Computerized Medical Imaging and Graphics 116 (2024) 102399

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Enhancing cancer prediction in challenging screen-detected incident lung nodules using time-series deep learning

Shahab Aslani [a,b], Pavan Alluri [c], Eyjolfur Gudmundsson [a], Edward Chandy [a], John McCabe [a], Anand Devaraj [d,e], Carolyn Horst [b,f], Sam M. Janes [b,f], Rahul Chakkara [c], Daniel C. Alexander [a,g], SUMMIT consortium, Arjun Nair [h], Joseph Jacob [a,b,*]

[a] Centre for Medical Image Computing, University College London, London, UK
[b] Department of Respiratory Medicine, University College London, London, UK
[c] MANAS AI, London, UK
[d] Department of Radiology, Royal Brompton and Harefield NHS Foundation Trust,, London, UK
[e] National Heart and Lung Institute, Imperial College London, London, UK
[f] Lungs for Living Research Centre, University College London, London, UK
[g] Department of Computer Science, University College London,, London, UK
[h] University College London Hospitals NHS Foundation Trust, London, UK

ARTICLE INFO

Dataset link: https://cdas.cancer.gov/learn/nlst/images/

Keywords:
Computer-aided diagnosis
Computed tomography
Lung cancer
Longitudinal study
Deep learning

ABSTRACT

Lung cancer screening (LCS) using annual computed tomography (CT) scanning significantly reduces mortality by detecting cancerous lung nodules at an earlier stage. Deep learning algorithms can improve nodule malignancy risk stratification. However, they have typically been used to analyse single time point CT data when detecting malignant nodules on either baseline or incident CT LCS rounds. Deep learning algorithms have the greatest value in two aspects. These approaches have great potential in assessing nodule change across time-series CT scans where subtle changes may be challenging to identify using the human eye alone. Moreover, they could be targeted to detect nodules developing on incident screening rounds, where cancers are generally smaller and more challenging to detect confidently.

Here, we show the performance of our Deep learning-based Computer-Aided Diagnosis model integrating Nodule and Lung imaging data with clinical Metadata Longitudinally (DeepCAD-NLM-L) for malignancy prediction. DeepCAD-NLM-L showed improved performance (AUC = 88%) against models utilizing single time-point data alone. DeepCAD-NLM-L also demonstrated comparable and complementary performance to radiologists when interpreting the most challenging nodules typically found in LCS programs. It also demonstrated similar performance to radiologists when assessed on out-of-distribution imaging dataset. The results emphasize the advantages of using time-series and multimodal analyses when interpreting malignancy risk in LCS.

1. Introduction

Lung cancer is the most common cause of cancer death in the world (WHO, 2022). Early detection of lung cancer using low-dose CT scans in lung cancer screening (LCS) studies allows timely intervention and treatment thereby reducing lung cancer mortality rates. This has resulted in defined lung cancer screening guidelines (Aberle et al., 2011; NLST, 2011; Black et al., 2014; Koning et al., 2020). The US National Lung Screening Trial (NLST) (NLST, 2011) and the Dutch-Belgian NELSON Trial (Koning et al., 2020), demonstrated overall reductions in lung cancer mortality of at least 20% (NLST, 2011; Koning et al., 2020).

The focus of LCS studies is the detection of pulmonary nodules and the assessment of nodule growth which may indicate the presence of early lung cancer. However, the majority of screen-detected nodules are either benign or have no bearing on a patient's prognosis. The probability of a lung nodule being malignant is currently determined, in combination with an individual's risk factors, in two ways: (1) at baseline using an assessment of size and associated characteristics (location, density, morphology); and (2) evolution of nodule characteristics – chiefly growth rate – on interval scans. Lung cancer screening studies function optimally when lung nodules are detected with high sensitivity. Subsequently, they are interpreted with high specificity

* Corresponding author.
E-mail address: j.jacob@ucl.ac.uk (J. Jacob).

https://doi.org/10.1016/j.compmedimag.2024.102399
Received 30 October 2023; Received in revised form 18 March 2024; Accepted 8 May 2024
Available online 20 May 2024

to achieve excellent discriminatory performance for predicting lung malignancy. Together, this allows the detection of high-risk nodules at an early curable stage.

Lung cancer screening generates huge volumes of CT imaging that require evaluation by radiologists. Yet in countries such as the UK, there remain national shortages of radiologists to evaluate screening CTs (RCR, 2021). The field of lung cancer prediction on CT scans using machine learning and deep learning algorithms has matured following the wide availability of large screening datasets for analysis. Algorithms can now be expertly trained with a breadth of examples of nodule types beyond that which an average radiologist would encounter over an entire career. There is hope that utilizing computer algorithms as objective diagnostic aids for radiologists when interpreting LCS CT imaging may result in faster and better reads of challenging screening cases.

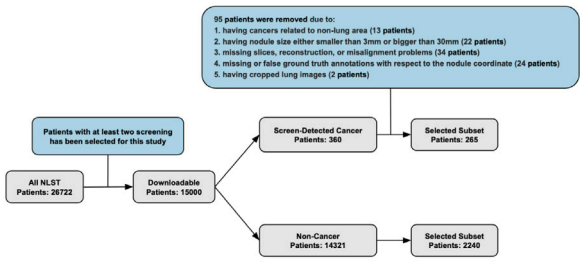
Study Rationale. In typical radiology workflows, when an abnormality is detected on a CT by a radiologist, prior imaging is sought to better understand how the lesion has changed in appearance over time. Evaluating time-series data to understand disease behaviour is a central tenet of radiology. It would, therefore, be expected that machines may also improve their performance in predicting malignancy by interrogating time-series data. Yet most previous studies have focused on analysing single time-point CT datasets acquired as part of lung cancer screening programs (Liao et al., 2019).

Studies examining malignancy risk prediction in LCS have typically indiscriminately assessed all nodules contained within screening cohorts. However, nodules subtypes can vary across the breadth of a screening study. Cancers seen at the first CT screening round are typically larger and easier to identify by human readers, with or without the aid of computers. However, cancers developing on incident screening rounds (pre-existing nodules or de-novo) are more challenging for a radiologist to distinguish from benign nodules. And yet the majority of cancers identified in lung cancer screening programs will be nodules developing across incident screening rounds. Support from computer tools to improve the classification of these challenging nodules would therefore be extremely beneficial as an assistive read for a radiologist in a screening setting. It is therefore necessary to define the performance of computer-based nodule classification system when applied to difficult cases. Lastly, no studies have integrated readily available clinical information with time-series image data when designing an automated system for lung cancer prediction. Yet this ignores potentially valuable information that is routinely collected in screening programs.

Therefore, in this study, we propose our Deep learning-based Computer-Aided Diagnosis model integrating Nodule and Lung imaging data with clinical Metadata Longitudinally (DeepCAD-NLM-L). The proposed model was trained on the NLST dataset to improve the prediction of lung cancer likelihood by utilizing time-series CT data (two or three longitudinal CTs per patient). Our model aggregates both lung-level and nodule-level information, thereby leveraging the advantages of both imaging data sources. We also combine clinical metadata with imaging data to aid lung cancer prediction.

