

Colorectal Cancer Detection Using Transformer-based Approach with Attention Mechanism

by

Showmen Sarker

19301188

Sadman Fardin

19301068

Saik Rahman

19101011

Md.Tanjimul Islam

19101613

Golam Sifat

20301478

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2022

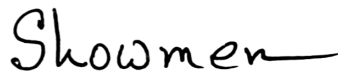
© 2022. Brac University
All rights reserved.

Declaration

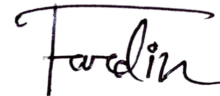
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Showmen Sarker
19301188




Sadman Fardin
19301068



Saik Rahman
19101011



Md. Tanjimul Islam
19101613



Golam Sifat
20301478

Approval

The thesis/project titled “Enhancing Security of Steganography Using Deep Learning” submitted by

Md.Tanjimul Islam (19101613)
Saik Rahman (19101011)
Sadman Fardin (19301068)
Showmen Sarker (19301188)
Golam Sifat (20301478)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23,2023 .

Examining Committee:

Supervisor:
(Member)



Md. Tanzim Reza
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)



Shakib Mahmud Dipto
Lecturer
Department of Computer Science and Engineering
University of Liberal Arts Bangladesh (ULAB)

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Image classification is the process of labeling and classifying pixels or vectors within an image according to preset rules. Classification can be done using spectral or textural features. Computer vision researchers focus on image classification, localization, segmentation, and object recognition. One of the biggest challenges is image classification. It's a foundation for various object recognition problems. Image classification applications are used in medical imaging, satellite object tracking, traffic management, brake light detection, and many more fields. Try to uncover more real-world photo categorization applications in our complete list of AI vision applications. "Maximum likelihood" and "minimum distance" are two popular training data-based picture categorization algorithms. The "maximum likelihood" classification analyzes the picture's textural and spectral indices' standard deviation and mean values to take advantage of statistical data. Using a normal distribution on each class's pixel data, the chance of each pixel belonging to each class is calculated. Many traditional statistical approaches and probabilistic relationships are also applied. The highest probability pixels are given to a group of characteristics. We used the Vision transformer's attention-based method to distinguish afflicted and healthy colons during our investigation. Our path has involved using various models to achieve the best result. We next compared CNN model findings to our chosen transformer model ViT16, which supports attention-based techniques. Colorectal cancer detection models include VGG16, VGG19, Resnet101, and Resnet 50. The results were then compared to our model ViT16. We chose the best Colorectal Cancer Detection model from the comparison. We compared results based on val_accuracy, val_loss, precision, recall, and f1_score to select the best model. The confusion matrix was another sign that the ViT-16 model worked well. In this report, ViT-16 had the top val_accuracy, val_loss, Precision, Recall, and f1 score, while ResNet101 ranked second. Thus, ViT-16, which uses the attention mechanism, is the best model for colorectal cancer detection.

Keywords: image classification; object recognition; traditional statistical methods; maximum likelihood; traffic management; categorization applications

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our Supervisor Tanzim Reza sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vi
List of Tables	vii
1 Introduction	1
1.1 Thoughts Behind The Prediction Model	1
1.2 Aims And Objectives	2
1.3 History Of Colorectal Cancer	2
1.4 Problem Statement	3
2 Literature Review	5
2.1 Literature Review	5
2.2 Research Objectives	7
3 Research Methodology	8
3.0.1 VGG-16	9
3.0.2 VGG-19	11
3.0.3 ResNet-50	13
3.0.4 ResNet-101	15
3.0.5 ViT Model	17
3.0.6 Swin Transformer	19
3.1 Dataset Collection	21
3.2 Data Pre-Processing	23
3.2.1 Data Labeling	23
3.2.2 Data Splitting	23
3.2.3 Dataset Class Distribution	24
3.2.4 Proposed ViT model	26
3.3 ViT Model Layers	28
3.3.1 Token Embedding	28
3.3.2 Multi-Head Attention	28

3.3.3	Position-wise Feed forward Network	28
3.3.4	Layer Normalization	28
3.3.5	Classifier	29
3.4	ViT Model Mathematical Operations	29
3.4.1	Linear Projection:	29
3.4.2	Positional Encoding:	29
3.4.3	Self-Attention Mechanism:	30
3.4.4	Softmax:	30
3.4.5	Loss Function:	30
4	Implementation Of Results	32
4.1	Findings	32
4.1.1	Colorectal Cancer Analysis With VGG-16	32
4.1.2	Colorectal Cancer Analysis With VGG-19	34
4.1.3	Colorectal Cancer Analysis With ResNet-50	36
4.1.4	Colorectal Cancer Analysis With ResNet-101	37
4.1.5	Colorectal Cancer Analysis With ViT-16	39
4.2	Result Analysis	40
4.2.1	Why ViT-16 Excels: Outperforming VGG16, VGG19, ResNet101 and ResNet50	46
5	Future Works And Discussion	48
6	Conclusion	51
	Bibliography	51

List of Figures

3.1	Architecture Of VGG16 Model	10
3.2	Architecture Of VGG19 Model	12
3.3	Architecture Of ResNet Model	17
3.4	ViT Model	18
3.5	Swin Transformer	19
3.6	Workflow Diagram	20
3.7	Colon Benign Tissue	21
3.8	Colon Adenocarcinoma	22
3.9	Dataset Class Distribution	26
3.10	Vision Transformer(Vision Transformer architecture — main blocks. First, image is split into fixed-size patches and flatten. Second, position embeddings is added, and resulting sequence of vectors is forwarded to a standard Transformer encoder.)	26
3.11	Transformer Encoder	27
3.12	GELU Activation Function Through Graph	31
4.1	Accuracy Curve of Vgg-16	33
4.2	Validation Loss Curve of Vgg-16	33
4.3	Confusion Matrix of Vgg-16	34
4.4	Accuracy Curve of Vgg-19	34
4.5	Validation Loss Curve of Vgg-19	35
4.6	Confusion Matrix of Vgg-19	35
4.7	Accuracy Curve of ResNet-50	36
4.8	Validation Loss Curve of ResNet-50	36
4.9	Confusion Matrix of ResNet-50	37
4.10	Accuracy Curve of ResNet-101	37
4.11	Validation Loss Curve of ResNet-101	38
4.12	Confusion Matrix of ResNet-101	38
4.13	Accuracy Curve of ViT-16	39
4.14	Validation Loss Curve of ViT-16	39
4.15	Confusion Matrix of ViT-16	40
4.16	Comparison Between Used Models Based On Accuracy	41
4.17	Comparison Between Used Models Based On Validation Loss	42
4.18	Comparison Between Used Models Based On Precision	43
4.19	Comparison Between Used Models Based On Recall	44
4.20	Comparison Between Used Models Based On F1-score	45

List of Tables

3.1	Table Of Pre-Processed Data	23
3.2	Table Of Data Splitting	24
4.1	Comparison Between Used Models	40

Chapter 1

Introduction

1.1 Thoughts Behind The Prediction Model

Cancer occurs when cells in the body divide at an accelerated pace. A mass, or tumor, forms as these rogue cells multiply. The term "cancer" is used to describe over a hundred distinct illnesses. It has a wide range of potential growth sites throughout the body. All living things are composed of cells. Each time the body requires more of a certain kind of cell, it triggers the growth and division of the relevant cell type. Cells often perish because they are too old or damaged. Once those cells die, other ones take their place. Colorectal cancer refers to cancer of the rectum or colon. These malignancies are known by several names, including colorectal cancer and colon cancers, depending on where they arise. Because of their similarities, rectal and colon cancer are often regarded as the same illness. The large intestine consists of the colon and rectum and is a portion of the digestive system (GI system). The colon, a tube-like muscular structure around 5 feet long, makes up the bulk of the large intestine. The passage of food through the colon is used to divide into several portions.[25] Water and salt absorbed from the small intestine's remnants are transported to the colon (small bowel). The rectum, the remaining 6 inches (15 cm) of the gastrointestinal system, receives trash after the colon. The anus is the route then taken. Sphincter muscles in the shape of a ring around the anus prevent waste from passing through until they relax. The colon and rectum are common sites for the development of colorectal cancers. The tissue growths you see here are called polyps. Some polyps, but not all, may catch the disease over time. The likelihood of a polyp turning malignant varies according to its kind. Polyp cancer spreads to the colon or rectum wall. There are multiple levels in the colon and rectum. Colorectal cancer develops in the mucosa and can spread to all layers of the colon.[22] Cancer cells in the walls may develop into blood or lymph vessels. They have the potential to spread to the surrounding nodes or to distant areas of the body. The Colorectal cancer stage is defined by how far it has gone and whether it has gone further than the rectum or colon. Machine learning and artificial intelligence can help in cancer prediction. Ai technology can detect existing tumors and pinpoint those who are at higher risk of developing the disease before it becomes an issue. This enables physicians to closely monitor these people and intervene promptly if necessary.

1.2 Aims And Objectives

A deep learning model for predicting the outcomes of colorectal cancer screenings is the focus of this study, which aims to apply "Image Classification" to that end. By focusing on the results of the transformer-based strategy, we hope to attract the attention of those who may benefit from the transformer-based strategy in the fight against colorectal cancer.[14] Researchers may be able to do better studies if they have a better understanding of the significance of qualities. Better results may be achieved if, in the future, researchers are able to determine the significance of particular characteristics.

1.3 History Of Colorectal Cancer

Adenoma polyps are aberrant cells that serve as a precursor for colorectal cancer by promoting epithelium tumorigenesis and crypt instability.[12] Colorectal cancer is characterized by a lengthy development period, typically spanning 10-15 years from the appearance of these precursor lesions. This gradual progression is marked by a series of phases accompanied by mutations in various repressors, leading to anomalies in cell control.[4] Both environmental factors and an individual's genetic makeup play critical roles in this complex cycle of causation and effects. Family history and extensive genetic research have been instrumental in identifying those most susceptible to colorectal cancer. Broader identification of individuals genetically predisposed to this disease, as well as early detection and screening of the general population, offers promising avenues for future control of colon cancer.[36] Lifestyle factors and personal adjustments are also vital in the initial prevention of colorectal cancer. The NPS (National Polyp Study) illuminated how potential adenomatous polyps could develop into malignant colorectal tumors.[44] To qualify for the study, participants needed to have one or more colorectal adenomas without signs of aggressive malignancy. These polyps were removed, and subjects were monitored for over six years through many types of screening methods, colonoscopy, and stool guaiac test. Interestingly, approximately 75% of colorectal cancer diagnoses occur in individuals without known risk factors. Those lacking known risk factors and aged 50 or older fall into the sporadic category. The remaining 25% represents vulnerable populations, including 5% with genetic nonpolyposis colorectal cancer (HNPCC), 1% with inflammatory bowel disease, and 1% with familial adenomatous polyposis (HNPCC). This category accounts for the other 15% - 20% of those highly at risk. Mutations in the APC gene are now considered pivotal in transforming normal colonic mucosa, initiating a cascade that may result in cancer. These changes, called somatic mutations, occur in an organism's cells and are not hereditary. Hyperproliferation and abnormal crypt foci may first appear as benign adenomas, evolving over time into malignant tumors.[38] Notably, only about one in twenty minor adenomas will progress to substantial adenomas, emphasizing the complexity of cancer progression.

1.4 Problem Statement

The process of categorizing images is a straightforward undertaking that aims to comprehend them in their entirety. We are attempting to determine what type of label would be most appropriate for this image so that we can put it away properly. The analysis of single-subject images is the most common application of the word "Image Classification," and it describes the process by which the term "Image Classification" is utilized. Object detection, on the other hand, is beneficial for researching more realistic circumstances in which numerous things may coexist in an image. Because object detection involves both classification and localization tasks, it is appropriate for this kind of research. The building of a model will be straightforward with image classification, which is another advantage of this strategy, along with its dependability. Appropriate for crucial endeavors such as the categorization of medical images. A task such as transfer learning, which requires less time and data to train for, often gets satisfactory results in the majority of the situations. The complicated mathematical computations that take place within the model are the source of the issue. Because training takes up a significant amount of time, we would require more sophisticated equipment and more powerful processors.

Under the heading "Image Classification," we are able to track down a component that is referred to as a "Attention Based Mechanism." The attention mechanism of Neural Networks is, in general, a fair approximation of the cognitive attention that is possessed by people. The primary goal of this feature is to bring more attention to the aspects of the data that are most significant while drawing less attention to the characteristics that are less relevant. This feature also has the secondary goal of reducing the amount of attention drawn to the characteristics that are less important.[40] Following this method is essential in order to avoid a crash of the system because of the fact that both people and computers have constrained quantities of working memory. When talking about "deep learning," the phrase "attention" can be thought of in the same way as a vector of significance values. This is because both concepts are related to "deep learning." When making predictions, the attention vector is used to make inferences about the level of relationship that exists between the various components.[24] These inferences are then used to create the predictions. These elements could be anything, from the pixels that comprise an image to the words that constitute an assertion.

Colorectal cancer is one of the numerous forms of cancer, but it is the one that causes the most deaths worldwide. Recent developments in deep learning have made it possible to make enhancements to detection using picture categorization. Despite their excellent accuracy in making predictions, the application of deep learning algorithms in clinical practise is no longer practical because these algorithms cannot be interpreted. Convolutional neural networks have started a revolution in the fields of computer vision, signal processing, and natural language processing. This is due to the fact that convolutional neural networks are capable of representing features in a superior manner. It has been demonstrated that using convolutional encoder-decoder designs is beneficial for position-sensitive applications such as semantic segmentation that contain numerous areas of interest (ROIs). The model has the potential to capture global to local information with the addition of additional layers and a

stride kernel. Because of their predetermined size and form, these kernels can only function well with a constrained range of inputs.

In both NLP and CV, the most cutting-edge technology utilises a transformer-based design that combines self-attention and long-range modelling. Tokens for the transformer pipeline are produced by the vision transformer (ViT), which are derived from these patches. The transformer has been shown to generate better results than the conventional CNN model while requiring only a few lines of additional code. However, in order to compete with a convolutional neural network, you would need a very expensive processing system and a very vast dataset. This is unachievable.

Because there is a lack of clinical data that has been adequately annotated, effective model creation in medical image processing is impossible, necessitating the usage of the ImageNet pretrained model. When it comes to the identification of medical regions-of-interest (ROI), the most popular techniques, such as ResNet and DenseNet, rely on their pretrained weights and only modify the upper layers for maximum performance on a certain number of specialised applications. However, you can only utilise the vast majority of these approaches with newly collected data sets. In order to facilitate downstream transfer learning, big annotated datasets will need to be decoupled from the pretrained model of ImageNet and other datasets through the use of fully supervised learning. On the other hand, self-supervised learning frameworks have lately garnered a lot of attention in the field of medical image analysis due to their ability to offer solid findings despite the absence of labelled datasets.[18] This is because these frameworks can train without direct supervision using their own data. In addition, there has been a lot of interest in the self-supervised learning paradigm in the field of medical picture analysis. Defining an appropriate proxy task from unlabeled data presents a significant challenge for self-supervised learning because it is dependent on the data. In spite of the fact that these proxy tasks have proven to be helpful in other settings, their performance in medical picture tasks has been significantly less satisfactory.

