A DEEP LEARNING AND EXPLAINABLE AI FRAMEWORK FOR LUNG CANCER DIAGNOSIS USING CT SCAN IMAGES

Kallukuri Vyshnavi
PG Scholar
Department of Computer Science & Engineering
JNTUA College of Engineering (Autonomous)
Ananthapuramu, Andhra Pradesh, India
kallukurivyshnavi@gmail.com

Dr. A. P. Siva Kumar
Professor
Department of Computer Science & Engineering
JNTUA College of Engineering (Autonomous)
Ananthapuramu, Andhra Pradesh, India
sivakumar.cse@jntua.ac.in

Abstract: The lungs play a vital role in the respiratory system, and detecting lung cancer at an early stage is important for improving survival rates and treatment effectiveness. The traditional approach to diagnosing lung cancer typically depends on radiologists manually interpreting images obtained from computed tomography (CT) scans, which may require significant time and is susceptible to human error. In order to tackle these issues, this study introduces an automated lung cancer diagnosis framework based on deep learning and Implementation of Explainable Artificial Intelligence (XAI) methods.

The proposed system employs a MobileNet architecture along with an LSTM (Long Short-Term Memory) network to learn feature representations and the model classifies CT images into three predefined classes: benign, malignant, and normal. To improve model interpretability, an Explainable Artificial Intelligence (XAI) component employing Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated to create visual heatmaps that reveal the key regions responsible for the model's predictions. The proposed framework improves transparency and interpretability, increasing clinician confidence in AI-driven diagnostic systems.

***Keywords*: Lung cancer Diagnosis, MobileNet, LSTM, Explainable AI, Grad-CAM, CT.**

1. INTRODUCTION

Lung cancer is among the most common causes of cancer-related mortality worldwide, resulting in a substantial number of deaths each year. The World Health Organization identifies it as a major public health challenge, primarily due to its high death rate and the fact that many cases are detected only at advanced stages [1]. Early recognition of lung cancer contributes substantially to better patient prognosis by enabling prompt diagnosis, informed treatment decisions, and improved therapeutic outcomes [2]. Nevertheless, early-stage lung cancer detection continues to be a difficult task because the abnormalities present in medical images are often subtle and complex, making them challenging to identify accurately.

Computed Tomography (CT) scans are extensively utilized in lung cancer screening and diagnosis because they offer detailed cross-sectional images of the lung structures. Despite its effectiveness, traditional diagnosis relies heavily on radiologists to manually interpret CT images. This process is computationally demanding in terms of time and effort and is also affected by inter-observer variability and diagnostic errors, especially when large imaging datasets are evaluated [3]. As a result, there is an increasing demand for automated diagnostic solutions that can support healthcare professionals in delivering accurate and timely clinical decisions.

Recent advances in deep learning have greatly enhanced the effectiveness and accuracy of medical image analysis systems. Convolutional Neural Networks (CNNs) have achieved outstanding performance in various computer vision applications, including image classification, image segmentation, and object detection [4]. Pre-trained networks such as MobileNet are commonly utilized because of their efficient design, which enables the extraction of meaningful high-level features while requiring relatively low computational resources [5]. These architectures are especially valuable in medical imaging applications, where achieving high diagnostic accuracy while maintaining computational efficiency is essential.

Moreover, Long Short-Term Memory (LSTM) networks are capable of modeling sequential data by capturing complex patterns and long-term dependencies present within the data [6]. By integrating CNN-based feature extraction with LSTM networks, hybrid architectures can leverage both spatial and sequential information, thereby improving classification performance in medical imaging tasks [7].

Although deep learning models can deliver high classification performance, their limited interpretability often reduces confidence in their use within clinical settings. This “black-box” nature poses a significant barrier to their adoption in healthcare systems. To address this challenge, Explainable Artificial Intelligence (XAI) methods have been developed to enhance the transparency and interpretability of model predictions [8]. Among various XAI approaches, Gradient-weighted Class Activation Mapping (Grad-CAM) is widely adopted for generating visual explanations by emphasizing the image regions that have the greatest impact on the model's predictions [9].