We compared the utility of time-series analyses of CTs using DeepCAD-NLM-L against nodule management algorithms utilized in the SUMMIT LCS study. The contemporary SUMMIT study CTs are acquired at higher spatial resolution and lower dose than the CTs contained within NLST. Our analysis, therefore, tested the performance of DeepCAD-NLM-L on data out-of-distribution to the data used to train DeepCAD-NLM-L. As part of this analysis, we also examined whether the sensitivity and specificity of human readers versus computer algorithms might provide complementary interpretations of nodule malignancy risk in a LCS setting. Finally, we tested how well a combined time-series model could classify the most challenging lung nodules developing on incident LCS rounds. Specifically, we evaluated classification performance on "small" to "intermediate" sized nodules (5–10 mm).

a



b

Variable	Description
Education	Level of the education completed
Age	Age
Race	Race
Gender	Male or female
Smoke 1	Pack year tertile
Smoke 2	Smoking status
Smoke 3	Total years of smoking
Smoke 4	Average number of cigarettes per day
Ethnicity	Is the participant of Hispanic/Latino heritage?

Fig. 1. NLST data selection strategy for our study. (a) Data from 15,000 NLST patients were available, with up to 3 time point CT scans per patient. All of the available screen-detected incident lung cancers (n = 360) in NLST were considered. 95 patients were excluded for reasons outlined above. 2240 control patients without cancer were randomly selected for analysis. (b) The lower table shows the clinical metadata analysed by our model.

Automated Lung Cancer Detection. Automated diagnosis of lung cancer using deep learning methods typically addresses either lung-level or nodule-level predictions. For lung-level prediction methods, the entire CT scan is used as an input to the model (Wang et al., 2019; Causey et al., 2019; Jiang et al., 2020; Gao et al., 2019). Nodule-level lung cancer prediction methods consist of two sequential stages: computer-aided detection (CADe), where the nodule is identified, followed by "diagnosis" (CADd), where a malignancy probability is assigned to the identified nodule and then a cancer/non-cancer label is assigned to the case (Liao et al., 2019; Li and Fan, 2020; Khosravan and Bagci, 2018; Ding et al., 2017; Kuan et al., 2017; Liao et al., 2019; Trajanovski et al., 2018; Ozdemir et al., 2019; Ardila et al., 2019; Zhu et al., 2018).

Nodule-level methods have been shown to be more accurate than lung-level models when predicting malignancy risk in candidate nodules. However, classifying malignancy risk in a nodule remains reliant on the ability of the model to first detect the nodule which in turn, relies on good quality training data. It is also possible that valuable information may be contained within tissue immediately surrounding the nodule itself. Evaluating features across the whole-lung could also provide improved contextual information about a nodule and potentially improve model discrimination of features of malignancy on an individual CT. Accordingly, our model was designed to consider nodule-level and whole lung-level features.

Related Work. Several studies have utilized deep learning-based models for lung cancer prediction. For models that just consider whole lung-level predictions, a 3D CNN network has been proposed to predict lung cancer in Jiang et al. (2020). Regarding nodule-level prediction methods, Liao et al. (2019) proposed an approach that included CADe and CADd systems. They used a 3D Faster R-CNN (Ren et al., 2016) as the nodule detection network. Later, a set of shallow 3D deep CNNs were used to extract features from candidate nodules and predict the malignancy score using a Leaky Noisy-OR approach (Pearl, 1988). Recently, Ardila et al. (2019) proposed an approach for lung cancer prediction combining lung-level and nodule-level predictions. The CADe system in this method is a 3D Inception CNN (Szegedy et al., 2015) combined with a Region Proposal Network (RPN) (Ren et al., 2016) to identify a set of candidate nodules on a participant's current and (and if present) prior CTs. The same 3D Inception CNN was used as a CADd system to calculate the malignancy score. Alternative features from imaging at the lung-level were incorporated and boosted the performance of the CADd system. Similar to Ardila et al. (2019) and

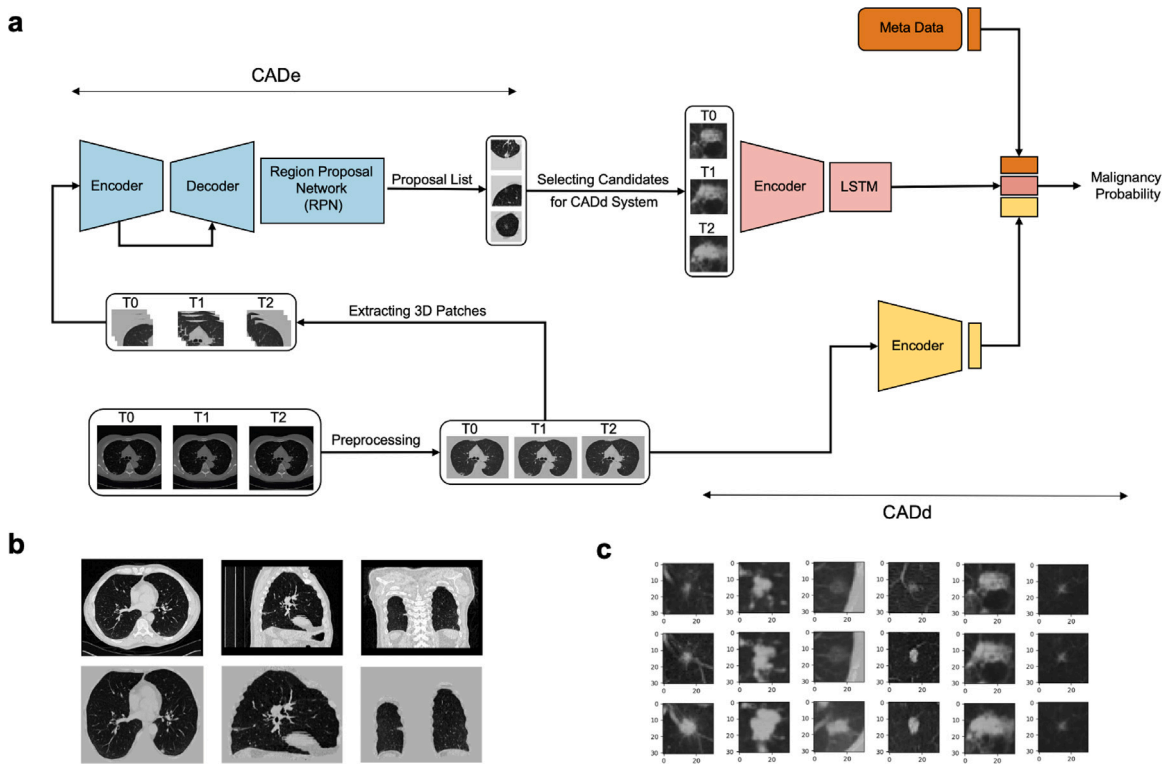


Fig. 2. General overview of the our DeepCAD-NLM-L method. (a) This model includes computer aided detection (CADe) and computer aided diagnosis (CADd) systems. In this architecture, the CADe system (blue) includes an Encoder–Decoder followed by a region proposal network. Input to the CADe system is segmented 3D patches from preprocessed longitudinal CT scans. The output is a set of nodule proposals. A single nodule that has the highest confidence score is selected as the candidate nodule. Three 3D CT patches around the candidate nodule are then extracted from all available CT time-points. A CADd system combines the extracted nodule-level (pink Encoder+LSTM) and lung-level (yellow Encoder) longitudinal features with metadata information to predict malignancy. For detailed information about the structure of CADe and CADd models, please refer to the Supplementary Appendix. (b) Image preprocessing. The top row of images show the original computed tomography scan in axial, sagittal, and coronal planes. Images in the second row show the preprocessed version of the corresponding image in the top row, with the chest wall having been cropped out. (c) The output of the patch extraction process for candidate nodules. Each column corresponds to a different time point for a single case. The first, second, and third rows are related to the T0, T1, and T2 time-points, respectively.

