MediScan LungGuard: Lung Cancer Detection Using Machine Learning

Rupashri Barik [1], Md Kaif Ansari [2], Khushi Kumari[3], Manjit Singh[4]
[1,2,3,4] *JIS College of Engineering, Kalyani, W. B., INDIA*
[1]*rupashri.barik@jiscollege.ac.in,* [2]*ansarimdkaif0@gmail.com,*
[3]*khushi80203@gmail.com,* [4]*manjitsinghjii1999@gmail.com*

Abstract:

Lung cancer is a prevalent and deadly disease that origin in the lungs. Early detection is crucial for saving lives, but mortality rates vary by gender and country. Raising awareness and encouraging screenings are essential to reduce these fatalities. However, early detection of lung cancer is challenging. Machine learning (ML) offers a promising solution by analyzing medical images, such as chest X-rays or CT scans, to detect suspicious lesions or nodules. These ML approaches can support radiologists in early diagnosis, potentially improving patient outcomes. The proposed method includes image enhancement, segmentation, and feature extraction to highlight areas of concern within lung images. Machine learning algorithms learn to identify patterns associated with lung cancer, enabling precise and automated detection. Results show high accuracy in identifying lung cancer, indicating that this method could be a valuable tool for healthcare professionals. The simplicity and efficiency of this approach make it a practical solution for early diagnosis, facilitating timely intervention and better patient care in the fight against lung cancer.

Keywords: *Lung cancer, Machine learning, Medical imaging, Image segmentation, Feature extraction, Image Classification*

1. Introduction:

Detecting lung cancer [1][2][3] using machine learning (ML) techniques is a vital area of research in healthcare. Lung cancer is a common and deadly disease worldwide, making early detection crucial for improving patient outcomes. ML methods offer a promising approach for the timely and accurate diagnosis of lung cancer by analyzing medical imaging and other clinical data. This paper explores the use of machine learning algorithms to detect lung cancer using various types of data, such as CT scans [5], X-rays [3], patient demographics, and potentially genomic information. The goal is to develop robust models that help healthcare professionals identify suspicious lesions or patterns that indicate lung cancer, facilitating early intervention and treatment planning.

This introduction provides an overview of the importance of early lung cancer detection, the challenges involved, and the potential of machine learning to address these challenges. It delves into different ML approaches, including deep learning methods, and discusses how these techniques can be used to achieve accurate and efficient lung cancer detection. Additionally, it highlights key datasets, model architectures, evaluation metrics, and ethical considerations associated with implementing ML-based systems in clinical settings.

The aim of this research is to advance lung cancer diagnostics by developing reliable ML models that complement existing screening methods and improve patient outcomes. By leveraging the power of machine learning, this study aims to enhance the efficiency and effectiveness of lung cancer detection, ultimately saving lives through early intervention and personalized treatment strategies. The project uses computers to help doctors detect lung cancer early by analyzing medical images and identifying potential problem areas. This is achieved through image processing, enabling computers to analyze images more effectively and identify areas of concern. Smart algorithms and machine learning techniques assist in detecting early-stage lung cancer patterns, creating a faster and more accurate diagnostic tool for doctors. This study aims to improve early detection through advanced image processing and the development of a more effective detection tool by exploring existing methods and technologies.

2. Background study:

Lung cancer detection is crucial for early treatment and better survival rates. Traditional methods include imaging techniques like X-rays and CT scans [4], which help doctors see the lungs' structure and identify suspicious areas. However, these methods can sometimes miss small details. Machine learning (ML) can enhance lung cancer detection by analysing vast amounts of data quickly and accurately.

ML algorithms can be trained to recognize patterns in medical images, such as nodules or tumors, that might indicate cancer. Techniques like convolutional neural networks (CNNs) are particularly effective for image analysis. These models learn to identify features in lung scans that differentiate between healthy and cancerous tissues.

