National College of Ireland

Identification and Classification of Cancer from Images using CNN and Vision Transformers

MSc Research Project
MSC Data Analytics

Aniket Rane
Student ID: 23299720

School of Computing
National College of Ireland

Supervisor: Prof. Shubham Subhnil

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aniket Rane
Student ID:	23299720
Programme:	MSC Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Prof. Shubham Subhnil
Submission Due Date:	28/01/2026
Project Title:	Identification and Classification of Cancer from Images using CNN and Vision Transformers
Word Count:	6426
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	22nd January 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identification and Classification of Cancer from Images using CNN and Vision Transformers

Aniket Rane
23299720

Abstract

Traditional methods of detecting cancer rely on invasive methods and tend to be often time-consuming, and early detection is crucial to increase the chances of survival for the patient. Deep learning (DL) is a subset of machine learning that utilizes neural networks that have a structure similar to the neurons inside a human brain and have the ability to learn complex patterns from data and this ability of deep learning models makes them suitable for tasks like image recognition, natural language processing, and speech recognition. In this study, the aim is to train deep learning models, specifically Convolutional Neural Networks (CNN's) and Transformer-based models. In this study, a comparison between these two models has been done on the basis of different metrics like accuracy, precision, recall and training time. The dataset that was used for the study is a multi-cancer dataset which was formed after combining a few different cancer datasets and is available on Kaggle. The dataset consists of images for eight different types of cancer (cervical, lung, colon, oral, kidney, breast, lymphoma and brain) and 26 subclasses. Our research focuses on lung, colon, oral, lymphoma, and breast cancer and their related subclasses and few CNN and Transformer models including the vision transformer (ViT) were trained on this dataset in order to test which type of models are most suitable and perform the best at accurately recognizing different cancer types. The dataset consists of 60,002 cancer images and the findings of this study might help future researchers in improving the patient outcomes and provide a direction for further research in this domain.

Keywords - Deep learning, ResNet, CNN, cancer, cancer classification, Lung cancer, Lymphoma, Oral cancer, accuracy, Vision Transformer, zero-shot, few-shots

1 Introduction

Cancer is one of the deadliest diseases around the world and is a leading cause of mortality. Around 15% of the deaths worldwide occur due to cancer. The highest percentage of deaths occurring due to cancer are due to Lung and Colon cancer. These two types of cancer have also accounted for nearly 2.8 million deaths in 2020 alone Hasan et al. (2023). This prevalence of these cancers and their impact on death rates poses a challenge for both the researchers and medical practitioners as early-detection is crucial to ensure a positive treatment outcome Patharia et al. (2024). Despite, the existence of many manual diagnostic techniques and visual examinations, early-detection is still a challenge as these traditional methods often tend to be time-consuming, are invasive in nature and lack the

capability of identifying multiple types of cancer simultaneously Kumar et al. (2024). Computed Tomography (CT) scan can help with early detection of lung cancer and in the resulting scans, lung cancer appears as a nodule. To detect these nodules Computer-aided diagnosis (CAD) systems are widely used but still classification remains a challenge for these Gai et al. (2024).

One of the leading causes for deaths among women is breast cancer Xu et al. (2025). It is more common among older women but can affect younger women too in some cases. The treatment options include surgery, radiotherapy, chemotherapy, and targeted therapies depending on the type and severity. The important sign of it is a breast lump and most of the lumps (9 in 10) are not cancerous but it's always best to get checked by a general practitioner (GP). Early detection can help to start the appropriate treatment at the early stages hence, improving survival rates.

Abnormal cell growths that occur when the cells on the lips or in the mouth mutate are termed as oral cancer. It is the most common form of head and neck cancer and usually affects people aged 60 and older. Oral Squamous Cell Carcinoma (OSCC) is the most prevalent subtype of it and typically presents as localized lesions across oral regions like the tongue, floor of the mouth, lips, etc.Tafala et al. (2025). It usually starts as a painless red or white patch, that thickens overtime and continues to grow. Symptoms include pain or difficulty swallowing, lumps in the neck, swelling or numbness in the mouth or lips. Treatment options include surgery, radiation therapy, targeted therapy, or chemotherapy, or a combination of these depending on the size, location, and spread of the oral cancer. Traditional diagnostic methods such as histopathology which involves the microscopic analysis of tissue samples is generally time-consuming and depends on the expertise of pathologists and remains the standard and widely used method for diagnosing oral-cancer Singh et al. (2024).

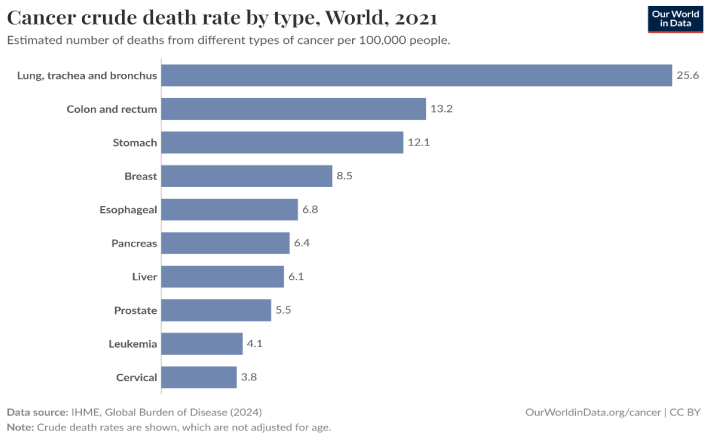

Figure 1: Deaths from different types of cancer per 100,000 people worldwide (Source: Our World in Data (2024))

When it comes to any type of cancer, early detection and accurate diagnosis is crucial. Early discovery of malignant tumours and precise classification could help both in defining an appropriate treatment and increasing the chances of survival of the patient Verma (2024). This is where deep learning could play an important role. To tackle the problem of the accuracy of diagnostics related to cancer Provath et al. (2023), researchers have carried out studies using deep learning models like Convolutional Neural Networks (CNN's). If we find a method that could help detect cancer in its very early stages, the survival rate

could be improved further, thus reducing the number of deaths. As the applications of deep learning continue to grow, current deep learning networks are very accurate and can easily learn features automatically from the data itself Hamida et al. (2021). For years now, CNN's have dominated the tasks related to computer vision, but the newer Vision Transformers (ViT's), when pre-trained on large datasets and then applied to the small or mid-sized datasets, attain better results than the CNN's and require substantially lower computational resources Dosovitskiy et al. (2021). Hence, doing a comparison between models like Vision Transformer (ViT) which is a transformer based model and are good for image data tasks related to classification and segmentation, and CNN's could help researchers understand which type of models are more relevant to these type of problems related to medical image classification and provide higher accuracy which is crucial in this field. This research aims to do this and would attempt to provide clarity on which type of deep learning models are more suitable, provide accurate results, and are more robust and provide higher accuracy for different cancer image types included in the dataset. This research would also aim at implementing the approaches of zero-shot and few-shots training for the cancer image dataset also collectively termed as internal transfer learning. This paper follows a structured approach where under Section 2 literature review of relevant research papers from the selected domain is performed. Section 3 outlines the methodology that was followed in order to successfully implement the selected techniques. Section 4 gives the overview of the followed process in the form of an architecture diagram. Section 5 discusses how the models were implemented. Section 6 outlines the different experiments that were performed with the selected models and finally, Section 7 provides the conclusions of the performed research and suggests some improvements that could be made in future research papers using similar approaches.

