Research Article

A Neural Network Model in Identifying Non-Small-Cell Lung Cancer through CT Scans

Srihari Subramanian

Jamnabai Narsee International School, Mumbai, India

ABSTRACT

Lung cancer is the leading cause of cancer-related deaths worldwide, causing approximately 1.8 million deaths in 2022 alone. Lung cancer is often detected through the use of Low-Dose Computed Tomography (LDCT) scans, which use small amounts of radiation to construct detailed pictures of regions in the body. This study aims to explore the ability of Artificial Intelligence, particularly Convolutional Neural Networks (CNNs), to detect lung cancer from CT scans. Using an online dataset consisting of 1000 images, a CNN was developed with ResNet50 as the base model used for feature extraction. The model achieved a validation accuracy of 98.78% and a testing accuracy of 97.53%. This showcases the proficiency of the model in detecting lung cancer. However, this was only when a binary classification system was implemented, where the model was made to simply determine the presence of cancer. The model faced great difficulty in distinguishing between the types of lung cancer: Adenocarcinoma, Squamous Cell Carcinoma, and Large Cell Carcinoma. Additionally, the presence of a small number of false negatives while testing shows the danger of relying on AI and demonstrates the necessity of further fine-tuning before practical use.

Keywords: Convolutional Neural Networks, Lung cancer, detection, Artificial Intelligence, CT scans

INTRODUCTION

Lung cancer was first described in 1761 by Morgagni GB, an Italian anatomist. However, the first literature review about lung cancer was only published in 1912 by Isaac Adler (2). The primary reason for cancer-related deaths around the world is lung cancer, which caused around 1.8 million deaths in 2022 alone. In 2022, there were approximately 2.48 million new cases of lung cancer, showcasing the large number of people affected every year (3). Due to the relatively high frequency of lung cancer cases compared to other types of cancer (4), it is crucial to detect lung cancer efficiently and accurately. Identifying and treating lung cancer in its early stages is vital for effective treatment and much better chances of survival. The primary method of identifying lung cancer is through LDCT (Low-Dose Computed Tomography) scans, which aid in identifying a solid mass, a nodule, or an abnormal area of tissue present in the lungs, which are all signs of cancerous growth. However, there are concerns regarding the accuracy of LDCT scans as they often lead to false

Corresponding author: Srihari Subramanian, E-mail: ssrisrisrihari@gmail.com.

Received April 13, 2025**; Accepted** May 14, 2025
https://doi.org/10.70251/HYJR2348.332533

positive results in cases where other lung problems, such as scarring and infections, are identified as cancer. This may lead to unnecessary stress (1), underscoring the need for machine learning models to verify the accuracy of a lung cancer diagnosis.

LITERATURE REVIEW

A Low-Dose Computed Tomography (LDCT) scan is defined by the National Cancer Institute as a procedure that uses a computer linked to an X-ray machine that gives off a very low dose of radiation to make a series of detailed pictures of areas inside the body (5). While attempting to diagnose lung cancer in patients, radiologists often look for signs such as pulmonary nodules, lung masses, and pleural effusions.

Pulmonary nodules are areas in the lungs that experience abnormal growth. These nodules are often symptomless and could be neglected. Pulmonary nodules are common occurrences and are described by radiologists to be present in 1 out of 3 chest CT scans. While they are not a clear sign of cancer, they are one of the lung abnormalities that radiologists look for when attempting to diagnose lung cancer. It is not possible to diagnose cancer by viewing a pulmonary nodule via a CT scan. Small nodules, which are smaller than 0.6 centimeters, and nodules that have smooth edges are likely to be benign or non-cancerous. However, doctors recommend a biopsy in cases where the Nodule is larger than 9 millimeters or based on several other risk factors such as smoking, exposure to radon, age greater than 65, or a family history of cancer (6.

