Automatic Generation of Interpretable Lung Cancer Scoring Models from Chest X-Ray Images

Michael James Horry
Center for Advanced modelling and Geospatial Information Systems
School of Information, Systems & Modelling, Faculty of Engineering and IT, University of Technology Sydney
Sydney, NSW 2007, Australia
mhorry@au1.ibm.com

Subrata Chakraborty
Center for Advanced modelling and Geospatial Information Systems
School of Information, Systems & Modelling, Faculty of Engineering and IT, University of Technology Sydney
Sydney, NSW 2007, Australia
subrata.chakraborty@uts.edu.au

Biswajeet Pradhan
Center for Advanced modelling and Geospatial Information Systems
School of Information, Systems & Modelling, Faculty of Engineering and IT, University of Technology Sydney
Sydney, NSW 2007, Australia
biswajeet.pradhan@uts.edu.au

Manoranjan Paul
Machine Vision and Digital Health (MaViDH)
School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW 2795, Australia
mpaul@csu.edu.au

Douglas P. S. Gomes
Machine Vision and Digital Health (MaViDH)
School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW 2795, Australia
dgomes@csu.edu.au

Anwaar Ul-Haq
Machine Vision and Digital Health (MaViDH)
School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW 2795, Australia
aulhaq@csu.edu.au

***Abstract*— Lung cancer is the leading cause of cancer death and morbidity worldwide with early detection being the key to a positive patient prognosis. Although a multitude of studies have demonstrated that machine learning, and particularly deep learning, techniques are effective at automatically diagnosing lung cancer, these techniques have yet to be clinically approved and accepted/adopted by the medical community. Most research in this field is focused on the narrow task of nodule detection to provide an artificial radiological 'second reading'. We instead focus on extracting, from chest X-ray images, a wider range of pathologies associated with lung cancer using a computer vision model trained on a large dataset. We then find the set of best fit decision trees against an independent, smaller dataset for which lung cancer malignancy metadata is provided. For this small inferencing dataset, our best model achieves sensitivity and specificity of 85% and 75% respectively with a positive predictive value of 85% which is comparable to the performance of human radiologists. Furthermore, the decision trees created by this method may be considered as a starting point for refinement by medical experts into clinically usable multi-variate lung cancer scoring and diagnostic models.**

Keywords—Artificial Intelligence, Machine Learning, Computer Vision, Lung Cancer, Malignancy Model, Explainable AI, Automatic Model Generation

I. Introduction

Lung cancer is the leading cause of cancer death worldwide [1] with over 2 million new cases in 2018 and rising. There is a long history of research into automated diagnosis of lung cancer from medical images using computer vision techniques encompassing linear and non-linear filtering [2], grey-level thresholding analysis [3], and, more recently, machine learning including deep learning techniques [4-7]. However, despite the many lab-based successes of computer vision medical image diagnostic algorithms, the actual approval and clinical adoption of these computer vision techniques in medical image analysis is very limited. As of September 2020 the U.S. Food and Drug Administration (FDA) has approved only 30 radiology related deep learning or machine learning based applications/devices of which only three utilize the X-ray imaging mode [8] with the subject of one being wrist fracture diagnosis (FDA DEN180005) and the other two being for pneumothorax assessment (FDA K190362 and K183182).

In contrast to the limited number of field applications relating to clinical use of machine learning, there exists a massive corpus of published research in this field. The Scopus database [9] returns over 700 results from 1988 to the present for a title and abstract search on ("Computer Vision" OR "Machine Learning" OR "Deep Learning" AND chest AND X-ray). The overwhelming majority of these papers have been authored in the past decade as shown in Figure 1. There is, however, a significant gap between this massive research and development effort and the lack of approved clinical applications for the automated diagnosis of lung cancer from CXR.

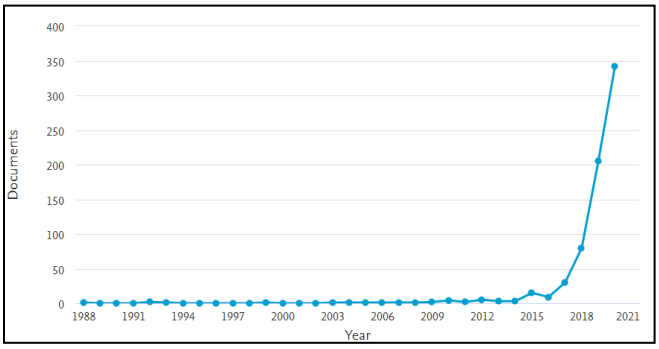

Fig. 1. Scopus bibliographic histogram relating to Machine Learning Chest X-Ray.

Driving this huge interest in medical computer vision research is a desire to provide tools to improve the productivity of medical clinicians. In the case of radiology, the objective of most research efforts in the field of automated medical imaging interpretation is to provide an automated "second reading" to assist radiologists with their workloads. This ambitious goal has arguably been met under lab conditions over the course of many studies [10]. However, the lack of any clinically approved multi-class Chest X-ray (CXR) based computer vision diagnostic tool evidences the difficulty of translating success in the lab to engineering that is useful outside lab-controlled conditions.

This paper takes a different approach to the application of machine learning to automated lung cancer diagnosis and stratification. Rather than aiming to provide an automated second reading, our objective is to autonomously create a range of reasonable, explainable decision tree models for lung cancer malignancy scoring. Our intention is that that these models can be used by the medical community as a data driven foundation for multivariate diagnostic scoring of lung cancer.

We extend the typical method of training a deep learning algorithm (usually a variant of the Convolutional Neural Network architecture) in lung nodule detection and classification into a two-step approach that combines deep learning multiple-feature extraction with automatic fitting of decision trees.

Firstly, we investigate lung pathology features that are closely associated with lung cancer then use these conditions to train a multi-class deep learning algorithm. Secondly, the score for each condition (score tuple) is inferenced from the trained model against an independent lung cancer CXR dataset for which malignancy scoring metadata is available. Finally, the score tuple is fitted to the malignancy data for each patient using a simple decision tree with the most accurate decision tree/s extracted as a base for future clinical/empirical study by the medical community.