Traditional imaging methods like X-rays are used for detecting lung nodules or masses, though they may not always provide detailed information about the nature of the lesions. Computed Tomography (CT) scans offer higher resolution than X-rays, allowing for better visualization of lung structures and abnormalities, and are commonly used for lung cancer screening and staging. Positron Emission Tomography (PET) scans, often combined with CT (PET-CT), use radioactive tracers to detect metabolic activity in tissues, helping identify areas of increased metabolic activity that may indicate cancerous lesions. Biopsy techniques, including Fine Needle Aspiration (FNA), Core Needle Biopsy, Bronchoscopy, and Surgical Biopsy, involve extracting cells or tissue samples for examination under a microscope. These methods provide varying degrees of comprehensiveness and invasiveness, from using thin needles to flexible tubes with cameras or direct surgical sampling. Cytological techniques such as Sputum Cytology and Bronchial Brushing and Washing involve examining cells collected from the airways or sputum under a microscope to detect abnormal changes suggestive of lung cancer. Molecular and genetic testing techniques like Next-Generation Sequencing (NGS), Fluorescence in Situ Hybridization (FISH), and Immunohistochemistry (IHC) allow for comprehensive analysis of genetic mutations, specific genetic abnormalities, and protein expressions in lung cancer cells, aiding in precise diagnosis and classification. Advancements in these detection techniques, particularly in molecular and genetic testing, have led to more precise diagnoses and personalized treatment approaches for lung cancer patients. Ongoing research continues to explore new technologies and methodologies to improve the accuracy and efficiency of lung cancer cell detection.

Convolutional Neural Network - A **Convolutional Neural Network (CNN)** is a type of Deep Learning neural network architecture commonly used in Computer Vision. Computer vision is a field of Artificial Intelligence that enables a computer to understand and interpret

the image or visual data. CNN [4] is the extended version of Artificial Neural Network (ANN) which is predominantly used to extract the feature from the grid-like matrix dataset. For example, visual datasets like images or videos where data patterns play an extensive role.

3. Proposed Methodology:

This section describes the proposed algorithm for detecting lung cancer cells. Fig 1 depicts the essential block diagram of the proposed model. It showcases the interconnected modules, starting from input sources, progressing through critical stages, and culminating in valuable outputs. This visual representation underscores the seamless integration of components and fostering efficiency. Overall, Fig 1 encapsulates the cohesive essence of the model's execution.

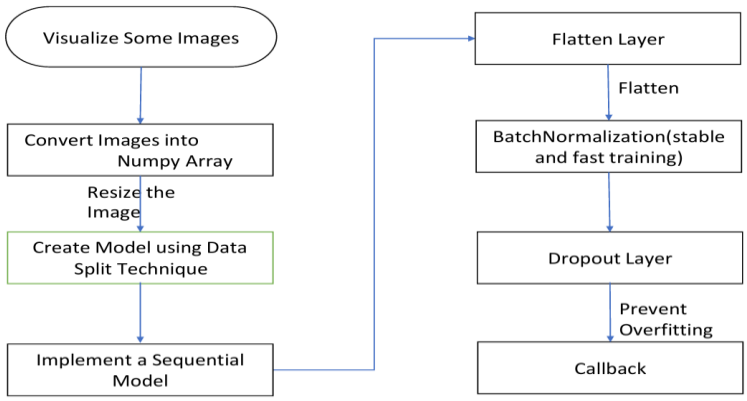

Figure 1. Block Diagram

Firstly, the process begins by examining CT scan images. These images are then converted into a data format known as a Numpy array. After this, the images are resized to a specific size. The next step involves using a technique called data splitting to create a model. This model is designed in a sequential manner, meaning it processes data in a specific order. To ensure stable and quick training, a layer called batch normalization is added. This layer helps in standardizing the inputs to the model, making training more efficient. Additionally, a dropout layer is included in the model. The dropout layer randomly ignores some connections in the neural network during training, which helps prevent overfitting. Overfitting occurs when the model becomes too focused on the training data and performs poorly on new, unseen data. Finally, callback is used. A callback is a function that monitors the model during training and performs certain actions, such as saving the best model or stopping training early if performance metrics meet specific criteria. Overall, these steps are part of how the new model operates to analyze CT scans effectively and efficiently.
The following algorithm outlines the module:

1. Visualize some images that have been provided to build the classifier for each class.
2. Convert the given images into Numpy arrays of their pixels after resizing them using the Numpy and Open-CV libraries.
3. Create a model to predict the highest probability class of an image by splitting the data using the train data split technique.
4. Use the TensorFlow library to build the CNN model. Implement a sequential model that contains three convolutional layers, followed by a flattened layer.
5. Add two fully connected layers along with the output from the flattened layer.

6. Include batch normalization to enable stable and fast training, along with a dropout layer to prevent overfitting.
7. Use a callback function to check whether the model is improving during training.

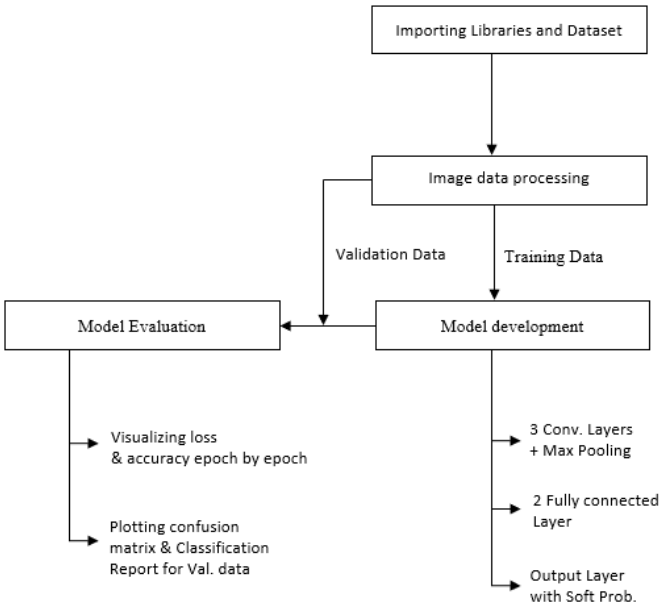

Figure 2. Process Flowchart

Fig 2 illustrates the proposed algorithm and how the proposed system will function. First, import the dataset containing lung cancer images and related information. Next, visualize the data using graphs and images to understand its distribution and characteristics. Then, prepare the data for training by cleaning it, resizing the images, and converting them into a suitable format. Set the hyperparameters, which control the training process, such as learning rate, batch size, and number of epochs. Develop the model by defining its architecture, specifying the different layers and how they connect. Compile the model by selecting the optimizer, loss function, and evaluation metrics to be used. Implement callbacks to monitor the model's progress and make adjustments as needed during training. Split the data into training and validation sets to ensure a proper evaluation of the model's performance. Proceed with training the model using the training data, and visualize the training history to track improvements and detect any issues. Finally, evaluate the model's accuracy and overall performance using the validation set to ensure it can correctly identify lung cancer.

Importing Dataset - The dataset used in this study has been sourced from Kaggle [7]. The dataset comprises 25,000 images categorized into three classes representing different lung conditions.

- Normal Class
- Lung Adenocarcinomas
- Lung Squamous Cell Carcinomas

The images for each class have been generated from 25,000 original images using Data Augmentation techniques. As a result of this augmentation, no further Data Augmentation methods will be applied to these images during the analysis.

Python is used to develop the model, with libraries such as Panda, Numpy, Matplotlib, Sklearn, OpenCV, and TensorFlow playing crucial roles in the process.

Image data processing

Here, the given images are converted into NumPy arrays of their pixels after resizing. This is because training a Deep Neural Network on large-size images is computationally expensive and time-consuming. The OpenCV and NumPy libraries in Python are used for this task. After converting the images into the desired format, they are divided into training and validation data sets. This division allows for the assessment of the model's performance. Adjustments to some hyperparameters can be made throughout the entire notebook based on this data preparation step.