2 Related Work

This section includes a comprehensive review of recent state of the art research studies done in this domain. Several research papers related to Lung and Colon cancer classification using CNN and ViT models have been reviewed in the 1st subsection. 2nd subsection discusses papers doing a comparison between CNN and ViT in the domain. The 3rd subsection highlights some challenges that were faced by previous research studies in this domain. And the 4th subsection reviews the work related to zero-shot and few-shots training.

2.1 Lung and Colon cancer

Deep Learning models have previously shown promising results when it comes to image classification tasks. They have also proved to be accurate and effective when it comes to detection of cancer from histopathological and CT images. Hasan et al. (2023) in their research applied four different deep learning models including vision transformers and they achieved excellent results with both CNN and ViT attaining a perfect accuracy of 100% with the ViT based model. Their research was based on the histopathological images for Lung and Colon cancer (LC25000 dataset) but, achieving 100% accuracy is something that's questionable in a real-world scenario and puts the model at question whether it generalizes well or not and indicates signs of overfitting. Also, the researchers have pointed out the need for their approaches to be validated on larger datasets. A comprehensive review of various researches carried out on lung and colon cancer image

datasets was done by Patharia et al. (2024) and they have highlighted the major challenges that are faced by the researchers when it comes to classification of lung and colon cancer using Deep Learning methods. The key challenges highlighted include overfitting issues because of deep learning models being trained on limited image datasets due to which the models are not able to generalize well on unseen data. Also, scarcity of annotated datasets hinders the development of robust algorithms. They have highlighted that a combined research involving images of multiple different cancer types would help in developing a model that generalizes well and hence, the research question for this study attempts to work in that direction. Hamida et al. (2021) applied several CNN models including ALEXNET, VGG, RESNET, DENSENET and INCEPTION models to colon cancer datasets namely AiCOLO, CRC- 5000, and NCT-CRC-HE- 100K and utilized the approaches of from-scratch training and transfer learning, in which case the strategy of from scratch training provided better results. They achieved an excellent accuracy of 99.5% with the SEGNET model. For implementing Transfer Learning, they utilized models pre-trained on the ImageNet dataset. Although, as highlighted by them, SEGNET increases computational costs and hence, emphasize that future research should focus on optimizing the computational costs. The newer ViTs could achieve better performance while utilizing lesser computational resources. The CNN architecture proposed by Verma (2024) achieved an overall accuracy of 99% which is exceptional and was able to properly differentiate between benign and malignant tissues of lung and colon cancer. They have pointed out to the fact that future research needs to be conducted on larger datasets and hence, suggest more robust models. Ochoa-Ornelas et al. (2025) implemented the strategy of transfer learning utilizing an EfficientNetB3 model pre-trained on the ImageNet dataset and configured the final classification layer. Further, they augmented the dataset by replacing 1,000 cancerous images with images from the Genomic Data Commons (GDC) Data Portal to ensure the model is able to generalize well. On this version of combined LC25000 (Lung and Colon cancer images) and GDC dataset they obtained an accuracy of 99.39%. They also tried implementing the model only on the GDC subset of images separately and achieved an accuracy of 96.71% which is excellent and also suggests that the model is able to generalize well. This research will aim to train a similar model from scratch and implement internal transfer learning and evaluate the model's performance on multiple cancer types to ensure model generalizability.

2.2 Comparative studies for CNN and ViT

A comparative study between CNN and ViT was carried out on a lung cancer dataset of 212 images by Gai et al. (2024) and achieved an excellent accuracy of 95.3% with the self-supervised learning approach and EfficientNet3D CNN model. In their case, the CNN performed better than ViT due to the smaller dataset size and have highlighted the need for performing future research based on larger and varied datasets in which case ViT's could potentially provide more accurate and robust results. A similar methodology of comparing the performance of ViTs with CNNs was followed by Manali and Demirel (2025) for classifying brain cancer images achieving the highest accuracy of 84.39% with the ViT based model which is good considering the complexity of brain MRI images. They could have achieved an even better accuracy by using the approach of internal transfer learning or few shots training.

2.3 Challenges faced in the domain

Prasad and Anbarasi L (2024) in their study "Acute Lymphoblastic Leukemia Subtypes Detection using Vision Transformer Model" applied a Vision Transformer based model for detecting ALL from microscopic blood smear images and concluded that ViTs perform better than CNN because of their attention mechanism while requiring less computational resources. They attained an exceptional accuracy of 98.01% using ViT which demonstrates the capability of Vision Transformers. They have highlighted that future researches should focus on larger datasets as it will help in ensuring model's reliability and hence, the purpose of this study. Another type of cancer that occurs in the lymph nodes is termed as Lymphoma and Aravind et al. (2024) leveraged 4 different CNN models and obtained the best results with the EfficientNetB0 attaining an accuracy of 96.2%. They have highlighted that the distinctive characteristics of histopathology images makes it important to implement deep learning models for accurate classification and implementing more powerful DL models would help in further improving the accuracy obtained. The few key challenges highlighted in their study are scarcity of public datasets related to certain types of cancer especially lymphoma, lack of sufficient data, and variability in pathology images.

Mehmood et al. (2022) utilized a pretrained neural network AlexNet and implemented an approach of 'Class Selective Image Processing (CSIP)' which involves improving the picture quality of underperforming classes and using this approach they saw a significant increase in accuracy from 89% initially to 98.4% after implementing CSIP. Their method was pretty unique and good and showcases its capability in helping to improve image quality of certain images which in return helps the model to accurately classify them but other CNN models like the EfficientNet have previously achieved similar accuracies for the LC25000 dataset without the need for CSIP.

2.4 Zero-shot and Few-shots training

Danilochkin et al. (2025) have pointed out that when it came to specialized image tasks like differentiating between muffins and chihuahuas wherein the objects in the dataset are similar and belong to the category of similar object detection, the researchers struggled to reduce the error rates. In order to work in this direction, the researchers of Danilochkin et al. (2025) implemeted CLIP based zero-shot and few shots training approaches. Basically, zero-shot training focuses on classifying unseen examples, whereas few-shot learning allows a minimal number of samples per class to be used in training Danilochkin et al. (2025). Using these approaches, they obtained excellent results, achieving an accuracy of 100% in the muffins vs. chihuahuas problem and an accuracy of 99.35% in the Cats vs. Dogs dataset. This approach could help in obtaining good results in the cancer image classification problem and hence, this study aims to do this.