This more holistic approach, which emphasises the importance of multiple features along with interpretability and human judgement in the creation of medical computer vision applications, may help overcome the hurdles that have held back the acceptance and widespread use of medical artificial intelligence algorithms in the field.

II. Related Work

A. Recent Work on Automated Lung Cancer Diagnosis

Although the Computed Tomography (CT) imaging mode has attracted most of the machine learning lung cancer diagnostic research, there are a number of studies that show the usefulness of CXR for this task. Best results against the very large Chest-Xray14 dataset [11] was achieved using a hybrid approach of deep learning combined with local feature extraction achieving an average AUC of 0.8097 for thirteen conditions (the "No Finding" class being excluded) [12] representing a significant improvement over the results of the Chest-Xray14 authors. Other recently successful approaches include incorporation of label dependencies via LSTM modules [13] and consideration of the relationship between pathology and location in the lung geometry [14]. Good results in lung nodule detection using deep learning have been achieved by teams using other datasets with a systematic survey for this research up to 2018 being provided by [15]

These studies tend to focus on lung nodules only, which is problematic for two reasons; firstly, a lung nodule is defined as measuring ≤ 3 cm in diameter [16] with larger nodules or masses typically ignored in these studies even though these may indicate more serious and likely malignant cancers, and secondly, most pulmonary nodules are benign [17]. These two factors combined would logically lead to existing deep learning systems tending to over diagnose benign conditions as lung cancer (where the study equates lung cancer to the presence of detected nodules) and under diagnose serious nodules and masses over 3 cm in diameter. Both consequences are obviously highly undesirable.

B. Recent Work on Automated Diagnostic Scoring from Medical Images

Interest in severity scoring from CXR images has received some recent focus due to the COVID-19 pandemic of 2020. A combination of deep learning feature extraction and logistic regression fit to severity has been shown to be predictive of the likelihood of ICU admission for COVID-19 patients [18]. A large number of papers on classification of lung nodules detected from the CT imaging mode have been published employing various deep learning techniques [19], however few such papers have been published for the CXR imaging mode. This is most likely due to the CXR imaging mode having lower sensitivity in comparison to the CT imaging mode [20], making nodule characterisation difficult using traditional segmentation and shape analysis techniques.

III. Data Sets

A. Data Sourcing

In order to achieve our objectives two datasets are needed. The first is a large corpus of labelled CXR data that can be used to train a deep learning classifier as a feature extraction component. The National Institute of Health (NIH) ChestX-ray14 dataset [21] provides over 100,000 such images being uniformly 1024 x 1024 pixels in a portrait orientation with both Posterior-Anterior (PA) and AP views. The second necessary dataset must comprise CXRs with malignancy metadata indicating whether lung cancer is present in the image and if so, whether the cancer is considered by expert radioloists to be benign or malignant. The Lung Image Database Consortium Image Collection (LIDC-IDRI) [22] CXR subset meets these requirements and additionally provides metadata desribing whether a malignant cancer originates in the lung or elsewhere. The LIDC-IDRI dataset has been manually labelled by four radiologists with access to corresponding patient CT scans. The label metadata has been provided at the patient level, meaning that there are some images provided where the nodule location is known and logged from the CT scan but not visible on the CXR image.

Normally, any such inconsistency between the dataset and it's labels would be problematic for a computer vision diagnosis since the image data would not support the binary label ground truth. Our more holistic classification should be relatively robust to this problem since the presence of visible nodules is only one of a number of features under consideration.

Although we could find no indication of the projection used for the LIDC-IDRI images, they appear to be posterior-anterior as indicated by ribs clearly being in front of the spinal column along with the unclear scapula which is obfuscated by the lung field.

B. Data Curation

The NIH data set contains 112,120 CXR images from 30,805 unique patients with a mix of both posteroanterior (PA) and anteroposterior (AP) projections. The dataset has been labelled using natural language processing to extract disease classes for each image from the associated radiology report, which the NIH dataset authors claim are of greater than 90% accuracy. Many of the images have a mix of disease classes. Since our objective is to achieve explainable lung cancer scores, we have restricted this study to images labelled with only a single disease class.

Not all of the disease classes included in the NIH data set are indicative of lung cancer. In order to either exclude or include the classes the simple rule was applied. If the literature noted a general indicative connection between lung cancer and the class in question then that class was to be extracted from the NIH set for further analysis. The only exception to this inclusion rule is the "No Finding" class was included to enrich the generated models with a contra indicator. This resulted in five classes of interest for this study being Atelectasis, Effusion, Mass, No Finding, and Nodule. Once filtered in this way the totals for images in this dataset are as included in Table I.

TABLE I. COUNT OF NIH IMAGES FOR SINGLE CLASSES

NIH Image Data (Single Class) Summary			
Classification	***Count***	***Extracted***	***Associated with Lung Cancer***
Atelectasis	2210	Y	Has been documented as a first sign of lung cancer [23].
Cardiomegaly	746	N	Not related to lung cancer although in rare cases misdiagnosed when underlying condition is mass in same geography of CXR [24].
Consolidation	346	N	Can sometimes accompany lung cancer but usually associated with pneumonia [25].
Edema	51	N	Can be a complication from treatment for lung cancer but does not indicate lung cancer [26].
Effusion	2086	Y	Can be caused by a build up of cancer cells and a common complication of lung cancer [27].
Emphysema	525	N	Has been linked as a risk factor for lung cancer but not an indication [28].
Fibrosis	648	N	Has been linked as a risk factor for lung cancer but not an indication [29].
Hernia	98	N	Has been mistaken for lung cancer but does not indicte lung cancer [30].
Infiltration	5270	N	Generic descriptor used informally in radiological reports and not actually an accepted lung disease classification.
Mass	1367	Y	A primary indication of lung cancer [26].
No Finding	39302	Y	By definition not lung cancer but included to enrich generated models with a counter-indicator.
Nodule	1924	Y	A primary indication of lung cancer [26, 31] with about 40% of nodules being cancerous.
Pleural_Thickening	875	N	This is often an indication of mesothelioma caused by exposure to asbestos. It is also a very common abnormal finding on CXR. It is not an indication of lung cancer [32].
Pneumonia	176	N	Often a compication of lung cancer [25] with 50-70% of patients developing a lung infection. Persistent pneumonia can lead to a diagnosis of lung cancer. Not typically used as indicator of lung cancer.
Pneumothorax	1506	N	Can be the first sign of lung cancer but this is rare [33].