Lung masses are also areas of abnormal growth in the lungs. However, they are much more severe than pulmonary nodules and are more often malignant(cancerous) than benign. Abnormal growth is considered a lung mass when it is more than 3 centimeters in diameter. However, lung masses do not guarantee the presence of lung cancer either, as they could be the result of other infections such as bronchitis or pneumonia. In rare cases, it may even be the result of Lipoid Pneumonia, which is the entry of fat particles into the lungs. Nodules and lung masses often appear as a white “spot” or oval on CT scans (7).

Another possible sign of lung cancer is pleural effusion. Pleural effusion is the build-up of fluid in the pleural cavity, the space between the lungs and the chest wall. In normal cases, cells in the pleura produce small quantities of fluid to reduce friction and to keep the tissues moist. This fluid is constantly drained and replaced by the lymphatic system. In the case of a malignant pleural effusion, cancer cells spread to the space between the pleural layers. These cells cause the body to produce too much pleural fluid, besides blocking the flow of lymph fluid in the pleural cavity (8. On a CT scan, pleural effusion appears as a dependent opacity with a lateral upward sloping of a meniscus-shaped contour, a hazy upward-sloping line along the edge of the lungs seen in a CT scan (9).

The primary cause of lung cancer is cigarette smoking, which is the reason for around 80-90% of lung cancer-related deaths. Other tobacco-related products, such as cigars and pipes, also contribute to lung cancer. Tobacco smoke is known to contain around 70 chemicals that are capable of causing cancer in humans. Cigarette smoking increases the chances of lung cancer by around 15 to 30 times (11). Lung cancer can also be caused by inhaling Radon, which is a radioactive gas that is capable of causing genetic and epigenetic alterations in tumor genomes, affecting genes in a way that could lead to the development of lung cancer (10). Exposure to Radon occurs when it enters a house through cracks and holes in the walls or the floor. Since Radon cannot be smelled, tasted, or seen, it is hazardous even to non-smokers (11).

There are three main types of non-small-cell lung cancer: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma is primarily caused by cancer that has evolved from mucus-producing cells in the mucosal glands. It is the most common subtype of lung cancer, accounting for over 40% of lung cancer cases. It also accounts for a majority of lung cancer cases in non-smokers. This subtype often occurs in the peripheral regions of the lungs, typically originating in the bronchi (12). Squamous cell carcinoma occurs in the central regions of the lungs. It is the subtype of lung cancer that is most commonly caused by smoking. This type of lung cancer is mainly found near the bronchi. The third type, large cell carcinoma, which is much less common and can occur in any region of the lungs, is difficult to treat due to its tendency to metastasize at rapid rates (13).

There are 5 stages of lung cancer in humans, which are classified based on the following variables: Primary Tumor (T), Nodal Involvement (N), and Distant Metastasis (D. Each of these variables has its respective stages. Primary tumor aims at classifying lung cancer in terms of the size of the tumor and whether or not it has spread to structures adjacent to the lungs, such as the pleura or the chest wall. Nodal Involvement classifies lung cancer in terms of the extent to which it has metastasized to lymph nodes present in the region of the cancer. Nodal involvement is severe when the cancer metastasizes to lymph nodes in the mediastinum, or the area between the lungs (14). Distant metastasis is based on whether or not the cancer has spread to distant lymph nodes or distant organs from the original organ with cancer (15).

The least severe stage of lung cancer is stage 0, which refers to a cancer in the lining of the lungs or bronchus. In this stage, the cancer has not metastasized to other areas. In the stage, the cancer is considered to be in situ or at the original location. Stage I lung cancer has metastasised within the lungs but not to other organs or lymph nodes. Stage II lung cancer has spread to lymph nodes. Lung cancer is also considered stage two if there are two or more tumors in the same lobe. Stage III occurs when the cancer has metastasised to external structures and lymph nodes. Stage IV involves metastasis to the other lung, pleural effusion, and may even involve the fluid around the heart. There is no current cure for stage IV lung cancer. Lung cancer must be detected quickly and accurately to ensure that treatment and recovery are possible (16).

