CP 468 – Artificial Intelligence

Spring 2024

Group 12

Project Summary Report

Member 1 Name: Laiba Ali **ID:** 169021077

Member 2 Name: Aliha Ali **ID:** 210184090

Member 3 Name: Arsalan Khan **ID:** 210862640

Member 4 Name: Rhyme Her **ID:** 20157720

Submission Date: August 2nd, 2024

All the listed members have contributed, read, and approved this submission.

Member 1 Name: Laiba Ali	**Signature:** Laiba.Ali
Member 2 Name: Aliha Ali	**Signature:** AA
Member 3 Name: Arsalan Khan	**Signature:** AK
Member 4 Name: Rhyme Her	**Signature:** RH

Abstract

Approximately 80% of lung cancer cases are classified as non-small cell lung cancer (NSCLC). NSCLC encompasses a diverse range of lung cancers that often spreads into other parts of the body, and is often undetectable until advanced stages. This project aims to develop a tool to aid radiologists in the early detection of NSCLC using Convolutional Neural Networks (CNNs). The tool integrates three CNN models to analyse Computerised Tomography (CT) scans of the lungs. Keras, TensorFlow, and Gradio are utilised for their prototyping, backend and frontend features. Through image processing and normalisation, this tool is able to demonstrate the ability of identifying cases of NSCLC given a pool of CT scans.

Introduction to the Lung Cancer Detection Problem

Lung cancer remains one of the leading causes of cancer-related deaths worldwide. Early detection is crucial for successful treatment, yet it remains challenging due to the subtle nature of early-stage tumours. Lung cancers fall into two categories encompassing different types: 1) small-cell lung cancer (SCLC), a cluster of dense cells that metastasizes, or spreads, rapidly in the lungs and 2) non-small-cell lung cancer (NSCLC), a large cluster of abnormal cells that slowly metastasizes to other parts of the body.

This project focuses on creating a tool to identify NSCLC in patients' CT scans. NSCLC's slow growth rate allows for a longer detection window, leading to opportunities for early treatment and improved outcomes. Additionally, NSCLC has a higher prevalence and includes several subtypes than SCLC. Automating the detection process aims to support radiologists and improve diagnostic accuracy.

This detection tool uses Computed Tomography (CT) scans. CT scans offer high-resolution imaging of internal body structures while avoiding interference from overlapping structures.

Convolutional neural networks (CNNs) are used to detect lung cancer from CT scans. The project aims to design a CNN from scratch, amalgamate the results of three different pre-trained models through transfer learning, and apply the learned experiences to contribute to the early detection and diagnosis of lung cancer.

Of Similar and Of Existing Solutions

CNNs have found use in the life sciences. Their presence involves medical imaging for detection and diagnosis. Their ability to process and adapt makes them suited for handling exponentially large data and the complexity present in biological data.

Similar applications of CNNs were present in Kaggle's 2017 Data Science Bowl. Inception and ResNet were utilised, with DICOM files being the main medium. It is noted a number of existing solutions did not implement VGG16 despite its advantages towards image classification.

Google AI has made strides in lung cancer detection with their research and development of Google AI Lung Cancer Detection, developing a deep learning model for cancer prediction. With the affirmation of radiologists, the AI demonstrated high accuracy and consistency through their use of a volumetric CNN that processes 3D CT scans.

Of the noted existing solutions, deep-learning techniques, CNNs, 3D-CNNs, and ensemble methods have demonstrated their efficacy in the life sciences. Their applications have shown to improve the accuracy and efficiency of lung cancer detection for medical imaging. This project aligns with these approaches by leveraging the pre-trained models and the ensemble strategy.

The project's model approach leverages the strength of different models in an ensemble to improve the detection of NSCLC from CT scans. Integrating the various models such as VGG16, ResNet50, and InceptionV3, the goal is to address the following issues:

1. Mitigate overfitting in the training data by combining different aspects of outputs to create generalised solutions,
2. Improving accuracy through the comparison of outputs to rule out outliers, and
3. Reducing bias from external factors by increasing diversity of learning.