To address class imbalance during training, the classes were under-sampled to 2000 examples of each class. This left the "Mass" and "Nodule" labels as minority classes with 1367 and 1924 samples respectively. This remaining imbalance was addressed in training by means of a weighted random sampler in the data loader.

Standard augmentations were applied only to the training NIH dataset with random rotation of 1 degree with expansion, and random horizontal flip. Vertical flipping was not used since CXR images are not vertically symmetrical. Training and testing were run with and without equalization. The images were then resized with a default size of 244 x 244 pixels.

The NIH dataset was split into an 80:20 training and validation pair resulting in 6641 images for training and 1661 images for validation. A set of 6085 images (conforming to the NIH recommended test set) was used as a holdout test set. Since these images were drawn from the recommended NIH test split there was no patient overlap between the data used for training and testing.

No curation or processing other than resizing and equalization matching the training set was applied to the LIDC-IDR dataset.

IV. MODEL DEVELOPMENT

A. Network Selection

Following experimentation with a number of classifiers including VGG-19 [34], AlexNet [35] , DenseNet-121 [36], ResNet-50 [37] and ResNext-50 [38], we found that DenseNet-121 network initialized with ImageNet [39] weights consistently gave the best results. In contrast to the ResNet variants, DenseNet did not seemingly benefit from staged training [40] thereby allowing for a simpler experimental setup. We therefore selected the DenseNet121 architecture for the study which is consistent with other studies relating to the use of deep learning classifiers on large CXR datasets [41-43].

We followed standard practice employed in transfer learning [44] and replaced the output fully connected layer (by default 1000 neurons) with the number of classification outputs required by the experiment being five. These five output nodes matched our five selected features being Atelectasis, Effusion, Mass, No Finding, and Nodule. After experimenting with trainable output head layers vs simply fine-tuning the entire model we found that the entire model fine-tuning approach worked best for DenseNet121 with AUC-ROC results comparable to state-of-the-art for this dataset in consideration that we have restricted classes and under-sampled (Table III).

The Adam optimizer [45] was used along with a cosine annealing learning rate scheduler with learning rate of 0.001. This scheduler was selected because during model testing and hyperparameter optimization, it was noticed that the model trained well with a more aggressive learning rate, leading to higher validation accuracy at a lower number of epochs.

Finally, we chose to test with both the standard binary cross entropy and focal loss [46] functions since inspection of the NIH dataset revealed that many images were objectively low quality (due to overexposure, improper patient position, presence of intrusive medical devices and differing focal distance) resulting in a number of sample images that could be considered to be adversarial to training. We hypothesised that a focal loss function might allow the model to better account for the adversarial images thereby allowing good results without curating these images out of the dataset manually at the risk of introducing sampling bias. The tests performed are summarised in table II below.

TABLE II. TEST IDENTIFICATION AND SUMMARY

Test ID	*Loss Function*	*Histogram Equalization*
A	BCE	TRUE
B	BCE	FALSE
C	FOCAL	TRUE
D	FOCAL	FALSE

The results of 10 training/holdout testing runs are shown in Figures 2 – 5 below.

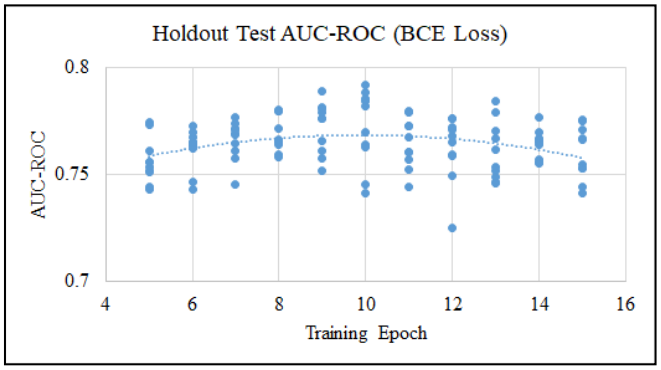

Fig. 2. Average AUC-ROC for 10 Round of Holdout Testing (Test A)

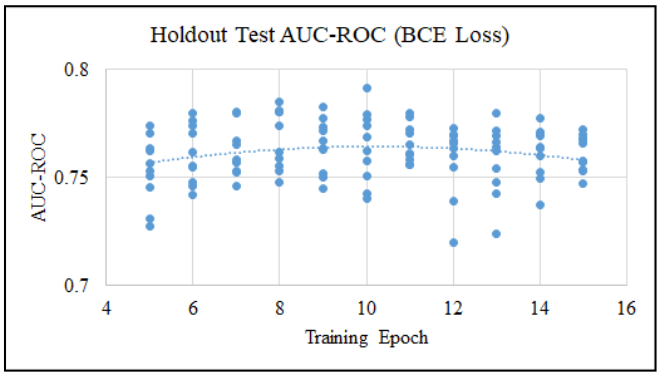

Fig. 3. Average AUC-ROC for 10 Round of Holdout Testing (Test B)

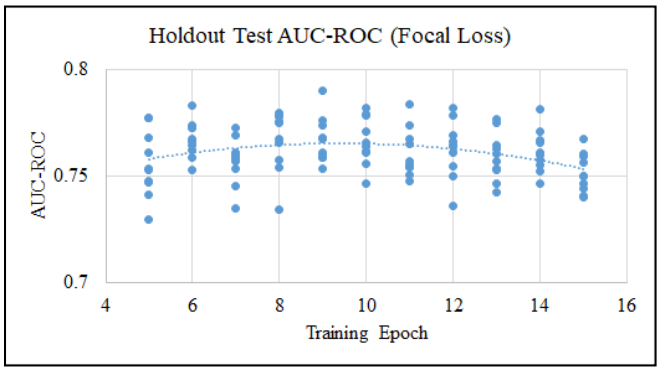

Fig. 4. Average AUC-ROC for 10 Round of Holdout Testing (Test C)

