IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2022, accepted March 20, 2022, date of publication March 24, 2022, date of current version April 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161954

Deep Learning-Based Long Term Mortality Prediction in the National Lung Screening Trial

YAOZHI LU[1,3], SHAHAB ASLANI[1,4], MARK EMBERTON[2], DANIEL C. ALEXANDER[1,3], AND JOSEPH JACOB[1,4]

[1]Centre for Medical Image Computing, University College London, London WC1V 6LJ, U.K.
[2]Division of Surgery and Interventional Science, University College London, London W1W 7TS, U.K.
[3]Department of Computer Science, University College London, London WC1E 6BT, U.K.
[4]Department of Respiratory Medicine, University College London, London WC1E 6BT, U.K.

Corresponding author: Yaozhi Lu (yz.lu@ucl.ac.uk)

This work was supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK [C23017/A27935], Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester.

ABSTRACT In this study, the long-term mortality in the National Lung Screening Trial (NLST) was investigated using a deep learning-based method. Binary classification of the non-lung-cancer mortality (i.e. cardiovascular and respiratory mortality) was performed using neural network models centered around a 3D-ResNet. The models were trained on a participant age, gender, and smoking history matched cohort. Utilising both the 3D CT scan and clinical information, the models can achieve an AUC of 0.73 which outperforms humans at cardiovascular mortality prediction. The corresponding F1 and Matthews Correlation Coefficient are 0.60 and 0.38 respectively. By interpreting the trained models with 3D saliency maps, we examined the features on the CT scans that correspond to the mortality signal. By extracting information from 3D CT volumes, we can highlight regions in the thorax region that contribute to mortality that might be overlooked by the clinicians. Therefore, this can help focus preventative interventions appropriately, particularly for under-recognised pathologies and thereby reducing patient morbidity.

INDEX TERMS Computed tomography, deep learning, lung, saliency map.

I. INTRODUCTION

A. OVERVIEW

Cardiac and respiratory illnesses are the leading causes of mortality globally [1], [2], especially amongst older age groups. The ageing global population means that greater numbers of patients with multimorbid conditions are utilising healthcare services with increasing frequency and for ever more complex problems. Responding to such growing healthcare needs requires cost-effective approaches for the early detection of disease. Early detection allows timely intervention before diseases become irreversible. In this study, a joint human-computer approach is proposed to identify imaging features on CT that are predictive of mortality in patients undergoing Lung Cancer Screening (LCS).

Annual Computed Tomography (CT) imaging in LCS studies has been demonstrated to be an effective screening tool for the early detection of lung cancer, reducing lung cancer mortality in the National Lung Screening Trial (NLST) [3]–[5]. In this study, CT scans from NLST are examined with a 3D-ResNet [6], [7] to predict imaging features associated with long-term mortality outcomes. To allow patients and clinicians to understand the morphological basis for the mortality signal on the CT image, it is imperative that the model is able explain the relationship between the clinical outcome data and the imaging labels. Saliency map methods [8], [9] are popular approaches to highlight relevant features on images that can explain deep learning models and localise contributing features to predictions. The visual interpretation of model classifications through saliency maps, for example when predicting long-term mortality, can aid radiologists in the identification of unsuspected pathology on a CT scan. In turn, this may allow early interventions that may prevent or delay the occurrence of adverse health events, thereby prolonging a patients' life expectancy. This is of particular relevance to LCS where patients have numerous comorbidities but where detecting lung cancer is typically the overarching aim of the radiologist.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhen Ren.

B. LITERATURE REVIEW

Lung cancer and Cardiovascular Diseases (CVD) share several similar risk factors including smoking (both active and passive) and exposure to fine particulates from air pollution [10]. Though the pathophysiological mechanisms differ, it has been shown that smoking leads to increased mortality risk from both lung cancer and CVD [10], [11]. It is therefore logical that patient cohorts enriched with heavy smokers, such as LCS studies (e.g.the National Lung Screening Trial (NLST)) [3]–[5] can be used for the prediction of CVD-related mortality.

Recently, attempts have been made to predict CVD-related mortality in the NLST cohort [12], [13]. These studies have adopted Deep Learning (DL) based approaches to analyse NLST Low-Dose CT scans (LDCT). As demonstrated in van Velzen *et al.* 2019 [12], a Convolutional Autoencoder (CAE) was trained to derive abstract image features and the features were fed into 3 separate classifiers to predict CVD-related mortality. The CAE encoded the automatically extracted 3D LDCT volume around the heart and exported the image features to the subsequent classifiers. The support vector machine classifier achieved performances in terms of Area under ROC curve (AUC) of 0.72. Though the study recognised the value of using clinical information for prediction, including handcrafted variables such as the Coronary Artery Calcium (CAC) score [14], [15] which is a known predictor of CVD, such information was not utilised in making predictions. Instead, the study demonstrated that it is possible to predict CVD-related mortality from LDCT scans alone.

Predicting CVD-related mortality was further improved by Guo *et al.* 2020 [13]. A multimodal approach was adopted in this study where models incorporated both LDCT imaging information and handcrafted features to make mortality predictions. Firstly, for imaging data, a dual branch CNN network was adopted. Two 2D ResNets [6] were used to analyse manually selected 2D axial slides at 2 different resolutions/magnifications: the whole lung region and an automatically cropped cardiac region. For the clinical data, manually derived metrics such as CAC, were used to train a linear support vector machine classifier. When the contributions of the imaging features and clinical data were optimised, this approach improved the AUC performance to 0.82. Both approaches show improvement over human performance in this regard. In fact, as reported by Guo *et al.*, visual inspection of the coronary artery calcium measure by a radiologist could only achieve performance with an AUC of 0.64. However, despite the improvement in classification performance, there are a few drawbacks that require further study. The 2D axial slices were chosen from the 3D LDCT volumes based on the visibility of the coronary artery and thus the task complexity for the CNNs was reduced considerably. Additionally, the amount of human effort in curating the imaging data means that the approach can not scale well to real-world clinical scenarios.

One way to lessen the burden on clinicians to provide image-derived variables to models would be to automate the CAC score acquisition. The deep-learning based approach proposed by van Velzen *et al.* 2020 [16] demonstrated the feasibility of scoring CAC and thoracic aortic calcification (TAC) automatically. More importantly, it had been shown that the model can be performed on CT scans derived using a variety of different CT imaging protocols.

