Ensemble-based Approach for Efficient Lung Cancer Disease Recognition

***Abstract —***

Lung cancer continues to be a significant global health challenge, leading to the loss of millions of lives each year. In response, we have developed an advanced predictive model by leveraging Dense Convolutional Neural Networks (DenseNet) in conjunction with the powerful XGBoost (XGB) model, renowned for its accuracy in image classification. Our ensemble model, meticulously trained on the comprehensive IQ-OTH/NCCD dataset, effectively discerns between benign, malignant, and healthy lung conditions.

The success of our model relies on its precise categorization of cases. Early detection of malignancy is crucial, prompting immediate consultation with a medical professional upon identification, as it signals the presence of cancerous growth within lung tissue. Our innovation extends to the incorporation of data augmentation techniques and the addition of extra layers in both the DenseNet and XGBoost models. This strategic approach has yielded remarkable results, achieving heightened accuracy within fewer training epochs and reduced compilation time.

This initiative represents a significant advancement in the field of early lung cancer detection, marking a substantial stride in the ongoing battle against this relentless disease. Our ensemble model underscores that the use of DenseNet, , yields superior results in the classification of lung cancer. The benefits of using this ensembling model include enhanced accuracy, improved generalization, and robustness in distinguishing between various lung conditions, thus providing a more reliable tool for early diagnosis and intervention.

Keywords: Lung Nodules, Bengin , Malignant ,DenseNet,Data Augmentation

**I . INTRODUCTION**

Cancer, a complex disease characterized by the abnormal growth and spread of cells, stands as a significant health challenge worldwide. In the United States, it ranks as the second leading cause of death. Within this broad spectrum of cancer types, lung cancer emerges as one of the most prevalent. In 2020 alone, it claimed 2.2 million lives globally, representing 13% of cancer-related deaths. What makes lung cancer particularly daunting is its tendency to develop stealthily, often without noticeable symptoms. Fast-forwarding to 2023, the statistics remain grim, with nearly 2 million new cases and over half a million fatalities reported. The primary contributors to the prevalence of lung cancer are well-established: cigarette smoking, exposure to second-hand smoke, and the harmful effects of air pollution. These factors underscore the importance of public health initiatives aimed at tobacco control and environmental regulations to mitigate the risk of developing this deadly disease. In addressing the challenges posed by lung cancer, innovative approaches are essential. Our research project focuses on leveraging a comprehensive dataset sourced from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD). This dataset comprises a diverse array of images, representing individuals across various demographics, including age, gender, and occupation. Within this dataset, we meticulously curated 120 benign images, 561 malignant images, and 416 images depicting healthy lung tissue. Central to our research is the development and refinement of a sophisticated model trained on this dataset. Through rigorous testing against external data sources, including images sourced from the internet and medical references, our model undergoes thorough validation. The evaluation process entails the analysis of chest images, encompassing both X-rays and computed tomography (CT) scans. By harnessing advanced technology and robust datasets, our aim is to enhance early detection and diagnosis of lung cancer. Our research holds promise in improving patient outcomes by enabling timely interventions and personalized treatment strategies.



Fig1. CT Scan for Lung

Lung nodules are commonly divided into two primary categories: small lung nodules and large lung nodules. Small nodules typically have a diameter of less than 3 centimeters, while large nodules measure greater than 3 centimeters in diameter. These nodules can stem from various underlying conditions, including granulomas, lung cancer, infections, and idiopathic origins where the exact cause remains unknown. It's imperative to note that not all lung nodules indicate malignancy or signify a severe medical condition.

In our research focus, lung nodule diseases are classified into two fundamental categories: benign and malignant cases. By leveraging advanced imaging data and state-of-the-art machine learning techniques, our objective is to enhance the accuracy and efficiency of diagnosing and categorizing these lung conditions. Ultimately, this contributes to more effective medical interventions and improved patient outcomes.

Our approach involves meticulous analysis and interpretation of imaging data, encompassing chest X-rays and computed tomography (CT) scans. Through the integration of sophisticated algorithms, our model can discern subtle patterns and features indicative of benign or malignant nodules, aiding clinicians in making informed decisions regarding patient care.