Chapter 2

Literature Review

2.1 Literature Review

Innovation has always been a cornerstone of medical fields. Due to the advancement in modern medical sectors, more effective methods of cure and treatment are available that may help humans to live a little longer. Different branches of medical fields are evolving and benefiting because of the technological advances of medical sector. Medical technology now can personalize medicine, diagnose disease, perform robotic surgeries that have opened up possibilities beyond what anyone thought was impossible. Cancer still remains the most feared illness in all over the world. The development in new technology in cancer are helping more and more people trying to treat their cancer effectively that may help them live longer and healthier than before. With then help of medical image processing the detection of cancer have achieved remarkable progress in the last few years.[33] Colorectal cancer is third most common cancer. These cancers are also called colon cancer. It is also responsible for highest amount of death. Some researchers used transformer-based approach to diagnose colorectal cancer previously. Authors of [29] worked to mark the region caused by the colorectal cancer in order to detect the cancer. They used a framework named CST that use a novel transfer learning protocol and a segment tusk depending on the transfer model. To use the decision more accurate they followed an image level decision approach based in an auto-encoder.[30] With the combination of autoencoder and transformer architecture they used a framework capable of performing multitask. Authors of [4] applied vision transformer to perform multiclass tissue classification. They compared vision transformer with compact convolutional transformer with the help of CRC histology image dataset consisting of 5000 images with eight category of tissue and gained higher accuracy.[4] Authors of [42] tried to evaluate the transformer-based approach more systematically. They evaluated six transformer models with the use of PAIP liver histopathological dataset.[50] Then they compared the models based on transformers with six major models based in CNN and show that the models that used transformers show better result. Some researchers used transformers to improve diagnosis in other types of cancers. Authors of [27] wanted to collect the information about the information about the relationship between multiple mammograms from a patient. The unregistered mammograms create the breast lesion feature and it is important to capture that feature.[28] To catch the relationship between the mammograms it is better to use the vision transformers than CNN because CNNs can not model long-range dependencies well. Authors

of [25] used the time-dependent features to get repeated images and capture the changes to identify the lesions.[26] In this case time-distance vision transformers are used to capture repeating images because a single image cannot provide the information of a lesion changing over time in case of lung cancer. In the year 2020, [6] conducted an investigation of a neural net (NN) structure that had been subjected to 10-fold cross-validation. In comparison to the other methods, the artificial neural network and ensemble learning strategy was able to achieve a better level of accuracy.[6] In addition to that, the models were verified using a dataset consisting of malignant mesothelioma. The use of genomic, transcriptomic, and histopathological data, amongst others, is facilitating the transition toward personalized medicine in cancer treatment, which is made possible by the increased availability of these data and their integration. When doing translational research or treatment procedures, it is necessary to invest a large amount of time and expertise in order to effectively utilize and comprehend a variety of high-dimensional data types. In addition, knowing individual data types is less time-consuming and requires less resources than comprehending many data types at once. Additionally, modeling techniques that are able to learn from just a large number of complicated components are required.[16] An experimental study [20] was conducted in 2021 using three genuine databases (hyperglycemia, heart, and cancer) obtained first from UCI repository.[21] The study was done out on real data. A host - based and network ensemble learning technique was proposed as a possible approach to the categorization of illnesses in this research. On the datasets pertaining to diabetes, heart disease, and cancer, respectively, the computational model obtained a respectable level of accuracy by achieving a score of 98.5, 99, and 100%. A CNN algorithm was provided in a more recent research [5] that was carried out in 2021 with the intention of predicting whether or not prostate cancer patients had metastases. The categorizing technique resulted in fruitful outcomes.[5] The neural model was successful in achieving an AUROC score that was, on average, 68percent. In a subsequent research [8] carried out in 2021, CNNs proved their substantial classification performance.[8] This experiment was the most current of its kind. In this particular piece of research, CNNs were used to produce forecasts about cancer. The research made use of an authentic collection that comprised 311 persons who had been diagnosed with cancer as the dataset that was employed in the study. CNNs were employed first for the purpose of determining the pertinent feature, and subsequently machine learning models like SVM and KNN were utilised for the purpose of diagnosing cancer in patients. Using histopathology pictures as input, a further effective hybrid deep learning technique for diagnosing prostate cancer was suggested in the year 2021.[19] This paper discussed a revolutionary approach to the picture segmentation process that was given the moniker RINGS (Rapid Registration of Glandular Structures). This procedure was more accurate than any other method that was considered state-of-the-art and reached a 90 percent success rate. In the first step of the process, a method called deep learning was used to determine the locations with increased mitotic activity could be identified. The SVM model was then used to provide a prediction about the ultimate tumor growth. The proposed method was successful in achieving an accuracy of 74 percent and greatly excelled all other methods that had been tried before. A groundbreaking study [19] analyzing photographs of skin lesions for the purpose of predicting skin cancer was carried out in the year 2021. In the course of the research study, a comprehensive analysis was carried out, and the most common

challenges associated with skin cancer diagnosis were uncovered. In addition, the performance of conventional models was analyzed in this study, and ensemble-based deep learning techniques were proposed with the objective of increasing the accuracy of predictions. Breast cancer prediction using ANN methods and the SEER dataset was the subject of another 2021 study. According to the findings of this research, preprocessing techniques have the potential to enhance cancer prediction results.[20] A cervical cancer prediction algorithm based on ensemble learning was developed by [19] in the year 2021. The dataset that was used in the research was obtained from the repository at UCI. In this particular investigation, KNN imputations were used to fill in the blanks left by missing data, and data balancing methods were also utilized. Due to the unbalanced nature of the dataset, the data had to be corrected by the use of the oversampling method. The most important risk indicators were determined via the use of a random forest feature selection. The ensemble architecture that was recommended in the research worked very well and got a score of 99 percent for its area under the curve (AUC).[15]

2.2 Research Objectives

The goal is to train a deep learning model that can use "Image Classification" to predict the results of a colorectal cancer screening. Our primary goal is to showcase the outcomes of the transformer-based approach to colorectal cancer through the attention mechanism.[4] When researchers better grasp the importance of attributes, they may conduct more effective investigations. If future studies can figure out what traits mean, better job might be done. Our try is to find out a optimum and efficient solutions for finding colorectal cancer cells and different types of models will be used to reach our goal.[11]

1. Classifying colorectal cancer using histopathology pictures CNN models like VGG-16, VGG-19, and ResNet101.
2. We will employ Whole Slide pictures of cancerous cells and non cancerous cells of colorectal cancer in the image classification process in order to determine which of the two types of cells is impacted.[35] In order to finish this procedure in a flawless manner, datasets that contain images will be needed.
3. In the not too distant future, explainable artificial intelligence (AI) will be employed in order to achieve higher quality detection and raise the degree of accuracy currently existing in the algorithm. This will be done in order to obtain higher quality detection.[7]
4. The conclusion will allow us to determine whether or not a cell has colorectal cancer traces. Additionally, if it detects any new traces, it will attempt to learn and incorporate them in the future for more precise findings.

Chapter 3

Research Methodology

In our attempts to find out the optimal result for colorectal cancer detection using transformer based approach's attention mechanism we will be using the following research methodology which has been represented using a flow diagram below. In the beginning of our research we have set up the goal of collecting data which will be used later. After collecting the data we will move onto the data pre processing part which includes Scaling augmentation. Moving on, using these pre processed data we will move onto Training Data State where we will different Deep learning model. For example, we will be using VGG16,VGG19,Resnet 50,RestNet 101 which are very effective regarding researches related to image processing only to compare the findings with our proposed model VIT. In the meantime we will also try to find out the optimal result by comparing the findings with other models as our goal is to help in cancer detection with top most accuracy. Testing and validated data will be sent to test and validate. At this state, the trained data will also be sent after being trained by deep learning models. The next state will be Comparison and analysis state. Here we will compare our tested ,trained and validated data with outside world which will test the accuracy of our researched model.[39] Then we will look forward to implement the preferred modeling Technique to achieve our goal. Finally ,the analyzed result will be ready. For the next step these newly obtained data will be sent to Test again to gain more and more accuracy and that is how we are planning to conduct the whole research.

We initially embarked on a project involving 10,000 images collected from various sources, with a focus on image classification. The dataset comprises 10,000 images in total, with a split of 8000 images for training and 2000 images for testing and validation.[9]

The transformer model has emerged as one of the most significant contributions to the field of deep learning and neural networks in recent years. Its primary function is in high-level natural language processing applications. RNN-based encoder-decoder architectures were previously used to incorporate attention to neural machine translation before the development of the Transformer model. By eliminating repetition and convolutions and replacing them with a self-attention mechanism, the Transformer model completely rethought the way in which attention is implemented.

Most cutting-edge NLP systems formerly depended on LSTMs and gated recurrent

units (GRUs) as examples of gated RNNs with additional attention mechanisms before the advent of transformers. In the same way, as RNNs employ attention processes, Transformers do as well, although they lack the recurring structure of RNNs. That is, without the RNNs' attention, the performance may be matched by the attention mechanisms alone, given enough training data. One type of deep learning model, known as a "transformer," uses a technique called "self-attention" to assign various weights of importance to different parts of the input data. The Encoder-Decoder architecture for machine translation with neural networks benefits significantly from the addition of an attention mechanism. In the field of Deep Learning, "attention" is a major concept.[32] This method is currently being used for various issues, including picture captioning and others. In the 1990s, attention-like processes were developed under the titles multiplicative modules, sigma pi units, and hyper networks. we will work on this method in our entire thesis.

3.0.1 VGG-16

VGG-16, an abbreviation denoting "Visual Geometry Group 16," stands as a formidable exemplar within the pantheon of deep convolutional neural network (CNN) architectures.[27] Its inception can be attributed to the pioneering work of the Visual Geometry Group at the University of Oxford, bearing the indelible mark of architectural ingenuity. This architectural marvel, first unveiled in the paper titled "Very Deep Convolutional Networks for Large-Scale Image Recognition" by Karen Simonyan and Andrew Zisserman in 2014, has etched its name in the annals of computer vision history.

In an academic exposition, it becomes imperative to unravel the exceptional facets that render the VGG-16 configuration noteworthy:

Layer Depth: VGG-16 distinguishes itself through its exceptional depth. Comprising a total of 16 weight layers, it constitutes a symphony of convolutional and fully connected strata. The sequential alignment of these layers, each meticulously fine-tuned, begets an architectural opulence that transcends prior conventions.

Uniform Convolutional Layers: One of the hallmarks of VGG-16 is its uniformity in architecture. Throughout its convolutional layers, it adheres steadfastly to 3x3 convolutional filters, which are densely stacked. This uniformity imparts a fine-grained approach to feature extraction, contributing to its exceptional performance.

Max-Pooling Layers: VGG-16's distinctive character lies in its consistent incorporation of max-pooling layers. These layers serve as architectural keystones, contributing to spatial down-sampling while preserving the salient features inherent in the input data. Their presence endows the network with robustness against spatial variance.

Fully Connected Layers: At the zenith of its architectural hierarchy, VGG-16 houses three fully connected layers, culminating in the output layer. These layers harmonize seamlessly to distill high-level abstractions from the hierarchical features extracted in the preceding convolutional layers.

Classification Prowess: An exceptional attribute of VGG-16 is its prowess in image classification. It was instrumental in achieving top positions in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2014. The network’s capacity to discern objects within images and classify them into a wide array of categories has solidified its legacy.

Transfer Learning: VGG-16’s architecture has catalyzed transfer learning paradigms. The pre-trained VGG-16 model, fine-tuned on specialized datasets, serves as a versatile starting point for diverse computer vision applications. Its convolutional layers are often adopted as feature extractors, allowing for rapid training of custom classifiers.

Computational Complexity: While VGG-16 exhibits exceptional performance, it is noteworthy that its depth entails substantial computational complexity. Training and inference times may be protracted, necessitating computational resources commensurate with its architectural grandeur.

In summation, VGG-16, bearing the imprimatur of the Visual Geometry Group, emerges as a paragon of architectural integrity and performance. Its exceptional depth, uniformity, and classification prowess have left an indelible mark in the realm of computer vision. The VGG-16 configuration serves as a beacon of transfer learning, enabling its profound architectural insights to illuminate a myriad of applications. Yet, it is vital to recognize that its computational demands are commensurate with its exceptional capabilities, emphasizing the need for computational resources that match its architectural grandeur.

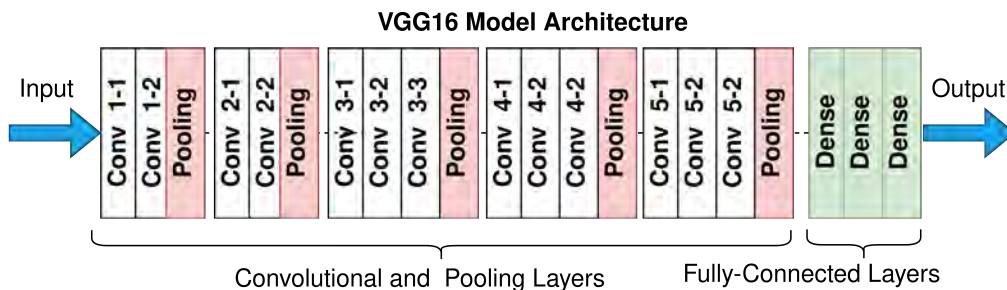


Figure 3.1: Architecture Of VGG16 Model [42]

3.0.2 VGG-19

VGG-19, as the nomenclature suggests, denotes the 19 weight layers that constitute this architectural gem.[26] It is a product of the distinguished Visual Geometry Group at the University of Oxford, and it emerged onto the scene with the seminal paper titled "Very Deep Convolutional Networks for Large-Scale Image Recognition," authored by Karen Simonyan and Andrew Zisserman in 2014.

Architectural Elegance: One of the salient aspects that immediately distinguishes VGG-19 is its architectural elegance. While its predecessor, VGG-16, bore the mantle of depth, VGG-19 extends this profundity by an additional three convolutional layers. This ensemble of 19 weight layers, meticulously arranged, orchestrates a symphony of feature extraction and abstraction.

Convolutional Consistency: An aspect of VGG-19's elegance lies in its consistency. It ardently adheres to 3x3 convolutional filters throughout its convolutional layers. This uniformity may appear simplistic, yet it underpins a nuanced approach to feature extraction. The densely stacked 3x3 convolutions engage in a hierarchical exploration of the input data, uncovering features of varying granularity.

Pooling Paradigm: In the VGG-19 architecture, max-pooling layers punctuate the convolutional strata, serving as architectural milestones. These layers introduce spatial down-sampling, reducing the dimensions of feature maps while preserving the essential features. The judicious inclusion of max-pooling layers imparts robustness against spatial variations—a hallmark of VGG-19's design.

Fully Connected Finale: At the apex of its architectural hierarchy, VGG-19 culminates in a trio of fully connected layers, culminating in the output layer. These fully connected layers harmoniously collaborate to distill intricate abstractions from the hierarchical features unearthed by the preceding convolutional cascade. It is within these layers that the network synthesizes high-level representations suitable for classification tasks.

Performance Pedigree: VGG-19's exceptional classification prowess merits special mention. In the realm of image classification, it earned accolades by securing top positions in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) of 2014. The network's ability to discern objects within images and categorize them into a vast spectrum of classes is emblematic of its performance pedigree.

Transfer Learning Torchbearer: A noteworthy facet of VGG-19 is its role as a torchbearer of transfer learning. The pre-trained VGG-19 model, honed on vast datasets, serves as a versatile foundation for myriad computer vision applications. Its convolutional layers, in particular, are often employed as feature extractors,

expediting the training of custom classifiers.

Computational Considerations: However, it is imperative to acknowledge that VGG-19’s architectural richness comes at the cost of computational complexity. Its depth and dense convolutional layers demand substantial computational resources. Training and inference may necessitate hardware configurations commensurate with its architectural grandeur.

In conclusion, VGG-19, an opus of the Visual Geometry Group, emerges as a quintessential embodiment of architectural depth and elegance. Its remarkable consistency, classification acumen, and transfer learning capabilities have etched its name in the annals of computer vision. However, the computational demands it imposes underscore the need for suitable hardware infrastructure to unlock its full potential

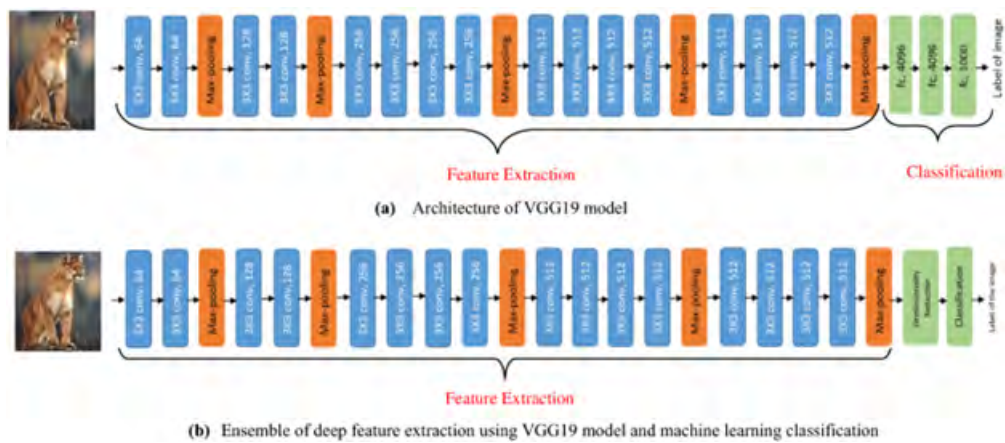


Figure 3.2: Architecture Of VGG19 Model

[47]

3.0.3 ResNet-50

ResNet-50, an abbreviation signifying "Residual Network with 50 layers," represents a pivotal milestone in the evolution of convolutional neural network (CNN) architectures.[23] This landmark innovation was ushered in through the seminal research titled "Deep Residual Learning for Image Recognition," authored by a distinguished consortium of researchers, including Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, in the year 2016. ResNet-50, building upon the foundation laid by its antecedent, ResNet-101, stands as an exemplar of architectural sophistication, boasting a neural network of substantial depth. This depth, in turn, bestows upon it unparalleled efficacy across an expansive gamut of computer vision tasks, prominently including image classification and object detection.