Venkadesh et al. (2023) proposed a nodule-level CADd approach using prior and current CTs. The architecture includes a three-dimensional nnU-Net (Isensee et al., 2021) performing volumetric segmentation of the nodules on both prior and current scans. The segmented areas are then stacked and fed into a nodule classifier for malignancy prediction. In Gao et al. (2019) a Long Short-Term Memory Model (LSTM) capable of learning both long-term and short-term dependencies between features was used. A Distanced LSTM allowed evaluation across irregularly sampled intervals though nodule-level features were not studied. It is important to mention that models proposed in Ardila et al. (2019), Venkadesh et al. (2023) and Gao et al. (2019) included patients where cancer was confirmed after the first screening round. Such cancerous nodules are likely to be larger and easier to identify than incident cancers that occur during incident screening rounds. Xu et al. (2019) developed a nodule-level prediction method in which the CADd system included a 2D CNN with a Recurrent Neural Network (RNN). This model used four different scans per patient: a baseline CT and CTs one, three and six months later. However, the proposed approach was not fully automated as it did not contain a CADe system.

To date, no published fully automated method has combined lung-level and nodule-level features with clinical metadata demographics in a longitudinal manner for malignancy estimation.

2. Material and method

Datasets. Our DeepCAD-NLM-L model was trained on a historic lung cancer screening dataset: The National Lung Screening Trial (NLST, 2011; Aberle et al., 2011) and tested on unseen data from NLST and separately on the SUMMIT Lung Cancer Screening Study (Horst et al., 2020).

NLST was a large randomized multi-centre LCS study in the United States in which 26,722 participants underwent three annual screens (T0, T1, and T2) using low-dose CT scans. If cancer was diagnosed on the first CT screen subsequent screening CTs were not performed. Fig. 1.a displays the NLST cohort analysed in our study. An individual CT was considered cancer-positive if the result of a biopsy or surgical resection was positive during the screening study year. An individual CT was considered cancer-negative if the patient was cancer-free on the follow-up screen (NLST, 2011; Aberle et al., 2011; Ardila et al., 2019). 679 patients had biopsy-confirmed screen detected (n = 360) and non-screen detected (n = 319) cancers. All NLST CT scans were acquired with the use of multidetector scanners with a minimum of four channels. The acquisition variables were chosen to reduce exposure to an average effective dose of 1.5 mSv as previously described (https://www.nejm.org/doi/full/10.1056/nejmoa1102873). The North London-based SUMMIT Lung Cancer Screening Study uses a commercial CADd system to identify lung nodules. Following the automated read, a radiologist accepts or rejects the CAD outputs and performs a second read to ensure that no nodules are missed. The SUMMIT Study aims to assess the implementation of LDCT for lung cancer screening in a high-risk population and to validate a multi-cancer early detection blood test (ClinicalTrials.gov identifier NCT03934866). SUMMIT participants with prevalent cancers (cancers diagnosed on the first screening CT) were excluded from our analysis as we aimed to focus on nodules that were challenging to diagnose, and where computer assistance would be most valuable. SUMMIT participants with at least two CT images acquired between April 2019-April 2020 were analysed. The nodule management plan in the SUMMIT study required participants with a suspicious nodule on CT (but no lung cancer diagnosis), to be referred to a multi-disciplinary team for further

investigation or to undergo CTs at 3 and 12 months unless an interim diagnosis of lung cancer was received (Horst et al., 2020).

Data Preprocessing. All CT scans were converted to Hounsfield Units (HU) (the quantitative scale for describing radiodensity) as per in Li and Fan (2020). Images were binarized by thresholding at −600 HU. Using the 2D/3D connected components and measuring their distances to the centre of the image, we extracted lung-connected domains. Erosion and dilation morphological operations were applied to divide the lung mask into right and left lungs. The convex hull of each lung was computed and using a dilation operation masks were combined to create a more accurate binary lung mask. The original image was clipped within the range [−1200 to 600 HU] and normalized to [0, 1]. The ultimate mask was used to segment the lung. All voxel values outside the lung were assigned a density of 170 HU corresponding to normal tissue density. The image was cropped in all three dimensions to retain just the lung in the image. Finally for each patient, we rigidly registered the later time-point CTs (T2 and T1) to the first time-point CT (T0) using FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson and Smith, 2001; Jenkinson et al., 2002). The result of the preprocessing pipeline on a single NLST CT acquired with different views is shown in Fig. 2.b.

Model Architecture. Our novel fully automated lung cancer prediction model considers time-series data and includes a CADe system to detect suspicious nodules and a CADd system to estimate a malignancy score. We combined lung-level and metadata features with nodule-level features to increase the performance of the CADd system. The general architecture of our method can be seen in Fig. 2.a.

Following the idea in Li and Fan (2020), our CADe nodule detection system includes a 3D U-net-like (Ronneberger et al., 2015) Encoder–Decoder for feature extraction followed by a 3D Region Proposal Network (Liao et al., 2019) that enables the model to generate proposals directly. The backbone architecture of the Encoder is a customized ResNet18 (He et al., 2016) combined with the mirrored version of the same network with reduced blocks as the Decoder. As suggested in Li and Fan (2020), the Encoder–Decoder is combined with squeeze-and-excitation blocks (Hu et al., 2018) to generate richer features with more contextual information to detect candidate nodules. Fig. 2.a shows the overall architecture of our CADe system. A more detailed structure of our CADe model in shown in the Supplementary Appendix.

To measure the malignancy score, our CADd system aggregates lung-level, nodule-level, and clinical metadata features extracted from longitudinal data. As a first step, a selection criteria is applied to the outputs of the CADe system that chooses a candidate nodule coordinate and extracts 3D patches from all time-point CTs at the corresponding coordinates (Please refer to Section 3 for more details). Then a 3D Encoder followed by a long short-term memory (LSTM) layer is used to extract nodule-level features from longitudinal input data. In the second step, the preprocessed 3D scans from all three CT time-points are aggregated to create a single 4D input. A 3D Encoder extracts features from the 4D input corresponding to the lung-level time-series information. In the next step, nodule-level and lung-level features are combined with demographic metadata features to predict the malignancy score. Both the 3D Encoders used in the CADd system are customized ResNet10 (He et al., 2016). Fig. 2.a demonstrates the architecture of our CADd system, with more details provided in the Supplementary Appendix.