METHODS AND MATERIALS

The dataset used to train the model to detect lung cancer is the Chest CT-Scan images Dataset from Kaggle (17). This dataset consists of 1000 images, which are divided into four classes: Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma, and normal. The percentages of images used for training, testing, and validation are 70%, 20%, and 10%, respectively. The software used to conduct exploratory data analysis and ultimately build the model is Google Colab. Figure 1 shows the first 9 images in the training data laid out in a 3 × 3 matrix along with their respective classes to provide examples about the data used in this study.

In most cases, it is relatively easy to distinguish between scans that contain signs of lung cancer and scans that do not. In normal cases, the lungs appear black as they mostly contain air, and air does not absorb the X-ray radiation emitted during a CT scan. However, denser areas such as regions with pleural effusion or solid lung masses

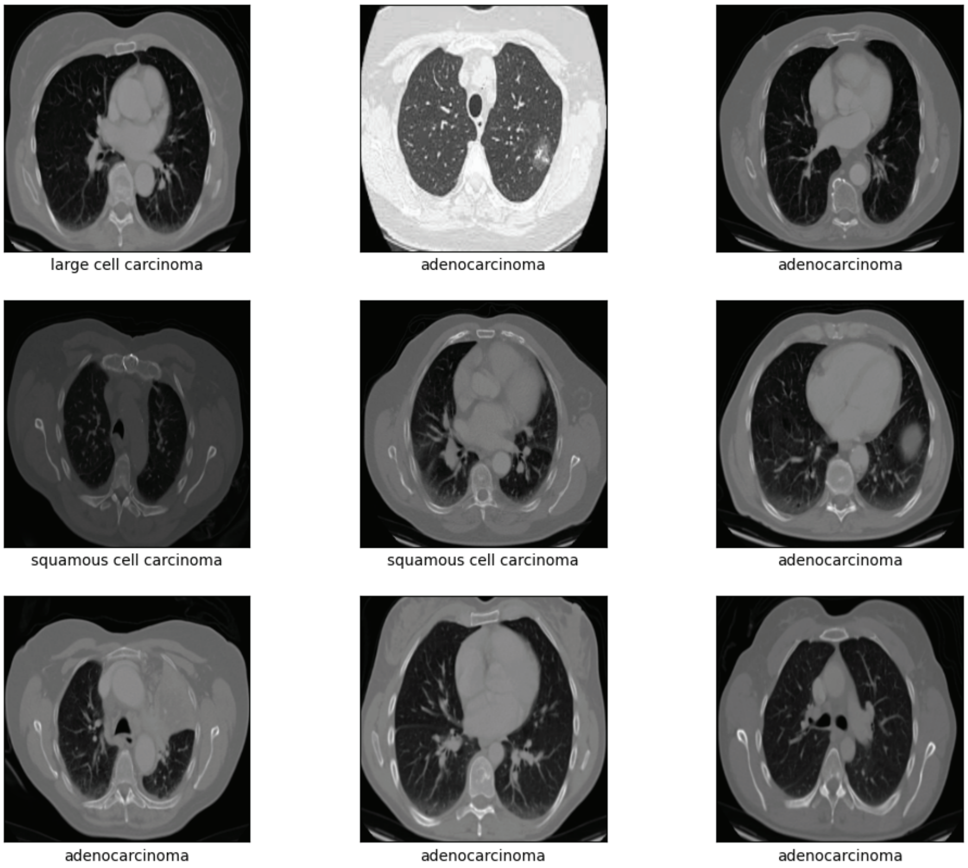

Figure 1. The first nine images from the training dataset as an example of the data used in this study. The various types of lung cancer are portrayed as labels below their respective images.

appear lighter on a CT scan as they absorb more radiation. This is shown in Figure 2 where the brighter region within the lungs could indicate cancer, while Figure 3 showcases a normal case where the lungs appear black.

The colorbar is present to help showcase the difference between the colors of a normal lung and a lung in which lung cancer has developed.

