Medical Image Classification Using Deep Learning Technique

MUHAMMAD TALLHA SALEEM

Enrollment No: 03-243221-009

Supervisor: Dr. Iram Noreen

A thesis submitted to the Department Computer Science, Faculty of Engineering Science, Bahria University Lahore campus in the partial fulfillment for the requirements of a Master degree in Computer science

March 2024

Approval for Examination

Scholar’s Name: Muhammad Tallha Saleem Registration No. 78833

Program of Study: MSCS

Thesis Title: Medical Image Classification Using Deep Learning Technique

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index ______% that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor’s Signature: ____

Date: ________________________

Name:

Author's Declaration

I, Muhammad Tallha Saleem hereby state that my MS thesis titled "Medical Image Classification Using Deep Learning Technique" is my own work and has not been submitted previously by me for taking any degree from this university "Bahria University, Lahore Campus" or anywhere else in the country/world.

At any time if my statement found to be incorrect even after my graduation the university has the right to withdraw/cancel my degree.

Name of Scholar: Muhammad Tallha Saleem

Date: ____________________

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled "Medical Image Classification Using Deep Learning Technique" is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw / revoke my MS degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: ________________________

Name of the Scholar: Muhammad Tallha Saleem

ACKNOWLEDGMENTS

First, I would like to thank my supervisor Dr. Iram Noreen for giving me time, ideas, guidance and encouragement. Thank you mam, for always being humble, calm and cooperative throughout the study and research work.

I want to show my gratitude to Dr. Ghulam Mustafa for being my unofficial supervisor. I really am thankful to you for providing me time, guidance, directions and valuable assistance during the thesis.

ABSTRACT

The research in medical image classification by the use of Deep Learning Techniques is a major trend these days as it serves best in medical diagnosis. Deep Learning Models help us to recognize the patterns and features with promising accuracy that task may be very difficult for medical practitioners due to time constraints and performing comparison with large datasets. As we are testing the maximum limit of technology and going exponentially in the field. Similarly this study was intended to reduce human effort by utilizing Deep Learning technologies in the field of medical by classifying Gastrointestinal (GI) images from an open source Dataset KVASIR V2 with all 8 classes. In most of the previous studies dataset was found as the main limitation as limited number of samples and classes were used. As compare to previous research this study, does image classification on increased number of classes with promising accuracy and identifies the suitable parameters. Initial results were carried of 3 CNN based models out of which two models EfficientNetV2B2 and VGG-16 were pretrained on ImageNet Dataset while the third model was AlexNet. Evaluation metrics reported in this study include accuracy, precision, recall, f1-measure including categorical accuracies and Confusion matrix. Based on above metrics we selected the best model EfficientNetV2B2 for further fine tuning. **The study achieved promising results in classification of GI Images based on pretrained model Efficient Net V2B2 with SGD optimizer by achieving training accuracy of 97.03%, validation accuracy 94.03%, while testing accuracy of 95.34%.** Transfer learning technique was tested by utilizing above two pre-trained models as a foundation, transfer learning proved to be great for substantial reductions in training time and processing resources beside AlexNet. Further utilizing Transfer learning leaded to better training and improved performance in classification of Gastrointestinal (GI) medical images. Lastly an application with GUI interface was built using Tkinter python library to better interact with image classification process.

Keywords: Medical Image Classification, Deep Learning, Endoscopic Images, Gastrointestinal (GI) images, KVASIR V2.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DL	-	Deep Learning
CNN / ConvNet	-	Convolutional Neural Network
GI	-	Gastrointestinal
AI	-	Artificial Intelligence
VGG	-	Visual Geometry Group
ACC	-	Accuracy
FC	-	Fully-Connected
ReLU	-	Rectified Linear Unit
CM	-	Confusion Matrix
AUC	-	Area Under the (ROC) Curve
ROC	-	Received Operating Characteristic

CHAPTER 1

INTRODUCTION

Among all cancers the gastric cancer is the 4th common type of cancer worldwide with 2nd most common cause of mortality. Around 80 thousands deaths are due to gastric cancer where Southern India, China, South-America and Easter-Europe have the high incidence of this disease [23]. According to a study [7], There is no national cancer registry in Pakistan having exact ratio of gastric cancer but yet a lot of gastric diseases are reported some are even leading to cancer. According to American Cancer Society [2] risk factors causing gastric cancer include age, gender, smoking, family history (genealogy), high salted diet, highly smoked and long preserved foods. Gastric cancer is taken as most dangerous in healthcare community as cancer cells in stomach may spread through lymphatic system or by bloodstream to other parts of the body easily, like in lungs and liver. The proposed research is about classifying endoscopic images of gastric and helps in the early detection of diseases which may lead to cancer including gastritis, dysplasia and ulcer.

Medical images is considered as the main source of disease diagnosis. For that reason Deep Learning and AI assisted detection systems are rapidly expanding in the medical sector. Implementation of DL and AI techniques in gastroscopy can play a role as third eye for endoscopists with greater performance and accuracy [36]. Early detection of disease plays a crucial factor in lowering the mortality rate from cancer. Due to the current modalities' of large data sets, ML is constrained, whereas DL is effective regardless of the data size also provide efficient results based on training and Techniques used. That’s why DL is seen as an improved scheme of machine learning as it provide better accuracy and flexibility [26]. The most widely known algorithm in Deep Learning domain for classification of images and processing is the Convolutional Neural Network (CNN). CNN is a prominent architecture for image classification and object detection too. With smart and more efficient image processing techniques CNN and models based on CNN architecture including ResNet, EfficientNet, AlexNet, VGG, INCEPTION, GoogLeNet are in mainstream for medical images processing [36].

1.1 Background to the Research

Medical image classification accuracy and efficiency could be improved by recent developments in the field of AI and deep learning. We can get benefits from technology in classification of Gastrointestinal (GI) endoscopic images by improving accuracy and efficiency. Thus training the model on larges dataset to learn features and applying to new unforeseen endoscopic images of gastric to classify will help us to test the models in an efficient way.

1.1.1 Architecture and working of the CNN Model for images classification

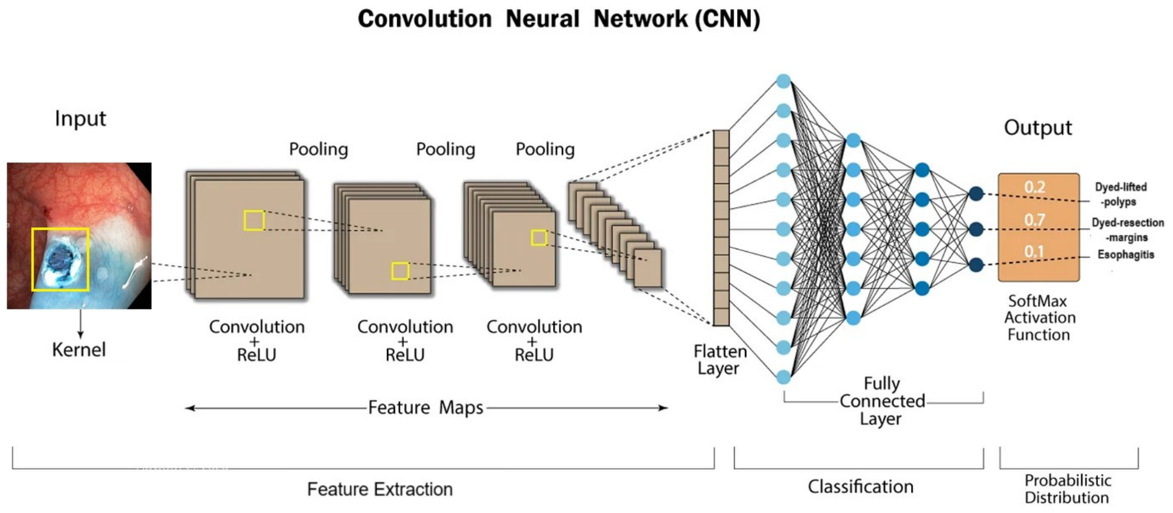


Figure 1.1: Architecture of a CNN Model [37]

In multiclass image classification, CNN models (Convolutional Neural Networks) work by learning and extracting features from images followed by sequence of convolutional, pooling, and FC layers. Here's a brief overview of how CNN models work in multiclass image classification:

1. Input Layer: The input layer of a CNN takes in the raw pixel values of the image.
2. Convolutional Layers: The Convolutional Layers comprise of filters and kernels that are convolved with input sample to extract features.
3. Activation Function: To add non-linearity to the model, an activation function (like ReLU or ELUs) is used after each convolutional operation.
4. Pooling Layers: By downsampling the feature maps acquired from the convolutional layers, pooling layers minimize the number of parameters in the network and its spatial dimensions.
5. Fully Connected Layer: One or more fully connected layers, working as a conventional neural network, take the flattened output of the final pooling layer and use it to learn more complex features and generate predictions.
6. Output Layer: This layer works by taking an input picture and generates class probabilities using an activation function (like SoftMax for multiclass classification).

The convolutional neural network (CNN) model learns its parameters during training by optimizing it using optimization methods such stochastic gradient descent to minimize a loss function, such as cross-entropy loss. For tasks involving the classification of images into many classes, the model is thus able to produce reliable predictions.

Multiclass image classification, object recognition, scene interpretation, and picture categorization are just a few of the many uses for convolutional neural network (CNN) models due to their ability to learn hierarchical features directly from input data [37].

1.2 Rationale / Research Gap / Significance of the research

The purpose of this study intended to reduce human effort by utilizing Deep

Learning Technologies. As researchers are continuously testing the limit of technology and we are going exponentially in the field. The field of research is not limited to just one domain, so for the same in this study we are going to test multiclass image classification in case of Gastro Intestinal images (through endoscopic images dataset). The goal of investigating deep learning's potential use in this setting is to make gastrointestinal (GI) images classification more efficient and accurate while decreasing the amount of human work required.

The utilization of deep learning technologies offers significant contributions to streamlining and enhancing gastrointestinal (GI) image classification within the context of endoscopic imaging. By eliminating the need for human intervention, deep learning algorithms—and CNNs in particular—have drastically improved the efficiency and accuracy of picture classification tasks by automatically learning complicated patterns and characteristics from raw image data. By leveraging deep learning, the process of classifying diverse and complex GI images can be automated, leading to increased efficiency and speed in identifying abnormalities or diseases within the gastrointestinal tract. Furthermore, deep learning models can continuously improve their performance through exposure to more data, contributing to the improvement of GI image classification accuracy over time. When applied to endoscopic imaging, deep learning has the ability to revolutionize the industry by providing faster and more accurate diagnoses, which in turn improves patient care and clinical results.