Training Implementation Details. If we consider a set of scans J, for each scan, we may have a set of nodules I. Each nodule has specific information regarding the volume coordinate (x_{ij}, y_{ij}, z_{ij}) and the diameter (r_{ij}). To train our CADe system, we needed the coordinate and size of the nodules on the scans. For the NLST dataset, (x_{ij}, r_{ij}) values representing the axial location and diameter of the nodules were provided. However, (y_{ij}, z_{ij}) values representing coronal and sagittal locations were missing. Therefore, to prepare the training datasets for our CADe system, we manually annotated 2000 single time-point scans in the NLST dataset. All selected scans have nodules size in the range of 5 mm–30 mm. In total, 3081 nodule locations were labelled (1–2 nodules per scan, on average) using ITK-SNAP software (Yushkevich et al., 2006). For training the CADe system, we divided the whole annotated dataset into training (70%), validation (15%) and test (15%) sets. Cases were specifically selected with respect to nodule size to ensure a wide distribution of nodule size in each dataset (please refer to the Supplementary Appendix for more information). As the input to the CADe system, we extracted random 3D patches of $128 \times 128 \times 128$ from preprocessed scans followed by additional data augmentation. We trained our model using the stochastic gradient descent optimizer with an initial learning rate of 0.001. The batch size was fixed at 8 and the maximum number of training epochs was 100 for all experiments. Focal loss function (Lin et al., 2017) was used to train the model for capturing more true positives amongst all nodule candidates.

To train our CADd model, as a first step, for each patient, we extract clinical metadata features comprising 9 variables (Fig. 1.b). Initially, we studied all available clinical metadata features to select the most informative attributes for lung cancer prediction. NLST clinical metadata contained detailed information with 273 variables, including patient demographics (age, education, ethnicity, gender, etc.), previous disease diagnosis (asthma, emphysema, heart disease, hypertension, etc.), smoking information (smoking years, pack-years, age at smoke onset, average number of cigarettes, age at trial randomization, etc.), personal and family history of cancer, lung cancer-related invasive procedure, work history, etc. All variables were recorded in numeric format, so we normalized variables using min–max normalization to maintain consistency between variables. Then, we implemented a feature selection algorithm using Random Forest to select the most informative clinical features for lung cancer prediction. As a result of this process, we selected nine features (Fig. 1.b), including the level of education completed, pack-year tertile, age, race, gender, ethnicity, smoking status at T0, total years of smoking, and the average number of cigarettes smoked per day.

In the second step, an independent 3D Encoder is designed to predict malignancy scores and extract features on time-series data based on lung-level information. For the input to the model, we concatenate all available preprocessed scans (T0, T1, and T2) of the patient as a 4D input. This model was trained using a stochastic gradient descent optimizer with an initial learning rate of 0.001, batch size of 8, and cross-entropy as the loss function. After finalizing the training procedure, we treat the model as a feature extractor by removing the last fully connected layer and extracting a feature vector including 512 nodes from the output of penultimate fully connected layer. We categorize these 512 features as lung-level time-series features.

In the last step, we established a 3D Encoder combined with additional layers, including an LSTM and two fully connected layers. Input to this model is the nodule-level information achieved from the CADe system. The output of the CADe system is a set of detected nodules $(x_{ij}, y_{ij}, z_{ij}, r_{ij}, p_{ij})$, where (x_{ij}, y_{ij}, z_{ij}) is the central coordinate, r_{ij} is the radius, and p_{ij} is the confidence score of a nodule i in scan j. We first extract this information related to detected nodules on the latest available time-point CT. Then from the set of potential nodules, a single candidate nodule that has the highest confidence score is selected. According to the central coordinate information of the selected candidate nodule, we extract three 3D patches of $64 \times 64 \times 64$ around the nodule from all available time-points. An example of extracted patches from all three time-points for the selected candidate nodule can be seen in Fig. 2.c. We use these patches as inputs to the model. To get the malignancy score, we add 512 features (lung-level) and 9 clinical features (metadata) extracted in steps 1 and 2 to the last fully connected layer of the model. We train our model using a stochastic gradient descent optimizer with an initial learning rate of 0.001, batch size of 16, and cross-entropy as the loss function. Our model was implemented in Python language[1] using Pytorch (Paszke et al., 2019). All experiments were done on a Nvidia Titan RTX 24 GB GPU.

[1] https://www.python.org.

Table 1
Performance of various models analysing a balanced NLST dataset.

Method	Features	CT time-points	ACC	SE	SP	F1	AUC
Support Vector Machine (Cristianini and Shawe-Taylor, 2000)	Metadata	–	0.63	0.57	0.70	0.61	0.63
Random Forest (Liaw et al., 2002)	Metadata	–	0.64	0.69	0.60	0.66	0.64
DeepCAD-NLM-S	Nodule + Lung + Metadata	T2	0.75	0.73	0.77	0.75	0.83
DeepCAD-N-L	Nodule	T0, T1, T2	0.72	0.87	0.57	0.75	0.71
DeepCAD-L-L	Lung	T0, T1, T2	0.78	0.64	0.94	0.75	0.81
DeepCAD-NM-L	Nodule + Metadata	T0, T1, T2	0.72	0.88	0.60	0.77	0.73
DeepCAD-LM-L	Lung + Metadata	T0, T1, T2	0.79	0.67	0.94	0.77	0.82
DeepCAD-NL-L	Nodule + Lung	T0, T1, T2	0.83	0.83	0.85	0.81	0.85
DeepCAD-NLM-L	**Nodule + Lung + Metadata**	**T0, T1, T2**	**0.85**	**0.84**	**0.87**	**0.85**	**0.88**

Statistics. Five statistical measures were used to evaluate and compare different versions of our DeepCAD model: Accuracy (ACC), Sensitivity (SE), Specificity (SP), F1-Score (F1), Area Under Receiver Operating Characteristic Curve (AUC). In all experiments 5-fold cross-validation was implemented and the results on the test sets are expressed as the average of 5-folds. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. To avoid biasing models towards a specific class, we used 0.5 as an optimal operating point.

3. Experimental evaluation

To measure the robustness of our DeepCAD-NLM-L architecture, four different experiments were implemented using the National Lung Screening Trial (NLST) (NLST, 2011; Aberle et al., 2011) and North London-based SUMMIT Lung Cancer Screening Study (Horst et al., 2020).

Experiment 1. In this experiment, we examined our architecture on a balanced dataset including 265 cancer cases equivalent to randomly selected 265 cancer-negative cases with at least two CT time-points (Fig. 1.a). The selected balanced dataset avoided our architecture becoming biased on one class alone (i.e. cancer). Our analyses compared single features or combinations of the three different features (nodule-level, lung-level and clinical metadata), in ablation studies of our model (Table 1). To examine the benefits of using longitudinal datasets, we compared the DeepCAD-NLM-L (longitudinal) model with a DeepCAD-NLM-S (single time point) model trained and tested on the latest time-point CT (T2) only [i.e. a DeepCAD model which omitted the LSTM layer (refer to Section 2)]. We also compared DeepCAD-NLM-L to malignancy classification using traditional machine learning methods such as a Support Vector Machine (SVM) and Random Forest Classifier (RFC) on clinical metadata. In all experiments 5-fold cross-validation was implemented. Table 1 shows the results of Experiment 1 on the NLST dataset. The DeepCAD-NLM-L method combining all available features (nodule-level, lung-level and clinical metadata) and time-points (T0, T1 and T2) outperformed all other models in sensitivity (0.84) with little compromise in specificity (0.87), resulting in the highest AUC (0.88) (Fig. 3.b).