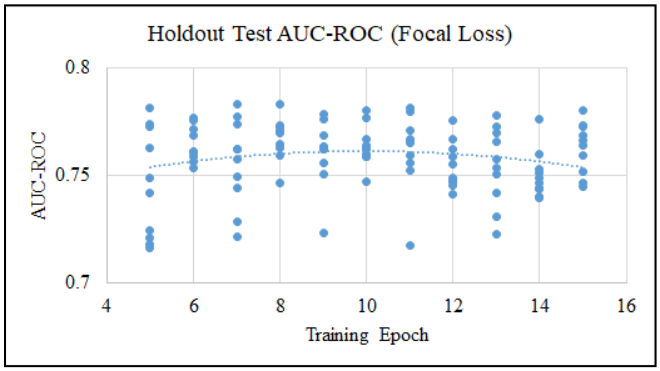

Fig. 5. Average AUC-ROC for 10 Round of Holdout Testing (Test D)

B. *Model Selection*

Inspection of the average AUC-ROC curves for the tested combinations of equalization and loss functions (figures 2-5 above) led us to hypothesise that the equalized BCE model corresponding to test A would be preferred for LIDC dataset inferencing and decision tree generation. This configuration, as shown in figure 2, was the most consistent performer with the majority of average AUC-ROC scores between 0.75 and 0.8 with only a single outlier returning an average AUC-ROC value below 0.73. Best AUC-ROC values for the extracted features for each tested configuration are shown in table III below with highest score for each feature in bold. Also included as a baseline are the AUC-ROC values for the same conditions from the original NIH paper [11] from their best reported classifier (ResNet50) and the most relevant state-of-the-art results from [47], noting that these papers also considered additional conditions in their multi-classifier and did not restrict the CXR images to PA projections meaning that results are not directly comparable.

We interpret our slightly improved scores compared to [11] and [47] as a result of filtering the dataset for our purposes as described above, thereby reducing the label noise in the dataset which has been estimated by some commentators as up to 10% [48]. We do note that other teams have achieved even better results than [47] by applying additional data sets [14] and hand-crafted shallow feature integration [12], however at best these results are on average only 1.4% improved upon ours. Our approach is that pure deep learning is enough to meet the objectives of this study with the significant benefits of simplicity, efficiency and adaptability to a federated model in the future.

From the model training AUC-ROC curves we hypothesised that the most accurate models for the inference stage of the experiment would be in the range of 9-11 epochs, corresponding to the peaks of the holdout testing polynomial fit curve.

TABLE III. AUC-ROC SCORES FOR NIH SUBSET

NIH Image Data (Single Class) Summary				
Configuration	***Atelectasis***	***Effusion***	***Mass***	***Nodule***
Test A (Epoch 8)	**0.779**	0.843	**0.821**	0.725
Test B (Epoch 8)	0.767	0.858	0.795	0.714
Test C (Epoch 9)	0.754	0.849	0.817	**0.739**
Test D (Epoch 10)	0.753	**0.869**	0.810	0.733
Wang *et al.*[11]	0.700	0.759	0.693	0.669
Yao *et al.*[47]	0.733	0.806	0.778	0.272

V. MALIGNANCY MODEL GENERATION

A. Method

157 CXR images from the LIDC dataset that included patient level diagnosis metadata were extracted from DICOM format into PNG format to match the classifier training data format. Of these, images 27 were classified with a diagnosis of "Unknown" and were excluded from further analysis. The remaining 130 records were categorised by the LIDC as follows:

TABLE IV. LIDC PATIENT LEVEL DIAGNOSIS METADATA SUMMARY

LIDC Image Diagnosis Summary		
Diagnosis	***Description***	***Number of Images***
1	benign or non-malignant disease	36
2	malignant, primary lung cancer	43
3	malignant metastatic	51

The DenseNet121 models were used to extract pathological feature scores for the LIDC images. Inferencing was performed from saved models of all training epochs and runs resulting in 150 models for each combination of equalization and loss function. A seven-column csv template was prepared containing columns for the Patient ID, placeholders for the five features of interest (including the "No Finding" class), and the diagnosis score 1 to 3 as determined by four experienced thoracic radiologists [49]. Diagnosis scores 2 and 3 were combined into a single malignancy class with 94 images. representing malignant diagnosis and thereby allowing for a binary separation. Values for "Atelectasis", "Effusion", "Mass", "No Finding" and "Nodule" were inferenced from the Densenet121 models as a score tuple and written to the placeholder columns to complete a data-frame of patients, inferenced feature scores and diagnosis label. The inference reference files are available as supplementary information.

The data-frame was then randomly split into an 80:20 Training/Testing set (resulting in 26 records reserved for holdout testing) before being used to fit a decision tree classifier with a limited maximum depth of 3 (in order to reduce overfitting due to the small sample size) fitting on an entropy criteria. The fit accuracy was captured and written to a CSV file, and the tree view generated was captured as an image file for any model with greater than 60% accuracy for further investigation of the associated confusion matrix and tree as a potentially useful diagnostic model. This process is summarised in Figure 6.

An experiment was scripted for all combinations of training epoch/round, loss function (BCE and Focal) and histogram equalization usage.