1.2.1 Theoretical Gap

Research in medical image classification encounters several challenges and gaps. In the first place, developing accurate and broadly applicable models is hindered by the lack of annotated medical pictures used to train deep learning models. In medical image classification research addressing bias and validating the performance of models with unseen data is also important but often overlooked challenges. Further using limited samples for the dataset training with appropriate techniques could result in great accuracy rates. However, the limited data lacks diversity, which makes the work done less than optimal.

The accurate and timely diagnosis of diseases is one of the important aspect in medical field. Treatment decisions are based on diagnosis of disease. Incidence of Gastric Cancer and subsequent mortality ratio is increasing at very high rate with more chance of its incidence as age increase as discussed in study [6] and [23]. In order to detect stomach abnormalities and cancer, doctors may use imaging techniques such as endoscopy, MRI, CT or CAT scans, and biopsies [29]. While Endoscopic images analysis is the most effective and widely used method to diagnose gastric issues causing cancer because of its high sensitivity. As accurate and enhanced analysis can diagnose gastric diseases at early stages the acute interpretation of endoscopic-images is a challenging task and there is also a challenge of inter observer variability and gastric experts[3].

More there is a limitation of endoscopic examinations in a specific time interval as examination take much time by the specialist further if data-size (endoscopic images to examine) is huge, it would be extremely difficult to diagnose efficiently and in timely manner. During analysis of endoscopic images missing rate was 22.2% even if there are two experts [29]. So use of DL Models to examine and classify the Medical Images will help out to overcome above limitations [26]. With the help of DL Models and AI, diagnosis of difficult cases in healthcare can be made easy as there is no need to examine all the images of all patients or of a huge dataset one by one rather detecting the images with probability of gastric diseases. The proposed research will help out researchers in healthcare in adoption of suitable model and techniques of endoscopic images classification on large datasets.

1.2.2 Contextual Gap / Analysis

To date, researchers are continuously digging through the latest trends of AI and deep learning and getting benefits from it by comprehensive implementations. The trend of AI in medical domain is spreading widely by studying and testing out the maximum results to be achieved. This can be achieved by working to define new and improve existing information processing and classification techniques. Thus fully utilizing AI and Deep Learning Techniques to maximize benefits from it. This study is intended to implement pretrained models for classification of medical images with improved number of classes, comparing and improving the classification results.

1.2.3 Methodological Gap / Analysis

The study [16] used a patent small dataset and few techniques by suggesting to enhance dataset, use of data augmentation in future research work and also suggested to use of cross validation, because there is a hazard of overfitting if the model is tested with only one test set, which in turn raises variance. So here Cross-validation helps out by demonstrating models capability in means of consistency with unseen data by allowing us to evaluate our model with additional data. Further few research have been conducted on using two or more Classification Models at the same time to use comparison measures in case of endoscopic images classification. Some studies used a small patent dataset while some studies used opensource dataset KVASIR V1 with up to 5 classification classes. In the proposed study we are going to use open source dataset namely KVASIR Version 2 (latest released version) with all 8 classification classes. In proposed research we are also going to run and find out more suitable model and parameters to achieve greater accuracy and validation scores upon large number of samples.

1.3 Problem statement

Implementation of Deep Learning Techniques in endoscopic images using Deep Learning comes up with some challenges like choosing suitable pertained model for Feature Extraction and Classification. Testing and improving the training, validation and testing accuracy on unseen data by utilizing transfer learning. Many studies have implemented different CNN models but Comparison of different Models and parameters is also required. So in the proposed research we are going to implement and will compare results of two or more CNN Models from EfficientNet, AlexNet, VGG-16, VGG-19, GoogleNet and ResNet. Another important problem is regarding dataset. Model training and results gets biased if dataset have low rate of variations. Some of the studies have used either small manually collected dataset or used an open-source dataset. In the proposed study we will be using endoscopic images of gastric using an open-source dataset from

KVASIR V2 [24] with 8 classes and validate results after training the model on unseen data.

1.4 Research Questions

Some of the questions that this investigation will try to address are as follows:
RQ1: What are suitable models, methodologies, parameters for classification of Gastrointestinal (GI) endoscopic images?
RQ2: Provide Comparison results of some CNN Models in classification of endoscopic images between different diseases of gastric endoscopic images?
RQ3: What accuracy, validation and test results can be achieved after training the model with suitable parameters?
RQ3: What accuracy, validation and test results can be achieved by utilizing Transfer Learning in medical imaging domain?
RQ4: What accuracy can be achieved after enhancing classification classes in case of gastric endoscopic images.

1.5 Objectives

The objective of study is to classify Gastrointestinal(GI) medical images achieving better accuracy with opensource dataset KVAIR V2 with eight categorical classes by utilizing Transfer Learning (using pretrained models). Further study includes implementation and observe evaluation measures by changing different hyperparameters and fine tuning the model. The study also aimed to provide comparison results between different CNN Models and finding out suitable parameters to achieve better training, validation and testing accuracy in case of medical image classification with increased number of classes.

1.6 Significance of the Study:

The study in medical image classification by the use of Deep Learning Techniques is a major trend these days as it serves best in medical diagnosis. Deep Learning Models help us to recognize the patterns and features with promising accuracy that task may be very difficult for medical practitioners due to time constraints and performing comparison with large datasets. As AI and Deep Learning technologies are evolved in every aspect of life and we are going exponentially the field so this study will also help to observe at how much extent we can rely on Deep Learning in case of medical images classification.

Using Deep Learning Techniques the process of image analysis is automated and models also helps us in classification of medical images and enhancing efficiency. This research have momentum towards the classification of endoscopic Gastrointestinal(GI) images by the use of deep learning techniques. The proposed method we will be using convolutional neural network-based models from Deep Learning to extract relevant information from the images and then classifies them into several groups, such as normal-cecum/pylorus, normal-z-line,dyed-lifted-polyps and so on depending on features of each class.

1.7 Thesis organization

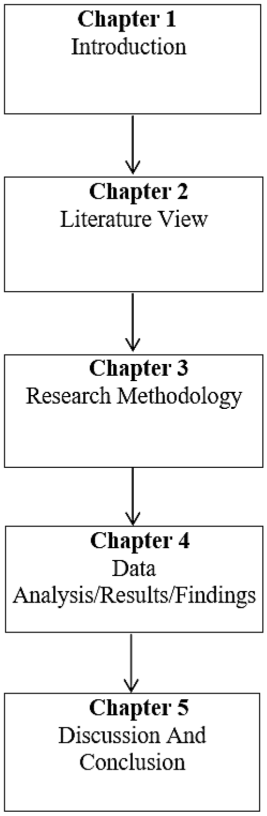


Figure 1.2: Thesis Organization

Figure 1.2 shows the thesis's organizational structure. The survey of the literature on medical image classification using deep learning techniques is covered in Chapter 2. It includes an overview of significant previous research works. The method of the research will be covered in Chapter 3. It will provide a detailed representation of the process of classification for gastrointestinal (GI) images. The discussion of findings and evaluations will be followed in chapter 4. It determines the entirety of the results, whether a hypothesis is agreed upon or not, and also compares the findings with those of earlier research. The results will be presented. Finally, chapter 5 presents the findings and makes recommendations for further research work.

1.8 Summary of Chapter 1

In this introductory chapter we started by highlighting the necessity of the study by presenting prevalence of the Gastric cancer as the second most frequent cancer in terms of death and the fourth most common cancer overall. The importance of early identification of diseases is highlighted in the chapter, which focuses on the significance of medical imaging, especially endoscopic pictures, in the diagnostic process. Chapter 1 describes how Convolutional Neural Networks (CNNs) are used as a primary tool for picture classification in medical image analysis, with an emphasis on Deep Learning (DL) and AI globally. The research seeks to fill gaps in current datasets and methodology by utilizing DL technology to improve the accuracy of GI endoscopic image classification. Along with the importance of the work in improving medical picture classification, important research problems are raised about appropriate models, methodology, and parameters for classification. At the end of the chapter, the thesis's structure is outlined. This structure consists of a literature review, methodology, discussion of the findings, and suggestions for further research.

CHAPTER 2

LITERATURE REVIEW

2.1 Review of the Literature

The current and ongoing research in the domain of Medical image classification helped to improve models in means of methodologies and techniques in endoscopic images classification. The study [33] proposed a DL framework and used pretrained models VGG-16, inception V3 and ResNet-50. Dataset was too small as 380 endoscopic images have been taken of each category where total category or classes was three normal, chronic erosive and ulcer gastric images. Classification accuracy for VGG-16 was 94.80%, for inception v3 it was 92.53%, and for ResNet it was 90.23%. That means accuracy improved as in previous study [33] got accuracy of 89% by use of CNN but limitation was the small dataset as not certain yet how models will perform on unforeseen or large data-set [33]. Study [22] used transfer Learning and classified 5 classes of gastric disease. Dataset was also a constraint here and data augmentation technique was implemented to enhance it. MobileNet V2 with ImageNet data-set was used and fine-tuned the model with small dataset of gastric endoscopic images. Accuracy achieved by study was 96.40% where evaluation measures also taken including accuracy, precision and recall.

Gastrointestinal abnormality using wavelet transform and DCNN of endoscopic image [9] used HyperKvasir dataset. The implementation was divided into 5 stages which includes image resizing, image preprocessing, image augmentation, Empirical Wavelet Transform and classification. EWT is another technique which helps to extract feature pattern from given dataset of images to improve the classification performance. HyperKvasir includes dataset of lower and upper Gastrointestinal where the research achieved the accuracy of 96.6%. But the study can be further analyzed by incorporating manual dataset while same DL model was implemented on 2 “levels with stages” so a few changes in model can further helps to achieve better results and performance comparison. Study [34] introduced Class Imbalance Loss an alternative to the loss functions to a Deep Neural Network (MobileNet-V2) where class imbalance technique worked on basics of assigning weights to each sample utilizing Gradient Descending in training process while also focusing on adjusting classifiers. Backpropagation algo used to update parameters of errors that are calculated by loss function. Study used manual and KVASIR dataset of gastric endoscopic images for training and implementation the achieved accuracy is 94.01%.