Hyperparameters:

In this study, several key hyperparameters for training the lung cancer detection model have been used. The images were resized to 256x256 pixels using OpenCV's `cv2.resize()` function, ensuring uniformity in image dimensions. Here, set aside 20% of the dataset for validation, with the remaining 80% used for training. The model was trained over 10 epochs, meaning it made ten complete passes through the entire training dataset. Additionally, a batch size of 64 was used, meaning the model's weights were updated after processing every 64 images.

Model Development

Starting from this step, the TensorFlow library will be used to construct the CNN model. The Keras framework within the TensorFlow library provides all the necessary functionalities to define the architecture of a Convolutional Neural Network (CNN) and train it using the data. A Sequential model to be implemented, consisting of three convolutional layers followed by max-pooling layers, a flatten layer to flatten the output of the convolutional layer, and two fully connected layers following the output of the flattened layer. BatchNormalization layers will be used for stable and fast training, along with a dropout layer to prevent overfitting. The final layer, known as the output layer, will provide soft probabilities for the three classes. The final Dense layer has 3 units (assuming it's a multi-class classification task with 3 classes) and uses the softmax activation function to output probability scores for each class.

Model Evaluation

After defining the model architecture, compile the model by specifying the optimizer, loss function, and metrics. Here the optimization algorithm is used during training (e.g., Adam). The loss function is used to compute the model's error (categorical cross-entropy for multi-class classification). The evaluation metrics to monitor during training and validation (e.g., accuracy). Callbacks are used to check whether the model is improving with each epoch or not. If not, then what are the necessary steps to be taken like ReduceLROnPlateau decreases learning rate further. Even then if model performance is not improving then training will be stopped by EarlyStopping. Some custom callbacks can also be defined to stop training in between if the desired results have been obtained early.

4. Results & Discussion:

The results of this study highlight the importance of early detection in combating lung cancer, a prevalent and deadly disease. The findings underscore the variability in mortality rates across genders and countries, emphasizing the need for targeted awareness campaigns and screening initiatives. The challenges of early detection are addressed through machine learning (ML), which analyzes medical images like chest X-rays or CT scans to pinpoint suspicious lesions indicative of lung cancer. By enhancing images and extracting key

features, ML algorithms facilitate precise and automated detection, supporting radiologists in improving patient outcomes.

The proposed method demonstrates promising accuracy in identifying lung cancer, showcasing its potential as a valuable tool for healthcare professionals. The simplicity and efficiency of this ML approach make it practical for early diagnosis, enabling timely intervention and enhanced patient care. Overall, the research contributes to advancing lung cancer diagnostics, paving the way for improved detection rates and ultimately, saving lives in the battle against this devastating disease.

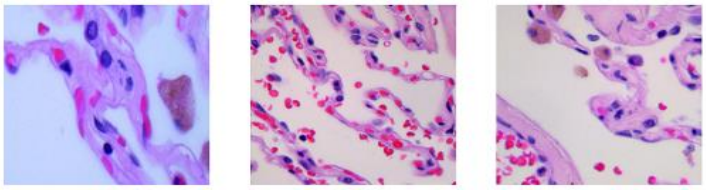

Figure 3. Images for lung_n category

The images shown in Figure 3 represent the lung_n category, which is one of the classes used in the study to train the machine learning model for lung cancer detection. These images capture various aspects of lung pathology associated with cancerous conditions. Upon close examination of the images, several key features stand out. For instance, certain images exhibit distinct patterns indicative of tumor growth or nodules within the lung tissue. These patterns may include irregular shapes, increased opacity, or defined edges, all of which are common characteristics observed in lung cancer cases.

Additionally, the images depict a range of severity levels, from early-stage lesions to more advanced tumor formations. This diversity in image characteristics underscores the complexity of lung cancer diagnosis and the importance of accurate detection methods.