In response to these challenges, this study presents a deep learning-driven framework for lung cancer diagnosis from CT scan images, enhanced with explainable artificial intelligence techniques to improve interpretability. The proposed framework utilizes a MobileNet-based model to efficiently extract relevant features from CT images, while an LSTM network performs classification into three categories: benign, malignant, and normal. Furthermore, a Grad-CAM-based explainability module is integrated into the system to produce heatmaps that identify the most influential regions responsible for the model's predictions. This methodology not only increases diagnostic performance but also improves model transparency, fostering greater confidence among clinicians in AI-supported healthcare systems.

In summary, the proposed framework seeks to combine the strong predictive capabilities of deep learning with enhanced interpretability, supporting faster, more reliable, and accurate clinical decision-making in medical diagnosis [10].

2. RELATED WORKS

Mohamed Hammad et al. [11] introduced a lightweight custom CNN framework combined with Explainable Artificial Intelligence (XAI) methods to facilitate lung cancer detection from CT scan images. The framework used Gradient-weighted Class Activation Mapping (Grad-CAM) to generate visual heatmaps that reveal the image regions driving the model's predictions, providing greater interpretability and fostering increased confidence among clinicians in AI-assisted diagnosis. The system also employs advanced data augmentation and regularization methods to enhance generalization, achieving an accuracy of approximately 93%. However, the study is limited by a relatively small dataset, lack of external validation, and increased computational complexity. Future enhancements may involve validating the model on larger and diverse datasets while exploring hybrid architectures integrated with automated segmentation techniques to enhance diagnostic performance.

Md. Imaran Hossain et al. [12] proposed a hybrid deep learning framework that integrates the LeNet-5 Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) networks to classify lung cancer from CT scan images. In this framework, the CNN extracts discriminative features from CT images, whereas the LSTM learns sequential dependencies within the extracted features, leading to improved classification accuracy. The proposed framework performs multi-class classification of CT scan images into benign, malignant, and normal categories, providing support for the early diagnosis of lung cancer by radiologists. Despite its effectiveness, the model exhibits variability across different clinical datasets and requires high-performance GPU resources for training. The authors recommend exploring ensemble deep learning approaches and integrating XAI techniques to enhance interpretability and robustness.

Multi-modal deep learning approaches have also been explored to enhance diagnostic performance. Mohamed Hosny et al. [13] developed an attention-enhanced Inception-ResNet architecture that combines information from multiple medical imaging modalities, including X-ray images, CT scans, and histopathological data. The model incorporated attention mechanisms and multi-scale feature extraction to focus on discriminative regions, achieving high accuracy across different datasets (up to 99.73%). While this approach demonstrated superior performance, it introduced increased computational complexity and required large-scale multi-modal datasets, which may not always be available in real-world clinical settings.

A predictive modeling approach was introduced by Abdulrahman Alzahrani [14] to overcome class imbalance in lung cancer datasets through the use of Conditional Tabular Generative Adversarial Networks (CTGAN) for synthetic data generation and dataset balancing. The proposed method integrates CTGAN with traditional machine learning classifiers, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), and reports a classification accuracy of about 98.93%. Although the model effectively improves performance on imbalanced datasets, it relies on structured data rather than imaging data and lacks interpretability. Future research directions highlight the importance of utilizing larger and more diverse datasets while incorporating explainable AI techniques to improve model transparency and reliability.

Mohamed and Ezugwu [15] developed a deep learning framework that utilizes multi-omics information, incorporating both mRNA and DNA features, to enhance lung cancer classification. The framework applies Principal Component Analysis (PCA) to reduce feature dimensionality, employs the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, and utilizes a CNN for classification, resulting in an accuracy of nearly 97%. Despite providing valuable biological insights, the approach is limited by the availability of multi-omics datasets, noise in public data, and the black-box nature of CNN models. The authors suggest incorporating explainable AI and transfer learning techniques in future work.

