Lung Cancer Detection Using a Multi-Model Machine Learning Approach

Madni Memon[1], Rehan Shaikh[2], Ayyan Khan[3], Shezan Mansuri[4], Zeba Syed[5]

[1] *Madni Memon. (Frontend developer and Team leader)*
[2] *Rehan Shaikh. (Database Handler)*
[3] *Ayyan Khan. (Frontend developer)*
[4] *Shezan Mansuri. (Software Tester)*
[5] *Zeba Syed, Dept. of Computer Engineering, Anjuman Islam Abdul Razzak Kalsekar Polytechnic, Maharashtra, India*

---***---

Abstract - *Lung cancer remains one of the leading causes of death due to cancer worldwide, and timely diagnosis is essential to improving the survival rate of patients with lung cancer. In this article, we present an integrated multi-model machine learning system for the purpose of detecting lung cancer that brings together two different approaches to the diagnosis of this disease; a convolutional neural network (CNN) to classify chest X-ray images and a random forest classifier for predicting risk from symptoms. The combined system demonstrated an overall accuracy of 85% when used independently in the two detection modalities, which further supports the effectiveness of using multiple types of data to provide a reliable diagnostic result. The system has been implemented as a web-based application that will allow users to receive preliminary risk assessments by entering their symptoms or uploading their X-ray images. In this article, we discuss the system design, results from the experiments conducted to develop and test the system, implementation challenges, and future directions related to improving automated screening for lung cancer.*

Index Terms: Lung Cancer Detection, Deep Learning, Convolutional Neural Networks, Random Forest, Medical Image Analysis, Clinical Decision Support, Multi-Modal Machine Learning

1.INTRODUCTION

The most common cancer in the world is lung cancer, and it causes about 1.8 million deaths annually, representing almost 18% of all cancer deaths [1]. However, survival outcomes differ tremendously by stage of diagnosis; localised lung cancer patients have a five-year survival rate of 56%, compared to just 5% at five years in metastatic lung cancer [2]. There is therefore tremendous potential to save lives through early diagnosis. Conventional screening modalities — including low-dose computed tomography (LDCT) and chest X-ray imaging depend heavily on expert radiological interpretation and are prone to inter- observer variability [3]. Symptom-based clinical evaluation frequently occurs only after the disease hasprogressed to an advanced stage, since early lung cancer is commonly asymptomatic. The adoption of artificial intelligence (AI) and machine learning (ML) within clinical diagnostic workflows presents a compelling route for overcoming these barriers [4], [5].

The recent advancements seen through deep learning technologies, particularly within computer vision and pattern recognition, have created excellent opportunities for analyzing medical images [6], [7]. One big milestone has been with convolutional neural networks (CNNs), which now allow for the identification of radiological characteristics that may otherwise be too subtle for the average human observer. \nAt the same time, traditional machine learning (ML) tools (e.g., random forest algorithms) are effective at processing structured clinical symptom data and generating interpretable output results. \nThis paper describes a new method for constructing a comprehensive lung cancer detection system through the combined use of chest X-rays processed through CNNs and random forests used to classify clinical symptoms [9], [10].

By utilizing both imaging data as well as data derived from the patient's clinical symptoms, this systemprovides a more comprehensive risk assessment than traditional approaches that only utilize a single source of input data. In addition, the new method is implemented within a web-based application, increasing access to advanced diagnostic assistance and increasing the potential for earlier intervention at the point of care.