Solution

Tools & Libraries

Tools Featured	
Name	**Description**
Keras	For prototyping, abstraction of operations/models, and model prediction. Provides sequential models for CNNs, layers for building CNNs, applications for pre-trained models, and data pre-processing.
Tensorflow	For experimentation and scalability. Provides resources for model prediction.
Gradio	For front-end interface. Provides a web interface for Keras and Tensorflow to facilitate interaction with the models.
Python Imaging Library (PIL)	For image processing. Provides the ability to open, manipulate, and save images.
OpenCV	For image resizing and normalisation. Provides consistency in formatting, and other image processing functionalities.
NumPY	For array manipulation. Provides support for multi-dimensional arrays and matrices alongside operations to manipulate them.
Pydicom	For support of DICOM files. Provides the ability to read, write, and modify DICOM files.
Matplotlib	For data visualisation. Provides `pyplot` for creating visualisations and reports in Python.
Sklearn	For machine learning and statistical models. Provides tools to evaluate metrics, classification reports, confusion matrices, and ROC curves.

Convolutional Neural Networks

1. Simple CNN

A straightforward CNN model that augments data for performance.

- Architecture →
  - Convolutional layers: Feature extraction from the input images.
  - MaxPooling layers: Downsample the spatial dimensions of the feature maps.
  - Fully Connected layers: Classification based on the extracted features.
  - Activation Functions: ReLU for hidden layers and Softmax for the output layer.
- Advantages →
  - Simple and easy to implement.
  - Effective for small to medium-sized image classification tasks.
  - Requires less computational resources compared to more complex models.
- Limitations →
  - Limited capacity to capture complex patterns in the data.
  - May not perform as well as deeper architectures on large and diverse datasets.

2. VGG16

A pre-trained CNN that excels at image classification tasks.

Architecture→

- Convolutional Layers: Uses multiple small (3x3) convolutional filters for capturing details.
- Depth: composed of 16 weight layers.
- Pre-trained on ImageNet: trained on a large dataset (ImageNet), allowing for transfer learning to adapt the model for cancer detection.
- Custom Top Layers: Additional fully connected layers to adapt the model for cancer detection.

Advantages→

- Strong feature extraction due to its depth and small filters.
- Transfer learning enables strong performance from less labelled data.
- Architecture is suited for image classification, boosting accuracy.

Limitations→

- Computationally intensive for resources and time.

- Depth and parameter count increases memory-intensity exponentially.

3. ResNet50

A pre-trained model for stronger feature extraction and improved performance on the classification task.

- Architecture →
    - Residual blocks: Allow training of deeper networks by mitigating the vanishing gradient problem.
    - Pre-trained on ImageNet: Utilises transfer learning to leverage previously learned features.
    - Custom top layers: Adapt the pre-trained model to the specific lung cancer classification task.
- Advantages →
    - Strong feature extraction capabilities due to deep architecture.
    - Improved performance on complex image classification tasks.
    - Benefits from transfer learning, reducing the need for large amounts of labeled data.
- Limitations →
    - Computationally intensive and requires more resources.
    - Longer training times compared to simpler models like the one stated above

4. InceptionV3

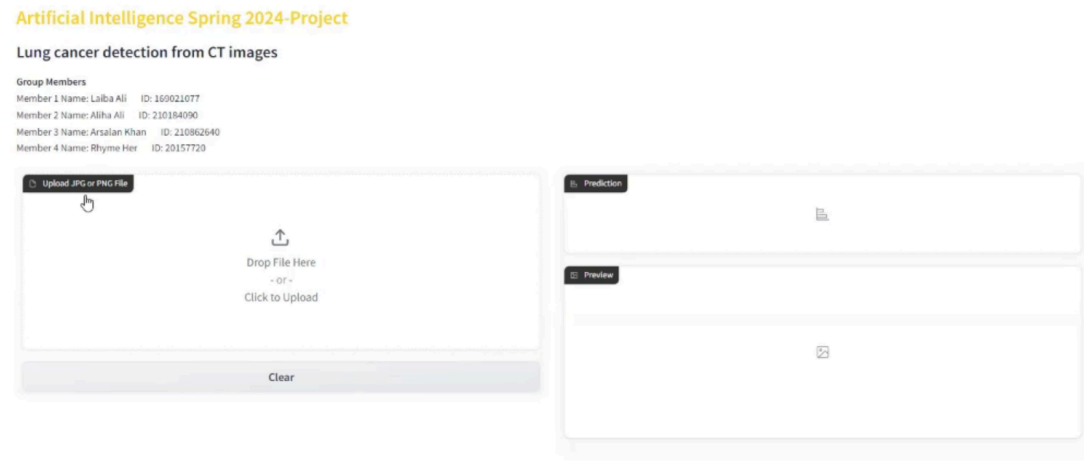
InceptionV3 is utilised for its image analysis to enhance dataset diversity and generalisation.