The techniques of zero-shot and few-shots training haven't been tried on the cancer image datasets in previous research studies. Hence, this study will aim to implement these approaches with CNN and Transformer based models including a ViT and evaluate their performance on the selected dataset. This would help researchers understand the effectiveness of these techniques in cancer datasets and allow them to work more in this direction. These techniques would also prove to be helpful considering the domain as the model would not have always seen many images that need to be accurately classified for effective treatment, and hence applying these would ensure the robustness required in this domain.

3 Methodology

This section explains the process followed and different steps that were carried out for performing Exploratory Data Analysis (EDA) on the dataset and outlines the techniques used for image pre-processing and model building.

3.1 Dataset

The dataset utilized for the purpose of this study was put together by Naren (2022) and it originally consisted of images for 8 main cancer types and 26 subclasses with 5000 images for each subclass in a cancer type. The total number of images included in the dataset were 130,000 but for the purpose of this study images for Lung, Colon, Breast, and Oral cancer were selected for classification and evaluation with the help of deep learning models. The number of images used for this study are 60,002 as the two subclasses for oral cancer consisted of 5001 images each and the other 10 subclasses under Lung, Colon, Breast, and Lymphoma cancer types had 5000 images per subclass. It is a very comprehensive and well-maintained cancer image dataset providing a wide range of images and opportunities for research in this field.

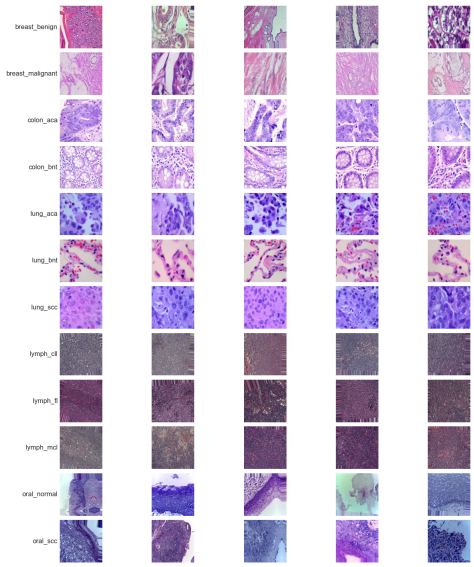

Figure 2: Image samples per subclass

Figure 2 represents five sample images from each subclass present in the dataset used for the study. Visually, we can see that there are some differences in the images from each subclass of different types of cancer and to confirm this a few metrics were calculated which will be discussed later on in this section.

3.2 Exploratory Data Analysis (EDA) for cancer images

Before starting with model building and evaluation, it is necessary to ensure that all the images present in the dataset have the same color channel format, image size, and also there must be a balance of number of images present in the various cancer image subclasses. To carry out this process, a few important checks were run. This involved checking for the folder level structure of the dataset so that accessing the images becomes easier if the same structure is followed for every cancer type. After inspection, it was found that the images were stored in the manner given below inside folders by subclass:

Dataset folder structure:
`dataset/` → `Folder for main cancer type/` → `Folder for subclass within the cancer type/` → `Image files for the subclass/`

Also, as mentioned earlier, each subclass consists of 5000 images except for the Oral cancer subclasses which consist of 5001 images each.
Following this, image sizes were checked and it was found that, all the images in the dataset had the same size of 512 x 512 pixels. The images also had the same three-channel 'RGB' color format which was a good sign.

In order to identify if there were duplicate images present in the dataset, "Perceptual hashing" was used. There are several different algorithms or methods for implementing perceptual hashing and the one that was used for this study is called "Perceptive hash" or "P-Hash". Perceptual hashing is used in various domains to identify similar, tampered, or duplicate images. P-Hash is a method in which Discrete Cosine Transformation (DCT) is used to extract sensitive data from images Aissi et al. (2024). What it does is it transforms images from spatial domain to frequency domain, where the lower frequencies represent the overall structure and higher frequencies represent the noise and the details in the image. Aissi et al. (2024) in their study carried out a comparison between six different perceptual hashing techniques and found out P-Hash algorithms evaluations to be reliable for image authentication.

Using P-Hash it was found that there were a total of 2674 duplicate images present in the dataset of which 2666 belonged to the Lung and Colon cancer subclasses, and 6 were in the breast_malignant oral_normal, and oral_scc subclasses with 2 in each of them, and the lymph_cll and lymph_fl had 1 duplicate image in each of them.

In Table 1 the distribution of duplicates per subclass is provided.

In Figure 3 the example of perceptual duplicates of an image is provided highlighting the effectiveness of this technique.

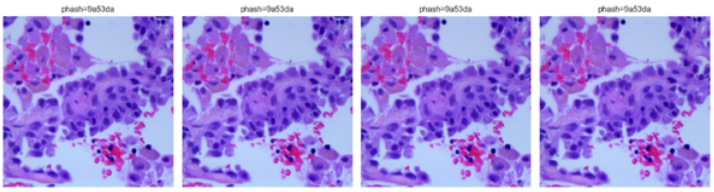

Figure 3: Example of images having same phash value (duplicates)

Table 1: Number of duplicates by subclass.

Main Cancer Type	Subclass	Number of duplicates
Lung	lung_scc	565
	lung_aca	558
	lung_bnt	544
Colon	colon_aca	517
	colon_bnt	482
Breast	breast_malignant	2
Oral	oral_normal	2
	oral_scc	2
Lymphoma	lymph_cll	1
	lymph_fl	1

3.3 Feature Embeddings

Feature embeddings were extracted from the images using a pre-trained ResNet50 model. Feature embeddings are nothing but the numerical representations of an image. These embeddings were extracted using a pre-trained ResNet50 model by removing the final classification head or layer of the model as the last layer before classification is the one that captures the semantic meaning of the images. These embeddings are calculated in such a manner that images from the same subclass would have similar embeddings while images from different classes would have dissimilar embeddings. The embeddings were then normalized in order to make further computational tasks easier. Once embeddings were calculated, they were cached to avoid repetitive calculations every time.

These feature embeddings were utilized to calculate intra-class (within the same class) and inter-class (between two different classes) similarities with the help of a function called as the (cosine_similarity) which is a part of the scikit-learn library in python to identify how different or similar the images present within one or between two different classes are. This step is crucial to understand whether the cancerous and non-cancerous images within a class and also between different classes look similar or are fairly different. This was done with 3 different sample sizes 20, 50 and 100 images per class to ensure that the similarity measurements are stable and do not biased by specific image subsets. It was found that the values did not vary much with sample size indicating consistency and reliability of the results. Heatmaps were plotted to better understand the mean inter class similarity scores wherein a heatmap was utilized per sample size.

Figure 4 shows the heatmaps for mean inter-class similarities by sample size and by class.

Figure 5 displays the bar chart for mean intra-class similarity by sample size.

The intra-class similarity values for the lung, colon, and lymphoma cancer subtypes suggest that the images within these cancer subtypes are very similar and hence the intra-class similarity scores are close to 0.7. Hence, it would be relatively easier for the model to learn such classes. Whereas, the intra-class similarity scores for breast and colon cancer subtypes suggest that the images are more diverse and hence it would be relatively difficult for the model to classify images of these subtypes.