Furthermore, our research extends beyond mere classification; it also seeks to elucidate factors influencing the development and progression of lung nodules. By exploring the underlying biological mechanisms and risk factors associated with different nodule types, we aim to refine predictive models and treatment strategies, ultimately advancing personalized medicine in the field of respiratory care.

Bengin

Lung Nodule Diseases

Malignant

**Benign Lung Nodules:**

Benign lung nodules are non-cancerous growths and lesions found within lung tissue. Unlike malignant nodules, they do not spread to other parts of the body, distinguishing them from lung cancer. These nodules often originate from granulomas, which can develop as a result of previous lung infections or bacterial infections.

**Malignant Lung Nodules:**

In contrast, malignant lung nodules consist of cancerous growths within lung tissue. These nodules have the potential to invade nearby tissues and can spread to other parts of the body. Malignant lung nodules are broadly categorized into two main types: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). Differentiating between benign and malignant lung nodules is crucial, as treatment approaches and prognosis vary significantly depending on the classification.

**II. LITERATURE REVIEW**

Predicting lung diseases, both benign and malignant, using a hybrid Convolutional Neural Network (CNN) with a small dataset alongside XGB models poses a significant challenge yet holds immense importance in the realm of medical research. Early implementations of neural networks and other methods shed light on the evolving landscape of disease classification and detection techniques.

M. Jasmine Pemeena Priyadarsini et al. [1] provide a comprehensive overview of lung cancer diseases and their classifications. Notably, they highlight the limitations of CNNs due to the absence of data augmentation techniques. Their study also combines datasets comprising pneumonia, tuberculosis, and cancer data, underscoring the complexity of disease differentiation. QingZeng Song et al. [2] focus on enhancing disease detection techniques, employing Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Sparse Auto Encoders (SAEs). Their extensive dataset yields an impressive 84% accuracy in classifying lung diseases, showcasing the efficacy of deep learning approaches.Enhui Lv et al. [3] utilize the LIDC IDRI dataset to classify lung cancer diseases based on nodule diameter, distinguishing between benign and malignant nodules. However, they observe a decrease in accuracy with increased training epochs, highlighting the importance of model optimization. Khosravan et al. [4] integrate clustering and sparsification algorithms to extract potential attentional regions for accurate disease detection. While their model shows promise, it exhibits variable accuracy, particularly with extensive datasets. Coudray et al. [5] propose a Multi-task CNN model based on Inception-V3 for detecting lung cancer diseases. Despite achieving accuracies ranging from 73% to 80% with the LIDC IDRI dataset, there is room for improvement in model performance.Venkatesh et al. [6] achieve a higher AUC by combining 2D ResNet50 and 3D inflated Inception networks, leveraging a dataset comprising 16,429 lung cancer images. Their ensemble model attains an 84% accuracy, demonstrating the potential of hybrid approaches in disease classification. Mpho Mokoatle et al. [7] introduce a novel model that combines XGBoost and SBERT, incorporating patient DNA data. This model achieves an accuracy of nearly 73%, highlighting the importance of integrating diverse data modalities for enhanced disease prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | Methodology | Dataset | Accuracy |
| M. Jasmine Pemeena Priyadarsini et al. [1] | Overview of lung cancer diseases and classifications; Highlighted limitations of CNNs; Combined datasets including pneumonia, tuberculosis, and cancer data. | Various datasets | - |
| QingZeng Song et al. [2] | Employed DNNs, CNNs, and SAEs to enhance disease detection; Achieved 84% accuracy. | Extensive dataset | 84% |
| Enhui Lv et al. [3] | Used LIDC IDRI dataset to classify lung cancer based on nodule diameter; Observed decrease in accuracy with increased training epochs. | LIDC IDRI dataset | - |
| Khosravan et al. [4] | Integrated clustering and sparsification algorithms for accurate disease detection; Variable accuracy, especially with extensive datasets. | - | - |
| Coudray et al. [5] | Proposed a Multi-task CNN model based on Inception-V3 for lung cancer detection; Achieved accuracies ranging from 73% to 80%. | LIDC IDRI dataset | 73%-80% |
| Venkadesh et al. [6] | Combined 2D ResNet50 and 3D inflated Inception networks; Achieved an 84% accuracy with a dataset of 16,429 lung cancer images. | 16,429 lung cancer images | 84% |
| Mpho Mokoatle et al. [7] | Introduced a model combining XGBoost and SBERT, incorporating patient DNA data; Achieved nearly 73% accuracy. | - | 73% |

