be published separately. Notably, as stated in the Methods ROC section, the mean ROC and AUC calculations across radiologists assume radiologist equivalence and permutation invariance. As a result, Test3 was annotated by 20 different readers, each assessing between five and 297 patients. The distribution of scan-patient counts per annotator is provided in Fig.S10. This dataset (named LIDC-annot-NLST501) is made publicly available with the publication of this paper.

Following the seniority classification proposed in⁶—which defines junior radiologists as those with $\leq$ 2 years postgraduate experience, mid-level seniors with 2–4 years, and seniors with $\geq$ 4 years—the 20 radiologists were distributed into the following

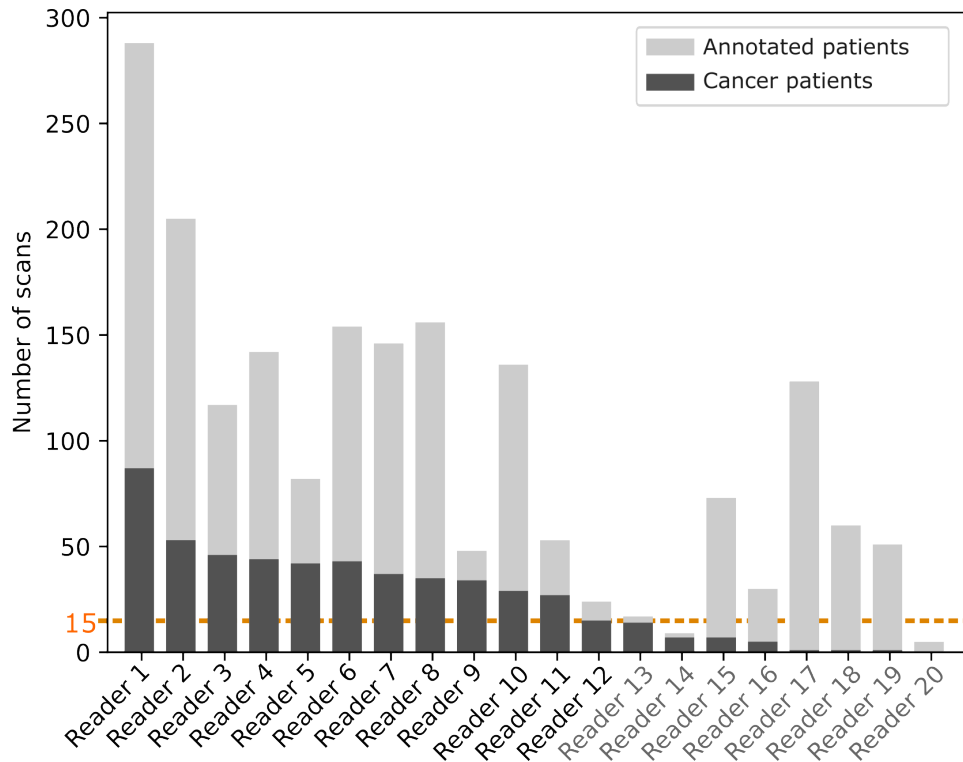


Figure S10. Number of annotated scans per reader: Readers are ranked by number of annotated cancer cases. The 12 readers who annotated at least 15 scans with cancer and retained for the performance comparison to single reader, are labeled in black; all others are labeled in grey.

cohorts: 8 juniors; 7 mid-level seniors; and 5 seniors. This distribution of postgraduate experience across the 20 readers, illustrated in Fig.S11, exhibited no significant correlation with the number of scans annotated per reader (correlation coefficient $\rho = -0.208$, $p=0.38$).

The annotation protocol was designed to align with and adhere to the LIDC-IDRI annotation protocol and methods, as the LIDC-IDRI dataset has become a benchmark in the field and its protocol was developed by a panel of expert radiologists^{7,8}. Radiologists were blinded to the NLST clinical data and instructed to detect nodules and provide a subjective assessment of the likelihood of malignancy for each nodule, assuming the scan originated from a 60-year-old male smoker. Malignancy score ranges from one to ten, with the following definition:

- 1: Definitely benign nodules and Benign Appearance or Behavior Nodules with a very low likelihood of becoming a clinically active cancer due to size or lack of growth;
- 2 to 3: Probably benign finding(s) - short term follow up suggested; includes nodules with a low likelihood of becoming a clinically active cancer;
- 4 to 5: Suspicious Findings for which additional diagnostic testing is recommended;
- 6 to 10: Very Suspicious Findings for which additional diagnostic testing and/or tissue sampling is recommended or necessary;

The assessment from 2 to 3, 4 to 5, and from 6 to 10 is purely subjective, representing the perceived degree of suspicion based on the visual features observed.

Lung-RADS® radiologist assessment and Ardila *et al.* model prediction: We use the data provided by Ardila *et al.*⁹ in their supplementary material, that comprises their model's predictions alongside Lung-RADS® 2019 assessments of lesions by six radiologists, both with and without prior TP information (see⁹ for the methodology). Comparisons with our model's predictions was carried out at the patient level on the same patient set. To further enhance comparability, in the minority of cases where Ardila *et al.* did not use the same TP (for non-cancer patients) or series (kernel) as our selection, we substituted with an alternate kernel from the same TP where available. If not, we used the scan from the closest available TP.

