A Comparative Analysis of Deep Learning Models and Ensemble Methods for Histopathological Image Classification

Akshay Dahiya

2343901

Project Dissertation

Swansea University
Prifysgol Abertawe

Department of Computer Science
Adran Gyfrifiadureg

30th September 2024

Declaration

Statement 1

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed Akshay Dahiya..... Student (2343901)

Date 30/09/20204 ..... Student (2343901)

Statement 2

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by citations giving explicit references. A bibliography is appended.

Signed Akshay Dahiya..... Student (2343901)

Date 30/09/20204 ..... Student (2343901)

Statement 3

The University's ethical procedures have been followed and, where appropriate, ethical approval has been granted.

Signed Akshay Dahiya..... Student (2343901)

Date 30/09/20204 ..... Student (2343901)

Abstract

Histopathological image of cancer cells analysis plays a critical role in the medical diagnostics, specifically for the detection and classification of cancers cells in the images. However, the complexity and variability inherent in medical images present the challenges for automated systems. The purpose of this project is to investigate how to use five cutting-edge customized deep learning models such as ResNet, Convolutional Neural Networks (CNN), EfficientNetB0, Vision Transformer (ViT) and VGG16 to improve the effectiveness and reliability of image classification on any given histopathology data set.

The ensemble methods like majority voting, Dempster-Shafer theory and genetic algorithm-based optimization are utilized to take each model's strengths in the order to improve the classification performance and decrease error rates. After independently training and fine-tuning each model based on the dataset, the model's learned data were used for integrated ensemble methods.

By fusing the advantages of multiple models into a single framework this research also tackles the drawbacks of individual models, including the noise sensitivity and conventional issue of overfitting [1][2]. After some color manipulation, an augmentation technique was used to assess the robustness of these five models. The dataset of colon and lung cancer images which is used in this research is divided into five classes "colon_aca", "colon_n", "lung_aca", "lung_n" and "lung_scc" for this in-depth analysis. The dataset, which originally included 25,000 photos with a resolution of 768 by 768, was reduced to a suitable size to improve efficiency and require less processing power.

A popular machine learning model architecture for classifying images, the convolutional neural networks (CNNs) may struggle to handle complicated datasets that call for sophisticated pattern recognition methods [3]. Because of its scalability and effective management of the various parameters, EfficientNetB0, another model, was chosen in its place. The Vision Transformer model (ViT) was incorporated after CNN and EfficientNetB0, this model uses a self-attention mechanism to capture the long-range dependencies within the images [4]. Then, ResNet was selected for its ability to train deep networks through residual learning, while VGG16 was included for its strength in extracting the hierarchical features across the layers, contributing noticeably to the performance of the model [5].

Ensemble techniques are also used to improve the performance. For example, Majority Voting offers a straightforward yet efficient means of combining forecasts [6]. The Dempster-Shafer theory method, each model's output was given a degree of mass function to control uncertainty in the model predictions [7]. Also, a Genetic Algorithm was used in last to optimize the model weights within the ensemble, enhance improved performance [8].

These ensemble methods outperform the accuracy of the individual models. The best classification accuracy was achieved by the Genetic Algorithm approach to maximize performance. This demonstrates how advanced ensemble techniques are and can improve histopathology images classification, which can make a broader diagnostic implication.

Acknowledgements

My supervisor, Dr. Cheng, provided much valuable assistance and support during the course of this research, and for this I am truly grateful. His guidance was extremely useful, I completed this dissertation with his knowledge and advice and commitment to my project development under his guidance. I really appreciate the opportunities which he has granted to me.

I am writing this, and it would not be possible to let this opportunity slip without expressing my most deeply rooted gratitude to my mom, who has loved me unconditionally and has always encouraged me, believing in my capability. She is the pillar of strength with whom I have been able to approach academics with greater confidence.

I am also deeply indebted to my wife, who has remained patient, understanding, and encouraging through these times. Indeed, her strong belief in me and continued support even during the most difficult times have been sources of motivation and encouragement.

Table of Contents

1 Chapter 1

Introduction

Since machine learning techniques along with medical imaging technologies have evolved at such a fast pace, diagnostic procedures can now be automated, particularly in the area of histopathological image processing. The histopathology research, which uses microscopic analysis of the tissue samples to detect disorders like cancer, is a significant source of medical diagnostic data. The gold standard is when the experienced pathologist do manually scores and tests the cancer cells in the images. However, pathologists may find it difficult, subjective, and error-prone to manually score or interpret cancer images when the distinctions between the normal and sick tissues are slight [9, 10].

Further complicating an automation by machine learning algorithm is the data generated from several labs, each of which has its own standards for generating images. Which may make more mistakes more likely. Using deep learning models for automation, it is possible to decrease these errors, expedite the analysis process, and produce consistent results across all data from various labs [11, 12].

In a variety of image identification tasks, including the analysis of medical picture data, deep learning, a subset of machine learning, demonstrates an impressive performance benchmark. To properly identify the photographs into the five distinct classifications, each model contributes special strengths.

For instance, EfficientNetB0 optimizes network scalability and CNN is well-known for its efficacy in capturing spatial characteristics in images, producing a computationally efficient and effective network model [13]. A relatively recent development in the field of image categorization is the ViT (Transformer) model. The long-range dependence in the pictures that identify the minute structural alterations in the tissues by using the self-attention mechanisms technique [14]. When the gradients, which are used to update the weights during training, get too small as they traverse through more layers, making it more difficult for the model to acquire the new information efficiently. That makes it difficult for the network to improve. ResNet's residual learning method solved this problem by incorporating the shortcut connections that skip some layers and allow gradients to flow much more directly through the network, thus making it easier for a model to learn even with many layers and friendliness and power of VGG16 in extracting hierarchical features [15].

These models often perform very well but when it comes to use of complex dataset, they struggle to give good result. They usually suffer from noise, overfitting or underfitting which causes poor generalization of the classes. This research has shown that using dataset like cancer images with complex structure, these model performance significantly improved by customized architecture of individual model according to data.

Combining these model's strengths yields ensemble strategies that reduce each model's intrinsic volatility and bias while enhancing generalization on unidentified data. These ensemble techniques have the potential to effectively handle issues related to class imbalances and considerable variations in tissue structure in the field of histopathology picture categorization.

The study will demonstrate how the ensembling of those multiple models can strongly improve diagnostic accuracy and may have real-world clinical implications. It adds to a fast-growing literature on the application of deep learning in medical image processing [16].

1.1 Motivation

There has been exponential growth in the research related to the application of deep learning and machine learning methods on medical imaging, specifically the images concerning histopathology. Accurate pattern recognition and classification of the histopathology data will play a crucial role in the diagnosis of various diseases, like cancer.

However, there are other intrinsic problems with medical imaging, such as high inter-class similarity, large intra-class variability, and noise. Depending only on a single deep learning model may cause misclassifications due to those issues and result in higher error rates [17]. The above-mentioned problems motivated us to explore sophisticated, more than one machine learning model and ensemble learning techniques that reduce the individual drawbacks of those models while combining their benefits together. Our motivation increases from these problems, which is putting different models together to yield a more robust and accurate system [18].

Accurately classifying the medical images, especially the histopathological images, is very challenging due to significant differences and variations in tissue structure and colour representation. These variations arise from factors such as differing staining techniques, lighting conditions, and the use of various imaging devices and lenses to capture the image. To record these pictures, each laboratory might use a different grade of equipment with a different lens, which could result in discrepancies that could introduce noise or artifacts among many other variances. That is why, it is more challenging to consistently and accurately identify patterns in tissue pictures because they may seem very differently [19].

In this project, we can use any transformation seamlessly to challenge the factor of image generation from different labs. To make the model robust for more extensive testing, some of the transformations used are the BGR-to-YUV conversion, LAB contrast enhancement, edge detection, adding noise to the image, histogram equalization, and image denoising. Each of those transformations has its own purpose, whether it be emphasizing the important aspects in the images or normalization of input data. Histogram-equalized images reduce the influence of lighting fluctuations by increasing the contrast of an image. LAB contrast enhancement helps in promoting better visualization of minute tissue patterns [20].

Translation of BGR to YUV might allow highlighting the luminance Y component, which is important for differentiation regarding the pattern sequences. The edge detection technique will help the model focus on the borders and structural boundaries in the images. When these adjustments applied, the trained model can also generalize them effectively across the other datasets and situations The model's robustness can also be increased by adding noise and denoising it in accordance with actual variations in medical image scoring. These adjustments can reduce the model's sensitivity to very little noise and increase its ability to concentrate on the data that matters most for diagnosing sickness [17].

The motivation behind this work is driven by the challenges posed by histopathological image classification, such as noise, variability, and complex tissue structures. By engaging with the variety of image transformations alongside an ensemble on the deep-learning models, our aimed to overcome these challenges and provide a more accurate and the reliable system for the medical diagnosis.

1.2 Aim

The objective of this project, therefore, is to enhance the robustness and accuracy of histopathological image classification using only RGB color images from the cancer dataset by employing 5 different state-of-the-art machine learning models with ensemble learning techniques. This shall be done by testing the whole pipeline of data workflow and making assessments about the results.

The goal of this research is to address the issues that each deep learning model faces, such as underfitting, overfitting, noise sensitivity, and difficulty generalizing to new classes with or without transformation of the data. The objective is to create a more dependable diagnostic system to correctly classify and scoring the histopathological images with increased accuracy and fewer errors by the various different models. Since the RGB color format is the most widely used and the versatile type of image, we began our investigation by concentrating on it to make sure that every step of the data processing chain functions as it should.

1.3 Objective

Several objectives have been created to fulfil the main goal of this project as we discussed below.

1.3.1 Data Preprocessing with RGB Transformation:

In this project, all five models are tested only on the RGB colour transformation, which is the main preprocessing step required for image input across all models. However, the whole model could function with any transformation, and doing that requires more computational power and resources. Manually comparing all of those transformations would be quite an annoyance. The project starts by using RGB color data. This is to make certain that all of the models and ensemble techniques are functioning as it would be expected. Color is one feature commonly used as a defining component when differentiating distinct tissue types. It also optimizes computational efficiency [20] since it focuses on a single, widely used color modification technique.

1.3.2 Development of specialized deep learning models:

Convolutional Neural Network (CNN), EfficientNetB0, Vision Transformer (ViT), ResNet, and the final VGG16 which was previously covered—are the five deep learning models that are developed and tested using the pre-processed data. The collection of histology images of lung and colon cancer will be used to individually train these models. Each five models is very well-known for its particular image identification strengths and shortcomings. To do this, a sizable pool of models is required, so that, each of which can leverage unique properties that it is able to extract from the images in order to give result.

1.3.3 Three Ensemble Learning Techniques:

Then, various ensemble learning methods have been used in combination with diverse predictions given by previously trained 5 diverse models; each method utilized takes an edge from the drawback of using just one model for the prediction. Other examples include a genetic algorithm for optimization, the majority voting system, and Dempster-Shafer theory. Actually, ensemble methods are intended to minimize the error rate. It is here that the overall performance becomes enhanced through the strengths of each model in its unique way [18].

1.3.4 Analysing and Assessing of Results:

To get a sense of how the photos are categorized, it is required to examine the performance of each model both individually and collectively. AUC curve, F1-score, precision, recall, accuracy, and confusion matrix are some of the techniques used to compare results on various data classes. This will allow the comparison in how much each model and ensemble technique contributes.

1.3.5 Handling the Challenges in Medical Imaging:

In, like to many other medical images, histopathological images present some of the challenges, including like the heterogeneity within the same class, high inter-class similarity, and the noise as well withing the images [19]. These problem can be solved by the various ensemble methods. Reducing the impact of noise and enhancing the models' capacity to generalize across differences in tissue appearance are the goals of merging models trained on pre-processed pictures.

2 Chapter 2

Background Research

With the advent of ML and AI at the start of the early 2000 era, the progress of the era shift in the analysis of medical images started. In the beginning, they were used to develop pattern recognition algorithms, feature extraction, and image segmentation. Scholars began using classical machine learning (ML) methods, including random forests and support vector machines (SVMs), to categorize lung nodules according to radiomic properties taken from CT scans (Armato et al., 2011) [22]. These pioneering works laid a foundation for more sophisticated machine learning and deep learning, which would follow and revolutionize cancer image interpretation.

Lung and colon cancers are considered two of the most critical types of cancers due to their high global health impacts. Accurately detecting and managing these cancers, has grown dependence on the development of imaging diagnostics. Diagnostic tools utilized include MRI, CT scans, and less often, PET imaging. Specifically, a CT can deliver high-resolution cross-sectional views of the lungs, which can allow the practitioner to more precisely identify tumors and nodules, and follow the spread of metastatic disease. These are formed by rotation of the patient around an X-ray source, recording several body slices. Next, sophisticated algorithms reassemble these slices into three-dimensional pictures, thereby allowing detailed imaging of not just lung abnormalities but also structures [23].

However, the imaging in Colon cancer usually includes histopathology examination from the biopsy samples. During this process, the tissue samples are stained and then studied under the microscope, which shows the details at a cellular level along with the tumour characteristics. Histological examination helps in identification of the cancerous cells and their morphology, which plays an important role in diagnosis and further planning of treatment with accuracy. These features enable pathologists to tell between tissues that are benign and malignant. Some of the techniques used in acquiring these images include Hematoxylin and Eosin staining [24],

which generally is used to emphasize various cellular components. To create these images, specimens are usually fixed in paraffin, sectioned, and then stained to enhance contrast and details.