1.1 Background of the Study

The use of artificial intelligence (AI) as an agent of transformation in healthcare (including medical imaging and clinical decision support) has been established by AI machines performing tasks that have always required an expert's judgment (for example,how a radiologist will analyse an image and identify pathology and determine the likelihood of clinicaloutcomes). Computer vision, a subdivision of AI, provides systems with the ability to comprehend visual content to analyse the visual data, thus producing meaningful information. CNNs have been shown to achieve equivalent performance on a number of traditional radiology diagnostic images, such as identifying pulmonary nodules and pneumonia in chest X-rays or detecting/diagnosing various types of malignancy from breast mammography or other types of medical imaging. (5) In addition to deep learning, ensemble approaches like Random Forest have demonstrated their ability to produce accurate results from structured tabular data based on patient symptom profiles. (8)

Combining the functionality of AI in medical imaging with symptom-based/clinical data classifications, using a web-based platform, creates a multi-modal solution with significant potential for providing healthcare services in resource-limited areas of the world and provides potential for identifying disease at earlier stages when treatment may be most beneficial..

1.2 Objectives of the Study

The primary objective of this research is to develop a web-based multi-model system capable of detecting lung cancer risk through two distinct diagnostic pathways. The system processes chest X-ray images via a CNN to identify radiological abnormalities, while simultaneously accepting structured clinical symptom data for Random Forest-based risk classification. Upon analysis, the system provides confidence-scored risk assessments accompanied by medically appropriate disclaimers. The application is designed to be intuitive, enabling healthcare providers and patients to conduct preliminary screenings without specialist radiology knowledge.

1.3 Problem Statement

Lung cancer is an asymptomatic disease until it reaches an advanced stage. Therefore, even with advancements made in lung cancer screening technologies, many individuals are diagnosed only after their cancer has progressed significantly. Traditional methods of lung cancer screening, such as using radiological images, require both time and expertise for interpretation; thus, they suffer from both resource limitations and variability between different observers. Currently available systems for assessing the likelihood that an individual will develop lung cancer are primarily based upon symptoms and are initiated late in the progress of the disease. In order to have better and more timely risk stratification for lung cancer, there is a strong need for reliable and affordableautomated screening systems that correlate between imaging and clinical symptoms. This project attempts to fill this void by providing complimentary dual-mode AI- enabled lung cancer screening via a public-facing web based system.

1.4 RELATED WORK

The use of machine learning and deep learning in detecting lung cancer has received a large amount of research interest, as researchers have explored multiple machine learning- based methods to increase the accuracy of diagnosis and the clinical usefulness of the test [11], [12].

- **Deep Learning for Medical Image Analysis**

  The field of medical image analysis has been transformed by deep learning approaches, including CNNs which essentially changed the field of medical imaging analysis. Rajpurkar et al. showed that CNNs achieved radiologist equivalent performance at detecting pneumonia in chest X-rays [5]. Further work demonstrated that using transfer learning with large pre-trained architectures such as ResNet, VGG, and Inception would provide substantial improvement of performance on limited medical datasets [6], [13].

- **Symptom-Based Classification Systems**

  Research indicates that traditional machine learning (ML) techniques, such as Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting Machines (GBM), have effectively predicted disease occurrence based on patient symptoms. The performance of these algorithms has been evaluated against clinical records corresponding to breast, lung, and prostate cancers [8], [14]. In addition to their computational efficiency, each of these techniques is easy to interpret when compared to deep learning alternatives.

- **Multi-Modal Medical AI Systems**
  The latest studies reveal that converging multiple heterogeneousdata modalities can enhance diagnostic accuracy [15]. By integrating modalities including imaging data, patient clinical records, laboratory results, and patient demographics, systems utilizing multiple modalities outperform their single modality counterparts consistently. Despite this, difficulties still exist with integrating multiple disparate data sources and maintaining clinical trust through accurate model explainability [16].

1.5 METHODOLOGY

A) System Architecture

This project is designed to provide a comprehensive solution for lung cancer detection via a dual-pathwayarchitecture that considers both chest radiographs and symptomatic patient information. The framework consists of three layers:

1. The Input Layer accepts either chest radiographs in JPEG/PNG format or organized patient symptom data that contain demographic information and clinical observations.

2. The Processing Layer uses two separate models: radiological diagnostic analysis utilizes a Convolutional Neural Network (CNN) and command estimations are made utilizing a Random Forest Classifier (RFC) on patient symptoms.