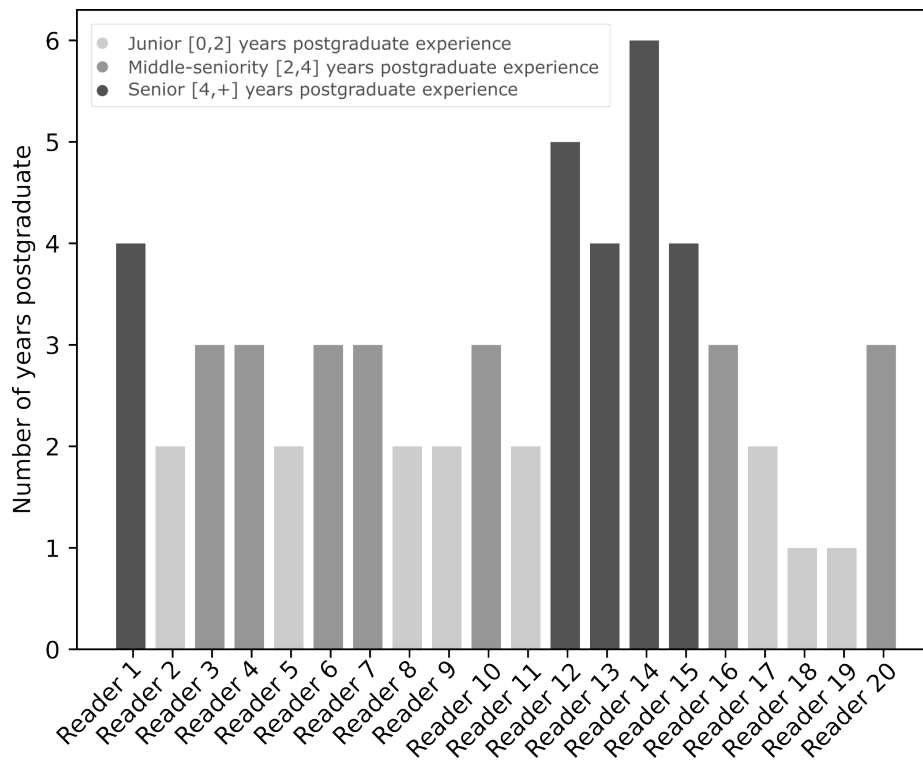


Figure S11. Reader experience and seniority: Number of postgraduate years of radiologist experience per reader, categorized into Junior, Middle-Seniority, and Senior.

Independent cohort GT annotation campaign: The IC was assembled from three data sources: images, paired with semi-structured clinical data collected by the AIR study (EU)¹⁰, in conjunction with images collected by two US data providers. To ensure a high degree of precision, the annotation protocol implemented a consensus truthing process based on three readers, and proceeded in three stages. However, this protocol could not be strictly identical to the NLST protocol due to discrepancies in available clinical information. Initially, two concurrent readers performed localization and measurement of all nodules and abnormalities, identifying whether or not these constituted lesions of interest (i.e., within the [4,40]mm range). When present, each reader was required to specify solidity and margins. For cancer patients, readers were required to classify the nodule as a candidate or confirmed cancer (or rule out that possibility) based on the available clinical data. For each patient, readers were to deliver no more than 15 lesions. A second stage addressed discrepancies among the initial concurrent readings. Following lesion pairing, an independent reader (hereafter adjudicator) reviewed each annotation conflict, confirming or rejecting nodules by location. Upon accepting a single-read nodule, the adjudicator also recorded their own assessment of lesion characteristics. A separate adjudicator performed an independent review of the lesion characteristics alone. As a result nodule location and characteristics were established by consensus of at least two of the three radiologists (2/3). In the final stage, all cancer-confirmed or candidate nodules, along with a representative sampling of benign nodules, underwent concurrent segmentation by the three readers. The GT segmentation was defined by a two-thirds consensus across the three segmentation masks. All readers were recruited and overseen by Ingedata, and were current practicing radiologists with a minimum of two years' experience interpreting chest CT. Adjudicators were selected on the basis of seniority.

0.0.1 Inclusion/exclusion

NLST inclusion/exclusion :

Exclusion of unavailable CT scans: The CT arm of NLST, excluding all X-Ray images, comprised 26,722 patients. Of these, only 12,810 patients—including 1,047 cancer patients and 11,763 non-cancer patients with nodules—were eligible for download without failure.

Kernel selection: For all those patients, for each of the 3 TP scans (T_{-1} , T_{-2} , T_{-3})—when available—only a single CT scan was selected among the multiple scans acquired for each TP with varying kernels and acquisition parameters (up to ten scans per patients per TP). The selection criteria for this single scan closely follows those established by Ardila *et*

al.⁹: prioritizing Hard Kernels (best signal-to-noise ratio for parenchymal nodule detection and characterization); eliminating localizers; excluding scans with insufficient resolution or number of images. The exact inclusion/exclusion rules for kernel selection are as follows⁹: (1) Exclude all localizers; (2) For each case, select the highest ranking kernel according to the following lists (no cases involve more than one manufacturer):

- Siemens: (1) B50f; (2) B45f; (3) B50s; (4) B40f; (5) B41s; (6) B60f; (7) B60s; (8) B70f; (9) B36f; (10) B35f; (11) B30f; (12) B31s
- GE: (1) LUNG; (2) BONE; (3) BODY FILTER/BONE; (4) STANDARD; (5) BODY FILTER/STANDARD; (6) SOFT; (7) EXPERIMENTAL7; (8) BODY FILTER/EXPERIMENTAL7
- Philips: (1) D; (2) C; (3) B; (4) A
- Toshiba: (1) FC51; (2) FC50; (3) FC52; (4) FC53; (5) FC30; (6) FC11; (7) FC10; (8) FC82; (9) FL04; (10) FC02; (11) FC01; (12) FL01

Scan Quality Control (QC) exclusion: Scan QC implements the following inclusion/exclusions:

- **Scan slice thickness:** scans with slice thickness greater or equal to 5 mm were excluded from the study.
- **Sufficient lung coverage:** whenever a scan did not have sufficient number of images, indicating incomplete coverage of the lung, we replaced it by the next best available based on kernel ranking in the previous list and scans without suitable replacement were excluded. The QC exclusion was performed manually on the set of scans suspected of covering less than 25 cm of the body (assuming lung size should be greater than 25 cm).
- **Prone position:** CT scans acquired in the prone position were excluded
- **Artifacts:** scans with artifacts in the lung altering the image quality were excluded.
- **Incorrect scan reconstruction:** scans with missing slices were excluded.

Parenchymal nodule and truthing inclusion/exclusion: As a result, each patient contributed a single selected, readable scan for each TP. In total 9,183 non-cancer cases and 695 cancer cases were submitted for annotation to Quantified Imaging (the contracted annotation company). For Train3 (CADx training set) and Test1, we included one additional exclusion criterion following the annotation, to retain only correctly localized and diagnosed malignant parenchymal nodules. Among cancer patients, 600 were identified as having at least one disambiguated malignant lesion: for 95 patients, it proved infeasible to detect the cancerous lesion on the scan. This was due to a number of factors, including: discrepancy between clinical diagnostic information and CT scan data (e.g., size or location); the time lapsed between the invasive diagnostic procedure confirming the cancer and the date of scan acquisition sometimes exceeded one year; and additional image quality issues (e.g., motion artifact) that could render the CT scan reading inconclusive.