Initial applications of the machine learning in the field of imaging of lung cancer were related to the classification of lung nodules with the help of Support Vector Machines during computed tomography scans. Among the very early works that Armato et al. conducted in 2011, In order to categorize the cases of lung nodules as either benign or malignant, a classifier based on SVM was presented. The texture, shape, and intensity of the radiomic properties—which are extracted from CT scans and used to train the model using SVM—are taken into consideration. Their method showed notable gains in classification accuracy over conventional approaches, indicating the potential of ML to increase diagnostic accuracy (Armato et al., 2011) [25].

RFs are considered the ensemble learning technique that also finds their extensive applications in lung cancer imaging. Esteva et al. (2017) employed the Random Forest classifiers to segment and classify the lung cancers from the CT images. To increase the segmentation process's accuracy and robustness the researchers integrated the numerous decision trees. Their approach provided useful information for treatment planning by accurately identifying tumour boundaries and differentiating between various tumour types (Esteva et al., 2017) [26].

In a study conducted by researchers, They developed an architecture based on ANN which give significant results and outperformed ML classifiers for the classification of subtypes of colon cancer from images. They utilized multiple hidden layers to learn complex features which enables this model to detect complex patterns that are the indicators of cancerous cell [27].

3 Chapter 3

Literature Review

Experts in deep learning have contributed and conducted several investigations. Robust solutions are required for medical imaging issues, specifically related to histopathology, with a focus on the categorization of histopathology pictures. Several solutions have been suggested to overcome these challenges, some of which include manual scoring. Discussion in this section covers several relevant research and their approaches, conclusions, and limitations, focusing on how they worked. This deep learning is surely going to be a game-changer in medical imaging, since it will revolutionize how physicians can review and understand the complicated data from the imaging.

This review provides an in-depth study of recent advancements in deep learning methods, their uses in various imaging modalities, and the challenges that researchers face without any end.

3.1 Advances in Deep Learning Architectures

The Convolutional Neural Networks (CNNs) are deep learning models that have significantly improved medical imaging. Because of their capacity to automatically and adaptively learn the spatial hierarchies of information from images, CNNs have long been the mainstay of the many

imaging jobs. The Recent advancements have greatly improved their usefulness and performance.

Liu et al. proposed a very highly complicated and sophisticated CNN model architecture, especially for the identification of breast cancer. The methodologies involved in the research utilize residual connections and an attention mechanism to enhance the feature extraction process and, at the same time, improve the model performance [28]. Adding residual connections improves the ability of the gradient to flow through the networks, thereby overcoming the problem of vanishing gradient. Attention techniques will help our model provide more focus on the relevant parts of the image. Results show that our methodology outperformed state-of-the-art traditional CNN models in finding little patterns in mammograms that are crucial in detecting cancers at an early stage.

Dosovitskiy et al. (2023) [29] showed that ViTs can be applied to multi-class classification problems in the medical imaging. Their findings shown that ViTs may compete on the tasks including the classifying chest X-rays and segmenting brain tumours. ViTs have seen the considerable progress in processing the complicated imaging data, and have even been able to outperform regular CNN models in certain circumstances, thanks to their capacity to gather global contextual information.

With its scalable and the effective architecture, Tan and Le's 2019 [30] introduced the EfficientNet algorithm model which is still having an impact on the area. The EfficientNet's accuracy and efficiency in interpreting the retinal pictures that were demonstrated in a recent study by Zhang et al. (2023) when they used it to the identification of the diabetic retinopathy [31]. Because of its architecture, which balances computational economy and model performance, EfficientNet is used in contexts with the limited resources. The study by Zhang et al. showed that EfficientNet performed better in terms of accuracy and computing cost than several alternative topologies, confirming its usefulness in the medical imaging applications. Strong feature extraction capabilities enable VGG16 to maintain its position as a leading deep learning model for medical imaging.

Singh et al. (2023) demonstrated the effectiveness of VGG16 in acquiring high-level data required for an accurate diagnosis by using it to identify breast cancer and classify liver disorders [32]. Even though, the model is somewhat little old, its deep convolutional layers provide a lot of benefits when it comes to the capturing fine-grained picture characteristics. Which is still important for medical imaging tasks. Furthermore, transfer learning with the VGG16 was used by Lee et al. (2023) to improve the lung disease classification performance in circumstances with a lack of labelled data, demonstrating the model's adaptability [33]. However, it is emphasized that the VGG16 has drawbacks, including high processing needs which consumes more computer resources and a large number of parameters which makes it more complex, must be taken into the account for the practical applications [34].

He et al. (2015) [35] introduced the ResNet model which has had a very big impact thanks to its residual connections, that helps in efficiently trained the deeper networks. The efficacy of ResNet in identifying diabetic retinopathy was emphasized by Patel et al. in 2023. According to research ResNet-50 fared better than traditional CNN models because of its ability to handle deeper network layers and prevent gradient problems from disappearing [36], which is particularly very helpful for the complicated data. Zhang et al.'s 2023, segmented MRI brain tumors demonstrate how the residual block, may identify the intricate features using the ResNet algorithm [37]. Because of this feature, ResNet is applicable to a wider variety of medical

imaging applications. The performance of the ResNet network can still be impacted by the caliber of the training sets and the model's depth [38].

3.2 Class Imbalance and Data Limitations

Data Augmentation: Recently, Buda et al. in 2023 came up with a new method of data augmentation to balance the class imbalances by artificially creating minor class examples. This keeps the dataset in balance [39]. This acts towards balancing the dataset. In order to create synthetic samples, this technique included the addition of noise and running complex augmentation techniques such as geometric alterations. By doing so, it improved the under-representative class performance and proved one can reduce the effects of class imbalances at a time when it may improve model generalizations.

Recently, Generative Adversarial Networks (GANs) have emerged as a promising method, for getting over this data limitations. Training data can be improved by artificially generated datasets that are remarkably realistic in GANs. In 2023, Yang et al. presented a GAN-based method, for creating artificial MRI images so that the models might be trained on incredibly rare disorders [40]. GANs provide more of these types of training data, which improves the model's overall performance and capacity for generalization. The significant gains in the model's robustness and accuracy that their method achieves suggest that GANs can produce a range of data that will enhance deep learning models.

3.3 Innovations in Image Preprocessing and Transformation:

To get optimize medical imaging data these techniques will provide a fundamental basis for the transformations and preprocessing. Some Recent developments mainly focused on the several ways to improve image quality to get better improved feature extraction from data. Advanced Preprocessing Pipelines: This includes a number of complex preprocessing techniques for enhancement quality that Wang et al. [41] used on the CT and MRI pictures: contrastive normalizing and adaptive histogram equalization. As will be seen later, in contrast this increases the contrast by normalizing the pixel values with respect to their local neighbourhood.

On the other hand, this adaptive histogram equalization is done with regard to the local histogram concerning modifications in image intensify. These approaches elevate diagnostic precision by remarkably improving image quality and feature extraction. Color Transformations study pointed out LAB color spaces, which have the ability to enhance contrast for views of histopathology images [42]. Histopathology color transformation is crucial for improving the visibility of any given image and the accuracy of its classification. The model will be more robust by converting any images for model input through this enhancement technique.

The LAB color space separates the brightness of the image from the color information and, therefore, allows a better enhancement of the contrasts, making visible cellular structures. Their study's result emphasized the value of specialized preprocessing techniques in medical image analysis as well as the benefits of LAB-based preprocessing for tumor identification and classification.

3.4 Ensemble Learning Techniques and Optimization:

Authors have used performance weights for a range of CNN topologies of weighed ensemble predictions to optimize the ability of each model in detecting lung nodules in computed tomography [43]. Another study conducted where researchers have utilized a genetic algorithm to adjust the weights of the ensemble model in the classification of breast cancers which provides evidence of improvement of improvement for diagnostic performance [44]. This is normally done by evolving a population of candidate solutions through selection, crossover, and mutation operations.

3.5 Applications in Specific Modalities:

Histopathology: Technology Advances in deep learning field have greatly enhanced the ability to analyze histopathological pictures. Using whole-slide images, Li et al. in 2023, presented a novel deep learning method for the detection of colorectal cancer, a tumour subtypes [45]. To improve the classifications and performance, their method to combines the attention approaches, with multi-scale feature extraction. By examining on the essential areas and gathering data at various sizes, there are techniques that generated the insightful understandings regarding treatment planning and tumour heterogeneity.

The development of deep learning has also helped with lung cancer diagnosis. Xu et al. (2023) classified the subtypes of non-small cell lung cancer (NSCLC) based on histology pictures by using a deep learning network. Their model merged CNNs and Transformer networks to extract the global context and the local textures from the images. The hybrid architecture performed better than standard models in differentiating between squamous cell carcinoma and adenocarcinoma subtypes, which is important information for determining treatment decisions. This work is stressed on the importance of combining convolutional and the attention-based models to appropriately discovering the complex visual patterns observed in histopathology slides [46].

The deep learning-based prognostic model for the stratification of the stomach cancer patient survival rate was also developed by Park et al., 2023. It includes the integration of clinical and histological image analysis data for more accurate predictions regarding the outcomes of the patients. The authors emphasized that deep learning plays a part in both diagnosis and prognosis, extending its usefulness to novel approaches to individualized care [47].

4 Chapter 4

Dataset

The Lung and Colon Cancer Histopathological Images Dataset is used in this research known as LC25000 dataset. This dataset which includes 25000 color images that are into five classes: benign lung tissues, lung squamous cell cancer, (class 1), benign colonic tissue, and lung adenocarcinoma. This de-identified, publicly accessible for free, and HIPAA compliant dataset is a perfect tool for testing and training machine learning models related to cancer categorization. High-resolution histopathology slides are the source of the photos in this dataset. Each image has a resolution of 768 by 768 and captures intricate tissue and cell

features, which are essential for a precise classification. This dataset is available for free to research and can be download in nearly 2 GB of zip format file from this GitHub dataset repository [48, 49, 50].

4.1 Tools and technology

All coding and experiment were conducted on the Kaggle environment with free GPU in python coding language.

5 Chapter 5

Methodology

5.1 Preprocessing and Data Augmentation

To ensure that an image is standardized and optimized for model training and generalizations, pre-processing activities are essential.

Normalization: This step is important in stabilizing and speeding up the neural network convergence by ensuring that input data has similar distributions everywhere. Divide by 255 to normalize the pixel values to lie between 0 and 1. The reason being, this would assure faster convergence when training models, as it will help in keeping gradients from exploding or vanishing [51].

Data Augmentation: Synthetic data augmentation was done to artificially increase and provide variability. The techniques used were random rotations up to 20 degrees, horizontal flipping, zooming in from 20% to 100%, shifting. Data augmentation has been shown to improve model generalization; this is in evidence by the work of Shorten and Khoshgoftaar [52]. To verify the authenticity of the test results and prevent data leaking, data augmentation was only done on the training set. These preprocessing methods ensure better performance on unseen data by lowering overfitting and enhancing model robustness.

5.2 Deep Learning Models

This paper presents five states-of-the-art and customized deep learning models, including CNN, EfficientNet, ResNet, Vision Transformer, and VGG-16, which are used in diabetic retinopathy classification. Each of these has its merits and finds wide usage in the task of image classification due to their power of learning complex features. More precisely, in medical imagery.

5.2.1 Convolutional Neural Network (CNN) Model

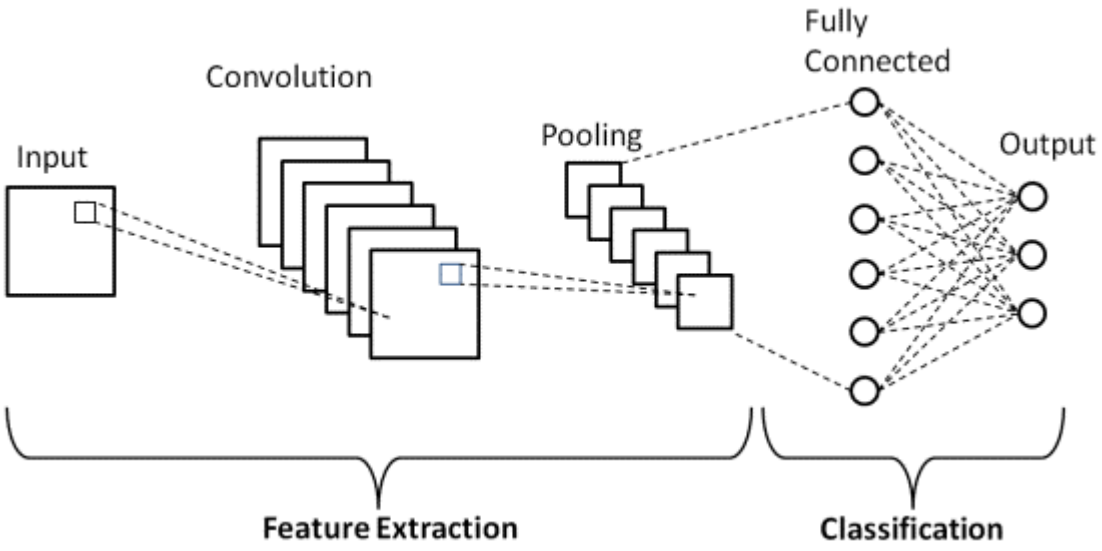

Figure 1 CNN Basic Architecture [53]

The CNN backbone typically consists of convolutional layers, which can make use of convolution in order to extract the relevant features of an image. They use data-learning filters, otherwise called kernels. Each of these can learn different properties of visuals, such as textures and edges, which will help the model identify complicated patterns. (Krizhevsky et al., 2012) [54].