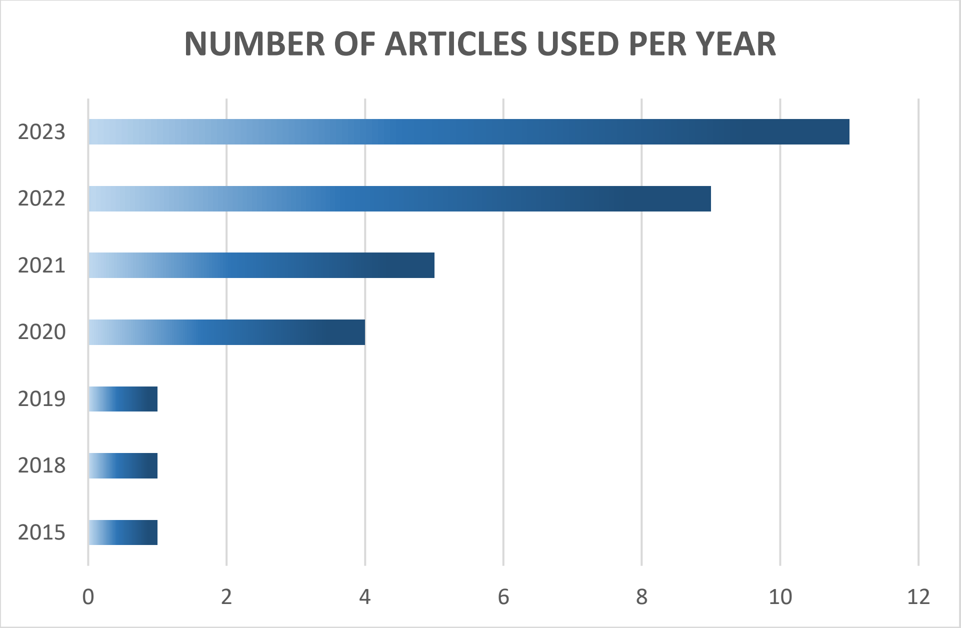


Figure 2.1: Number of articles used per year

Study [3], [6] and [21] are based on theoretical research focusing on current technologies and advancements in field of medical and endoscopic images classification. Below literature review table contains author, model used, techniques and accuracy information.

The figure below illustrates frequency of different Models and Techniques implemented so far from chosen recent years studies in the classification, accuracy and performance evaluation of medical images.

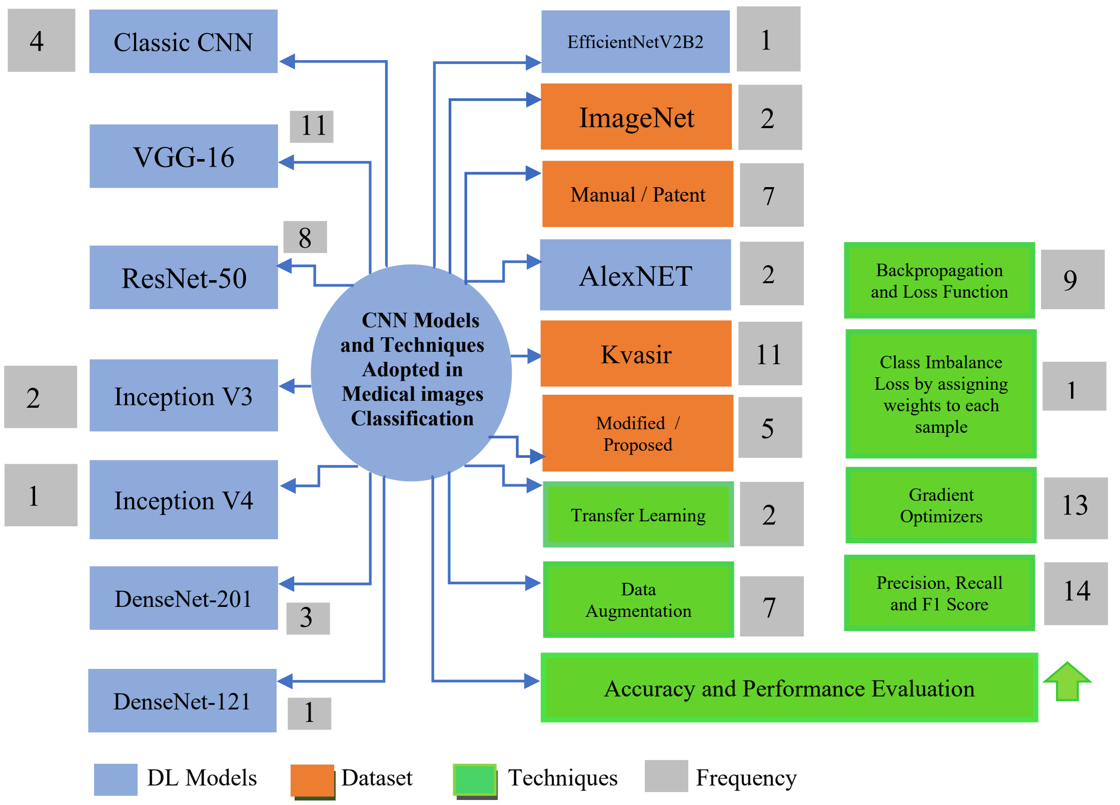


Figure 2.2: Frequency Diagram for CNN Models and Techniques Adopted in Medical images Classification

A literature review table comparing models, methodologies, datasets, limitations, contributions, and assessment measures across multiple research is also provided in the table below.

Literature Review Table

Table 2.1: Comparison of model and techniques for Medical Image Classification

Model / Technique Name	Dataset Used	Limitation	Contribution	Evaluation Measures
VGG-16, ResNet-50	Locally Collected Samples	Small Dataset. performance and results may vary if dataset is enhanced. Further fine tuning can improve performance and accuracy of model	Classified 3 classes of endoscopic gastric images and achieved 94.80% accuracy	Accuracy, Confusion Matrix
MobileNet v2 Used Transfer Learning with ImageNet Dataset	Locally Collected Samples	Dataset Limitation. performance and results may vary if dataset is enhanced	Classified 5 classes of endoscopic gastric images and achieved 96.40% accuracy	Accuracy, Precision and Recall
CNN, Data Augmentation	3591 samples collected from tertiary hospital	Dataset Limitation. performance and results may vary with increased dataset and classes	Classified 4 classes of endoscopic gastric images and achieved 94.1% accuracy	Accuracy, precision, recall, and F-measure
VGG-Net, ResNet pretrained on Imagenet, Transfer Learning	Locally collected 787 samples of gastrointestinal images	Too Small Dataset. performance and results may vary if dataset increased	Classified 3 classes of gastrointestinal by achieving 90% accuracy through ResNet	Accuracy

Model / Technique Name	**Dataset Used**	**Limitation**	**Contribution**	**Evaluation Measures**
Customized Model derived from ResNET-50 Localization, Segmentation	Locally Collected Samples	The limitation of this research includes small dataset and limited number of classes	Classified 2 classes normal and EGC. Segmentation done based on ROI and pixel level labeling. Achieved accuracy of 98%	Accuracy, Precision, Recall, Specificity and F1-score.
ResNet-152, Generalization with Grad-CAM(Heat Map)	KVASIR open-source dataset	- - -	Achieved a good accuracy 93.46% with large dataset.	Accuracy, Confusion Matrix
Siamese Neural Network, Distance metric-based technique, (regularization and augmentation)	Manual collected samples from video frames	Dataset limitation as a total of 115 images trained	Achieved accuracy of 92% for wireless endoscopic images	F1-score, AUC, and accuracy

2.2 Summary of the chapter

Recent Advancements in endoscopic image classification methodologies and approaches are the primary emphasis of this chapter, which explores the present landscape of medical image classification. The use of deep learning frameworks and models including ResNet-50, MobileNet V2, Inception V3, Efficient Net, and VGG-16, this chapter examined a number of research studies. The use of different dataset of Gastrointestinal Medical images studied, manual and opensource datasets like KVASIR, GastroEffNet also mentioned. The main limitation found as most of the previous studies used small datasets with limited number of samples. The use of data augmentation and transfer learning for the classification of stomach disorders was also covered in light of prior research. Illustration done about the frequency of different models and techniques implemented in recent years for medical image classification. A literature review table comparing models, methodologies, datasets, limitations, contributions, and assessment measures across multiple research is also provided in the chapter.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Epistemology

The intended study follows a positivist paradigm for gastrointestinal images classification using Deep Learning. The positivist paradigm relies on the belief that knowledge can be discovered through **empirical observation and scientific method**. The type of learning is supervised in which labeled data is provided to the model and model predicts the class of sample followed by its training on labeled data. In this context, the use of Deep Learning for classification is rooted in the collection and analysis of large datasets of gastrointestinal images to identify patterns and features that can be used for accurate classification. In this study we focused on empirical observation, systematic experimentation, and the objective analysis of data, which are essential for advancing the field of medical imaging and diagnosis.