For nodule definition, we adhered to the guidelines of the Fleischner society¹¹. Our study focuses exclusively on parenchymal nodules, so it includes/excludes the following:

- **Solid and part-solid inclusion:** patients with solid and part-solid malignant nodules were included; patients with pure GGO malignant lesions were excluded.
- **Parenchymal inclusion:** patients with parenchymal malignant nodules were included; patients with mediastinal or hilar masses/malignant lesions were excluded.
- **Diameter range inclusion [4,30]mm:** for the tests sets, patients with malignant nodules in the range of [4,30]mm diameter on the segmentation GT were included, all patients with malignancies outside this range were excluded. For the training sets, a broader inclusion range of [3,40]mm diameter was considered, in order to prevent performance degradation and enhance generalizability.

In the original lesion classes ontology of NLST introduced in previous section, this inclusion corresponds to a lesion selection of type (*sect ab desc*) 51, 53, 62, and 50. As a result of these inclusion/exclusion criteria based on intended use, and the union of the Train3 and Test1, we obtained a cohort of 448 cancer patients and 9,183 non-cancer patients.

For Train1 (the CADe training set) only patients with purely ground glass nodules were excluded, resulting in 543 cancer patients and 7,156 non-cancer patients.

TP selection: for Train1 (CADe) all available TPs were selected. For Train3 and all tests sets, among all available TPs for each cancer patient, we select the scan acquired at time of diagnosis, if available, or otherwise the nearest scan acquired

within one year prior to the diagnosis confirmation. For non-cancer patients, the earliest available TP was selected to maximize follow-up duration and ensure cancer-free status, while also increasing the sample size of non-cancer in LCS programmes due to the high dropout rates frequently observed during such programmes.

For longitudinal analyses and for similar reasons, for non-cancer patient, only the first two consecutive TPs available (approximately one year apart) were selected. For cancer patients, only those with scans acquired within [0, 12] months prior to time of diagnosis (T_{-1}), and [12, 24] months prior to time of diagnosis (T_{-2}) were included. Accordingly, patients diagnosed with cancer during the first year of the LCS programme.

LIDC inclusion/exclusion :

In line with previous studies, we only considered nodules annotated by at least four radiologists with a diameter greater than or equal to 3 mm.

For Train3 (CADx training): LIDC-IDRI also provides some additional clinical information for a subset of patients, including biopsy and resection results for some cancer cases, and follow-up data for some non-cancer cases. Using these data, we were able to isolate 77 cancer cases with biopsy/resection histopathological confirmation, and 36 non-cancer cases, i.e., consistent with our NLST inclusion protocol. These confirmed patients were incorporated into Train3.

For Train2 (CADx pretraining set): For each nodule, a radiologist rated malignancy on a scale from 1 to 5. We included all patients and lesions (non GG0, [3, 40]mm mean diameter range) that were not included in Train3 (i.e., not confirmed by biopsy). The final malignancy labels were obtained by taking the median value of all radiologists' ratings. Nodules with a median of 3 were excluded, as they could not be reliably classified as benign or malignant. This selection results in a total of 656 nodules, of which 275 were labelled as malignant (54%), and 268 as benign (46%) with weak annotation GT.

For Train1 (CADE training): we included all patients and lesions (non GG0, [3, 40]mm mean diameter range) applying the same method and rules as Train2.

IC inclusion/exclusion :

The IC population was compiled from three sources. The first consisted of 133 patients (18 cancer patients) enrolled in the French AIR¹⁰ that investigates the predictive power of blood biomarkers for lung cancer in a LCS population with COPD, following the clinical inclusion criteria provided. The second dataset included 276 US patients (92 cancer cases) purchased from the data provider SEGMed. 412 patients (199 cancer patients), also from the US, acquired via Gradient; of these, 66 patients (36 cancer cases) were reserved for a separate study. The third comprised 412 patients (199 cancer cases), also from the US, acquired via the data provider Gradient; of these 66 patients (36 cancer cases) were reserved for a separate study. Each subset underwent a two-stage screening: first, verification for compliance with our clinical protocol; second, quality control of imaging data. Following this, scans were annotated and disambiguated. The inclusion criteria used were the following:

- aged 50 to 80 years;
- current or former smoker (≥ 20 pack-years).
- underwent patient screening and surveillance as part of a LCS programme; adhering to LCS guidelines.
- underwent LDCT screening following classification as high-risk for lung cancer.

Benign status was verified as follows:

- Evidence of nodule stability for at least 12 months, documented either by imaging (follow-up scans showing no progression or regression), and/or radiological reports, explicitly stating stability or reduction in size.
- Negative tissue sampling within 12 months of LDCT screening scan.

Cancer status was verified as follows:

- with proof of cancer within 12 months of a LDCT screening exam. Either from extraction of histopathology reports, or from mention of primary lung cancer in radiological reports (or corresponding codes from International Classification of Diseases for lung neoplasm—excluding carina, hilum, mediastinum, and bronchus cancers, or specific to lung cancer cell types).
- additional information eligible to derive cancer diagnosis, such as mention of diagnostic procedures (Cytopathology, Biopsy, etc.), mention of treatment (resection, excision, chemo/radiotherapy, etc.), or relevant patient aftercare follow-up.

The various steps of the data selection process, as illustrated in Fig.7 of the article are: **Clinical QC exclusion:** the exclusion criteria are as follows:

- history of lung resection;
- pacemaker or other intra-thoracic indwelling metallic medical devices that interfere with CT acquisition;
- prior inclusion (patients/images) in AI model development;
- patients with only mediastinal cancer(s);
- patients with only ground glass cancer(s);
- patients with nodules, solid or part-solid, >30 mm (masses).

Scan QC exclusion: Each selected image was manually reviewed for consistency with the corresponding electronic case report forms. Scans exhibiting any of the following were excluded:

- slice thickness >3 mm;
- missing slices;
- partial coverage of the lung;
- artifacts;
- contrast injection.

Patients without any image suitable for analysis, were excluded. **Truthing exclusion:** cases were rejected when the radiologist responsible for establishing the GT was unable to resolve ambiguity regarding either the location or malignancy of suspicious lesions. In some instances, lesions observed in presumed benign cases appeared highly suspicious, prompting a verification process before definitive inclusion. These cases were flagged for clarification with the original data providers. If the site could not confirm the lesion's status—whether malignant or benign—the patient was excluded during truthing. **Intended use exclusion:** the following exclusion criteria were applied:

- patients with only mediastinal cancer(s);
- patients with only ground glass cancer(s);
- patients with nodules, solid or part-solid >30 mm (masses).