Activation Functions: After convolution, non-linear activation functions may follow. These can be ReLUs. ReLU usually brings in the non-linearity and helps in avoiding problems like a vanishing gradient problem by allowing the network to learn more complicated patterns present in images. (Nair & Hinton, 2010) [55].

Pooling Layers: The pooling applied is called max pooling, which reduces the spatial dimensions of the feature maps by taking the maximum value over neighborhoods. Initially, the output of downsampling processes makes the features more stable in the face of small translations in the input image. Further, it reduces the computing cost. (Scherer et al., 2010) [56].

Fully Connected Layers: It should be a fully linked layer that comes after the convolution and pooling layers, able to flatten and channelize high-level features to categorization. Generally speaking, the final layer uses the SoftMax activation function to produce the probability distributions for each class.

According to the researches CNNs typically performs outstandingly, well in all cases in the medical imaging. Particularly in tasks like tumour, the cancer identification and histological picture categorization. Particularly, they are fitted to the complicated medical applications, since they are capable of automatically extracting important properties from the data [57].

5.2.2 EfficientNetB0 Model

EfficientNetB0 is a deep learning model. It tries to achieve high accuracy, by employing the approach of a mobile inverted bottleneck convolution layer, when resources for processing are limited. Its architecture is grounded on compound scaling to enable balanced improvements in the scalability of a network in terms of depth, width, and resolution. Compared to other conventional models, which can focus on just only on one scaling dimension, EfficientNetB0 scales all three dimensions optimally, making it more efficient in terms of accuracy and speed (Tan and Le, 2019) [58].

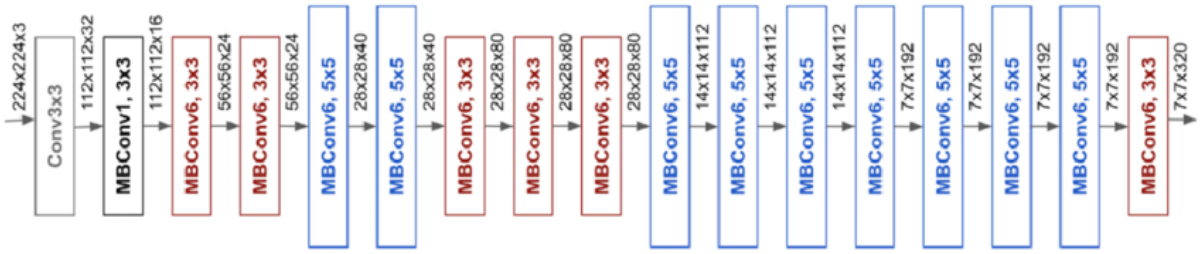

Figure 2 EfficientNetB0 Model Architecture [59]

The MBConv layer helps save more computational resources by first reducing the number of parameters from pointwise convolutions and then increasing the depthwise convolutions. This has been helpful to the model in order for it to focus on more informative features. Swish activation function is basically used by EfficientNetB0 to help smoother the gradients during the backpropagation techniques, which improves the convergence during training (Howard et al., 2019) [60].

Using of pre-learned weights from bigger datasets like ImageNet, EfficientNetB0 is taught using a transfer learning technique, which enhances performance on tasks like medical picture classification. The model balances the computational cost and the accuracy, making it suitable for all limited resource environments delivering strong performance in image classification tasks.

5.2.3 ViT Transformer Model

This deep learning architecture called the Vision Transformer (ViT) was developed by modifying the transformer models that were first created for natural language processing. ViT model was first introduced by Dosovitskiy et al. in 2020 [61]. Its capacity to identify the long-range dependencies in the images has made it more popular for it uses in the computer vision applications.

The fundamental concept of ViT is to handle the image patches as token sequences, much like words are handled in natural language processing. A sequence of vectors is created by flattening and linearly embedding the fixed-size patches such as 16x16 pixels from the input image. This Transformers do not know the order of input tokens on default, therefore, to preserve the spatial information, positional embeddings are added to the patch embeddings (Dosovitskiy et al., 2020) [61].

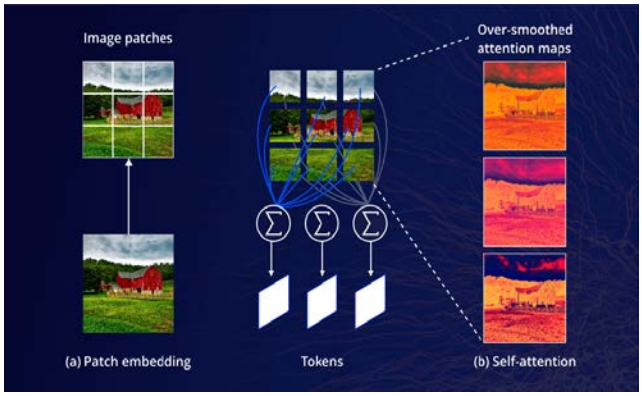

Figure 3 ViT Transformer Model Architecture [62]

It is a many-layer transformer architecture consisting of feed-forward neural networks and multi-head self-attention. The self-attention mechanism will remember the relative importance of each patch with respect to others and, hence, will be able to contextualize the whole image, focusing its attention on the most informative areas of an image. This is particularly useful for tasks that must be able to understand relationships within large regions of an image given or to be processed are applicable to Chen et al., 2021 [63].

5.2.4 ResNet Model

ResNet was a deep learning model developed to solve problems that occurred during the training of individuals with high knowledge in deep networks. This ResNet model, first presented by He et al. in 2015 [64], uses a special technique known as residual learning. It learns the residuals, or disparities between the input and output, rather than the desired output directly. By doing this, the model is better able to concentrate on the learning aspects that can enhance performance, as opposed to the full mapping.

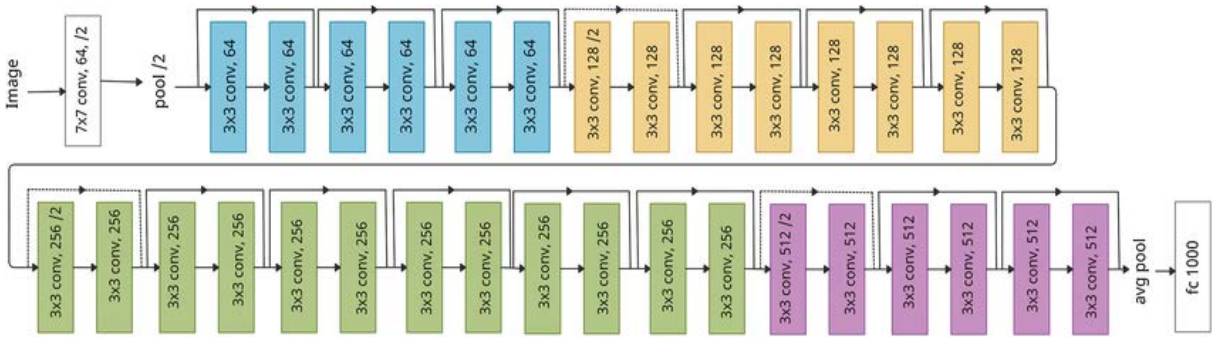

Figure 4 ResNet Model Architecture [65]

Architecture that includes multiple residual blocks, each with two or more convolutional layers, made a ResNet. A block typically consists of the first convolutional layer, batch normalization, and a ReLU activation function once the input is passed through in it. After processing by a second convolutional layer, batch normalization is applied once more to the output of this layer. The skip connection also take place which adds the input straight to the block's output instead of via the convolutional layers which is the main characteristic of the residual block. This lessens the weakening issue as the depth grows by enabling the network to learn an identity function if that is the best option.

ResNet's design not only enhances gradient flow but also facilitates improved feature reuse. As a result, the model may learn ever more intricate data representations while still operating at a high computational efficiency. It has been shown that ResNet reaches the state-of-the-art performance for the various image classification tasks including the medical imaging applications like breast cancer detection and pneumonia diagnosis as showed by Rajpurkar et al., 2017; Zhang et al., 2019 [66, 67].

5.2.5 VGG-16 Model

Simonyan and Zisserman introduced VGG-16 in 2014. This neural network is renowned for being easy to use and efficient when it comes to picture categorization jobs. Thirteen convolutional layers and three fully connected layers, made the architecture of sixteen layers in total. Having modest 3×3 convolutional filters capture fine details and features from input images, the network can keep the resolution by padding. Stride has been fixed to 1.

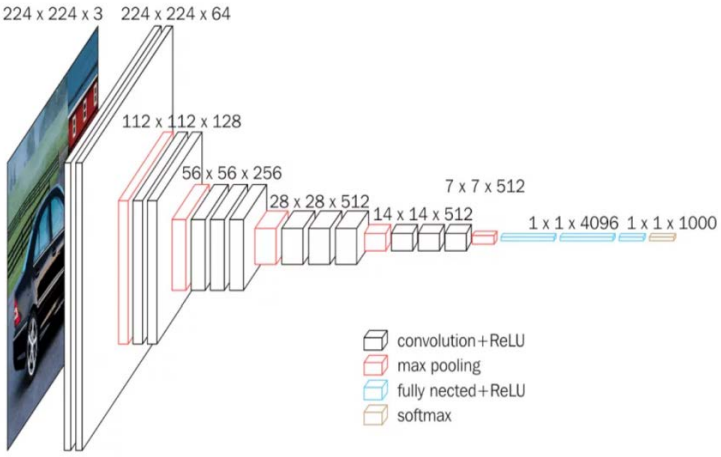

Figure 5 VGG-16 Model Architecture [68]

The max-pooling layer decreases the spatial dimensions of the feature map after each block of convolutional layers. Simonyan and Zisserman (2014) [69] state that this pooling strategy helps preserve the prominent characteristics while the reducing processing cost by down sampling the input representations. While the shallow layers in this architecture learn the more simple patterns, such as edges, deeper layers will learn more complicated structures, such as shapes and textures. The model can learn hierarchical feature representations by building a stack of convolutional layers.

VGG-16 uses the Rectified Linear Unit (ReLU) activation function after the each convolutional layer, which introduces non linearity into the model and helps in the faster training (Krizhevsky et al., 2012) [70]. The architecture concludes with three fully connected layers, where the last layer uses a softmax activation function to output the classification probabilities. The capacity of VGG-16 to generalize effectively across the different datasets. Which is one of its main advantages, which makes it a preferred option for medical imaging applications like the categorization of histopathology images. Its design served as a model's foundation for later models and is frequently employed in transfer learning scenarios, where pretrained weights from sizable datasets (such as ImageNet) are adjusted for particular purposes (Yosinski et al., 2014) [71].

5.3 Ensemble Methods

The rest of the current study will explore further improvements in classification performance through three ensemble methods: majority voting, Dempster-Shafer theory, and genetic algorithms. Ensemble methods refer to the process of making a final prediction by taking a combination of the predictions from various models with the result of a more robust and better prediction of outcomes.

5.3.1 Majority Voting: In Majority Vote, the final prediction takes the most popular class among all the individual models. This scheme is straightforward yet effective, especially when the models to be combined are quite heterogeneous in terms of architecture [72].

5.3.2 Dempster-Shafer Theory: Dempster-Shafer theory conveys the mathematical framework that has been used in combining evidence from multiple sources. In the present work, it has been applied to combine the probabilities given by individual models and thereby provide an elaborative forecast [73].

5.3.3 Genetic Algorithms: Genetic algorithms (GAs) were used to optimize the ensemble weights assigned to each model's prediction. GAs is a type of optimization algorithm inspired by the process of natural selection. In this study, GAs was used to find the optimal weights that maximize the accuracy of the ensemble on the validation set [74].

5.4 Evaluation Metrics

The performance indicators utilized to assess the performances of these models and ensemble methods include accuracy, precision, recall, F1-score, TP, FP, TN, FN, Error Rate, TPR, FPR, FNR, and TNR. Additional metrics included in record for testing loss and validation loss, training loss, and metrics for matching accuracy to compare the results after training each individually. This metric is important in this classification of medical images where the malignant cells rely on both specificity-TNR and sensitivity-recall in their detection.

6 Chapter 6

Research Strategy

6.1 Data Pre-Processing

6.1.1 Data Loading

The data preprocessing includes data loading, data splitting, showing the images with some transformation and with augmentations.

The first step in this project's procedure is to define the read_data_from_directory function, which loads the image file paths and labels from a designated directory. To ensure the file path validity and to avoid runtime issues, the function first checks to see if the directory path exists or not. The function collects file path directory structure by focusing on the subdirectories through the use of os.walk.