Experiment 2. To further examine the robustness of the model trained in experiment 1, we tested the DeepCAD-NLM-L model on 2240 unseen NLST cases (265 cancer, 1975 non-cancer) to evaluate model performance on a sample size more representative of lung cancer screening programs. Performance, expressed as the average of 5 cross validation folds on the test datasets was: 72%, 83%, 72% and 84%, for accuracy, sensitivity, specificity and AUC, respectively.

Experiment 3. To evaluate the trained DeepCAD-NLM-L architecture on contemporaneous lung cancer screening imaging data, that was out-of-distribution to the data used to train DeepCAD-NLM-L, we analysed data from participants in the contemporary SUMMIT LCS study. The NLST data used to train DeepCAD-NLM-L was acquired almost 20 years ago. Interval improvements in imaging have meant that todays CTs, such as seen in the SUMMIT LCS study have a much narrower slice thickness (<1 mm in SUMMIT vs 2.5 mm typically seen in NLST) and are performed at lower-dose using iterative reconstruction techniques compared to filtered back projection CTs in NLST. To further challenge DeepCAD, SUMMIT cases with varying time intervals between CTs were evaluated. The NLST data used to train DeepCAD-NLM-L had standard intervals of a year between CTs. The selected SUMMIT test subset comprised 89 consecutive cases (18 cancer and 71 non-cancer) with two or more time-point CTs. These included: baseline and 3 month follow up CTs (n = 30), baseline and 12 month follow up CTs (n = 33); baseline, 3 month and 12 month follow up CTs (n = 26).

The performance of the DeepCAD-NLM-L was compared with that of the SUMMIT nodule management (Creamer et al., 2023) protocol. This comparison was made to enable an understanding of the practical impact and potential complementarity for a deep learning system that might integrate with radiologist CT reads. For the purposes of this comparison, a scan was given a label of "cancer" when the SUMMIT radiologist had indicated an urgent patient referral for a suspected lung cancer was required (according to the SUMMIT nodule management protocol).

Table 2 shows the performance of DeepCAD-NLM-L, and the SUMMIT nodule management protocol on the SUMMIT dataset. DeepCAD-NLM-L showed good sensitivity (0.83) despite being trained on NLST data that was very different to that contained within the SUMMIT study. Radiologists applying the SUMMIT nodule management protocol achieved perfect specificity as the protocol was the diagnostic benchmark. Taken together, these findings indicate that employing a computer algorithm could heighten sensitivity in analysing nodules during incident screening rounds within LCS programs. Subsequent radiologist reads could offer high specificity, resulting in an optimal combined performance for LCS. The results indicate the advantages that could be gained in LCS programs when computer algorithms are combined with nodule management algorithms.

Experiment 4. In this final experiment, our aim was to compare the performance of our DeepCAD-NLM-L architecture with that of two thoracic radiologists. Specifically, this comparison was conducted on a subset of the NLST dataset containing the types of lung nodules typically identified during incident screening rounds, which are the most challenging for a radiologist to interpret. The challenging nodules were selected based on size (5–10 mm nodules being the hardest for radiologists to classify as benign versus malignant) and on morphology. Regarding morphology, a nodule with spiculated borders (due to malignant cells extending within pulmonary interstitial tissue) raise suspicions of malignancy. However, a similar appearance is also seen in benign infectious/inflammatory nodules. In contrast, smooth margins to a nodule with well-defined borders (non-spiculated) cannot exclude malignancy as up to 20% of primary lung cancers also have this appearance (Zhao et al., 2014). Accordingly we ensured that the 5–10 mm nodules were evenly divided into spiculated and non-spiculated nodules. And that equal numbers of benign and malignant nodules were found amongst the spiculated and non-spiculated subsets.

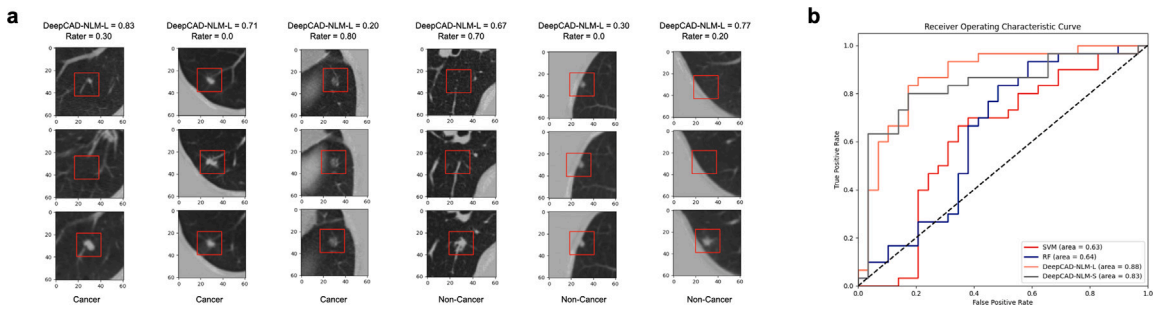


Fig. 3. Some results obtained with analysis of the NLST dataset. (a) Output classification results for malignancy prediction for six different patients in the NLST dataset. Each column corresponds to a patient; each row relates to a separate CT time-point (from top to bottom: T0, T1, and T2). The red bounding box contains the candidate nodule. The malignancy prediction output of our model and rater (average of rater1 and rater2) is shown above the top image. The true nodule diagnosis is shown under the bottom image. (b) Comparison of the AUC curve of DeepCAD-NLM-L with other approaches in one of the folds of Experiment 1 using the NLST dataset.

Table 2
Comparison of DeepCAD with radiologist performance on SUMMIT lung cancer screening data. NMA refers to the nodule management algorithm.

Method	ACC	SE	SP	F1	AUC
SUMMIT NMA	0.93	0.66	1.00	0.80	–
DeepCAD-NLM-L	0.75	0.83	0.73	0.58	0.80

For this experiment, as our DeepCAD-NLM-L model had been previously trained on all eligible cancer cases, the model had to be retrained. CTs containing challenging nodules which had previously been part of the DeepCAD-NLM-L model training dataset (Experiment 1) were removed. DeepCAD-NLM-L was then retrained on a subset of cases where at least three CT time-points were available (85 cancer cases with non-challenging nodules and 180 non-cancer cases). We hypothesized that interpretation of difficult nodules by DeepCAD-NLM-L and radiologists would show a complementary sensitivity and specificity which together would provide an optimal assessment of malignancy risk in a LCS setting.

Evaluation of the newly trained DeepCAD-NLM-L model was performed on a hold out dataset of 25 unseen cancer cases and 75 unseen non-cancer cases which had not been part of the training dataset. Comparison with human performance involved having two radiologists each with an average of 10 years clinical experience independently assess each set of scans in the test dataset. To simulate clinical practice, for each patient, the radiologists read the longitudinal series of scans side-by-side, and assigned a per-patient diagnosis of cancer or non-cancer, together with measure of diagnostic certainty expressed as a percentage. Both the DeepCAD-NLM-L longitudinal models and the radiologists read the scans in two ways: first, using all three time-points, and second, using only the first and last time-points (T0, T1 and T2). To avoid recall bias, the radiologists performed the reading exercise four weeks apart. Whilst the time-points of the CTs were known to the radiologists, the scans were read in a random order.