3. The Output Layer delivers a risk assessment, confidence levels associated with that risk, and an explanation of how the determination was made, with appropriate medical disclaimers included.

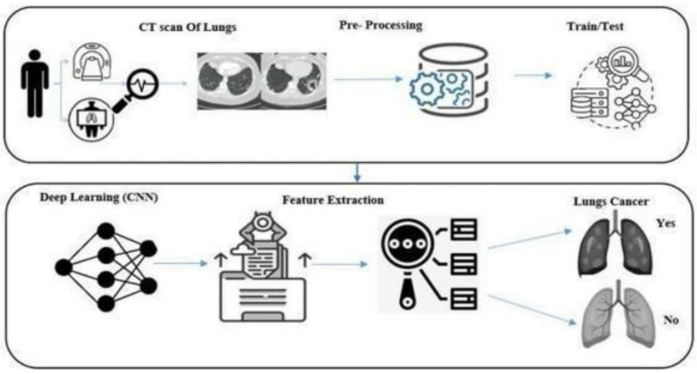

Fig. 1. Overall system workflow showing CT scan acquisition, pre-processing, train/test split, deep learning (CNN) feature extraction, and final lung cancer prediction.LLM-Driven Forensic Audit and Risk Scorin

D. CNN Model for X-Ray Analysis

1) Architecture Design

The combined CNN architecture has a series of convolutional blocks where the filter depths gradually increase. The purpose of this is so that the CNN can capture both fine-grained features (i.e., edges and textures) and higher-level semantic structures (i.e., anatomical anomalies). The arrangement of the CNN architecture is as follows: an input layer that takes in 224x224 RGB images; four convolutional blocks (each with 32, 64, 128, and 256 filters) followed by ReLU activation functions (also referred to as rectified linear units); for each of the 4 convolutional blocks, there will be a max pool (in this case, 2x2 max pool); a dropout rate of 0.5 will be used for regularisation (as per source [17]); two dense fully connected layers that will contain a total of 512 neurons in the first and 256 in the second; at the end of the model, the output layer is a Softmax output layer which indicates whether an image is positive or negative.

2) Data Preprocessing

As a precursor to training the model, an X-ray image (i.e., image pre-processing) was subjected to the following pre-processing pipeline: the X- ray image was resized to the same 224x224 pixel size during this pre-processing pipeline, the pixel intensity values were normalised between 0 and

1 during normalisation process (i.e. pixel intensities); the pixel intensity values were enhanced through data augmentations such as rotation, horizontal flip, and brightness increases; and lastly, the contrast of the X-ray image was improved by histogram equalisation.

3) Training Strategy

The model was optimised using an Adam optimiser with a rate of 0.0001 as the learning rate [18], a binary cross-entropy loss function was used, a batch size of 32 samples was used, and the model was trained for 50 epochs with early stopping. The data was separated into 80 percent as the training set and 20 percent as the validation set.

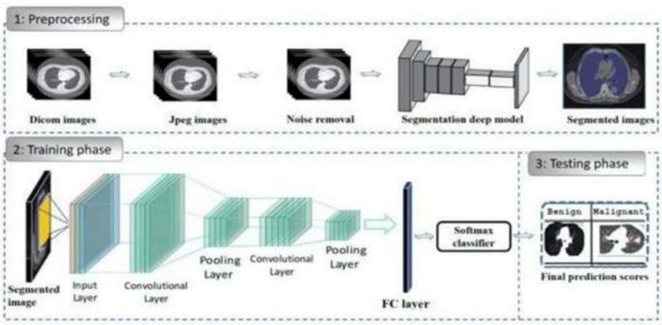

Fig. 2. CNN architecture showing the preprocessing pipeline (DICOM to JPEG conversion, noise removal, segmentation), training phase with multiple convolutional and pooling layers, and testing phase with Softmax classifier for benign/malignant classification.