Within the purview of scholarly discourse, it becomes imperative to expound upon the salient attributes that constitute the ResNet-50 configuration with the following elucidation:

Input Layer: The ResNet-50 architecture adheres to convention, receiving RGB (Red, Green, Blue) images as input data, conforming rigorously to the industry-standard image size of 224x224 pixels.

Convolutional Layers: The initial strata of the architecture consist of conventional convolutional layers, augmented serially by critical constituents, notably batch normalization and rectified linear unit (ReLU) activation functions. These foundational layers orchestrate the extraction of elemental, low-level features intrinsic to the input images.

Building Blocks: ResNet-50 introduces the seminal concept of residual blocks, distinguished by their capacity to transcend the vanishing gradient dilemma—a profound innovation underpinning the training of exceptionally deep neural networks. Each of these enigmatic residual blocks encapsulates two or more layers of convolutional operations. Moreover, they encompass ancillary components, including batch normalization, the crucible of ReLU activations, and most significantly, the pivotal skip connections.

The conceptual core of this innovation lies in the notion that, instead of seeking to directly approximate the target output, the neural network is empowered to discern the differential—commonly referred to as the "residual"—between its present predictive approximation and the coveted target. This residual is harmoniously amalgamated with the input to engender a holistic, comprehensive representation. ResNet-50, bearing a totemic legacy, harmonizes numerous strata of these enigmatic residual blocks, imparting to it the profound depth characterizing this architecture.

Bottleneck Blocks: ResNet-50 avails itself of bottleneck building blocks—a testament to architectural ingenuity and computational efficiency. These ingenious modules commence their symphony with an inaugural 1x1 convolutional layer, conducting a dimensionality reduction symphony. Subsequently, they ascend to a crescendo of 3x3 and 1x1 convolutional layers, orchestrated meticulously for the extraction of discerning, high-level features.

The discerning inclusion of bottleneck blocks meticulously pares down the surfeit of model parameters and computational overhead while preserving the sacred tenets of architectural depth.

Skip Connections: Skip connections, known for their heroic role in vanquishing the vanishing gradient challenge, are meticulously orchestrated within ResNet-50. These conduits facilitate the unhindered flow of gradients, thereby ameliorating the tribulations associated with the training of exceedingly deep networks.

The brilliance of these skip connections is manifest as the output of a preceding stratum finds its harmonious fusion with the ensuing stratum through the simplicity of addition.

Pooling Layers: ResNet-50 gracefully ushers in the symphony of average pooling layers. These layers serve as cartographers of spatial information, adeptly downsizing the expansive topography of feature maps, all while preserving the saliency of distinguishing spatial features.

Fully Connected Layer: The apogee of the ResNet-50 architecture culminates in a global average pooling layer—a crucial intermediary in the trajectory leading to the ultimate destination—a fully connected layer. In the context of classification tasks, this fully connected layer converges toward the categorical finish line, bearing allegiance to the venerable softmax activation. This activation imparts its essence to multi-class classification endeavors with consummate grace.

For object detection tasks, the final layer embarks upon a transformative metamorphosis, morphing into a bespoke construct meticulously tailored to the exigencies of regression or precise localization.

Output Layer: The denouement of the ResNet-50 narrative is scripted by the opulent proclamation of an output layer. This layer mirrors the number of classes germane to the classification conundrum at hand. In the illustrious context of ImageNet classification, this layer hosts a pantheon of 1,000 neurons, reverberating in harmony with the 1,000 ImageNet classes with unwavering fidelity.

In resolute synthesis, ResNet-50, poised at the vanguard of architectural virtuosity, emerges as a paragon of neural network configuration. It harnesses the transcendental essence of skip connections and residual blocks to orchestrate the disciplined tutelage of profound neural networks. ResNet-50's architectural depth and structural ingenuity coalesce to render it an indomitable sentinel, an indispensable luminary, spanning a rich tapestry of computer vision mandates, foremost among them, image classification and object recognition

3.0.4 ResNet-101

ResNet-101, denoting "Residual Network with 101 layers," represents a significant advancement in deep convolutional neural network architecture. It emerged through the seminal work titled "Deep Residual Learning for Image Recognition" authored by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in 2016. ResNet-101 builds upon the foundation laid by the original ResNet-50 architecture, offering increased depth and, consequently, enhanced performance across various computer vision tasks, including image classification and object detection.

In a scholarly context, the salient characteristics of the ResNet-101 architecture can be elucidated as follows:

Input Layer: ResNet-101 conventionally takes RGB images as input, adhering to the common image size standard of 224x224 pixels.

Convolutional Layers: The initial layers comprise standard convolutional layers, sequentially followed by batch normalization and rectified linear unit (ReLU) activation functions. These layers serve the primary purpose of extracting rudimentary, low-level features from the input image.

Building Blocks: ResNet-101 introduces the concept of residual blocks, or skip connections. Each residual block encompasses two or more convolutional layers.

The foundational idea behind these blocks is their ability to learn residual functions, streamlining the training of exceedingly deep networks. Instead of directly learning the desired output, the network learns the discrepancy between the desired output and the current prediction—the residual. This residual is then incorporated by addition into the input of the block.

Within ResNet-101, numerous stacked residual blocks are employed. The structural components of these blocks encompass convolutional layers, batch normalization, ReLU activation, and the pivotal skip connections.

Bottleneck Blocks: ResNet-101 adopts bottleneck building blocks, notable for their computational efficiency. These blocks incorporate a 1x1 convolutional layer for dimensionality reduction, subsequently succeeded by 3x3 and 1x1 convolutional layers for feature extraction.

The adoption of bottleneck blocks effectively reduces the number of model parameters and computational overhead while preserving network depth.

Skip Connections: Skip connections play a pivotal role in mitigating the vanishing gradient problem and enhancing the training of exceedingly deep networks. In ResNet-101, these connections skip over one or more residual blocks, allowing for the seamless flow of gradients.

The output of a preceding layer is directly added to the output of a subsequent layer through these skip connections.

Pooling Layers: ResNet-101 implements average pooling layers to downsample feature maps spatially. This operation contributes to the extraction of salient spatial features.

Fully Connected Layer: The ultimate layers of ResNet-101 comprise a global average pooling layer, succeeded by a fully connected layer. For classification tasks, the fully connected layer typically employs softmax activation, facilitating multi-class classification.

For object detection tasks, this final layer may be substituted with layers tailored for regression or localization.

Output Layer: The output layer corresponds to the number of classes relevant to the classification task at hand. In the context of ImageNet classification, this layer comprises 1,000 neurons to accommodate the 1,000 ImageNet classes.

In summation, ResNet-101 stands as an emblematic deep neural network architecture that harnesses skip connections and residual blocks to train deep networks efficaciously. Its architectural depth and structural ingenuity render it a potent choice across an array of computer vision applications, particularly those centered on image classification and object recognition.

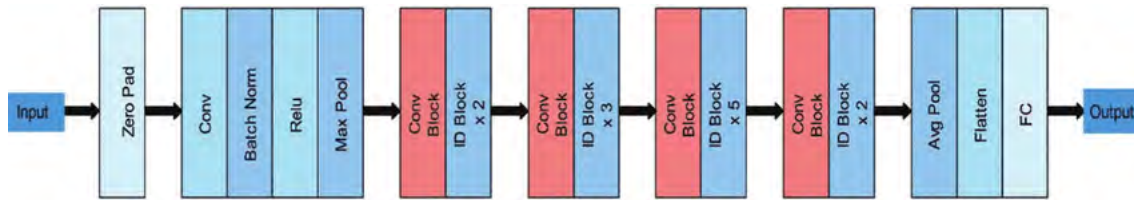


Figure 3.3: Architecture Of ResNet Model [46]

Few Examples Of Transformer Based Approach:

3.0.5 ViT Model

The Vision Transformer (ViT) is an innovative deep-learning model architecture that revolutionizes the field of computer vision. The Transformer design was first utilized for natural language processing activities, but Alexey Dosovitskiy and his team at Google Research introduced the ViT model in 2020 to efficiently process photos and carry out image identification tasks.[13] For many years now, the go-to method for computer vision problems has been convolutional neural networks (CNNs). Nevertheless, Vision Transformers present a viable alternative that has demonstrated exceptional performance, outperforming even the most advanced CNN models in terms of computational effectiveness and precision.

The fundamental concept underlying Vision Transformers involves partitioning the input image into patches of a predetermined size, followed by linear embedding of these patches into a vector space of reduced dimensions. Positional embeddings are added to retain spatial information about the relative positions of these patches within the image.[29] The patch embeddings, along with the positional embeddings, are processed through a series of transformer encoder layers, where the self-attention mechanism captures long-range dependencies between patches. A surprising development in the computer vision field is the Vision Transformer, which differs from more conventional methods like CNNs in a number of important ways. One of its key advantages is its scalability, allowing efficient handling of large images. This versatility makes ViT suitable for a wide range of tasks and datasets with varying sizes, making it an invaluable tool for diverse applications. Furthermore, Vision Transformer (ViT) models provide the advantage of end-to-end learning, eliminating the requirement for manual engineering of features. By learning directly relevant features from the data, this model becomes more adaptive and accurate, while also saving precious time and effort for researchers and engineers.[1] Additionally, the global context captured by the self-attention mechanism enables Vision Transformers to recognize complex patterns and dependencies within the images.[34] Their superior performance in many computer vision tasks, like image classification, object identification, and image segmentation, can be attributed in large part to their in-depth comprehension of the relevant concepts involved. In summary, the Vision Transformer model represents a significant breakthrough in computer vision. Its scalability, end-to-end learning capability, and ability to comprehend global context

make it a compelling choice for image recognition tasks, with demonstrated impressive real-world applications. As the field of artificial intelligence continues to evolve, Vision Transformers remain an exciting and promising area of research, poised to lead to even more remarkable advancements in the future.[25]

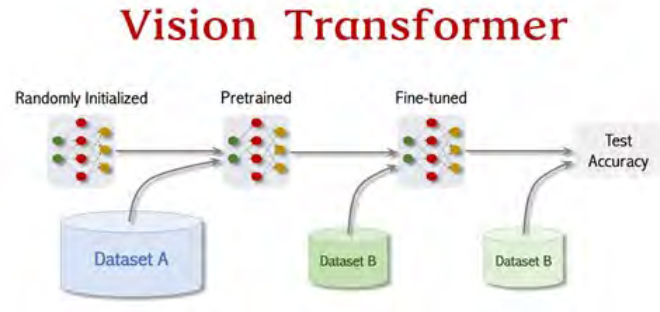


Figure 3.4: ViT Model
[41]

3.0.6 Swin Transformer

The Swin Transformer is a variant of the Vision Transformer (ViT) that has been specifically developed for the field of computer vision. The Swin Transformer and the conventional Vision Transformer both employ the foundational transformer architecture, which was originally designed for language problems but subsequently modified for computer vision purposes.[29] The Swin Transformer sets itself apart by substituting the conventional multi-head self-attention module with a shifted windows module within a Transformer block. The present study proposes a novel methodology that employs a hierarchical technique with shifted windows to effectively handle picture patches, hence offering notable benefits for processing high-resolution photos. As a result, the Swin Transformer demonstrates remarkable performance when applied to larger images, rendering it a very appropriate option for a range of computer vision applications.[10]

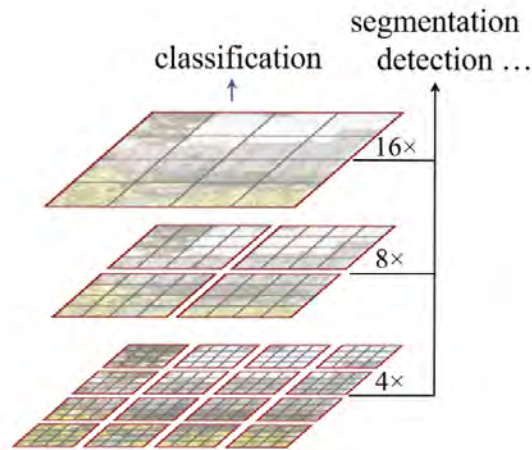


Figure 3.5: Swin Transformer
[45]

Workflow Diagram:

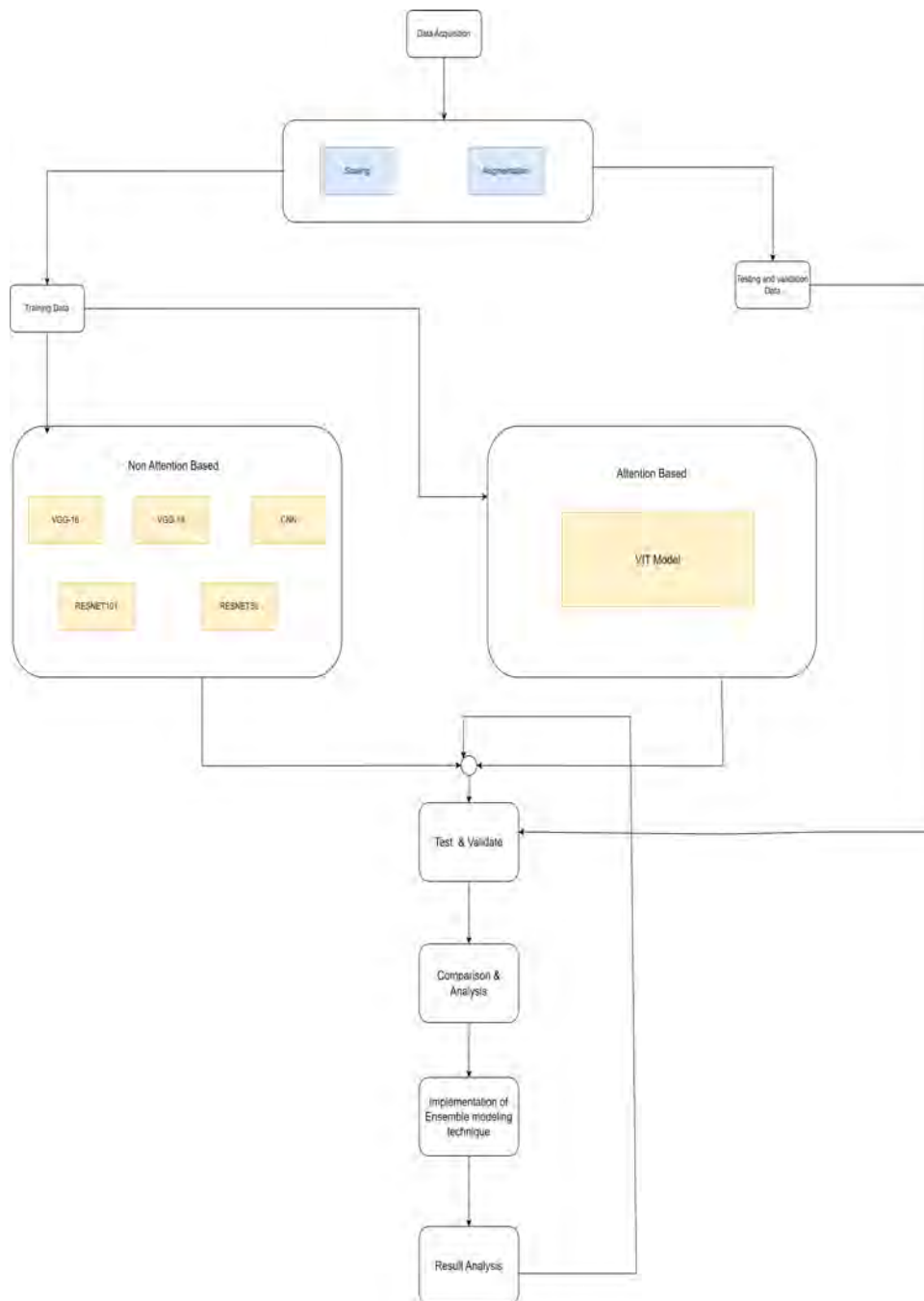


Figure 3.6: Workflow Diagram

3.1 Dataset Collection

For this particular model, we have decided to use the Colon Cancer Histopathological Images dataset. This dataset consists of 10,000 histopathological images, each of which falls into one of two categories.[2] JPEG is the file format used for each image, and its dimensions are exactly 768 by 768 pixels. Our dataset was available for download as a zipped file with a size of 1.85 gigabytes and is named LC25000.zip. After unzipping, the primary folder named “colon-image-sets” will have two subfolders within it. These subfolders are named colon benign tissue and colon adenocarcinoma.