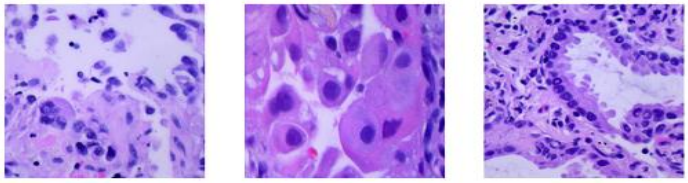

Figure 4. Images for lung_aca category

Figure 4 represents the lung_aca category, showcasing a distinct set of images related to lung adenocarcinoma (lung_aca), a common type of lung cancer. These images provide a visual representation of the histopathological features associated with lung_aca, aiding in the development of accurate diagnostic models. Upon analysis of the images, certain characteristics specific to lung_aca become evident. These may include glandular formations, mucin production, and varying degrees of cellular differentiation. These features are crucial in differentiating lung_aca from other lung cancer subtypes and benign lung conditions.

The diversity within the lung_aca category is notable, as the images depict different stages of tumor development, ranging from well-differentiated adenocarcinomas to more aggressive and invasive forms. This variability underscores the challenges in lung cancer diagnosis and emphasizes the need for robust machine learning models capable of accurately identifying and classifying these diverse patterns.

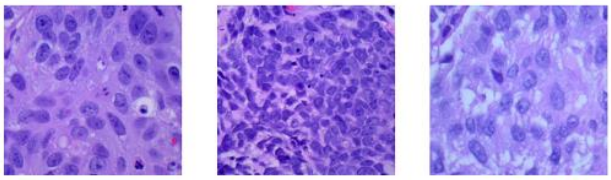

Figure 5. Images for lung_scc category

Figure 5 displays images belonging to the lung_scc category, representing squamous cell carcinoma (lung_scc) in this study. These images capture specific histopathological features associated with this type of lung cancer. Upon visual examination, distinctive characteristics of lung_scc become apparent in these images. These may include keratinization, intercellular bridges, and atypical squamous cells, all of which are hallmark features of squamous cell carcinoma. These visual cues are essential for accurately identifying and distinguishing lung_scc from other lung cancer subtypes and non-cancerous conditions.

These three are images shown have some special meaning in medical hostile, such as: Small cell lung cancer (SCLC) About 10% to 15% of all lung cancers are SCLC. This type of lung cancer tends to grow and spread faster than NSCLC. In most people with SCLC, the cancer has already spread beyond the lungs at the time it is diagnosed. Since this cancer grows quickly, it tends to respond well to chemotherapy and radiation therapy. "ACA lung" refers to adenocarcinoma of the lung, which is a type of non-small cell lung cancer (NSCLC). Adenocarcinoma is the most common form of lung cancer, accounting for about 40% of all lung cancer cases. It typically originates in the outer regions of the lungs, in cells that secrete mucus and other substances. This category includes tumors (≤3 cm) with small invasive components (≤5 mm) and generally has an excellent prognosis. The CNN model developed contains about 33.5 million parameters. This extensive parameter count and the model's complexity contribute to its high-performance capabilities, making it suitable for real-world applications.

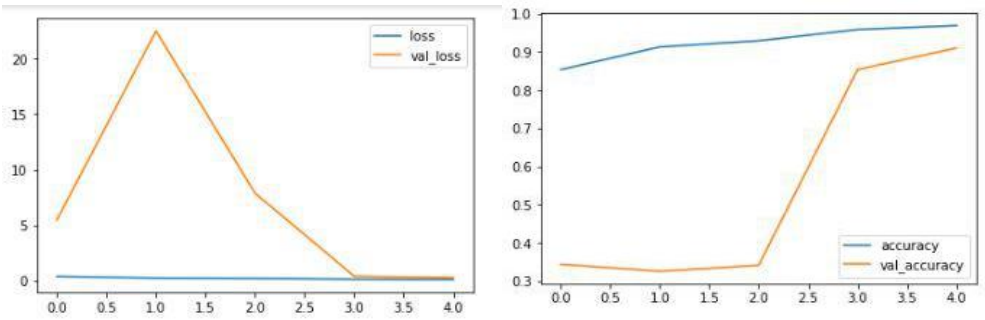

Figure 6. Training and validation accuracy graph

Figure 6 represents the Training and Validation accuracy graph with each epoch. Based on the graphs, it's clear that the model hasn't overfitted the training data. This is evident from the minimal difference between the training and validation accuracy.