Other than the direct prediction of mortality, the NLST cohort has also been analysed for long-term mortality risk stratification. In Lu *et al.* 2019 [17], a 2D CNN (inception-v4 architecture) was used to analyse chest radiographs in the NLST cohort. The study used Gradient-weighted Class Activation Maps (GradCAM) [9] to isolate the contributing imaging features. As illustrated by the classification maps [17], the deep learning model tended to focus on the cardiac region to search for cardiovascular and respiratory mortality signals. The results demonstrate that the proposed approach can stratify long-term mortality risk, and help identify patients that might benefit the most from preventative interventions.

More recently, various studies have attempted to incorporate medical imaging features together with temporal features [18], [19] for diagnosis purposes. By combining 2D CNN and RNN (Recurrent Neural Network), the joint model outperformed the base CNN models in both studies. In addition, by explicitly address the irregular followup interval which is common under clinical setting, Gao *et al.* 2020 [19] improved the performance of a standard LSTM (Long Short-Term Memory) network in evaluating the malignancy of pulmonary nodules. These demonstrate the feasibility of using a deep learning based approach to monitor the disease progression. More importantly, the findings illustrate the superior performance of using time-series data over a single snapshot in making the diagnosis.

C. OBJECTIVES

This study aims to use both 3D LDCT lung volumes and readily available clinical data from NLST to make long-term mortality predictions, with minimum human input. More importantly, we aim to use saliency maps to identify contributing features on CT images that link to long-term mortality, particularly non-lung-cancer-related mortality. This can help radiologists and clinicians diagnose pathology that might not be obvious at first glance, or help corroborate the presence of subtle but important damage. Such an approach, highlighting neglected features on CT imaging in lung cancer screening populations may speed up decision making for radiologists and clinicians and help optimise patient care.

The key contributions of the study are as follows:

- A 3D deep-learning-based approach to predict long-term outcome based on imaging of the thorax.
- Propose an approach to identify areas of concern on CT volumes that by allowing preventative interventions can improve healthcare efficiency and utilise personalised intervention to improve patient quality of life.

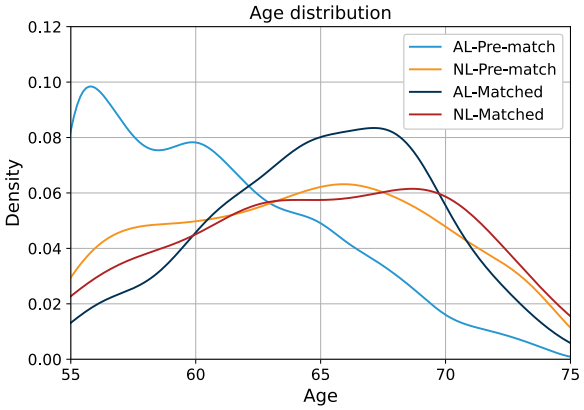
II. METHODOLOGY

A. THE NATIONAL LUNG SCREENING TRIAL

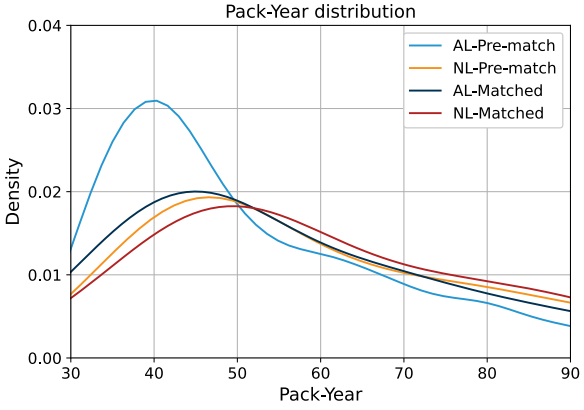
The National Lung Screening Trial was a multi-centered lung cancer screening study conducted in the US from 2002 to 2007 [3]–[5]. 53,454 heavy smokers, aged 55-74 years, who were at high risk for developing lung cancer were recruited. They were randomly assigned to either the low-dose CT (LDCT) branch (26,772 participants) or the chest radiography branch (26,732 participants) of the study. Three annual screening attendances for CT imaging, denoted as T0-T2, were conducted. The NLST study's primary end-point was mortality. The participants' survival status and the cause of death were ascertained through evaluation of death certificate ICD-10 (International Classification of Diseases, 10th edition) codes. In the LDCT branch, the main cause of death (as of the end of 2009) was cardiovascular disease (26.1%), followed by lung cancer (22.9%) and then other types of cancer (22.3%) [4]. In 2015, the survival status of the cohort was updated. In our study, the latest LDCT scan (i.e. T2 screening) and accompanying clinical data, were used to predict long-term mortality in the NLST population. The participant's survival status in 2015 was chosen as the ground-truth label.

B. DATASET SELECTION

Based on the patients' mortality status in 2015, the NLST dataset can be grouped into 3 classes: lung-cancer-induced mortality (LC), non-lung-cancer-induced mortality (NL), and alive (AL). The first group (i.e. LC) was withheld for later use to evaluate the ability of saliency maps models to localise known imaging features associated with patient mortality. For patients in the second group, only cases where the cause of death related to cardiac (ICD-10 codes: I10-I52) or respiratory diseases (ICD-10 codes: J00-J99) were selected. Furthermore, only cases with all three screenings timepoints (i.e. T0, T1 and T2) and with CT scan thickness in the axial plane of no more than 2.5mm were kept. The former criterion aimed to avoid bias introduced by patients who left the trial early for unknown reasons. 873 eligible NL cases were identified from the metadata of the dataset based on the above mentioned criteria. The corresponding CT images were reviewed by a radiologist to exclude cases with imaging artefacts and anatomical biases. As we had limited specialist's time for this manual review process, only 180 of the eligible NL cases (n=873) were randomly selected for analysis. To ensure a homogeneous dataset, scans with streak artefacts in the cardiac region (n=12) and patients with severe forms of thoracic spinal scoliosis (n=2) were excluded, resulting in a study population of 166 cases labelled as dying of either cardiac or respiratory death. The selected NL mortality cases were age, gender, and smoking history matched in a 1:2 ratio with a control population of participants alive at the end of the 2015 follow up period (the AL class). The final study population therefore comprised 498 cases, a third of whom (n=166) had died of cardiac or respiratory causes and n=332 remained alive in 2015.



(a) Age distribution.