Table 3 summarizes the results of Experiment 4 on the NLST dataset. DeepCAD-NLM-L showed good sensitivity (0.92) whilst the radiologists showed good specificity (0.72–0.77). The findings suggest that an initial analysis of screening CTs with DeepCAD-NLM-L followed by an evaluation of concerning nodules with radiologists might optimize nodule management in LCS studies. Importantly, the similar performances of both the radiologists and DeepCAD-NLM-L when assessing two as opposed to three CT time-points, suggests that an intermediate annual time-point CT does not influence malignancy prediction. Fig. 3.a depicts the malignancy probability classification of DeepCAD-NLM-L and the radiologists for six subjects.

4. Discussion and conclusion

In this paper, we propose a fully automated pipeline for lung cancer prediction from CT scans. Our DeepCAD-NLM-L model encompasses a nodule detection and malignancy prediction system that combines lung-level, nodule-level, and clinical metadata information to increase prediction performance. Importantly, by leveraging valuable information contained within time-series CT data, our model achieves improved prediction of the likelihood of lung cancer in the most challenging lung nodule subtypes. Our model also demonstrates complementary performance when compared to radiologist interpretation of incident lung nodules emphasizing the importance of integrating human and computer intelligence in LCS programs.

The diagnosis of lung cancer on imaging by radiologists has evolved iteratively over the past 100 years. For mid- or late-stage lung cancer, a confident diagnosis can be made on single time-point imaging. However, as the possibility of stage-shifting cancer diagnosis with earlier detection has become apparent, radiologists have focused on studying changes in the morphology of smaller nodules over time to better distinguish benign from malignant lesions. Longitudinal nodule evaluation underpins lung cancer screening programs and is essential to reduce morbidity in LCS studies from unnecessary investigation of benign lesions and the avoidance of missing a cancer diagnosis.

However to date, most computer algorithms assessing malignancy risk in lung nodules only study single time-point imaging. Radiologists would be remiss if they ignored pertinent historical patient information in the form of old imaging when evaluating lung nodules. Intuitively, one would imagine that computer algorithms would improve malignancy estimation of early cancers by considering any available longitudinal imaging. This concept formed the central premise of our study and was confirmed in the finding that a DeepCAD-NLM-L model evaluating two or three time-point CTs demonstrated better performance than the same model that only utilized a single time-point CT (Table 3). The benefits of using a time-series model for cancer prediction were also emphasized in the results obtained when DeepCAD-NLM-L assessed SUMMIT study data. SUMMIT data was inherently different in composition (image quality and reconstruction, radiation dose, and CT time intervals) compared to that used to train DeepCAD-NLM-L.

Our analyses have also highlighted the limitations when focusing solely on nodule-level or lung-level features. The DeepCAD-N-L model that only considered nodule-level features had a sensitivity of 87% in Experiment 1 but had limited specificity (57%). The DeepCAD-L-L model that utilized lung-level features conversely had a specificity of 94% with a sensitivity of 64%. Combining lung-level and nodule-level features would appear to optimize the necessary trade-off between sensitivity and specificity confirmed in DeepCAD-NL-L, with a higher F1 measure of 81% (Table 1). Our model with multiple time points and all available features (DeepCAD-NLM-L) performed better than the single time-point version (DeepCAD-NLM-S) in all metrics evaluated. These results suggest that using time-series data with additional information can improve prediction.

The proposed DeepCAD-NLM-L model utilizes a LSTM layer. The primary motivation for integrating LSTM layers into the DeepCAD-NLM-L model for lung-level time-series feature extraction derives from their adeptness at capturing long-term dependencies within sequential data, a critical aspect of robust time-series analysis (Lu et al., 2023).

Table 3
Comparison of performance of several models: our DeepCAD method, and two radiologists when assessing challenging nodules in NLST dataset.

Method	Features	CT Time-points	ACC	SE	SP	F1	AUC
Rater 1	–	T0, T1, T2	0.72	0.72	0.72	0.56	0.78
Rater 2	–	T0, T1, T2	0.74	0.60	0.77	0.53	0.75
Rater 1	–	T0, T2	0.71	0.72	0.71	0.55	0.76
Rater 2	–	T0, T2	0.74	0.64	0.77	0.55	0.76
DeepCAD-NLM-L	Nodule + Lung + Metadata	T0, T1, T2	0.71	0.92	0.64	0.61	0.77
DeepCAD-NLM-L	Nodule + Lung + Metadata	T0, T2	0.71	0.92	0.64	0.61	0.76

While RNNs present an alternative, they often face the issue of gradient vanishing (Mozer, 1991), mainly when dealing with limited 3D longitudinal imaging datasets during training. Furthermore, in comparing attention layers with LSTM layers, our findings highlighted the LSTM's advanced model performance, mainly due to their ease of training and ability to effectively navigate hyperparameter optimization. Attention layers increase the number of parameters significantly and, consequently, are more prone to overfitting (particularly when data as in the current study is relatively limited). As a result, the performance of attention layers will eventually decrease when applied to unseen dataset.

Prior studies evaluating lung nodules have generally had a "blunderbuss" approach to nodule datasets, by focusing on all nodules, rather than indeterminate and challenging cases. In doing so, the performance of such models is potentially artificially inflated by the simultaneous inclusion of both easily dismissed nodules (nodules that are too small and would not result in any meaningful intervention even if classified as cancer at an earlier time-point) and nodules that are clearly cancerous.

The lung cancer prediction models in Gao et al. (2019), Ardila et al. (2019) and Venkadesh et al. (2023) reported average AUC values of 82.5% (sensitivity = 61.6%; F1 = 70.8%), 87.3% (sensitivity = 64.7%; specificity = 95.2%) and 92.5% for cases where cancer was diagnosed in the first two years of screening, respectively. However, DeepCAD-NLM-L demonstrated comparable performance metrics to Gao et al. (2019), Ardila et al. (2019) and Venkadesh et al. (2023) when evaluating challenging incident nodules alone (cancer cases diagnosed in the first three years of screening) with AUC value of 88% (sensitivity = 84%; specificity = 87%) with a high resultant F1 score (85%).

Importantly, comparisons between DeepCAD-NLM-L and the models proposed in Gao et al. (2019), Ardila et al. (2019) and Venkadesh et al. (2023) are not simple like-for-like comparisons. Our work began with a vision of aiming to apply deep-learning models to medical problems with the greatest clinical need. In the case of lung cancer screening, the acute need remains assistance for radiologists in the confident characterization of cancers identified on incident CTs, thereby minimizing false positive and negative reporting. We therefore curated a subset of indeterminate-size spiculated and non-spiculated nodules (5–10 mm), precisely the types of nodules that consume a disproportionate amount of radiologist interpretation time. As our study premise has not been considered previously there is paucity of relevant prior work (delineating malignancy in the most challenging subset of nodules) to which we can reasonably compare our model performance.

When we used DeepCAD-NLM-L to assess a curated subset of the most challenging lung nodules found in lung cancer screening programs, our model correctly identified cancerous nodules when present on longitudinal CTs. One would expect the discriminatory ability of the DeepCAD-NLM-L model to distinguish cancers from non-cancers to be markedly improved compared to single time-point trained models (DeepCAD-NLM-S); in other words, DeepCAD-NLM-L should have a high specificity. Such a high specificity would, in turn, allow automated triage of scans with a high probability of lung cancer to urgent lung cancer referral.