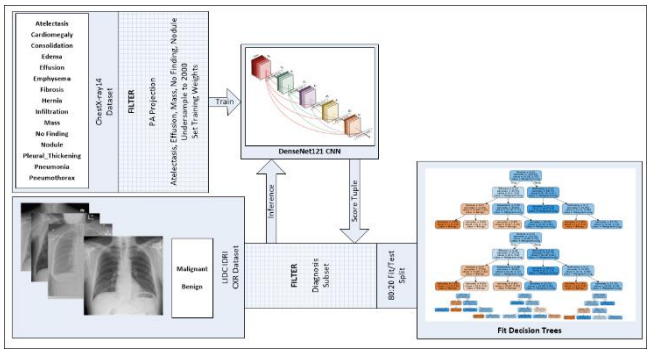

Fig. 6. Automatic Model Generation Process.

VI. RESULTS

The majority of fitted decision trees achieved accuracy of 60% or greater with the combination of focal loss without equalization proving to be the most consistent with 92% of trees meeting this criterion. The most accurate decision tree generated using the method described achieved 85% accuracy with positive predictive value of 83% for test C using a combination of focal loss function and histogram equalization. All tests achieved best accuracy greater than 81%, with sensitivity in the range 85-100% and specificity in the range 29-75% which (accounting for sensitivity/specificity trade-off) is consistent with studies showing human radiologist performance in detecting symptomatic lung cancer from CXR to have sensitivity of 54-84% and specificity of 90% [50]. Our results also confirmed our hypothesis relating to the utility of the focal loss function for the NIH data set on an accuracy/consistency measure. A summary of the experiment is below in table V with the set of test metrics shown in table VI.

TABLE V. SUMMARY OF AUTOMATICALLY GENERATED DECISION TREE MODELS

Test ID	***Loss Func***	***Eq***	***Trees having Accuracy ≥ 60% (%)***	***Best Accuracy (%)***	***Best Epoch***
A	BCE	TRUE	88	81	12
B	BCE	FALSE	91	81	6
C	FOCAL	TRUE	91	85	11
D	FOCAL	FALSE	92	81	12/14

TABLE VI. SUMMARY OF TEST METRICS

Test ID	***Acc (%)***	***Sensitivity***	***Specificity***	***Precision (PPV)***	***FP Rate (%)***	***F1 (%)***
A	81	1.0000	0.286	0.792	71	88
B	81	0.8462	0.750	0.846	25	85
C	85	1.0000	0.429	0.826	57	91
D	81	0.9231	0.6259	0.8000	38	86

Before interpreting the candidate decision tree models for discussion, we investigated the confusion matrices for the highest accuracy results for each experiment to understand the predictive value of each model.

A. *Model Analysis*

The confusion matrices for the highest accuracy result/s of each test shown in figures 7-10 below:

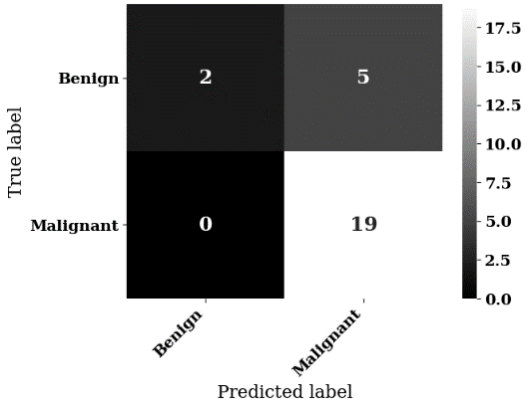

Fig. 7. Confusion Matrix for Test A

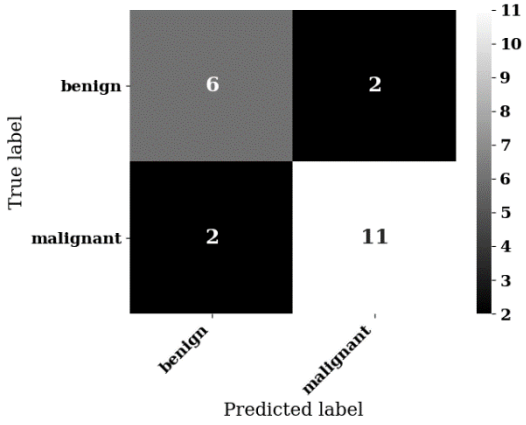

Fig. 8. Confusion Matrix for Test B.

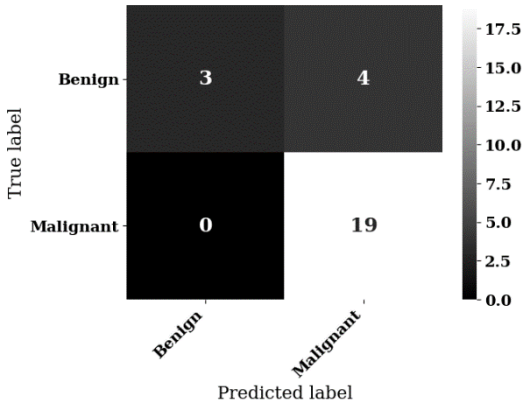

Fig. 9. Confusion Matrix for Test C

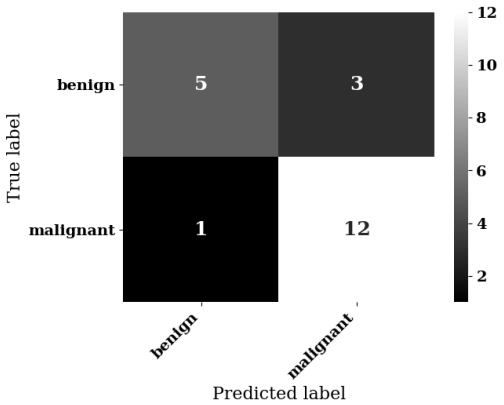

Fig. 10. Confusion Matrix and Classification Summary for Test D

Inspection of the confusion matrices for these best tests in figures 7-10 above show that the models tended to overclassify as "Malignant" resulting in a very high recall for this class at the expense of a high false positives rate for the "Benign" test samples. This can be interpreted as resulting from the small number of benign samples (36) available in the LIDC dataset in comparison to the malignant samples (94). Note that in the medical context false positives are preferable to false negatives since requisite follow-up radiology such as CT scans combined with other clinical indicators will achieve more accurate diagnosis [51] and eliminate the false positive. On the other hand, a false negative result can lead to a missed diagnosis and inaction, which is particularly problematic in the case of lung cancer where early detection has been shown to significantly improve outcomes [52].