Beyond classification, recent studies have increasingly emphasized the lung cancer detection and staging. Alaa Wehbe et al. [16] developed a comprehensive deep learning framework that integrates YOLOv8 for identifying lung cancer subtypes and a customized TNM classification model for determining cancer stages. The detection framework attained a mean Average Precision (mAP) of 97.1%, indicating strong performance and suitability for real-time clinical deployment. In addition, the TNM staging component achieved a classification accuracy of 98%, demonstrating the strong performance and reliability of the integrated framework. Although the model demonstrates high performance, it does not provide sufficient interpretability, which can make it difficult for clinicians to understand the reasoning behind its predictions and confidently rely on its outcomes. Future improvements should focus on incorporating explainable AI techniques to enhance transparency and clinical reliability.

3. PROBLEM STATEMENT

Lung cancer continues to be a major cause of death worldwide, largely because of the difficulties associated with its early and accurate detection. Despite the widespread use of CT imaging for lung cancer assessment, its manual evaluation by radiologists can be labor-intensive, subject to variability, and susceptible to diagnostic errors. Existing deep learning approaches have improved classification performance however, their lack of interpretability limits their adoption in clinical practice. Therefore, there is a need for an automated and reliable framework that can accurately classify lung cancer from CT images while providing explainable insights to support clinical decision-making and enhance trust in AI-based diagnostic systems.

4. EXISTING SYSTEM

Manual evaluation of Computed Tomography (CT) images by radiologists remains the primary approach for diagnosing lung cancer. However, the process is resource-intensive, time-consuming, and susceptible to human error, particularly when interpreting a high volume of medical images.

To address these challenges, existing systems have transitioned from traditional machine learning approaches with handcrafted features to advanced deep learning techniques, especially Convolutional Neural Networks (CNNs), have facilitated automatic feature learning from images, leading to enhanced classification accuracy and overall performance. Nevertheless, despite their effectiveness, these models often lack transparency and interpretability in their decision making process.

Disadvantages:

1. Lack of Interpretability: Many deep learning-based models lack the ability to provide clear and understandable explanations for their predictions, which reduces their reliability and acceptance in clinical practice.

2. Risk of Human Error: Manual interpretation of CT images can be subjective and variable, increasing the likelihood of misdiagnosis, particularly with large datasets.

5. PROPOSED METHODOLOGY

This research presents an automated lung cancer detection framework that combines deep learning methods with explainable artificial intelligence techniques for the analysis of CT scan images. The proposed approach follows a structured pipeline that includes preprocessing the input data, extracting deep features with a pre-trained MobileNet model, performing classification using an LSTM network, and generating visualization through Gradient-weighted Class Activation Mapping (Grad-CAM). Fig.1 shows the proposed methodology workflow.

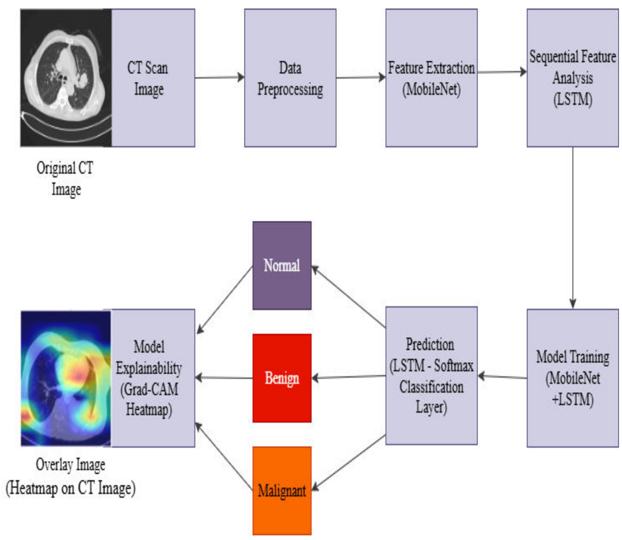

Fig-1: Proposed Methodology Workflow

5.1 Dataset

The Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) dataset was obtained during of 2019 was taken from kaggle website. It consists of CT scan images from 110 patients, including both lung cancer cases and healthy individuals. A total of 1,097 CT image slices are included in the dataset, grouped into three classes**:**

