1. Introduction

1.1 Dataset review

Of the two main categories of lung cancer, non-small cell lung cancer (NSCLC) is the most common, making up 80-85% of all cases. Within NSCLCs, the three most common subtypes are: adenocarcinoma, squamous cell carcinoma and large cell carcinoma.

Adenocarcinoma:

The most common form of lung cancer accounts for an estimated 40-50% of all cases and is the leading cause of cancer death in the United States. Due to the amount of carcinogens present, smoking tobacco is the primary risk factor associated with lung adenocarcinoma diagnosis, both through primary and secondary inhalation.

When viewing a CT scan, common signs of adenocarcinomas include:

- Peripheral masses in the outer regions of the lung, particularly those which look uneven or spiculated.
- Ground-glass opacity - hazy, whitish regions in the lung - which may be indicative of early stage adenocarcinoma.
- Pleural effusion - fluid accumulation near the lining of the lung - appears to be large white, opaque masses.

Squamous cell carcinoma:

The second most common form of lung cancers, squamous cell carcinoma (SCC) cases make up ~30% of cases. Whilst smoking tobacco is associated with 80% of lung cancer cases in men, and 90% in women, SCC cases are most strongly linked with smoking compared to the other subtypes. Squamous cell growths are commonly found either in the left or right bronchus, or in more central areas of the lungs.

When viewing a CT scan, common signs of SCCs include:

- Centrally located masses, near the main airways or the hilum.
- Cavitation - an air-filled space surrounded by tissue.

Large-cell carcinoma:

Less common than adenocarcinomas, large-cell carcinoma (LCC) makes up 10-15% of NSCLC cases. It is characterised by aggressive and rapid growths compared to other forms of lung cancers, and is often harder to detect in early stages. Due to its less defined characteristics compared to adenocarcinomas and squamous cell carcinomas, LCCs are defined as “undifferentiated non-small cell carcinomas”, which is an encompassing term to account for non-small cell carcinomas which cannot be categorised as either adenocarcinoma or squamous cell carcinoma.

When viewing a CT scan, common signs on LCCs include:

- Large, peripherally located masses due to the rapid growth rate.
- Pleural effusion, similar to adenocarcinoma.

LCCs can be harder to distinguish from adenocarcinomas and SCCs, and are often diagnosed via lack of features unique to adenocarcinomas (ground-glass opacity) and SCCs (cavitation).

As each carcinoma exhibits unique features, deep learning techniques such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) may be used to classify such images. This paper will focus on comparing the performances of the two types of models.

Additionally, baseline models of each type will be compared to pre-trained models to determine the impact of using pre-trained datasets in cancer detection.

1.2 Convolutional Neural Networks

A Convolutional Neural Network is a type of deep learning algorithm, which is used for visual analysis. Common applications of CNNs include image recognition or object detection such as facial recognition or medical image analysis.
CNNs utilise a series of layers, which detect different features of an image input, starting with simple features in the initial layers such as lines and textures, and using learned patterns in subsequent layers to capture features of increased complexity.
A CNN contains three types of layers: convolutional layers, pooling layers and a fully-connected layer.

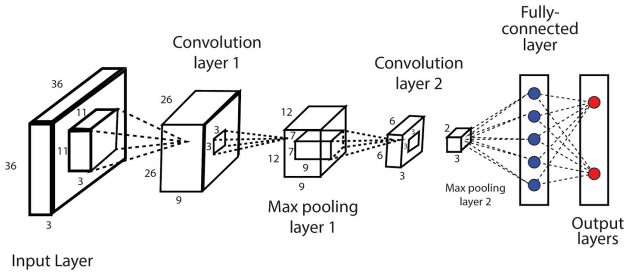

Figure 1: Basic architecture of a CNN

Convolutional layers are the fundamental building blocks of CNN models, and is where the majority of computation occurs. Within this layer, a filter or kernel - a small matrix of weights - is used to move across an image to detect the presence of specific features. The kernel slides over the image's entire width and height over multiple iterations, and calculates a dot product between the pixel values and the kernel's weights. In doing so, the input is transformed from an image to a set of convolved features, each of which represents the presence and intensity of a specific feature at various points in an image. CNNs often include multiple stacked convolutional layers to progressively learn new and more complex information.