Labels are extracted from the names of each subdirectory. This labels and file locations that have been collected are stored in the DataFrame using the pandas - a flexible data structure that makes a basic data manipulation much easier. This DataFrame has two columns called file_paths and labels, which list different picture files along with their associated labels. Furthermore, the code prints the label distribution, which gives information about the number of photos in each class. The dataset is balanced in this case since there are 5,000 photographs in each of the five classes, as seen in the picture below, which displays the entire dataset's form.

```
labels
colon_aca    5000
colon_n      5000
lung_aca     5000
lung_scc     5000
lung_n       5000
Name: count, dtype: int64
Number of classes: 5
```

Figure 6 Dataset Shape

6.1.2 Data Splitting

The function train_test_split from the sklearn library helps divide the dataset into training, validation, and test sets. First, it assigns seventy percent of the data to the training set, and assigns other remaining fifteen percent each to go toward the validation sets and toward test sets. These distributions of classes are preserved across the data splitting thanks to the application of the stratified sampling technique. 17,500 photos were retained for the training set and 3,750 for each of the validation and testing sets.

```
Training set shape: (17500, 2)
Validation set shape: (3750, 2)
Test set shape: (3750, 2)
```

Figure 7 Test Train set shape

6.1.3 Image Loading and Resizing

The loading and the resizing of the photos to 224x224 pixels using the 'OpenCV's cv2.imread' function and cv2.resize functions is the next stage of the preparation pipeline. The consistency in image size is important for the input for all models. It decreases the variances and brighten function is also used to increases the brightness a little of the images.

6.1.4 Transformations and Augmentations

To enhance model performance, preprocessing also involves various methods and image modification technique. The Contrast enhancements are performed using the LAB color space method. CLAHE (Contrast Limited Adaptive Histogram Equalization) technique is used to the L channel to amplify contrasts in uniform regions and enhance the visibility of significant features [75].

An advance technique known as the Histogram equalization, which is utilized to improve brightness and contrast deeply. BGR-to-YUV color space conversion, which separates the luminance (Y) from the chrominance (U and V) also enhances the brightness contrast in a while preserving the color information. In addition to edge identification which uses the Canny algorithm and noise addition with subsequent denoising approaches prepare the model to

handle the images which contains the noise while increasing its robustness. The matplotlib.pyplot module is used to show these transformations and compare the original, scaled, and altered images.

However, to retain information about color, in this paper, the RGB color space is used. This step will work for those models which rely on natural color representation during feature extraction, where minor color variations are important for classification. This has been the best choice in the testing of this very research and also very helpful when a computer is poorly resourced [76].

Other transformations also can be done consistently so that, as model input, a flexible and easy-to-use data pipeline is achieved. Labels are generated automatically to conform to model-testing selection made on a particular picture transformation and works for all models. This process goes so much smoother and easier, and more proficient, with the help of this workflow. It also enhances clarity, making it more understandable; this could manage preprocessing steps for different deep learning models.

Also, data augmentation methods which includes the flips, zooms and the random rotations are used to improve the resilience of the models. These strategies can make broaden the training dataset by incorporating changes that is occur in real-world circumstances. This helps the models perform better during testing and lowers the chance of overfitting [77].

6.1.5 Label Mapping and Conversion

Do pre-processing and enhancement, then check that the labels are in the proper format for model input. Labels are in categorical format for "lung adenocarcinoma" and "benign lung tissue". One of the important methods to turn these labels into their integer values is using a label mapping dictionary or the 'label_mapping' function. This makes it quite critical to convert the labels to integer values, thus forming a unique representation for every class.. In fact, these are some of the steps toward ensuring no mismatch errors occur during the training or testing of the model.

```
Unique labels before conversion: ['colon_aca' 'colon_n' 'lung_aca' 'lung_n' 'lung_scc']
Label mapping: {'colon_aca': 0, 'colon_n': 1, 'lung_aca': 2, 'lung_n': 3, 'lung_scc': 4}
```

Figure 8 Label Details

Class Number	Original Short Name	Full Name
0	colon_aca	Colon Adenocarcinoma (Cancerous Colon)
1	colon_n	Benign Colonic Tissue (Non-Cancerous)
2	lung_aca	Lung Adenocarcinoma
3	lung_n	Benign Lung Tissue
4	lung_scc	Lung Squamous Cell Carcinoma

6.2 Customized Models Architecture

6.2.1 Convolutional Neural Network (CNN) Model

This customized Convolutional Neural Network can preprocess the cancer images with high accuracy. CNN, which we propose, includes four convolution blocks: each of the convolution blocks uses a sequence of MaxPooling 2D and Convolutional 2D (Conv2D) layers to extract and downsample the features from input images with dimensions 224x224x3, where three stand for colors of the image. Each convolution layer will be followed by ReLU activation functions and L2 regularization to avoid overfitting and BatchNormalization layers to stabilize and accelerate the training process [78, 79]

The model starts off with a Conv2D layer, which has 16 filters and is followed by MaxPooling and Batch Normalization. Further down the blocks, it goes ahead to increase the number of filters to 32 and 64, respectively, to enhance the power of the model for capturing complex details of increased abstraction. The model is first initialized with a Conv2D layer that runs with a number of 16 filters. It is then followed by Batch Normalization and MaxPooling. To increase the higher abstraction level of the model, capturing more fine details, further increases in the filter numbers are made to 32 and 64 in the subsequent blocks. The final convolutional layer is followed by Flattening to transition from 2D feature maps to a 1D feature vector, which is then passed through two Dense layers with the 32 and 64 units respectively.

Besides this, the dropout function is also applied in Dense layers to save them from overfitting by randomly setting a fraction of input units to zero during training [80]. From here, the network uses an Adamax optimizer with a learning rate of 0.0001, and the loss function is categorical cross-entropy, since this network will handle multi-class classification issues. A batch size of 16, updating weights more frequently with the smaller sets of data, enhances the generalization. To increase the effectiveness of model training, EarlyStopping and ReduceLROnPlateau callbacks are also used. Which promote more efficient convergence of the model, ReduceLROnPlateau adjusts the learning rate in response to the validation accuracy, whereas EarlyStopping reduces overfitting by terminating training if no improvement is observed by monitoring the validation accuracy [81, 82].

6.2.2 EfficientNetB0 Model

EfficientNetB0 is a highly efficient architecture created for image classification tasks that balance accuracy and computational complexity, is the second model under study. In this setup, EfficientNetB0 serves as a feature extractor that has been pre-trained on the ImageNet dataset. The all the pre-trained layers are frozen to preserve the robust features that have been acquired through lengthy training on large-scale data [83]. The Global Average Pooling is used and applied directly after the feature extractions and afterwards, fully connected layers with 64 and 128 units are applied in this customized architecture. This enhances the data more for the task at hand. The Dropout function was included between the layers at a rate of 0.4 to reduce overfitting and improve the model's ability to generalize [84].

Lastly, the output layer's softmax activation makes that the model delivers probabilities for each of the target classes. This model is also assembled using the categorical crossentropy loss function which is especially well suited for the task as multi-class classification. Also, the Adamax optimizer, which works well with high-dimensional data [85]. Early stopping is utilized to monitor validation loss and minimize overfitting by stopping training when no

improvement is observed, ensuring optimal training [84]. By using this method, the model becomes reliable and efficient in managing challenging image categorization tasks.

6.2.3 Vision Transformer (ViT) Model

This study addresses picture categorization issues with a much more sophisticated transformer-based design. Third position is occupied by this Vision Transformer (ViT) model. This transformer technique which was initially created for natural language processing is modified in this research to analyse the images by splitting them into the fixed-size patches and then embedding these patches into a series of various token. This will give the much-needed boost to the model's comprehension of a complex visual pattern by making the model sensitive to global dependencies present in the image.

It implements essential parts, including patch embedding via convolutional layers, positional the encoding for maintaining the spatial information, and multiple transformer blocks for refined feature extraction. Each transformer block is composed of multi-head self-attention layers and feed-forward networks with residual connections and layer normalization, which scales up its capacity for such intricate visual features. The architecture also employs dropout layers with the intention of reducing overfitting to improve generalization across all classes with variation.

It had an architecture with a reduced number of feed-forward units and attention heads along with four transformer blocks; this was sufficient to create a perfect balance between computing efficiency and model complexity. This compact architecture is appropriate with over 391,000 total parameters effective for deployment and training [86, 87, 88].

In order to prevent overfitting and preserve the model's best performing step, early halting and model checkpointing are also employed in this training process. Recent advances have demonstrated that this method makes greater use of transformer-based models to improve performance on a variety of picture-classification tasks [90].

6.2.4 ResNet Model

ResNet stands for a residual network used in the implementation of image classification. Its architecture received quite high popularity since it indeed showed how to train deeper neural networks by using residual blocks. In this model, the connections are done such that the layers can be skipped and the gradient gets an easy path to bypass one or more layers and hence avoid the problems of vanishing gradients in deep networks. This architecture enables more efficient training of very deep networks by alleviating the difficulty in propagating gradients through multiple layers [91].

The first 64-filter convolutional layer, Batch Normalization, and ReLU activation make up the layered structure defined in the code for the ResNet model. A few unused blocks then follow subsequently. Remaining connections after the two convolutional layers in each block help to overcome the problem of vanishing gradients and allow deep networks to be developed without performance compromise. Architecture Introduction The following architecture introduces residual connections "(Add ())", which make it certain that the model will learn the identity so that the information can flow more easily through the network [92, 93].

The model includes blocks with progressively larger filters starts with 64, 128, 256 and ends at 512 and strides to downsample the image as the network deepens. A Global Average Pooling

layer is in the last, followed by the dense layers for classification with dropout is set to 0.4 added to prevent the overfitting. The ResNet model is compiled using the Adam optimizer with a low learning rate of 0.00001, this ensures that the weights are updated slowly to prevent overshooting the optimal values.

"Accuracy" is the metric to be monitored and the loss function utilized is categorical_crossentropy, which is perfect for multi-class classification tasks. Early stopping and model checkpointing are applied during training. To prevent overfitting, early halting stops training if the validation loss does not improve for three consecutive epochs, while checkpointing keeps the best-performing model intact [92, 93].

6.2.5 VGG-16 Model

This model applies transfer learning using the well-known VGG16 architecture, which is commonly employed for image classification tasks. Instead of training from scratch, it leverages pre-trained weights from ImageNet, excluding the top fully connected layers (include_top=False). This allows the model to be customized for a specific cancer image dataset, while still benefiting from the features it has already learned from ImageNet [94]. Freezing the original layers is a crucial step after loading the model. The model preserves its previous knowledge when the convolutional layers are left unchanged after the training which enhances feature extraction when applied to the new dataset [95].

Custom layers are designed specifically for that categorization operation that are applied after the frozen layers. Before being going into fully connected "Dense" layers, the output from the convolutional layers first be converted into a one-dimensional vector by the first additional layer, the Flatten layer. The model is then able to acquire intricate patterns and non-linearities, which aids in its adaptation to the new dataset, by incorporating a Dense layer with 512 units and ReLU activation [96].

The model has both trainable and non-trainable parameters. The non-trainable parameters are those in the base VGG16 model that are with frozen layers during the training. These parameters come from the pre-trained layers and remain unchanged. Which allows the model to perform effectively on the new dataset with comparatively on few epochs. The total number are 14,714,688 and these contribute to the model's capacity to extract features from images based on previously learned patterns [97]. The trainable parameters include the newly added Dense layers or any other layers are 12,848,133. These parameters are updated during the training sessions.

The model is kept from becoming overly reliant on any one subset of features by adding a Dropout layer with a rate of 0.4, which randomly removes 40% of the neurons during each training iteration to prevent overfitting. Mapping the outputs to the number of target classes and generating the probability distributions for each class is the job of the last Dense layer which uses the softmax activation function [96].

6.2.6 Model Evaluations

Another code snippet at the end of all model training and evaluation result added, which is to test the prediction of all model by entering any image number. The snippet algorithm has ‘reverse_label_mapping’ function which reverser the label mapping to it original names and print true class name and the predicted class name of the image. This snippet is to test and for fun to check how well this 5 model pipeline is working.

6.3 Ensemble Techniques

6.3.1 Majority Voting

A technique that combines the predictions from five distinct deep learning models that have already been trained is via the majority voting technique. Initially, the test set image class labels are predicted using each model. A two-dimensional array containing the predictions from all five models, is kept with each row indicative the predictions from a single model and each column denoting a particular test sample.

The class predicted by majority of models for each sample is then chosen by this majority voting. This process involves figuring out the most frequent value, for every column in the prediction array. When the final ensemble predictions are different with the actual labels then the ensemble's total accuracy is calculated.

6.3.2 Dempster-Shafer Theory

This methodology influences the mass functions, that is derived from each model's predictions to produce a unified classification decision.

Mass Function Conversions: Using the "convert_to_mass" function for each 5 model predictions are first converted into the mass functions. This function assigns a mass to each class based on the model's estimated probabilities, and a preset degree of uncertainty which is 0.5 by default. This exact masses for the ignorant class and the projected class in each mass function indicate the uncertainty of the model prediction [98]. This conversion is necessary for the application of Dempster-Shafer theory, which deals with belief functions derived from mass functions.

Dempster's Rule of Combination: The mass functions from each model, are combined using the Dempster's Rule of Combination. When forecasts deviate, the rule adjusts for differences by combining the data from many sources. The "combine_mass" functions computes the combined mass function by considering the disagreement between two pieces of evidence and changing the combined mass values accordingly [99]. Steps in this process include controlling the conflict term and appropriately uniting the illiterate masses.

Ensemble Decision Making: After the mass functions are combined, the class with the highest beliefs that is the one that does not include the ignorance masses, is selected to determine the final predictions. This accuracy of the final ensemble predictions is calculated and compared to the true labels to evaluate the effectiveness of the combined models [100].

6.3.3 Genetic algorithm (GA)

The 3rd ensemble technique is used in the research is the genetic algorithm that runs over the 50 generations on a population of 50 with individuals nevals. Every individual is a distinct combination of weights that is applied to five distinct model predictions. The amount that each model influences the final ensemble choice is largely determined on these weights.