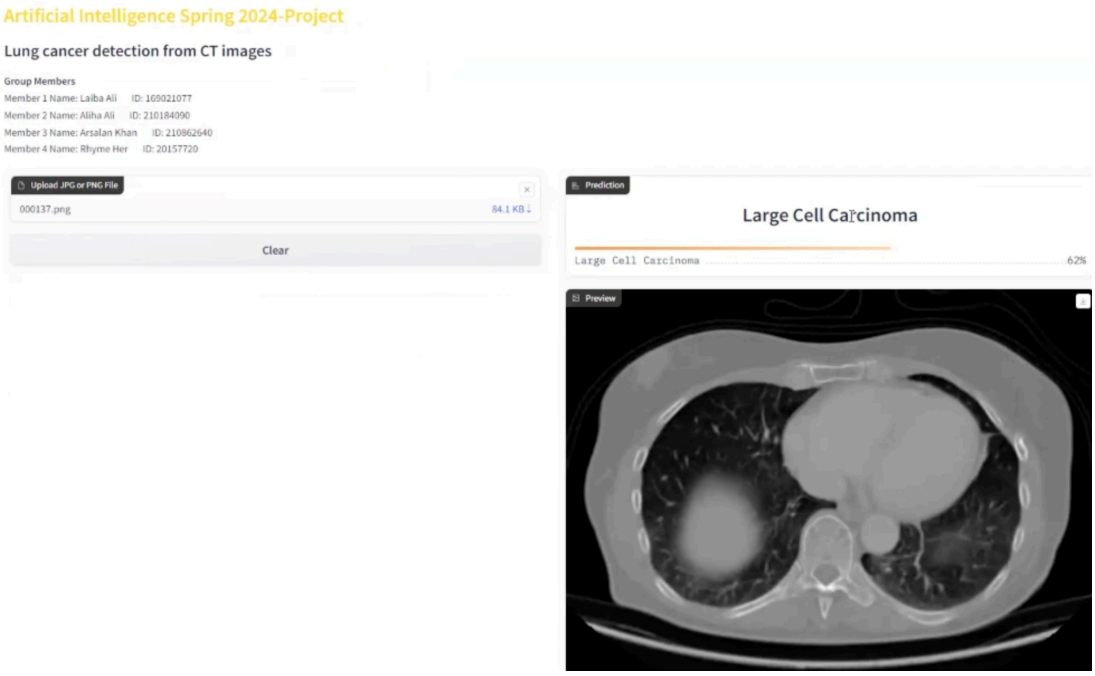
- Architecture →
    - Inception modules: Combine multiple convolutional filter sizes to capture diverse spatial features.
    - Pre-trained on ImageNet: Leverages transfer learning for effective feature extraction.
    - Custom top layers: Tailored to the lung cancer classification task.
- Advantages →
    - Efficienti in ensemble capturing spatial hierarchies in images.
    - Versatile in handling different scales of features.
    - Benefits from transfer learning, improving generalisation.
- Limitations →
    - Complex architecture requires significant computational resources.

- Prone to overfitting if not properly regularised.

Gradio

Gradio functions as the front-end interface, streamlining the use of the pre-trained models. Its architecture provides the user interface for uploading and classifying images. Gradio utilises the ensemble of the three pre-trained models, averaging the predictions to improve overall accuracy. The use of Gradio increases computational load due to multiple model predictions and, in return, increases classification accuracy by combining the strengths of various models. Below are screenshots of the interface:





Analysis of Methodology

Data Description

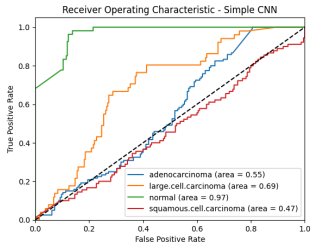
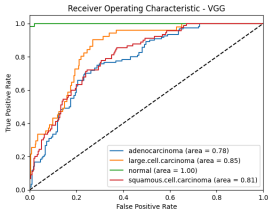
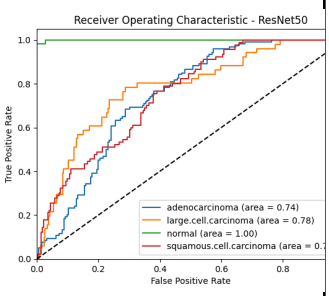
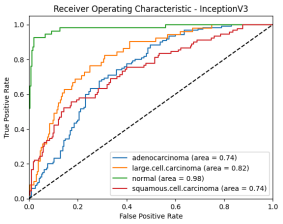
Sourced from Kaggle, the dataset of CT scans is composed of: 464 training files (70%), 72 validation files (10%), and 315 testing files (20%). File types accepted for the model requirements are JPG, PNG, and DICOM, with the latter being excluded in the used dataset. Images are resized, normalised, and augmented through the use of rotation, flipping, and zooming to improve model generalisation.