i. Normal - 416
ii. Benign - 120
iii. Malignant - 561

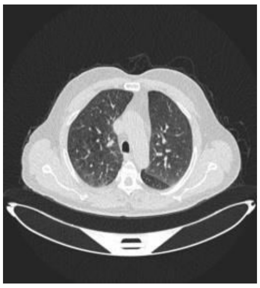

Fig-2: Normal

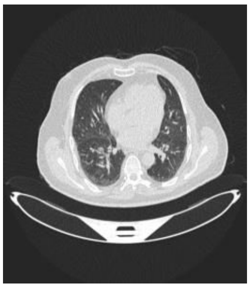

Fig-3: Benign

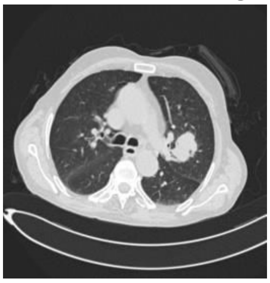

Fig-4: Malignant

5.2 Data Preprocessing

Data preprocessing plays a vital role in preparing CT scan images, ensuring that they are appropriately formatted and optimized for efficient deep learning model training. The following preprocessing steps are applied:

Image Resizing: All CT scan images are standardized by resizing them to a uniform resolution of 224 × 224 pixels. This is necessary because the pretrained MobileNet model requires inputs of a consistent size. Resizing reduces the computational complexity while preserving essential image features.

Dataset Splitting: An 80:20 ratio is employed to divide the dataset into training and testing portions.

Class Label Encoding: The class labels (benign, malignant, and normal) are encoded into numerical values and further converted into one-hot encoded vectors to enable multi-class classification for softmax classification layer.

Class	Label	One-Hot Vector
Benign	0	[1,0,0]
Malignant	1	[0,1,0]
Normal	2	[0,0,1]

Table-1: Class Label Encoding values

5.3 Model Development

5.3.1 Feature Extraction using MobileNet

Feature extraction converts raw CT scan images into meaningful representations for classification. In this study, a pretrained MobileNet model is used due to its lightweight architecture and computational efficiency. The architecture employs depthwise separable convolutions to minimize computational complexity while preserving strong performance.

The model is initialized with pre-trained ImageNet weights, and the top classification layers are excluded (include_top=False), allowing the network to serve as a feature extractor through transfer learning. All layers are frozen during training to retain learned features and to prevent overfitting of the model.

The extracted feature maps are processed using a Global Average Pooling layer, which converts them into a fixed-length feature vector. This vector captures essential spatial information while reducing the number of parameters, providing an effective input for the subsequent LSTM-based classification layer.

5.3.2 Sequential Feature Learning using LSTM

The MobileNet feature extractor is combined with a Long Short-Term Memory (LSTM) network to improve the representation and learning of extracted features. While MobileNet effectively captures spatial features from CT images, it lacks the ability to model dependencies among extracted features. Therefore, an LSTM layer is introduced to learn sequential relationships within the feature space.

The pre-trained MobileNet network, configured with frozen weights and without its final classification layers, is employed to extract deep image features. These features are subsequently transformed into a one-dimensional feature vector using a Global Average Pooling layer. This feature vector is reshaped into a sequential format to serve as input to the LSTM network.