Interestingly, in our experiments, this proved to be only partially true. While the DeepCAD-NLM-L model outperformed its single time-point (DeepCAD-NLM-S) and traditional machine-learning model counterparts (SVM and RF), it could not match the performance of individual radiologists on the NLST dataset, nor a rigorous nodule management protocol on the SUMMIT dataset. Conversely, radiologists showed good specificity. This suggests that a composite approach whereby DeepCAD-NLM-L pre-reads time-series CTs and highlights nodules of concern for definitive evaluation by a radiologist could represent an effective screening workflow. With some lung cancer screening programs considering imaging at 2-yearly intervals, it was reassuring to note that DeepCAD-NLM-L with two CT time-points performance was maintained when the second time-point CT (T1) was omitted from time-series analyses in Experiment 4 (Table 2).

The comparable performance of the DeepCAD-NLM-L model against human radiologists in lung cancer prediction may relate to several potential factors. DeepCAD-NLM-L can rapidly process and analyse longitudinal imaging and non-imaging datasets more efficiently and comprehensively than is likely to be possible by humans. This allows the algorithm to consider a broader range of information and extract relevant features more effectively for lung cancer prediction. Furthermore, the model is powerful at identifying relevant and specific patterns within datasets (imaging and non-imaging) that may be invisible to human observers. Specifically with regard to cancerous nodules, the model may be able to determine subtle details and variations in nodule texture, shape, homogeneity and density, which might not be discerned with the human eye.

Our study had several limitations. In our pipeline, one candidate nodule was specified for analysis from all the potential nodules generated by the CADe system. This could constrain our CADd system in situations when the model selects a noncancerous nodule as the candidate nodule in cases containing other cancerous nodules. To mitigate this problem, we aim to develop a new pipeline that uses a selection of candidate nodules as input to the CADd system.

DeepCAD-NLM-L also focuses on the detection of lung nodules which is not synonymous with lung cancer detection. Lung cancers can be located in regions other than the lung itself. An important example of this is small cell lung cancer which may be entirely contained within the mediastinum and therefore not detected by lung nodule detection systems. A future aim would be to incorporate information from mediastinal lung reconstruction kernals when assessing patient level cancer risk.

In conclusion, in this paper, we show that the combination of different levels of features in the DeepCAD-NLM-L model including clinical metadata and imaging data at the lung and nodule-level across longitudinal time-point CTs provides a good estimation of malignancy particularly for incident screening nodules that are challenging for a radiologist to interpret. DeepCAD-NLM-L shows complementary performance metrics of sensitivity and specificity when compared to nodule management algorithms emphasizing the role such tools may have in rationalizing the assessment of lung cancer screening CTs.

Code availability

Under the specific conditions the code can be made available to those contacting the corresponding author.

Ethical statements

Anonymized data retrospectively analysed in this study was acquired with full informed consent from all subjects and/or their legal guardian(s) by the NLST and SUMMIT studies. Analysis of data from

the SUMMIT study was approved by University College London research ethics committee and the Leeds East Research Ethics Committee: 20/YH/0120. The analysis of data in NLST was approved by UCL Research Ethics Committee:15401/002. All research was performed in accordance with the Declaration of Helsinki. Illustrations and Tables

CRediT authorship contribution statement

Shahab Aslani: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Pavan Alluri:** Data curation. **Eyjolfur Gudmundsson:** Data curation. **Edward Chandy:** Data curation. **John McCabe:** Data curation. **Anand Devaraj:** Formal analysis, Writing – review & editing. **Carolyn Horst:** Data curation, Writing – review & editing. **Sam M. Janes:** Formal analysis, Resources, Writing – review & editing. **Rahul Chakkara:** Data curation, Supervision. **Daniel C. Alexander:** Formal analysis, Writing – review & editing. **SUMMIT consortium:** Data curation. **Arjun Nair:** Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Joseph Jacob:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shahab Aslani reports financial support was provided by Cancer Research UK. Joseph Jacob reports financial support was provided by Wellcome Trust Clinical Research Career Development Fellowship. This work was supported by Cancer Research UK (C68622/A29390). The authors thank the National Cancer Institute for access to NCI's data collected by National Lung Screening Trial (NLST). The statement contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. J.J and this research was supported by Wellcome Trust Clinical Research Career Development Fellowship 209553/Z/17/Z. For the purpose of open access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission. This project, J.J, E.G, SH.A, SM.J, A.N and DC.A were also supported by the NIHR UCLH Biomedical Research Centre, UK. The SUMMIT Study is funded by GRAIL LLC. through a research grant awarded to SM.J as Principal Investigator. SM.J is supported by CRUK programme grant (EDDCPGM/100002), and MRC Programme grant (MR/W025051/1). SM.J receives support from the CRUK Lung Cancer Centre of Excellence (C11496/ A30025) and the CRUK City of London Centre, the Rosetrees Trust, the Roy Castle Lung Cancer foundation, the Longfonds BREATH Consortia, MRC UKRMP2 Consortia, the Garfield Weston Trust and University College London Hospitals Charitable Foundation. SM.J's work is supported by a Stand Up To Cancer-LUNGevity- American Lung Association Lung Cancer Interception Dream Team Translational Research Grant and Johnson and Johnson (grant number: SU2C-AACR-DT23-17 to S.M. Dubinett and A.E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C.This work was partly undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme.

J.J reports consultancy fees from Boehringer Ingelheim, F. Hoffmann-La Roche, GlaxoSmithKline and NHSX. J.J reports advisory boards in Boehringer Ingelheim and F. Hoffmann-La Roche. J.J reports lecture fees from Boehringer Ingelheim, F. Hoffmann-La Roche, and Takeda. J.J reports grant Funding from GlaxoSmithKline, Wellcome Trust, Microsoft Research and Gilead Sciences. J.J reports patents UK patent application numbers 2113765.8 and GB2211487.0. J.J was supported by Wellcome Trust Clinical Research Career Development Fellowship 209553/Z/17/Z. SM.J has received fees for advisory board membership in the last three years from Bard1 Lifescience. SM.J has received grant income from GRAIL Inc. SM.J is an unpaid member of a GRAIL advisory board. SM.J has received lecture fees for academic meetings from Cheisi and Astra Zeneca. SM.J's wife works for Astra Zeneca. SM.J reports fees from Astra-Zeneca, Bard1 Bioscience, Achilles Therapeutics, and Jansen unrelated to the submitted work. SM.J received assistance for travel to meetings from Astra Zeneca to American Thoracic Conference 2018 and from Takeda to World Conference Lung Cancer 2019 and is the Investigator Lead on grants from GRAIL Inc, GlaxoSmithKline plc and Owlstone. A.N reports advisory fees from Part-funded by UCLH Biomedical Research Centre (BRC), National Institute for Health Research (NIHR); Member of Advisory Board, Aidence, Artificial Intelligence BV; Co-Investigator, Integration and Analysis of Data Using Artificial Intelligence to Improve Patent Outcomes with Thoracic Diseases (DART) study; Scientific Advisory Board, iDx Lung Trial; Collaborator/Advisory fees Merck Sharp and Dohme (MSD) (UK) Limited; Speaker Fees, Astra Zeneca (AZ) UK Ltd. A.D's disclosures are fees from Boehringer Ingelheim, Roche, Brainomix, and Vicore. P.A and R.C are the founder of ManasAI, an AI SaaS company specializing in predictive models. P.A and R.C. own equity in ManasAI. SH.A, E.G and DC.A is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. E.CH, J.M and C.H have no competing interests to declare. The remainder of the SUMMIT consortium declare no competing interest with regard to the current manuscript.