Time Point Selection: For cancer patients, the annotated scan (TP) is the nearest LDCT scan acquired within 12 months before the date of diagnosis. In benign cases, the annotated scan is the earliest available LDCT for LCS. Scan selection followed NLST criteria: original axial; prone position; most lung-like kernel with slice thickness closest to 1 mm; at most 5 mm; and with less than 1,000 slices. Note that any deviation from the above criteria triggered a query to the originating site (e.g., for missing information). Any such queries (whether clinical or imaging related) remaining unresolved at the time of analysis led to the exclusion of the patient.

Data analysis

Competition Performance Metric (CPM):

The CPM averages sensitivity over a predefined list of FPs/scan thresholds. It was designed to enable consistent comparison by summarizing performance across multiple FROC OPs, and was notably used for the LUNA16 challenge¹². The list of FP/scan thresholds used for LUNA16 challenge is [1/8, 1/4, 1/2, 1, 2, 4, 8], and we use this definition in order to compare with LUNA16 challenge results. We adopted the same FP/scan thresholds used in LUNA16—[1/8, 1/4, 1/2, 1, 2, 4, 8]—for direct comparability with LUNA16 results.

0.0.2 Subgroup analysis:

As shown in Fig.2 of the article, the NLST and IC datasets exhibit very different slice thickness distributions, owing to NLST comprising older scans typically acquired at greater slice thickness. To account for this, we applied dataset-specific stratifications: [0.5, 1.5]mm, [1.5, 2.3]mm, [2.3, 3.5]mm for NLST; and [0.5, 0.8]mm, [0.8, 1.5]mm, [1.5, 3]mm for IC.

Due to insufficient sample size for performance assessment, particularly in the IC cohort ($n = 2$), Canon and Philips were not reported in the subgroup analysis. Additionally, 'Siemens healthineer' and 'SIEMENS' were pooled together.

Sharp, average, and soft kernel subgroups used in Fig.2 of the article were defined according to manufacturers-provided specifications:

- Extra Sharp siemens kernel list: ["I70f"; '1'], ["I70f"; '2'], ["I70f"; '3'], ["I80s"; '1'], ["I80s"; '2'], ["I80s"; '3'], ["I80s"; '4'], ["I80s"; '5'], "B70f", "B70s", "B75f", "B75h", "B75s", "B80f", "B80s", "B90s", "I70f", "I80s", "Hr68f"]
Extra Sharp GE kernel list: ["BONE", "Bone2", "BONEPLUS2", "BODY FILTER/STANDARD", "BODY FILTER/STANDARD", "BONEPLUS", "BODY FILTER/BONE", "BODY FILTER/EXPERIMENTAL"]
Extra sharp toshiba list: ["FC55", "FC56", "FC59", "FC65", "FC57", "FC58", "FC80", "FC81", "FC82", "FC86"]
- Sharp Siemens list: ["I50f", "B50s", ["I50f"; '1'], ["I50s"; '2'], "I50s", ["I50s"; '1'], "B50f", ["I50f"; '2'], ["I50f"; '3'], ["I50s"; '3'], ["BI54f"; '2'], ["BI54f"; '1'], ["Br54d"; '2'], ["BI54d"; '1'], "BI54d", ["BI54d"; '2'], ["BI54d"; '3'], "Br54d", ["Br54d"; '1'], ["BI54f"; '3'], "BI54f", ["Br54d"; '3'], ["BI56f"; '1'], ["BI56f"; '2'], ["BI56f"; '3'], ["BI57d"; '1'], ["BI57d"; '2'], ["BI57d"; '3'], ["BI57f"; '1'], ["BI57f"; '2'], ["BI57f"; '3'], ["BI60f"; '1'], ["BI60f"; '2'], ["BI60f"; '3'], ["BI64d"; '1'], ["BI64d"; '2'], ["BI64d"; '3'], ["BI64f"; '1'], ["BI64f"; '2'], ["BI64f"; '3'], ["BI64f"; '4'], ["BI64f"; '5'], ["Br51f"; '1'], ["Br51f"; '2'], ["Br51f"; '3'], ["Br58f"; '1'], ["Br58f"; '2'], ["Br58f"; '3'], ["Br59d"; '1'], ["Br59d"; '2'], ["Br59d"; '3'], ["Br59f"; '1'], ["Br59f"; '2'], ["Br59f"; '3'], ["Br60f"; '1'], ["Br60f"; '2'], ["Br60f"; '3'], ["Br64d"; '1'], ["Br64d"; '2'], ["Br64d"; '3'], ["Br64f"; '1'], ["Br64f"; '2'], ["Br64f"; '3'], ["Br64f"; '4'], ["Br64f"; '5'], "B60f", "B60s", "B65f", "B65s", "BI56f", "BI57d", "BI57f", "BI57s", "BI60f", "BI64d", "BI64f", "Br51f", "Br58f", "Br59d", "Br59f", "Br60f", "Br64d", "Br64f", "Tx60f"]
Sharp GE list: ["CHEST", "CHST", "HD Lung", "HD LUNG", "LUNG"]
Sharp Toshiba list: ["FC52", "FC53", "FC53", "FC31", "FC50", "FC51", "FC51", "FC24", "FC30", "FC17", "FC14", "FC18", "FC19", "FC02", "FC08", "FC10", "FC11", "FC12", "FC13", "FC13-H", "FL01", "FL03", "FL04"]
- Smooth Siemens kernel list: ["B45s", "B45f", "Br49d", ["Br49d"; '3'], "Br49f", ["Br49d"; '2'], ["Br49d"; '1'], ["Br49f"; '3'], ["Br49f"; '1'], ["Br49f"; '2'], ["B44d"; '1'], ["B44d"; '2'], ["B44d"; '3'], ["B44f"; '1'], ["B44f"; '2'], ["B44f"; '3'], ["Bf37f"; '1'], ["Bf37f"; '2'], ["Bf37f"; '3'], ["Br32d"; '1'], ["Br32d"; '2'], ["Br32d"; '3'], ["Br32f"; '1'], ["Br32f"; '2'], ["Br32f"; '3'], ["Br36d"; '1'], ["Br36d"; '2'], ["Br36f"; '1'], ["Br36f"; '2'], ["Br36f"; '3'], ["Br36s"; '1'], ["Br36s"; '2'], ["Br36s"; '3'], ["Br38f"; '1'], ["Br38f"; '2'], ["Br38f"; '3'], ["Br40d"; '1'], ["Br40d"; '2'], ["Br40d"; '3'], ["Br40f"; '1'], ["Br40f"; '2'], ["Br40f"; '3'], ["Br44f"; '1'], ["Br44f"; '2'], ["Br44f"; '3'], ["Bv36d"; '3'], ["Bv40d"; '1'], ["Bv40d"; '2'], ["Bv40d"; '3'], ["Bv40f"; '1'], ["Bv40f"; '2'], ["Bv40f"; '3'], ["I26f"; '1'], ["I26f"; '2'], ["I26f"; '3'], ["I30f"; '1'], ["I30f"; '2'], ["I30f"; '3'], ["I30s"; '1'], ["I30s"; '2'], ["I30s"; '3'], ["I31f"; '1'], ["I31f"; '2'], ["I31f"; '3'], ["I31s"; '1'], ["I31s"; '2'], ["I31s"; '3'], ["I40f"; '1'], ["I40f"; '2'], ["I40f"; '3'], ["I41f"; '1'], ["I41f"; '2'], ["I41f"; '3'], ["I41s"; '1'], ["I41s"; '2'], ["I41s"; '3'], ["I41s"; '4'], ["I44f"; '1'], ["I44f"; '2'], ["I44f"; '3'], "B08f", "B08s", "B10f", "B10s", "B19f", "B19s", "B20f", "B20s", "B25f", "B26f", "B29f", "B29s", "B30", "B30f", "B30s", "B31f", "B31s", "B35f", "B35s", "B36d", "B36f", "B36s", "B39f", "B39s", "B40f", "B40f", "B40s", "B41f", "B44d", "B44f", "B46f", "B46s", "Bf37f", "Br32d", "Br32f", "Br36d", "Br36f", "Br36s", "Br38f", "Br40d", "Br40f", "Br44f", "Bv40d", "Bv40f", "I26f", "I30f", "I30s", "I31f", "I31s", "I40f", "I41f", "I41s", "I44f", "T20f", "T20s", "Tr20f"]
Smooth GE kernel list: ["STANDARD", "Detail2", "Detail", "Veo", "EXPERIMENTAL", "SOFT"]
Smooth Toshiba kernel list: ["FC01"]