The reshaped feature vectors are processed by an LSTM layer with 128 units, enabling the network to capture meaningful feature relationships that enhance class separability and improve multi-class classification performance. The feature vector obtained from the LSTM layer is provided as input to a dense layer equipped with a softmax activation function, enabling accurate classification across multiple classes. The MobileNet–LSTM framework leverages the advantages of both convolutional feature extraction and sequential pattern learning, leading to enhanced performance in lung cancer classification.

```python
def create_mobilenet_lstm_model():
    print("Building MobileNet + LSTM model...")

    base_model = MobileNet(weights='imagenet', include_top=False, input_shape=(224, 224, 3))
    print("MobileNet base model loaded.")

    # Freeze the layers of MobileNet for transfer learning
    for layer in base_model.layers:
        layer.trainable = False

    # Create the feature extraction layer
    feature_extractor = base_model.output
    feature_extractor = layers.GlobalAveragePooling2D()(feature_extractor)

    # LSTM layer - reshape for sequence input
    lstm_input = layers.Reshape((1, feature_extractor.shape[1]))(feature_extractor)
    lstm_layer = layers.LSTM(128, return_sequences=False)(lstm_input)

    # Fully connected layer for classification
    dense_layer = layers.Dense(3, activation='softmax')(lstm_layer)

    # Create the model
    model = models.Model(inputs=base_model.input, outputs=dense_layer)
    print("Model built successfully!")
```

Fig-5: Model Development Using MobileNet-LSTM

5.3.3 Classification Layer

The features refined by the LSTM layer are subsequently fed into a dense layer, which performs the classification task. The dense layer contains three output nodes, each representing one of the three lung cancer categories. The dense layer processes the high-level features obtained from the LSTM to generate class-specific scores.

The output scores are processed using a softmax activation function to generate normalized probability values for each target class, enabling the final classification decision. The classification layer thus serves as a mapping mechanism between learned features and class probabilities, which are later utilized during training for optimization and during prediction for final decision-making.

Mathematical representation of softmax activation function:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

Where, $P(y_i)$ is the output probability, z_i is the input value of i^{th} neuron and, j value from 1 to 3.

5.4 Model Training

The proposed framework is trained using preprocessed CT scan images standardized to a resolution of 224 × 224 pixels, facilitating efficient model learning. Batch-wise data loading is performed using an ImageDataGenerator, with 80% of the data allocated for training and the remaining 20% for validation. The model is configured with the Adam optimizer using a learning rate of 0.0001 and employs categorical cross-entropy as the loss function, while accuracy is used to assess its performance.

The model is trained for up to 10 epochs. However, it was observed that the model converged early, with validation performance stabilizing around the 7th epoch. To prevent overfitting, an Early Stopping technique monitors the validation loss and restores the weights from the epoch with the highest validation performance. The trained model is subsequently employed for prediction and performance analysis.

5.5 Prediction

After training, the model is used to classify unseen CT images by applying the preprocessing steps. The processed image is passed through the MobileNet–LSTM model, which outputs class probabilities using the softmax function. The final classification result corresponds to the class that receives the maximum predicted probability. The final output generated by the model corresponds to one of the three classes: Benign, Malignant, or Normal.

$$\hat{y} = \arg max_i \; P(y_i|x)$$

Where, $P(y_i|x)$ is the predicted probability of class i of input image x.

The prediction outputs can be further analyzed using Grad-CAM to identify the image regions that most strongly influence the model's decisions.

5.6 Model Explainability using Grad-CAM

To enhance the transparency of the proposed framework, Grad-CAM is incorporated as an interpretability technique that reveals the reasoning behind the model's decisions. This technique determines the regions of lung CT images that have the greatest impact on the predicted class by examining the gradients propagated through the convolutional layers of the MobileNet feature extractor. The gradients are averaged to compute significance weights, which are then applied to the corresponding feature maps to produce a localization heatmap. This generated heatmap visually identifies the areas of the image that contribute most significantly to the classification outcome.