The dataset includes images of the following NSCLC types:

- Adenocarcinoma: cells lining lungs are cancerous.
- Large Cell Carcinoma: clusters of cancerous cells spread throughout lungs.
- Squamous Cell Carcinoma: cancerous cells found in middle lung structures (i.e. bronchi).[1]
- Normal: healthy lungs of no concern.

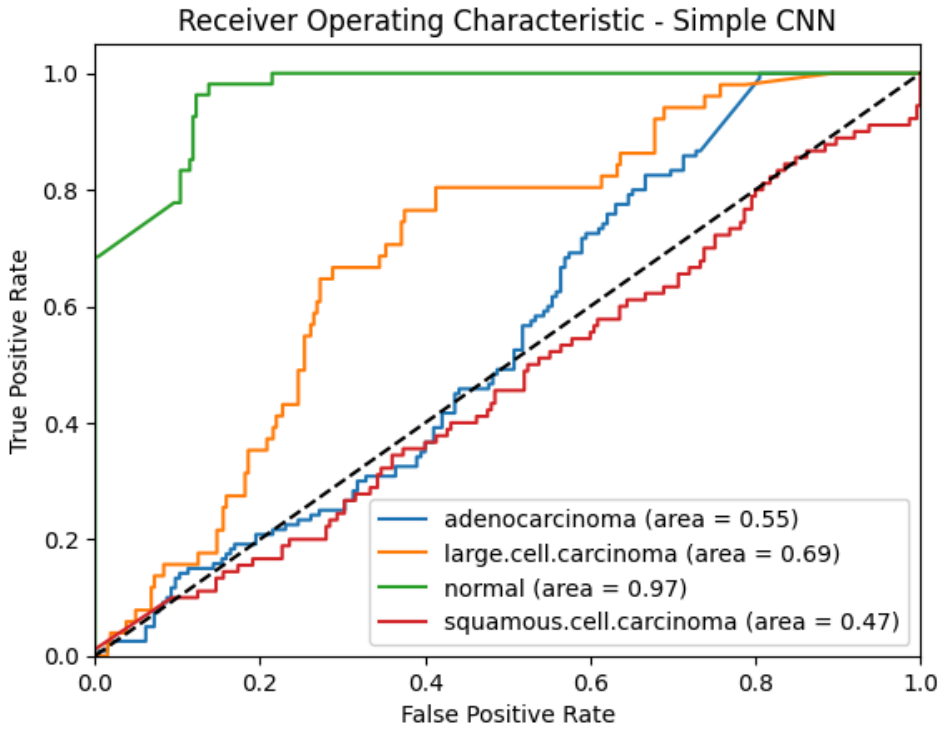
[1] Rachael Zimlich, “Large Cell Lung Carcinoma: Symptoms, Treatment, and Outlook,” ed. Rena Goldman et al., Healthline, June 5, 2024, https://www.healthline.com/health/lung-cancer/large-cell-carcinoma, section. "What is large cell lung carcinoma?"