When we filter these results for lowest false positive rate then the results from test B clearly represents the best balance between accuracy, sensitivity and specificity with scores of 81%, 85% and 75% respectively with test D results also reasonable. Tests A and C achieved high accuracy by means of sensitivity to the malignant class at the expense of specificity for the benign class, resulting in a high false positive rate. These tests (A and C) correspond to the two tests utilizing histogram equalization and we suspect that the equalization process may reinforced benign features in the CXR image which were then mistaken for nodules or masses by the classifier, thereby reducing the separability of the malignant and benign classes.

We consider test B to have produced the best results with accuracy, sensitivity and specificity consistent with the medical literature for human radiological performance [53] and showing that this model is good at separating malignant samples under the constraints of the small dataset.

B. *Generated Decision Tree Interpretation*

Example decision trees corresponding to these results are shown in figures 11 to 14 below. These decision tree models explain the scores achieved by each test using a decision path and serve to illustrate the result of the end-to-end technique presented in this paper. Due to the small sample size of the LIDC dataset used for inference, it is not possible to claim that these decision tree models are clinically viable. However, even on this small dataset the results achieved are reasonable and could be expected to be greatly improved with additional inferencing samples. For example, the tree model for Test A (fig 11) indicates (at the third level of the tree) that a high value for "Mass" with a value for "Effusion" between 0 and 28% is associated with Malignancy whereas a low value for "Mass" along with a high value for "No Finding" is associated with a "Benign" classification.

Recalling that Test B yielded our best results, we expected the associated tree model to provide the most meaningful insights into the relationship between our selected pathological features and lung cancer malignancy. This proved to be the case, with most paths matching our understanding of the lung cancer condition. For example, a high value of "Atelectasis" coupled with a high value for "Mass" yields a malignant diagnosis. A low value for "Atelectasis" along with a high value for "Effusion" also leads to a malignant diagnosis. Low scores for "Atelectasis" and "Effusion" lead to a benign diagnosis as well as low a low score for "Mass" along with a high score for "Effusion", which we interpret as non-cancer related effusion caused by conditions such as Pneumonia.

The decision tree generated for Test C associates a high score for "Nodule" with Malignancy but is otherwise counter-intuitive with respect to the "No Finding" and "Mass" classes. One interpretation of this may be that the histogram equalization process applied to this test may have reduced the separability of the "Mass" and "No Finding" features by boosting non-malignant features of the "No Finding" class.

Finally, the decision tree generated for Test D associates a high score for "No Finding" with a benign diagnosis unless a high score for "Effusion" and a low score for "Atelectasis" are present. Conversely a low score for "No Finding" along with a high score for "Atelectasis" is associated with malignancy unless the "Atelectasis" score is moderate (between 0.2 and 0.24) indicating a benign condition.

Interestingly, both tests B and D associated "Effusion" with malignancy which could indicate that these models were sensitive enough to automatically detect the build-up of fluid and cancer cells between the chest wall and lung associated with malignant lung cancer known as Malignant Pleural Effusion [54].

In general, we found the higher levels of the generated decision trees to correspond to our understanding of the lung cancer condition, with the lower levels of the tree being less consistent and sometimes counter intuitive. We interpret this as a result of our small inferencing dataset not providing enough samples for the lower levels of the tree to reliably fit.

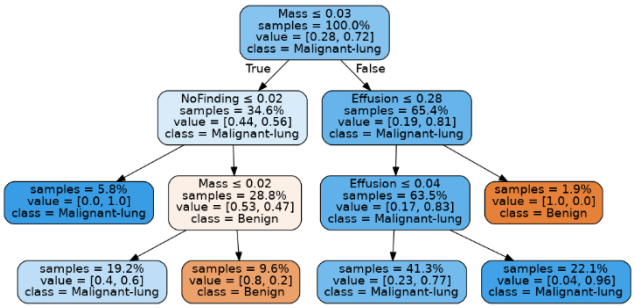

Fig. 11. Automatically Generated Decision Tree for Test A

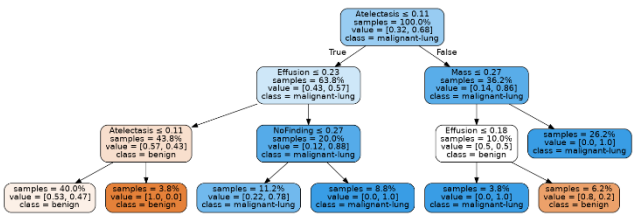

Fig. 12. Automatically Generated Decision Tree for Test B

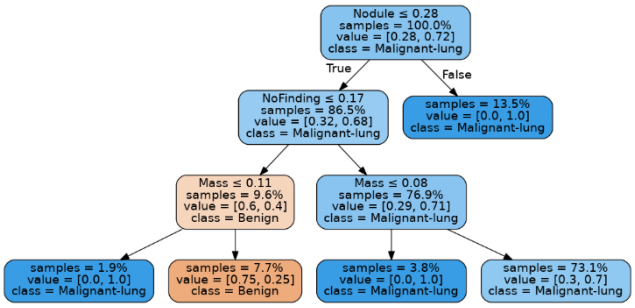

Fig. 13. Automatically Generated Decision Tree for Test C

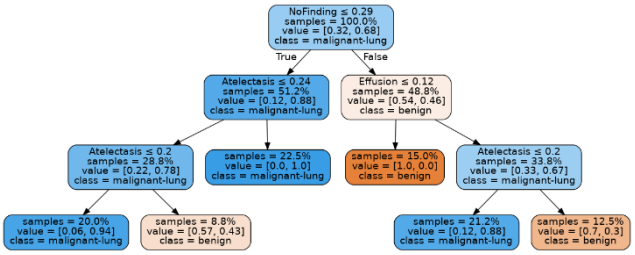

Fig. 14. Automatically Generated Decision Tree for Test D

VII. Discussion and Future Directions

We have described a method whereby interpretable decision trees fitted to multiple features extracted from CXR images (using deep learning) have been used to separate an independent lung nodule dataset into benign or malignant classes with good results.