Following each convolutional layer is a pooling layer. Whilst the pooling layer also iteratively slides over the image, the main function of a pooling layer is to reduce dimensionality whilst retaining critical pattern information. While some data is lost, the pooling layer helps to improve model efficiency and prevents overfitting by decreasing the number of data points in the input.
The most common type of pooling is max pooling, which retains the maximum value within a certain kernel size, while discarding all other values.

The fully-connected layer is the final layer of a CNN model, and functions to classify images based on the features extracted from the previous layers. In this layer, each neuron is connected to a corresponding neuron in the following layer, which is not the case in the other layer types. CNNs typically contain 1-3 fully-connected layers within an architecture.

1.3 Vision Transformers

Vision Transformers (ViT) are relatively new architectures which were created in 2020 by Google Research. They are an adaptation of the Transformer architecture, which was predominantly used in natural language processing applications. Dosovitskiy et al. (2020) highlight the high performances of ViTs when trained on large datasets, and applied to small to medium sized tasks, stating its ability to outperform the traditional CNN models which are the current benchmark.

Vision Transformers operate by dividing an image into multiple patches of a fixed size. These patches are then flattened and linearly embedded, similar to token embedding in NLP which are then processed by the transformer. Positional encodings are added to the patch embeddings to keep information regarding the position in the original image because transformers don't naturally understand or process sequential data. Before being processed by the transformer, a class token is added which is used to aggregate and hold information from the other patches during processing.

The transformer itself is made up of multiple layers, each containing multi-head self-attention mechanisms and position-wise feed-forward networks. Self-attention mechanisms enable each patch to interact with every other patch. For each patch, the model calculates how much it should pay attention to other patches when understanding a part of the image, to help gather contextual information. A position-wise feed-forward network helps to enhance feature representation of each token.

Following the transformer layers, the representation of the class token is fed into a neural network - often just a linear layer - to produce final class predictions.

Vision Transformers are usually pre-trained on large datasets, with fine-tuning on smaller datasets enabling the model to adapt to specific requirements.

2. Literature Review

The field of medical imagery has been significantly improved following the creation and implementation of deep learning technologies such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These architectures have the ability to increase the precision and effectiveness in the diagnosis of cancer, which is essential for providing patients with adequate and quality care.

2.1 CNNs in Medical Imaging

CNNs have been essential to the automation of medical image analysis since their introduction. They are renowned for their ability to handle a wide variety of imaging tasks including pathological classification, organ segmentation and cancer detection (Mohammed et al., 2024) due their ability in capturing complex data patterns in images. Additionally, Mohammed et al. (2024) emphasize how CNNs have the potential to reduce diagnostic time and human error through the automation of image processing.

2.2 Vision Transformers in Medical Imaging

Shisu et al. (2024) describe the innovative approach of ViTs, involving processing images as a sequence of patches. In doing so, the positional context of each patch is can analyzed to capture complex patterns in images. This method is effective in diagnostic scenarios where contextual relationships within the image is crucial, such as tumor detection within lungs.

2.3 Baseline and Pre-trained Models

The use of baseline and pre-trained models have become a focal point in discussion regarding the optimization of machine learning applications in medical imaging. Baseline models require extensive training from scratch, but can be tailored to specific datasets and contexts. In contrast, pre-trained models can capitalize on previously learned features from extensive datasets, helping to enhance their effectiveness and efficiency in new, related tasks. For example, pre-trained models such as ResNet and DenseNet have been adapted to improve medical imaging performance by providing a larger and more robust feature set for diagnostic tasks, even with limited additional training (Cai et al., 2020; Mohammed et al., 2024). Currently, training takes a large amount of computational power, making the training of baseline models on context-specific datasets to the same extent of pre-trained models difficult and ineffective without significant resources.