Quality Measure

	SimpleCNN	VGG16	ResNet50	InceptionV3
Confusion Matrices	[37 83 0 0] [11 40 0 0] [8 13 33 0] [29 61 0 0]	[44 64 0 12] [0 49 0 2] [0 0 53 1] [24 39 0 27]	[111 9 0 0] [30 21 0 0] [1 0 53 0] [75 15 0 0]	[87 31 1 1] [15 35 1 0] [3 1 50 0] [54 26 5 5]
Classification Report	Adenocarcinoma: **Precision** 0.44 **Recall** 0.31 **F1-score** 0.36 Large Cell Carcinoma: **Precision** 0.20 **Recall** 0.78 **F1-score** 0.32 Normal: **Precision** 1.00 **Recall** 0.61 **F1-score** 0.76 Squamous Cell Carcinoma: **Precision** 0.00 **Recall** 0.00 **F1-score** 0.00	Adenocarcinom: **Precision** 0.65, **Recall** 0.37, **F1-score** 0.47 Large Cell Carcinoma: **Precision** 0.32, **Recall** 0.96, **F1-score** 0.48 Normal: **Precision** 1.00, **Recall** 0.98, **F1-score** 0.99 Squamous Cell Carcinoma: **Precision** 0.64, **Recall** 0.30, **F1-score** 0.41	Adenocarcinoma: **Precision** 0.51, **Recall** 0.93, **F1-score** 0.66 Large Cell Carcinoma: **Precision** 0.47, **Recall** 0.41, **F1-score** 0.44 Normal: **Precision** 1.00 **Recall** 0.98, **F1-score** 0.99 Squamous Cell Carcinoma: **Precision** 0.00 **Recall** 0.00 **F1-score** 0.00	Adenocarcinoma: **Precision** 0.55, **Recall** 0.72, **F1-score** 0.62 Large Cell Carcinoma: **Precision** 0.38, **Recall** 0.69, **F1-score** 0.49 Normal: **Precision** 0.88, **Recall** 0.93, **F1-score** 0.90 Squamous Cell Carcinoma: **Precision** 0.83, **Recall** 0.06, **F1-score** 0.10
ROC/AUC				
Training and Validation Accuracy	Struggles identifying "squamous cell carcinoma."	Adept at identifying "large cell carcinoma."	Excels at detecting "adenocarcinoma." Fails in "squamous cell carcinoma."	High performance on "adenocarcinoma." Struggles with "squamous cell carcinoma."
Evaluating Ensemble for Accuracy and Robustness	Omitted from ensemble.	Provides high-level feature extraction and robust performance on various image features.	Adds the ability to recognize fine-grained details and patterns that simpler models might miss.	Helps in generalising the model to diverse image features and patterns.

Training Curve Results

Simple CNN



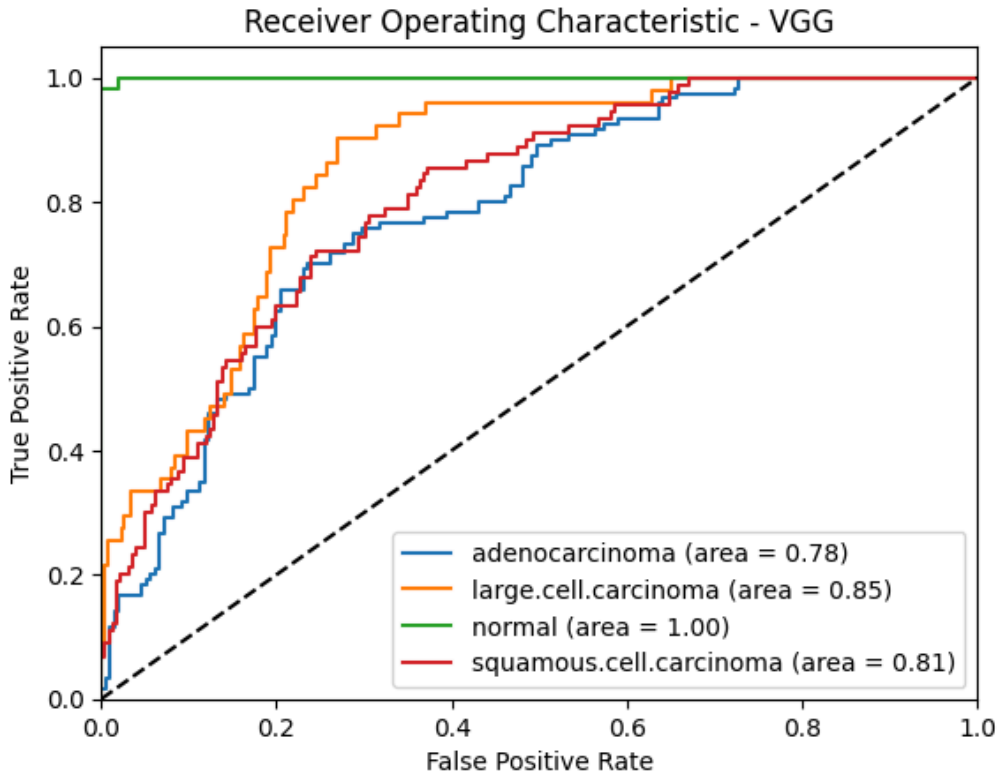
Accuracy: 34.92%

Loss: 1.3088

Confusion Matrix:

- Adenocarcinoma: 37 TP, 83 FN
- Large Cell Carcinoma: 11 TP, 40 FN
- Normal: 8 TP, 13 FN
- Squamous Cell Carcinoma: 29 TP, 61 FN
    - **Observations:** The model has high precision for the "normal" class and performs well on "large cell carcinoma."