Model

The number of parameters for each DNN submodule is reported in Fig.8 of the article. For reproducibility¹³, the main hyperparameters of all DNN submodules are listed in TableS1.

Table S1. Models hyperparameters: List of the main hyperparameters for each DNN sub-modules. LR stands for Learning Rate, SGD stands for pytorch's mini-batch Stochastic Gradient Descent, MNO stands for Momentum and Nesterov Optimization, *Mini-batch size.

Module Hyperparameters	LR	LR schedule	Optimizer	Momentum	Batch size	Epochs
Lung segmentation	1e-4	None	Adam	Defaults	2	100
Nodule detection (CADE)	1e-2	None	SGD	MNO	4*	50
Nodule segmentation (CADE)	1e-3	None	Adam	Defaults	32	100
3D CNN (CADx)	1e-4	None	Adam	Defaults	3	200
2D CNN (CADx)	1e-4	None	Adam	Defaults	80	300
Full-CT scan (CADx)	1e-5	None	AdamW	Defaults	8	300

Lung segmentation module

Our approach leverages chest CT scans which are preprocessed using a unified pipeline to ensure data uniformity. Volumes are resampled to $128 \times 128 \times 128$ voxels using trilinear interpolation, and Hounsfield Units (HU) intensity values are clipped to $[-1024, 150]$, following conventional lung windowing protocols. These values are then normalized to the $[0, 1]$ range. Output masks are subsequently resampled to the original dimensions through trilinear interpolation. Our model is an adaptation of the 3D SegResNet architecture¹⁴, originally designed for 3D brain tumor segmentation. This relatively lightweight model was chosen to accommodate a manually re-annotated subset of the LUNA16 public dataset¹², which provides chest CT scans annotated with lung masks. Subset 0 is used for training and validation, while subset 1 is reserved for testing. The architecture employs residual units in an encoder-decoder design, mitigating gradient instability during training. Empirical optimization led to the selection of a 13-dimensional initial feature space. Training hyperparameters are listed in Table S1. We use a SoftDice loss function, derived from the Dice similarity coefficient, to capture segmentation accuracy by comparing probabilistic predictions to ground-truth annotations. Small, isolated misclassified regions are removed by retaining only the largest connected component for each predicted class. This refinement proves particularly beneficial when processing low-quality input scans, yielding substantial gains in segmentation accuracy.

Nodule detection module (CADe)

The CAdE architecture builds upon nnDetection¹⁵ that implement a RetinaNet architecture¹⁶ and leverages voxel-level segmentation ability. This one-stage, anchor-based detector, is favoured for its computational efficiency. At its core, the RetinaNet model combines ResNet for deep feature extraction and feature pyramid networks (FPNs) to construct a multi-scale feature level. The layers were adapted for 3-dimensional input volumes. The network has two main output branches: detection and segmentation. In turn, the detection branch includes two smaller branches: a classifier (predicting object class inside a bounding box); and a regressor (predicting bounding box offsets for each anchor). These sub-branches are trained using Binary Cross-Entropy loss and generalized IoU loss, respectively. The segmentation branch predicts the voxel-wise volume mask and is trained with Dice and Cross-Entropy loss. Adaptive Training Sample Selection is employed to match anchors with bounding boxes, while hard negative mining balances positive and negative anchors in each mini-batch, maintaining a 1/3 positive and 2/3 negative anchor ratio. Non-Max Suppression (NMS) filters overlapping predictions for the same object, prioritizing predictions at the center of a patch. Due to GPU memory limitations, cropped patches of CT scans are used as input for each training iteration. During inference, an overlapping sliding window is applied across the whole volume, with NMS handling overlapping predictions. Weighted Box Clustering is employed to consolidate predictions from multiple models or test-time augmentations. Training hyperparameters are detailed in Table S1. To enhance the robustness of the model and improve generalization across input variability, the training applies data augmentation techniques, such as elastic deformation, scaling, rotation, mirroring along axes, and adjustments in brightness.

Patch extraction and filter in lung module (CADe)

The full volume was resampled to a predefined spacing of $0.625 \times 0.625 \times 0.625$. Detected candidates were recomputed to match the new spacing, and candidates without any IoU overlap with the lung mask were filtered out. The conservative threshold of 0 for IoU was chosen to minimize the risk of missing nodules located at the lung boundaries (e.g., pleura). A patch of size $64 \times 64 \times 64$ centered on each candidate was then extracted for segmentation and characterization.