2.4 Challenges

Despite their advantages, the integration of CNNs and ViTs into everyday practice produces its own challenges, such as the need for substantial computational resources and access to a large database of correctly annotated images. The reliance on extensively annotated data may be reduced by developing more efficient models or exploring semi-supervised learning approaches (Cai et al., 2020; Mohammed et al., 2024) which is a key focus of future research.

3. Methodology

3.1 Data Acquisition

The dataset was obtained from Kaggle, and contained 1,000 labelled CT scans in total, which were split into train (613), validation (72) and test (315). Within each of these categories images were sorted by each of the four class labels: normal, adenocarcinoma, large-cell carcinoma and squamous cell carcinoma.

3.2 Preprocessing

Images were resized to **224x224** to ensure a fair comparison between VGG and the baseline model, standardising the input resolution so that both models are trained on similar levels of detail.. The size also aligns with VGG's pre-trained configuration, facilitating consistent transfer learning. By controlling for input size, we can focus on comparing the models' architectures rather than differences in image resolution or computational load.

Data augmentation was also utilised by flipping images horizontally, and added back into the dataset. This helps to both increase the diversity of the training data and increase the size of the dataset without the need for additional images. Ideally, this will aid the model to generalise better during training. Similarly, minor variations to the dataset were implemented such as rotations, which helps to increase training size without removing key patterns within the CT scans.

Pixels were also rescaled from 0-255 to 0-1 to normalise the data, helping to decrease training times. This was done so the model is able to handle gradients more efficiently during backpropagation.

3.3 Modelling

3.3.1 Baseline CNN

To evaluate the effectiveness of the VGG model, a baseline CNN was created as a benchmark to compare with on the Kaggle Dataset. The baseline model utilizes several convolutional layers, starting with an initial layer containing 32 filters of 3x3 kernels each, and employing ReLU activation. The filter sizes increase with each layer up to 512, allowing for more complex patterns to be captured at each level. This is key to cancer diagnosis, as the presence of cancer in CT scans may be hard to detect, meaning a high level of granularity is required. Each convolutional layer is followed by batch normalization and 2x2 max pooling. This helps to lower computational requirements and training time.

Following these layers, a dense network segment helps to flatten the output to prepare for classification. The dense network utilizes a 512-unit layer with a 50% dropout rate, which is standard for image classification tasks, and ends with a softmax layer which classifies the output into one of the four categories.

The Adam optimizer is used to compile the model as it is known for its computationally efficient learning rate.

3.3.2 Pre-trained CNN

The chosen pre-trained CNN model was the VGG-16, due to its powerful feature extraction capabilities alongside its versatility. VGG-16 is a CNN model developed by Karen Simonyan and Andrew Zisserman from the Visual Geometry Group at the University of Oxford, and is trained on over 14 million images across 1000 categories in the ImageNet dataset. Here, VGG-16 has leveraged transfer learning to detect and classify chest cancer in CT scans. Additionally, VGG-16's relatively simple architecture enables fine-tuning on smaller datasets compared to more modern and larger architectures.

VGG-16 contains 21 layers, consisting of 16 trainable layers and 5 non-trainable layers. There are 13 convolutional layers, which use 3x3 convolutional filters, with 5 max-pooling layers dispersed every 2-3 convolutional layers. Each max-pooling layer makes use of a 2x2 kernel with a stride length of 2 to reduce dimensionality and improve efficiency. Following the convolutional and max-pooling layers are 3 fully-connected layers. The first layer takes the flattened output from the first 18 layers, and connects to 4096 neurons with ReLu activation to allow for non-linear transformations. The second layer is similar in format to the first, also consisting of 4096 neurons with ReLu activation. The third contains 1000 neurons, corresponding to the 1000 categories of the ImageNet Dataset. Here, softmax activation is used instead of ReLu to output a probability distribution which can be easily interpreted.