From the heatmap for the inter-class similarity scores, it could be concluded that the images within lymphoma class subtypes are the most similar followed by lung and colon subclasses. The colon_aca and colon_bnt subclasses also show a high degree of similarity

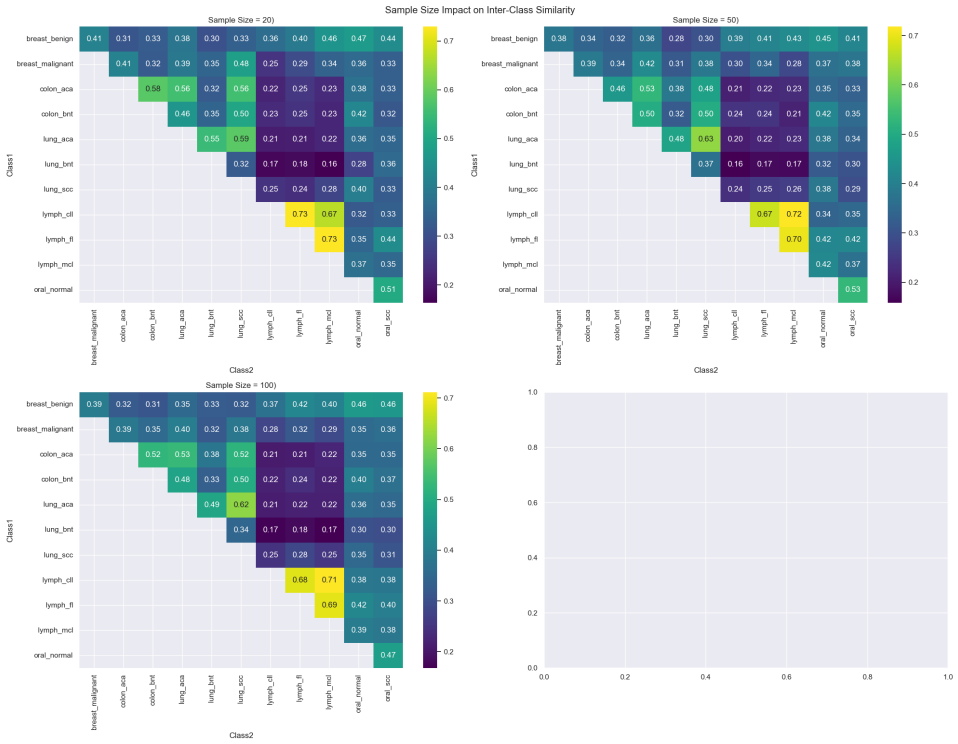

Figure 4: Mean Inter-class similarity by sample size

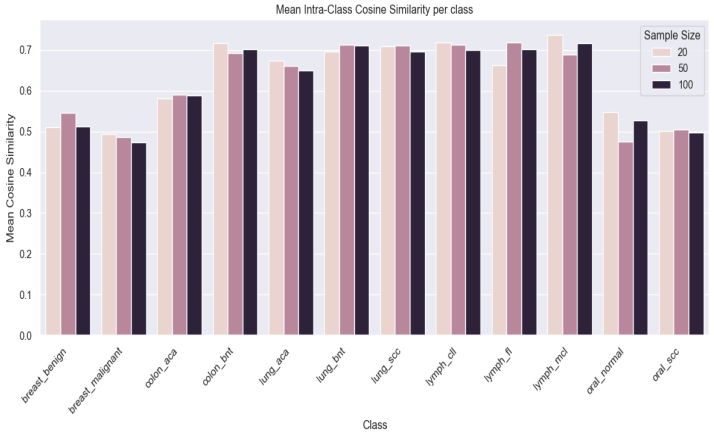

Figure 5: Mean Intra-class similarity by sample size

to lung_aca and lung_scc subtypes. The rest of the image subclasses are pretty diverse and distinguishable from each other.

Following this, diversity index was also calculated for every cancer subtype, using k-means clustering with 5 clusters per class to group images into visually coherent subclusters. The reason behind this is that a highly diverse class would naturally split into multiple well separated clusters. Intra-cluster and inter-cluster similarities were then calculated to compactness and separation between subgroups respectively. The diversity index was calculated using the formula given below:

Diversity Index:

$$Diversity\_Index = \frac{Mean\_Inter - Cluster\_Distance}{Mean\_Intra - Cluster\_Distance}$$

Table 2 gives the diversity index values for each cancer subtype.

Table 2: Diversity Index values for cancer subtypes.

Subclass	**Diversity Index**
lung_scc	1.176427
lung_aca	1.129750
lung_bnt	1.145590
colon_aca	1.158916
colon_bnt	1.204711
breast_malignant	1.157637
breast_benign	1.102379
oral_normal	1.133182
oral_scc	1.127789
lymph_cll	1.237348
lymph_fl	1.194717
lymph_mcl	1.209644

A higher diversity index indicates that a subclass is more distinct from other classes relative to its internal variability. All the subtypes in the selected dataset exhibited diversity indices 1 pointing towards good inter-class variability. The subclass breast_benign shows slightly lower separation whereas lymph_cll has more clear separation.

Before training the models, the images were pre-processed in order to ensure model robustness and training efficiency. All the images in the dataset were standardized to 224 x 224 pixels from the original 512 x 512 pixels using padding to preserve the aspect-ratio of the images taking into consideration that this is a medical images dataset. 224 x 224 pixel size works well for all the selected CNN (ResNet34, EfficientNetB2) and Transformer models (ViT_Base_16, Deit_Small_Patch16_224). During training, data augmentations were applied using the FastAI aug_transforms pipeline, including random zoom, lighting variation, warping, and image flips. These augmentations improve generalization considering the natural variability of histopathological images. Too harsh augmentations were avoided to maintain image quality as this is a medical image dataset hence, harsh augmentations would tamper image quality. Given in Figure 6 is an example of the original image and the different augmented versions of it.

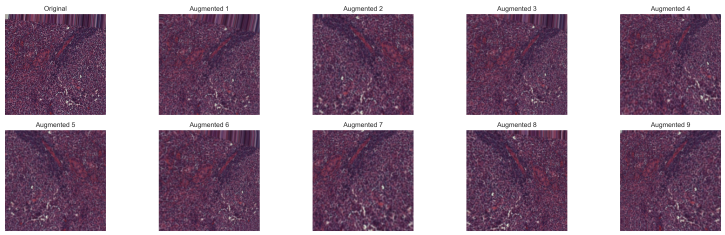

Figure 6: Original images vs the augmented versions of the same

3.4 Approaches for model training

Following the pre-processing of images, two different approaches were utilized for model training.

3.4.1 Conventional training

To address the research question, two training methods were implemented. Conventional model training involves training of selected models on the entire dataset. For implementing this approach all the models were trained from scratch and each model architecture was initialized with random weights.