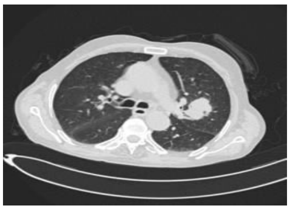

Fig-6: Original CT Scan Image

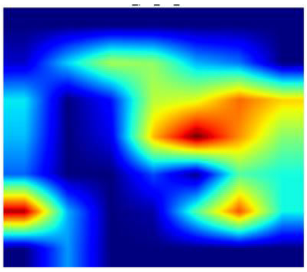

Fig-7: Grad-CAM Generated Heatmap

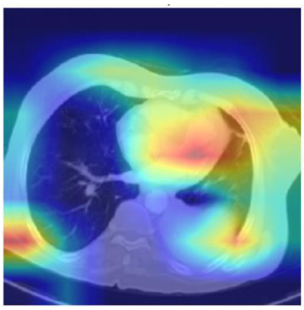

Fig-8: Overlay Image

After resizing, the generated heatmap is overlaid onto the original CT scan, providing a visual representation of the areas that influenced the model's prediction. These explanations allow medical professionals to validate whether the model is attending to meaningful lung regions, thereby improving model transparency and supporting reliable clinical decision-making.

6. RESULTS AND DISCUSSION

A deep learning model combining MobileNet and LSTM has been proposed for the classification of lung cancer using CT scan images with Grad-CAM visual explainability. The MobileNet architecture effectively extracts deep spatial features, while the LSTM layer enhances feature representation by capturing dependencies among the extracted features. This integrated approach allows the model to deliver strong performance in the multi-class classification of lung conditions i.e., normal, benign, and malignant cases.

The confusion matrix serves as an important evaluation tool, offering class-wise insights into prediction outcomes through TP, TN, FP, and FN metrics. Higher values concentrated on the diagonal indicate that the model has accurately classified more samples.

The model achieves high reliable performance in recognizing and classifying malignant cases. Similarly, the normal class is accurately predicted, with only a few misclassifications. However, some misclassification is observed in the benign class, where some of the benign images are incorrectly classified as normal.

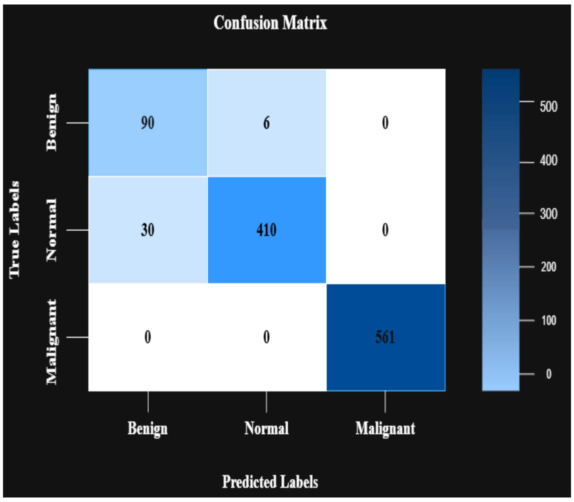

Fig-9: Confusion Matrix (Benign, Normal, Malignant Cases)

The experimental findings show that the proposed model attains a training and validation accuracies of 93% and 85%, reflecting its strong learning ability and effective generalization. The model is highly effective in identifying malignant cases, supporting early lung cancer detection and timely medical intervention. Although some misclassifications occur between benign and normal classes due to their similar visual characteristics, the overall performance remains reliable.