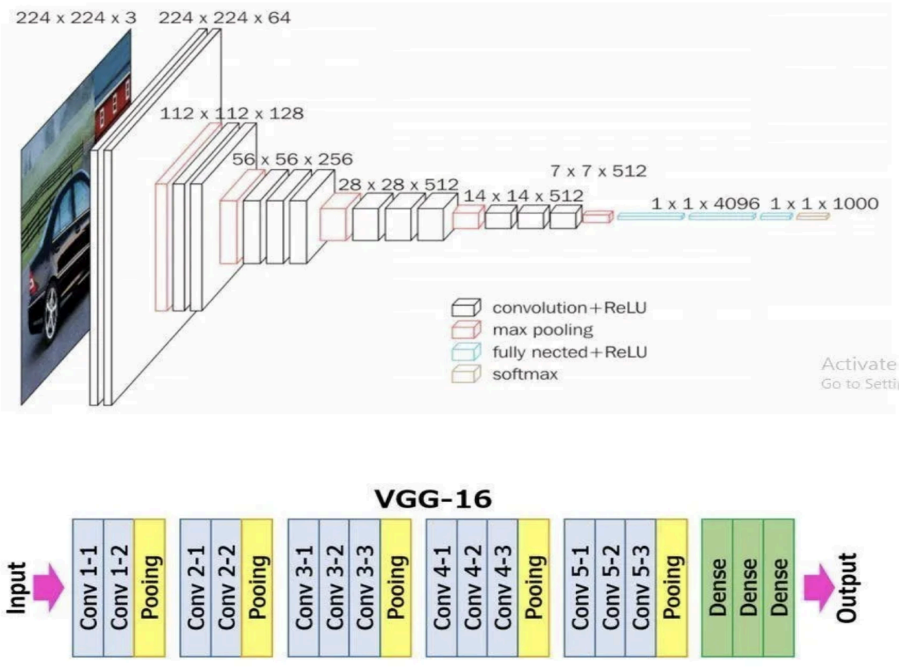

Figure 2: Visualisation of the VGG-16 architecture

3.3.3 Baseline ViT

A baseline ViT model was also created to compare with the pre-trained ViT. The baseline model is a Hybrid Vision Transformer (HybridViT), which initially extracts features using CNN layers to capture local patterns before analyzing the patterns in a broader context from the Vision Transformer. A hybrid model was also chosen because it requires less training data compared to a pure Vision Transformer, which is ideal due to the limited number of training instances, even with augmentation.

The model utilizes the pre-trained ResNet18 for the CNN layers, which are then reformatted into sequences of patches and treated as individual tokens by the Vision Transformer. This was done to maximize the efficiency of the model, whilst ensuring a baseline comparison can be made with the Vision Transformer.

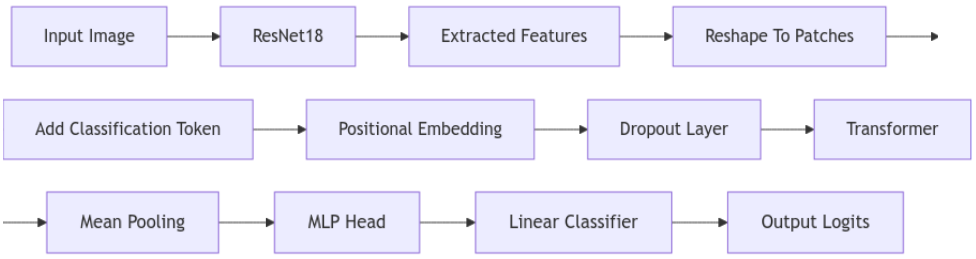

Figure 3: Hybrid ViT Architecture

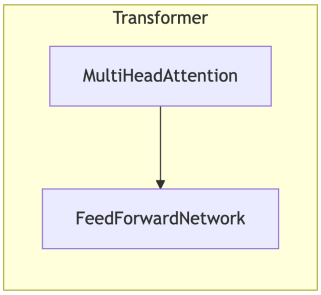

Figure 4: Transformer Architecture

3.3.4 Pretrained ViT

The ViT_B_16 model from PyTorch splits images into 16x16 pixel patches. It then processes these patches similarly to a regular transformer model. The model is usually pretrained on the ImageNet dataset in a supervised way, which learns to classify images across many classes.