3.2 Research approach

The proposed research work is divided in various components, each of which is described in detail below:

Research Methodology diagram

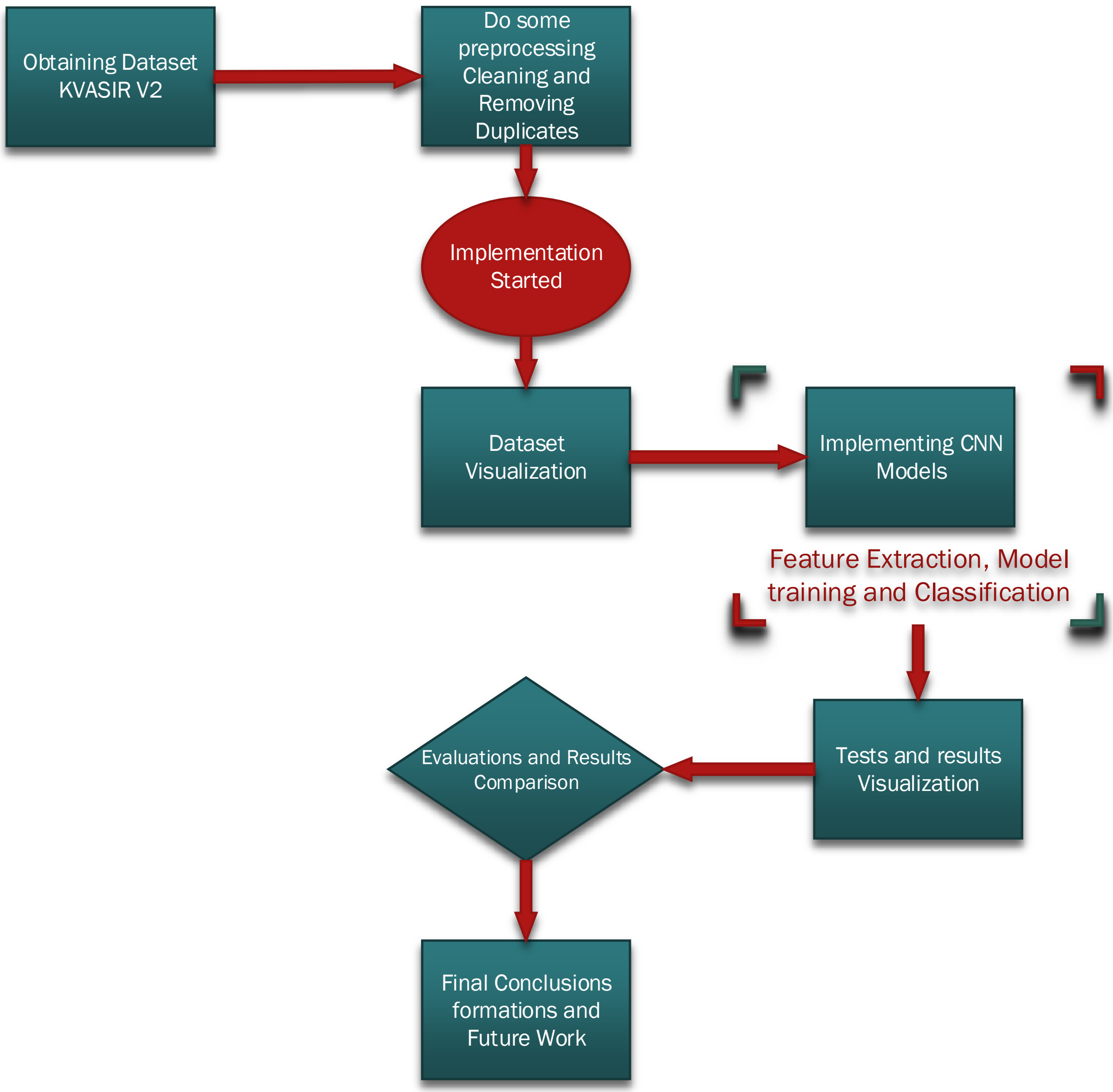

Obtaining Dataset KVASIR V2
Do some preprocessing Cleaning and Removing Duplicates
Implementation Started
Dataset Visualization
Implementing CNN Models
Feature Extraction, Model training and Classification
Tests and results Visualization
Evaluations and Results Comparison
Final Conclusions formations and Future Work

Figure3.1: Research Methodology diagram

3.2.1 Transfer Learning

Transfer learning is highly beneficial for classifying images using the pretrained models on larger datasets. By utilizing a pre-trained model as a foundation, transfer learning allows for substantial reductions in training time and processing resources. By leveraging knowledge garnered from a larger and more diverse dataset, transfer learning may leads to improved performance in classifying medical images too. TL also addresses the challenge of data scarcity, as it allows the utilization of knowledge from larger datasets, thus mitigating the impact of limited data. Additionally, the adaptability of pre-trained models used in transfer learning makes them well-suited for various types of visual recognition tasks, including medical images classification. For that reason, in this study we are going to implement and analyze the performance measurement utilizing transfer learning on KVASIR-V2 Dataset for classification of gastrointestinal diseases.

3.3 Research Strategy

This empirical study starts with dataset obtaining and preprocessing, then we will delve into implementation stage. The correctness of the model will be tested using pre-trained models based on training, validation, and testing findings. The best model will be chosen and fine tuning will be performed to further enhance the accuracy. The KVASIR V2 GI images dataset is well suited to be passed to CCN Models for classification. Timeframe is being followed as per institutes policy:

3RD SEMESTER ACTIVITY: PHASE 1 (SPRING 2023)

	WEEKS															
Activity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Literature review	■	■	■	■					Mid Term						Exam preparation	Exam week
Data collection					■	■	■									
Data preprocessing								■		■	■					
Analysis and predicting												■	■	■		

4TH SEMESTER ACTIVITY: PHASE 2 (FALL 2023)

Activity	WEEKS															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Proposed model building	■	■						Midterm preparation	Mid Term					Exam Preparation	Exam preparation	Exam week
Testing and Validation			■	■	■											
Model evaluation						■	■			■						
Final documentation											■	■	■			

3.4 Population and Sampling

In the context of an open source dataset like our case KVASIR V2 is incorporated containing 8 classes of gastrointestinal diseases, In this context, "the population" means all of the gastrointestinal photos that make up the dataset. The sampling process involves selecting a subset of images from this population for training, and testing and validation to test models performance on unseen data.

The KVASIR V2 dataset contains a large and diverse collection of gastrointestinal images, representing a wide range of conditions, anatomical structures, and pathologies. The population in this case would encompass all the images included in the dataset, which could potentially number in 8000, reflecting the variety of gastrointestinal conditions and abnormalities. Having a large enough and diverse sample that represents the population well improves the validity and reliability of the research and the models performance. Each of the three categories—training, validation, and testing—will be trained as the ratio of: 70%, 20%, 10% then 80%, 15%, 5% and lastly 85%, 10% and 5%, respectively and results will be analyzed. The data will be further passed to CNN model for training and classification.

3.5 Dataset Collection and Data Analysis / Preprocessing

In this research, an open source dataset of endoscopic images “KVASIR v2” by [24] is incorporated. As per author V2 was released in 2019 and it has more annotations and went also trough another round of quality control by medical experts. The dataset contains 8 classes where each class contains 1000 samples. In total the dataset have 8000 images. Data preprocessing is done to remove additional bounding boxes around images while 2 to 3 samples from each class were removed in preprocessing which contains either extreme noise, high distortions, low quality or seems to have low impact to contribute for feature extraction. Further to balance the dataset an equal no of 3 images were removed to reduce chances of model’s biasness. The final dataset contained 997 samples of each class with a total sample size of 7976.

3.5.1 Feature extraction

For feature extraction and classification two pretrained Deep Learning Models EfficientNet V2 B2, VGG-16 were used while results also studied on AlexNet. Purpose is to use pretrained models, studying accuracy, fine tuning the suitable model on different optimizers and hyperparameters, then comparing the results.

3.5.2 Model Selection for training (Classifier)

Two pretrained CNN Models from EfficientNet V2 B2, VGG-16 are incorporated and results will be compared while results also studied on AlexNet. By utilizing pre-trained models that were originally developed on large-scale datasets like ImageNet, medical picture categorization can benefit from transfer learning. ImageNet dataset contains millions of annotated images across numerous categories, and the pre-trained models have learned to extract a diverse set of features from visual data. When applied to medical image classification, these pre-trained models serve as feature extractors to capture meaningful patterns and structures within the images. This method presents several advantages, including the **reusability of features** learned from general visual data, which can be relevant to medical images, thereby **reducing the training time and computational resources** required to achieve optimal performance. Additionally, transfer learning can enhance the generalization of medical image classifiers, particularly when working with small or limited medical image datasets. We can enhance the precision and effectiveness of medical image classification models through the application of transfer learning by customizing and refining these pre-trained models to the unique features of medical image datasets. Additionally, various optimization techniques including as SGD, Adam, and Adagrad will be implemented. SGD, which stands for stochastic gradient descent, is commonly used in ML and DL applications to determine which model parameters yield the most accurate predictions and actual results. Different Techniques will be implement to get predicted outcome including Batch Size, Learning Rate, Drop rate where Early Stopping to avoid overfitting.

3.5.3 Result Generation and Evaluation

The predicted results will be visualized, results will be compared and performance evaluation will be done based upon training, validation and testing accuracy. Further precision, recall and F1 measure will be analyzed to better understand results.

3.5.4 Tools to be used

1. **Google Colab for Research Implementation,**

   Google Colab currently holds Python version 3 and uses Google Compute Engine at backend (T4 and A100 GPU's), We used Colab Pro version to faster the training and convergence process with access from 15 to 40GB of GPU Memory and 12.7 to 83.5GB of RAM with storage size of 78.2 Gb.

2. **Mendeley for Reference Management**
3. **MS Word for Documentation**
4. **MS Visio for Charts or Flow Diagrams**

3.6 Research Ethics

3.6.1 Privacy And Confidentiality

The research in medical domain is critical in its nature as ensuring the privacy and confidentiality of patients data. As we incorporated KVASIR-V2 dataset which is an opensource dataset and was released in 2019. The authors of the dataset didn't added the personal information of the patients and allowed its use for the purpose of research.

3.6.2 Informed consent

The authors allowed the use of dataset for research purpose only and with condition to place source reference [24].

3.6.3 Transparency in Reporting

Complete transparency in reporting methods, procedures, and results is followed in this research. I want to make this study more credible and reliable by giving specifics on the methodology utilized, the steps taken, and the outcomes got. Encouraging other scholars to reproduce and validate the findings while accurately assessing the study's validity and potential bias.

.

3.7 Summary of Chapter 3

Here in the research methods chapter, we explored the theoretical foundations of the study, with an emphasis on the positivist paradigm that is based on scientific method and empirical observation. This study incorporated pre-trained Deep Learning models to identify gastrointestinal images using supervised learning, which makes use of labeled data. In the context of medical imaging datasets using an increased number of classes, transfer learning is emphasized as a crucial method for improving classification accuracy by utilizing pre-trained models. From collecting and preprocessing dataset to developing and testing models and evaluating their efficacy, this chapter lays out the whole research method that was implemented over the course of two semesters. Investigated here are

CNN-based models for feature extraction that make advantage of Transfer Learning, such as EfficientNetV2B2 and VGG-16. With an emphasis transfer learning and optimization methods like SGD, Adam, and Adagrad, the criteria for selecting models are detailed. Methods of evaluation measures, the process of result generation, and ethical considerations pertaining to patient privacy and reporting transparency elaborated in this chapter. Google Colab, Microsoft Word, Microsoft Visio, and Mendeley are among the tools used in research implementation, documentation and reference management.