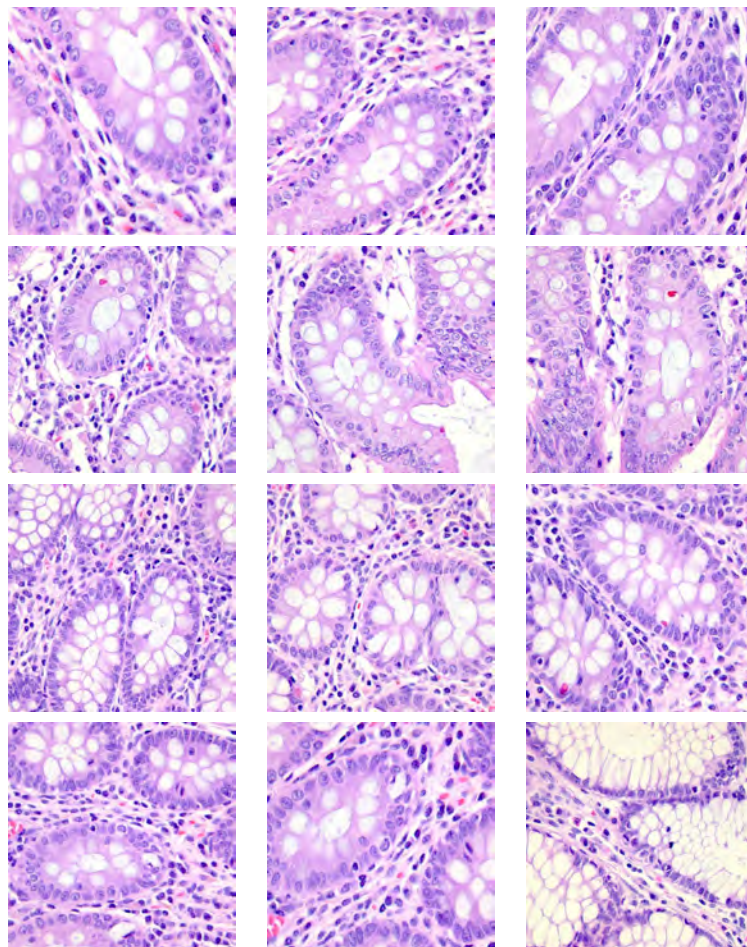


Figure 3.7: Colon Benign Tissue

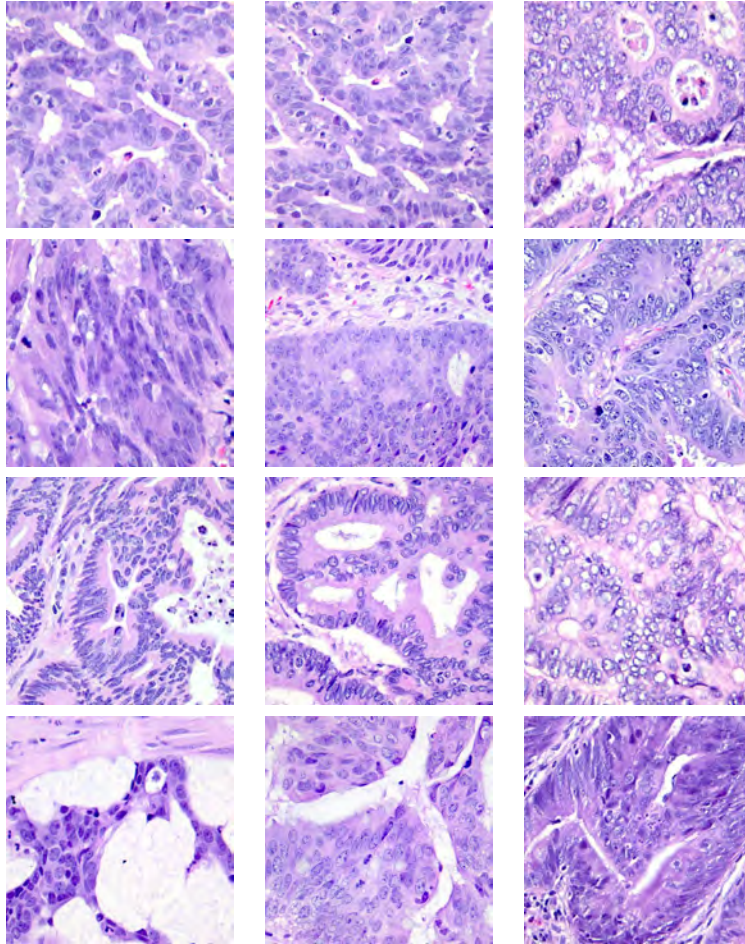


Figure 3.8: Colon Adenocarcinoma

The “colon-image-sets” subfolder has 5000 images of colon adenocarcinoma in the “colon-aca” sub folder and 5000 images of benign colonic tissues in the “colon-n” sub folder. The images were generated from an original sample that was HIPAA-compliant and validated.

3.2 Data Pre-Processing

The first stage of model development is data pre-processing. It reroutes the unprocessed data collected from many sources into valuable insights. The raw data often contains inconsistencies and outliers, as well as duplicates, therefore it must be cleaned up before it can be used. Data may prevent the model from producing the expected result and lead to a significant loss of knowledge.[6] In addition, the input of target form is standard in most Deep Learning algorithms. That form of the dataset is required for training to begin.[14] The dataset on colon cancer is documented using photos. One of our datasets comprises images for training, another contains images for validation, and the third includes images for testing. These datasets were generated using two subsets of the original dataset, which contained approximately 10,000 photographs. For the sake of both training and validating our model, we make use of each dataset on its own. After pre processing the data, we chose, where height=224, width =224 and channel= 3.

Data Segment	Parameters	Value
Training Data	Rescale	224
	Validation Split	0.8
Validation Data	Rescale	224
	Validation Split	0.1
Testing Data	Rescale	224
	Validation Split	0.1

Table 3.1: Table Of Pre-Processed Data

3.2.1 Data Labeling

We were able to break down the Colon Cancer data set into more than 3 distinct sections. There were .jpeg files with descriptive names inside each component.

3.2.2 Data Splitting

At the same time that the data was being pre-processed, it was also being segmented or partitioned. In an 8:1:1 ratio, we divided the total dataset into 80 percent for training, 10 percent for validation, and 10 percent for testing. In addition, validation was performed on 10 percent of the training data. As a result, about 8000 images have been utilized in the training process, 1000 images have been used for validation, and about 1000 images will be used in the testing phase.

Data Segments	Percentage	Total Images	Parameters	Values
Training	80% of total data	8000	Class Mode	Categorical
			Subset	Training
			Batch Size	32
Validation	10% of training data	1000	Class Mode	Categorical
			Subset	Validation
			Batch Size	32
Testing	10% of total data	1000	Class Mode	Categorical
			Subset	Testing
			Batch Size	32

Table 3.2: Table Of Data Splitting

3.2.3 Dataset Class Distribution

General Description: In the context of this investigation, the dataset of 10,000 histopathology pictures, serving as the basis for the identification of colorectal cancer. The dataset has a noteworthy attribute of possessing a balanced distribution, since it is evenly partitioned between two distinct groups or categories of tissue types.

Image Specifications: The images in the collection adhere to the JPEG format and were initially captured with dimensions of 768 pixels by 768 pixels. To adhere to the computing requirements of the project, the dimensions of the photographs have been adjusted to 224 pixels for both height and width. Furthermore, it is important to note that every image is comprised of three distinct color channels.

Taxonomy of Classes: The dataset has been systematically arranged into two main classes that are clinically significant, indicating two distinct variants of colon tissue.

1. Tissues of Benign Colonic Nature
2. Adenocarcinomatous Colon Tissues

Designation of Class Labels: In order to optimize the computational procedures, a distinct label has been allocated to each of the aforementioned classes to facilitate their identification in machine learning tasks.

1. Label for Class 0: Tissues of Benign Colonic Nature
2. Label for Class 1: Adenocarcinomatous Colon Tissues

Enumeration of Class Samples: A noteworthy feature of the dataset is the balanced allocation of image samples among both groups, as seen in the following:

1. Class 0 (Benign Colonic Tissues): The dataset has a total of 5,000 photos.
2. Class 1 (Adenocarcinomatous Colon Tissues): Additionally, the dataset has a total of 5,000 photos. The concept of proportional distribution across classes refers to the equitable allocation of resources or opportunities based on the relative size or importance of different classes or groups within a given system or context. This approach ensures that each class or A notable characteristic of the dataset is its equitable distribution. Specifically class 0 encompasses half of the entirety of the dataset. Class 1 also comprises the remaining 50%.

Segmentation for Model Training and Evaluation: The dataset has been meticulously partitioned into several subsets, each serving a specialized purpose, namely training, validation, and testing. This partitioning adheres to a distribution ratio of 8:1:1. As a result, the training dataset consists of around 4,000 photographs per class, whereas the validation and testing sets have approximately 500 images for each class.

Consequences for the Analytical Model: Due to the balanced allocation of classes, the researchers anticipate that there will be no need for employing specialist methodologies commonly used to tackle class imbalance, such as oversampling underrepresented classes or modifying loss functions.

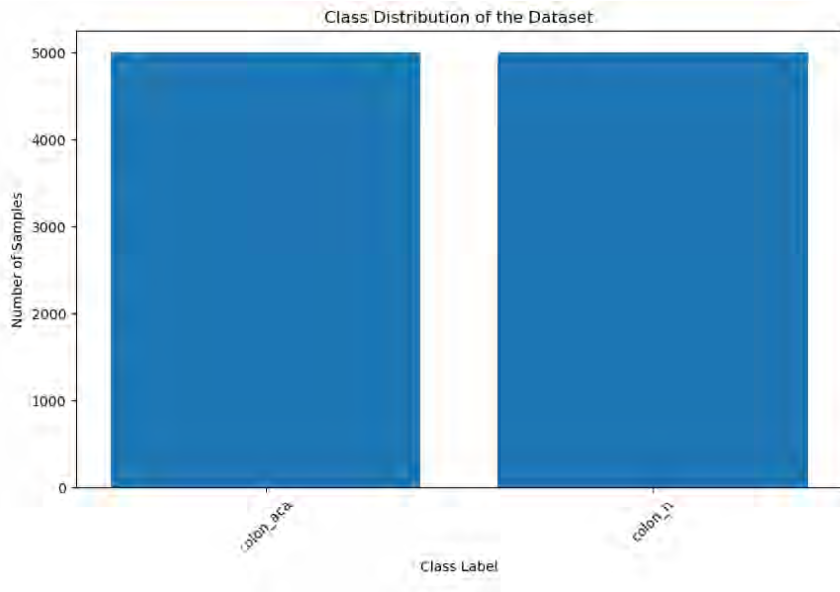


Figure 3.9: Dataset Class Distribution

3.2.4 Proposed ViT model

ViT (Vision Transformer) is a type of deep learning model that is designed to process visual data, such as images and videos. It is based on the transformer architecture, which was originally developed for natural language processing tasks. ViT models are trained to process image data by breaking the image into a grid of non-overlapping patches, and then treating each patch as a token in a sequence. These tokens are then processed by the transformer architecture, which learns to extract meaningful features from the image data. ViT models have been shown to achieve state-of-the-art performance on a wide range of image classification and object detection tasks, and have been used in many real-world applications.

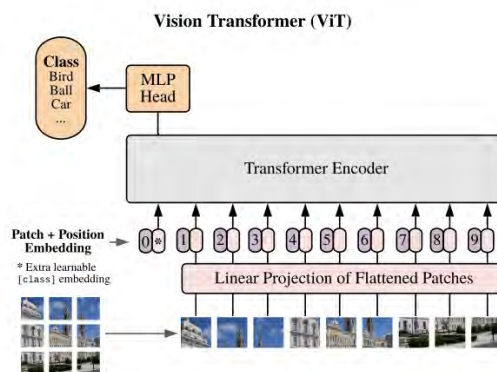


Figure 3.10: Vision Transformer(Vision Transformer architecture — main blocks. First, image is split into fixed-size patches and flatten. Second, position embeddings is added, and resulting sequence of vectors is forwarded to a standard Transformer encoder.)

Transformer Encoder

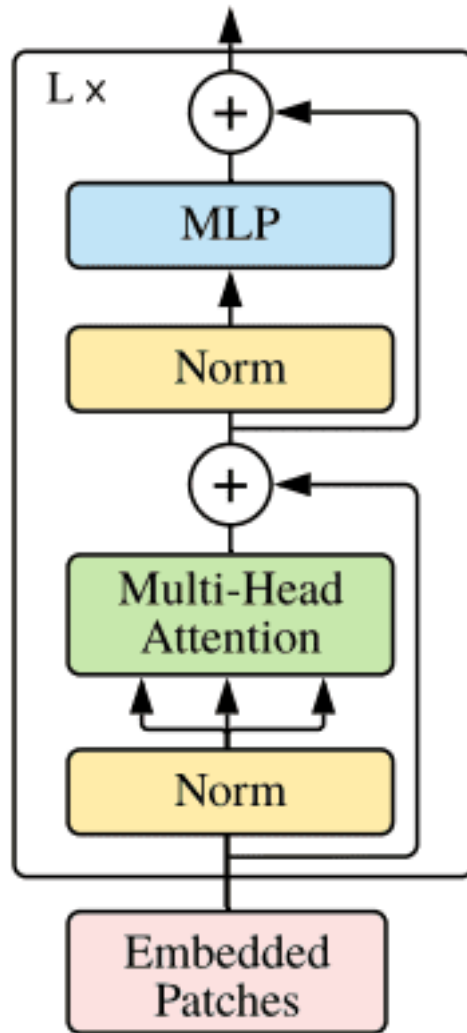


Figure 3.11: Transformer Encoder
[48]

3.3 ViT Model Layers

The Vision Transformer (ViT) model is composed of multiple layers, including Token Embedding, Multi-Head Attention, Position-wise Feed forward Network, Layer Normalization, Classifier etc.

3.3.1 Token Embedding

Patching and Embedding: Token Embedding commences by slicing the input image into fixed-size patches, transforming visual information into a sequence of tokens. These patches are then projected into a high-dimensional space using a linear transformation.[17] This transformation acts as a bridge between the spatial domain of images and the vector space that the Transformer operates. **Positional Encoding:** In addition to the patch embeddings, a positional encoding is added to provide information about the spatial arrangement of patches.[5] Unlike text, where sequence order conveys inherent meaning, images require this added information to maintain spatial coherence.

3.3.2 Multi-Head Attention

Attention Mechanism: The utilization of Multi-Head Attention enables the model to selectively concentrate on various spatial relationships present within an image. It employs multiple attention heads to simultaneously explore various types and scales of relationships between patches. This technique facilitates the model in evaluating the significance of various regions, hence including a wide range of visual characteristics.[44]

Parallel Processing: By employing multiple heads, the attention mechanism can be parallelized, leading to substantial efficiency in training and inference. This aspect allows the ViT to scale effectively with both the size of the model and the input data.

3.3.3 Position-wise Feed forward Network

Non-linear Transformation: Positional Feed-Forward Network (FFN) applies non-linear transformations at each position of the sequence. By using activation functions like ReLU (Rectified Linear Unit), the FFN introduces non-linearities, allowing the model to learn more complex and abstract features from the data.

Information Refinement: By sequentially applying linear transformations and non-linear activation functions, the FFN refines and integrates information passed from the attention layers, further abstracting the understanding of the image.

3.3.4 Layer Normalization

Stabilizing Learning: Layer Normalization is crucial for maintaining stability during the training process. By normalizing activations across features, it helps in reducing internal covariate shift, allowing for higher learning rates and accelerating convergence.