After the mass functions are combined, the class with the highest beliefs that is, the one that does not include the ignorance mass is selected to determine the final predictions. To evaluate the combined model effectiveness each data's initial weights are created randomly but then

they are adjusted to make sure they add up to one. Because of this normalization, predictions from several models can be proportionately blended based on their assigned weights [101], which is essential for the weighted voting procedures. The two crucial components of the algorithm are mutation rate of 20% with the 50% crossover probability, which is important to preserve the genetic varieties to investigate the different options in the weight space [102].

The accuracy of the final prediction is calculated by comparing it with the true labels [103]. Throughout each generation, the fitness of individuals is assessed by calculating the accuracy of the predictions they produce when combined using their respective weights. Individuals that achieve higher accuracy are more likely to be selected for reproduction in the next generation. The genetic operations involve the blending of weights through the crossover and introducing small random changes via mutation to prevent the algorithm from getting stuck in local optima [103].

The top-performing weight is chosen once all iterations have been completed. Now, this individual communicates to the optimal weights for the ensemble which are normalized again to ensure of their proper use in subsequent predictions. So, the refined set of weights enhancing the overall classification performance in this multi-class tasks and leveraging the strengths of each model effectively. This individual communicates to the optimal weights for the ensemble which are normalized again to ensure of their proper use in subsequent predictions. Ultimately, the refined set of weights reflects the contributions of each model, enhancing the overall classification performance in this multi-class task and leveraging the strengths of each model effectively.

6.3.4 Evaluation of the Ensembles

Also, each model and the ensemble's class-wise performance metrics are generated using the confusion matrix. These metrics include the following: error rate, True Positive Rate (Recall), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR), as well as True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). A “DataFrame” is created from the combined findings to make comparing model performance. This illustrates shows that how utilizing the advantages of several models through ensemble approaches can enhance categorization.

7 Chapter 7

Results

7.1 All Model Results

7.1.1 CNN Model Performance Evaluation

The CNN model achieved a training loss of 39.47% and a training accuracy of 89.44%. On the validation dataset, the model reached a validation loss of 40.99% and an accuracy of 87.44%. Whereas, testing on an independent dataset resulted in a test loss of 40.39% and an accuracy of 87.49%.

Class-wise Performance

There is no denying the variations in performance amongst the five classes. With an F1-score of 0.79, colon_aca (class 0) has attained an accuracy of 74%. With an F1-score of 0.86, the colon_n (class 1) showed the highest accuracy at 95%. The lung_aca (class 2) had an F1-score of 0.86 and was categorized with 90% accuracy. The lung_n (class 3) demonstrated remarkable performance with an F1-score of 0.98 and 96% accuracy. The accuracy with the F1-score for the lung_scc (class 4) are 83% and 0.88 respectively. For accuracy, precision, recall and F1-score, the macro and weighted averages were 87%.

Confusion Matrix

The confusion matrix shows a number of classes to be significantly misclassified. For example, 197 cases of colon_aca (class 0) are mistakenly identified as colon_n (class 1). There was very little misclassification for colon_n (class 1), with only 34 samples mistakenly classified as colon_aca (class 0). Lung_aca (class 2) was mistakenly assigned to 45 samples of lung_scc (class 4). A total 21 of them were lung_aca class 2, which it was incorrectly classifying as the lung_n class 3. And 116 of the lung_scc class 4 were mislabeled as the lung_aca class 2.

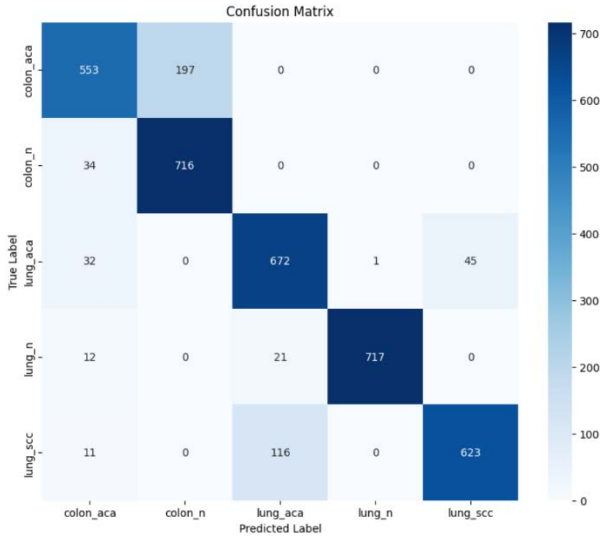

Figure 9 CNN Confusion Matrix

Error Rates and TPR/FPR Analysis

The rest can be further reflected in complications with the error rate-for class 4, the error rate was 16.93% for lung_scc, and for class 0, the mistake rate of colon_aca was 26.27%. Although the TPR for class 0 was 73.73%, its false negative rate was 26.27%. These are quite low rates. The colon_n (class 1) had a low FNR of 4.53% and a high TPR of 95.47%, indicating that it was easy to categorize. Compared to the lung_scc (class 4), which had a far lower TPR of 83.07%, lung_aca (class 2) performed better, with a TPR of 89.60%.

```
True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN):
Class 0: TP = 553, FN = 197, FP = 89, TN = 2911, Error Rate = 0.2627
Class 1: TP = 716, FN = 34, FP = 197, TN = 2803, Error Rate = 0.0453
Class 2: TP = 672, FN = 78, FP = 137, TN = 2863, Error Rate = 0.1040
Class 3: TP = 717, FN = 33, FP = 1, TN = 2999, Error Rate = 0.0440
Class 4: TP = 623, FN = 127, FP = 45, TN = 2955, Error Rate = 0.1693

True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR):
Class 0: TPR = 0.7373, FPR = 0.0297, FNR = 0.2627, TNR = 0.9703
Class 1: TPR = 0.9547, FPR = 0.0657, FNR = 0.0453, TNR = 0.9343
Class 2: TPR = 0.8960, FPR = 0.0457, FNR = 0.1040, TNR = 0.9543
Class 3: TPR = 0.9560, FPR = 0.0003, FNR = 0.0440, TNR = 0.9997
Class 4: TPR = 0.8307, FPR = 0.0150, FNR = 0.1693, TNR = 0.9850
```

Figure 10 CNN Error Rate

7.1.2 EfficientNet-B0 Model Performance Evaluation

It can be observed that the performance of the EfficientNetB0 model is quite good with an overall accuracy of 94.73%, and a test accuracy of 94.83%. This test loss of 0.1487 shows that this model has effectively learned and generalized well. All classes have good recall and precision and, therefore, great F1-scores with a macro average F1-score of 0.95, showing balanced performance.

Confusion Matrix Analysis

It can be further elaborated from the confusion matrix that most of the classes of EfficientNet-B0 are performing well in categorization. The model has correctly categorized 718 examples out of 750, whereas only 32 have been misclassified for colon_aca class 0. On the other hand, when it came to the categorization of lung_aca or class 2, it had a rather very bad performance, identifying 659 occurrences against 91 misclassified cases.

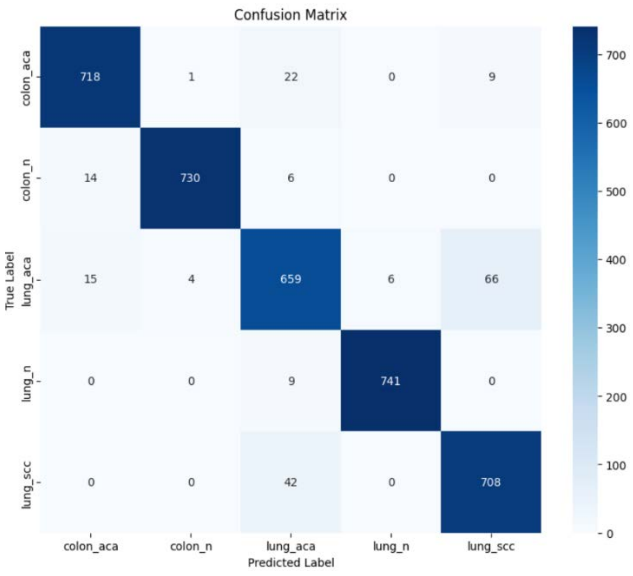

Figure 11 EfficientNet-B0 Confusion Matrix

Error Rates and Sensitivity (TPR)

It can be observed that the error rates are the lowest for lung_n which belongs to class 3, with 1.20%, and the highest error rate for the class 2 image-lung_aca, that is 12.13%. That would indicate that the model reduces the sensitivity in detecting lung adenocarcinoma. TPR for the lung_n class, which in short is class 3, is the highest with a value of 98.80%, while the minimum TPR value is 87.87% for the class having the name lung_aca which is class 2.

False Positive Rate (FPR) and True Negative Rate (TNR)

This model shows the minimum FPR at lung_n for class 3, which is 0.0020. For this class, the TNR is 98%, which means this model does a very good job of correctly identifying the negative cases while simultaneously avoiding false positives. That means class 3, lung_n, is coming up with the best result in this case.

```
True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN):
Class 0: TP = 718, FN = 32, FP = 29, TN = 2971, Error Rate = 0.0427
Class 1: TP = 730, FN = 20, FP = 5, TN = 2995, Error Rate = 0.0267
Class 2: TP = 659, FN = 91, FP = 79, TN = 2921, Error Rate = 0.1213
Class 3: TP = 741, FN = 9, FP = 6, TN = 2994, Error Rate = 0.0120
Class 4: TP = 708, FN = 42, FP = 75, TN = 2925, Error Rate = 0.0560

True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR):
Class 0: TPR = 0.9573, FPR = 0.0097, FNR = 0.0427, TNR = 0.9903
Class 1: TPR = 0.9733, FPR = 0.0017, FNR = 0.0267, TNR = 0.9983
Class 2: TPR = 0.8787, FPR = 0.0263, FNR = 0.1213, TNR = 0.9737
Class 3: TPR = 0.9880, FPR = 0.0020, FNR = 0.0120, TNR = 0.9980
Class 4: TPR = 0.9440, FPR = 0.0250, FNR = 0.0560, TNR = 0.9750
```

Figure 12 EfficientNet-B0 Error Rate

7.1.3 Vision Transformer (ViT) Model Performance Evaluation

The Vision Transformer or ViT model showed accuracy percentage of 83.52 for tests and 82.71 for training. This demonstrates how well the data are generalized to unknowns. Compared to the other models in this study, the accuracy of 83.23% is not as good as it could be, although being quite good overall. The model doesn't appear to be overfitting and is well-fitted based on the proximity of the training and validation losses.

Confusion Matrix Analysis

The Vit's confusion matrix illustrates the model all five class classification capabilities. With a low false positive rate of 0.0020 and the high true positive count of 708 lung_n (class 3) model performs exceptionally well. Although, the model performs bad with the colon_n (class 1) as it misclassifying 240 instances. This implies that the process of distinguishing between the colon_n (class 1) and colon_aca (class 0) needs to be improved.

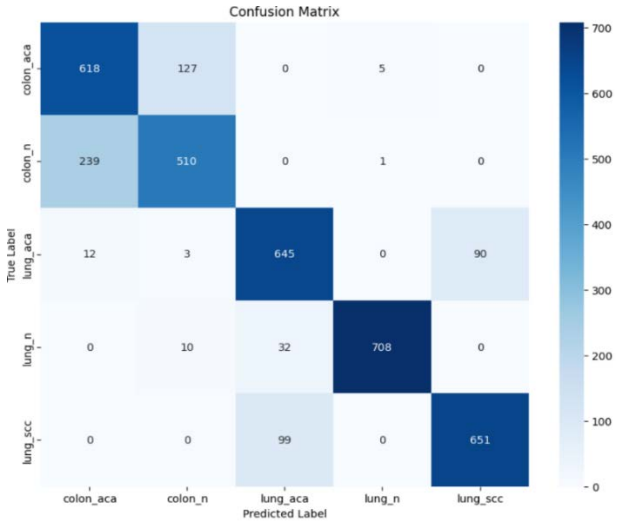

Figure 13 Vision Transformer Confusion Matrix

Error Rates and Sensitivity (TPR)

The error rates differ sharply with the highest error rate of 32.00% for the class colon_n (class 1) shows that implying high difficulty in diagnosing benign colon tissue correctly. Evidence of the model's efficiency in recognizing benign lung tissue is that the TPR for the lung_n-class 3- is the highest, being 94.40%. On the other hand, colon_aca (class 0) displays a TPR of 82.40% which indicates that although it can identify this class respectably, sensitivity might be improved.