VGG16



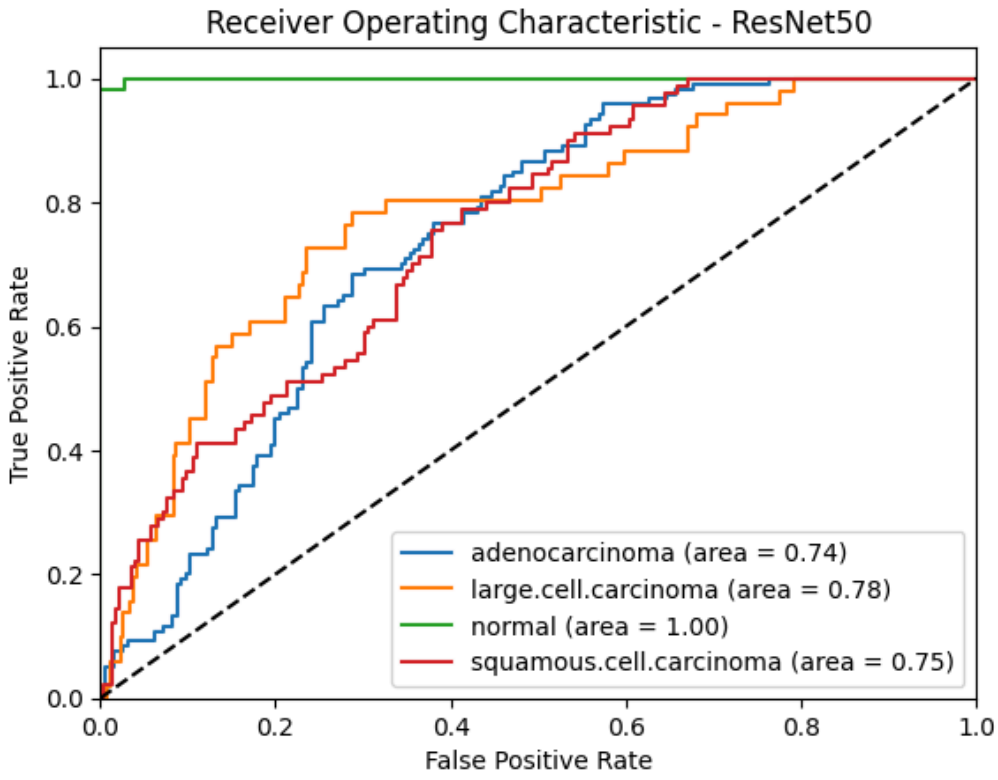
Accuracy: 54.92%

Loss: 1.0669

Confusion Matrix:

- Adenocarcinoma: 44 TP, 64 FN
- Large Cell Carcinoma: 0 TP, 49 FN
- Normal: 0 TP, 0 FN
- Squamous Cell Carcinoma: 24 TP, 39 FN
    - **Observations:** The model has high precision for the "normal" class and performs well on "large cell carcinoma.