This study focuses on the effectiveness of a machine learning algorithm in detecting lung cancer from CT scans. A machine learning algorithm was used to analyse CT scans. Firstly, various filters were applied to the CT scan images. A ResNet50 base model was used to extract the prominent features of the image and make it ready for the model to classify the data. This was followed by dense layers, which are often used after the base model to extract the features. The dense layers consist of neurons that connect to the neurons of the preceding layers. The neurons had a strong dropout of 0.5 to prevent the model from overfitting, as it generally did. Through the addition of a dropout layer, overfitting is prevented by randomly shutting down outputs from neurons. This is achieved by setting input units to 0 (18). Figure 4 shows the structure of the model.

However, the model continued to overfit. The training

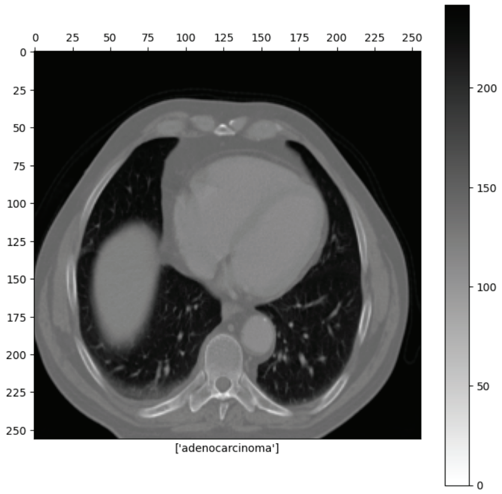

Figure 2. The hazy white spot on the left lung indicates a malignant lung mass, which is a sign of cancer. Therefore, the image has been classified as adenocarcinoma.

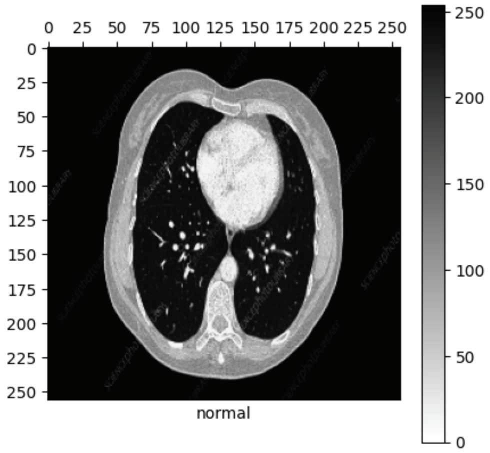

Figure 3. In this figure, the lungs have no sign of lung cancer as they appear black on the CT scan with no dense areas. Therefore, it has been classified as 'normal'.

