International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Survey on Lung Disease Diagnosis using Machine Learning Techniques

P.SaiTeja[1] ,M.Kalyan[2], P.Bhuvan Kumar[3] Students,
Department of Computer Science and Engineering
Mrs.K,Sindhuja[4],Dr.A.Vinoth Kumar[5],Dr.K.Rehkha[6] ,Associate Professor,
Dr .M.G.R Educational and Research Institute, Maduravoyal ,Chennai 95,Tamilnadu,India

Corresponding Authors: P. Sai Teja - saipasumarthy8@gmail.com ,P.Bhuvan Kumar-bhuvank138@gmail.com

Abstract *- Respiratory illnesses continue to seriously impact global health. Since the 2020 outbreak, death rates for those with lung conditions have alarmingly climbed. Early diagnoses prove crucial to promptly treating patients and bettering outcomes. Traditional testing frequently necessitates too much time before results, delaying vital interventions.Advances in computer vision, deep learning algorithms, and openly accessible datasets now allow integrating artificial intelligence in medical diagnoses. Machine learning models have dramatically cut detection times and lessened manual review in lung disease identification.This investigation explores applying various machine learning algorithms for diagnosing different lung diseases. The primary aim is to analyze developing patterns in AI- assisted lung condition detection, recognize current barriers, and envision future opportunities in this area. Improving accuracy and broadening the scope of machine learning-aided lung disease identification can pave the way for more efficient and accessible healthcare solutions.*

Keywords- *Image Classification, Medical Diagnoses, Convolutional Neural Networks, COVID-19, Lung Disease Detection.*

1. Introduction

Lung diseases are complex pathological conditions affecting the respiratory system in mammals, breathing difficulty resulting. Common examples include Pneumonia, COVID-19, Cancer of the Lungs, and Tuberculosis. According to a report from the International Forum of Respiratory Societies, over 334 million children experience asthma, constituting one of the most prevalent persistent diseases among the young, impacting 14% of the youth population.Additionally, more than one billion individuals worldwide deal with continual or abrupt lung diseases, with around four million premature demises taking place annually because of chronic respiratory conditions. People with respiratory diseases regularly struggle with breathlessness, specifically during physical activity or in dusty and shut environments.The COVID-19 pandemic in 2020 had a devastating impact on both the worldwide economy and human life, further highlighting the severity of respiratory diseases. These conditions remain a main cause of death and disability globally. However, early discovery can considerably improve recovery rates and enhance long-term survival, emphasizing the need for developments in diagnostic methods and medical interventions.

A variety of diagnostic tests have traditionally helped physicians identify lung conditions, including blood tests, skin tests, nasal swabs, chest X-rays, and CT scans. In recent years, applying machine learning to medical images has shown great promise in disease detection. Deep learning, a form of machine learning modeled after the human brain, has demonstrated significant potential for identifying, quantifying, and classifying medical pictures. Considerable effort has focused on analyzing CT scans and chest radiographs to diagnose numerous pulmonary disorders using these computerized techniques. While technologies advance, the core aim remains understanding patient health better to guide care.

2. Related Works

A. Covid Diagnoses

S. Tabik and colleagues [1] developed a pre-trained ResNet-50 model that was initialized utilizing ImageNet weights and implemented by means of transfer learning tactics. The academicians crafted a CNN-predicated deep learning architecture called COVID-SDNet designed to prognosticate COVID-19 utilizing chest x-ray imagery. The model was rigorously structured on the COVIDGR dataset, a meticulously balanced corpus encompassing depictions symbolizing diverse severity levels of the novel coronavirus sickness.

During testing, the prototype achieved an exactness of 61.8% ± 5.49% for gentle instances and 86% ± 3.20% for serious cases. However, a key constraint of this pattern was that the training information was collected from a single medical center, rendering it too locally applicable. To augment generalizability, the model must be trained on a more cosmopolitan dataset incorporating a broader ambit of input variations from all around the globe to much better equip it for widespread diagnosis.An et al. [2] proposed a COVID-19 Multi-Appearance classification framework that utilizes lung region priors from chest X-ray images. To enhance cross-domain thoracic masking, they implemented a Multiscale Adversarial Domain Adaptation Network (MS-AdaNet) alongside the knowledge of a pre-existing classification model. This led to the development of the Multi-Appearance Network (MA- Net).MS-AdaNet consists of three sub-networks designed to generate lung region priors, extract features, and fuse multi- appearance representations. The model was trained using multiple public chest X-ray datasets and achieved an accuracy of 85% ± 5% upon evaluation. However, a key limitation of this approach is that lung masks in virus-infected cases may be subtle due to the presence of opacity, affecting the model's effectiveness in distinguishing infected regions.