Nodule segmentation module (CADe)

The model is trained on CT scan crops of size $64 \times 64 \times 64$ with a fixed spacing of $0.625 \times 0.625 \times 0.625$ around the centers of the bounding boxes of the nodule detection. Since lesion segmentation is done on a 3D patch, some lesions may extend beyond the patch size. Only true positive detected boxes are chosen to train the model, which means each training sample contain a corresponding mask. Based on this training set, we trained ten separate models, deploying a 10-fold cross-validation strategy to improve the generalizability of the model. The same method is applied for the test set, which means only true detected boxes are used for evaluation. This network was based on a 3D-U-Net¹⁷ architecture, which is able to process the entire 3D patch as input, and output a 3D mask of matching dimensions. We normalized HU intensity scale to the $[0, 1]$ range, then the patch is rotated, and flipped. The addition of several noises (e.g., Gaussian and Gibbs noise) promotes robustness. Dice loss was used to optimize the model's objective. All training hyperparameters are provided in Table S1. The resulting masks are subsequently used for radiomic and 3D-morphomic computation, and for volume or diameter measurements.

Diameter measurement module (CADe)

This module computes the diameter for each segmented mask obtained from the detection stage. The major axis is defined as the longest arc on the convex hull of the 2D mask taken on the slice exhibiting the largest area. The minor axis is defined as the longest segment included in the convex hull of the mask orthogonal to the long axis on the same slice. In cases where the 3D segmentation mask comprises multiple connected components, only the largest component by volume is considered. The resulting diameter is used as an input filter for all subsequent (characterization) processes to exclude all detections that fall

outside the [4, 40]mm range. This filtering step not only considerably reduces computation time and charge of the product, but also decreases FPs, thereby improving both characterization performance and training time.

Nodule characterization module (CADx)

CT scans are first resampled to a uniform voxel dimension $0.625 \times 0.625 \times 0.625$ (mm) using bicubic spline for the CT images and Nearest Neighbors for masks. A basic HU normalization is applied, consisting of a percentile crop (below 1% and above 99%) followed by a normalization to ensure that numerical values in the input data fall within the [0, 1] interval. Each of the 15 models is then trained on 15 different training subsets of Train3, each comprising all malignant findings and one of the 15 non-overlapping, randomly split and equally sized partitions of the benign findings. This procedure increased classifier diversity and considerably reduced training times, while improving the overall accuracy, by counterbalancing the very large imbalance of the training set. Each of the training sets was then further split into training (80%) and tuning (20%) sets.

3D Densenet module: The 3D model takes as input a patch centred on a nodule and outputs a malignancy score for that nodule during both training and inference. For each detection, a 3D patch centered on the corresponding bounding box of dimensions $64 \times 64 \times 64$ is extracted. The model architecture is a shallow 3D DenseNet39 (with 39 layers) inspired by the work of Yuan *et al.*¹⁸. It consists of three Dense-S Blocks, each containing four Dense Conv Blocks, with a growth rate of 32 and dropout rate ($p=1/2$). The main difference between a 2D DenseNet¹⁹ and its 3D counterpart lies in the replacement of 2D convolution and 2D pooling layers with their 3D equivalents. Basic data augmentations is applied at random during training, including flips and 'light' affine transforms (with probability = 0.5, translate range = (5, 5, 5) voxels, rotate range = $(2\pi, 2\pi, 2\pi)$, using reflection padding and excluding scaling transformations, as malignancy is not considered scale-invariant). As with 2D models, 3D models are trained using a focal loss function. Training hyperparameters are provided in Table S1.

2D CNN module: For each detected nodule, 64×64 patches were extracted from all slices containing target nodules (in axial view). The model consists of three convolutional layers, one max pooling layer and one mean pooling layer, producing a learned feature vector as output. A learned bilinear transform is applied between the feature vector of a given patch and the average feature vector of all patches in the lesion, in order to produce a contextualized feature vector. A final dense layer outputs a two-dimensional vector, where the second coordinate represents the malignancy likelihood of the input nodule. To break linearity, the network also includes ReLU layers, and to enforce and enhance generalizability, it uses dropout layers and batch normalization applied prior to each pooling layer. This model is trained using a focal loss function to address the high imbalance in the dataset. Nodule-wise inference is made by computing a malignancy score for each patch within the nodule and then taking the maximum score among them as the final output, yielding a single malignancy score for each nodule of a given patient. Training hyperparameters are given in TableS1.

XGBoost classifiers module: Following^{20,21}, 111 radiomic features are extracted (see [pyradiomics](#) for detailed description), including:

- 17 Morphological features (margins, spiculation, sphericity): 3D shape-based (one additional redundant feature of "volume" computed by counting the voxels volumes), including multiple different diameters calculus, and nodule surface area, surface to volume ratio, sphericity among other interesting features.
- 94 texture/luminance features (for attenuation, solid, part solid, ground glass, calcification): 19 first order Statistics of HU, 24 Gray Level Cooccurrence Matrix (GLCM), 16 Gray Level Run Length Matrix (GLRLM), 16 Gray Level Size Zone Matrix (GLSZM), 5 Neighboring Gray Tone Difference Matrix (NGTDM), 14 Gray Level Dependence Matrix (GLDM).

The Pyradiomics documentation details each radiomic feature, all computed in or from the mask of segmentation in the $(64 \times 64 \times 64)$ voxels 3D CT patches of the lesions (within the masks for the textural features). Additionally, our model extracts 11 3D-morphomics as described in²² that proved to be very efficient for malignancy classification. The resulting 122 3D-morphomics and radiomics are then used to train and test a gradient boosted decision tree classifier (XGBoost²³, [DMLC XGBoost](#)). Training: XGBoost hyper-parameters are tuned on the NLST train-tuning set with a 80-20% split using Bayesian descent on a binary logistic loss (hyperopt library). The explored space and corresponding range is: 'eta': loguniform $[1E-5, 0.1]$; 'max depth': [1, 2, 3, 4, 5]; 'gamma': loguniform $[1E-2, 7]$; 'alpha': loguniform $[1E-8, 1E-2]$; 'lambda': loguniform [1, 4]; 'colsample bytree': uniform [0.5, 0.7]; 'colsample bylevel': uniform [0.5, 1]; 'min child weight': qloguniform [1, 30, 1]; 'subsample': qloguniform [0.1, 0.7, 0.05]; 'scale pos weight': uniform [1, imbalance]; 'num boost round': quniform [100, 1000, 30]. The number of estimators is set to 180 with initialization at default XGboost hyperparameters. The final model

ultimately used for inference is then re-trained on the full train set with optimal parameters with a learning rate of 0.01 and 1,000 estimators.