CHAPTER 4

DATA ANALYSIS/RESULTS/FINDINGS

4.1 Descriptive & Demographic analysis

As we incorporated KVASIR-V2 dataset which is an opensource dataset sourced from Vestre Viken Health Trust, a Norwegian hospital covering 26 municipalities around Norway. As per one of the author V2 was released 2019. It has more annotations and went through another round of quality control by medical experts. The dataset includes images from patients of different ages, genders, and ethnic backgrounds and from patients from various geographic regions, which can provide insights into any geographical variations in the prevalence of medical conditions. However due to privacy and ethical conditions the authors of the dataset didn’t included these personal information like age, gender and so on. Further these personal information was also not required for deep learning models for the classification purpose.

4.1.1 Preprocessing and Dataset Visualization

The dataset is quite robust as it contains 8000 samples in total with 8 classes, each class have 1000 samples based on features and characteristics of specific Gastrointestinal disease. The fact that each class has exactly 1000 samples suggests a well-balanced dataset, which can be beneficial for Deep Learning models. All samples were in jpg format. In the stage of preprocessing, we once again manually checked each sample based on the quality of image, extreme noise ratio or dull samples, sample scaling issues or and duplicates if any. Most of the samples were on similar scale based on resolution around 720*576. During manual preprocessing some sample were cropped to remove black bounding boxes around the images. Based on quality of images, extreme noise or having low impact in feature extraction, we removed around 2 to 3 samples from each class. Finally to balance the dataset an equal no of 3 sample were selected to be removed from each class. So the final dataset size we obtained after manual preprocessing was 7976. Means each class have an equal size of samples which is 997. Visualization of dataset can be seen in the figure below after loading the dataset on pretrained models.

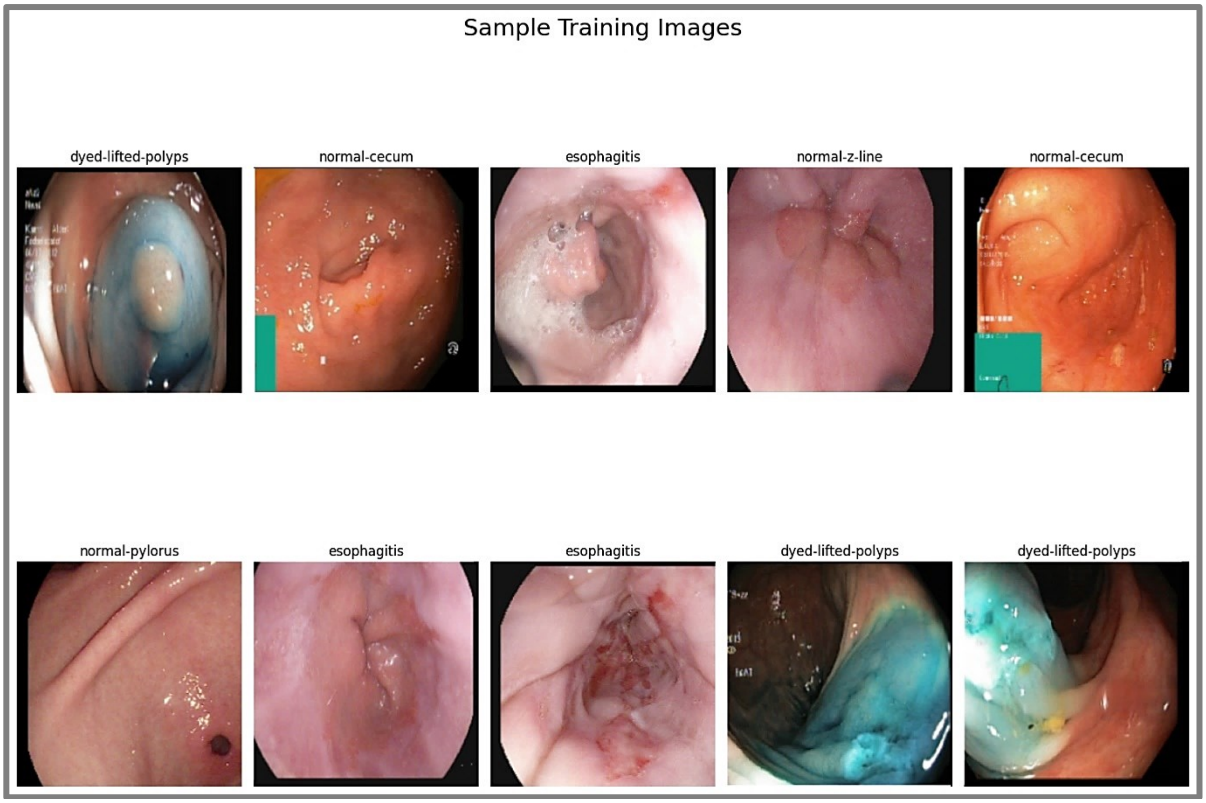


Figure 4.1: Visualization of Data Samples

The histogram of the training dataset is provided below. An equal no of samples from each class were passed to the model to avoid biasness factor and achieve fair results:

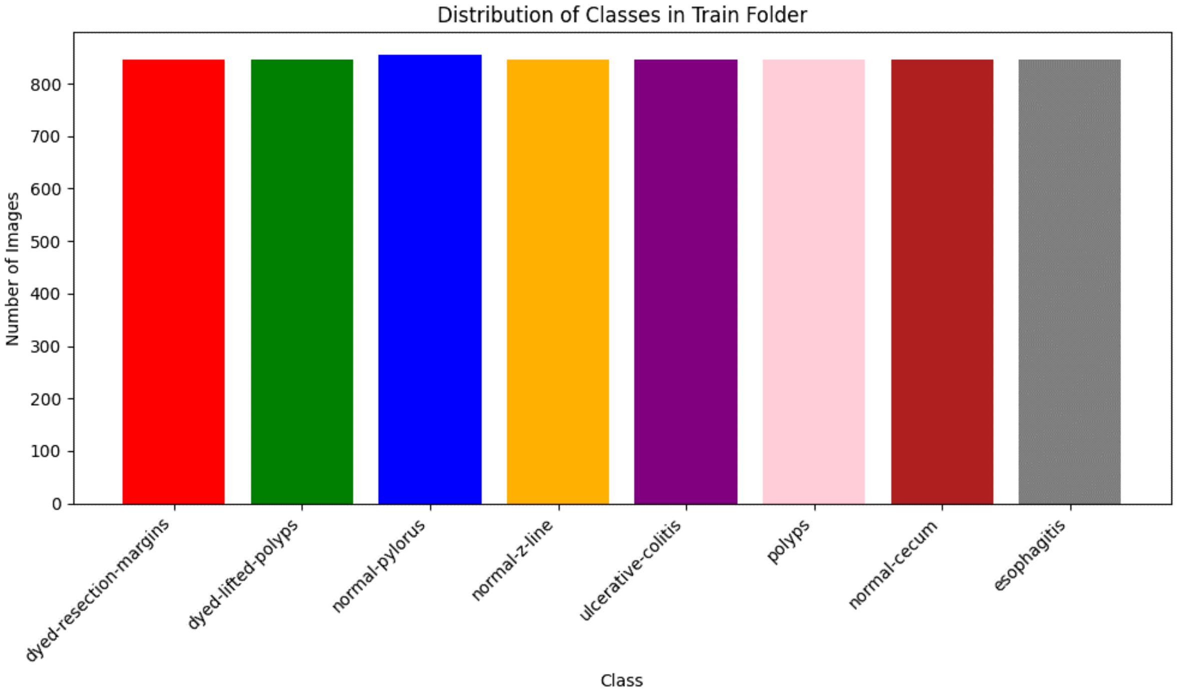


Figure 4.2: Histogram / Distribution of Classes for training

The label mappings and 8 classes present in the datasets include, 1st class corresponds to dyed-lifted-polyps, 2nd class corresponds to dyed-resection-margins, 3rd class resembles esophagitis, 4th class corresponds to normal-cecum, 5th class corresponds to normal-pylorus, 6th class resembles normal-z-line, 7th class corresponds to polyps, and 8th class corresponds to ulcerative-colitis. These mappings indicate the classification of different conditions or features within the datasets, which are used for training validation and testing purposes in DL models.

The use of data augmentation is a crucial aspect in medical image processing for deep learning and computer-assisted diagnosis, since the use of these methods might produce biased outcomes. This is due to the fact that the data being generated is not authentic, as data augmentation entails making copies of the initial medical photos in order to diversify the dataset and there is a concern in research community that using generated data could lead to unfair results, as the augmented data may not accurately represent the true distribution of medical images [38]. Following above reasons we also not incorporated some more techniques like image enhancement to reserve the originality of the data.

4.2 Reliability and validity analysis

We started the implementation by using three pretrained models: VGG16, AlexNet, and EfficientNetV2B2. After preprocessing, we formed the final benchmark dataset, which consists of 7976 images belonging to eight classes. Initially as a general rule of thumb the train, validation and test split ratio is carried out to be 70%, 20% and 10%, the training set comprises 70% of the data while the validation set accounts for 20%, and the remaining 10% is used for testing the model's accuracy and performance on unseen samples. We assessed the model's performance after training, validation, and testing based on the model's learning experience, which includes the accuracy during training, validation, and testing, and also its losses. Support, f1-score, accuracy, and recall are also assessed. Since sigmoid or logistic regression used in binary classification whereas for our multiclass problem softmax activation function was used in all three models.