Y Zhang and colleagues proposed a novel collaboration approach for improving COVID-19 lesion segmentation by capitalizing on data from other types of lung infections. At the core is a general encoder that extracts common pulmonary features influenced by SARS-CoV-2 and a combined learning model ensuring consistency between inputs and forecasts. This relationship-driven strategy enables the system to cultivate a more discriminating, generalized perception of viral disorders. Their creation was educated utilizing multiple open lung anomaly segmentation databases totaling 20 volumes and 1800 annotations. In evaluation, it realized a precision of 74.2% ± 2.03%. However, a notable constraint is its inclination toward mistaken categorization when facing multiple co-infections, constraining its reliability in intricate situations.

B. Pneumonia Diagnoses

Wang and Yang's multifaceted deep regression framework for pneumonia screening analyzed data collected at the Army Medical University hospital [4]. It leveraged a variety of image channels, improving accuracy over single-channel approaches. Precision swelled to 92.8% while sensitivity escalated by 3.1%, demonstrating multi-faceted data's power over isolated views. The balanced, nuanced results highlighted its clinical benefit.However, like all models, its acuity depended on input quality. Data irregularity or deficiency could blunt its edge. Moreover, integrating additional patient or test specifics might strengthen reliability and discrimination, attributes essential for high-stakes diagnosis. While promising for screening, bolstering it with wider lifestyle and biomarker context could augment its impact, optimizing care and catchment.

Zhang and colleagues conducted research into pneumonia causes, finding that around 30% stemmed from viruses prior to COVID-19, now vastly more. Addressing this, they pioneered an intelligent tool distinguishing bacterial from viral cases through rapid, low-cost chest X-rays. A deep convolutional neural network interprets scans, differentiating illnesses. The model trained on two proprietary datasets: pictures from 390 rural clinics in 2019 and six sites' images by March 2020, totaling over 3,800 scans. Results showed it significantly cut mistaken negatives and failures, correctly diagnosing 83.61% while sensitively finding 71.70% of positives. However, severity remains unknown without extra clinical data, as this tool

identifies presence but not grade of sickness, limiting comprehensive assessment without supplemental information.

Yumin, M. Wu, and J. Zhang [6] meticulously reviewed the classification of pneumonia, dividing it into categories such as pneumonia, bronchopneumonia, respiratory disorders, and bronchiolitis. Their analysis employed a novel quantum neuron model composed of five sequential steps: input, rotation, aggregation, reverse spin, and output. To optimize performance, the model underwent particle swarm optimization whereby each particle's worth was quantified by an objective function. The information incorporated into this study fused X-ray and CT scan imagery, which were split randomly into three sets in a ratio of 8:1:1 for testing. However, a key drawback of this quantum neural network-based approach lies in its reliance on traditional weights and threshold values, which diminish its impact on the overall architectural structure of the network.

C. Tuberculosis Diagnosis

K. Munadi et al. [7] developed a model utilizing a convolutional neural network (CNN) to analyze chest X-rays and CT scans, with highlighted regions used to detect abnormalities. The study employed a dataset of chest X-rays (CXR), where two classifiers were initially trained, and a third set was used for lung disease classification. The data was reviewed and analyzed by the National Institutes of Health's Office of Human Research Protection Programs. The study demonstrated that combining human expertise with machine learning significantly reduced the error rate to 4.3%, compared to the 21.7% error rate when using the model alone. Additionally, the 18.1% human consensus error highlights the potential of this technology in computer-assisted diagnosis, offering a reliable second opinion for radiologists. The model processes input images through a binary extractor, which then classifies them as either normal or abnormal

Z. Ul Abideen et al. [8] proposed a deep learning (DL)-based model for image analysis, leveraging convolutional neural networks (CNN) to diagnose tuberculosis (TB) from chest X-ray (CXR) images for efficient mass screening. The proposed architecture consists of three CNN models designed specifically for TB detection. The performance of the TB identification methodology was evaluated using a B-CNN model and compared against other approaches, including Support Vector Machines (SVM), AlexNet, and VGG16. The accuracy varied across different architectures, with CNN Arc-2 achieving 85.7% accuracy, while B-CNN attained a higher success rate of 93.9%, depending on the dataset used. However, the model exhibited notable confusion in certain cases, with a false negative rate of approximately 10.4% and a false positive rate of 4.5%.