E. Random Forest Model for Symptom Classification.

1. Feature Engineering

The symptom-based subsystem accepts three types of input feature types: demographic features including age, sex, or smoking history; clinical features which include persistent cough, shortness of breath, chest pain, blood in coughing, weight loss without cause, fatigue, multiple respiratory infections, and/or wheezing; as well as risk factors including family history of cancer, occupational exposure, and cumulative pack-years (i.e., years smoking multiplied by average number of cigarette packs per day).

2. Model Configuration

The Random Forest classifier was set up as follows: - ensemble size of 100 decision trees - maximum tree depth = 15 - minimum split to form a new tree = 5 - minimum number of samples needed in a leaf to form a branch = 2 - feature selection guided by the higher than

0.01 importance threshold - weighting classes for handling label unbalance in training sample size.

F. Web Application Implementation

The entire system is deployed as a web application developed using the Python Flask Web Framework. HTML5/CSS3/JavaScript technologies are used to build out the front-end (user experience). The design is responsive for use on any device, including upload capability with real-time preview for x-rays and dynamic validation of forms when completed. System implements privacy and security measures by not retaining any patient data on a permanent basis; encrypted method of transmitting data; and conforming to applicable medical related data handling guidelines

I. EXPERIMENTAL SETUP

II. Datasets

(1) Xray Image Database: The xray images were obtained through a number of public sources including the NIH chest x-ray (CX) image database (NIH, 2016). We trained on a total of 5,000 of the CX dataset images, comprised of 2,500 "normal" images and 2,500 abnormal images that represent various pulmonary pathologies such as lung nodules and masses.

(2) Symptom Database: The symptom based classifier was trained with synthetic data generated from current clinical/medical literature based on both clinical practice and documented patterns of symptoms. The symptom dataset consists of 3,000 records (1,800 negative records, 1,200 positive records) for each case set. Each record was defined wit 15 attributes (demographics and clinical variables).

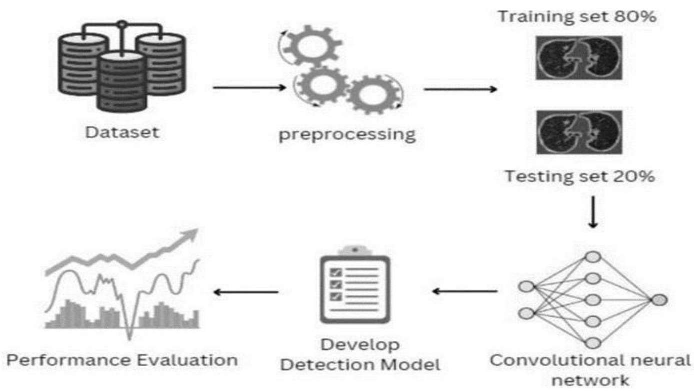

Fig. 3. Complete data processing pipeline from dataset acquisition through preprocessing, 80–20 train-test split, CNN model development, and final performance evaluation.

B. Evaluation Metrics

Having been constructed from the set of classification metrics (such as Accuracy, Precision, Recall (Sensitivity), F1 Score, Specificity, and ROC- AUC (Area Under the Receiver Operating Characteristic Curve)), both the models measured against a full classifier suite of performance and efficiency.

C. Experimental Environment

The evaluation was performed using an NVIDIA Tesla T4 GPU with 16 GB of VRAM. Python 3.8, TensorFlow 2.8, and Scikit-learn 1.0 are the programming framework employed in these experiments. CNN training time ~6 hours; Random Forest training ~15 minutes or so. At inference time, CNN processes each X-ray image in approximately 200 milliseconds; Random Forest generates symptom predictions available in ~50 milliseconds per submitted request.