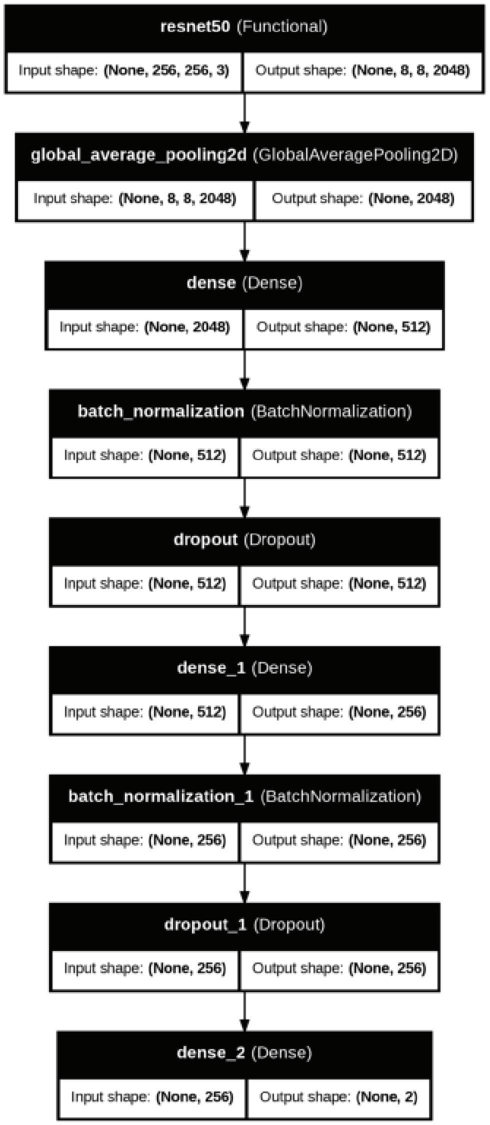
resnet50 (Functional)
Input shape: (None, 256, 256, 3)
Output shape: (None, 8, 8, 2048)
global_average_pooling2d (GlobalAveragePooling2D)
Input shape: (None, 8, 8, 2048)
Output shape: (None, 2048)
dense (Dense)
Input shape: (None, 2048)
Output shape: (None, 512)
batch_normalization (BatchNormalization)
Input shape: (None, 512)
Output shape: (None, 512)
dropout (Dropout)
Input shape: (None, 512)
Output shape: (None, 512)
dense_1 (Dense)
Input shape: (None, 512)
Output shape: (None, 256)
batch_normalization_1 (BatchNormalization)
Input shape: (None, 256)
Output shape: (None, 256)
dropout_1 (Dropout)
Input shape: (None, 256)
Output shape: (None, 256)
dense_2 (Dense)
Input shape: (None, 256)
Output shape: (None, 2)

Figure 4. This figure shows the structural architecture of the model, including the dense and dropout layers.

accuracy was high while the validation accuracy remained at around 60%. This was corrected through a variety of methods that ensured that the model was robust and did not overly depend on the training images provided. Firstly, an image data generator was used to augment the training images through rotation, flipping, height and width shifts. This ensures that the model is prepared for a wide variety of CT scan images, which may not all follow the same pattern. Secondly, due to the high-class imbalance in the dataset, class weights were calculated and higher weights were provided to the smaller class (normal), so that the model attends to it more while training. Additionally, the learning rate of the model was changed dynamically while training to prevent overfitting, and the model was made to stop early when validation accuracy ceased to improve to ensure that the model with the best weights was preserved.

RESULTS

In the dataset, the images were separated into four classes: Adenocarcinoma, Squamous Cell Carcinoma, Large Cell Carcinoma, and Normal. However, while using all four classes, the model experienced low accuracy while classifying the images. The validation accuracy was around 48%, with a peak at 51.39%, which is undesirable. This was due to the high difficulty in distinguishing between the three types of lung cancer. This was likely due to a high similarity in features between the various types of cancer. Location is a big indicator that enables us to distinguish between the types of lung cancer. However, in some cases location may be misleading, such as when lung adenocarcinoma occurs in the primary bronchi. In this case it may be mistaken for squamous cell carcinoma (19). In reality, it is very difficult to distinguish between the various types of Non-Small-Cell Lung Cancer solely using CT scans (20). In most cases, radiologists employ different means to distinguish between types of lung cancer such as biopsy. Therefore, to improve the accuracy, the folders containing images of the three types of cancer were merged into one to create a binary classification system where the model was made to simply detect cancer and not establish the type of cancer present in the CT scan. In this case, the same model was more successful and had a validation accuracy of 98.78%, which was most common across epochs. The validation accuracy was able to reach 100% in some epochs on the validation dataset. Figure 5 shows the accuracy and the validation accuracy across epochs. The model was then tested on a testing dataset consisting of 325 images of which 8 were misidentified, demonstrating a testing accuracy of around 97.53% which was the highest accuracy observed across various tests. This is shown in a confusion matrix in Figure 6.