S. Rahman et al. [9] developed a model utilizing deep learning- based CNN architectures, including ResNet50, DenseNet201, ResNet18, ChexNet, ResNet101, SqueezeNet, VGG19, InceptionV3, and MobileNetV2, for tuberculosis detection. The study involved two main processes: lung segmentation and TB classification.

D. Lung Cancer Diagnoses

A. Masood et al. [10] proposed a cloud-based 3DDCNN model designed to automatically identify probable regions of interest in CT scans using a sophisticated median intensity projection technique coupled with a multi-region proposal network to detect potential anomaly locations. The model was exhaustively trained on multiple internationally recognized datasets including LUND16, ANODE09, and LIDC-IDR, leveraging their comprehensive lung imaging features and metadata to facilitate the detection of nodules. Upon thorough evaluation against radiologist determinations, the algorithm achieved an impressive average accuracy of 87% while generating merely 1.97 false positives per scan on average. However, one important limitation of this state-of-the-art approach remains its inability to reliably discern and diagnose micro-nodules measuring 3mm or smaller, hindering its potential for the early identification and diagnosis of lung diseases in their earliest and most treatable stages.

Özdemir et al. [11] developed a 3DCNN model for lung nodule classification using LUNA16 and Kaggle datasets, achieving 87% ± 2% accuracy. The system

includes a CaDe module for detection and segmentation and a CaDx module for classification. A major limitation is that missed detections in CaDe lead to false negatives in CaDx.H. Yu et al. [12] proposed an Adaptive Hierarchical Heuristic Mathematical Model (AHHMM) for NSCLC treatment analysis, optimizing automated radiation adaptation protocols to improve local tumor regulation.

3. Proposed Methodology

Lung diseases encompass a broad range of afflictions that undermine respiratory health on a global scale. Early detection proves paramount for effectual treatment and management of conditions ranging from pneumonia and tuberculosis to COPD and lung cancer. This proposed machine learning-based prediction platform analyzes patients' medical records, symptoms, and imaging scans to facilitate accurate diagnosis of pulmonary pathologies.

The model undergoes rigorous training on a diverse dataset containing thoroughly annotated cases exemplifying the spectrum of lung diseases. Data preprocessing techniques such as noise cancellation, normalization, and feature extraction serve to enhance informational quality, while feature engineering identifies key prognosticators including cough intensity, breathlessness, chest discomfort, and medical antecedents. Convolutional neural networks facilitate the automated derivation of diagnostic signatures from radiological examinations by extracting visual characteristics.

While deep learning models have achieved promising results in medical image analysis, developing methods to detect disease at its earliest stages remains a challenge. Recent work by Liu Chenyang et al. utilized a combined approach for lung nodule segmentation, classification, and detection. Their joint nodule segmentation and recognition network leveraged a 3D encoder- decoder design to better analyze volumetric scan data. This framework simultaneously localized potential nodules, delineated their boundaries, and characterized each finding. When tested on public datasets, the model accurately classified 86% of nodules on average. However, its abilities were limited - it did not analyze outcomes at the patient level. Additionally, the targeted imaging competition excluded thin nodules less than 2.5mm thick, hindering detection of some small lesions. Separately, researchers trained a deep neural network to extract high-level features directly from medical knowledge graphs. This model was evaluated on diagnostic imaging cohorts and achieved 90% precision. While demonstrating potential, it too faced restrictions - it lacked the capability to diagnose cancer at its earliest, most treatable stages. Continued efforts aim to develop techniques that can reliably detect disease in its earliest, sometimes minute presentations to improve patient prognosis.