The model outputs logits for each image, which are then converted to class probabilities using softmax. The predicted class for each image is determined by taking the class with the highest probability.

4. Results

4.1 CNN Result

Model Configuration	Test Accuracy
Baseline CNN	45.4%
Pretrained CNN with augmented images	88.5%

Table 1: CNN Test Accuracy

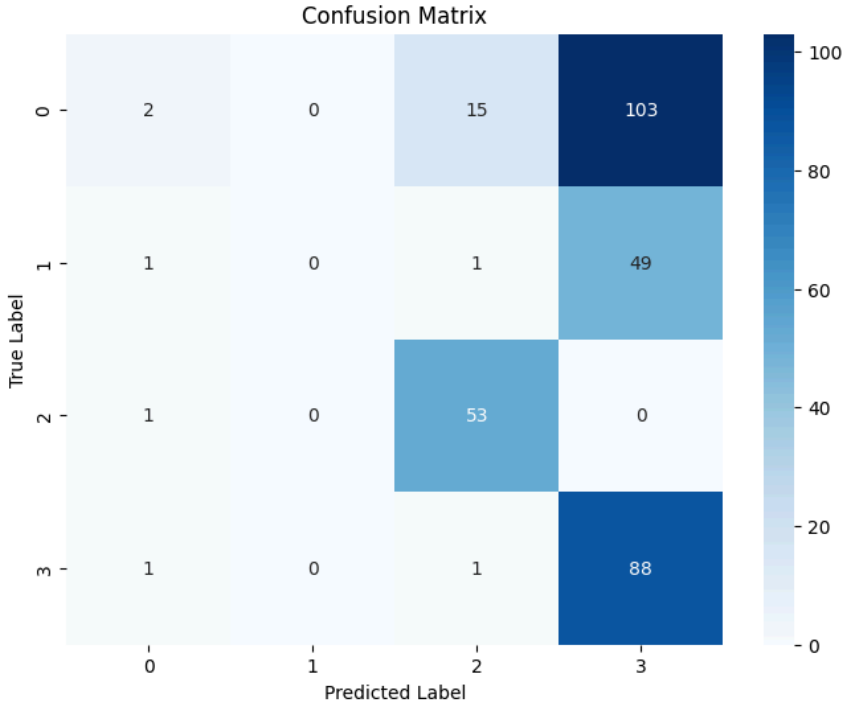

Figure 5: Baseline CNN model

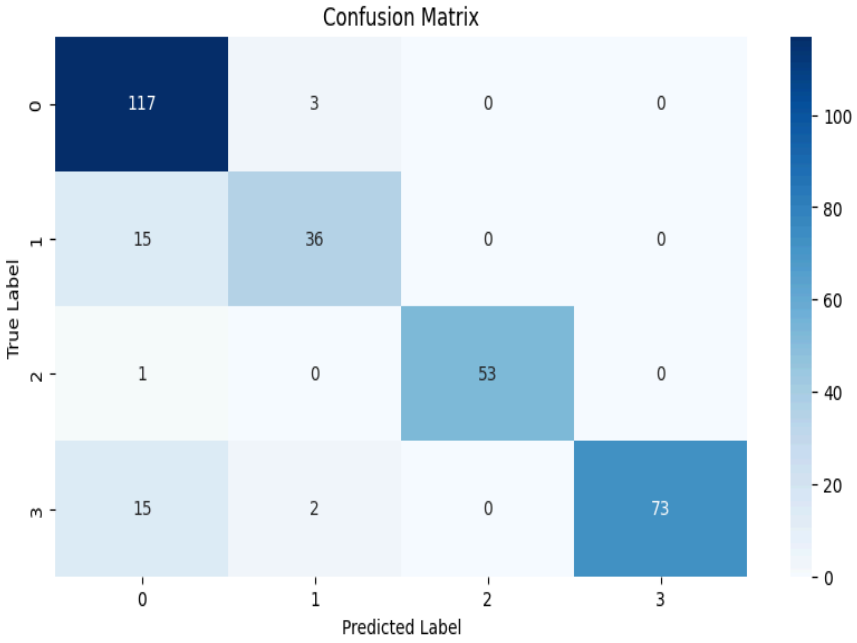

Figure 6: Pretrained CNN model

4.2 ViT Result

Model Configuration	**Test Accuracy**
Pretrained ViT	43.49%
Pretrained ViT with Augmented Images	33.33%
Pretrained ViT with Equalized Images	45.08%