ResNet50



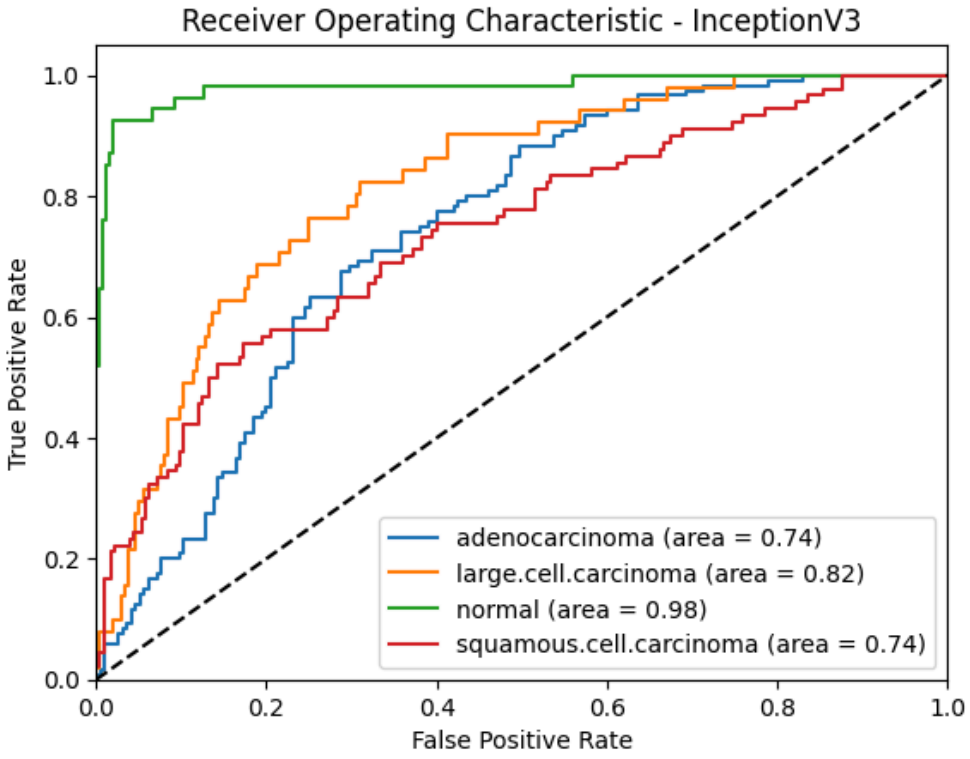
Accuracy: 58.73%

Loss: 1.0037

Confusion Matrix:

- Adenocarcinoma: 111 TP, 9 FN
- Large Cell Carcinoma: 30 TP, 21 FN
- Normal: 1 TP, 0 FN
- Squamous Cell Carcinoma: 75 TP, 15 FN
    - **Observations:** The model excels in detecting "adenocarcinoma" and "normal" classes but fails in "squamous cell carcinoma."

InceptionV3



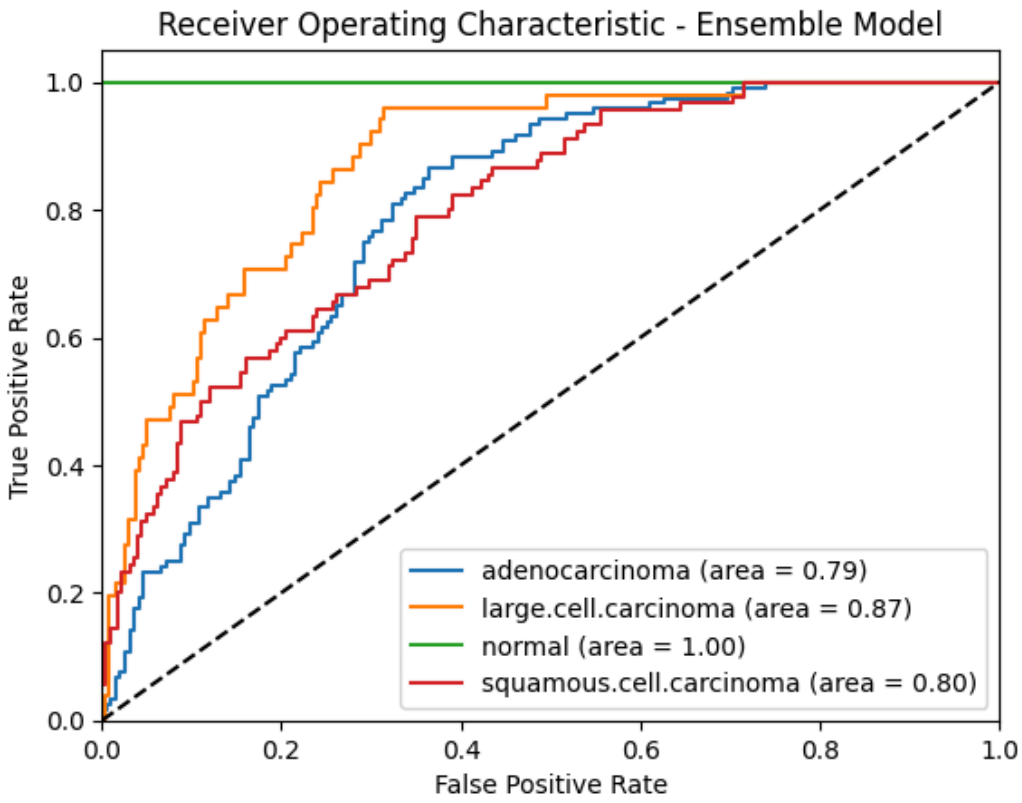
Accuracy: 56.19%

Loss: 1.1073

Confusion Matrix:

- Adenocarcinoma: 87 TP, 31 FN
- Large Cell Carcinoma: 15 TP, 35 FN
- Normal: 3 TP, 1 FN
- Squamous Cell Carcinoma: 54 TP, 26 FN
    - **Observations:** High performance on "adenocarcinoma" and "normal" but struggles with "squamous cell carcinoma.