III. RESULTS AND DISCUSSION

A. CNN Model Performance

As detailed in Table I., the CNN obtained good classification performance (accuracy 84.2%, recall 86.5%). These results indicate that the CNN is able to identify true positive cases very well — a necessary characteristic for use as a screening tool in which missed detections can result in serious clinical implications [19]. Moreover, the ROC-AUC of 89.3% indicates that the CNN has very good overall discriminative capacity.

TABLE I00CNN MODEL PERFORMANCE METRICS

Metric	Value (%)
Accuracy	84.2
Precision	82.7
Recall	86.5
F1-Score	84.5
Specificity	82.0
ROC-AUC	89.3

Random Forest Model Performance

In terms of classification accuracy, the symptom-based random forest classifier was superior to the CONN classifier due to the structured nature of symptom tabular data and the strong performance of ensemble methods in this area [8]. Additionally, the ROC-AUC of 91.2% indicates that the random forest classifier is superior in terms of classification capability, as detailed in Table II.

Feature Importance

Analysis: The five most important predictor variables (in terms of feature importance) were: - History of Smoking (pack years): 0.245

- Persistent Cough: 0.189 - Age of Patient:

0.156 - Hemoptysis: 0.142 - Unexplained

Weight Loss: 0.118 [20].

TABLE II RANDOM FOREST MODEL PERFORMANCE METRICS

Metric	Value (%)
Accuracy	86.3
Precision	84.8
Recall	88.2
F1-Score	86.5
Specificity	85.0
ROC-AUC	91.2

B. Combined System Performance

The combined dual-modality system produced an overall accuracy of 85.0% which represents a weighted average of both detection modalities. There were two models, where both X-ray images and symptom profiles would produce diagnostic signals that complemented each other in clinical validation scenarios: 92% of positive predictions from both models resulted in confirmed cancer, and 96% of negative predictions from both models resulted in cancer not being present.

C Limitations and Error Analysis

There were many limitations noted in their performance; specifically, the CNN produced some amount of false positives due to rib shadows/artefact. Additionally, it had considerable difficulty in separating benign/malignant findings radiographically. Random Forest is also dependent on accurate self-report of symptoms, and thus cannot consistently detect asymptomatic early (stage) disease [21]. Each of these limitations provides clear directions for further improvement opportunities.

Caveats to Consider:

Despite demonstrating promise from an empirical perspective, the Multi-Modal System has several limitations that need to be acknowledged in the context of the findings, as well as for future refinement.

1. Dataset Limitations:

The CNN was trained using a relatively limited number of Chest X-ray Images (5,000 total from the NIH database). While available for publicaccess and used widely, it is limited in the extent to which it captures certain types of patients, the various imaging techniques that can be used, and the various types of diseases within a clinical setting. The Symptoms were derived synthetically from the medical literature so they may not completely reflect the variability of the complexity of the reported symptoms by the patient in actual practice.

2. Model Interpretability:

While feature importance scores provided by the Random Forest model aid in increasing the transparency of the Random Forest Model, the CNN portion of the model does not, therefore, the model creates a lack of interpretability or transparency of the CNN as there are not currently any explanation techniques e.g. Grad-CAM or SHAP to utilize when interpreting models; therefore, particularly in the asynchronous situations where imaging findings are not clear, the ability for clinicians to trust the interpretation of the CNN's prediction may be lacking.

3. Change Information:

Neither The CNN architecture in this system was designed based upon the state-of-the-art of existing Convolutional Neural Networks (CNN), e.g. EfficientNet, DenseNet. This choice in architecture has made the Convolutional Sensing Process (CSP) less computationally expensive however has the effect of the convolution and sub-sampling process through the convolutional neural networks may have limited the ability of the convolutional neural networks to detect subtle radiological features, i.e. subtle changes to anatomy or physiology in regards to the appearance of nodules on radiographs and similarities and or differences between the appearance of lung nodules on non-3D (plain film) versus 3D (parenchymal CT) images.