Table 2: Pre-trained ViT Test Accuracy

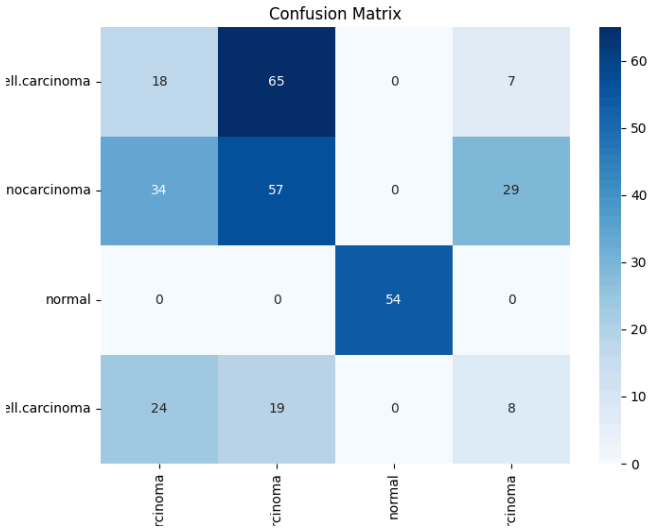

Figure 7: Pretrained ViT Confusion Matrix

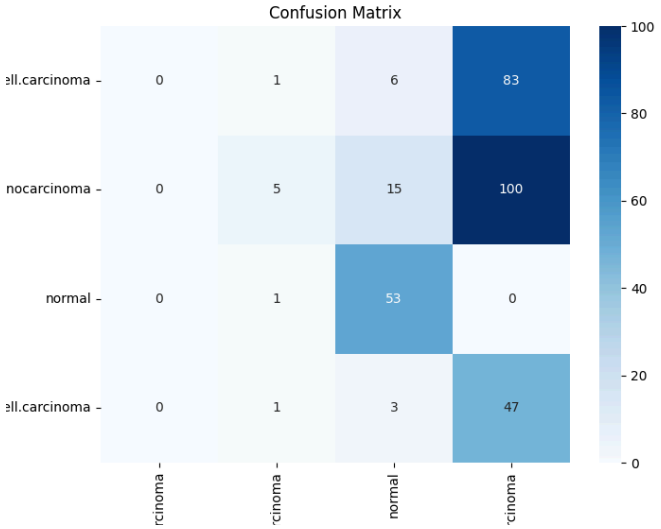

Figure 8: Pretrained ViT with Augmented Images Confusion Matrix

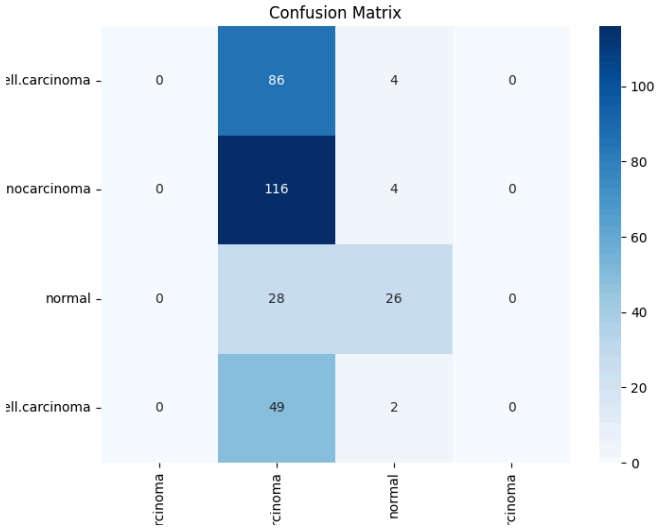

Figure 9: Pretrained ViT with Equalized Images Confusion Matrix

4.3 Hybrid ViT Results

Model Configuration	**Test Accuracy**
Hybrid ViT	70.16%
Hybrid ViT with Augmented Images	35.87%
Hybrid ViT with Equalized Images	54.60%