Regularization Effect: Besides improving optimization, Layer Normalization also has a slight regularization effect, making the model more robust to overfitting, especially when training on smaller datasets.

3.3.5 Classifier

High-level Abstraction to Prediction: The Classifier serves as the decision-making component of the ViT, converting high-level abstractions into concrete predictions, such as class labels or bounding boxes. **Fine-tuning for Specific Tasks:** Depending on the particular computer vision task, the classifier’s architecture may vary. Because it is so flexible, the ViT can be fine-tuned for a wide range of uses, including image classification, object detection, and also semantic segmentation, among others.

Vision Transformer (ViT) presents an integration of effectively coordinated layers and mechanisms, each contributing uniquely to the model’s ability to understand and interpret visual data. From spatially encoding the input image through Token Embedding to recognizing intricate patterns via Multi-Head Attention, followed by abstracting and refining information through Position-wise Feed Forward Networks, and stabilizing the learning process with Layer Normalization, the ViT culminates in the Classifier that translates these complex representations into actionable insights. Its innovative architecture offers a glimpse into the future of computer vision, where the lines between traditional image-processing models and text-based Transformer models blur. The synergistic operation of these components not only attests to the model’s efficacy across various visual works but also lays the groundwork for future exploration and adaptations in the continuously evolving field of artificial intelligence.

3.4 ViT Model Mathematical Operations

3.4.1 Linear Projection:

If we consider ‘A’ to be the input image divided into patches, and ‘lw’ to be the learnable weight matrix for the linear projection then we can say that $\text{patch_embedding} = A \cdot lw$

3.4.2 Positional Encoding:

Another important mathematical operation. Here if we consider ‘position’ to be the position of the token in the sequence and ‘dim’ to be the dimension of embedding then following operations happen,

$$\text{PositionalEmbedding}(\text{position}, 2i) = \sin\left(\frac{\text{position}}{10000^{\frac{2i}{\text{dim}}}}\right)$$

$$\text{PositionalEmbedding}(\text{position}, 2i + 1) = \cos\left(\frac{\text{position}}{10000^{\frac{2i}{\text{dim}}}}\right)$$

Here PositionalEmbedding represents the two components of the positional encoding (even position and odd position)

3.4.3 Self-Attention Mechanism:

The self-attention mechanism involves three linear transformations: Query, Key, and Value. Let Q_{ry} , k , and V_l be the learnable weight matrices for the query, key, and value projections, respectively. Also, let dim_k be the dimension of the query/key vectors.[36] The dot-product attention scores A for each token is computed as,

$$A = \text{softmax}\left(\frac{(\text{Patch_embedding} \cdot Q_{ry}) \cdot (\text{Patch_embedding} \cdot k)^T}{\sqrt{\text{dim_k}}}\right)$$

Here, $(\text{Patch_embedding} \cdot Q_{ry})$ and $(\text{Patch_embedding} \cdot k)^T$ represent the query and key vectors for each token in the sequence.

3.4.4 Softmax:

In image classification tasks, global representation of the image (often associated with the “CLS” token) is fed into a fully connected layer, followed by a SoftMax activation. Let W_{cls} be the learnable weight matrix and b_{cls} be the bias vector for the classification head. The SoftMax output probabilities for different classes can be computed as:

$$\text{Probabilities} = \text{softmax}((\text{CLS_token_representation} \cdot W_{cls}) + b_{cls})$$

3.4.5 Loss Function:

For supervised learning, the commonly used loss function is categorical cross-entropy. Let x_{t} be the ground truth one-hot encoded label vector, and x_{prd} be the predicted probability vector for the classes. The categorical cross-entropy loss can be calculated as:

$$Loss = - \sum (x_{\text{true}} \cdot \log(x_{\text{prd}}))$$

The Vision Transformer (ViT) model, which is a type of transformer-based architecture, has gained popularity in computer vision tasks. The ViT model does not use conventional activation functions like ReLU (Rectified Linear Unit) as commonly seen in traditional convolutional neural networks (CNNs). Instead, the ViT model uses the GELU (Gaussian Error Linear Unit) activation function. GELU Activation Function: The GELU activation function is a smooth approximation of the rectifier (ReLU) function and is defined as follows: The GELU activation function helps to address the vanishing gradient problem that can occur during the training of deep

neural networks. It has been shown to be effective in transformer-based models like ViT, as well as other architectures like BERT (Bidirectional Encoder Representations from Transformers).[21] The GELU activation function is widely used in transformer-based models and has shown to be effective in various natural language processing (NLP) and computer vision tasks.

Function:

$$\text{GELU}(x) = 0.5x \cdot \left(1 + \tanh \left(\frac{\sqrt{2}}{\sqrt{\pi}} \cdot (x + 0.044715 \cdot x^3) \right) \right)$$

$$\text{GELU}(x) = 0.5x \left(1 + \tanh \left(\sqrt{2/\pi}(x + 0.044715x^3) \right) \right)$$

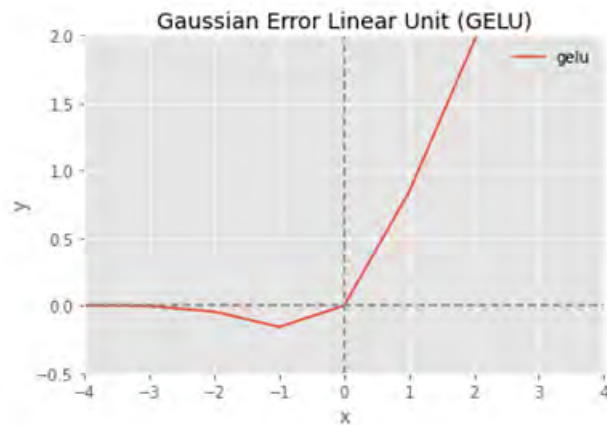


Figure 3.12: GELU Activation Function Through Graph

Chapter 4

Implementation Of Results

In this particular part we will be focusing on the results that have been found using different models in case of detecting Colorectal cancer detection. We will also be looking to find the most accurate result using different parameters for each models. Later the results from different models will be compared with each other so that we can find the optimum result.

4.1 Findings

Generally Deep learning / Attention based mechanism focuses on finding the optimum result to maximize efficiency. Accuracy, Precision, Recall, f1- score ,Confusion matrix etc are the parameters based on which we can scan whether a model is good enough to server the best interest of different purposes.

4.1.1 Colorectal Cancer Analysis With VGG-16

In the research, we have found that VGG-16 achieved an accuracy of 78.30%, with a validation loss of 0.6868. In terms of precision and recall, the model attained a precision of 0.78 and a recall of 0.79. Additionally, the F1 score was calculated to be 0.78. Furthermore, a confusion matrix has been included for this selected model.

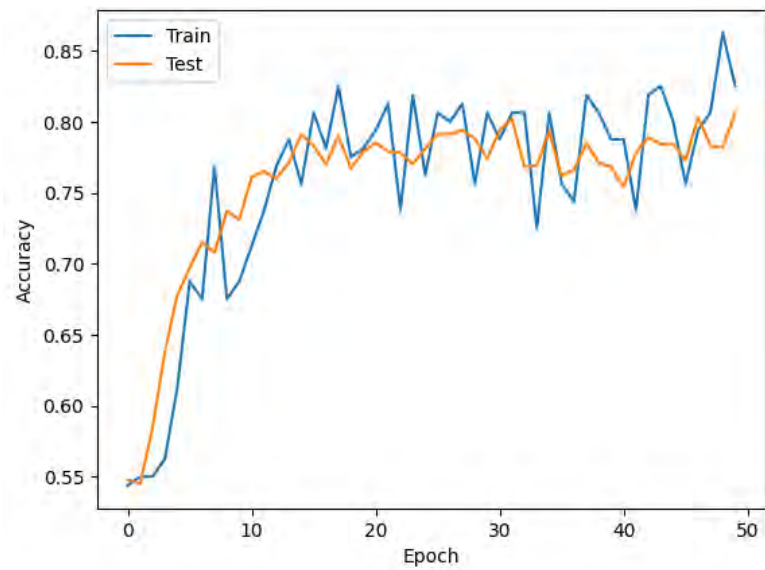


Figure 4.1: Accuracy Curve of Vgg-16

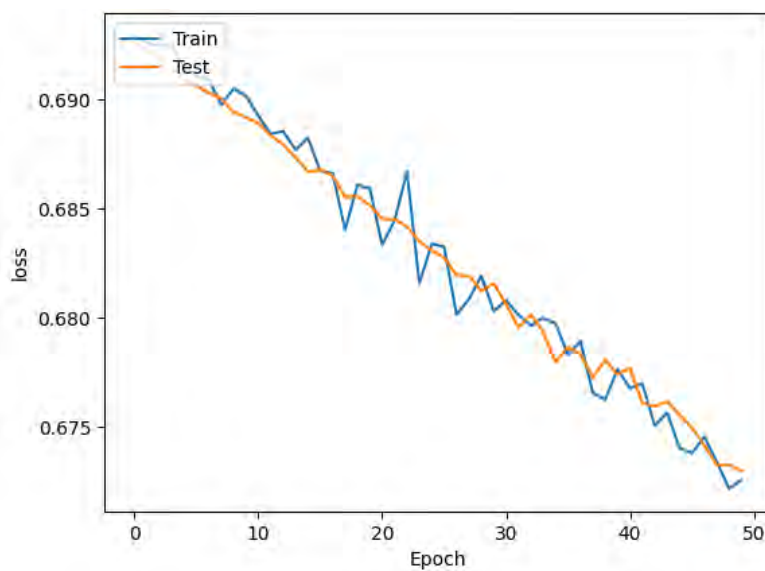


Figure 4.2: Validation Loss Curve of Vgg-16

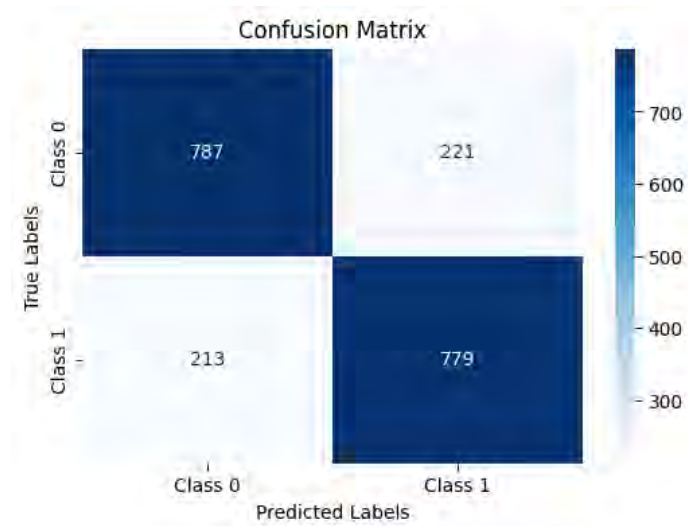


Figure 4.3: Confusion Matrix of Vgg-16

4.1.2 Colorectal Cancer Analysis With VGG-19

In the research, we have found that VGG-19 achieved an accuracy of 74.80%, with a validation loss of 0.6892. In terms of precision and recall, the model attained a precision of 0.74 and a recall of 0.75. Additionally, the F1 score was calculated to be 0.75. Furthermore, a confusion matrix has been included for this selected model.

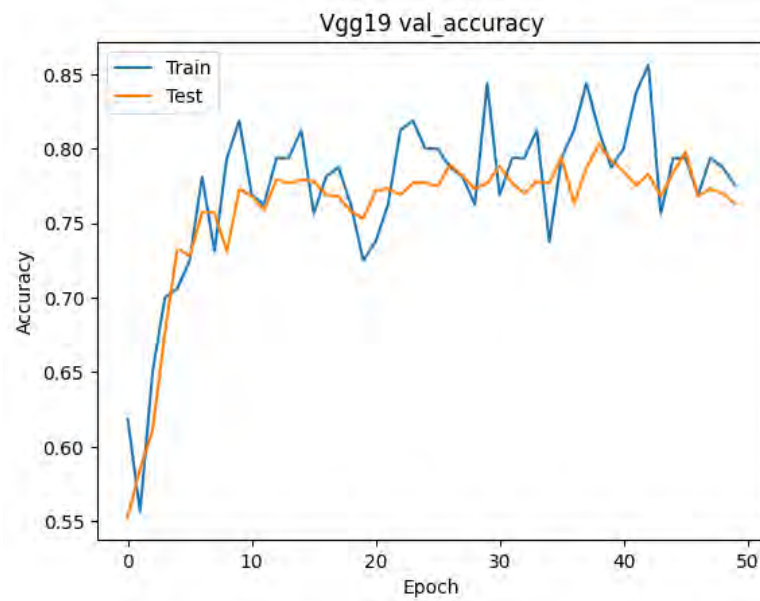


Figure 4.4: Accuracy Curve of Vgg-19

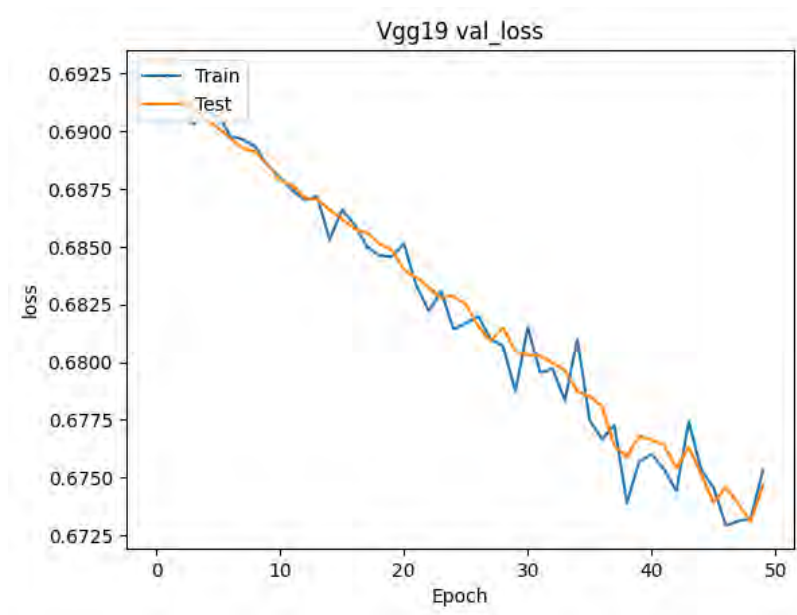


Figure 4.5: Validation Loss Curve of Vgg-19

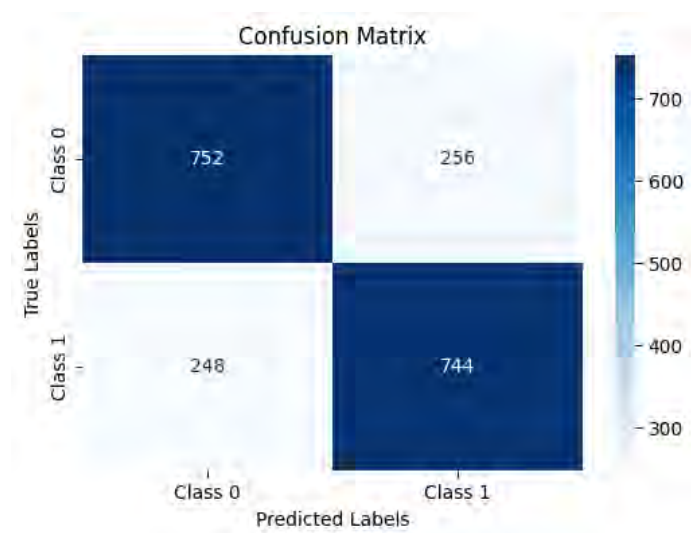


Figure 4.6: Confusion Matrix of Vgg-19

4.1.3 Colorectal Cancer Analysis With ResNet-50

In the research, we have found that ResNet50 achieved an accuracy of 95.5%, with a validation loss of 1.2646. In terms of precision and recall, the model attained a precision of 0.95 and a recall of 0.96. Additionally, the F1 score was calculated to be 0.95. Furthermore, a confusion matrix has been included for this selected model.