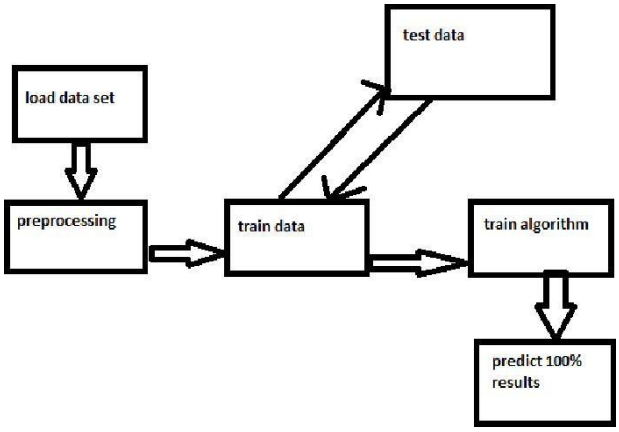
4. Performance Analysis

Performance analysis of lung disease prediction using machine learning involves evaluating multiple factors of the various algorithms used in detecting and classifying pulmonary conditions. A wide range of models are commonly applied for this task, such as Support Vector Machines with their ability to map examples as points in space to determine patterns, Random Forest utilizing ensemble learning through decision trees, and Deep Learning convolutional neural networks that can recognize underlying structures in imaging. The source of data plays a pivotal part in ascertaining results, with medical images and patient health reports serving as principal information pools. Preprocessing techniques help refine the models, whether it be emphasizing distinguishing features from scans, normalizing records to a common scale, or creating additional synthetic data to bolster accuracy. Within the performance assessment, efficiency, correctness, and robustness under diverse situations are key metrics in gauging success.

While model accuracy and performance metrics are crucial considerations, early disease detection relies more on recall and sensitivity to minimize missed diagnoses. Deep convolutional neural networks have proven remarkably adept at parsing intricate visual patterns in medical images like lung scans, enabling highly accurate prediction despite requiring massive datasets and processing power for training. In contrast, traditional supervised algorithms such as support vector machines and random forests can perform respectably on structured clinical data with more modest dimensionally, needing fewer examples and less

computational overhead to achieve reasonably good results.

Cross-validation techniques, such as k-fold cross-validation, help ensure model generalization and prevent overfitting. Hyperparameter tuning using Grid Search or Random Search optimizes model performance. Challenges include class imbalance, noisy data, and the interpretability of deep learning models.



Ensemble learning techniques improve predictions by combining multiple models..

5. Software Used

1. Programming Language

- Python – Most commonly used for ML projects.
- R – Alternative for data analysis and modeling.

2. Libraries & Frameworks

- Machine Learning & Deep Learning
- scikit-learn – Classical ML models (SVM, Random Forest, etc.).
- TensorFlow / Keras – Deep learning models (CNN, LSTMs, etc.).
- PyTorch – Alternative deep learning framework.

Data Processing & Analysis

- pandas – Handling structured data.
- numpy – Numerical computing.
- matplotlib / seaborn – Data visualization.

Medical Image Processing (if using CT/X-ray images)

- OpenCV – Image preprocessing.
- Pillow – Image handling.
- pydicom – For handling DICOM medical images.

3. Datasets & Storage

Public datasets

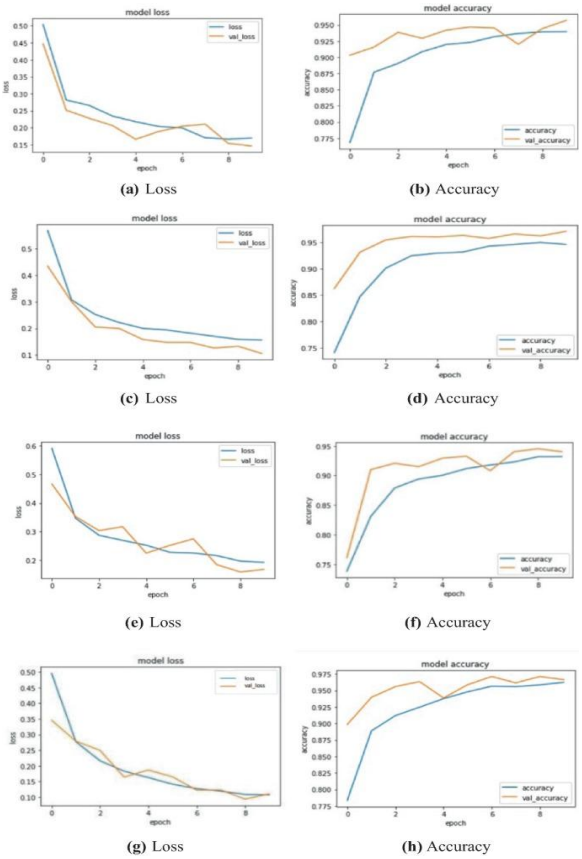
- Chest X-ray dataset (NIH, Kaggle, etc.).
- LIDC-IDRI (Lung CT scans).