Nodule predictions ensembling and calibration module (CADx)

As an output of the malignancy classifiers, we have for each lesion, 15 3D model predictions, 15 2D model predictions and 15 XGBoost predictions. Average prediction per model: The first step of the ensembling computes the mean of each class of models to obtain three mean predictions. Optimal convex domain ensembling: The second step consist in a stacking ensemble of the three predictions²⁴ (convex domain affine optimization) using [Hyperopt](#) library trained on the training set, that derive the three optimal coefficients for those models reflecting each model's predictive power: mean 3D-CNN=92.38%; mean 2D-CNN=1.84%; mean XGBoost=5.78%. The lesion level intermediate prediction of malignancy is then computed as the weighted sum of those three mean predictions.

Calibration: The final step involved calibrating the lesion level prediction using Netcal logistic calibration optimization model by Küppers *et al.*¹ ([GitHub repository](#)). Logistic calibration temperature scaling provided superior calibration performances. The calibrated ensembled prediction demonstrated significantly higher AUC than any of the three mean models alone.

0.0.3 Full-CT scan characterization module (CADx)

This model is designed to capture contextual features at both patient level and at lesion level, e.g., emphysema, pleural wall thickness, lesion multiplicity, inflammations, mediastinal and hilar lesions (although the latter are beyond the scope of the present study). It aims at improving diagnostics accuracy by reducing patient and lesion level errors. Data preprocessing: The model operates on 3D chest CT scans. Each CT scan is first cropped around the center of the bounding box determined by its lung segmentation. The scans are then resampled to a uniform volume size of $256 \times 256 \times 200$ voxels by nearest neighbor down-sampling, preserving the original spacing along the z-axis by allowing for cropping or padding. The HU intensity values are cropped below -1020 and above 300, consistent with a standard lung windowing, then rescaled to the [0, 1] interval, excluding the lowest and highest 1% of intensity values. Model : The model we use in this version of the product is a variant of Sybil, developed by Mikhael *et al.*²⁵. Initially designed to predict cumulative cancer risk over one to six years, we adapted the output of the model to fit a binary criterion (patient-level “cancer” vs. “non-cancer”) at T_{-1} . The model relies solely on 3D cropped CT scans as input for inference and whereas during training it additionally uses 3D masks of detected nodules. This flexibility allows for the model to enforce attention to lesion localizations during training, thereby enhancing performance²⁵, while limiting inference costs. Optimization: Training relies on a dual-term loss, combining a Focal loss term which controls the proximity of the predicted score to the binary GT, and an attention loss term ensuring localization of detected nodules. Unlike the original implementation, we remove the loss term predicting the side of the lung where the potential cancer has or will grow. The network is initialized with a Resnet-18 encoder pretrained on the Kinetics-400 dataset. Due to the model's complexity relative to the number of data instances, affine augmentations are performed during training. Training hyperparameters are provided in Table S1.

Nodule prediction update and calibration module (CADx)

The ensembling of scan-level prediction combines the output from five different elementary scan predictions : four full volume models prediction and one lesion based prediction (the maximum of the predictions over all detection per scan). Rescaling: As a preliminary preprocessing step, all the five predictions are linearly rescaled to the [0, 1] interval such that the minimum prediction becomes 0 and the maximum 1. The parameters governing these linear transformations are obtained on the training set, and stored in our model as learned parameters. Optimal convex domain ensembling: mirroring the nodule level, the second step employs a stacking ensemble using the same method, to derive the five optimal coefficients for those models representing their respective predictive power:

- Nodule (max) model = 24.89%
- Full-CT scan model 3 = 23.10%
- Full-CT scan model 1 = 21.41%
- Full-CT scan model 2 = 16.36%
- Full-CT scan model 4 = 14.24%

The intermediate scan level prediction is calculated as the weighted sum of those five predictions. Calibration: As for nodule level, the final step consisted in calibrating the ensembled patient prediction using exactly the same methodology. This ensembled and calibrated scan prediction provided significantly higher AUC than any of the five elementary model alone. Nodule prediction update: This last step refines nodule-level malignancy prediction by adjusting them based on the final patient-level prediction. Such correction accounts for broader contextual cancer cues (emphysema...) that are beyond the scope

of lesion level prediction which infer malignancy from only a very small and localized region of the image. The correction is computed by the algorithm as the delta between between the final patient prediction and the maximum lesion-level prediction for a patient, expressed as a percentage of the maximum lesion:

$$\Delta \text{ correction} = \frac{\text{Final scan prediction} - \text{Max nodule prediction}}{\text{Max nodule prediction}}$$

The final updated lesion prediction was then obtained as:

$$\text{Final nodule prediction} = \text{Intermediate nodule prediction} + \Delta \text{ correction} \times \text{Intermediate nodule prediction}.$$

This update-correction slightly improves lesion-level performance compared to nodule-level prediction alone.

Multiple detections reduction module and diameter measurement (CADx)

Owing to the detection model output, identical finding may occasionally be detected multiple times across multiple patches. The deduplication eliminates such redundant detections through a two steps process:

- *Clean finding segmentations*: (1) remove connected components with diameter < 4 mm; (2) keep only the connected component closest to the patch center, removing others. Patches were centered on the detected bounding box center. Ensuring connected components were centered in the patch guaranteed recovery of all detected findings.
- *Merge cleaned findings*: (1) group findings whose masks intersected; (2) merge finding groups into their union; and (3) assign the highest malignancy score among merged findings to the resulting finding.

This procedure guarantees that segmentations of reported findings do not overlap with one another, and that the merged finding receives the maximal malignancy prediction. For pairing purposes, the detection bounding box of the merged finding is set to the bounding box of the merged segmentation. The segmentation bounding box is preferable for pairing since the segmentation may contain multiple findings detected independently with smaller detection boxes. The updated diameter extraction method operates on deduplicated findings: the diameter extraction method used following the nodule segmentation module remains unchanged. Then the SAX is extracted directly from the segmentation, guaranteeing it lies on the segmentation border. Whether the segmentation itself or its convex hull is used does not affect the extraction of the long axis (LAX), which remains unchanged. This new constraint—that the extremities of the SAX must lie on the segmentation contour—, produces the following edge case: the SAX may be too short to intersect with the LAX. In such instances, the SAX segment is extended until it intersects with the LAX. Additionally, differing from the previous method for diameter extraction and for aesthetic considerations: LAX and SAX are extracted from the largest 2D component on the slice with the greatest area.