Table 3: Hybrid ViT Test Accuracy

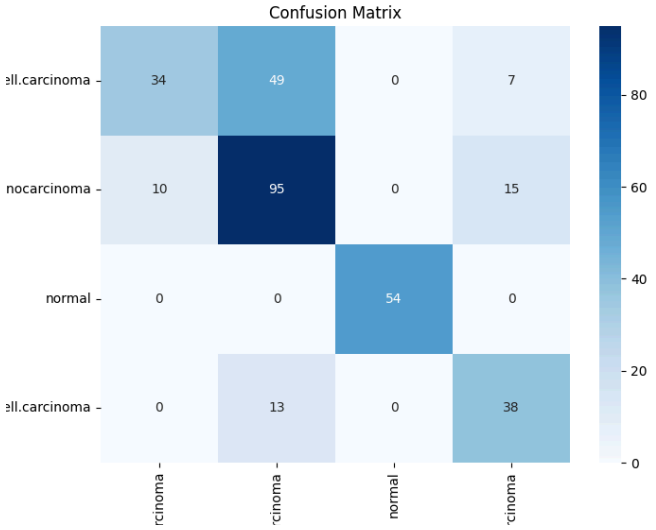

Figure 10: Hybrid ViT Confusion Matrix

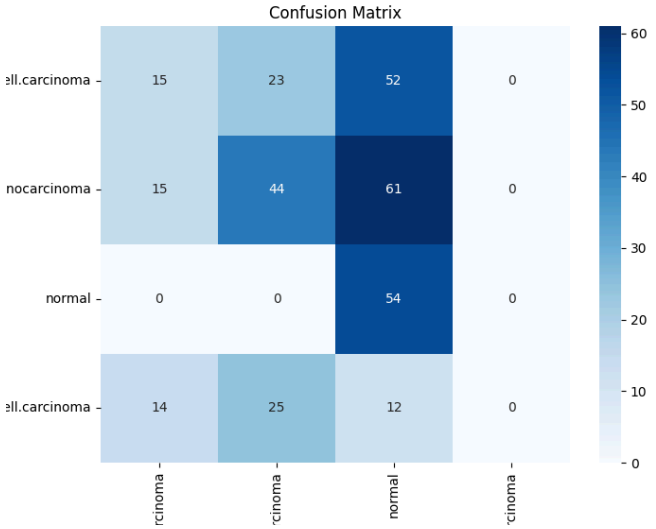

Figure 11: Hybrid ViT with Augmented Images Confusion Matrix

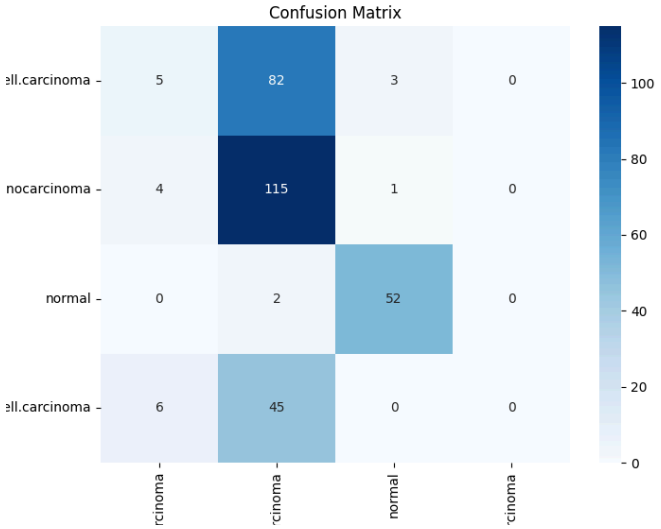

Figure 12: Hybrid ViT with Equalized Images Confusion Matrix

Recall

	Original Images	Augmented Images	Augmented and Equalized Images
Large-cell carcinoma	37.8%	16.7%	5.6%
Adenocarcinoma	79.2%	36.6%	95.8%
Normal	100%	100%	96.3%
Squamous cell carcinoma	74.5%	23.5%	0%