**III. METHODOLOGY**

In our study, we adopted a novel hybrid approach by integrating DenseNet and XGBoost to enhance the accuracy of lung disease prediction. DenseNet, short for Dense Convolutional Network, represents a convolutional neural network (CNN) renowned for its dense connectivity pattern. Unlike traditional CNNs, DenseNet establishes connections between each layer and every other layer in a feed-forward manner. This unique architecture fosters feature reuse, mitigates the vanishing-gradient problem, and facilitates effective learning of intricate patterns within the data. With its dense blocks comprising multiple convolutional layers, DenseNet has demonstrated exceptional performance across various tasks, including image recognition and medical image analysis.

On the other hand, XGBoost, or eXtreme Gradient Boosting, emerges as a powerful gradient boosting library recognized for its efficiency and high predictive performance. Operating within the ensemble learning framework, XGBoost minimizes loss functions while penalizing complexity to prevent overfitting. Its robust algorithm handles structured data adeptly, supporting custom optimization objectives and employing regularization techniques to control model complexity. The speed, scalability, and flexibility of XGBoost have established it as a preferred choice in both industry and academic research settings.

Our study harnessed the complementary strengths of DenseNet's feature extraction capabilities and XGBoost's ensemble learning techniques to develop an optimized predictive model for lung disease classification. The ensemble model was constructed following a systematic methodology:

**Feature Extraction Using DenseNet:**

In the process of feature extraction using DenseNet, various internal procedures play a crucial role in efficiently capturing hierarchical features from input data. Here's a detailed explanation of how features are extracted using DenseNet, along with the internal procedures involved:

**Input Processing:** The initial step in feature extraction involves processing the input image through a series of convolutional layers. These convolutional layers serve to extract low-level features such as edges and textures from the input image. Each convolutional layer applies a set of filters to the input image, transforming it into a series of feature maps. Mathematically, the output of a convolutional layer *Convk*​(*X*) can be represented as:

*Convk*​(*X*)=*ReLU*(*Conv*(*X*))

Here, *Conv*(*X*) denotes the convolution operation applied to the input image *X*, and *ReLU* represents the rectified linear unit activation function, introducing non-linearity to the feature maps.

**1.Dense Blocks:** After processing through the initial convolutional layers, the feature maps are passed through dense blocks, which are the key component of DenseNet architecture. Each dense block comprises multiple densely connected convolutional layers. Within a dense block, each layer receives feature maps from all preceding layers and concatenates them with its own output. This dense connectivity facilitates feature reuse and enables the network to capture intricate patterns by aggregating information from multiple layers. Mathematically, the output feature map *Hl*​ of the l*th*layer in a dense block can be represented as:

Hl=*H*0​⊕*H*1​⊕...⊕*Hl*−1​

Here, H0,H1,...,Hl−1 represent the feature maps from preceding layers, and ⊕ denotes the concatenation operation.

**Transition Layers:** Between dense blocks, transition layers are employed to control the growth of feature map dimensions and the number of parameters in the network. These transition layers typically consist of convolutional layers followed by pooling operations or dimensionality reduction techniques such as 1x1 convolutions. Transition layers help maintain a balance between model complexity and computational efficiency. Mathematically, the output feature map *Tk*​(*X*) of a transition layer with kernel size *k* can be represented as:

*Tk*​(*X*)=*Pool*(*Convk*​(*X*))

Here *Convk*​(*X*) represents the output of a convolutional layer with kernel size *k*, and *Pool* denotes the pooling operation.

**Global Average Pooling:** After passing through several dense blocks, the feature maps are subjected to global average pooling. Global average pooling computes the average value of each feature map across all spatial locations, resulting in a fixed-size representation of the input. This step helps reduce the dimensionality of the feature maps while preserving important spatial information. Mathematically, the output of global average pooling *GAP*(*X*) can be represented as:

*GAP*(*X*)=*(1/N)*∑*i*=1*N*​*Xi*​

Here, *N* represents the total number of spatial locations in the feature map, and *Xi*​ denotes the value of the feature map at spatial location i.