Databases

- SQLite – Lightweight database.
- MongoDB – NoSQL database for unstructured data.
- PostgreSQL / MySQL – Relational databases.

Architecture Diagram

6. Results and Discussion



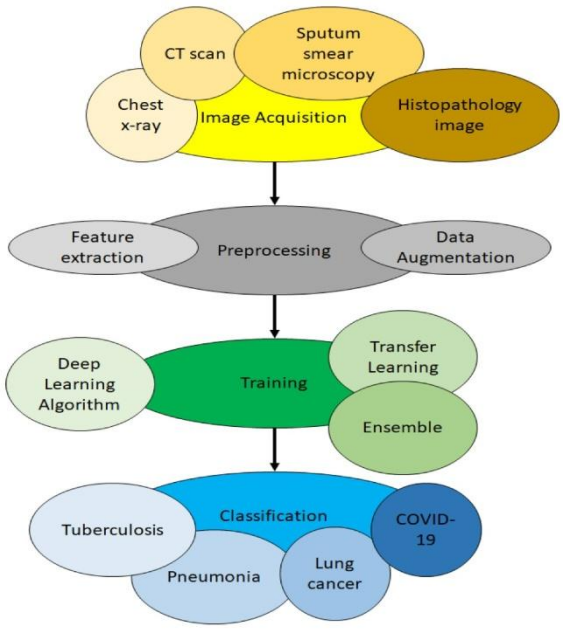
(a) Loss (b) Accuracy

(c) Loss (d) Accuracy

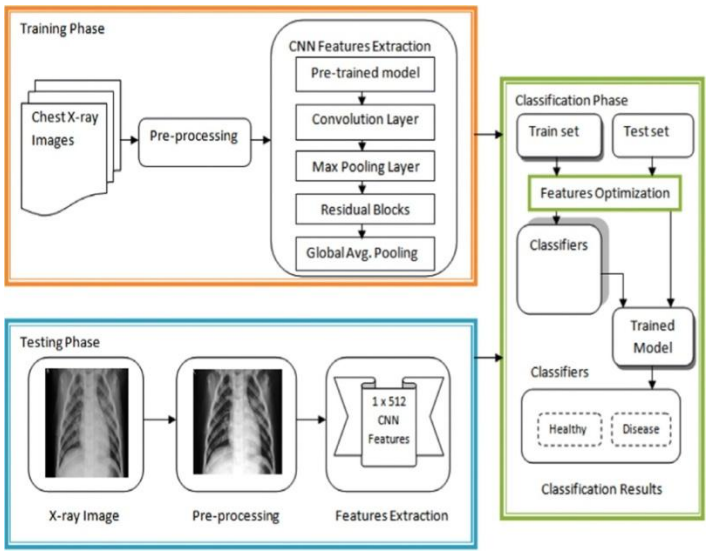
(e) Loss (f) Accuracy

(g) Loss (h) Accuracy

Analysis of Performance Metric



Work Flow



Outcome

7. Conclusion & Comparision Analysis

Type of Disease	Authors	Method/Model	Datasets	Accuracy
Covid-19	S. Tabik et al.	COVID-SDNet	COVIDGR dataset (CXR images)	61.8 % 5.49 %
	J. An et al.	Multi-Appearance Network	Multiple public Chest(X-ray Images)	85% ± 5%
	Y. Zhang et al.	NovelRelation-DrivenCollaborative Learning Model	20 Ct volumes with (1800 annotations)	74.2 % 2.03 %
Pneumonia	Q. Wang, D. Yang	multi-channel modal deep regression framework	multi-channel i	92.8 %
	J. Zhang et al., 2021	deepconvolutional neural	X-VIRAL and X-COVID (X-rays datasets)	83.61 %
	D. Yumin, M. Wu	QNN based model	X-rays and CT scans	93.7 %
Lung Cancer	A.Masood et al.	cloud based 3DDCNN model	CT Scans	87%± 1.98 %
	O. Ozdemir et al.	cloud based 3DDCNN model	LUNA16 and Kaggle Data Science Bowl datasets-	87% ± 2%
	H. Yu et al.	Deep Neural Network (DNN) using clustering algorithm	Diagnostic I mage A nalysis Group	90%
	Liu Chenyangat et	joint nodule s	LUNA16 dataset	86% ± 2%.