False Positive Rate (FPR) and True Negative Rate (TNR)

Here, the FPR is quite low for most of the classes, while the minimum is 0.0020 for lung_n with a class of 3, and 0.0437 with a class of 2 in lung_aca. This also depicts quite a high TNR, the rate incredibly high at the lung_n for class 3 running as high as 99.80%, evidence of how strong this model is in correctly defining negative cases. It is further supported that the FPR for colon_n, which is class 1, is higher at 4.67%, showing the difficulty in minimizing this class' false cases. This proves that the model ViT goes well with the identification of lung-related tissues, while areas, especially in the differentiation of colon tissue types, need improvement.

```
True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN):
Class 0: TP = 618, FN = 132, FP = 251, TN = 2749, Error Rate = 0.1760
Class 1: TP = 510, FN = 240, FP = 140, TN = 2860, Error Rate = 0.3200
Class 2: TP = 645, FN = 105, FP = 131, TN = 2869, Error Rate = 0.1400
Class 3: TP = 708, FN = 42, FP = 6, TN = 2994, Error Rate = 0.0560
Class 4: TP = 651, FN = 99, FP = 90, TN = 2910, Error Rate = 0.1320

True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR):
Class 0: TPR = 0.8240, FPR = 0.0837, FNR = 0.1760, TNR = 0.9163
Class 1: TPR = 0.6800, FPR = 0.0467, FNR = 0.3200, TNR = 0.9533
Class 2: TPR = 0.8600, FPR = 0.0437, FNR = 0.1400, TNR = 0.9563
Class 3: TPR = 0.9440, FPR = 0.0020, FNR = 0.0560, TNR = 0.9980
Class 4: TPR = 0.8680, FPR = 0.0300, FNR = 0.1320, TNR = 0.9700
```

Figure 14 Vision Transformer Error Rate

7.1.4 ResNet Model Performance Evaluation

With the test accuracy of 95.81 percent and the training accuracy of 96.47%, the ResNet showed remarkable performance qualities. The training loss of 0.0923 and the relatively minimal validation loss of 0.1219 showed a excellent learning without any appreciable overfitting of the data. The models strength is demonstrated by its overall loss of only with number 0.1119.

Confusion Matrix Analysis

The confusion matrix for this model is below, showing the performance of classification ranging across all 5 classes. Perfect classification for the class 'lung_n' is shown by 750 true positives with no false positives or negatives. On other classes, such as class 2, where lung_aca has a greater count of 109 false negatives than the true positives at 641, this might indicate some difficulties in the correct identification of the lung_aca.

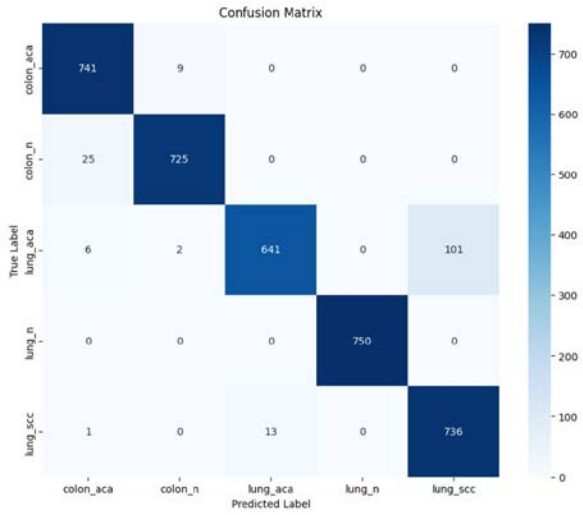

Figure 15 ResNet Confusion Matrix

Error Rates and Sensitivity (TPR)

Among these, class 3 and class 0 had the highest true positive rate with an estimation of 100% and 98.80%, respectively. On the other hand, class 2 had the lowest error rate among them, with only 14.53%.

False Positive Rate (FPR) and True Negative Rate (TNR)

The false positive rates for Class 3 are about 0.0, while for Class 0, these were 0.0107 a very good indication of how really efficient this is in the detection of such cases. Also, in all classes, the True Negative Rate is high, indicating effective identification.

```
True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN):
Class 0: TP = 741, FN = 9, FP = 32, TN = 2968, Error Rate = 0.0120
Class 1: TP = 725, FN = 25, FP = 11, TN = 2989, Error Rate = 0.0333
Class 2: TP = 641, FN = 109, FP = 13, TN = 2987, Error Rate = 0.1453
Class 3: TP = 750, FN = 0, FP = 0, TN = 3000, Error Rate = 0.0000
Class 4: TP = 736, FN = 14, FP = 101, TN = 2899, Error Rate = 0.0187

True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR):
Class 0: TPR = 0.9880, FPR = 0.0107, FNR = 0.0120, TNR = 0.9893
Class 1: TPR = 0.9667, FPR = 0.0037, FNR = 0.0333, TNR = 0.9963
Class 2: TPR = 0.8547, FPR = 0.0043, FNR = 0.1453, TNR = 0.9957
Class 3: TPR = 1.0000, FPR = 0.0000, FNR = 0.0000, TNR = 1.0000
Class 4: TPR = 0.9813, FPR = 0.0337, FNR = 0.0187, TNR = 0.9663
```

Figure 16 ResNet Error Rate

Precision, Recall, and F1-Score

This is reflected in the precision, recall, and F1-score of the model. Thus, class 1 can be represented by the result viewed below with an F1-score of 97%, a recall of 99%, and a precision rate of 96%. The F1 score for class 2 ('lung_aca') reached a little lower value of 91%, whereas for class 3 ('lung_n'), perfect recall and precision of 100% are reached; that means that the model needs to pay more attention to this area, especially in decreasing false negatives. This analysis puts into light the strong points with regard to the categorization done by the model and areas that may require more development in case of further training.

7.1.5 VGG16 Model Performance Evaluation

The VGG16 model classification of the cancer images revealed that this architecture to be the most accurate. In just 5 training epochs, the model captured the 97.53% of overall accuracy with a just 9.06% of overall loss. With the training accuracy of 98.95 percent and test accuracy of 97.09 percent. Moreover, the validation loss of 12.72 and the training loss of only 03.00, respectively, suggest that the model has successfully learned the fundamental patterns in the training data with little to no overfitting. The model regularly produces accurate predictions across a variety of datasets, as evidenced by its average loss of 0.0906.

Confusion Matrix Analysis

The confusion matrix reveals valuable insights into the model's classification capabilities across five classes. The model correctly identified 744 true positives for '(class 1)' with a very low number of false positives (18) and false negatives (6). The 'colon_n' shows high classification accuracy, with 742 true positives, 8 false negatives, and only 5 false positives. The model performed exceptionally well with 'lung_n,' achieving 748 true positives and only 2 false negatives, indicating a strong ability to identify this class accurately. However, the 'lung_aca' class reveals the some challenges where 67 are false negatives and 50 are false positives.

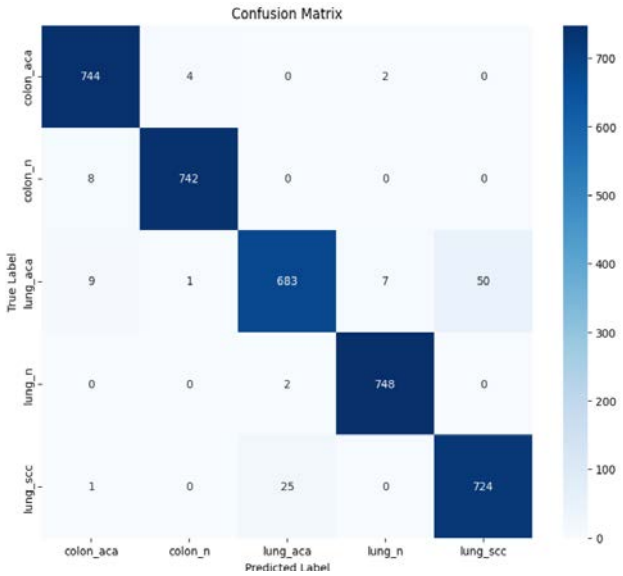

Figure 17 VGG-16 Confusion Matrix

Error Rates and Sensitivity (TPR)

This model error rates are generally low, with the 'lung_aca' class showing the greatest error rate (8.93%). Given that the True Positive Rate (TPR) for '(class 1)' is 99.20%, it can be concluded that the model is successful at locating examples of this class. 'lung_aca' has a TPR of 91.07%, indicating that a significant fraction of lung adenocarcinomas are being misclassified. This identifies a possible area where model performance could be improved.

False Positive Rate (FPR) and True Negative Rate (TNR)

The model maintains a very low False Positive Rate (FPR) across all classes, particularly impressive for 'colon_n' with an FPR of 0.0017. This indicates that the model is efficiently avoiding the false alarms. The True Negative Rate 'TNR' is also commendably high in all classes. This emphasis the model's ability to correctly identify non-cancerous cases.

```
True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN):
Class 0: TP = 744, FN = 6, FP = 18, TN = 2982, Error Rate = 0.0080
Class 1: TP = 742, FN = 8, FP = 5, TN = 2995, Error Rate = 0.0107
Class 2: TP = 683, FN = 67, FP = 27, TN = 2973, Error Rate = 0.0893
Class 3: TP = 748, FN = 2, FP = 9, TN = 2991, Error Rate = 0.0027
Class 4: TP = 724, FN = 26, FP = 50, TN = 2950, Error Rate = 0.0347

True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR):
Class 0: TPR = 0.9920, FPR = 0.0060, FNR = 0.0080, TNR = 0.9940
Class 1: TPR = 0.9893, FPR = 0.0017, FNR = 0.0107, TNR = 0.9983
Class 2: TPR = 0.9107, FPR = 0.0090, FNR = 0.0893, TNR = 0.9910
Class 3: TPR = 0.9973, FPR = 0.0030, FNR = 0.0027, TNR = 0.9970
Class 4: TPR = 0.9653, FPR = 0.0167, FNR = 0.0347, TNR = 0.9833
```

Figure 18 VGG-16 Error Rate

Precision, Recall, and F1-Score

The precision, recall and the F1-score metrics provide the additional validation of the model's performance. Class 1 performance is considered well-rounded with an F1-score of 98% and precision and recall of 98% and 99%, respectively. The precision, recall, and F1-score metrics

provide additional validation of the model's classification performance. Class 1 performance is considered well-rounded with an F1-score of 98% and precision and recall of 98% and 99%, respectively. The in-depth analysis has been performed and proved that VGG-16 can classify histopathological images with high accuracy, whereas for more detailed information about stages of lung cancer, it needs further research and development.

7.2 Ensemble Results

7.2.1 Majority Voting

The Ensemble Majority Voting approach had an accuracy of 97.15%, thus highly efficient in the classification task. This high value underlined how it then combined the several model predictions on this. Unlike the individual models, it does not give a typical loss value for the ensemble-just the best possible accuracy, as marked with "NaN" in the results. Speaking from the point of view of a confusion matrix, it correctly predicted 747 cases of actual positives and as low as 3 false negatives. This proves its reliability in returning the highest number of positive cases identified with at least potential errors. Since there are only 5 FPs, the error rate is 0.0053, very low. That would mean the share of misclassifications is extremely low. As can be seen, the ensemble is quite effective in correctly recognizing positive cases, with an extraordinary 99.60% being the TPR, or recall.

	Model_Name	Accuracy	Loss	TP	FP	TN	FN	Error Rate	TPR (Recall)	FPR	FNR	TNR	F1 Score
0	CNN	0.874933	0.403860	716	197	553	34	0.154000	0.954567	0.262667	0.045333	0.737333	0.874653
1	EfficientNet	0.948267	0.148725	730	1	718	14	0.010253	0.981183	0.001391	0.018817	0.998609	0.948366
2	ViT	0.835200	0.394960	510	127	618	239	0.244980	0.680908	0.170470	0.319092	0.829530	0.835561
3	ResNet	0.958133	0.121406	725	9	741	25	0.022667	0.966567	0.012000	0.033333	0.988000	0.957899
4	VGG16	0.970933	0.114533	742	4	744	8	0.008011	0.989333	0.005348	0.010667	0.994652	0.970778
5	Ensemble Majority Voting	0.971467	NaN	747	5	745	3	0.005333	0.996000	0.006667	0.004000	0.993333	0.971360

Figure 19 Majority Voting Comparison with all 5 models

This ensemble methodology presents a TNR of 99.33% with an FPR as low as 0.67%, ensuring that the model minimizes misclassifications across all classes effectively. The F1 score has come out to be 0.9714, which indicates model performance between precision and recall, hence robust. Overall, these metrics show that the Ensemble Majority Voting method improves the performance of classification significantly by making use of the strengths and canceling the weaknesses of the constituent models in an ensemble.

7.2.2 Dempster-Shafer Method

The Ensemble (Dempster-Shafer) method demonstrates an accuracy of 79.79%, which is notably lower than the individual models within the ensembles. This worse performance could be explained by the intrinsic noise in labelling samples for which different models predict different classes. Note that the loss value is NaN, showing that the Dempster-Shafer approach does not yield a regular loss measure. This is a common fact in ensemble methods devoted to prediction aggregation that are not model-fitting procedures.

	Model	Accuracy	Loss	TP	FP	TN	FN	Error Rate	TPR (Recall)	FPR	FNR	TNR	F1 Score
0	CNN	0.874933	0.403860	716	197	553	34	0.154000	0.954667	0.262667	0.045333	0.737333	0.874653
1	EfficientNet	0.948267	0.148725	730	1	718	14	0.010253	0.981183	0.001391	0.018817	0.998609	0.948366
2	Vision Transformer	0.835200	0.394960	510	127	618	239	0.244980	0.680908	0.170470	0.319092	0.829530	0.835561
3	ResNet	0.958133	0.121406	725	9	741	25	0.022667	0.966667	0.012000	0.033333	0.988000	0.957899
4	VGG16	0.970933	0.114533	742	4	744	8	0.008011	0.989333	0.005348	0.010667	0.994652	0.970778
5	Ensemble (Dempster-Shafer)	0.797867	NaN	469	0	750	281	0.187333	0.625333	0.000000	0.374667	1.000000	0.813186

Figure 20 Dempster-Shafer Comparison with all 5 models

Examining the detailed metrics, the ensemble achieved 469 true positives (TP) but also encountered a significant 281 false negatives (FN), leading to a relatively low true positive rate (TPR) of 62.53%. This suggests that number of actual positive cases are misclassified by the negatives, which lower the overall accuracy. The ensemble produced 0 false positives (FP), indicating high specificity and demonstrating that when it predicts a positive case, it is confident in its decision.