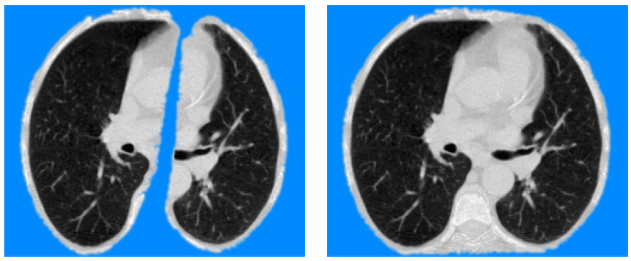
(b) Pack-Year distribution.

FIGURE 1. Pack-Year and age distribution of the matched dataset.

As a result of NLST's inclusion criteria of having a prolonged smoking history (i.e. minimum pack-year of 30), the pack-year distribution in the study population is skewed to the right which is illustrated in Fig. 1. The non-lung-cancer mortality class (NL) was older and had a greater smoking history than the control class (AL). To allow equivalent matching, criteria were relaxed to have a tolerance of ± 10 pack-year and ± 5 year age difference. As illustrated in Fig. 1, the pack-year and age distribution of the control cases were shifted to the right to satisfy the matching criteria.

C. LUNG CT VOLUME PRE-PROCESSING

The pre-processing approach used in Liao *et al.* [20] was adopted for this study. For each axial CT slice, the image was filtered with a Gaussian filter and a -600 Hounsfiled Unit (HU) threshold was applied to create the binarised slice. Absolute size and eccentricity of connected components were then used to filter out small components and imaging noise. The resulting 3D volumes were then filtered by their size (0.68 - 7.5 L) and distance to the centre of the scan. The results were joint to create the approximate lung mask. Morphological transformations, i.e. erosion, dilation, and convex-hull calculation, were performed to further separate the result into the left and right lung masks.



(a) From the original method. (b) From the modified method.

FIGURE 2. Comparison of the two pre-processing methods.

The Hounsfield Unit (HU) range was clipped to an interval −1200 HU to 600 HU. The range was linearly normalised to the interval 0 and 255. The regions outside the masks corresponding to the surrounding tissue were filled with an average value of 170. The pre-processing pipeline was applied over all axial layers to extract the lungs in three dimensions.

Given that cardiovascular disease contributes to majority [4] of the mortality seen in the NLST dataset, it was felt important to preserve the cardiac region. Thus, an additional convex-hull calculation was performed on the joint lung masks to recover the CT information overlying the heart. The comparison between the original approach and the modified approach on the same axial slice is illustrated in Fig. 2. The region in blue corresponds to tissue outside the masks which was given an average intensity of 170. As is evident from Fig. 2(b), the modified pipeline preserves the cardiac region during pre-processing.

D. MACHINE LEARNING MODELS

A two-tier approach was adopted to examine the effectiveness of different types of data, i.e. medical imaging and clinical data, in making mortality predictions. Firstly, non-deep-learning models were used to evaluate classification performance using only clinical data. Secondly, deep-learning-based methods were applied to the medical images to predict mortality. Finally, to assess whether the clinical information complements the CT scans in making mortality predictions, both clinical and imaging data were combined.

1) NON-DEEP-LEARNING MODELS

The Support Vector Machine (SVM) classifier (Model A), the Gradient Boosting Machine (GBM) classifier (Model B), and the Random Forest (RF) Classifier (Model C) were used to analyse the clinical data in the NLST dataset. The clinical data contained patient demographics, previous disease diagnosis, age, smoking history, and pack-year etc. A grid search with cross-validation was used to optimise the parameter settings for the three models. In addition to performing long-term mortality classification, the tree-based methods have the benefit of impurity-based feature selection. 11 features, tabulated in Table 1, were selected to complement the CT imaging when making mortality predictions in the later part of the study. The 4 principal components encapsulating time information consistently ranked among the most important mortality-predicting feature for both tree-based models.

TABLE 1. Selected clinical features.

Feature	Description
BMI	Body mass index in kg/m^2
gender	Patient's gender
diagchas	Pre-NLST diagnosis of childhood asthma
diagemph	Pre-NLST diagnosis of emphysema
diaghear	Pre-NLST diagnosis of heart diseases or heart attack
diaghype	Pre-NLST diagnosis of hypertension
invaslc	Lung-cancer related invasive procedure
PCA 1-4	Principal components of the patient's smoking history. The underlying features are: smoking years, pack-year, age at smoke onset, age at trial randomisation, and average number of cigarettes per day

2) DEEP-LEARNING

The deep-learning-based models in this study were based on a 3D implementation of the ResNet [6]. The pre-trained weights from Chen et. al. [7] were used for transfer learning purposes. The models were originally trained for medical imaging (CT images) segmentation tasks and had been shown effective in performing pulmonary nodule classification through transfer learning.

Two variants of the neural network were investigated: (1) the CT volume-only model (Model D), and (2) the multimodal model combining clinical and imaging data (Model E). They differ in the input data utilised and the corresponding ResNet-based architectures are illustrated in Fig. 3 and 4. The 3D ResNet backbone together with the pre-trained weights were used to analyse the input 3D CT volume. To transfer the pre-trained model to long-term mortality prediction in this study, the 3D ResNet backbone's learning rate was reduced to 1/10 of the rest of the network. The output from the ResNet backbone was converted to a 1D tensor after passing through a global adaptive average pooling layer. In the CT-scan-only model (i.e. Fig. 3), the tensor was passed through 2 fully connected layers (including dropout with $p = 0.5$ and ReLU activation) with 512 and 32 neurons respectively. By contrast, in the multimodal version (i.e. Fig. 4), the clinical data (11 by 1 tensor) was concatenated with the output from the first fully connected layer before passing through the second layer with 43 neurons. Applying the concatenation at the later layer had the aim of providing more weight to the clinical information.

To counter the class imbalance in the 1:2 matched dataset in this study, a weighted random sampler, which assigns a sampling probability inversely proportional to the class size was utilised. This approach attempts to produce, on average, balanced batches during training. Cross entropy loss was used as the loss function for the binary classification.

For the multimodal model as illustrated in Fig. 4, the two branches of the network were trained separately. The imaging branch of Model E inherited the weights from Model D and the training of the CNN portion was not optimised while training Model E. In contrast, the clinical data branch was optimised during training and the clinical measures were

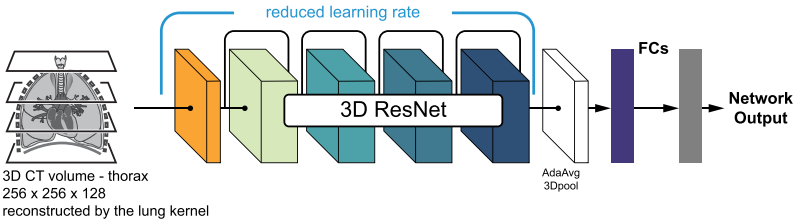


FIGURE 3. 3D ResNet architecture for transfer learning (Model D). CT scans alone are passed to the network.