3.4.2 Internal Transfer Learning

To further analyze how well do the models perform on unseen data when there are very less training samples, the approach of Internal Transfer Learning was implemented. This involves two phases:

a) Zero-shot training: Under zero-shot training the model first learns from only one subclass of a cancer type and then it is tested on the remaining subclasses from that same cancer type to analyze whether it is able to classify unseen samples.

b) Few-shots training: Now, the zero-shot pre-trained model was loaded and was trained on only 5% of images from the dataset and then tested on the remaining 95% images in one go.

This methodology allowed to assess both full-data learning and adaptability of the model within the same dataset.

3.5 Models selected

Four different image classification models were trained for implementing the two mentioned training approaches.

1. **ResNet34:** ResNet34 is a deep convolutional neural network consisting of 34-layers. The ResNets (Residual Networks) were originally developed to address the issue of vanishing gradients. They use "shortcut connections" which is added to every pair of 3 x 3 filters which allows them to flow directly to earlier layers He et al. (2016).

2. **EfficientNetB2:** This model architecture was introduced by Tan and Le (2019) and uses a technique called as compound scaling to efficiently scale up the model.

3. **vit_base_patch16_224:** This is a transformer based model architecture and they were first introduced by google in 2020. The ViT's work in such a way that each image is converted to tokens and each token uniquely represents a local area of the image.

4. **deit_base_patch16_224:** This model architecture is from the data efficient image transformers family. These architectures rely on a distillation based token that replaces the usual distillation for transformers Touvron et al. (2021)

4 Design Specification

This section outlines the specifications of the proposed methodology and also gives a description of how the proposed methodology was implemented.

4.1 System Architecture

The proposed framework implements two different deep learning methodologies for model training as mentioned in Section 3.4. Figure 7 shows how the proposed methodology was implemented in a systematic manner. It involved taking raw cancer images and then preprocessing them to ensure that all the images had the same color channels, and image size. Image augmentations were performed to ensure model robustness and then the four models outlined in section 3.5 were trained using the selected two approaches as displayed in the system design diagram and evaluated on the basis of various metrics.

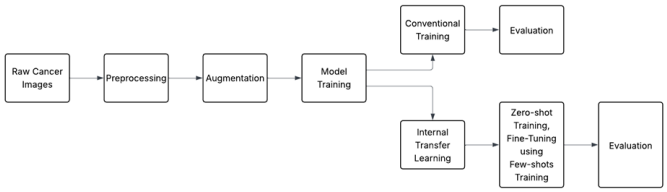

Figure 7: System data flow architecture

4.2 Training Optimization

1. LR Finder: Automated learning rate detection was utilized using the LR Finder from FastAI module. And since, the models were trained from scratch, conservative learning rate was used as suggested_lr/5 for stable training.

2. Batch: The images were fed to the model in batches of 32 images.

3. Optimizer: AdamW was used as the optimizer which is default in FastAI.

4. Loss Function: CrossEntropyLossFlat() was the used loss function.

5. Precision: Mixed precision (FP16) was used for faster training.

6. Epochs: The models were trained for 10 epochs under conventional training and 5 epochs for Internal Transfer Learning.

4.3 Metrics

The models were evaluated using Accuracy, Precision, Recall, and F1-Score as the primary metrics along with confusion matrices and Loss curves. These metrics are based on the following confusion matrix terms: True Positive (TP) (correctly predicted positive), True Negative (TN) (correctly predicted negative), False Positive (FP) (actually negative but predicted as positive), False Negative (FN) (actually positive but predicted as negative).

1. Accuracy specifies the overall correctness of the model and is calculated as:
   Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
2. Precision tells whether out of all the predictions made as positive how many were actually positive and is given as:
   Precision = $\frac{TP}{TP+FP}$
3. Recall shows the models ability to detect positive cases and is calculated as:
   Recall = $\frac{TP}{TP+FN}$
4. F1-Score score provides balance between precision and recall and is useful when there is class imbalance and is calculated as:
   F1-score = $2 * \frac{Precision*Recall}{Precision+Recall}$

4.4 Training Setup

The models were trained on a local machine (laptop) with the following hardware specifications:

1. RAM: 16GB DDR5 RAM
2. CPU: Intel Core Ultra 7 155H with 16-Cores
3. GPU: NVIDIA GeForce RTX 4060 Laptop GPU with 8GB VRAM.

The software requirements are as specified:

1. PyTorch, FastAI for deep learning model training.
2. OpenCV, PIL for images and computer vision.
3. Pandas, NumPy for pre-processing
4. matplotlib, seaborn for visualization

5 Implementation

5.1 Model Training process

This section describes the final implementation of the discussed CNN and Transformer models: ResNet34, EfficientNet-B2, ViT_Base_patch16_224, and DeiT_small_patch16_224 using the approaches of conventional training and Internal Transfer Learning using zero-shot and few-shots training. For the implementation, the final classifier head was modified with the same configuration for all the models for evaluating their performance on the cancer image dataset.

1. Data Pre-processing Module: Here the raw cancer images are loaded and preprocessed using resizing, normalization, and augmentations.

2. Model Architecture Module: The four specified models were implemented with the same customized classifier head architecture optimized for cancer images to ensure model robustness and make the performance comparison more justifiable.

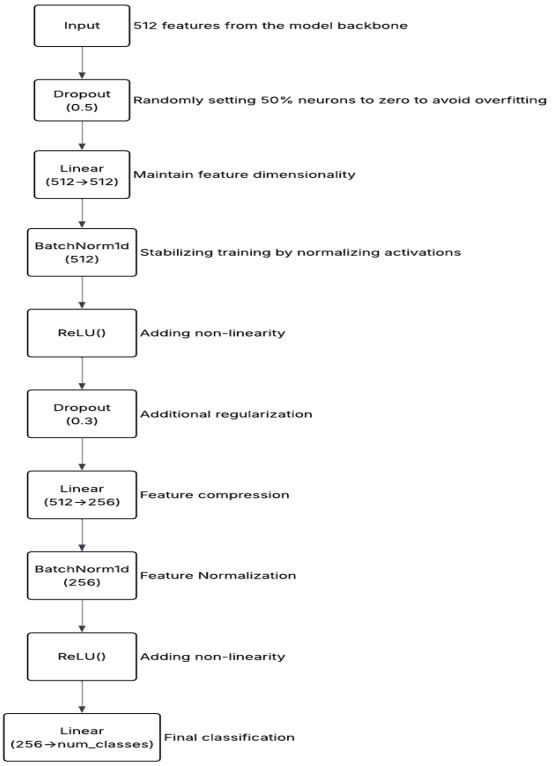

Figure 8: Classifier head setup

Figure 8 displays the classifier head for all the selected models.

3. Conventional Training: The models were trained without pre-training by initializing them with random weights and trained on 80% of the images and were validated on the remaining 20% of the cancer images.