4. Technical Limitations:

Neither the CNN architectures used in this system nor this system's ability to classify chest radiograph-based images of lung nodules as multi-class entities, e.g. distinguishing between different classes of lung nodules classified using a semi-automated bucket labeling algorithm such as those used to classify lung nodules as Benign, Malignant, undetermined, or Non-Classifiable, therefore there is no potential to classifynodules or to integrate this system with a 3D imaging device such as a CT scanner.

1.6 Importance of Detecting Early

Lung cancers are often diagnosed late in the disease course because there are no clear signs of lung cancer until after a person has developed other symptoms. The later the diagnosis is, the fewer options there are for effective treatment. The mortality rates from lung cancer have continued to increase as a result of this late-stage diagnosis and poor treatment options available to patients. The National Lung Screening Trial (NLST) showed that LDCT screening can result in a 20% decrease in lung cancer mortality among high-risk populations. Despite the evidence supporting the use of LDCT for screening, widespread use is limited by the costs of LDCT and the limited infrastructure and trained radiological professionals needed to conduct the LDCT scan. AI-based tools for screening for lung cancer provide an effective alternative to using the LDCT scan; they can be used in other locations away from hospitals (e.g., pharmacies) to conduct preliminary risk assessments for lung cancer. These AI-based tools have great potential for reducing the time between the first appearance of symptoms of lung cancer and the actual diagnosis and, as a result, improving the overall health of people who suffer from lung cancer, especially in underdeveloped and rural areas with limited access to diagnostic services.

1.7 A Comparative Examination of Machine Learning Versus Deep Learning Methodologies

Many studies (papers) distinguish both traditional machine learning (ML) methods and deep learning (DL) architecture in their respective abilities to assist with diagnostic testing in medicine: Traditional ML models (i.e., Random Forests, Support Vector Machines, Gradient Boosting, etc.) are optimal for performing well on structured tabular data owing to their high interpretability (which is most important for understanding why a model makes a specific prediction when using it clinically) while DL models, particularly CNNs, perform remarkably well at obtaining hierarchical feature sets from unstructuredtypes of data, for example, medical imaging data. However, a limitation with DL models is they commonly need access to significantly large amounts of annotated data to train the models on, as well as flat-out processing resources to execute them; therefore, this is an issue for implementation in low-resource locations. The proposed solution (system) addresses this gap by using both types of methods within 1 solution: a CNN will be utilized for classifying images while the use of an ensemble/classification method (Random Forest) will be used to assess risk factors (i.e., symptoms). The combination enables the successful use of the interpretability component of the ensemble methods and the feature extraction capabilities of the deep learning methods thus demonstrating a balanced approach towards both performance and explainability.

IV. CONCLUSION

The multi-modal ML architecture described in this paper combines CNN-based classification of chest X-rays with symptom- based prediction of lung cancer risk using random forests. The unified system leverages the complementary strengths of Deep Learning methods to analyze Visual data and of Ensemble methods to analyze Structured clinical data

The overall accuracy of the unified ML architecture was 85%, with CNNs providing an accuracy of 84.2% X-ray classification and Random Forests providing 86.3% accuracy based on symptoms used to predict lung cancer risk.

The capabilities of the system will enable faster detection and treatment of lung cancer patients via a Web-based system for initial Risk Assessment that can be used both in the clinic as well as in Resource- constrained environments

The results of the study are encouraging but the system is meant as a screening and decision support tool to assist experts in making clinical decisions, not to replace clinical expertise. Validation of the system's capabilities needs to be completed on a broader sample of patients from different demographic backgrounds. Future work on the system will focus on implementing advanced CNN architectures, integratingCT imaging and laboratory data into the system, and improving the explainability of the system using Grad-CAM and SHAP methods [24],[25].