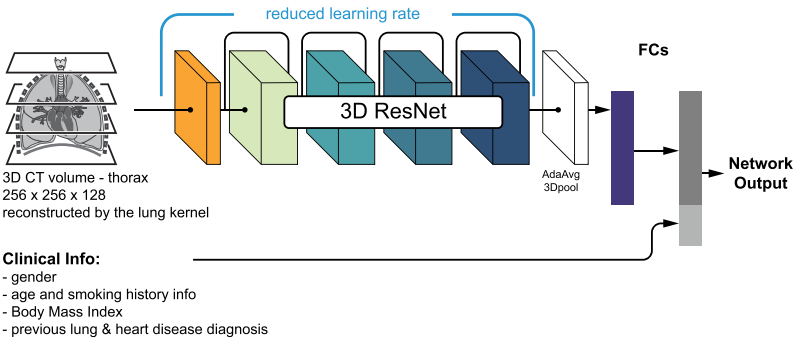


FIGURE 4. 3D ResNet architecture for transfer learning (Model E). Both CT scans & clinical data are passed to the network.

normalised to the range between 0 and 1. The alternative training strategy where the two branches, i.e. medical imaging and clinical data, are jointly trained were found to result in inferior mortality prediction performance.

E. SALIENCY MAPS

The saliency map methods were used to interpret the neural networks and to localise the contributing features to the predictions. As discussed earlier, performing saliency-map-based checks on the withheld lung cancer mortality class provided confidence in the trained models. In a clinical setting, saliency map visualisation of pathological regions on the CT contributing to mortality risk are crucial for clinician interpretation of potentially opaque neural network classifications. Saliency maps can also focus the attention of clinicians on potentially neglected or unrecognised structures on CT that are contributors to morbidity and mortality. This is particularly important in a setting where the detection of lung cancer typically takes primacy.

Common saliency map methods include the gradient approach [8], Gradient-weighted Class Activation Mapping (GradCAM) [9], Guided Backpropagation [21], and Guided GradCAM [9] etc. As elaborated in Adebayo et el. 2018 [22], the Guided Backpropagation and Guided GradCAM approaches tend to generate saliency maps that are independent of the data and model parameters, and cannot be relied upon to explain the model's class prediction. The gradient approach and GradCAM are free from such shortcomings and are the approaches adopted in this study. Though both approaches can highlight relevant features for model predictions, the GradCAM approach tends to have lower resolution than the gradient approach, and thus can be limited when focusing on subtle/small features.

To filter out noise, thresholding was applied to the vector extracted from the last convolutional layer before it was extrapolated in 3D to the same dimension as the input CT volume. The rationale of applying thresholding before rather than after the interpolation step was to maximise the localisation capability of the resulting GradCAMs.

F. EXPERIMENTS

For both the machine learning and deep learning models, a grid search approach was adopted to tune the hyper parameters. The hyperparameter value that produced the best AUC value was chosen. The values and settings for the key parameters are tabulated in Table 2.

For the deep learning based approach, the 3D CT volumes were interpolated into a 256 by 256 by 128 volume before being passed into the neural networks. The 10-layer version of the 3D ResNet backbone was adopted in this study as it was felt that deeper models might only lead to marginal performance improvement. The Adam algorithm [23], which is an adaptive learning rate optimisation algorithm, was implemented to train the network with an initial learning rate of 1E-4. The training was regularised by L2 regularisation (with weight decay parameter of 5E-3). The networks were trained over 1,000 epochs.

To assess the performance of the models, 5-fold cross-validation was performed. Of the 498 cases in the study dataset, 48 (9.6%) were reserved for validation. The remaining 450 cases were used as training (72.3%) and testing (18.1%) sets and were split into 5 folds of 90. The relative class composition (i.e. 1:2 dead vs alive cases) was

TABLE 2. Key hyperparameters.

Model	Key Parameter	Value /Option
Support Vector Machine	Kernel	Linear
	C	10
Gradient Boosting Machine	Learning rate	0.2
	Maximum Depth	4
	Number of boosting stages	16
Random Forest	Number of trees in the forest	32
	Maximum depth	8
	Minimum number of samples required to split a node	5% (of the dataset)
3D ResNets	ResNet depth	10 layers
	Dropout probability	0.5
	Optimiser	Adam
	Learning rate	1.00E-04
	Weight decay	5.00E-03

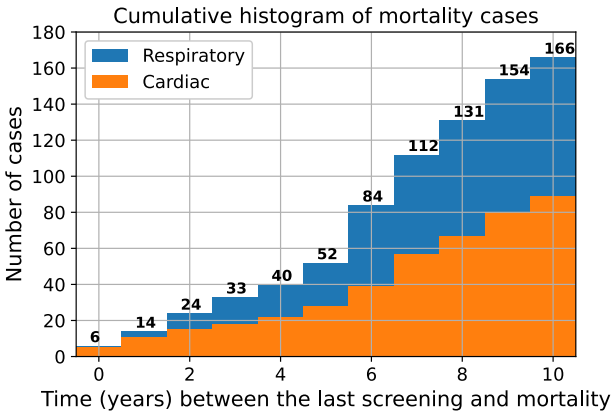


FIGURE 5. Cumulative distribution of the survival time.

maintained in all the subsets. The selection of cases in the cross-validation folds was kept the same for all the models used. In each of the 5 cross-validation experiments, a machine-learning /deep-learning model was trained and evaluated. 39 lung-cancer-induced mortality (LC) cases with malignant nodules were held out for later saliency map visualisation.

To compare the performance with related studies [12], [13], the models were assessed with the Area Under the Curve (AUC) metric. The average value and the standard deviation of the AUC values from the 5-fold cross validation were used to gauge the overall performance of the models. Additionally, other performances metrics such as sensitivity, specificity, F1 scores, and Matthews Correlation Coefficient (MCC) [24] were also examined.