Ensemble Model



Accuracy: 54.29%

Loss: Not explicitly measured

Confusion Matrix:

- Adenocarcinoma: 65 TP, 47 FN
- Large Cell Carcinoma: 1 TP, 45 FN
- Normal: 0 TP, 0 FN
- Squamous Cell Carcinoma: 38 TP, 28 FN
    - **Observations:** The ensemble model improves some aspects but still struggles significantly with "squamous cell carcinoma."

Executive Summary and Findings

Leveraging transfer learning, the pre-trained models obtained improvements in detection, culminating in an ensemble that combines each of their strengths.

The following findings summarise key findings of the various models used:

1. Simple CNN Model: highlights the limitation of simpler models in handling complex datasets. Achieved an accuracy of 37.14%
2. Pre-trained Models :benefited from transfer learning, improving feature extraction and performance. The models VGG16, ResNet50, and InceptionV3 achieved an accuracy of 54.92%, 58.73%, and 56.19% respectively.
3. Ensemble Model: generalizable classification performance demonstrated benefits of ensemble learning. Achieved an accuracy of 54.29%.

In comparison to the Keras model, utilising the pre-trained models expedited the increase in performance as their depth and features are suited to handle complex tasks. While accuracy improvements were minimal in the ensemble's use, the detection of all cancer classes increased. To compensate for these advantages, the demand for resources and processing time increased alongside overfitting. The presence of overfitting causes the models to require human intervention in fine-tuning weights for generalisation.

Closing

The ensemble model demonstrates the successful application of convolutional neural networks for lung cancer detection. The combination of pre-trained models and the ensemble approach in this tool improved its classification accuracy for non-small cell lung cancer.

It is in the project's interest to have this tool be of aid to radiologists for accurate diagnosing, early detection, and ultimately the increase of survivability in patients.

Bibliography

AJ_Buckeye, Cukierski, W.,Josette_BoozAllen, Kriss, J., Nilofer, Sullivan, J., O'Connell, M. "Data Science Bowl 2017." Kaggle. 2017. https://kaggle.com/competitions/data-science-bowl-2017.

Ardila, Diego, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, et al. "End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography." Nature News, May 20, 2019. https://www.nature.com/articles/s41591-019-0447-x

Hancock, M.C., and J.F. Magnan. "Lung nodule malignancy classification using only radiologist quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods". *SPIE Journal of Medical Imaging 3*, no.4 (2016): 044504. doi: 10.1117/1.JMI.3.4.044504

Li, P., S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. "A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx)." The Cancer Imaging Archive, 2020. doi:10.7937/TCIA.ABC123XYZ (Cancer Imaging Archive Wiki)

"LIDC-IDR (Lung Image Database Consortium and Image Database Resource Initiative) Dataset. A completed reference database of lung nodules on CT scans (LIDC-IDRI)." The Cancer Imaging Archive. https://www.cancerimagingarchive.net/collection/lidc-idri/

"Lung Cancer - NHS." NHS choices, n.d. https://www.nhs.uk/conditions/lung-cancer/.

Mohamed H. "Chest CT-Scan images dataset." Kaggle, 2020. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

Radiological Society of North America (RSNA) and American College of Radiology (ACR). "What Are the Benefits of CT Scans?" Radiologyinfo.org. https://www.radiologyinfo.org/en/info/safety-hiw_04.

Vaz, Joel Markus, and S Balaji. "Convolutional Neural Networks (Cnns): Concepts and Applications in Pharmacogenomics." Molecular diversity, August 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8342355/

Zimlich, Rachael. “Large Cell Lung Carcinoma: Symptoms, Treatment, and Outlook.” Edited by Rena Goldman, Sara Giusti, Afton DeLucca, and Siobhan DeRemer. Healthline, June 5, 2024. https://www.healthline.com/health/lung-cancer/large-cell-carcinoma.