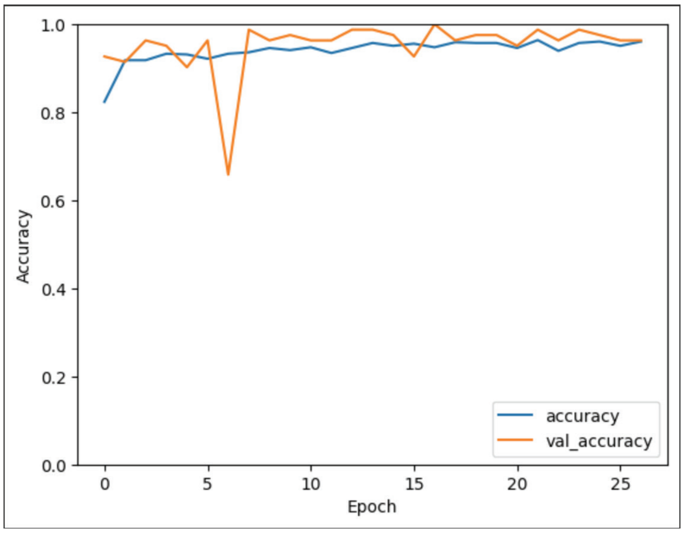

Figure 5. The figure shows the accuracy and the validation accuracy of the model across epochs. In general, the model appears to do well with accuracies consistently above 80%.

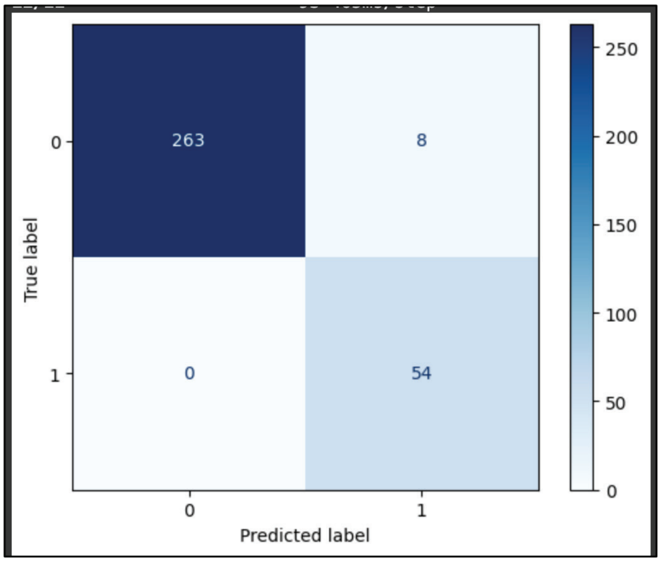

Figure 6. The confusion matrix shows the results of using the model on 325 testing images. The label 0 represents the 'Cancer' class whereas the label 1 represents the 'normal' class. The model is able to identify normal cases accurately, eliminating false positives. However, there are 8 false negatives, which are far more dangerous.

CONCLUSION

The model demonstrated its capability in identifying the presence of lung cancer from CT scans. However, the 8 images which were misidentified in the testing dataset were all false negatives. False negatives are far more dangerous than false positives, as undiagnosed lung cancer can be extremely dangerous if left alone for too long. Additionally, due to the usage of a single overall dataset, which was already cleaned up, the model may have become overly comfortable with one particular type of data. This leads to overfitting and possibly a lower accuracy when practically applied in a real-world scenario. This could be avoided by using a greater quantity of data from multiple datasets. However, due to time constraints and difficulties in procuring CT scan datasets to use for training, this could not be completed.

FUTURE DIRECTION

The model proved to be successful in identifying lung cancer in individuals through CT scans. However, the inability of the model to distinguish between various types of lung cancer is, at the moment, very difficult to solve. Additionally, while the accuracy was high, the risk of false negatives is always a major problem because the few images that were inaccurately predicted were cancerous cases classified as normal. False negatives are a major threat as the lung cancer could have advanced to an incurable stage before it is finally detected.

Furthermore, there were limitations such as GPU usage limits on Google Colab, which hindered progress as each epoch took approximately 15 minutes to run, which was not ideal when multiple epochs had to be run to fine-tune the model. At the moment, AI in lung cancer detection appears to be an increasingly popular topic for exploration due to various studies, such as the study by Geetu Lakshmi and P. Nagaraj, which also reported high accuracy using a ResNet model to detect lung cancer (21).