The tools used in the study was developed using Python (v3.8.5). The non-deep-learning models (e.g. SVM) in this study were implemented through the Scikit-Learn library (v0.23.2) while the neural networks were implemented through the PyTorch library (v1.7.1). Each neural network was trained with a batch size of 24 on a single Nvidia RTX 8000 GPU (with 48GB of memory) on the UCL Computer Science cluster. The average training time for each network model was less than 24 hours.

III. RESULTS

The median age in dataset used in this study was 66 and the median pack-year history was 54. Among the mortality cases, as shown in Fig. 5, the time difference between the latest screening point Ii.e. T2) and the time to patient death ranged from 0 to 10 years. Overall, 53.6% (89 patients) of the NL deaths were caused by cardiovascular diseases whilst the remainder (46.4%, 77 cases) were due to respiratory illnesses.

The performance metrics of the models, in terms of the average performance of the 5-fold cross-validation, are tabulated in Table 3. Models A-C, the non-deep-learning models were only trained with clinical data. Of the two neural network models, Model D was trained on CT images while Model E was trained on both medical imaging and clinical data. The corresponding ROC curves from the neural network models are grouped according to the type of input used and illustrated in Fig. 6. Similarly, the precision-recall curves are presented in Fig. 7. The composition of the cross-validation folds was maintained across the models.

Among the non-deep-learning-based methods, the Random Forest model (Model C) achieved the best performance albeit limited with an average F1 score of 0.53 and an average AUC of 0.58. This suggests that the clinical data analysed in this study has limited utility in making long-term mortality predictions. The available clinical data was primarily collected to inform recruitment eligibility into the NLST and therefore might not have contained the most relevant prognostic variables. The deep-learning-based methods utilising medical images for classification performed better than non-deep-learning methods. The neural network model (Model D) trained with the 3D CT volume achieved a mean F1 of 0.61 and a mean AUC of 0.73.

IV. DISCUSSION

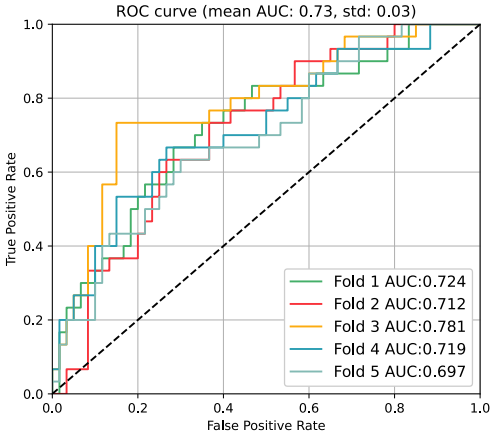
A. LONG TERM MORTALITY PREDICTION

Though the clinical data denoting time information, i.e. PCA 1-4 in Table 1, were ranked as the most important feature in mortality prediction by the tree-based models, their inclusion in Model E did not contribute to an improvement in performance over Model D. This was also true for clinical features encoding prior clinical diagnoses in patients. It is possible that the clinical variable of patient age, or more specifically biological age, might already be embedded in the imaging features, in the morphological appearances of the bones. In fact, as demonstrated recently by Raghu *et al.* [25], the imaging information in chest radiographs can be used to predict the biological age of the patients and be used to predict patients' survival. Therefore, the introduction of chronological age through the clinical data offers limited additional information to the networks.

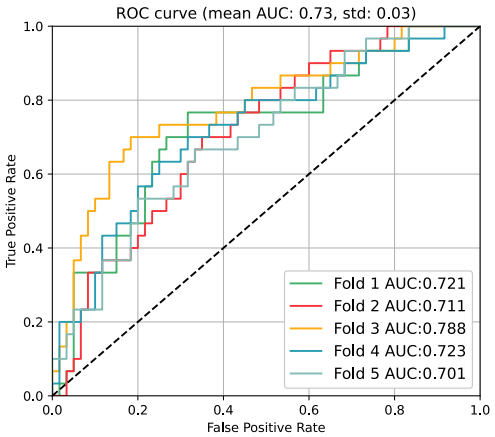
As shown in Table 3, the performance of the current deep learning models for cardiovascular and respiratory mortality prediction, is on par with the cardiovascular mortality prediction models from van Velzen *et al.* 2019 [12] which has an average AUC of 0.72 and standard deviation of 0.07. Other performance metrics, i.e. sensitivity, specificity, and F1 score,

TABLE 3. Performance metrics.

Model	Input	AUC (std)	Precision	Sensitivity	Specificity	F1 score	MCC
(A) Support Vector Machine	clinical data	0.57 (0.05)	0.40	0.66	0.48	0.49	0.13
(B) Gradient Boosting Machine	clinical data	0.58 (0.04)	0.42	0.53	0.62	0.46	0.15
(C) Random Forest	clinical data	0.58 (0.03)	0.38	0.84	0.33	0.53	0.18
(D) 3D ResNet-10	CT scan	0.73 (0.03)	0.54	0.71	0.68	0.61	0.37
(E) 3D ResNet-10	CT scan & clinical data	0.73 (0.03)	0.56	0.63	0.74	0.60	0.38

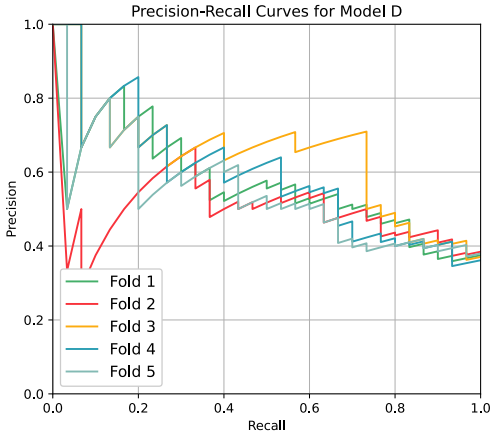


(a) ROC curves of the CT-volume-only network (Model D).

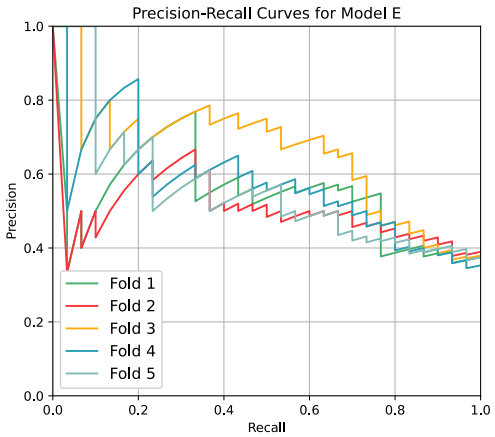


(b) ROC curves of the multimodal network (Model E).