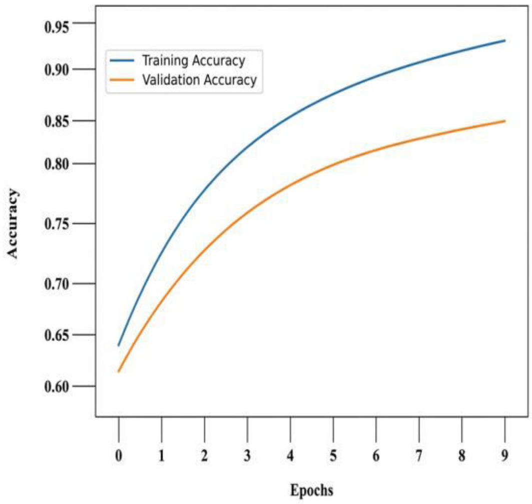

Fig-10: Training vs Validation Accuracy Graph

Model Interpretability: To enhance transparency, Grad-CAM visualization is applied to highlight important regions influencing model predictions. The heatmaps demonstrate that the model focuses on relevant lung areas, improving trust in its outputs.

7. CONCLUSION

The proposed Framework for Lung Cancer Diagnosis is effectively extracts deep spatial features using MobileNet, while the LSTM layer enhances feature representation by capturing dependencies among the extracted features. This hybrid architecture contributes to effective and robust multi-class classification of lung conditions. The obtained results indicate that the model achieves **93%** training accuracy and **85%** validation accuracy, reflecting its strong learning capability and ability to generalize effectively. The proposed model shows excellent capability in recognizing malignant cases, thereby supporting early lung cancer detection and enabling timely therapeutic intervention.

Furthermore, the integration of Grad-CAM improves the interpretability of the model by highlighting the regions of CT images that influence the predictions. This adds transparency to the decision-making process, which is essential for medical Professionals. Future work can focus on classifying different stages of lung cancer (Stages I–IV) for more detailed clinical analysis.

REFERENCES

[1] World Health Organization, “Cancer Fact Sheet,” 2023.

[2] Rebecca L Siegel, Kimberly D Miller, Ahmedin Jemal, S. H. Takacs, “Cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, 2020.

[3] Armato et al., “LIDC-IDRI: A public lung image database,” *Medical Physics*, 2011.

[4] Krizhevsky et al., “ImageNet classification with deep CNNs,” *NeurIPS*, 2012.

[5] Andrew G. Howard, Menglong Zhu, Bo Chen, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", 2017.

[6] Hochreiter & Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, 1997.

[7] Shin et al., “Deep CNNs for medical image analysis,” *IEEE TMI*, 2016.

[8] Gunning, “Explainable Artificial Intelligence (XAI),” DARPA, 2017.

[9] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", *International Journal of Computer Vision*, vol. 128, 2020.

[10] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, Clara I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis", vol. 42, 2017.

[11] Mohamed Hammad, Mohammed ElAffendi, Ahmed A. Abd El-Latif, Abdelhamied A. Ateya, Gauhar Ali, Pawel Plawiak, “Explainable AI for lung cancer detection via a custom CNN on CT images”, *scientific reports*, Apr. 2025.

[12] Md. Imaran Hossain , Shoaib Ahmad, Sakib Hasan, “Lung cancer detection and classification using LeNet-LSTM model on computed tomography images”, Multidisciplinary Science Journal, Oct. 2024.

[13] Mohamed Hosny, Ibrahim a. elgendy, Mousa ahmad Albashrawi, "Multi-Modal Deep Learning for Lung Cancer Detection Using Attention-Based Inception-ResNet", IEEE, vol. 13, 2025.

[14] Abdulrahman Alzahrani, “Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning”, IEEE, vol.13, Feb. 2025.

[15] Tehnan I. A. Mohamed, Absalom EL-Shamir Ezugwu, “Enhancing Lung Cancer Classification and Prediction With Deep Learning and Multi-Omics Data”, IEEE, vol. 12, Apr. 2024.

[16] Alaa Wehbe, Silvana Dellepiane, Irene Minetti, "Enhanced Lung Cancer Detection and TNM Staging Using YOLOv8 and TNMClassifier: An Integrated Deep Learning Approach for CT Imaging", IEEE, vol. 12, Sep. 2024.