Data availability

The NLST dataset in a publicly available dataset and can be requested through the official procedure of the study: https://cdas.cancer.gov/learn/nlst/images/. The SUMMIT is an on going study and it cannot be made publicly available due to confidentiality.

Acknowledgements

This work was supported by Cancer Research UK (C68622/A29390). The authors thank the National Cancer Institute for access to NCI's data collected by National Lung Screening Trial (NLST). The statement contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. J.J and this research was supported by Wellcome Trust Clinical Research Career Development Fellowship 209553/Z/17/Z. For the purpose of open access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission. This project, J.J, E.G, SH.A, SM.J, A.N and DC.A were also supported by the NIHR UCLH Biomedical Research Centre, UK. The SUMMIT Study is funded by GRAIL LLC. through a research grant awarded to SM.J as Principal Investigator. SM.J is supported by CRUK programme grant (EDDCPGM/100002), and MRC Programme grant (MR/W025051/1). SM.J receives support from the CRUK Lung Cancer Centre and the CRUK City of London Centre, the Rosetrees Trust, the Roy Castle Lung Cancer foundation, the Longfonds BREATH Consortia, MRC UKRMP2 Consortia, the Garfield Weston Trust and University College London Hospitals Charitable Foundation.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2024.102399.

References

Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., et al., 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. 365 (5), 395–409.

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. 25 (6), 954–961.

Black, W.C., Gareen, I.F., Soneji, S.S., Sicks, J.D., Keeler, E.B., Aberle, D.R., Naeim, A., Church, T.R., Silvestri, G.A., Gorelick, J., et al., 2014. Cost-effectiveness of CT screening in the National Lung Screening Trial. N. Engl. J. Med. 371, 1793–1802.

Causey, J.L., Guan, Y., Dong, W., Walker, K., Qualls, J.A., Prior, F., Huang, X., 2019. Lung cancer screening with low-dose CT scans using a deep learning approach. arXiv preprint arXiv:1906.00240.

Creamer, A.W., Horst, C., Dickson, J.L., Tisi, S., Hall, H., Verghese, P., Prendecki, R., Bhamani, A., McCabe, J., Gyertson, K., et al., 2023. Growing small solid nodules in lung cancer screening: safety and efficacy of a 200 mm3 minimum size threshold for multidisciplinary team referral. Thorax 78 (2), 202–206.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press.

Ding, J., Li, A., Hu, Z., Wang, L., 2017. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 559–567.

Gao, R., Huo, Y., Bao, S., Tang, Y., Antic, S.L., Epstein, E.S., Balar, A.B., Deppen, S., Paulson, A.B., Sandler, K.L., et al., 2019. Distanced LSTM: time-distanced gates in long short-term memory models for lung cancer detection. In: International Workshop on Machine Learning in Medical Imaging. pp. 310–318.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Horst, C., Dickson, J.L., Tisi, S., Ruparel, M., Nair, A., Devaraj, A., Janes, S.M., 2020. Delivering low-dose CT screening for lung cancer: a pragmatic approach. Thorax 75 (10), 831–832.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18 (2), 203–211.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Jiang, H., Gao, F., Xu, X., Huang, F., Zhu, S., 2020. Attentive and ensemble 3D dual path networks for pulmonary nodules classification. Neurocomputing 398, 422–430.

Khosravan, N., Bagci, U., 2018. S4ND: Single-shot single-scale lung nodule detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 794–802.

Koning, H.J., van der Aalst, C.M., de Jong, P.A., Scholten, E.T., Nackaerts, K., Heuvelmans, M.A., Lammers, J.-W.J., Weenink, C., Yousaf-Khan, U., Horeweg, N., et al., 2020. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N. Engl. J. Med. 382 (6), 503–513.

Kuan, K., Ravaut, M., Manek, G., Chen, H., Lin, J., Nazir, B., Chen, C., Howe, T.C., Zeng, Z., Chandrasekhar, V., 2017. Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge. arXiv preprint arXiv:1705.09435.

Li, Y., Fan, Y., 2020. DeepSEED: 3D squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection. In: IEEE 17th International Symposium on Biomedical Imaging. ISBI, pp. 1866–1869.

Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE Trans. Neural Netw. Learn. Syst. 30, 3484–3495.

Liaw, A., Wiener, M., et al., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lu, Y., Aslani, S., Zhao, A., Shahin, A., Barber, D., Emberton, M., Alexander, D.C., Jacob, J., 2023. A hybrid CNN-RNN approach for survival analysis in a Lung Cancer Screening study. Heliyon 9 (8).

Mozer, M.C., 1991. Induction of multiscale temporal structure. Adv. Neural Inf. Process. Syst. 4.

NLST, 2011. The national lung screening trial: Overview and study design. Radiology 258 (1), 243–253.

Ozdemir, O., Russell, R.L., Berlin, A.A., 2019. A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans. IEEE Trans. Med. Imaging 39 (5), 1419–1429.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. Vol. 32, pp. 8024–8035.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., ISBN: 1558604790.

RCR, 2021. Royal college of radiologists: Clinical radiology UK workforce. URL https://www.rcr.ac.uk/media/30dhjeh2/clinical_radiology_census_report_2021.pdf.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Trajanovski, S., Mavroeidis, D., Swisher, C.L., Gebre, B.G., Veeling, B.S., Wiemker, R., Klinder, T., Tahmasebi, A., Regis, S.M., Wald, C., et al., 2018. Towards radiologist-level cancer risk assessment in CT lung screening using deep learning. arXiv preprint arXiv:1804.01901.

Venkadesh, K.V., Aleef, T.A., Scholten, E.T., Saghir, Z., Silva, M., Sverzellati, N., Pastorino, U., van Ginneken, B., Prokop, M., Jacobs, C., 2023. Prior CT improves deep learning for malignancy risk estimation of screening-detected pulmonary nodules. Radiology 308 (2), e223308.

Wang, J., Gao, R., Huo, Y., Bao, S., Xiong, Y., Antic, S.L., Osterman, T.J., Massion, P.P., Landman, B.A., 2019. Lung cancer detection using co-learning from chest CT images and clinical demographics. In: Medical Imaging: Image Processing. pp. 365–371.

WHO, 2022. Cancer report. URL https://www.who.int/news-room/fact-sheets/detail/cancer.

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R.H., Aerts, H.J., 2019. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin. Cancer Res. 25 (11), 3266–3275.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128.

Zhao, Y.R., Heuvelmans, M.A., Dorrius, M.D., van Ooijen, P.M., Wang, Y., de Bock, G.H., Oudkerk, M., Vliegenthart, R., 2014. Features of resolving and nonresolving indeterminate pulmonary nodules at follow-up CT: the NELSON study. Radiology 270 (3), 872–879.

Zhu, W., Liu, C., Fan, W., Xie, X., 2018. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In: IEEE Winter Conference on Applications of Computer Vision. WACV, pp. 673–681.