As one of the purpose of this study was to achieve a greater accuracy while utilizing the whole dataset with eight classes. So we first implemented the said three models, evaluated the results and measurements and chosen the best model for fine tuning. an excellent architecture for vision models that is currently accessible. The 3x3 filter convolution layers, stride=1, and consistent use of the same padding are one of the most noticeable aspects of VGG16. Other notable features include maxpool layers, 2x2 filter, stride= 2, and the usage of 3x3 filter. The layout of the max pool and convolution layers is fixed over the whole design. At the very end of the model, there are two fully connected (FC) layers, and then the output is handled by a SoftMax. The "16" in VGG16 indicates that it has sixteen layers of varying densities. With almost 138 million parameters, this network is extensive [39]. The maximum accuracy on VGG16 without fine-tuning and using Adam optimizer was achieved as 88% training, 84% validation and 85% test accuracy at 40 epochs by using Adam optimizer. We further tested models performance by changing optimizer to Adagrad, Adagrad optimizer is a popular optimization algorithm used in image classification tasks. It is particularly effective when it comes to changing the learning rate for specific parameters depending on their previous gradients. This implies that the learning rate will be lower for often changed parameters and greater for less frequently updated ones. This adaptive learning rate mechanism allows Adagrad to efficiently handle sparse data and non-uniform features, which are common characteristics of image data. As a result, With Adagrad's assistance, training convergence can be accelerated, leading to more accurate picture classification models. [40]. While changing optimizer to Adagrad we achieved 89.54% training, 82.92% validation and 83.09% test set accuracy but testing, training and validation loss was too high, 25.22%, 49.33% and 51.74% respectively. The learning rate was throughout 0.001 percent so that model can better converge while training. Further we tested pretrained VGG16 model utilizing SGD optimizer. SGD (Stochastic Gradient Descent) is considered more suitable than GD (Gradient Descent) for large datasets or when there are limited computational resources because of its faster and more efficient parameter updates. While GD calculates the average gradient of the entire training dataset for updating model parameters, which can be slow and memory-intensive for large datasets, Using the gradient of just one training example at a time, SGD adjusts the parameters. This allows for quicker iteration through the dataset and more efficient use of computational resources, making it an ideal choice for situations where memory or computing power is limited. Therefore, SGD's ability to make incremental updates using individual training examples makes it well-suited for large-scale datasets and resource-constrained environments [40]. The results for training, validation and testing accuracy was 77.33%, 80.03%, and 76.88% respectively at 40 number of epochs. The training loss found to be high while model also depicted a large no of parameters as non-trainable.

Subsequently we carried out our tests on pretrained EfficientNetB2V2 on ImageNet database, and this model can classify images 1000 objects classes. By incorporating the effectiveness of the EfficientNet model, EfficientNetB2V2 further creates a robust CNN architecture with additional enhancements to achieve superior performance in image recognition tasks. Version 2 of EfficientNet comes with increased training speed parameter efficiency by using a blend of scaling (depth, width and resolution) and neural architecture search methods [39].

This model utilizes a combination of depth-wise separable convolutions, mixed-scale feature fusion, and advanced regularization techniques to optimize the trade-off between model size and accuracy. The scalability of EfficientNetB2V2's design allows it to accomplish outstanding performance in a variety of computer vision tasks, such as object recognition, classification and segmentation process of image [39]. The initial accuracy on training, testing and validation dataset was promising and convergence speed found to be good also. The training accuracy was achieved 98% upon Adams optimizer with 40 epochs, validation accuracy 91.76% while testing accuracy was 91%. The results was promising but yet validation and training loss was high 39.84% and 44.76% respectively. Next we tested out EfficinetNetB2V2 utilizing Adagrad optimizer, Training the model upon 40 epochs we achieved the training accuracy of 91.27%, validation accuracy 89.91% and testing accuracy 90.62%. Yet the loss was high, training loss was 22.15%, validation loss 23.78% and testing loss was 23.52%.

AlexNet model with Adams optimizer comes with training accuracy of 71.48% on 60 epochs while validation and testing accuracy was also low 71.31% and 70.87% with a critical loss results. Changing optimizer to Adagrad and SGD also depicted very undesirable results. The below table shows the initial results of experiments with adjusted number of epochs, epochs were stopped when further validation accuracy was being dropped even if training accuracy is improving as this scenario leads to overfitting where model shows great training accuracy while too bad accuracy for test or validation data.

Table 4.1: Initial experimental results of Model training

Model	**Optimizer**	**No. of Epochs**	**Training Accuracy**	**Training loss**	**Validation Accuracy**	**Validation loss**	**Test set Accuracy**	**Test Loss**
VGG16 Pre-trained on ImageNet	Adam	40	88%	28%	84%	46%	85%	50%
VGG16 Pre-trained on ImageNet	Adagrad	60	89.54%	25.22%	82.92%	49.93%	83.09%	51.74%
VGG16 Pre-trained on ImageNet	SGD	40	77.33%	54.69%	80.03%	48.71%	76.88%	54.92%
Efficient Net V2B2 Pre-trained on ImageNet	Adams	40	98.00%	5.10%	91.76%	39.84%	91.23%	44.76%
Efficient Net V2B2 Pre-trained on ImageNet	**Adagrad**	**40**	**91.27%**	**22.15%**	**89.91%**	**23.78%**	**90.62%**	**23.52%**
Efficient Net V2B2 Pre-trained on ImageNet	**SGD**	**40**	**95.60%**	**11.80%**	**92.61%**	**18.53%**	**93.63%**	**17.85%**
AlexNet	Adams	60	71.48%	63.08%	71.31%	63.93%	70.87%	65.01%
AlexNet	Adagrad	40	59.97%	89.75%	68.53%	72.99%	68.25%	73.71%
AlexNet	SGD	40	66.21%	66.21%	72.74%	60.70%	70.50%	60.09%

4.3 Experimental/Factor analysis

As from the initial experiments we chosen Efficient Net V2B2 as our base model with SGD Optimizer following the model's results and further delved into fine tuning the model by adjusting hyperparameters and evaluation measures, the architecture of accepted model Efficient Net V2B2 is provided below:

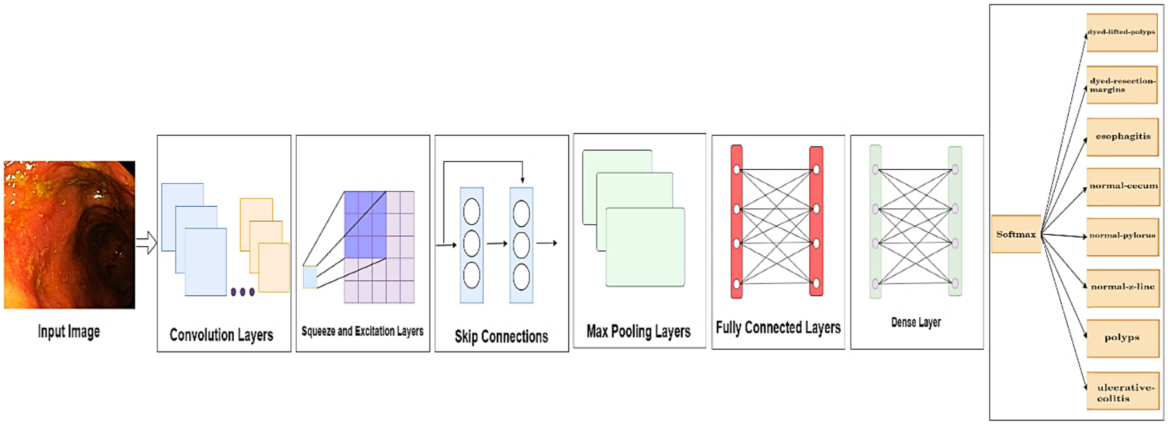


Figure 4.3: Architecture of Chosen Model Efficient Net V2 B2 [41]

While further tunning and adjusting the hyperparameters of the proposed model we utilized batch size of 32 so that 32 samples can passed through model training at a time, the batch size was taken into consideration following the total size of sample and dataset. Image width and height passed to model with different ratios like 128*128, 224*224, 256*256, 300*300 and 320*320 during parameter settings. In our final proposed model Train, validation and testing ratio was set as 85% training, 10% validation and 5% as test set. We also checked models performance by changing training model with 70% training data, 20% validation and 10% as test set data and also 80% training, validation at 15% and 5% for test set. The learning rate was set to default at 0.001 so that model can better converge the dataset and training process. During the online preprocessing 20% zoom range was set to achieve focus and center of the samples. The purpose of this technique was also to better avoid the bounding boxes and extra information. Shuffling was set for the train dataset so that model during training does not go biased. Model was loaded with weights of ImageNet dataset and then trained on KVASIR V2 dataset by utilizing Transfer Learning. Since sigmoid or logistic regression used in binary classification whereas for our multiclass problem SoftMax activation function was used in all three models.

Further we studied the maximum accuracy achieved and evaluation measures of our proposed model Efficient Net V2B2 with two optimizers SGD and Adams after implementation of above techniques. Utilizing Adams optimizer provided the maximum of training accuracy as 98.32% with 4.56% loss, validation accuracy of 93.18% with 25.91% loss while testing accuracy on unseen data was 94.36% with 24.39% loss. **The final proposed model Efficient Net V2B2 with SGD optimizer achieved promising results as training accuracy of 97.03%, with 7.92%loss, validation accuracy 94.03% with 17.69%, while testing accuracy of 95.34% with 16.10% loss.** The detailed results, confusion matrix and evaluation measures are presented in section 4.5.

The metrics we used for assessment include: accuracy, precision, recall, f1-score, and the confusion matrix. Relative to the total number of positive predictions generated by the model, precision is the measure of how many of those forecasts turned out to be accurate. This metric evaluates how well the favorable predictions fared. A high precision specifies that the model is good at correctly identifying true positives and not misclassifying negatives as positives [42].