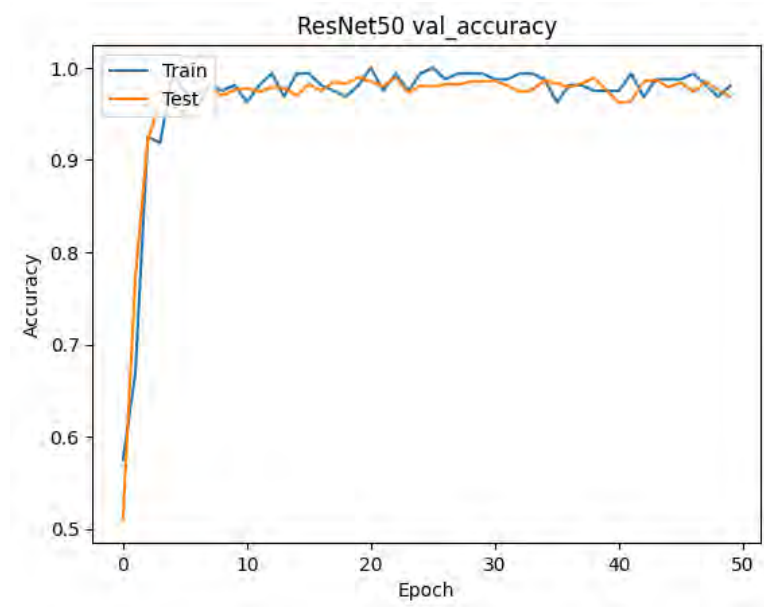


Figure 4.7: Accuracy Curve of ResNet-50

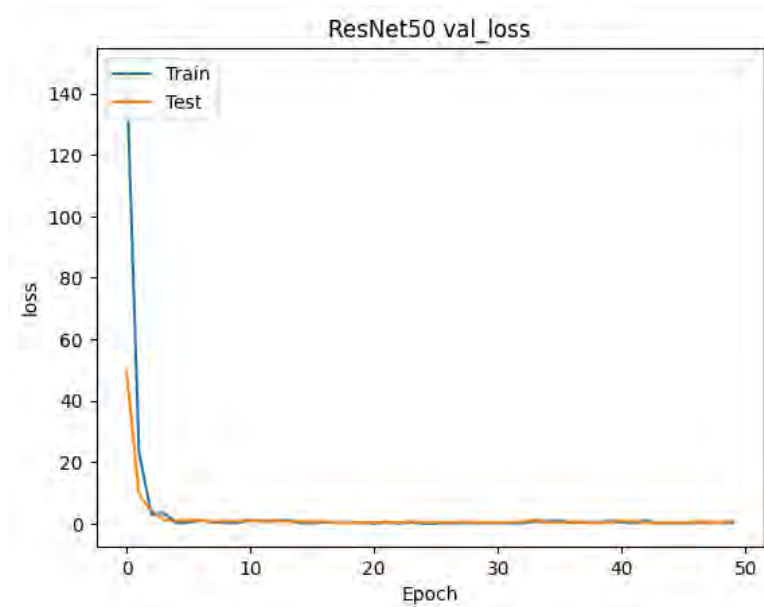


Figure 4.8: Validation Loss Curve of ResNet-50

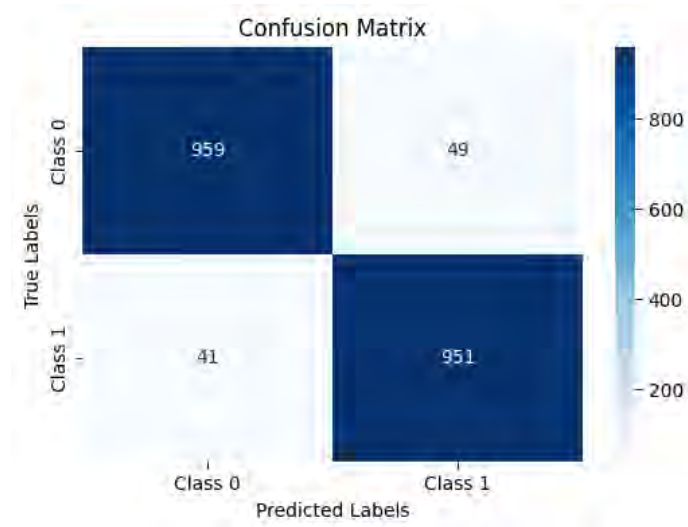


Figure 4.9: Confusion Matrix of ResNet-50

4.1.4 Colorectal Cancer Analysis With ResNet-101

In the research, we have found that ResNet101 achieved an accuracy of 97.9%, with a validation loss of 0.8477. In terms of precision and recall, the model attained a precision of 0.97 and a recall of 0.98. Additionally, the F1 score was calculated to be 0.98. Furthermore, a confusion matrix has been included for this selected model.

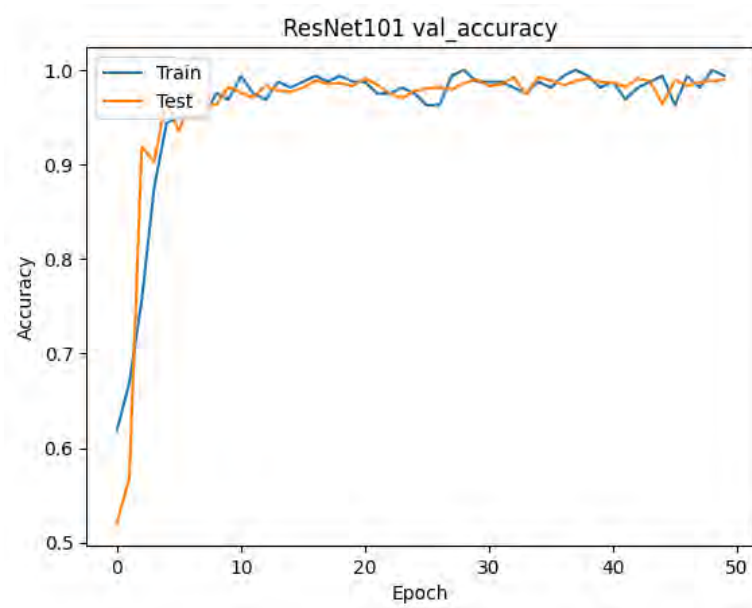


Figure 4.10: Accuracy Curve of ResNet-101

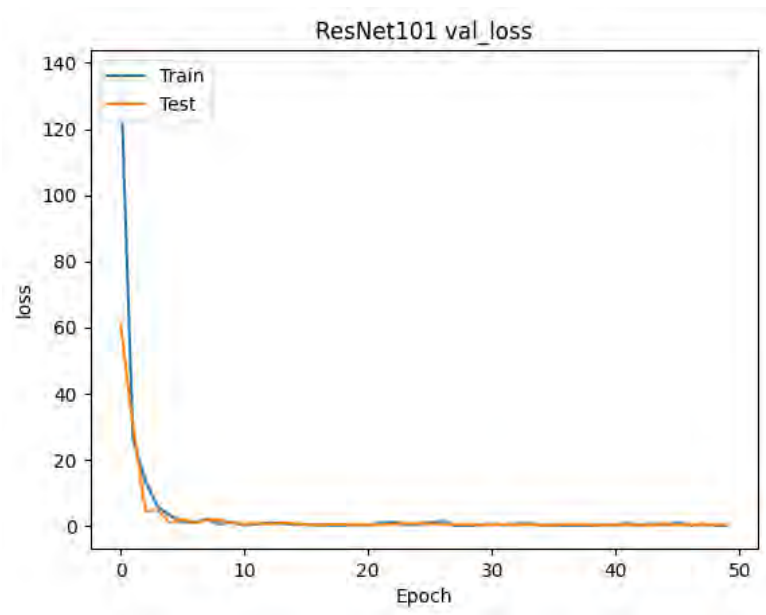


Figure 4.11: Validation Loss Curve of ResNet-101

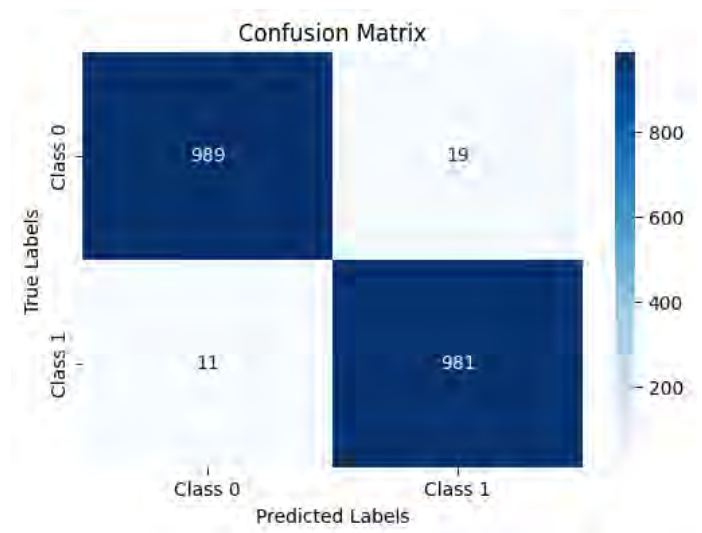


Figure 4.12: Confusion Matrix of ResNet-101

4.1.5 Colorectal Cancer Analysis With ViT-16

In the research we have found that ViT-16 has a val-accuracy of 99.04%, val-loss of 0.0201 which deals with accuracy. In the meantime precision 0.99, recall 0.99, f1-score 0.99 has also been found out. Moreover, Confusion matrix has also been added to for this selected model.

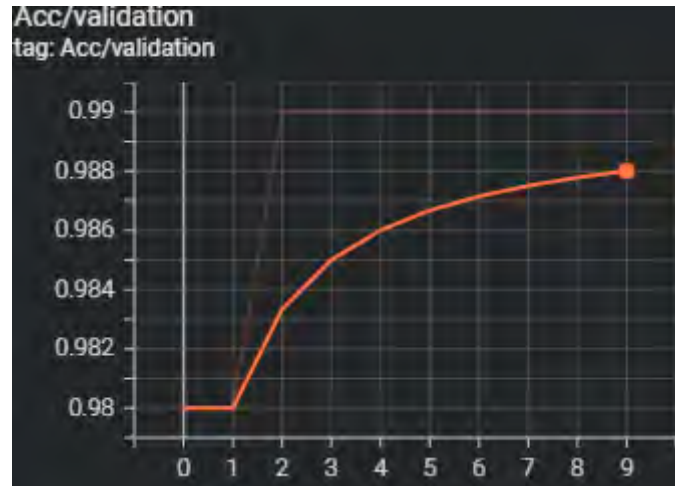


Figure 4.13: Accuracy Curve of ViT-16

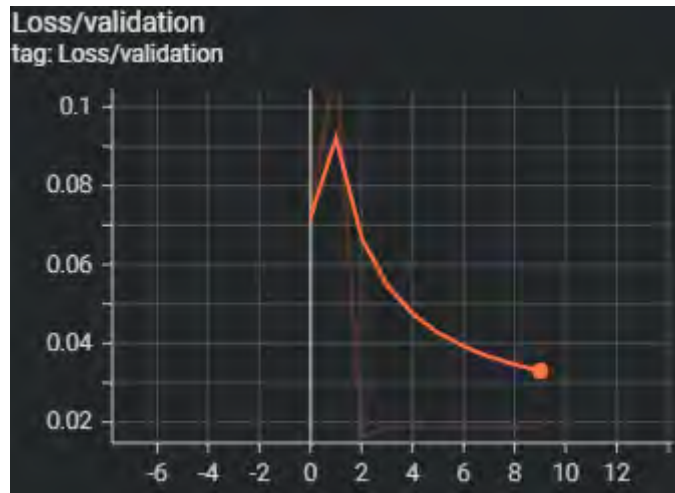


Figure 4.14: Validation Loss Curve of ViT-16

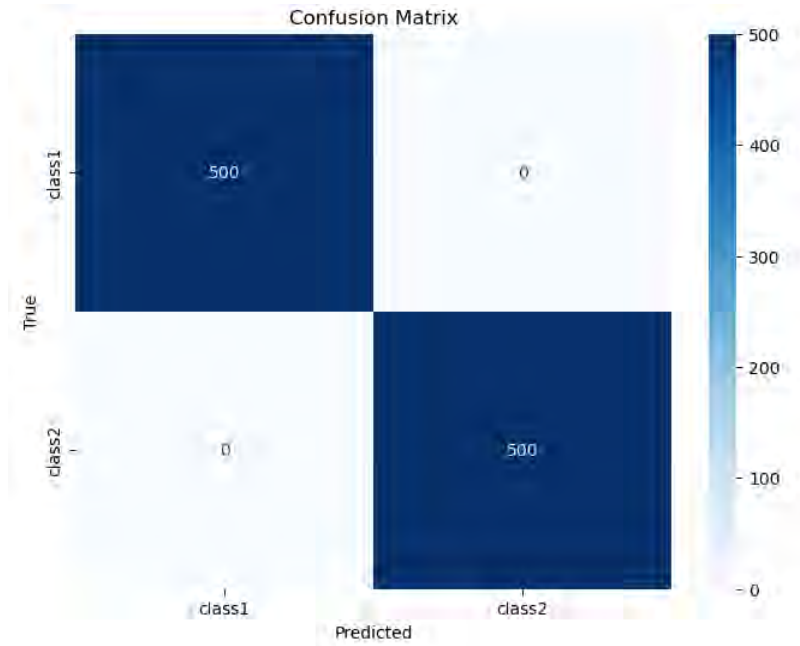


Figure 4.15: Confusion Matrix of ViT-16

4.2 Result Analysis

Architecture	val_accuracy (%)	val_loss	precision	recall	f1-score
VGG16	78.30	0.6868	0.78	0.79	0.78
VGG19	74.80	0.6892	0.74	0.75	0.75
ResNet-50	95.05	1.2646	0.95	0.96	0.95
ResNet-101	97.90	0.8477	0.97	0.98	0.98
ViT-16	99.04	0.0201	0.99	0.99	0.99

Table 4.1: Comparison Between Used Models

Comparison Between Used Models Based On Accuracy: First of all, we can see that VGG16 has an accuracy of 78.30%, while VGG19 has an accuracy of 74.80%. Moving onto the next three models, we can see a significant change compared to the first two models. In the case of ResNet-50, it achieves an accuracy of 95.50%, representing almost a 17% increase compared to the accuracy of the first two models. The following two models, ResNet-101 and ViT-16, have an accuracy of 97.90% and 99.04%, respectively. ViT-16 clearly outperforms all the other models with the highest accuracy, making it the preferred choice for tasks like colorectal cancer detection.

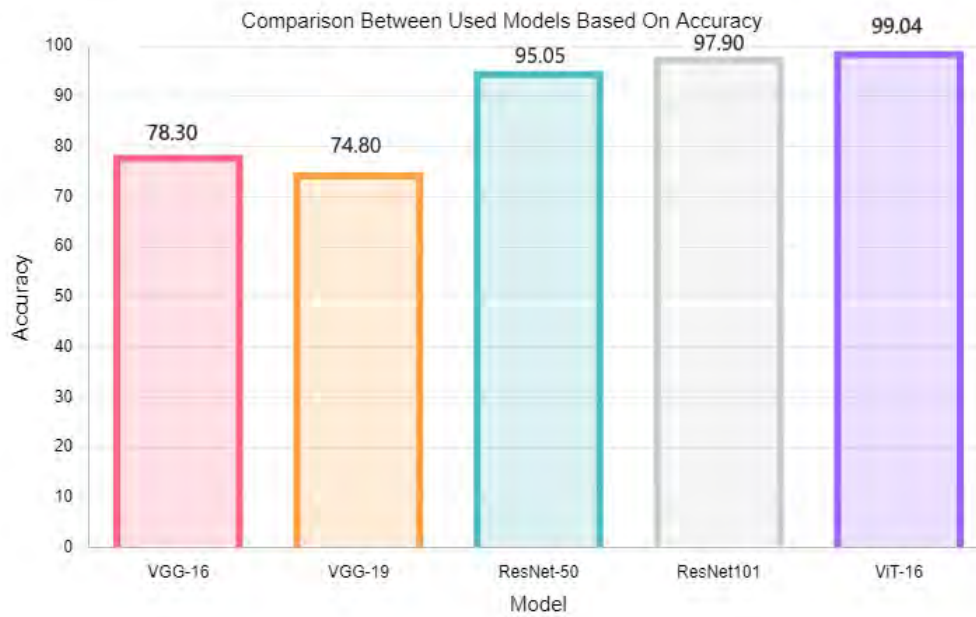


Figure 4.16: Comparison Between Used Models Based On Accuracy

If we draw a comparison between CNNs and Vision Transformers (ViT), we might be able to understand why Vision Transformers tend to have better accuracy values. Several reasons could explain this phenomenon, such as parameter efficiency, scalability, fewer architectural heuristics, and transfer learning capabilities. ViT-16, as a member of the Vision Transformer family, is a strong candidate for transfer learning tasks, especially when dealing with limited data, thanks to its accuracy of 99.04%. In terms of scalability, ViT-16 stands out as it can efficiently handle both small and large images by adjusting the number of attention heads and layers, whereas the other models, including VGG16, VGG19, ResNet-50, and ResNet-101, are not as versatile in this regard. Another noteworthy point is that ViT-16 tends to be more parameter-efficient compared to CNNs like VGG and ResNet, as it achieves superior performance with fewer parameters. In conclusion, when considering value accuracy, ViT-16 emerges as the optimal choice among the models listed, surpassing VGG16, VGG19, ResNet-50, and ResNet-101.