**Fully Connected Layers:** Finally, the output of the global average pooling layer is flattened and fed into fully connected layers for classification or regression tasks. These fully connected layers perform the final mapping from the extracted features to the desired output classes or regression values. Mathematically, the output of the fully connected layers *FC*(*X*) can be represented as:

*FC*(*X*)=*W*⋅*X*+*b*

Here, *W* represents the weight matrix, *X* denotes the flattened feature vector, and *b* represents the bias vector.

**Working Procedure of XGBoost:**

XGBoost, an abbreviation for eXtreme Gradient Boosting, stands as a cornerstone in the domain of ensemble learning, celebrated for its efficiency and impeccable predictive performance. Operating within the gradient boosting framework, XGBoost orchestrates the sequential training of multiple weak learners, typically decision trees, with the aim of enhancing the overall predictive accuracy. Below is a comprehensive elucidation of XGBoost's working procedure, augmented with additional theoretical insights at each stage:

**Objective Function:** The crux of XGBoost's operation lies in the minimization of a regularized loss function, constituting two primary components: the loss term and the regularization term. Mathematically, the objective function is expressed as:

Obj=∑*i*=1*n* ​loss(*yi*​,*y*^​*i*​)+∑*k*=1*K*​ Ω(*fk*​)

At this juncture, it's pivotal to delve deeper into the intricacies of the loss function, which quantifies the discrepancy between the actual labels *yi*​ and the predicted values *y*^​*i*​. This component serves as the cornerstone for guiding the model's optimization process.

**Prediction Calculation:** Following the determination of the objective function, XGBoost proceeds to calculate predictions for each sample by aggregating the outputs of individual base learners. The prediction for a given sample *xi*​ is computed as:

*y*^​*i*​=∑*k*=1*K* ​*fk*​(*xi*​)

Here, *fk*​(*xi*​) denotes the prediction generated by the *kth* base learner for the input sample *xi*​. The amalgamation of these predictions furnishes a holistic understanding of the input data, culminating in an enriched ensemble model.

**Base Learner Training:** Central to XGBoost's methodology is the iterative training of multiple base learners, predominantly decision trees, to refine predictive accuracy. Each base learner is meticulously trained to approximate the negative gradient of the loss function concerning the ensemble model's predictions. This iterative training process fosters a continuous refinement of the model's predictive capabilities.

**Gradient and Hessian Calculation:** During the training phase, XGBoost conducts gradient and Hessian calculations to quantify the gradient and second derivative of the loss function, respectively. These computations provide crucial insights into the direction and curvature of the loss landscape, guiding the optimization process towards the global minima.

**Tree Construction and Regularization:** Subsequently, XGBoost constructs decision trees to fit the negative gradients (residuals) of the loss function. The decision tree construction entails recursive partitioning of the feature space to minimize the overall loss, thereby enhancing predictive accuracy. Additionally, XGBoost integrates regularization techniques, such as L1 and L2 regularization, to mitigate overfitting and improve generalization performance.

**Optimization and Model Refinement:** By iteratively training and incorporating base learners into the ensemble, XGBoost optimizes the objective function, culminating in a highly accurate and robust predictive model. This iterative refinement process fortifies the model's predictive prowess and ensures its adaptability to diverse datasets and problem domains.