DECLARATION OF CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Cui JW, Li W, Han FJ & Liu YD. Screening for lung cancer using low-dose computed tomography: concerns about the application in low-risk individuals. Translational lung cancer research. 2015; 4 (3): 275–286. https://doi.org/10.3978/j.issn.2218-6751.2015.02.05
2. Kloecker G. History of lung cancer. Kloecker G, Arnold SM, Fraig MM & Perez CA. (Eds.), *Lung Cancer: Standards of Care*. McGraw-Hill. 2021. https://hemonc.mhmedical.com/content.aspx?bookid=2965&sectionid=250114244
3. World Cancer Research Fund. (n.d.). *Lung cancer statistics*. Available from: https://www.wcrf.org/preventing-cancer/cancer-statistics/lung-cancer-statistics (accessed on 2025-1-17)
4. Britannica T. Editors of Encyclopaedia (2025, January 15). *lung cancer. Encyclopedia Britannica*. https://www.britannica.com/science/lung-cancer (accessed on 2025-1-19)
5. *NCI Dictionary of Cancer Terms*. (n.d.). Cancer.gov. Retrieved on January 19 2025, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/low-dose-ct-scan
6. *Pulmonary nodules*. (2025, November 13). Cleveland Clinic. . Available from: https://my.clevelandclinic.org/health/diseases/14799-pulmonary-nodules (accessed on 2025-1-19)
7. *Treatment for a lung mass | RWJBarnabas Health NJ*. (n.d.). RWJBarnabas Health. . Available from: https://www.rwjbh.org/treatment-care/cancer/types-of-cancer/lung-thoracic-cancer/lung-mass/#:~:text=A%20lung%20mass%20is%20an,cases%2C%20lung%20masses%20are%20cancerous (accessed on 2025-1-19)
8. Du Cancer CCS. *Fluid buildup on the lungs (pleural effusion)*. Canadian Cancer Society. (2022, May). Available from: https://cancer.ca/en/treatments/side-effects/fluid-buildup-on-the-lung-pleural-effusion#:~:text=Cancer%20cells%20cause%20the%20body,lung%20cancer (accessed on 2025-1-19)
9. Lababede O, MD. *Pleural effusion imaging: practice essentials, radiography, computed tomography*. (2022, August 23). Available from: https://emedicine.medscape.com/article/355524-overview (accessed on 2025-1-19)
10. Choi JR, Koh SB, Kim HR, Lee H & Kang DR. Radon Exposure-induced Genetic Variations in Lung Cancers among Never Smokers. Journal of Korean medical science. 2018; 33 (29): e207. https://doi.org/10.3346/jkms.2018.33.e207
11. Lung cancer risk factors. Lung Cancer. (2024, October 15). Available from: https://www.cdc.gov/lung-cancer/risk-factors/index.html#:~:text=Smoking-,Cigarette%20smoking%20is%20the%20number%20one%20risk%20factor%20for%20lung,of%20more%20than%207%2C000%20chemicals (accessed on 2025-1-19)
12. Myers DJ & Wallen JM. *Lung adenocarcinoma*. In StatPearls. Treasure Island, FL: StatPearls Publishing. 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519578/ (accessed on 2025-1-19)