Comparison Between Used Models Based On Validation Loss: Secondly, let's examine the performance of VGG16, VGG19, ResNet50, ResNet101, and ViT-16 in terms of val_loss. ViT-16 continues to stand out with the lowest val_loss of 0.0201, indicating its remarkable performance in accuracy and quick convergence for colorectal cancer detection.

Comparison Between Used Models Based On Validation Loss

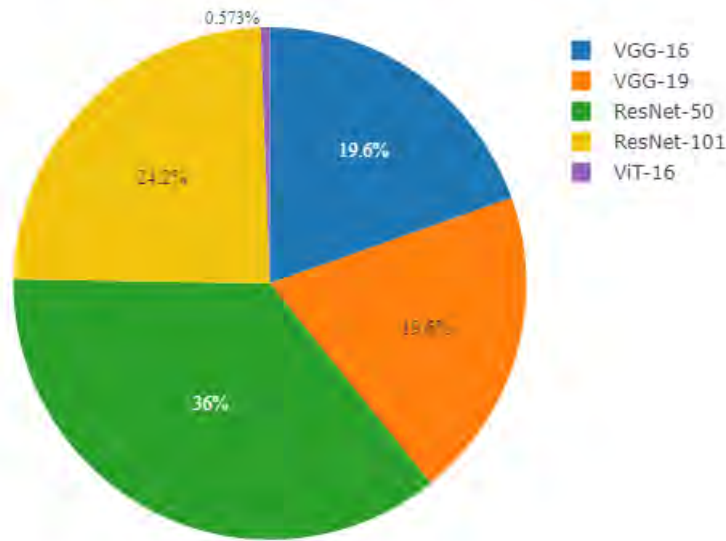


Figure 4.17: Comparison Between Used Models Based On Validation Loss

Following closely, VGG16 now holds the second spot with a val_loss of 0.6868, showcasing competitive results. VGG19 is right behind with a val_loss of 0.6892, making it a viable alternative.

In contrast, ResNet101 displays a higher val_loss of 0.8477, suggesting that it may need more training or might not capture complex data patterns as effectively.

ResNet50, with an accuracy value of 1.2646, presents different metrics for evaluation, and its performance should be analyzed more comprehensively.

In summary, ViT-16 remains the top performer with the lowest val_loss, while VGG16 and VGG19 offer competitive options. ResNet101 and ResNet50 have their unique characteristics, requiring further scrutiny for the given colorectal cancer detection task.

Comparison Between Used Models Based On Precision: Thirdly, Precision is an important metric in classification tasks that measures the ability of a model to make correct positive predictions. The precision values for these five models are: VGG16=0.78, VGG19=0.74, ResNet50=0.95, ResNet101=0.97, and ViT-16=0.99.

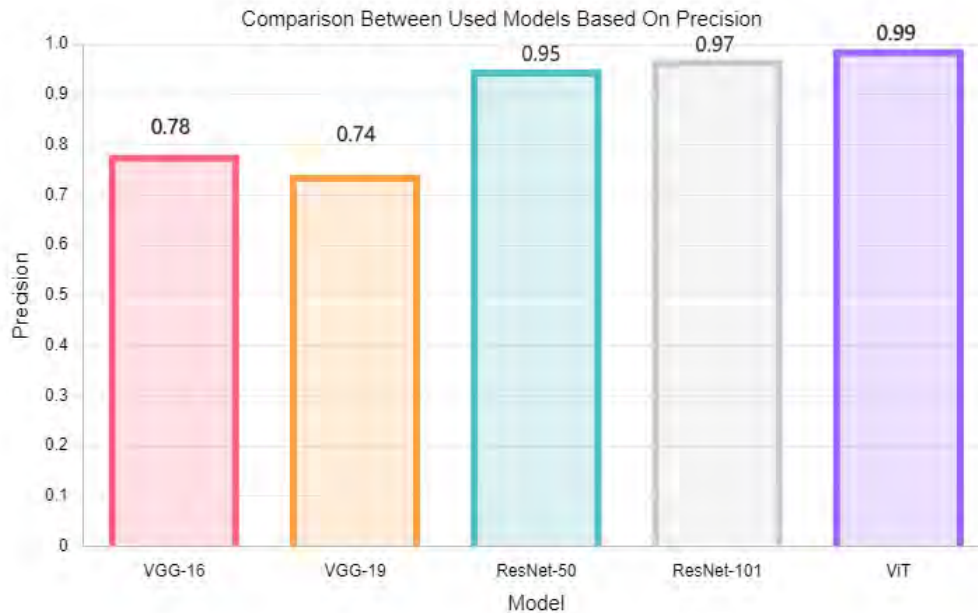


Figure 4.18: Comparison Between Used Models Based On Precision

First of all, ViT-16 consistently achieves the highest precision among all the models, with a precision value of 0.99, indicating that it has a strong ability to correctly classify positive cases. This suggests that ViT-16 is well-suited for tasks where precision is crucial, such as medical diagnoses. Following closely behind are ResNet-50 and ResNet-101, with precision values of 0.95 and 0.97, respectively. These models also exhibit high precision, slightly lower than ViT-16 but still significantly higher than VGG16 and VGG19. They are reliable for positive class predictions, which makes them suitable for applications where precision is important.

However, VGG16 and VGG19 still have the lowest precision among the group, with precision values of 0.78 and 0.74, respectively. While they are capable of making positive predictions, they are not as precise as the other models. This suggests that they might have more false positives in their predictions. To summarize, ViT-16

achieves the highest precision, followed by ResNet-101 and ResNet-50. VGG16 and VGG19 have lower precision values in comparison. Therefore, if precision is a critical metric for a colorectal cancer detection task, ViT-16 is the preferred model, followed by the ResNet models.

Comparison Between Used Models Based On Recall: Recall, a crucial metric in classification tasks, assesses a model’s ability to correctly identify all relevant instances of positive cases. Looking at the specific recall values for five different models, we can see distinct performance disparities. VGG16 records a recall of 0.79, while VGG19 follows closely with 0.75. In contrast, ResNet-50 achieves a significantly higher recall of 0.96, and ResNet-101 surpasses it with an even more impressive 0.98. Finally, ViT-16 stands out with an exceptional recall of 0.99. These values paint a clear picture of the models’ efficacy in capturing positive instances, with ViT-16 and the ResNet models taking the lead.

When comparing the performance of VGG16, VGG19, ResNet-50, ResNet-101, and ViT-16 in terms of recall:

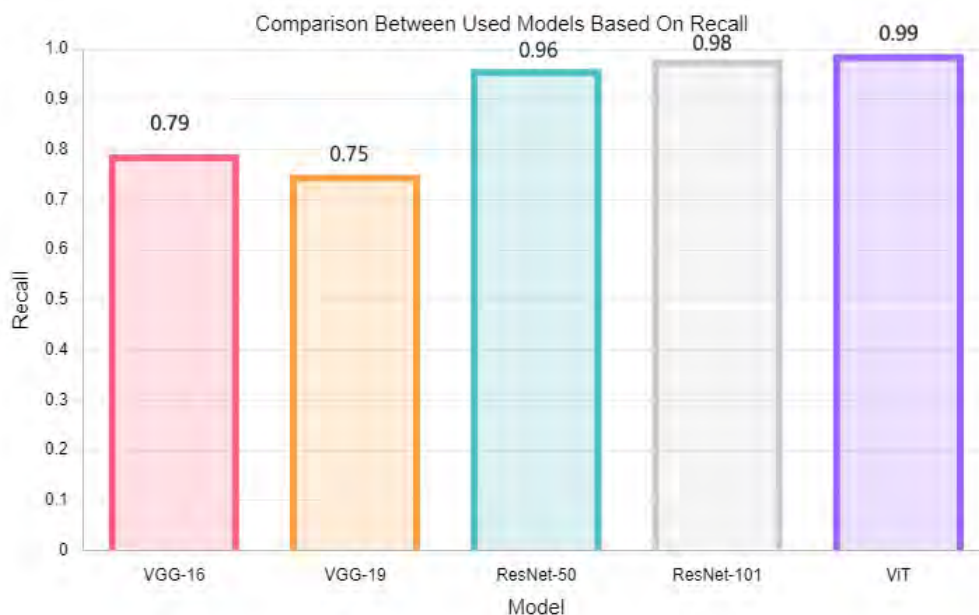


Figure 4.19: Comparison Between Used Models Based On Recall

ViT-16 consistently emerges as the top performer in terms of recall. Its impressive recall score of 0.99 underscores its robustness in accurately identifying positive instances. ViT-16 is particularly well-suited for tasks where a high recall rate is critical, such as medical diagnoses, where missing positive cases can have dire consequences. On the other hand, ResNet-50 and ResNet-101 also exhibit strong recall rates, albeit slightly trailing ViT-16. These models excel in correctly identifying positive cases, making them excellent choices for applications where recall is a pivotal performance metric.

In contrast, VGG16 and VGG19 lag behind the other models, displaying lower recall values. While they are still competent at detecting positive cases, their ability to

capture all relevant positive instances falls short when compared to the ResNet and ViT models.

In summary, ViT-16 leads the pack with the highest recall, followed by ResNet-101 and ResNet-50. VGG16 and VGG19 have lower recall values in comparison. Therefore, if maximizing recall is crucial, especially in tasks like colorectal cancer detection, ViT-16 emerges as the preferred model, with ResNet models as strong contenders, and VGG models as less favorable options.

Comparison Between Used Models Based On F-1 Score: When considering classification tasks that demand a balance between precision and recall, or when dealing with datasets where class imbalances are a concern, the F1 Score becomes a crucial metric. It offers a concise summary of a model's overall performance in such tasks. Looking at Table 4.1, we can see distinct F1 Score values for five different models: VGG16 with an F1 Score of 0.78, VGG19 at 0.75, ResNet-50 at 0.95, ResNet-101 at 0.98, and ViT-16 leading the pack at 0.99. The transition from VGG to Convolutional Neural Networks (CNNs) reveals a significant shift in performance. ViT-16 and ResNet models (50 and 101) show similar F1 Scores.

Now, let's examine the comparative performance of VGG16, VGG19, ResNet-50, ResNet-101, and ViT-16 concerning the F1 Score:

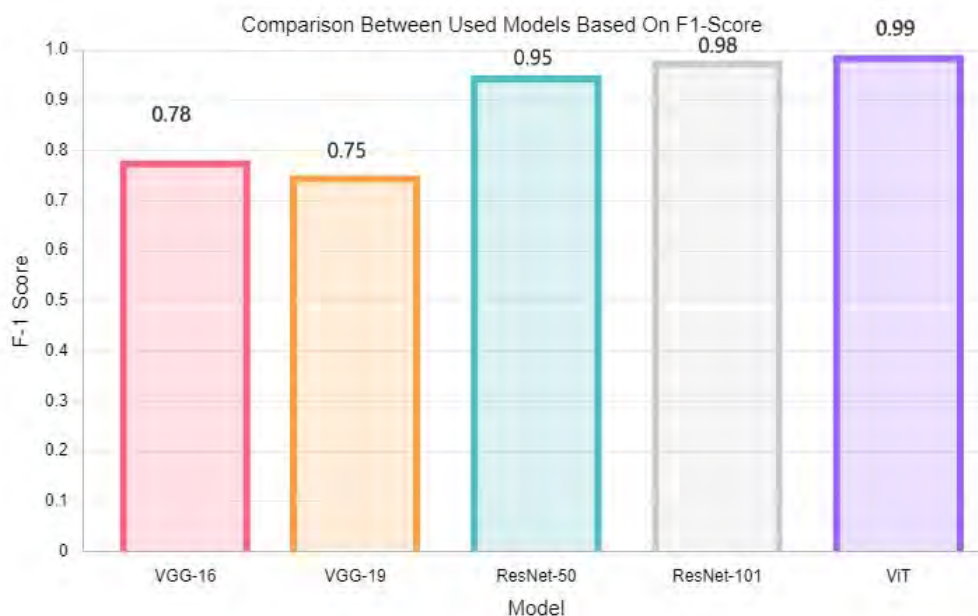


Figure 4.20: Comparison Between Used Models Based On F1-score

VGG16 and VGG19, in contrast to the other models, exhibit lower F1 Scores, suggesting a trade-off between precision and recall. While they perform reasonably well, they might not achieve the same balance as the ResNet and ViT models. Moving on to ResNet-50 and ResNet-101, these models also display strong F1 Scores, slightly trailing ViT-16 but significantly surpassing VGG16 and VGG19. They strike an impressive balance between precision and recall, making them suitable for applications where the F1 Score plays a critical role. However, ViT-16 consistently attains the highest F1 Score among all models, emphasizing its robust balance between precision and recall. This implies that ViT-16 is the preferred choice for tasks where achieving a high F1 Score is crucial, such as medical diagnoses, where both false positives and false negatives bear significant consequences.

4.2.1 Why ViT-16 Excels: Outperforming VGG16, VGG19, ResNet101 and ResNet50

The ViT-16 model demonstrates exceptional performance and surpasses its competitors. In the field of computer vision, the choice of a model architecture plays a crucial role in achieving outstanding results. The ViT-16, also known as the Vision Transformer with a 16-layer depth, has exhibited notable advantages when compared to traditional architectures such as VGG16, VGG19, ResNet101, and ResNet50. Let us examine the characteristics that position ViT-16 as the indisputable leader in this field.

Exceptional Accuracy and Generalization: The validation accuracy achieved by ViT-16, reaching a value of 99.04%, may be regarded as quite extraordinary. This achievement showcases the system's remarkable ability to discern intricate patterns inside images, hence showcasing its skill in the recognition and classification of objects. When evaluating the aspect of accuracy, it becomes apparent that ViT-16 outperforms its competitors, hence emphasizing its superior capacity to understand complex features included in the dataset.

Minimal Validation Loss: The ViT-16 model exhibits remarkable generalization capabilities, as seen by its validation loss of just 0.0201. The model's ability to effectively forecast outcomes, even given previously unseen data, is evidenced by the observed small loss value, indicating a high level of precision. In contrast, other models such as VGG16 and VGG19 have difficulties in efficiently decreasing losses, suggesting their restricted ability to acquire knowledge from the provided dataset.

Precision and Recall Mastery: The ViT-16 model demonstrates a high accuracy score of 0.99, indicating its ability to provide accurate positive predictions while effectively reducing the occurrence of false positives. This particular attribute is of paramount significance in scenarios where exactness is of ultimate relevance, such as in the fields of medical imaging or autonomous driving. Additionally, it is noteworthy

thy to mention that the ViT-16 model exhibits a recall score of 0.99, signifying its remarkable capacity to accurately identify almost all pertinent occurrences within the given dataset. This is a factor that warrants examination. The aforementioned attribute plays a crucial role in activities such as object detection and anomaly identification, hence making a substantial contribution to their total worth.

Impressive F1-Score: In this instance, the F1-score of 0.99 indicates that the ViT-16 model effectively achieves a favorable equilibrium between accuracy and memory use. This assertion is substantiated by the presence of a designated nomenclature for the model. The aforementioned measure effectively facilitates the categorization of various elements, rendering it a versatile and valuable option for several applications.

The exceptional performance of ViT-16 may be attributed to its remarkable accuracy, little loss, and impressive precision and recall values. The competence of computer vision in accumulating complicated visual features positions it as the leading contender for challenging jobs in this field. Significant advancements have been achieved as a result of the utilization of VGG16, VGG19, ResNet101, and ResNet50. The ViT-16 model represents a significant advancement in the field of deep learning for computer vision, as evidenced by its pioneering outcomes and expanded applicability in domains like as image analysis and object identification. This development signifies a transformative shift in the landscape of computer vision research.