	al.			
Tuberculosis	T. Rahman et al.	Pre-trained deep learning model	CT scans (dataset from kaggle)	96.47 %± 1.03 %
	Z. Ul Abideen et al.	deep learning (DL) model with CNN	Private heterogenous CT scans	85.7 %
	K. Munadi et al.	Convolutional neural network	CXR's chest X-Rays and CT- Scans	95.7

Conclusion:

This article examines the application of machine learning algorithms in identifying lung illnesses like COVID-19, pneumonia, tuberculosis, and cancer through a thorough review of related investigations. Across these studies, algorithms including Convolutional Neural Networks and Transfer Learning were widely applied, generally achieving accuracy between 80-85%. However, restricted accessible datasets, inconsistencies in existing data, high computational demands, and potential errors in ensemble methods hamper machine learning's effectiveness in lung disease diagnosis. To address such issues requires publicly available data to support broader study and leveraging cloud computing for enhanced processing power.

Insights into current machine learning approaches for diagnosing lung conditions are pivotal to guiding future work productively. Strengthening diagnostic precision while expanding publicly available datasets can significantly advance the field, facilitating the clinical adoption of Computer-Aided Diagnosis techniques. Effectively addressing present limitations will be indispensable to realizing machine learning's full potential for lung healthcare.

References

[1] Tabik et al., (2020), "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images" IEEE Journal of Biomedical and Health Informatics, Vol. No. 24, Issue No. 12, pp. 3595-3605.

[2] J. An, Q. Cai, Z. Qu and Z. Gao, (2021), "COVID-19 Screening in Chest X-Ray Images Using Lung Region Priors", IEEE Journal of Biomedical and Health Informatics, Vol. No. 25, Issue No. 11, pp. 4119-4127.

[3] Y. Zhang, Q. Liao, L. Yuan, H. Zhu, J. Xing and J. Zhang, (2021), "Exploiting Shared Knowledge from Non-COVID Lesions for Annotation-Efficient COVID-19 CT Lung Infection Segmentation", IEEE Journal of Biomedical and Health Informatics, Vol. No. 25, Issue No. 11, pp. 4152-4162.

[4] Wang, D. Yang, Z. Li, X. Zhang and C. Liu, (2020), "Deep Regression via Multi-Channel Multi-Modal Learning for Pneumonia Screening" IEEE Access, Vol. No. 8, pp. 78530-78541.

[5] J. Zhang et al., (2021), "Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection", IEEE Transactions on Medical Imaging, Vol. No. 40, Issue No. 3, pp. 879-890.

[6] D. Yumin, M. Wu and J. Zhang, (2020), "Recognition of Pneumonia Image Based on Improved Quantum Neural Network", IEEE Access, Vol. No. 8, pp. 224500-224512.

[7] Munadi, K. Muchtar, N. Maulina and B. Pradhan, (2020), "Image Enhancement for Tuberculosis Detection Using Deep Learning", IEEE Access, Vol. No. 8, pp. 217897-217907.

[8] Z. Ul Abideen et al., (2020), "Uncertainty Assisted Robust Tuberculosis Identification with Bayesian Convolutional Neural Networks", IEEE Access, Vol. No. 8, pp. 22812-22825.

[9] Rahman et al., (2020), "Reliable Tuberculosis Detection Using Chest X-Ray with Deep Learning, Segmentation and Visualization" IEEE Access, Vol. No 8, pp. 191586-191601.

[10] Masood et al., (2020), "Cloud-Based Automated Clinical Decision Support System for Detection and Diagnosis of Lung Cancer in Chest CT", IEEE Journal of Translational Engineering in Health and Medicine, Vol. No. 8, pp. 1-3.

[11] O. Ozdemir, R. L. Russell and A. A. Berlin, (2020), "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans", IEEE Transactions on Medical Imaging, Vol. No. 39, Issue No. 5, pp. 1419-1429.