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, measures how many predictions were correct relative to the total number of positive occurrences in the data. The accuracy with which the model can detect all positive cases is evaluated by this metric. When the recall of a model is high, it means that it successfully captures the majority of positive cases in the data[42].

$$Recall = \frac{TP}{TP + FN}$$

Accuracy and recall are harmonically averaged to get the F1 score. It offers a unified metric that takes precision and recall into account. When evaluating the model's overall performance, the F1 score can be used to account for both false positives and false negatives [42]. It can be expressed as:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The accuracy, on the other hand, tells you how well the model performed overall; it is the proportion of samples that the classifier correctly identified [42]. When calculating precision, the following formula is used to determine it:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

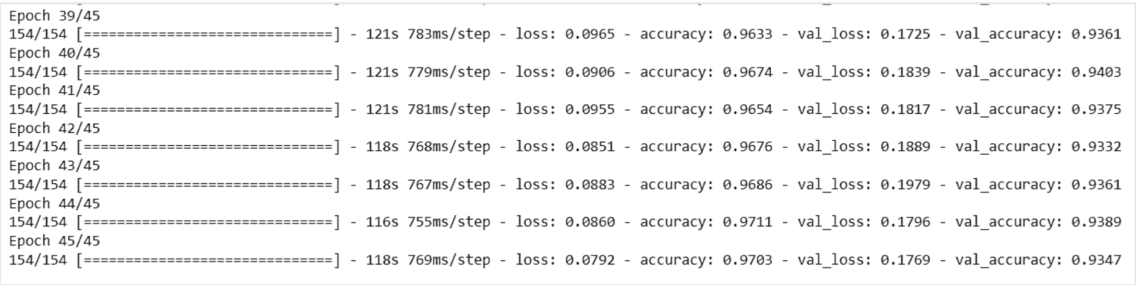
The results of above assessment metrics and confusion matrix is provided in section 4.5

4.4 Model fitness test

In line with our research questions, we thoroughly examined the study's hypotheses and the fitness of our models. We aimed to find the best convolutional neural network (CNN) model for GI endoscopic image classification by conducting extensive experiments with several models, such as VGG16, AlexNet, and EfficientNetV2B2. Based on our research, EfficientNetV2B2 showed remarkable performance, especially after being fine-tuned using the SGD optimizer. Our suggested model demonstrated resilience in picture classification across various gastrointestinal sample types, hitting 97.03% during training and 95.34% during testing. Additionally, F1-score, recall, and precision all pointed to very effective positive case identification. The model's performance in correctly classifying GI images was significantly improved by implementing Transfer Learning, which made use of pre-trained models. In order to tackle the challenges of medical picture classification in the gastrointestinal domain, our suggested model was evaluated using a variety of metrics. These included recall, accuracy, precision, and F1-score; the confusion matrix and ROC curves demonstrated the model's effectiveness and reliability. The detailed results against evaluations are mentioned in the next section.

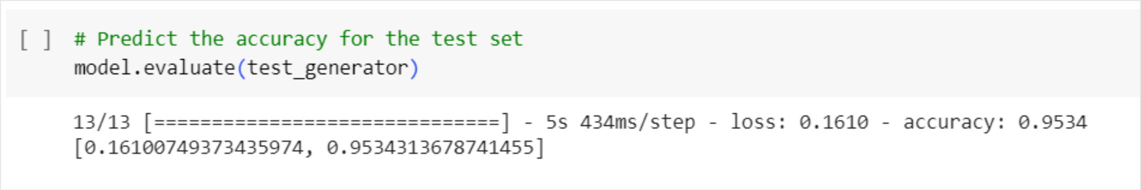
4.5 Result/Findings of hypotheses/experiments

The final proposed model Efficient Net V2B2 with SGD optimizer achieved promising results as training accuracy of 97.03%, with 7.92% loss, validation accuracy 94.03% with 17.69% loss, while testing accuracy of 95.34% with 16.10% loss. The detailed results, Confusion Matrix and evaluation measures are presented below:



```
Epoch 39/45
154/154 [==============================] - 121s 783ms/step - loss: 0.0965 - accuracy: 0.9633 - val_loss: 0.1725 - val_accuracy: 0.9361
Epoch 40/45
154/154 [==============================] - 121s 779ms/step - loss: 0.0906 - accuracy: 0.9674 - val_loss: 0.1839 - val_accuracy: 0.9403
Epoch 41/45
154/154 [==============================] - 121s 781ms/step - loss: 0.0955 - accuracy: 0.9654 - val_loss: 0.1817 - val_accuracy: 0.9375
Epoch 42/45
154/154 [==============================] - 118s 768ms/step - loss: 0.0851 - accuracy: 0.9676 - val_loss: 0.1889 - val_accuracy: 0.9332
Epoch 43/45
154/154 [==============================] - 118s 767ms/step - loss: 0.0883 - accuracy: 0.9686 - val_loss: 0.1979 - val_accuracy: 0.9361
Epoch 44/45
154/154 [==============================] - 116s 755ms/step - loss: 0.0860 - accuracy: 0.9711 - val_loss: 0.1796 - val_accuracy: 0.9389
Epoch 45/45
154/154 [==============================] - 118s 769ms/step - loss: 0.0792 - accuracy: 0.9703 - val_loss: 0.1769 - val_accuracy: 0.9347
```

Figure 4.4: Model training results and Epochs



```
[ ]  # Predict the accuracy for the test set
     model.evaluate(test_generator)

     13/13 [==============================] - 5s 434ms/step - loss: 0.1610 - accuracy: 0.9534
     [0.16100749373435974, 0.9534313678741455]
```