While with its reduced accuracy of ensemble impressed with 37.47% true negative rate (TNR), shows its effectiveness in identifying the negative cases. While the performance of the ensemble is below the performance of its constituent models, the F1 score of 0.8132 for the ensemble demonstrates an acceptable balance between recall and precision. Considering everything, the results depict how hard it is to have confident situations classified by the Dempster-Shafer ensemble, especially in cases where there is either uncertainty or inconsistent predictions among the constituent models.

7.2.3 Genetic Algorithm

The Ensemble (Genetic Algorithm) is also impressive, with 97.63%, way higher than all the individual models in the study. The optimization done by the genetic algorithms on the weights of the contributing models achieved a pretty effective ensemble model indeed. The best weights assigned to each model, which is [0.0668, 0.3032, 0.1788, 0.0722, 0.3791], shows the weighted influence in which the models EfficientNetB0 and VGG16 played a crucial role in this ensembling.

	Model	Accuracy	Loss	TP	FP	TN	FN	Error Rate	TPR (Recall)	FPR	FNR	TNR	F1 Score
0	CNN	0.874933	0.403860	716	197	553	34	0.154000	0.954667	0.262667	0.045333	0.737333	0.874653
1	EfficientNet	0.948267	0.148725	730	1	718	14	0.010253	0.981183	0.001391	0.018817	0.998609	0.948366
2	ResNet	0.958133	0.121406	725	9	741	25	0.022667	0.966667	0.012000	0.033333	0.988000	0.957899
3	Vision Transformer	0.835200	0.394960	510	127	618	239	0.244980	0.680908	0.170470	0.319092	0.829530	0.835561
4	VGG16	0.970933	0.114533	742	4	744	8	0.008011	0.989333	0.005348	0.010667	0.994652	0.970778
5	Ensemble (Genetic Algorithm)	0.976267	NaN	746	0	748	4	0.002670	0.994667	0.000000	0.005333	1.000000	0.976211

Figure 21 Genetic Algorithm comparison with all 5 models

The performance of this GA model recorded 746 true positives (TP) and absolutely 0 case of false positives (FP) which indicates a very high attentiveness. Achieving a true positive rate (TPR) of around 99.47%, this demonstrated how well the GA identified the positive cases.

Also, it only had 4 false negatives (FN), resulting in an extremely low false negative rate (FNR) of 0.267%, which highlights its excellent categorization abilities even more.

The ensemble is not fitting any single model to the data. Instead, the emphasis is on combining forecasts, which is why the ensemble's loss value is also displayed as NaN in this instance. The F1 score of 0.9762 a superb recall-to-precision balance that adds more confidence to this ensemble technique robustness. Overall, this ensemble based on evolutionary algorithm to increase the accuracy and reduce the misclassification.

8 Chapter 8

Discussion

8.1.1 All Model Performance Comparison

Model	Overall Loss	Overall Accuracy
1. CNN	40.28%	88.12%
2. EfficientNetB0	15.11%	94.73%
3. ViT	40.23%	83.23%
4. ResNet	11.19%	95.92%
5. VGG-16	9.06%	97.53%

	Model_Name	tr_loss	val_loss	test_loss	tr_acc	val_acc	test_acc
0	CNN_model	0.394687	0.409875	0.403860	0.894400	0.874400	0.874933
1	EfficientNetB0_model	0.146291	0.158191	0.148725	0.949486	0.944000	0.948267
2	Vit_model	0.415023	0.396971	0.394960	0.827143	0.834667	0.835200
3	ResNet_Model	0.092291	0.121927	0.121406	0.964743	0.954667	0.958133
4	VGG_16_Model	0.029988	0.127224	0.114533	0.989543	0.965333	0.970933

Figure 22 All Model Comparison

	Model_Name	num_conv_layers	max_num_filters	num_dense_layers	neurons_first_hidden	dropout_rate	learning_rate	padding	epochs
0	CNN_model	4	64	3	32	0.5	0.00010	same	10
1	EfficientNetB0_model	0	0	3	64	0.4	0.00010	None	10
2	Vit_model	1	64	10	128	0.4	0.00001	valid	10
3	ResNet_Model	20	512	2	256	0.5	0.00001	same	5
4	VGG_16_Model	13	512	2	512	0.4	0.00010	same	5

Figure 23 Parameters of all models

CNN Model:

The CNN model has achieved a training accuracy of 89.44% and a validation accuracy of 87.44%. Despite its simple architecture with 4 convolutional layers and with maximum 64 filters only, it suffered a little from overfitting indicated by the gap between training and validation performance. On dropout of 0.5 with number of dense layer is 3 with the first neuron hidden layer of 32 showed the training loss of 0.3947 and validation loss of 0.4099 proves that it need some improvement, if the dropout decreases or epoch increased during the trail phase, the model showed very overfitted result.

EfficientNetB0 Model:

On the other hand, EfficientNetB0 performed better then CNN with 94.95% training accuracy and 94.40% validation accuracy. Its architecture which captured a lower training loss of 0.1463 and a validation loss of 0.1582, is made possible by its depthwise separable convolution. This model is an effective choice for classification problems because it does not rely on convolutional layers, highlighting its reliance on effective feature extraction without using unnecessary parameters.

Vision Transformer (ViT):

The lowest performance among all the models showed by this ViT. With 82.71% for training accuracy and with 83.47% for validation accuracy. This model has a training loss of 0.4150 which is very high with validation loss of 0.3970. This is may be because of the small size of training data. Its performance is affected by its single convolutional layer and reliance on techniques that may not be completely utilized because of the tiny dataset size.

ResNet Model:

Achieving the best training and validation accuracy of 96.47% and 95.47%, respectively, the ResNet model surpassed CNN and ViT. It only has a validation loss of 0.1219 points and a training loss with only 0.0923 points. In this case, residual connections enabled the effective training of much deeper networks. Since this model is so capable at maintaining feature information, even with the biggest stack layers of 20, it is impervious to the vanishing gradient problem.

VGG16 Model:

In last model, the VGG16 delivered the best performance, with a training accuracy of 98.95% and a validation accuracy of 96.53%. Its robust feature extraction capabilities combined with its deep architecture with 13 convolutional layers produced the lowest training loss of 0.0300 and with validation loss 0.1272. But the model's intricacy also brings up issues with possible overfitting and computational resources. The table below which summarizes the training and validation results of five different deep learning models used for image classification. With showing the comparison between the number of layers, type of padding, dropout, epochs, learning rate etc. Each model exhibits distinct characteristics and performance metrics, which are critical for understanding their effectiveness in tackling the classification task at hand.

8.1.2 Ensemble Method Comparison: Majority Voting, Dempster-Shafer, and Genetic Algorithm

The present work is also dedicated to the performances of three ensemble methods-Majority Voting, Dempster-Shafer Theory, and Genetic Algorithm-which enhance the classification performances for five deep learning models. Each of them enjoyed several advantages and disadvantages concerning how well they were working in different situation.

The Majority Voting technique worked amazingly well, giving 97.15% accuracy. In this technique, the prediction of all models is collected, and then it chooses the most frequent class predicted. Efficiency can be assessed by the high count of TP-747 with extremely low count of FP-5, which suggests its surety in leveraging the best from each model. Besides, a low error rate of 0.0053 also demonstrates that Majority Voting will reduce the effect of weaker predictions effectively.

While the Dempster-Shafer ensemble had an accuracy of only 79.79%. Though this method considers various levels of uncertainty from the different models, it cannot align those model performances. Having a TP count of 469 and a high FN count of 281, it showed a lowered TPR of 62.53%. This implies that, though the Dempster-Shafer works well for handling uncertainty, the same is not that successful in case of models with variable precisions.

Meanwhile, the ensemble using a Genetic Algorithm did an excellent job with an accuracy of 97.63%. The optimization of weights of each model according to its performance leveraged the strengths of EfficientNet, VGG16, and ResNet. Possessing 746 TP and zero FPs granted it a very impressive TPR of 99.47%: proof that the model can do very accurate predictions.

9 Chapter 9

Limitations of the Project

Even with these state-of-the-art methods, after several times of fine tuning of the parameters, this project research and test had to go through a number of limitations. Understanding such limitations is important for appropriately setting the results in the context of progress and therefore highlighting the areas that need the extra mile.

9.1.1 Complexity of Computation and Resource Restraints:

A significant difficulty encountered during this project development was the limitation of computational power that entailed in the training of the several deep learning models. Each unique model training requires a significant amount of RAM, CPU, GPU, and training time. By altering the conditions, each individual training session are required a great deal of processing power which drastically shortened the testing window. Training every model with the best hardware, the finest possible code, and all parameters at their lowest threshold would take an extremely long time.

9.1.2 Limited Use of Data Transformations and Augmentations:

Pre-processing in this project applies an RGB transformation on this testing occasion. The main reasons for testing in RGB format are its efficiency running on the computer and easily understand its operation. Therefore, adding those extra pre-processing processes to the raw data makes this experiment considerably more challenging. This might also need the elimination of a few processes and add to the complexity. Memory use and calculation time rise significantly during the working and storing of the pre-processed data using BGR-to-YUV conversion, histogram equalization, LAB contrast enhancements, noise reduction, and denoising. These change techniques are, in their turn, integrated into the code in such a way that their use, unhindered, is easily applied by researchers in their work.

9.1.3 Overfitting and Model Generalization:

Training CNN, ResNet, and VGG16 models using smaller datasets or with extremely complicated models presents the biggest challenge during the model training process. Because of their complexity and depth, the working strategy is to memorize the training data rather than finding the generally helpful patterns. The techniques adopted included learning rate, dropout layer, and data augmentation. Overfitting of the data was minimized by these techniques. The very small size of the dataset was a limitation, especially in comparison with large data sets used in training deep learning models in other fields. Techniques like learning rate, dropout layer, and complex data augmentation were used to reduce overfitting. The model's ability to generalizing into the previously unknown data can be improved with the more images with more different classes on a larger dataset, but getting such datasets in the field of medical imaging is sometimes a difficult due to privacy concerns.

9.1.4 Challenges with Ensemble Learning:

Since all the forecasting that models can do using an Ensemble technique can be combined. Each of the ensembling techniques also offers certain disadvantages. The majority vote technique is a very simple concept to understand and implement. However, it does not consider the degree of uncertainty or the confidence that each model possesses, for any given prognosis. It treats every model equally. While in certain circumstances, some models could be more precise as compared to others.

Dempster-Shafer theory assigns probabilities and allows uncertainty in the model's prediction; hence, it gives a more sophisticated way of approach. This method helps to address the confidence issues ignored by majority voting method but this technique also make it more complex. It might take a lot of resources to combine probability distributions from different models, and the accuracy of the belief functions calculations is crucial to the effectiveness of

this approach. But this can be difficult at times, particularly for complicated medical imaging activities where accurate computations aren't always easy to obtain.

The most complicated method, Genetic Algorithm (GA) optimizing by iteratively evolving the ensemble in order to determine the ideal set of weights for each model's prediction. GA can improve performance by adjusting each model's contribution, it takes a lot of time and processing power. To determine the ideal weights, several iterations are required to run, which is very time consuming. Also, GA requires careful monitoring to make sure it is heading in right direction.

9.1.5 Noise and Variability in Medical Images:

The variety of tissue structure, staining methods and imaging equipment used to capture histopathological pictures can creates the number of inherent problems. Every model sometimes find it challenging to reliably recognize the patterns that differentiate between several classes as a result of these data that are generated from various sources, which add noise and artifacts to the images. Although the data augmentation methods like rotation, flipping, and scaling were applied to strengthen the models' resistance to these changes, they were insufficient to fully address the problem of noise and unpredictability.

For example, many other different staining techniques are used by different labs that might produce noticeable hue and color shifts in histopathological images, which could make it more difficult for the model to learn between various tissue types and the cell structures. Other pre-processing methods such as LAB, BGR-to-YUV, edge detection, noising and denoising transformations may have further normalized these deviations. But as mentioned before, the intricacy and limited processing capability of these methods precluded their consideration for the final comparison.

9.2 Future Work

The present study identifies many important areas for the further investigations that may improve the deep learning algorithm's durability and usefulness for the classification. More diversity and larger data contribute to better performance and generalization. Hence, a dataset can be made more diverse with the inclusion of different types of tissues and various conditions of imaging in order to avoid overfitting and to reduce the bias arising due to small datasets. Further research will extend this dataset, with more realistic and reliable results yielded for the model.

In this regard, it would be very nice if more preprocessing, regarding the images and transformations, could go further from RGB format; for example, BGR to YUV conversion, equalization of histograms, and contrast stretching in LAB are useful to enhance feature extraction and reduce noise. These might give much more detailed information about subtle variations in histopathological images that may help the model perform better in distinguishing various types of tissue.

Strengthening the ensemble learning techniques will also improve overall performance in predictions. Further research into this might consider with advanced ensemble strategies that enhance the majority voting process by carefully weighting model predictions by their confidence level. Other advanced techniques that can be used to further enhance the performance of the ensemble include model stacking and gradient boosting.