FIGURE 6. Performance comparison (ROC curve).



(a) Precision-recall curves from Model D.



(b) Precision-recall curves from Model E.

FIGURE 7. Performance comparison (Precision-recall curve).

were not available in the literature for comparison. While the benchmark study used an trained convolutional autoencoder to pass the extracted features to a SVM classifier, our models are trained in an end-to-end fashion. More importantly, the main motivation for identifying the respiratory mortality signal and visualising relevant areas on the CT image with saliency maps is to enhance clinical interpretability and confidence in our deep learning model predictions.

More recently, the cardiovascular mortality prediction study from Guo *et al.* [13] where a dual-ResNet was trained on manually selected 2D CT slices achieved a mean AUC of 0.76 and standard deviation of 0.10. When outputs from a SVM trained on handcrafted measures such as the CAC score, were tuned with the output from the dual-ResNet, the model achieved a mean AUC of 0.82. The performance improvement brought about by the handcrafted measures in Guo *et al.* [13], highlights the benefits of combining imaging features with known CVD mortality predictors. Therefore, for future work, we aim to develop an automated approach to derive these markers from the CT scans and use them in the multimodal prediction approach (i.e. Model E). In contrast to models from Guo *et al.* [13], the models in this study adopt an automated approach and thus lessens the burden on radiologists who would otherwise have to select CT slices of interest manually.

B. CONFIRMATION USING SALIENCY MAPS

To evaluate the trustworthiness of the trained models, the saliency maps of the trained models were evaluated on

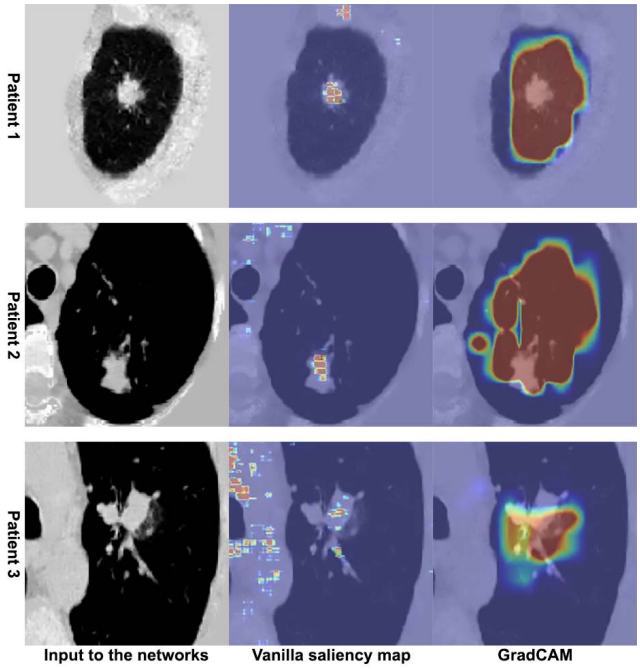


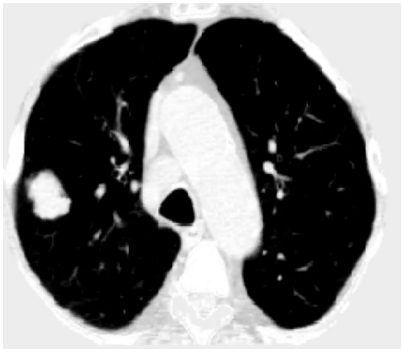
FIGURE 8. Sanity check (in axial view) on the saliency maps in CT volumes with cancerous lesion.

previously held out lung cancer mortality cases using the gradient approach [8] and GradCAM [9]. Given that in these patients it was the lung cancer that accounted for patient mortality, the heatmaps, if performing appropriately would be expected to highlight the malignant nodules in the lung. 3 cases are illustrated in Fig. 8. The original inputs to the network, together with their saliency map overlays, are presented. The heatmaps are colored such that red denotes high activation suggesting the local imaging features contribute more to the network's prediction. From the heatmaps it can be seen that the network can identify the cancerous nodules in the lung and are using these features appropriately when making mortality predictions.

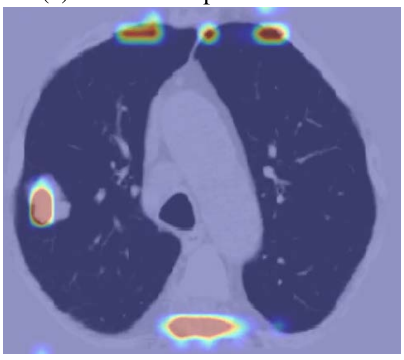
However, the current approach is not without potential limitations. Fig. 9 illustrates the GradCAM approach where the model has highlighted the malignant lesion in the right lung. The model has also highlighted the vertebra and the rib cage, which might intuitively have no direct link to cardiac and pulmonary-related mortality. Yet bone mineral density can act as a good indicator of patient health, particularly in patients with chronic lung diseases such as chronic obstructive pulmonary disease [26], [27]. Furthermore, various studies [28], [29] have demonstrated an association between low bone material density and coronary artery calcification. Accordingly, the signal of low bone material density shown using the GradCam approach may represent a relevant surrogate signal indicating both cardiovascular and respiratory-related mortality.

C. LIMITATION

It is important to know how the models, which are trained on the relatively dated CT scans from the NLST dataset [3] will



(a) Patient 4: input to network.



(b) Patient 4: GradCAM.

FIGURE 9. Saliency map with mixed signal.

perform on more recent lower radiation dose and higher resolution CT scans. It is expected that there will be challenges due to the improvement in CT scan qualities. However, to the best of the authors' knowledge, there is currently no publicly available all-cause mortality dataset with more recent scans. Additionally, the nature of this long-term mortality prediction study dictates the requirement for a sufficiently long follow-up period which some of the more recent screening trials have not yet reached. Thus, further evaluation of the approach will resume when the data from the more recent studies such as NELSON [30] become publicly available.

V. CONCLUSION AND FUTURE WORKS

The results shown in our study lead us to the following conclusions and directions for future studies.

1) The current method shows reasonable performance in predicting long-term mortality in the NLST dataset with an AUC value of 0.73. This illustrates the feasibility of performing mortality predictions from 3D CT scans, without handcrafted features as demonstrated in the literature. Additionally, the average testing time[1] per CT scan is approximately 0.10 second which makes is feasible to use in a clinical setting.
2) The use of saliency maps shows promise as an aid for clinicians' and radiologists' in identifying neglected regions of the CT that might associate with mortality.