4. Internal Transfer Learning: As mentioned previously to implement the process of internal transfer learning first **zero-shot training** was implemented which involves filtering one specific cancer type. The selected cancer type is then further split into training subclass and testing subclasses where the model is trained exclusively on a single subclass and tested on the completely unseen subclasses. Following this **few shots training** is implemented which involves training the model on only 5% of images by loading the weights from zero-shot and then testing on the remaining 95% of unseen data.

5.2 Tools and Technologies

Python was used for the entire process of Image pre-processing to model evaluation. FastAI was used as the primary module for pre-processing and also implementing deep

learning models. It uses PyTorch under the hood. Tensorboard logging was used for storing various Exploratory data analysis (EDA) charts and model training metrics and loss curves. CUDA 12.1 was used for GPU acceleration for model training.

6 Evaluation

This sections highlights the performance, metrics, and classification accuracy obtained by the trained models. This section also compares the models on the basis of various performance metrics for both the training approaches.

6.1 Experiment - 1: ResNet34

Table 3: ResNet34 performance metrics

Training method	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Training time**
Conventional Training	93.46%	93.85%	93.52%	93.44%	11m 5s
Internal Transfer Learning					
Lung cancer					
Zero-shot	50.50%	16.84%	33.33%	22.38%	56s
Few-shots	98.33%	98.36%	98.35%	98.33%	4s
Lymphoma					
Zero-shot	50.50%	25.25%	50.00%	33.55%	57s
Few-shots	68.50%	69.07%	68.59%	68.33%	7s
Oral cancer					
Zero-shot	100%	100%	100%	100%	53s
Few-shots	100%	100%	100%	100%	3s
Breast cancer					
Zero-shot	100%	100%	100%	100%	55s
Few-shots	100%	100%	100%	100%	4s
Colon cancer					
Zero-shot	100%	100%	100%	100%	52 secs
Few-shots	100%	100%	100%	100%	3s

From Table 3, we can see that, in the case of conventional training, ResNet34 provided an excellent accuracy of 93.46%. This suggests that the model was able to effectively learn about various features and accurately classified more than 93% of the images which is not bad considering the fact that the model was run on a laptop GPU for only 10 epochs.

The confusion matrix in Figure 9 shows that the model is more able to more accurately classify the images for Breast, Colon, and Lung cancer subclasses but struggles a bit with Lymphoma and Oral cancer images.

In the case of internal transfer learning for different cancer types, the model provided excellent performance with few-shots training for Lung cancer achieving an accuracy of 98.33%. For Lymphoma, the performance was good where it provided an accuracy of 68.5% but here also the model struggles a bit as in the case of conventional training for classifying Lymphoma image subclasses. This highlights how the model is able to learn from even a small set of images and able to accurately classify them. The overall training time in the case of Internal transfer learning was significantly less as compared to conventional training.

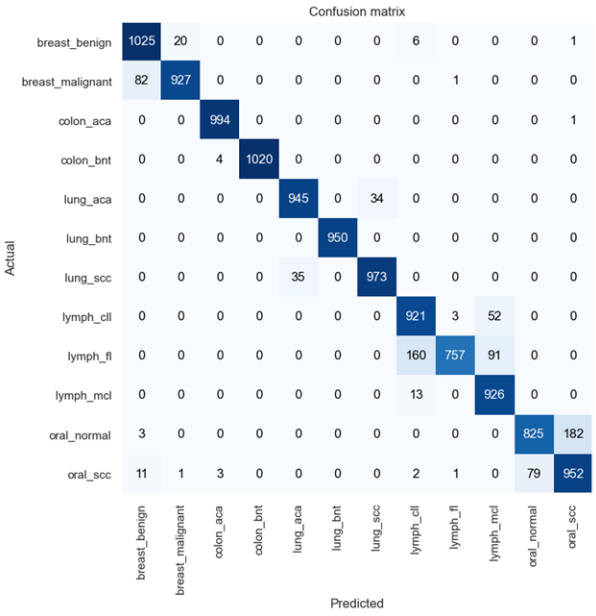

Figure 9: ResNet34 Confusion Matrix

In the case of Oral, Breast, and Colon cancer subtypes the model provided 100% because these cancer types have only two different subclasses and since, it is trained on only subclass it was able to accurately classify that the test subclass is not the one that it saw during training.

6.2 Experiment - 2: EfficientNet_B2

Table 4: EfficientNetB2 performance metrics

Training method	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Training time**
Conventional Training	90.65%	91.11%	90.68%	90.65%	13m 1s
Internal Transfer Learning					
Lung cancer					
Zero-shot	50.50%	25.25%	50.00%	33.55%	1m 4s
Few-shots	85.83%	88.04%	85.96%	85.65%	8s
Lymphoma					
Zero-shot	50.50%	25.25%	50.00%	33.55%	1m 11s
Few-shots	50.54%	25.27%	50.00%	33.57%	12s
Oral cancer					
Zero-shot	100%	100%	100%	100%	1m 56s
Few-shots	100%	100%	100%	100%	6.6s
Breast cancer					
Zero-shot	100%	100%	100%	100%	1m 2s
Few-shots	100%	100%	100%	100%	4s
Colon cancer					
Zero-shot	100%	100%	100%	100%	2m 17s
Few-shots	100%	100%	100%	100%	8s

From table 4, it could be seen that the EfficientNet_B2 model also provided a good performance achieving an accuracy of 90.65% for the conventional training. But, the

performance isn't as good as the ResNet34 CNN network and also training time in the case of EfficientNet_B2 model is comparatively more.

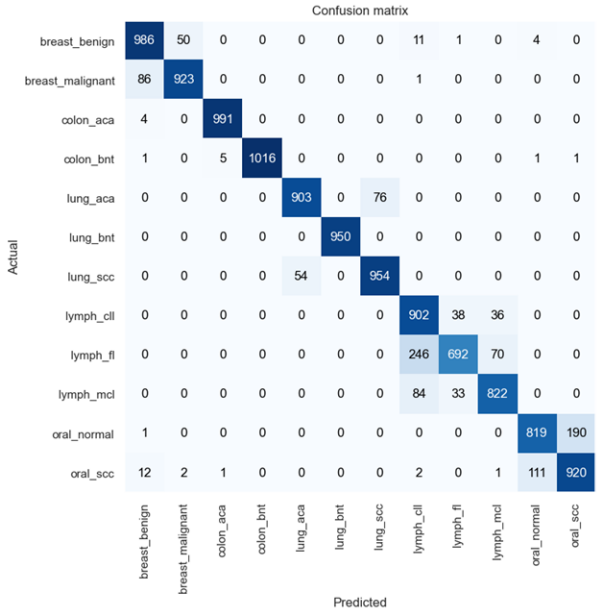

Figure 10: EfficientNetB2 Confusion Matrix

The confusion matrix for EfficientNet_B2 given in Figure 10 suggests that the model struggles more when it comes to classifying Lymphoma and Oral cancer subclasses as considered to subclasses from the other 3 cancer types. When compared to ResNet34's performance, the number of classes misclassified are even higher but the performance is similar for Lung and Colon cancer subtypes.