13. *What is lung cancer? | Types of lung cancer*. (2024, January 29). American Cancer Society. Available from:https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html (accessed on 2025-1-25)
14. Mountain CF. Staging of lung cancer. *The Yale journal of biology and medicine*. 1981; 54 (3): 161–172.
15. *NCI Dictionary of Cancer Terms*. (n.d.-b). Cancer.gov. Retrieved on January 25 2025 from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/distant-metastasis
16. *Lung cancer*. Cleveland Clinic. (2022, October 31). Available from: https://my.clevelandclinic.org/health/diseases/4375-lung-cancer (accessed on 2025-1-25)
17. Hany M & Omarhanyy. *Chest CT-Scan Images* [Dataset]. Kaggle. 2021. Available from: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images (accessed on 2025-2-8)
18. Team K. (n.d.). *Keras documentation: Dropout layer*. Retrieved on 8th march 2025, Available from: https://keras.io/api/layers/regularization_layers/dropout/#:~:text=The%20Dropout%20layer%20randomly%20sets,over%20all%20inputs%20is%20unchanged. (accessed on 2025-3-8)
19. Xu Z, Ren H, Zhou W & Liu Z. ISANET: Non-small cell lung cancer classification and detection based on CNN and attention mechanism. *Biomedical Signal Processing and Control. 2022*; 77: 103773. https://doi.org/10.1016/j.bspc.2022.103773
20. Liu H, Jiao Z, Han W & Jing B. Identifying the histologic subtypes of non-small cell lung cancer with computed tomography imaging: a comparative study of capsule net, convolutional neural network, and radiomics. *Quantitative imaging in medicine and surgery*. 2021; 11 (6): 2756–2765. https://doi.org/10.21037/qims-20-734
21. Geethu Lakshmi G & Nagaraj P. Lung cancer detection and classification using optimized CNN features and Squeeze-Inception-ResNeXt model. *Computational Biology and Chemistry. 2025*; 117: 108437. https://doi.org/10.1016/j.compbiolchem.2025.108437

APPENDIX

This is the final model that provided me with the best results through various iterations of testing different models:

```
base_model = tf.keras.applications.ResNet50(

  include_top=False, weights=”imagenet”, input_
  shape=(256, 256, 3)

)

base_model.trainable = False

model = Sequential([

  base_model,

  GlobalAveragePooling2D(),

  Dense(512, activation=’relu’),

  BatchNormalization(),

  Dropout(0.5),

  Dense(256, activation=’relu’),

  BatchNormalization(),

  Dropout(0.3),

  Dense(2, activation=’softmax’)

])


model.compile(

  optimizer=Adam(learning_rate=3e-4),

  loss=’categorical_crossentropy’,

  metrics=[‘accuracy’]

)
```

```
datagen = ImageDataGenerator(

  rescale=1./255,

  rotation_range=10,

  width_shift_range=0.1,

  height_shift_range=0.1,

  shear_range=0.1,

  zoom_range=0.1,

  horizontal_flip=True,

  fill_mode=”nearest”

)


x_train = datagen.flow_from_directory(

  directory=”/content/gdrive/MyDrive/Data 2/train”,

  target_size=(256, 256),

  batch_size=32,

  class_mode=”categorical”,

  shuffle=True

)

x_val = datagen.flow_from_directory(

  directory=”/content/gdrive/MyDrive/Data 2/valid”,
    target_size=(256, 256),
    batch_size=32,
    class_mode=”categorical”,
    shuffle=True
)
cw = compute_class_weight(
    class_weight=”balanced”,
    classes=np.unique(x_train.classes),
    y=x_train.classes
)
cw_dict = dict(enumerate(cw))


early_stopping = EarlyStopping(monitor=’val_accuracy’, patience=10, restore_best_weights=True, verbose=1)
reduce_lr = ReduceLROnPlateau(monitor=’val_accuracy’, factor=0.5, patience=7, min_lr=1e-6, verbose=1)


history = model.fit(
    x_train,
    validation_data=x_val,
    class_weight=cw_dict,
    epochs=30,
    callbacks=[early_stopping, reduce_lr],
    verbose=1
)
# second trial with a lower learning rate to finetune the model
model.compile(optimizer=Adam(learning_rate=1e-5), loss=’categorical_crossentropy’, metrics=[‘accuracy’])
history_tune = model.fit(
    x_train,
    validation_data=x_val,
    class_weight=cw_dict,
    epochs=10,
    callbacks=[early_stopping, reduce_lr],
    verbose=1
)
```