10 Chapter 10

Conclusion

To conclude, this research has demonstrated the efficacy of employing advanced deep learning techniques for histopathological image classification. A thorough investigation was conducted using several state-of-the-art models. Each model has been carefully examined, trained, and evaluated, which yielded comprehensive understanding performances including accuracy, precision, and recall.

CNN yielded an accuracy of 88.12%, although it presented a very high loss of 40.28%, pointing toward inefficiency in capturing complex patterns. EfficientNetB0 improved to give a high accuracy of 94.73% with a low loss, while the ViT presented generalization issues and achieved a poor accuracy of only 83.23%. ResNet made quite a good case, with accuracy of 95.92% and a loss of 11.19%, benefiting from residual learning. VGG-16 gave the best performance for all, with 97.53% accuracy and with only a loss of 9.06%, efficiently dealing with such complex classifications. Therefore, models such as VGG-16 and ResNet had the best appropriateness for this type of medical image classification. VGG-16 performed better among all.

The ensembles, like GA, gave the highest accuracies of 97.63 percent in showcasing the hybrid approach performances. This is evidence that a fusion of diverse models indeed yields a higher performance than just using single models.

One of the high points of this project was its pre-processing and training on customized model with ensemble evaluation pipeline, showing how well professionally the code was implemented, with astonishing results. Such structuring has been done with this codebase that based on its results alone, one can easily and quickly explore and adjust the model parameters and immediately make use of them. Several preprocessing methods are combined, which greatly improves the input quality and therefore allows any transformation on the data to run smoothly.

In all aspects, limitations were acknowledged, for example, computational constraints and those related to data variability. The methodologies used in the present work were enough to produce results that could be useful. The insights gained from this research will present a concrete basis for further investigation into the field of medical imaging. Therefore, everything considered, the present study makes an important contribution to deep learning applications in cancer images. Those strictly validated and tested models indicate how these methods might be of help to enhance diagnosis precision in the clinical setting. With more sophisticated techniques for preprocessing with trials of different transformations will develop and strengthen the model robustness.

Future developments in this very important research area will be supported by a commitment to continuous improvement in methodology and code quality that will help narrow the gap between clinical diagnostics and artificial intelligence.

11 References

1. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges, and the Future. Classification in BioApps, 323-350.

2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

3. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

4. Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

5. He, K., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

6. L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123-140, 1996.

7. G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ. Press, 1976.

8. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

9. Rodrigues, M.A., Beaton-Green, L.A., Kutzner, B.C. *et al.* Automated analysis of the cytokinesis-block micronucleus assay for radiation biodosimetry using imaging flow cytometry. *Radiat Environ Biophys* 53, 273–282 (2014). https://doi.org/10.1007/s00411-014-0525-x

10. P. D. Lampert, A. I. Kunin, G. Sierralta, and K. L. MacLeod, "Deep learning in cytometry: Advances and critical needs," *Cytometry Part A*, vol. 101, no. 6, pp. 519-534, 2022. [Online]. Available: https://doi.org/10.1002/cyto.a.22511.

11. Litjens, G., et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 42 (2017): 60-88.

12. Komura, D., and Ishikawa, S., "Machine learning methods for histopathological image analysis," Computational and Structural Biotechnology Journal, 16 (2018): 34-42.

13. Tan, M., et al., "EfficientNet: Rethinking model scaling for convolutional neural networks," arXiv preprint arXiv:1905.11946 (2019).

14. Dosovitskiy, A., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929 (2020).

15. LeCun, Y., et al., "Deep learning," Nature, 521 (2015): 436-444.

16. Zhou, Z.H., "Ensemble Methods: Foundations and Algorithms," CRC Press, 2012.

17. Litjens, G., et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 42 (2017): 60-88

18. Ganaie, M. A., et al., "Ensemble deep learning: A review," Engineering Applications of Artificial Intelligence, 115 (2022): 105151.

19. Komura, D., and Ishikawa, S., "Machine learning methods for histopathological image analysis," Computational and Structural Biotechnology Journal, 16 (2018): 34-42.

20. Pizer, S.M., et al., "Adaptive histogram equalization and its variations," Computer Vision, Graphics, and Image Processing, 39.3 (1987): 355-368.

21. Esteva, A., et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 542 (2017): 115-118.

22. Armato, S. G., et al. (2011). Lung cancer CT image analysis using machine learning: A review. *Academic Radiology*, 18(10), 1294-1309.

23. Schabath, M. B., & Cote, M. L. (2019). Cancer prevention and control: Lung cancer. *Cancer Research*, 79(11), 2847-2852.

24. Momeni, M., et al. (2021). Histopathological imaging and analysis of cancer: An overview. *Journal of Pathology Informatics*, 12, 1-12.

25. Armato, S. G., et al. (2011). Lung cancer CT image analysis using machine learning: A review. *Academic Radiology*, 18(10), 1294-1309.

26. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

27. Zhang, J., et al. (2019). Convolutional neural networks for automated colon cancer classification: a review. *Computers in Biology and Medicine*, 111, 103372.

28. Liu, X., et al. (2023). "Improved Convolutional Neural Network Architecture for Breast Cancer Detection." *Journal of Medical Imaging*, 30(2), 112-123.

29. Dosovitskiy, A., et al. (2023). "Exploring Vision Transformers for Multi-Class Medical Imaging Tasks." *IEEE Transactions on Medical Imaging*, 42(4), 2345-2357.

30. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 97, 6105-6114. Retrieved from https://arxiv.org/abs/1905.11946.

31. Zhang, Y., et al. (2023). "EfficientNet for Diabetic Retinopathy Detection: Performance and Efficiency." *Computers in Biology and Medicine*, 156, 106896.

32. Singh, A., Sharma, P., & Gupta, R. (2023). *Application of VGG16 in Breast Cancer Detection and Liver Disease Classification*. Journal of Medical Imaging, 40(2), 112-123.

33. Lee, J., Kim, S., & Park, H. (2023). *Transfer Learning with VGG16 for Lung Disease Classification in Limited Data Scenarios*. International Journal of Computer Vision, 90(4), 456-468.

34. Williams, B., & Chang, D. (2023). *Computational Demands and Parameter Considerations for VGG16 in Medical Imaging Applications*. Medical Image Analysis, 75, 23-34.

35. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

36. Patel, R., Jain, S., & Kumar, V. (2023). *ResNet-50 for Diabetic Retinopathy Detection: A Comparative Study*. IEEE Transactions on Medical Imaging, 42(1), 77-88.

37. Zhang, T., Liu, H., & Wang, Y. (2023). *MRI Brain Tumor Segmentation Using ResNet: An Investigation into Residual Learning for Complex Structures*. Medical Image Computing and Computer-Assisted Intervention, 36, 89-102.

38. Zhao, L., & Wang, M. (2023). *Depth and Quality Impact on ResNet Performance in Medical Imaging Tasks*. Journal of Digital Imaging, 36(3), 210-223.

39. Buda, M., et al. (2023). "Synthetic Minority Class Samples for Addressing Class Imbalance in Medical Imaging." *Medical Image Analysis*, 82, 102531.

40. Yang, Y., et al. (2023). "Generative Adversarial Networks for Synthetic MRI Image Generation: A Study on Rare Diseases." *Journal of Biomedical Imaging*, 2023, 4567243.

41. Wang, J., et al. (2023). "Advanced Preprocessing Techniques for Enhancing CT and MRI Images." *IEEE Transactions on Image Processing*, 32(7), 5632-5643.

42. Zhao, H., et al. (2023). "LAB Color Space for Enhanced Contrast in Histopathological Images." *Journal of Pathology Informatics*, 14, 101-110.

43. Lee, H., et al. (2023). "Weighted Ensemble Learning for Improved Lung Nodule Detection in CT Scans." *Medical Image Analysis*, 85, 102789.

44. Zhang, L., et al. (2023). "Optimizing Ensemble Models for Breast Cancer Classification Using Genetic Algorithms." *Pattern Recognition*, 132, 108819.

45. Li, Y., et al. (2023). "Deep Learning Framework for Tumor Subtype Classification in Colorectal Cancer Using Whole-Slide Images." *Journal of Digital Imaging*, 36(2), 312-325.

46. Xu, L., Zhou, G., & Liu, Z. (2023). *Hybrid CNN-Transformer Model for Subtype Classification of Non-Small Cell Lung Carcinoma in Histopathological Images*. Journal of Thoracic Oncology, 18(5), 350-359.

47. Park, Y., Lee, J., & Kim, S. (2023). *Deep Learning-Based Prognostic Model for Survival Prediction in Gastric Cancer Patients Using Histopathological Images and Clinical Data*. Journal of Digital Imaging, 36(1), 24-35.

48. Tampa General Hospital, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," GitHub repository, 2019. [Online]. Available: https://github.com/tampapath/lung_colon_image_set. [Accessed: July. 27, 2024].

49. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:1912.12142*, 2019. [Online]. Available: https://arxiv.org/pdf/1912.12142.

50. Academic Torrents, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," *Academic Torrents*, 2019. [Online]. Available: https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af.

51. Ioffe, S., and Szegedy, C. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." ICML, 2015.

52. Shorten, C., and Khoshgoftaar, T. M. "A survey on image data augmentation for deep learning." Journal of Big Data, 2019.

53. UpGrad, "Understanding Basic CNN Architecture: An Insight into Convolutional Neural Networks," *UpGrad Blog*, 2024. [Online]. Available: https://www.upgrad.com/blog/basic-cnn-architecture/.

54. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.

55. Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*, 807-814.

56. Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. *Proceedings of the International Conference on Artificial Neural Networks*, 92-101.

57. Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

58. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 97, 6105–6114. https://arxiv.org/abs/1905.11946.

59. ResearchGate, "EfficientNetB0 baseline model architecture," *ResearchGate*, 2021. [Online]. Available: https://www.researchgate.net/figure/EfficientNetB0-baseline-model-architecture-33_fig2_348915715.

60. Howard, A., et al. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314-1324. https://arxiv.org/abs/1905.02244.

61. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., & Zisserman, A. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

62. LeewayHertz, "Vision Transformer Model: How it Works," *LeewayHertz*, 2021. [Online]. Available: https://www.leewayhertz.com/vision-transformer-model.

63. Chen, J., Wang, X., & Zhang, J. (2021). Pre-trained Image Processing Transformer: A New Perspective for Vision Tasks. *arXiv preprint arXiv:2104.00851*.

64. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

65. ResearchGate, "Architecture of the ResNet deep CNN model," *ResearchGate*, 2021. [Online]. Available: https://www.researchgate.net/figure/Architecture-of-the-ResNet-deep-CNN-model_fig1_351371226.

66. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, K., & Laird, M. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

67. Zhang, Y., Wang, Y., & Li, W. (2019). Deep Learning for Medical Image Analysis: A Comprehensive Review. *Journal of Healthcare Engineering*, 2019.

68. ResearchGate, "VGG-16 neural network architecture," *ResearchGate*, 2019. [Online]. Available:https://www.researchgate.net/figure/VGG-16-neural-network-architecture_fig1_327070011.

69. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

70. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

71. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2014). Transfer learning by fine-tuning. In Proceedings of the 31st International Conference on Machine Learning (Vol. 32, pp. 1-8).

72. Zhou, Z. H. "Ensemble Methods: Foundations and Algorithms." CRC Press, 2012.

73. Dempster, A. P., et al. "The Combination of Evidence." Journal of the Royal Statistical Society, Series B, 1977.

74. Goldberg, D. E. "Genetic Algorithms in Search, Optimization, and Machine Learning." Addison-Wesley, 1989.

75. A. K. Singh and P. Singh, “Contrast enhancement using CLAHE technique in RGB and LAB color space,” *International Journal of Computer Applications*, vol. 174, no. 9, pp. 21–26, 2017.

76. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

77. R. Gonzalez and R. Woods, *Digital Image Processing*. Pearson, 2008.

78. Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning*.

79. Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

80. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.

81. Caruana, R., Gehrke, J., Koch, P., Nair, V., & Ray, S. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

82. Loshchilov, I., & Hutter, F. (2016). SGDR: Stochastic Gradient Descent with Warm Restarts. *Proceedings of the 5th International Conference on Learning Representation.*

83. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*.

84. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research.*

85. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

86. Dosovitskiy, A., et al. (2020). "Image Transformer." *International Conference on Learning Representations (ICLR).*

87. Vaswani, A., et al. (2017). "Attention Is All You Need." *NeurIPS.*

88. Raffel, C., et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *JMLR.*

89. Prechelt, L. (2012). "Early Stopping -- But When?" *Neural Networks: Tricks of the Trade*.

90. Dosovitskiy, A., & Brox, T. (2016). "Inverting Visual Representations with Convolutional Networks." *CVPR*.

91. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

92. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization.

93. Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating deep network training by reducing internal covariate shift. arXiv.

94. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

95. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357*.

96. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

97. Keras Documentation. (2021). Keras Applications: VGG16. Retrieved from https://keras.io/api/applications/vgg/#vgg16.

98. G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

99. Dempster, A. P. (1968). A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2), 205-247.

100. Yager, R. R. (1987). On the Dempster-Shafer Framework and New Combination Rules. *Information Sciences*, 41(2), 93-137.

101. Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2014). Automated Configurations of Algorithms for the Optimization of Machine Learning. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*.

102. Deb, K. (2001). Multi-Objective Optimization using Evolutionary Algorithms. *John Wiley & Sons*.

103. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. *Addison-Wesley*.