In the case of Internal Transfer Learning also the accuracy for Lung and Lymphoma cancer types is significantly lower in comparison to what was achieved with ResNet34. Also, the same situation has occured with cancer types having exactly 2 subclasses werein the zero-shot and few-shots accuracy is exactly 100% as the model is able to classify that the images that it was trained on vary from the evaluation set of images. Although, the accuracy for those subclasses was the same the training time required was considerably more.

6.3 Experiment - 3: ViT_Base_Patch16_224

From Table 5 it could be concluded that in the case of conventional training the vision transformer model ViT Base required the highest training time till now but it also provided better performance in comparison to the two CNN models reaching an accuracy of 96.10%. The precision and recall values also are very close to the accuracy metric and suggest that it provides a more balanced performance at classifying positive and negative classes.

The confusion matrix for the vision transformer model shows that it provides the best classification performance and a balanced performance for all the different cancer types. The ViT model is also able to more accurately classify the subclasses for Lymphoma and Oral cancer which the CNN models struggled a bit to classify.

When taking into consideration the accuracy achieved for zero-shot and few-shots training by the ViT model in comparison to the two CNN models, it could be said that

Table 5: ViT_Base_Patch16_224 performance metrics

Training method	Accuracy	Precision	Recall	F1-score	Training time
Conventional Training	96.10%	96.14%	96.12%	96.13%	17m 20s
Internal Transfer Learning					
Lung cancer					
Zero-shot	50.50%	25.25%	50.00%	33.55%	1m 23s
Few-shots	98.76%	98.76%	98.77%	98.76%	8s
Lymphoma					
Zero-shot	50.50%	25.25%	50.00%	33.55%	1m 19s
Few-shots	72.04%	76.99%	71.81%	70.55%	8s
Oral cancer					
Zero-shot	100%	100%	100%	100%	1m 19s
Few-shots	100%	100%	100%	100%	4s
Breast cancer					
Zero-shot	100%	100%	100%	100%	1m 19s
Few-shots	100%	100%	100%	100%	4s
Colon cancer					
Zero-shot	100%	100%	100%	100%	1m 18s
Few-shots	100%	100%	100%	100%	4s

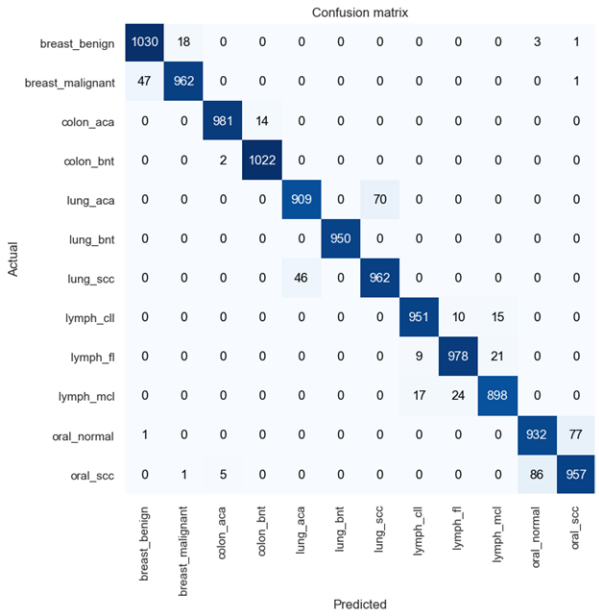

Figure 11: ViT Base Confusion Matrix

it outperforms both the ResNet34 and EfficientNet_B2 models. It also provided the best few-shots accuracy for Lymphoma till now signifying that transformer models are very capable and could be useful in a real-life scenario. Also, the training time required for few-shots training the ViT model is only marginally more by a few seconds which is acceptable as it provides the best classification performance.

6.4 Experiment - 4: DeiT_Small_Patch16_224

Table 6: DeiT_Small_Patch16_224 performance metrics

Training method	Accuracy	Precision	Recall	F1-score	Training time
Conventional Training	94.81%	94.85%	94.84%	94.84%	11m 40s
Internal Transfer Learning					
Lung cancer					
Zero-shot	50.50%	25.25%	50.00%	33.55%	52s
Few-shots	97.54%	97.58%	97.56%	97.54%	6s
Lymphoma					
Zero-shot	50.50%	25.25%	50.00%	33.55%	54s
Few-shots	50.54%	25.27%	50.00%	33.57%	6s
Oral cancer					
Zero-shot	100%	100%	100%	100%	48s
Few-shots	100%	100%	100%	100%	3s
Breast cancer					
Zero-shot	100%	100%	100%	100%	50s
Few-shots	100%	100%	100%	100%	4s
Colon cancer					
Zero-shot	100%	100%	100%	100%	49s
Few-shots	100%	100%	100%	100%	3s

The Data Efficient Image Transformer (DeiT) based model provided a similar performance as the ResNet34 CNN model in the the case of conventional training but it struggled in the case of Internal transfer learning especially for the Lymphoma subclasses achieving an accuracy of 50.54% for the few-shots training.

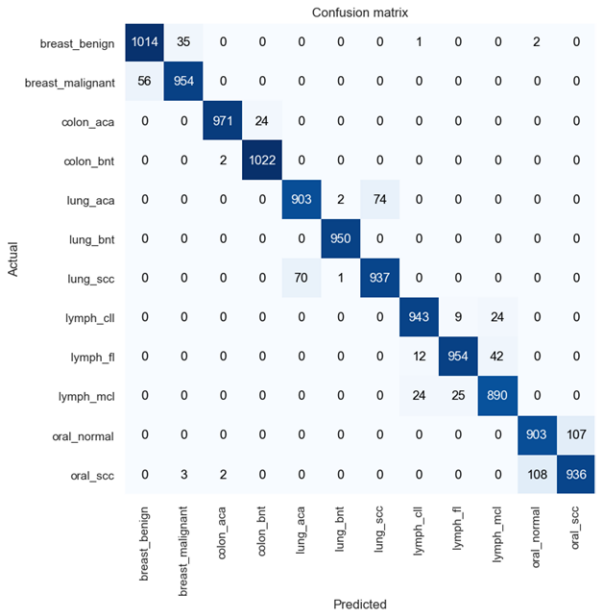

Figure 12: DeiT Small Confusion Matrix

In terms of classification during conventional training, it could be seen from the confusion matrix in Figure 12 that the DeiT model performed well and provided a balance of

performance between the ResNet34 and ViT Base based model where it required training time similar to the ResNet34 and achieved an accuracy higher than the ResNet but lower than the ViT base model.

The very low performance of DeiT in terms of few-shots training for Lymphoma shows that the DeiT based model struggles to learn complex features from images when there is less data available for training and hence accurately classifies merely 50% of the test images.

6.5 Two subclasses versus more than two

In the case of zero-shot and few-shots training for cancer types having only two subclasses, the models exhibited an accuracy of 100% as the model was trained on one subclass and then tested on the other subclass which is why the model was easily able to distinguish that the subclass it was being tested on is not the one that it was trained on. Zero-shot and few-shots training worked as expected in the case of cancer types having more than two subclasses as the model had to classify multiple different image subclasses which it hadn't seen earlier while testing in one go. This makes it a bit difficult for the model to classify as it hasn't learned a lot about the features present in those image subclasses.