Table 4: Recall metrics

5. Discussion

5.1 CNN

For the baseline CNN model, its shallow architecture may not be well-suited for a complex task like predicting cancer types from CT scans. Cancer detection from CT scans requires capturing subtle, fine-grained patterns, and this often demands deeper and more sophisticated models to effectively capture small but crucial details in the images. To address these challenges, feature refinement techniques and architectural enhancements were introduced to a pretrained model combined with VGG-16 model to improve feature learning and overall performance.

In the enhanced pretrained model, a Squeeze-and-Excitation (SE) block is incorporated to increase the network's representational power by adaptively recalibrating the importance of different feature channels. This SE block enables the model to focus more on relevant channels for the specific task, dynamically emphasizing important features while suppressing less useful ones. As a result, this selective weighting improves feature selection and boosts model accuracy.

Additionally, a Residual Block with SE is used to create skip connections, which allow information to bypass certain layers. These skip connections help alleviate the vanishing gradient problem, which can hinder learning in deep networks, and support the development of deeper architectures. Consequently, these enhancements enable the model to learn more complex patterns, leading to a substantial increase in accuracy from 45.4% to 88.5%.

5.2 ViT

Based on the results, the Hybrid ViT with unprocessed images achieves the highest performance, followed by the Hybrid ViT with equalized images. This suggests that neural networks like Vision Transformers can map input data to meaningful feature representations without requiring extensive preprocessing.

Data augmentation is performed through random rotation, horizontal flipping, adjustments to color properties, and resizing. Theoretically, these techniques help make the model more robust to varying features. However, Hybrid ViT performs worse on augmented images compared to the original images, indicating that Hybrid ViT is not robust enough to handle varying features effectively. When histogram equalization is applied to the augmented images, the test accuracy improves. This suggests that histogram equalization can enhance Hybrid ViT's robustness to images with changing features.

Hybrid ViT is especially good at predicting the normal class. For the recalls, normal class attains 100% for original and augmented images, and 96.3% for equalized images. Hybrid ViT performs well for adenocarcinoma and squamous cell carcinoma when it is trained with the original data. The recalls for these two classes are high, being 79.2% and 74.5% respectively. Although Hybrid ViT trained with equalized and augmented images attains higher test accuracy than solely augmented images, it does not learn to classify squamous cell carcinoma, with a recall of 0.

In the scenario of lung cancer, both false positives and false negatives should be avoided. Patients with normal chest CT scans might be classified as having lung cancer. A model attaining balanced recall and precision for all classes is preferred.

6. Conclusion

From the results, we can see that CNN performs better than ViT. This is likely due to the dataset size, as ViT models typically require larger training datasets. Takahashi et al. (2024)

highlight that Vision Transformers require larger datasets for effective training compared to CNNs, which can be critical in situations with limited data. Our dataset only has 1012 images for 4 classes, and only 617 of them are used for training. Takahashi et al. (2024) mention that the reason ViTs demand more data is that CNNs can detect local spatial relationships, such as edges and textures, while ViTs rely solely on their self-attention mechanism, which does not prioritize local features.

References

Clinical characteristics and treatments of large cell lung carcinoma: a retrospective study using SEER data - PMC (nih.gov)

Lung Adenocarcinoma - StatPearls - NCBI Bookshelf (nih.gov)

Pleural effusion | Radiology Reference Article | Radiopaedia.org

https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network

Convolutional Neural Networks Explained (CNN Visualized)