Supplementary References

References

1. Fabian Kupperts, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate Confidence Calibration for Object Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1322–1330, June 2020. Conference Name: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) ISBN: 9781728193601 Place: Seattle, WA, USA Publisher: IEEE.
2. S. Gregory Jennings, Helen T. Winer-Muram, Mark Tann, Jun Ying, and Ian Dowdeswell. Distribution of Stage I Lung Cancer Growth Rates Determined with Serial Volumetric CT Measurements. *Radiology*, 241(2):554–563, November 2006.
3. Dong Ming Xu, Hester Gietema, Harry de Koning, René Vernhout, Kristiaan Nackaerts, Mathias Prokop, Carla Weenink, Jan-Willem Lammers, Harry Groen, Matthijs Oudkerk, and Rob van Klaveren. Nodule management protocol of the NELSON randomised lung cancer screening trial. *Lung Cancer (Amsterdam, Netherlands)*, 54(2):177–184, November 2006.
4. N. Becker, E. Motsch, M. L. Gross, A. Eigentopf, C. P. Heussel, H. Dienemann, P. A. Schnabel, M. Eichinger, D. E. Optazait, M. Puderbach, M. Wielpütz, H. U. Kauczor, J. Tremper, and S. Delorme. Randomized Study on Early Detection of Lung Cancer with MSCT in Germany: Results of the First 3 Years of Follow-up After Randomization. *Journal of Thoracic Oncology*, 10(6):890–896, June 2015.
5. Nancy A. Obuchowski, Brandon D. Gallas, and Stephen L. Hillis. Multi-Reader ROC studies with Split-Plot Designs: A Comparison of Statistical Methods. *Academic radiology*, 19(12):1508–1517, December 2012.

6. Zilong He, Yue Li, Weixiong Zeng, Weimin Xu, Jialing Liu, Xiangyuan Ma, Jun Wei, Hui Zeng, Zeyuan Xu, Sina Wang, Chanjuan Wen, Jiefang Wu, Chenya Feng, Mengwei Ma, Genggeng Qin, Yao Lu, and Weiguo Chen. Can a Computer-Aided Mass Diagnosis Model Based on Perceptive Features Learned From Quantitative Mammography Radiology Reports Improve Junior Radiologists' Diagnosis Performance? An Observer Study. *Frontiers in Oncology*, 11:773389, December 2021.
7. Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.-Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Castele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, 38(2):915–931, February 2011.
8. The TCIA team. Annotation Protocol of the LIDC-IDRI Lung screening database, 2020.
9. Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, June 2019.
10. Sylvie Leroy, Jonathan Benzaquen, Andrea Mazzetta, Sylvain Marchand-Adam, Bernard Padovani, Dominique Israel-Biet, Christophe Pison, Pascal Chanez, Jacques Cadranet, Julien Mazières, Vincent Jounieaux, Charlotte Cohen, Véronique Hofman, Marius Ilie, Paul Hofman, and Charles Hugo Marquette. Circulating tumour cells as a potential screening tool for lung cancer (the AIR study): protocol of a prospective multicentre cohort study in France. *BMJ Open*, 7(12):e018884, December 2017.
11. Alexander A. Bankier, Heber MacMahon, Thomas Colby, Pierre Alain Gevenois, Jin Mo Goo, Ann N.C. Leung, David A. Lynch, Cornelia M. Schaefer-Prokop, Noriyuki Tomiyama, William D. Travis, Johnny A. Verschakelen, Charles S. White, and David P. Naidich. Fleischner Society: Glossary of Terms for Thoracic Imaging. *Radiology*, 310(2):e232558, February 2024. Publisher: Radiological Society of North America.
12. Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S. N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M. J. de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T. M. C. Manders, Alexander Sónora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M. Schaefer-Prokop, Ernst T. Scholten, Luuk Scholten, Miranda M. Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C. A. Zuidhof, Bram van Ginneken, and Colin Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, December 2017.
13. Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J. W. L. Aerts. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, October 2020. Publisher: Nature Publishing Group.
14. Andriy Myronenko. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijff, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 311–320, Cham, 2019. Springer International Publishing.
15. Michael Baumgartner, Paul F. Jäger, Fabian Isensee, and Klaus H. Maier-Hein. nnDetection: A Self-configuring Method for Medical Object Detection. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 530–539, Cham, 2021. Springer International Publishing.
16. Paul F. Jaeger, Simon A. A. Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H. Maier-Hein. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection, November 2018. arXiv:1811.08661 [cs].

17. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
18. Changan Yuan, Yong Wu, Xiao Qin, Shaojie Qiao, Yonghua Pan, Ping Huang, Dunhu Liu, and Nan Han. An effective image classification method for shallow densely connected convolution networks through squeezing and splitting techniques. *Applied Intelligence*, 49, October 2019.
19. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, July 2017. ISSN: 1063-6919.
20. Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J. W. L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, October 2017.
21. Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5, June 2014. 00015.
22. Elias Munoz, Pierre Baudot, Van-Khoa Le, Charles Voyton, Benjamin Renoust, Danny Francis, Vladimir Groza, Jean-Christophe Brisset, Ezequiel Geremia, Antoine Iannessi, Yan Liu, and Benoit Huet. 3D-Morphomics, Morphological Features on CT Scans for Lung Nodule Malignancy Diagnosis. In Sharib Ali, Fons van der Sommen, Bartłomiej Władysław Papież, Maureen van Eijnatten, Yueming Jin, and Iris Kolenbrander, editors, *MICCAI Cancer Prevention Through Early Detection*, Lecture Notes in Computer Science, pages 3–13, Singapore, 2022. Springer Nature Switzerland.
23. Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery.
24. Ammar Mohammed and Rania Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774, February 2023.
25. Peter G. Mikhael, Jeremy Wohlgend, Adam Yala, Ludvig Karstens, Justin Xiang, Angelo K. Takigami, Patrick P. Bourgooin, PuiYee Chan, Sofiane Mrah, Wael Amayri, Yu-Hsiang Juan, Cheng-Ta Yang, Yung-Liang Wan, Gigin Lin, Lecia V. Sequist, Florian J. Fintelmann, and Regina Barzilay. Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 41(12):2191–2200, April 2023.