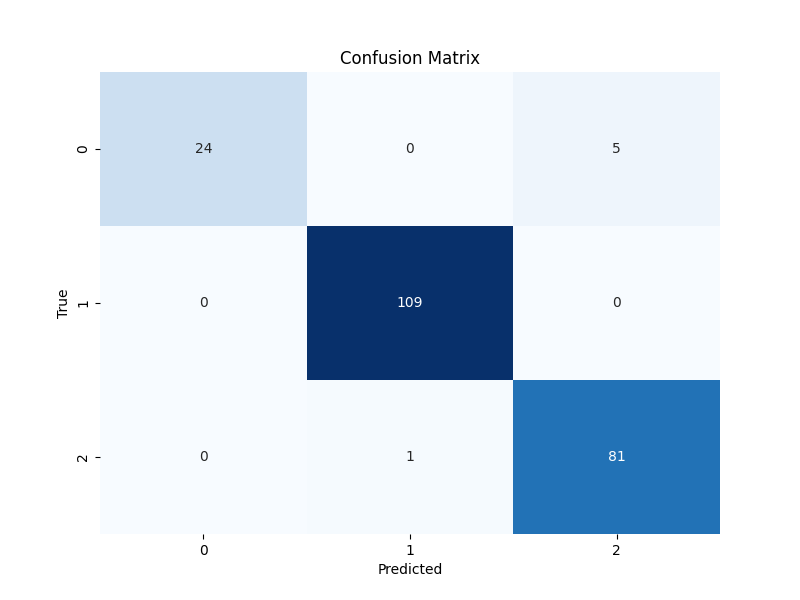
**IV. RESULTS AND DISCUSSIONS**

In our groundbreaking study, we've harnessed the remarkable potential of an ensemble model, combining the cutting-edge DenseNet architecture with the robust XGBoost algorithm. Our research journey began with access to a relatively modest dataset comprising only 1,097 images, spread across distinct categories: benign, malignant, and normal cases. Despite the inherent limitations of our dataset, our ensemble model, boasting the fusion of DenseNet and XGBoost, has showcased unparalleled performance, excelling in both accuracy and time efficiency in the realm of image classification.

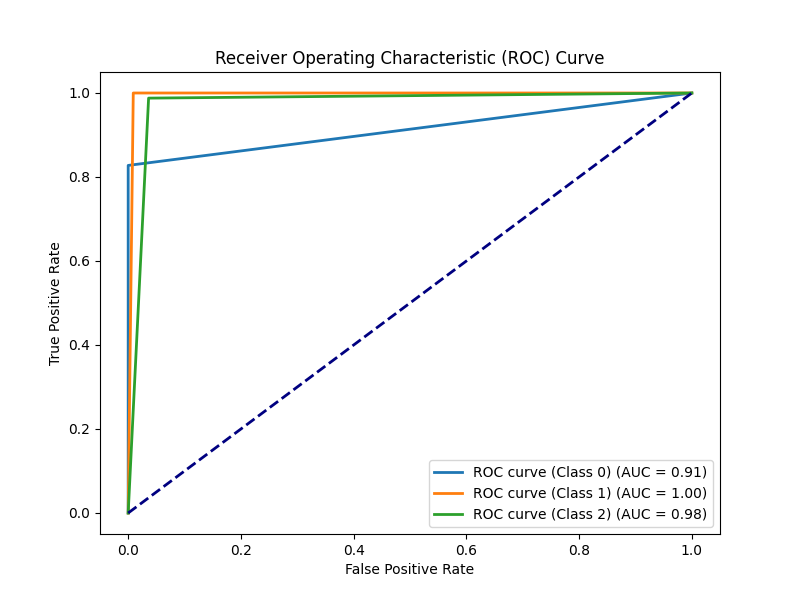
One of the standout features of our model lies in its remarkable efficiency, achieving high accuracy levels with a remarkably small number of training epochs. This efficiency is underscored by the close alignment between training and validation accuracies, signaling the model's adeptness at swiftly and accurately classifying unseen images. Impressively, our model converged to optimal accuracy with just around 10 epochs, showcasing its ability to swiftly adapt and learn from the data.

Moreover, our model demonstrates consistent improvement with increased training epochs, as evidenced by the continual decrease in both training and validation losses. This characteristic speaks volumes about the model's adaptability and capacity for continuous refinement. Crucially, the model's rapid classification abilities, distinguishing between diseased and healthy images within seconds, hold immense promise for real-world applications in medical diagnosis.