The described results show the potential of hybrid computer vision techniques in automatically generating diagnostic scoring models from CXR image data. Interestingly, the model from Test A shows apparent relationship between malignant lung cancer originating from the lung and the presence of Effusion and Mass on the LIDC CXR images, with Mass scores over 3% and Effusion scores between 4 and 28% resulting in a Malignant classification of 22.1% of the samples. Similarly, in Model A the No Finding class correctly leads to a Benign classification for a score greater than 2%. This could indicate that Model A was sensitive enough to automatically detect the build-up of fluid and cancer cells between the chest wall and lung associated with malignant lung cancer known as Malignant Pleural Effusion [54]. Whilst none of the generated models could be considered to be fit for clinical purposes at this stage, the fact that a number of models were generated that contain reasonable insights suggests that the techniques used here are able to capture important pathological information and are worthy of further refinement.

Using a novel combination of machine deep learning and decision tree analysis, we have automatically generated lung cancer diagnostic models that are capable of stratifying lung cancer patients into benign/malignant categories with best accuracy approaching 85% and best PPV of 83% for the malignant class**.** False positives tend to be high for all models driven by relatively poor sensitivity to the benign classification with best recall for this class being only 43%. Given the limitations of the datasets used, and in particular the small size of the LIDC dataset used for diagnostic ground truth, these results suggest that with additional data and further refinement this method could potentially be used to develop

useful clinical methods to assist in the diagnosis and scoring of lung cancer. Examples of such future refinements include:

- Sourcing of additional datasets for deep-learning training and decision tree fitting;
- Employ techniques to improve signal-to-noise ratio to improve deep learning accuracy and generalization;
- Testing with targeted clinical validation;
- Implementation in a federated learning framework to assure data security and provide a scalable pathway to the acquisition of a broad-base dataset supporting continuous model improvement.

ACKNOWLEDGMENT

The 1st author would like to thank his employer IBM Australia for ongoing support in relation to the preparation of this paper.

REFERENCES

[1] "Lung cancer statistics." https://www.wcrf.org/dietandcancer/cancer-trends/lung-cancer-statistics (accessed 8 December, 2020).

[2] H. Yoshimura, M. L. Giger, K. Doi, H. MacMahon, and S. M. Montner, "Computerized scheme for the detection of pulmonary nodules: A nonlinear filtering technique," (in English), *Investigative Radiology,* Article vol. 27, no. 2, pp. 124-129, 1992, doi: 10.1097/00004424-199202000-00005.

[3] Paulus and F. L. Gaol, "Lung Cancer Diseases Diagnostic Asistance Using Gray Color Analysis," in *2010 Second International Conference on Computational Intelligence, Modelling and Simulation*, 28-30 Sept. 2010 2010, pp. 355-359, doi: 10.1109/CIMSiM.2010.79.

[4] X. Li *et al.*, "Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection," *Artificial Intelligence In Medicine,* vol. 103, 2020, doi: 10.1016/j.artmed.2019.101744.

[5] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction from Chest X-ray Images Using the Deep Learning Approach," in *11th Biomedical Engineering International Conference, BMEiCON 2018*, 2019: Institute of Electrical and Electronics Engineers Inc., doi: 10.1109/BMEiCON.2018.8609997.

[6] J. Mendoza and H. Pedrini, "Detection and classification of lung nodules in chest X-ray images using deep convolutional neural networks," (in English), *Computational Intelligence,* Article 2019, doi: 10.1111/coin.12241.

[7] P. Gang *et al.*, "Dimensionality reduction in deep learning for chest X-ray analysis of lung cancer," ed: IEEE, 2018, pp. 878-883.

[8] B. Stan, D. Pranavsingh, and M. Bertalan, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database," *npj Digital Medicine,* vol. 3, no. 1, pp. 1-8, 2020, doi: 10.1038/s41746-020-00324-0.

[9] "Scopus - Document search." Elsevier. https://www.scopus.com/search/form.uri?display=basic (accessed 7 December, 2020).

[10] B. Liu *et al.*, "Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect," (in English), *Journal of Cancer Research and Clinical Oncology*, Review vol. 146, no. 1, pp. 153-185, 2020, doi: 10.1007/s00432-019-03098-5.

[11] X. a. P. Wang, Yifan and Lu, Le and Lu, Zhiyong and Bagheri, Mohammadhadi and Summers, Ronald, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," vol. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), ed, pp. 3462--3471.

[12] T. Ho and J. Gwak, "Multiple Feature Integration for Classification of Thoracic Disease in Chest Radiography," *Applied Sciences-Basel,* vol. 9, no. 19, 2019, doi: 10.3390/app9194130.

[13] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," arXiv:1710.10501, 2017.

[14] S. Gündel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to Recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Cham, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds., 2019// 2019: Springer International Publishing, pp. 757-765.

[15] C. Qin, D. Yao, Y. Shi, and Z. Song, "Computer-aided detection in chest radiography based on artificial intelligence: A survey," (in English), *BioMedical Engineering Online,* Review vol. 17, no. 1, 2018, Art no. 113, doi: 10.1186/s12938-018-0544-y.

[16] L. Anna Rita *et al.*, "Lung nodules: size still matters," *European respiratory review*, vol. 26, no. 146, 2017, doi: 10.1183/16000617.0025-2017.

[17] M. Sánchez, M. Benegas, and I. Vollmer, "Management of incidental lung nodules <8 mm in diameter," *Journal of thoracic disease*, vol. 10, no. Suppl 22, pp. S2611-S2627, 2018, doi: 10.21037/jtd.2018.05.86.

[18] D. Gomes et al., "MAVIDH Score: A COVID-19 Severity Scoring using Chest X-Ray Pathology Features", arXiv:2011.14983, 2020.