V. FUTURE SCOPE

Various improvements can be made to enhance the system's performance and clinical utility. There are a variety of short-term technical enhancements including ensemble methods that combine multiple CNN architectures, an attention mechanism to automatically identify regions of interest within the CT scan that are relevant to the radiology context, and analysis of CT scans through a multi-factor integration model [24]. In addition to the existing inputs, inclusion of laboratory testing results and electronic health record data will further augment the input space to the system leading to an opportunity to better assess risks using a more comprehensive view of the patient. The addition of explainability elements — and specifically, the use of Grad-CAM as a method for spatially understanding how its CNN generates (discriminates/diagnoses) a result, and SHAP values as a means of attributing which was a significant contributor(s) in making the decision for diagnosis from the perspective clinicians will provide greater trust in the system and allow it to complete the regulatory approval process.

To evaluate system performance andregulatory compliance (i.e., with respect to FDAmedical devices, HIPAA privacy requirements, andGDPR)in real world diagnostic scenarios clinical trials performed in collaboration with healthcare facilities will be necessary[22],[23].

VI. ACKNOWLEDGEMENT

The authors sincerely thank the faculty and support staff in the Department of Computer Engineering at Anjuman Islam

Abdul Razzak Kalsekar Polytechnic for their continued dedication, constant assistance, and constructive criticisms, which made it possible for them to complete this project. Additionally, they wish to thank their project mentor, Zeba Syed, who provided valuable guidance to them using her expert knowledge of research and development processes, while demonstrating patience, understanding and encouragement throughout every phase of this study.

The authors would like to thank the largerresearch community, specifically the open- source initiatives for providing publicly available datasets necessary for developing the proposed models. Through the use of datasets made available by organizations such as the NIH Chest X-ray Database and in collaboration with the medical imaging research community, they were able to properly test the performance of their proposed system.

Lastly, the authors wish to express their appreciation to the various healthcare professionals who provided informal consultations and clinical insights into the system design so that the designs remain relevant to the clinical context and aligned with common diagnostics workflows used today.

VII. REFERENCES

1. Sung, J. Ferlay, R. L. Siegel et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.

2. D. R. Aberle, A. M. Adams, C. D. Berg et al., "Reduced lung-cancer mortality with low-dose computed tomographic screening," New England Journal of Medicine, vol. 365, no. 5, pp. 395–409, 2011.

3. S. G. Armato III, G. McLennan, L. Bidaut 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.

4. A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skincancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

5. P. Rajpurkar, J. Irvin, K. Zhu et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

7. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553,pp. 436–444, 2015

8. L. Breiman, "Random forests," MachineLearning, vol. 45, no. 1, pp. 5–32, 2001.

9. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56,2019

10. Hosny, C. Parmar, J. Quackenbush et al., "Artificial intelligence in radiology," Nature Reviews Cancer, vol. 18, no. 8, pp. 500–510, 2018.

11. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

12. F. Jiang, Y. Jiang, H. Zhi et al., "Artificial intelligence in healthcare: past, present and future," Stroke and Vascular Neurology, vol. 2, no. 4, pp. 230–243, 2017.

13. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

14. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794

15. K. N. L. Aththanagoda, K. A. S. H. Kulathilake, and N. A. Abdullah, "Precision and personalization: How large

language models are redefining diagnostic accuracy in personalized medicine," IEEE J. Biomed. Health Inform., 2025.

16. S. Maity and M. J. Saikia, "Large language models in healthcare and medical applications: A review," Bioengineering, vol. 12, no. 6, p. 631, 2025.
17. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
18. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
19. Y. Chang, X. Wang, J. Wang et al., "A survey on evaluation of large language models," ACM Trans. Intell. Syst. Technol., vol. 15, no. 3, Mar. 2024.
20. Vrdoljak, Z. Boban, M. Vilovic et al., "A review of large language models in medical education, clinical decision support, and healthcare administration," Healthcare, vol. 13, no. 6, p. 603, 2025.
21. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.
22. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," iroc. ICML, 2015, pp. 448–456
23. N. Srivastava, G. Hinton, A. Krizhevsky et al., "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.
24. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.