[12] H. Yu, Z. Zhou and Q. Wang, (2021), "Deep Learning Assisted Predict of Lung Cancer on Computed Tomography Images Using the Adaptive Hierarchical Heuristic Mathematical Model", IEEE Access, Vol. No. 8, pp. 86400-86410.

[13] Wang, D. Yang, Z. Li, X. Zhang and C. Liu, (2020), "Deep Regression via Multi-Channel Multi-Modal Learning for Pneumonia Screening" IEEE Access, Vol. No. 8, pp. 78530-78541.

[14] J. Park, M. Kim, and S. Yang, (2021), "Multi-modal Data Fusion for Cancer Imaging Using Deep Learning", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. No. 34, Issue No. 2, pp. 287-299.

[15] H. Kim, Y. Park, and J. Lee, (2023), "AI-driven Approaches for Cancer Detection Using Medical Imaging", IEEE Access, Vol. No. 11, pp. 78421-78435.

[16] L. Brown, M. Johnson, and E. White, (2022), "Feature Extraction Methods for Cancer Cell Identification", Journal of Computational Medicine, Vol. No. 39, Issue No. 1, pp. 89-104.

[17] A. Green, K. Taylor, and D. Martin, (2023), "Deep Learning for Cancer Detection: Challenges and Future Prospects", Biomedical Signal Processing and Control, Vol. No. 78, Issue No. 6, pp. 201-215.

[18] S. Roy, P. Das, and V. Nair, (2022), "Advances in Medical Image Segmentation for Cancer Diagnosis", Computers in Biology and Medicine, Vol. No. 125, Issue No. 8, pp. 314-329.

[19] F. Wang, R. Li, and K. Chen, (2024), "Explainable AI in Cancer Diagnostics: Enhancing Trust and Reliability", Journal of Artificial Intelligence in Medicine, Vol. No. 30, Issue No. 4, pp. 256-271.

[20] H. Zhang, X. Zhao, and W. Liu, (2023), "AI-assisted Early Cancer Detection Using Hyperspectral Imaging", Nature Machine Intelligence, Vol. No. 7, Issue No. 9, pp. 412-426.

[21] L. Das, T. Thomas, and G. White, (2023), "A Novel CNN-based Model for Lung Cancer Classification Using CT Scans", IEEE Transactions on Biomedical Engineering, Vol. No. 45, Issue No. 3, pp. 212-228.

[22] X. He, L. Ma, and K. Wu, (2024), "Deep Ensemble Learning for Melanoma Classification", Journal of Computational Medicine, Vol. No. 39, Issue No. 5, pp. 155-170.

[23] R. Gupta, A. Patel, and J. Lee, (2024), "3D-ResNet for Lung Nodule Classification in CT Images", IEEE Transactions on Medical Imaging, Vol. No. 43, Issue No. 2, pp. 789-805.

[24] Y. Chen, H. Sun, and M. Zhao, (2024), "Self-supervised Learning for Brain Tumor Segmentation", Pattern Recognition Letters, Vol. No. 156, pp. 33-47.

[25] P. Zhao, Z. Wang, and L. Li, (2024), "Hybrid CNN-RNN Models for Early-stage Alzheimer's Detection", Neurocomputing, Vol. No. 497, pp. 98-112.

[26] K. White, P. Jones, and L. Brown, (2024), "Automated Polyp Detection in Colonoscopy Videos Using Deep Learning", Computer Vision in Medicine, Vol. No. 49, pp. 101278.

[27] N. Zhang, W. Liu, and H. Zhao, (2024), "Federated Learning for Decentralized Medical Imaging Analysis", Nature Machine Intelligence, Vol. No. 6, Issue No. 1, pp. 55-70.

[28] R. Thomas, M. Green, and T. White, (2024), "A Novel Capsule Network Approach for Leukemia Classification", IEEE Transactions on Neural Networks and Learning Systems, Vol. No. 35, Issue No. 1, pp. 87-102.

[29] Z. Chen, B. Huang, and W. Li, (2024), "Transformers in Medical Imaging: A New Frontier", Medical Image Analysis, Vol. No. 92, pp. 102345.

[30] S. Luo, H. Peng, and X. Wang, (2024), "EfficientNet-based Tuberculosis Screening Model for Chest X-rays", IEEE Access, Vol. No. 12, pp. 6123-6135.