[19] G. Zhang *et al.*, "An Appraisal of Nodule Diagnosis for Lung Cancer in CT Images," (in English), *Journal of Medical Systems*, Review vol. 43, no. 7, 2019, Art no. 181, doi: 10.1007/s10916-019-1327-0.

[20] J. T. Heverhagen *et al.*, "Lung Nodule Detection by Microdose CT Versus Chest Radiography (Standard and Dual-Energy Subtracted)," ed: American Roentgen Ray Society, 2015.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chest x-ray8 : hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097-2106.

[22] S. G. Armato Iii *et al.*, "Data From LIDC-IDRI," T. Team, Ed., ed: The Cancer Imaging Archive, 2015.

[23] U. Pokharel, S. Shrestha, and P. Neupane, "Atelectasis of Lung as a First Sign of Lung Cancer: A Case Report," *Journal of Lung Cancer*, vol. 1, pp. 1-3, 2016.

[24] S. Revannasiddaiah, B. Bhardwaj, S. P. Susheela, and S. B. Hiremath, "Radiographic illusion of cardiomegaly resulting from a pulmonary blastoma in a patient imaged for evaluation of chronic bronchitis," *BMJ Case Reports*, vol. 2013, no. jul25 1, 2013, doi: 10.1136/bcr-2013-010179.

[25] M. Sayako *et al.*, "Clinical features of primary lung cancer presenting as pulmonary consolidation mimicking pneumonia," *Fujita medical journal*,

vol. 2, no. 1, pp. 17-21, 2016, doi: 10.20407/fmj.2.1_17.
[26] W. Abouzgheib and R. P. Dellinger, *Pulmonary complications in cancer patients*. Springer International Publishing, 2016, pp. 191-202.
[27] P. Ioannis, K. Ioannis, M. P. Jose, W. R. Bruce, and T. S. Georgios, "Malignant pleural effusion: from bench to bedside," *European respiratory review*, vol. 25, no. 140, pp. 189-198, 2016, doi: 10.1183/16000617.0019-2016.
[28] J. Zulueta, "Emphysema and Lung Cancer: More Than a Coincidence," *Annals of the American Thoracic Society*, vol. 12, no. 8, pp. 1120-1121, 2015, doi: 10.1513/AnnalsATS.201506-360ED.
[29] T. Karampitsakos *et al.*, "Lung cancer in patients with idiopathic pulmonary fibrosis," *Pulmonary pharmacology & therapeutics*, vol. 45, pp. 1-10, 2017, doi: 10.1016/j.pupt.2017.03.016.
[30] J. V. Lodhia, S. Appiah, P. Tcherveniakov, and P. Krysiak, "Diaphragmatic hernia masquerading as a pulmonary metastasis," *Annals of the Royal College of Surgeons of England*, vol. 97, no. 2, pp. e27-e29, 2015, doi: 10.1308/003588414X14055925060758.
[31] "Pulmonary Nodules - Health Encyclopedia - University of Rochester Medical Center." https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=22&contentid=pulmonarynodules (accessed 10 December, 2020).
[32] S. Akira *et al.*, "Pleural thickening on screening chest X-rays: a single institutional study," *Respiratory research*, vol. 20, no. 1, pp. 1-7, 2019, doi: 10.1186/s12931-019-1116-9.
[33] S. Cicėnas and V. Vencevičius, "Spontaneous pneumothorax as a first sign of pulmonary carcinoma," *World journal of surgical oncology*, vol. 7, no. 1, p. 57, 2009, doi: 10.1186/1477-7819-7-57.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2015.
[35] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017, doi: 10.1145/3065386.
[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," vol. 2017-, ed: Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2261-2269.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.
[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2016.
[39] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
[40] T. Shermin, S. W. Teng, M. Murshed, G. Lu, F. Sohel, and M. Paul, "Enhanced Transfer Learning with ImageNet Trained Classification Layer," vol. 11854, ed: Springer, 2019, pp. 142-155.
[41] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," 2019.
[42] H. Pham, T. Le, D. Ngo, D. Tran, and H. Nguyen, "Interpreting Chest X-rays via CNNs that Exploit Hierarchical Disease Dependencies and Uncertainty Labels," arXiv:1911.06475, 2020.
[43] I. Allaouzi and M. Ben Ahmed, "A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases," *IEEE Access*, vol. 7, pp. 64279-64288, 2019, doi: 10.1109/ACCESS.2019.2916849.
[44] "Transfer Learning for Computer Vision Tutorial - PyTorch Tutorials 1.7.1 documentation." https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html (accessed 10 December, 2020).
[45] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412:6980, 2017.
[46] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," 2016.
[47] L. Yao, E. Poblenz, B. Covington, and K. Lyman, "Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions," arXiv:1803:07703, 2018.
[48] M. B. Ivo, N. Hannes, G. Michael, K. Tobias, and S. Axel, "Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification," *Scientific reports*, vol. 9, no. 1, pp. 1-10, 2019, doi: 10.1038/s41598-019-42294-8.
[49] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Medical Physics*, vol. 38, no. 2, pp. 915-931, 2011, doi: 10.1118/1.3528204.
[50] S. Bradley *et al.*, "Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review," 2019.
[51] M. Feng, X. Yang, Q. Ma, and Y. He, "Retrospective analysis for the false positive diagnosis of PET-CT scan in lung cancer patients," *Medicine*, vol. 96, no. 42, pp. e7415-e7415, 2017, doi: 10.1097/MD.0000000000007415.
[52] S. S. Birring and M. D. Peake, "Symptoms and the early diagnosis of lung cancer," *Thorax*, vol. 60, no. 4, p. 268, 2005, doi: 10.1136/thx.2004.032698.
[53] S. Bradley *et al.*, "Sensitivity of chest X-ray for lung cancer: systematic review," pp. 1). bjgp18X696905. ISSN 0960-1643, 2018.
[54] R. Dixit *et al.*, "Diagnosis and management options in malignant pleural effusions.(Review Article)(Report)," *Lung India*, vol. 34, no. 2, p. 160, 2017, doi: 10.4103/0970-2113.201305.