Chapter 5

Future Works And Discussion

The domain of artificial intelligence (AI) in the realm of medical imaging, namely in the diagnosis of colorectal cancer, exhibits significant potential, although also necessitates more paths for study and advancement. Future research should prioritize the improvement of generalization capabilities in artificial intelligence models, namely focusing on the Vision Transformer (ViT) and Swin Transformer. The aforementioned models have exhibited exceptional performance. Nevertheless, it is possible to further optimize their capabilities by exploring methodologies such as transfer learning and fine-tuning on bigger and more varied datasets. The increased capacities resulting from this expansion will render them applicable to a broader array of medical imaging endeavors, exceeding the realm of colorectal cancer diagnosis. The subject matter of multi-modal fusion has considerable potential for future scholarly investigations. The integration of other data sources, like as patient metadata, genetic information, or clinical history, has the potential to offer significant contextual information for artificial intelligence (AI) models. The integration of visual data with other modalities necessitates the creation of fusion algorithms that can proficiently utilize a variety of information sources to enhance the precision and dependability of diagnostic outcomes.[3] The incorporation of interpretable artificial intelligence (AI) is of great importance in ensuring the effective implementation of AI-driven diagnostic tools within healthcare environments. Future research should prioritize the advancement of methodologies aimed at elucidating the underlying principles that govern the predictions generated by these models. The primary aim is to augment comprehension and uptake among healthcare professionals. The incorporation of interpretability inside artificial intelligence (AI) systems not only fosters a sense of confidence, but also empowers healthcare professionals to make informed decisions by leveraging the insights offered by AI.[43] Further research is required to explore data augmentation methods that are especially designed for medical imaging. One potential approach for addressing class imbalance concerns and improving the model's capacity to manage differences in picture quality and patient demographics involves the generation of synthetic data. The use of effective data augmentation techniques holds significant promise in enhancing the model's versatility and dependability. Efforts should be focused on enhancing the efficiency of model inference to achieve operational capabilities in real-time or near-real-time. Efficient usage of models on edge devices is crucial for the effective transfer of research findings into practical use, hence enhancing the accessibility of artificial intelligence for point-of-care applications. The research conducted in this particular domain possesses the ca-

capacity to yield substantial progress in terms of the efficiency and efficacy of AI-driven diagnostic systems. The essential relevance is in ensuring the robustness and integrity of AI-driven medical systems. In order to improve the field, it is recommended that future study place a higher emphasis on the examination of adversarial assaults and defenses. Additionally, there is a need for the development of more sophisticated approaches to detect and mitigate model biases. The proficient implementation of these procedures is crucial for mitigating the potential vulnerabilities of AI systems and guaranteeing their capacity to provide impartial and dependable outcomes. The establishment of a strong partnership between scholars specializing in artificial intelligence and practitioners in the healthcare industry has great importance. The establishment of interdisciplinary collaboration is crucial to guarantee that artificial intelligence (AI) models align with the specific needs and standards of the healthcare sector. The scope of this partnership should encompass the procurement of data, the formulation of models, and the validation of those models within clinical settings. Complying with the ever-evolving rules and standards, namely the criteria set out by the Food and Drug Administration (FDA), holds utmost importance when contemplating the incorporation of artificial intelligence in the healthcare sector. To enhance the acceptability and endorsement of AI systems for clinical utilization, it is advisable for forthcoming research endeavors to prioritize the alignment of these systems with regulatory standards. This entails the examination and resolution of matters pertaining to the safeguarding of data privacy, ethical deliberations, and the promotion of openness. Examining the wider implications for public health regarding the application of artificial intelligence in the identification of colorectal cancer is a complex undertaking.[31] This study encompasses an examination of the cost-effectiveness, patient outcomes, and population-level advantages associated with the implementation of artificial intelligence (AI) technology inside healthcare systems. The importance of performing these evaluations cannot be overstated in terms of comprehending the practical ramifications of integrating artificial intelligence into the healthcare sector. In conclusion, it is imperative for future research endeavors to tackle the obstacles associated with the scaling of artificial intelligence models. This is necessary to effectively manage extensive datasets including millions of pictures, given the ongoing growth in both the magnitude and intricacy of these datasets. To successfully accomplish this task, it is imperative to devise distributed training methodologies and model architectures that possess the capability to effectively handle and examine substantial quantities of medical imaging data. In summary, the application of artificial intelligence (AI) in the identification of colorectal cancer has considerable promise, as it presents avenues for enhancing the efficiency, interpretability, and applicability of models in clinical settings.[37] The optimization of artificial intelligence (AI) in enhancing cancer detection and improving patient outcomes necessitates the crucial collaboration of computer scientists, medical practitioners, and regulatory entities. The examination of these educational trajectories will provide a noteworthy academic contribution to the ongoing progress of artificial intelligence in the healthcare field, particularly in its transformational influence on the identification of colorectal cancer.

In the pursuit of advancing colorectal cancer detection through artificial intelligence, a promising avenue of research lies in the exploration of a hybrid model combining the strengths of Vision Transformer (ViT-16) and Swin Transformer. Both these transformer-based models have individually showcased remarkable proficiency in

image recognition tasks, setting the stage for a potentially groundbreaking amalgamation. The envisioned hybrid model seeks to harness the distinctive features and learning capabilities of ViT-16 and Swin Transformer, aiming to uncover synergies that could further enhance the accuracy and robustness of colorectal cancer diagnosis from medical images. Delving into this hybrid approach necessitates meticulous fine-tuning and validation on diverse and extensive datasets, ensuring the model's adaptability and generalization across varied clinical scenarios. This venture into creating a hybridized model holds the potential to push the boundaries of medical imaging diagnostics, offering a novel perspective in the detection methodologies and possibly uncovering intricate patterns unobservable by singular models. Moreover, the exploration of such a hybrid model underscores the commitment to continuous innovation in AI-driven healthcare solutions, striving for optimized performance and improved patient outcomes through the integration of diverse technological advancements. The development and subsequent evaluation of this hybrid model represent a pivotal step in the ongoing quest to refine and enhance the application of artificial intelligence in colorectal cancer detection, contributing significantly to the evolution of diagnostic precision and reliability in the medical field.

Developing this innovative hybrid model, which leverages the unique capabilities of ViT-16 and Swin Transformer, is our primary objective for future research endeavors. This pursuit signifies our commitment to advancing the field of AI in colorectal cancer detection, aiming to achieve unprecedented levels of diagnostic accuracy and reliability.

Chapter 6

Conclusion

The transformer architecture has evolved in the recent years and many different branches have been identified. The transformer model has now become one of the main highlights in deep learning and deep neural networks. The attention mechanism helps to improve the performance of encoder-decoder model for machine translation. Mainly the idea of the attention mechanism is to handle the inputs with more flexible manner. Previously, the attention mechanism was being implemented by RNN-based encoder-decoder architecture. The transformer model revolutionized the implementation of attention relying solely on self-attention mechanism. The whole field of computer vision is being transformed by the transformers. Also, the transformer related researches are undergoing a rapid growth in the medical image analysis field. By using the attention mechanism in the transformer-architecture we hope to show the outcome of colorectal cancer diagnosis. We intend to use image classification in order to predict the outcome of colorectal cancer that may help the doctors to provide a better outcome. The study results should be verified using new diverse image datasets in the further studies.

Bibliography

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, “Lung and colon cancer histopathological image dataset (lc25000),” 2019. arXiv: 1912.12142v1 [eess.IV].
- [3] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, “Digital pathology and artificial intelligence,” *The Lancet Oncology*, vol. 20, no. 5, e253–e261, 2019.
- [4] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, “Skin lesion segmentation in dermoscopic images with ensemble deep learning methods,” *IEEE Access*, vol. 8, pp. 4171–4181, 2020.
- [5] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, “A comprehensive review of deep learning in colon cancer,” *Computers in Biology and Medicine*, vol. 126, p. 104003, 2020.
- [6] A. Tsirikoglou, K. Stacke, G. Eilertsen, M. Lindvall, and J. Unger, “A study of deep learning colon cancer detection in limited data access scenarios,” *arXiv preprint arXiv:2005.10326*, 2020.
- [7] T. L. T. Vuong, D. Lee, J. T. Kwak, and K. Kim, “Multi-task deep learning for colon cancer grading,” in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, Jan. 2020, pp. 1–2.
- [8] J. Abdollahi, B. Nouri-Moghaddam, and M. Ghazanfari. “Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases).” (2021), [Online]. Available: <http://arxiv.org/abs/2103.08182>.
- [9] N. Adaloglou. “How the vision transformer (vit) works in 10 minutes: An image is worth 16x16 words.” (Jan. 2021), [Online]. Available: <https://theaisummer.com/vision-transformer/>.
- [10] D. S. Berman and A. L. Smith, “Challenges and limitations of using transformers in medical imaging,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 2934–2945, 2021.
- [11] T. L. Chaunzwa, A. Hosny, and Y. e. a. Xu, “Deep learning classification of lung cancer histology using ct images,” *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.
- [12] S. Choi and J. Han, “Integrating transformer and convolutional networks for enhanced colorectal cancer detection,” *Computers in Biology and Medicine*, vol. 130, p. 104204, 2021.

- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16 x 16 words: Transformers for image recognition at scale,” in *Proceedings of the International Conference on Learning Representations*, Virtual Conference, May 3–7, May 2021, pp. 3–7.
- [14] S. Gupta and M. K. Gupta, “A comparative analysis of deep learning approaches for predicting breast cancer survivability,” *Archives of Computational Methods in Engineering*, pp. 1–17, 2021.
- [15] S. Gupta and M. K. Gupta, “Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm,” *The Computer Journal*, 2021.
- [16] A. B. Hamida, M. Devanne, J. Weber, C. Truntzer, V. Derangère, F. Ghiringhelli, *et al.*, “Deep learning for colon cancer histopathological images analysis,” *Computers in Biology and Medicine*, vol. 136, p. 104730, 2021.
- [17] J. He, Q. Wang, Y. Zhang, H. Wu, Y. Zhou, and S. Zhao, “Preoperative prediction of regional lymph node metastasis of colorectal cancer based on 18f-fdg pet/ct and machine learning,” *Annals of Nuclear Medicine*, vol. 35, no. 5, pp. 617–627, 2021.
- [18] J. H. Kim, H. W. Chang, and S. W. Lee, “Early detection of colorectal cancer using machine learning algorithms,” *Computers in Biology and Medicine*, vol. 129, p. 104182, 2021.
- [19] M. Masud, N. Sikder, A. A. Nahid, A. K. Bairagi, and M. A. AlZain, “A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,” *Sensors*, vol. 21, no. 3, p. 748, 2021.
- [20] R. Nateghi, H. Danyali, and M. S. Helfroush, “A deep learning approach for mitosis detection: Application in tumor proliferation prediction from whole slide images,” *Artificial Intelligence in Medicine*, vol. 114, p. 102048, 2021.
- [21] R. Patel and V. Reddy, “Clinical applications of transformer-based machine learning models: Focus on colorectal cancer,” *Journal of Clinical Oncology*, vol. 39, no. 24, pp. 2712–2721, 2021.
- [22] D. Sui, K. Zhang, W. Liu, J. Chen, X. Ma, and Z. Tian, “Cst: A multitask learning framework for colorectal cancer region mining based on transformer,” *BioMed Research International*, vol. 2021, 2021.
- [23] C. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2021. arXiv: 1905.11946.
- [24] F. Wessels, M. Schmitt, E. Kriehoff-Henning, *et al.*, “Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer,” *BJU International*, vol. 128, no. 3, pp. 352–360, 2021.
- [25] P. L. Williams and E. Thompson, “Data augmentation techniques for colorectal cancer detection in medical imaging,” *Journal of Healthcare Engineering*, vol. 2021, p. 8564132, 2021.
- [26] E. Wulczyn, D. F. Steiner, M. Moran, *et al.*, “Interpretable survival prediction for colorectal cancer using deep learning,” *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–13, 2021.

- [27] M. A. E. Zeid, K. El-Bahnasy, and S. E. Abo-Youssef, "Multiclass colorectal cancer histology images classification using vision transformers," in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, IEEE, Dec. 2021, pp. 224–230.
- [28] X. Chen, K. Zhang, N. Abdoli, P. W. Gilley, X. Wang, H. Liu, *et al.*, "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, 2022.
- [29] R. Kumar, A. Sharma, and S. Varma, "Transformer-based architectures for histopathological image analysis in colorectal cancer," *Medical Image Analysis*, vol. 70, p. 102115, 2022.
- [30] T. Z. Li, K. Xu, R. Gao, *et al.*, "Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography," *arXiv preprint*, 2022, arXiv:2209.01676.
- [31] X. Liu, H. Zhang, and Y. Chen, "Self-attention generative adversarial networks for medical image analysis," *Neural Networks*, vol. 36, pp. 122–131, 2022.
- [32] C. Nguyen, Z. Asad, R. Deng, and Y. Huo, "Evaluating transformer-based semantic segmentation networks for pathological image segmentation," in *Medical Imaging 2022: Image Processing*, SPIE, vol. 12032, Apr. 2022, pp. 942–947.
- [33] M. Reyes, R. Meier, and S. Pereira, "Transfer learning with transformers for cancer imaging: Techniques and applications," *Frontiers in Oncology*, vol. 12, p. 453, 2022.
- [34] M. Roberts and T. Schmidt, "The impact of data augmentation on transformer models in detecting colorectal cancer from pathology slides," *Machine Learning in Medicine*, vol. 9, no. 1, pp. 11–25, 2022.
- [35] A. S. Sakr, N. F. Soliman, M. S. Al-Gaashani, P. Pławiak, A. A. Ateya, and M. Hammad, "An efficient deep learning approach for colon cancer detection," *Applied Sciences*, vol. 12, no. 17, p. 8450, 2022.
- [36] J. Smith and H. Lee, "Attention mechanisms in medical imaging: A survey," *Medical Image Analysis*, vol. 68, p. 101987, 2022.
- [37] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," *Expert Systems with Applications*, p. 117695, 2022.
- [38] Y. Zhang, S. Wang, and Z. Dong, "A meta-analysis of deep learning for detecting colorectal cancer from endoscopy images," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 22, 2022.
- [39] F. Zhou, Y. Zhao, and X. Li, "Deep learning methods in the diagnosis of colorectal cancer using computed tomography images: A survey," *Computer Methods and Programs in Biomedicine*, vol. 204, p. 106020, 2022.
- [40] T. S. Brown, A. B. Patel, and J. Clark, "Attention-based transformers for pathological tissue classification," *Journal of Medical Systems*, vol. 47, no. 3, pp. 99–109, 2023.

- [41] “DeepLearning.” (Sep. 2023), [Online]. Available: <https://github.com/wangshusen/DeepLearning>.
- [42] “Hands-on transfer learning with keras.” (Sep. 2023), [Online]. Available: <https://www.learndatasci.com/tutorials/hands-on-transfer-learning-keras/>.
- [43] T. Johnson and K. Williams, “Transformer models for medical image classification: A comparative study,” *Journal of Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 253–267, 2023.
- [44] W. Lee, J. Kim, and Y. Song, “Multi-modal attention mechanisms in medical image analysis: A survey,” *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 1–16, 2023.
- [45] “Swin transformer.” (Sep. 2023), [Online]. Available: <https://paperswithcode.com/method/swin-transformer>.
- [46] “Unknown title.” (Sep. 2023), [Online]. Available: https://www.google.com/url?sa=i&url=https%3A%2F%2Fmdpi-res.com%2Fd_attachment%2Fappls%2Fappls-13-03072%2Farticle_deploy%2Fappls-13-03072.pdf%3Fversion%3D1677503466.
- [47] “Vgg19 u-net implementation in tensorflow.” (Sep. 2023), [Online]. Available: <https://idiotdeveloper.com/vgg19-unet-implementation-in-tensorflow/>.
- [48] “Vision transformer.” (Sep. 2023), [Online]. Available: <https://theaisummer.com/vision-transformer/>.
- [49] “Vit - hugging face model documentation.” (Sep. 2023), [Online]. Available: https://huggingface.co/docs/transformers/model_doc/vit.
- [50] M. Salvi, M. Bosco, L. Molinaro, and et al., “A hybrid deep learning approach,” *Journal Name*, vol. Volume, no. Issue, Pages, Year.