[1]Tested on a workstation with Intel i7-10700k 3.80 Ghz CPU, Nvidia RTX3080 GPU (10GB GPU memory), and Ubuntu 18.04.06 LTS OS.

This approach may facilitate the planning of personalised preventative interventions.

3) Given that cardiovascular diseases contribute to a significant portion of the non-lung-cancer mortality in the NLST dataset, it is reasonable to hypothesise that model performance can be improved by providing additional information from the cardiac region. One way to achieve this is by combining the mediastinal-kernel reconstructed cardiac CT volume with the lung-kernel-reconstructed lung volume used in this study. Accordingly, as a next step, we aim to develop a dual branch network for the imaging data. Additionally, we aim to explore an automated approach to calculate handcrafted measures, such as the CAC, to complement the automated image analysis in making mortality predictions.
4) In this study, the chronologically most recent scan (i.e. T2) in the NLST dataset was used to predict long-term mortality. The inclusion of the two earlier annual scans may add value by providing information on the progression of diseased sites in the lung. Such information can help identify patients who have more rapidly progressive diseases who would benefit more from early preventative interventions. To do so, we aim to explore longitudinal lung CT registration using all three screening time point CTs (i.e. T0, T1, and T2).
5) In addition to predicting long-term mortality outcomes, it would be worth exploring the quantification of damage in structures shown to be important on the saliency maps. Analysing quantitative variables in Cox Regression models to predict survival time may be an alternative way of identifying interpretable prognostic imaging biomarkers in LCS participants to facilitate risk stratification and personalised management in high-risk populations.

ACKNOWLEDGMENT

The authors would like to thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

REFERENCES

[1] G. A. Roth, "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017," *Lancet*, vol. 392, no. 10159, pp. 1736–1788, 2018, doi: 10.1016/S0140-6736(18)32203-7.

[2] T. Vos, S. Lim, C. Abbafati, K. Abbas, and M. Abbasi, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020, doi: 10.1016/S0140-6736(20)30925-9.

[3] D. R. Aberle, "The national lung screening trial: Overview and study design," *Radiology*, vol. 258, no. 1, pp. 243–253, Jan. 2011, doi: 10.1148/radiol.10091808.

[4] D. Aberle *et al.*, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England J. Med.*, vol. 365, no. 5, pp. 395–409, 2011, doi: 10.1056/NEJMoa1102873.

[5] D. R. Aberle, S. DeMello, C. D. Berg, W. C. Black, B. Brewer, T. R. Church, K. L. Clingan, F. Duan, R. M. Fagerstrom, I. F. Gareen, C. A. Gatsonis, D. S. Gierada, A. Jain, G. C. Jones, I. Mahon, P. M. Marcus, J. M. Rathmell, and J. Sicks, "Results of the two incidence screenings in the National lung screening trial," *New England J. Med.*, vol. 369, no. 10, pp. 920–931, 2013, doi: 10.1056/NEJMoa1208962.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[7] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," 2019, *arXiv:1904.00625*.

[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/iccv.2017.74.

[10] *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General*, Nat. Center Chronic Disease Prevention Health Promotion, Centers Disease Control Prevention, Atlanta, GA, USA, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK53017/

[11] C. A. Pope, R. T. Burnett, M. C. Turner, A. Cohen, D. Krewski, M. Jerrett, S. M. Gapstur, and M. J. Thun, "Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: Shape of the exposure–response relationships," *Environ. Health Perspect.*, vol. 119, no. 11, pp. 1616–1621, Nov. 2011, doi: 10.1289/ehp.1103639.

[12] S. G. van Velzen, M. Zreik, N. Lessmann, M. A. Viergever, P. A. de Jong, H. M. Verkooijen, and I. Išgum, "Direct prediction of cardiovascular mortality from low-dose chest CT using deep learning," *Proc. SPIE Med. Imag. Image Process.*, vol. 10949, Oct. 2019, Art. no. 109490X, doi: 10.1117/12.2512400.

[13] H. Guo, U. Kruger, G. Wang, M. K. Kalra, and P. Yan, "Knowledge-based analysis for mortality prediction from CT images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 457–464, Feb. 2020, doi: 10.1109/JBHI.2019.2946066.

[14] P. C. Jacobs, M. J. A. Gondrie, Y. van der Graaf, H. J. de Koning, I. Isgum, B. van Ginneken, and W. P. T. M. Mali, "Coronary artery calcium can predict all-cause mortality and cardiovascular events on low-dose CT screening for lung cancer," *Amer. J. Roentgenol.*, vol. 198, no. 3, pp. 505–511, Mar. 2012, doi: 10.2214/AJR.10.5577.

[15] C. Chiles, F. Duan, G. W. Gladish, J. G. Ravenel, S. G. Baginski, B. S. Snyder, S. DeMello, S. S. Desjardins, R. F. Munden, and N. S. Team, "Association of coronary artery calcification and mortality in the national lung screening trial: A comparison of three scoring methods," *Radiology*, vol. 276, no. 1, pp. 82–90, 2015.

[16] S. G. M. van Velzen, N. Lessmann, B. K. Velthuis, I. E. M. Bank, D. H. J. G. van den Bongard, T. Leiner, P. A. de Jong, W. B. Veldhuis, A. Correa, J. G. Terry, J. J. Carr, M. A. Viergever, H. M. Verkooijen, and I. Iágum, "Deep learning for automatic calcium scoring in CT: Validation using multiple cardiac CT and chest CT protocols," *Radiology*, vol. 295, no. 1, pp. 66–79, Apr. 2020, doi: 10.1148/radiol.2020191621.

[17] M. T. Lu, A. Ivanov, T. Mayrhofer, A. Hosny, H. J. W. L. Aerts, and U. Hoffmann, "Deep learning to assess long-term mortality from chest radiographs," *JAMA Netw. Open*, vol. 2, no. 7, Jul. 2019, Art. no. e197416, doi: 10.1001/jamanetworkopen.2019.7416.

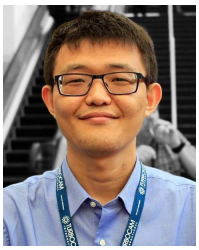
[18] S. Gheisari, S. Shariflou, J. Phu, P. J. Kennedy, A. Agar, M. Kalloniatis, and S. M. Golzan, "A combined convolutional and recurrent neural network for enhanced glaucoma detection," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Dec. 2021, doi: 10.1038/s41598-021-81554-4.