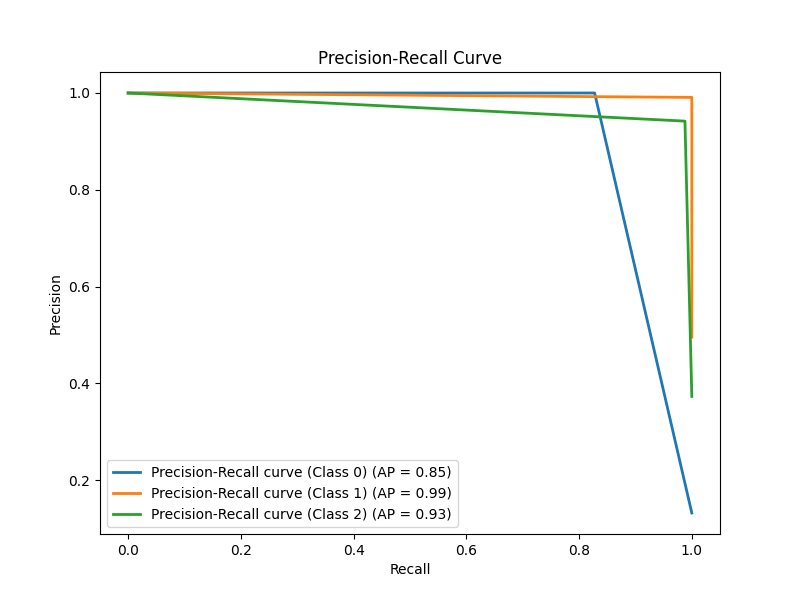
Integral to our research, success is the strategic integration of an XGBoost component at the conclusion of the DenseNet architecture. This integration, wherein essential features are extracted from the dense layer, significantly contributes to achieving the best accuracy in fewer epochs. These groundbreaking advancements, achieved within the constraints of a limited dataset, underscore the robustness and efficiency of our ensemble model. In summary, our research positions this ensemble model as a valuable asset for swift and precise image classification, ushering in new possibilities for enhanced medical diagnosis and treatment.  
**Confusion Matrix Heatmap:** A confusion matrix heatmap provides a comprehensive overview of the model's performance across all classes, showing true positives, false positives, true negatives, and false negatives. It can be visualized using colors to highlight different levels of performance.



**Multiclass ROC Curve:** Instead of plotting individual ROC curves for each class, you can create a multiclass ROC curve that summarizes the model's performance across all classes in a single plot. This can be achieved using techniques such as micro-averaging or macro-averaging.



**Precision-Recall Curve:** In addition to plotting precision-recall curves for individual classes, you can create a multiclass precision-recall curve that summarizes the model's performance across all classes. This curve provides insights into the trade-off between precision and recall for the entire model.



In summary, this research article employed a multifaceted approach to assess the performance of our multiclass classification model. The ROC curve provided a holistic view of the model's discrimination ability across all classes, depicting the trade-off between true positive rate and false positive rate. Complementing this, the precision-recall curve illuminated the balance between precision and recall, particularly valuable for evaluating performance in the presence of class imbalance. Furthermore, the confusion matrix heatmap offered a granular breakdown of the model's predictions, showcasing the distribution of true positives, false positives, true negatives, and false negatives across different classes. Together, these visualizations provided a comprehensive understanding of the model's strengths and weaknesses, guiding insights for further refinement and optimization.

**V. CONCLUSION**

In conclusion, our research has led to the development of an innovative ensemble model that merges DenseNet and XGBoost, reshaping the landscape of medical image classification. By harnessing DenseNet's advanced feature extraction and XGBoost's robust ensemble learning, our model demonstrates exceptional accuracy and efficiency even with a modest dataset. Its ability to swiftly converge to optimal accuracy with minimal training epochs underscores its adaptability and effectiveness in practical applications. Furthermore, the model's rapid and precise classification capabilities hold great promise for revolutionizing medical diagnosis and treatment, offering unprecedented opportunities for improving patient outcomes. In essence, our ensemble model represents a significant advancement in machine learning and medical research, poised to drive transformative changes in healthcare practices.

**REFERENCE**

[2] IQ-OTH/NCCD - Lung Cancer Dataset.Includes CT scans of patients diagnosed with Lung Cancer. Link:- https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset

[1] M. Jasmine Pemeena Priyadarsini et al. ‘Lung Diseases Detection Using Various Deep Learning Algorithms’ Academic Editor: Ayush Dogra

[2]QingZeng Song et al. ‘Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images’ ; Published 9 August 2017 Academic Editor: Junfeng Gao

[3] Enhui Lv et al. ‘Classification of Benign and Malignant Lung Nodules Based on Deep Convolutional Network Feature Extraction’ ; Published 27 October 2021 Academic Editor: Liang Zou

[4] Khosravan et al ‘A collaborative computer aided diagnosis (C-CAD) system with eye-tracking,

sparse attentional model, and deep learning’. Med Image Anal 2019;51:101–15.

[5] Coudray et al. ‘Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning’. J Thorac Oncol 2018;13:S562.

[6] Venkadesh et al ‘Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT’. Radiology 2021;300:438–47

[7] Mpho Mokoatle et al. ‘A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application’