6.6 Discussion

Along with conventional training, this research aimed at implementing the approach of zero-shot and few-shots training. This approach could prove to be useful especially in scenarios wherein the image that the model has to classify whether it belongs to cancerous or non-cancerous subclass and isn't an image from a subclass that the model has seen during training. This could improve detection rate and hence, could possibly improve the patient's chances of survival.

From the various experiments performed, it was found that the latest CNN architectures like ResNet34 and EfficientNet_B2 are very capable but are often outperformed by the later transformer based models like the ViT_Base_Patch16_224. These models require slightly more training time but the difference is not much and when it comes to medical domain like this accuracy needs to be given more priority over other factors. This study highlights how the vision transformer models outperform the CNN models. More further improvements could be made to the classifier head and larger transformer models could be implemented like the ViT-Large which this research couldn't implement because of GPU memory constraints as the experiments were run on an 8GB laptop GPU. This could help improve the accuracy attained in the case of complex cancer types like the Lung and Lymphoma cancer types using the approaches of both traditional and internal transfer learning using zero- and few-shots training.

7 Conclusion and Future Work

The aim of this research was to implement the approach of zero-shot and few-shots training for the domain of cancer image classification and find out which type of models provide a better performance.

In conclusion, this research study found that the vision transformer model does outperform the CNN based models in terms of accuracy in both the conventional training and internal learning approaches. The vision transformer model (ViT_Base_Patch16_224)

provided a significant improvement in classification performance in comparison to the ResNet and EfficientNet architectures providing a considerable improvement in classification accuracy and precision.

The accuracies achieved in the case of internal transfer learning for cancer types with more than 2 subclasses are decent and highlight how these approaches with the help of the vision transformer model could be utilized in real-world scenarios where the model sometimes might have to classify images that it has never seen before. This would help to detect cancer accurately in medical settings where some cancer types are not that prevalent and much training data isn't available.

This helps in successfully answering the research the question for this study that Yes, the vision transformer model outperforms CNN based ResNet and EfficientNet models and provides excellent accuracy even when trained on a few images. This study also sufficiently highlights why the approach of zero-shot and few-shots training could be useful in this subject area.

Future research studies could work more in the direction of Internal transfer learning and work with more cancer types having 3 or more subclasses which would help in further improving model performance and robustness thereby allowing to more critically evaluate the obtained results. Also, more larger datasets could be used along with more advanced and larger models with the help of a more powerful GPU.

References

Aissi, M. B., El Hazzat, S. and Yakhlef, M. B. (2024). Evaluation of perceptual hashing algorithms for digital image authenticity, *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)*, IEEE, pp. 1–7.

Aravind, K. R., Kumar, K. A., Reddy, V. V. D. and Yadlapalli, P. (2024). Leveraging convolutional neural networks for classifying lymphoma using histopathological images, *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, Bangalore, India, pp. 1–6.

Danilochkin, G., Ensary, F. and Tuba, E. (2025). Muffins vs. chihuahuas revisited: Exploring zero-shot and few-shot clip-based methods for similar object classification, *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, pp. 1–6.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale, *ICLR 2021 · The Ninth International Conference on Learning Representations*. Accessed: 10 June 2025.
URL: *https://arxiv.org/pdf/2010.11929/1000*

Gai, L., Xing, M., Chen, W., Zhang, Y. and Qiao, X. (2024). Comparing cnn-based and transformer-based models for identifying lung cancer: Which is more effective?, *Multimedia Tools and Applications* **83**(20): 59253–59269.

Hamida, A. B., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G. and Wemmert, C. (2021). Deep learning for colon cancer histopathological images analysis, *Computers in Biology and Medicine* **136**: 104730.

Hasan, M., Rahman, M. S., Islam, S., Ahmed, T., Rifat, N., Ahsan, M., Gomes, R. and Chowdhury, M. (2023). Vision transformer-based classification for lung and colon cancer using histopathology images, *2023 International Conference on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, pp. 1300–1304.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Kumar, Y., Shrivastav, S., Garg, K., Modi, N., Wiltos, K., Woźniak, M. and Ijaz, M. F. (2024). Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification, *Scientific Reports* **14**(1): 25006.

Manali, D. and Demirel, H. (2025). Comparison of vision transformer with convolutional neural networks for brain cancer classification, *2025 IEEE 17th International Conference on Computer Research and Development (ICCRD)*, Shangrao, China, pp. 78–84.

Mehmood, S., Ghazal, T. M., Khan, M. A., Zubair, M., Naseem, M. T., Faiz, T. and Ahmad, M. (2022). Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing, *IEEE Access* **10**: 25657–25668.

Naren, O. S. (2022). Multi cancer dataset.
URL: *https://www.kaggle.com/dsv/3415848*

Ochoa-Ornelas, R., Gudiño-Ochoa, A., García-Rodríguez, J. A. and Uribe-Toscano, S. (2025). A robust transfer learning approach with histopathological images for lung and colon cancer detection using efficientnetb3, *Healthcare Analytics* **7**: 100391.

Our World in Data (2024). Cancer crude death rate by type, world, 2021. Accessed: 10 June 2025.
URL: *https://ourworldindata.org/grapher/cancer-crude-death-rate-by-type*

Patharia, P., Sethy, P. K. and Nanthaamornphong, A. (2024). Advancements and challenges in the image-based diagnosis of lung and colon cancer: A comprehensive review, *Cancer Informatics* **23**.

Prasad, P. and Anbarasi L, A. (2024). Acute lymphoblastic leukemia subtypes detection using vision transformer model, *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, pp. 1413–1418.

Provath, M. A.-M., Deb, K., Dhar, P. K. and Shimamura, T. (2023). Classification of lung and colon cancer histopathological images using global context attention based convolutional neural network, *IEEE Access* **11**: 110164–110183.

Singh, A., Gill, K. S., Kumar, M. and Rawat, R. (2024). Transforming oral cancer care with convolutional neural networks, *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Vol. 1, IEEE, pp. 238–243.

Tafala, I., Ben-Bouazza, F.-E., Edder, A., Manchadi, O. and Jioudi, B. (2025). Deeppatchnet: A deep learning model for enhanced screening and diagnosis of oral cancer, *Informatics in Medicine Unlocked* **57**: 101658.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks, *International conference on machine learning*, PMLR, pp. 6105–6114.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention, *International conference on machine learning*, PMLR, pp. 10347–10357.

Verma, G. (2024). Deep learning-based classification of lung and colon cancer using a proposed cnn architecture, *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Bengaluru, India, pp. 863–867.

Xu, Q., Adam, A., Abdullah, A. and Bariyah, N. (2025). Advanced deep learning approaches in detection technologies for comprehensive breast cancer assessment based on wsis: A systematic literature review, *Diagnostics* **15**(9): 1150.