Table 1. Confusion Matrix for the validation data

	Predicted A	Predicted B	Predicted C
Actual A	910	4	73
Actual B	0	946	31
Actual C	0	162	874

Table 1 represents the Confusion Matrix for the validation data provides a visual representation of proposed model's performance in predicting different classes. It is The confusion metrics and classification report using the predicted labels and true labels.
the diagonal cells indicate correct predictions, while off-diagonal cells represent incorrect predictions. This matrix helps assess the accuracy of the proposed model across different classes and identify any areas where misclassifications may occur.

Table 2. Classification Report for the Validation Data

	Precision	Recall	F1-score	support
lung_n	1.00	0.92	0.96	987
lung_scc	0.85	0.97	0.91	977
lung_aca	0.89	0.84	0.87	1036
Accuracy			0.91	3000
Macro avg	0.91	0.91	0.91	3000
Weighted avg	0.91	0.91	0.91	3000

Table 8 represents the Classification Report for the Validation Data and presents a detailed summary of the model's performance in classifying different categories. It includes metrics such as precision, recall, F1-score, and support for each class.

- Precision measures the accuracy of positive predictions.
- Recall calculates the proportion of actual positives that were correctly predicted.
- F1-score provides a balance between precision and recall.
- Support indicates the number of instances for each class.

This report represents how well the model performs for each class and understanding its strengths and weaknesses in classification tasks.

5. CONCLUSION:

The use of machine learning (ML) techniques for detecting lung cancer cells represents a significant advancement in medical imaging and oncology. ML-based approaches can improve the accuracy, efficiency, and precision of lung cancer diagnosis and treatment, offering several benefits and opportunities. Early detection of lung cancer is possible through the use of machine learning, which can help prevent further complications. Nowadays, machine learning and image processing are being used in the medical field by all countries, whether developed, underdeveloped, or developing. The CNN model is considered to be the most effective system for detecting cancer cells.
The performance of the simple CNN model is impressive, with an f1-score exceeding 0.90 for each class, indicating a 90% correctness rate in predictions. By utilizing Transfer Learning Techniques, the Simple CNN model utilizes pre-trained parameters that have been trained on millions of datasets for weeks using multiple GPUs. Leveraging such advanced techniques could unlock even greater accuracy and reliability in predictions, paving the way for more effective applications in real-world scenarios.

REFERENCES:

[1] *Shanti L. Kadachha, Pinal J. Pate, "Cancer Detection Using Modified Watershed", International Journal of Engineering Research & Technology (IJERT).*

[2] *Sudha. V, Jayashree. P, "Lung Nodule Detection in CT Images using Lung Nodule Detection in CT Images using Thresholding and Morphological Operations", International Journal of Emerging Science and Engineering (IJESE)*

[3] *Shweta Suresh Naik, Dr. Anita Dixit, "Cancer Detection using Image Processing and Machine Learning", International Journal of Engineering Research & Technology (IJERT).*

[4] *Suneetha Davuluri, D. Rathna Kishore, "Cancer Clumps Detection using Image Processing Based on Cell Counting and Artificial Neural Network Techniques", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019.*

[5] *Shubhpreet Kaur and Gagandeep Jindal, "Watershed Segmentation of Lung CT Scan Images For Early Diagnosis of Cancer", International Journal of Computer and Electrical Engineering Vol. 3, No. 6, December 2011.*

[6] *Bhagyashri G. Patil, Prof. Sanjeev N. Jain, "Cancer Cells Detection Using Digital Image Processing Methods", International Journal of Latest Trends in Engineering and Technology (IJLTET).*

[7] *Lung and Colon Cancer Histological images, https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images*