[19] R. Gao, Y. Tang, K. Xu, Y. Huo, S. Bao, S. L. Antic, E. S. Epstein, S. Deppen, A. B. Paulson, K. L. Sandler, P. P. Massion, and B. A. Landman, "Time-distanced gates in long short-term memory networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101785, doi: 10.1016/j.media.2020.101785.

[20] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-OR network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3484–3495, Nov. 2019, doi: 10.1109/TNNLS.2019.2892409.

[21] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2018, pp. 9525–9536.
[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.
[25] V. K. Raghu, J. Weiss, U. Hoffmann, H. J. W. L. Aerts, and M. T. Lu, "Deep learning to estimate biological age from chest radiographs," *JACC, Cardiovascular Imag.*, vol. 14, no. 11, pp. 2226–2236, Nov. 2021, doi: 10.1016/j.jcmg.2021.01.008.
[26] J. Jaramillo, C. Wilson, and D. Stinson, "Reduced bone density and vertebral fractures in smokers. Men and COPD patients at increased risk," *Ann. Amer. Thoracic Soc.*, vol. 12, no. 5, pp. 648–656, May 2015, doi: 10.1513/AnnalsATS.201412-591OC.
[27] N. Campos-Obando, M. C. Castano-Betancourt, L. Oei, O. H. Franco, B. H. C. Stricker, G. G. Brusselle, L. Lahousse, A. Hofman, H. Tiemeier, F. Rivadeneira, A. G. Uitterlinden, and M. C. Zillikens, "Bone mineral density and chronic lung disease mortality: The Rotterdam study," *J. Clin. Endocrinol. Metabolism*, vol. 99, no. 5, pp. 1834–1842, May 2014, doi: 10.1210/jc.2013-3819.
[28] N. Ahmadi, S. Mao, F. Hajsadeghi, B. Arnold, S. Kiramijyan, Y. Gao, F. Flores, S. Azen, and M. Budoff, "The relation of low levels of bone mineral density with coronary artery calcium and mortality," *Osteoporosis Int.*, vol. 29, no. 7, pp. 1609–1616, 2018, doi: 10.1007/s00198-018-4524-7.
[29] P. A. Marcovitz, H. H. Tran, B. A. Franklin, W. W. O'Neill, M. Yerkey, J. Boura, M. Kleerekoper, and C. Z. Dickinson, "Usefulness of bone mineral density to predict significant coronary artery disease," *Amer. J. Cardiol.*, vol. 96, no. 8, pp. 1059–1063, Oct. 2005, doi: 10.1016/j.amjcard.2005.06.034.
[30] C. A. van Iersel, H. J. de Koning, G. Draisma, W. P. T. M. Mali, E. T. Scholten, K. Nackaerts, M. Prokop, J. D. F. Habbema, M. Oudkerk, and R. J. van Klaveren, "Risk-based selection from the general population in a screening trial: Selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON)," *Int. J. Cancer*, vol. 120, no. 4, pp. 868–874, Feb. 2007, doi: 10.1002/ijc.22134.



YAOZHI LU received the B.Eng. degree in mechanical engineering from the University College London, in 2014, and the M.Sc. and Ph.D. degrees in mechanical engineering from Imperial College London, in 2015 and 2020, respectively. He did his Ph.D. research on the alternative passage divergence phenomenon in commercial jet engines with the Rolls-Royce Vibration University Technology Centre, Imperial College London. He became a Chartered Engineer (CEng), in 2021. He is currently a Research Fellow at the Centre for Medical Image Computing, University College London. His research interests include medical image computing, deep learning, computer vision, and aeroelasticity.

SHAHAB ASLANI received the M.Sc. degree in electrical and electronic engineering (medical imaging curriculum) from Dokuz Eylul University (DEU), in 2015, and the Ph.D. degree in electrical and electronic engineering (pattern analysis and computer vision curriculum) from the Italian Institute of Technology (IIT-PAVIS), in 2019. He is currently a Research Fellow at the Centre for Medical Image Computing (CMIC), University College London (UCL), working on early lung cancer prediction. His main research interests include medical image analysis, medical imaging computing, machine learning, and deep learning.

MARK EMBERTON was appointed as the Dean of the UCL Faculty of Medical Sciences, in 2015. He is currently a Professor of interventional oncology at UCL. He is also an Honorary Consultant Urologist at the University College Hospitals NHS Foundation Trust and the Founding Pioneer of The Charity Prostate Cancer U.K. He specializes in the implementation of new imaging techniques, nanotechnologies, bio-engineering materials, and non-invasive treatment approaches, such as high intensity focused ultrasound and photo-dynamic therapy. He is also involved in teaching within UCL and the London and South East Urological Training Scheme. His research interests include improving the diagnostic and risk stratification tools and treatment strategies for prostate cancer (PCa). He is a founding partner of London Urology Associates. In addition to being a member of various urological and medical organizations, such as the American Association of GenitoUrinary Surgeons, the British Association of Urological Surgeons, and the European Association of Urology.

DANIEL C. ALEXANDER received the B.Sc. degree in mathematics from the University of Oxford, in 1993, and the M.Sc. and Ph.D. degrees in computer science from the University College London (UCL), in 1994 and 1998, respectively. He worked as a Postdoctoral Researcher with the University of Pennsylvania, until 2000, when he returned to London to take up an academic position. He became a Full Professor, in 2010, and the Director of CMIC, in 2015. He is currently the Director of the UCL Centre for Medical Image Computing (CMIC), UCL, and a Professor of imaging science with the Department of Computer Science, UCL. His expertise is in computational modeling, machine learning, imaging, and image analysis. He is currently an Associate Editor of *Magnetic Resonance in Medicine* and a Senior Fellow of the International Society for Magnetic Resonance in Medicine.

JOSEPH JACOB received the M.D. degree (research) from Imperial College under Prof. David Hansell with Royal Brompton Hospital, in 2016. He qualified in medicine from Imperial College, worked at Médecins Sans Frontières for two years, and completed radiology training with Kings College Hospital, London, U.K. His research with the Centre for Medical Image Computing, University College London, centers on computational image analysis of the lungs and heart on CT imaging. He has coauthored over 80 articles, won national and international awards for his work. He was awarded the 2017 Best Thesis Prize by the National Heart and Lung Institute. He was awarded a five-year Wellcome Trust Clinical Research Career Development Fellowship for his M.D. degree, in 2018.