Figure 4.5: Test Dataset Accuracy

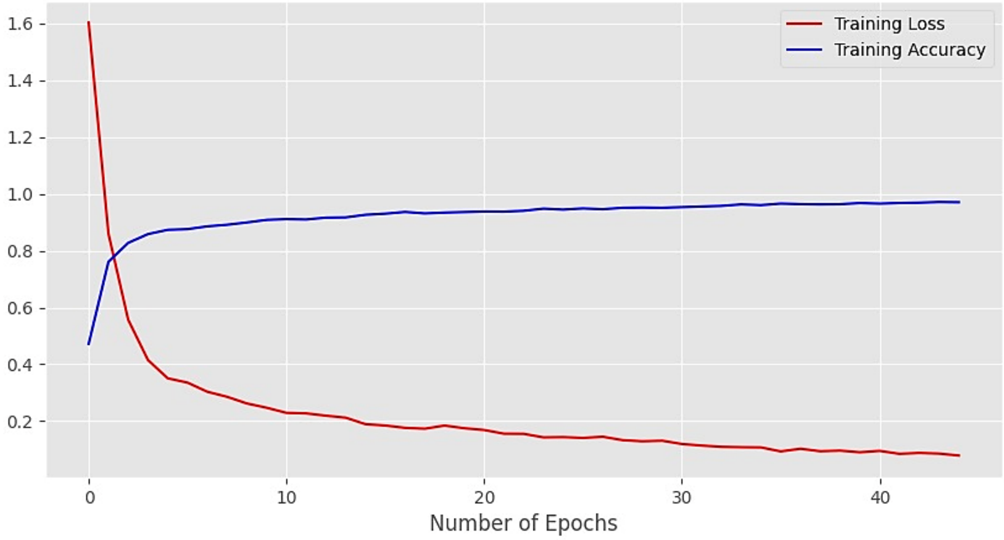


Figure 4.6: Plot of Training Accuracy and Training Loss

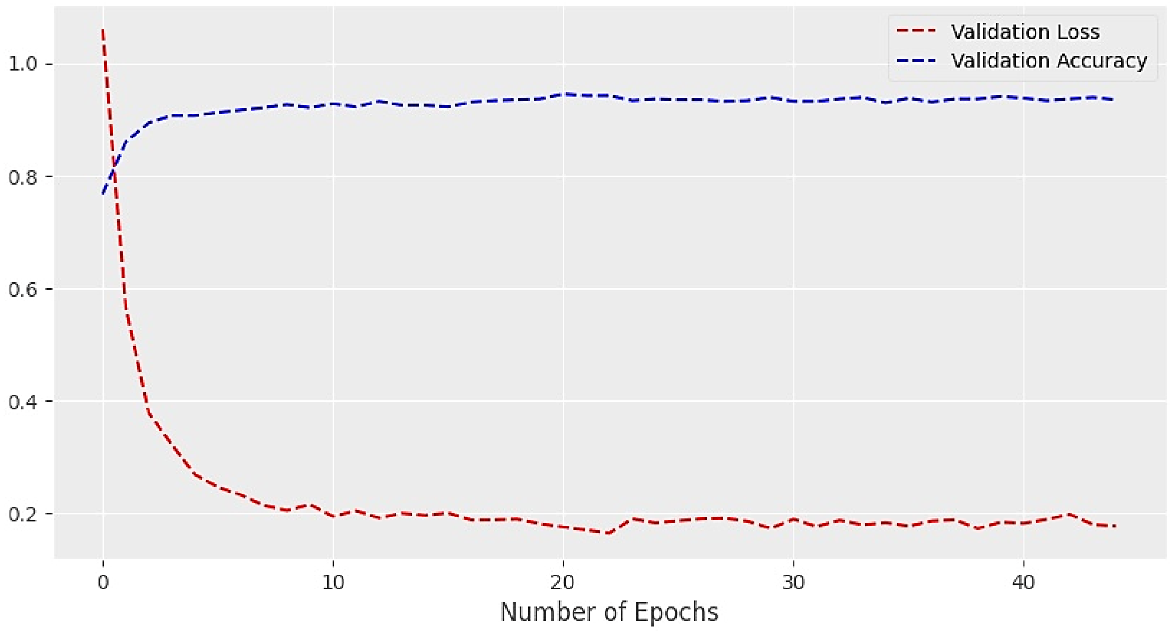


Figure 4.7: Plot of Validation Accuracy and Loss

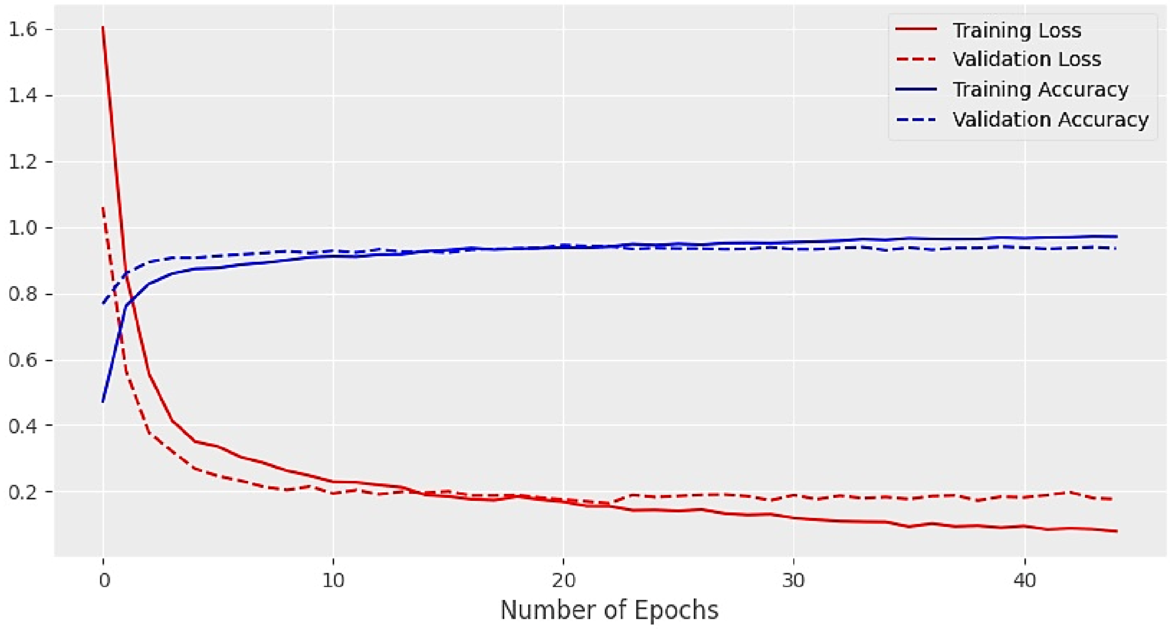
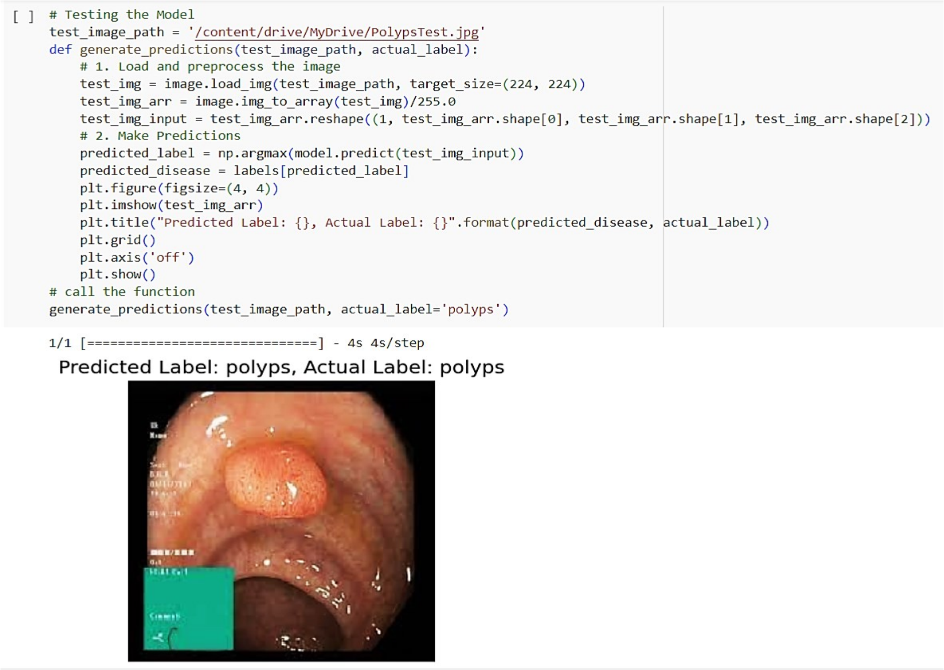


Figure 4.8: Plot of the Accuracy and Error combined

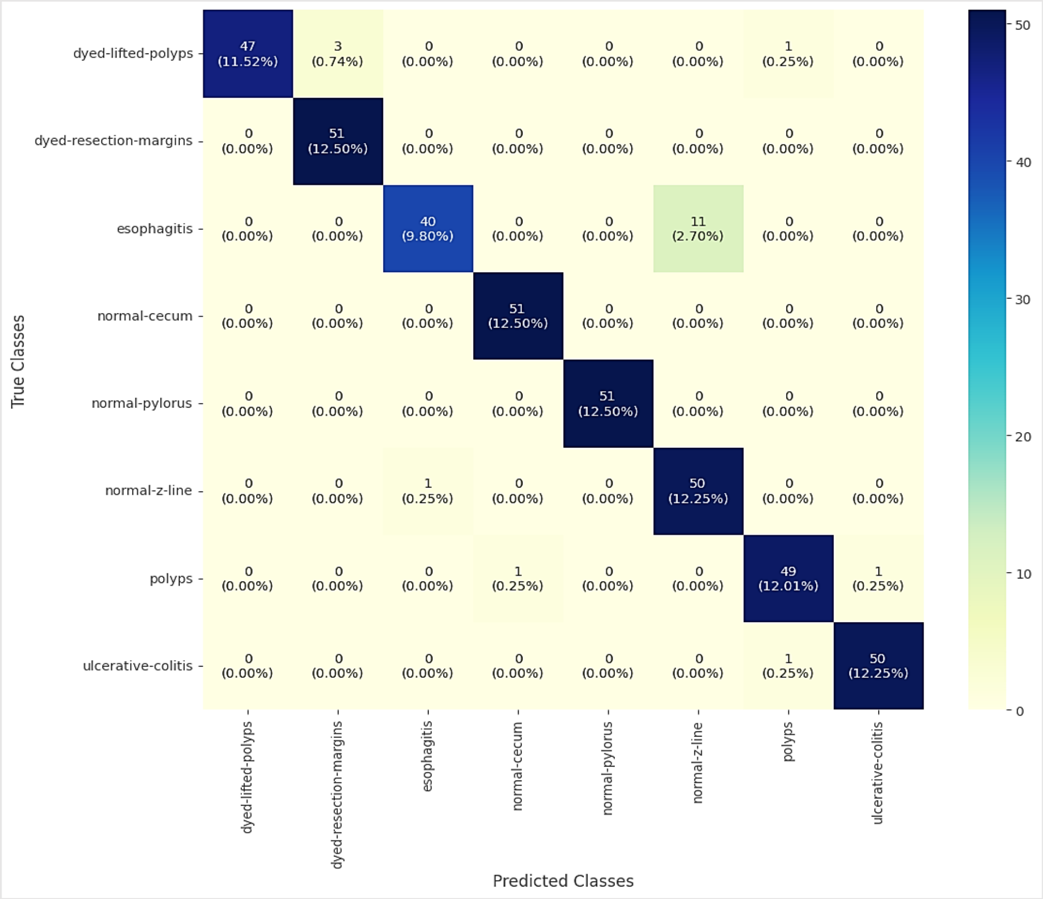


```
[ ]  # Testing the Model
     test_image_path = '/content/drive/MyDrive/PolypsTest.jpg'
     def generate_predictions(test_image_path, actual_label):
         # 1. Load and preprocess the image
         test_img = image.load_img(test_image_path, target_size=(224, 224))
         test_img_arr = image.img_to_array(test_img)/255.0
         test_img_input = test_img_arr.reshape((1, test_img_arr.shape[0], test_img_arr.shape[1], test_img_arr.shape[2]))
         # 2. Make Predictions
         predicted_label = np.argmax(model.predict(test_img_input))
         predicted_disease = labels[predicted_label]
         plt.figure(figsize=(4, 4))
         plt.imshow(test_img_arr)
         plt.title("Predicted Label: {}, Actual Label: {}".format(predicted_disease, actual_label))
         plt.grid()
         plt.axis('off')
         plt.show()
     # call the function
     generate_predictions(test_image_path, actual_label='polyps')
```

```
     1/1 [==============================] - 4s 4s/step
```

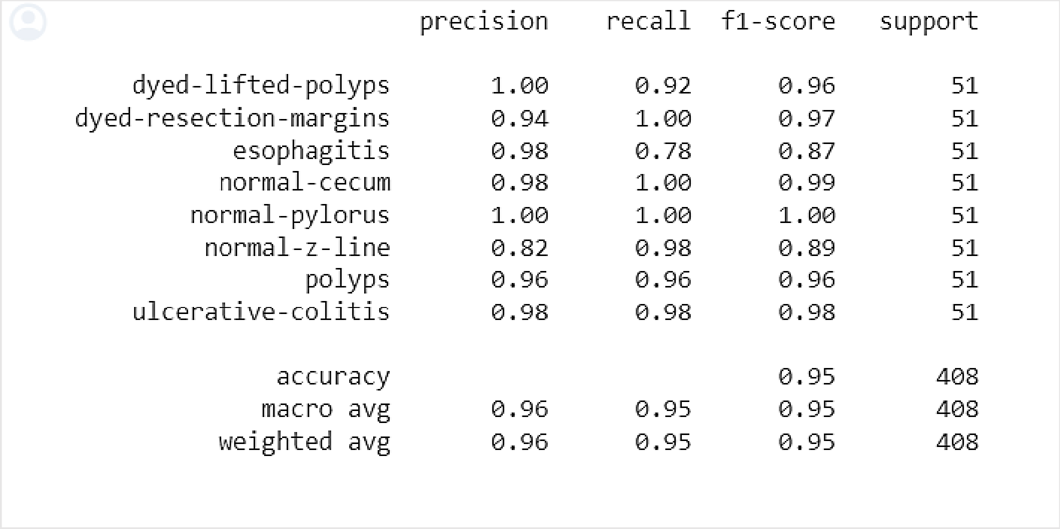
Predicted Label: polyps, Actual Label: polyps

Figure 4.9: Model Testing result on unseen sample



	dyed-lifted-polyps	dyed-resection-margins	esophagitis	normal-cecum	normal-pylorus	normal-z-line	polyps	ulcerative-colitis
dyed-lifted-polyps	47 (11.52%)	3 (0.74%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.25%)	0 (0.00%)
dyed-resection-margins	0 (0.00%)	51 (12.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
esophagitis	0 (0.00%)	0 (0.00%)	40 (9.80%)	0 (0.00%)	0 (0.00%)	11 (2.70%)	0 (0.00%)	0 (0.00%)
normal-cecum	0 (0.00%)	0 (0.00%)	0 (0.00%)	51 (12.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
normal-pylorus	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	51 (12.50%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
normal-z-line	0 (0.00%)	0 (0.00%)	1 (0.25%)	0 (0.00%)	0 (0.00%)	50 (12.25%)	0 (0.00%)	0 (0.00%)
polyps	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.25%)	0 (0.00%)	0 (0.00%)	49 (12.01%)	1 (0.25%)
ulcerative-colitis	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.25%)	50 (12.25%)

Figure 4.10: Confusion Matrix of model: Ratio of actual vs predicted classes



```
                         precision    recall  f1-score   support

     dyed-lifted-polyps       1.00      0.92      0.96        51
dyed-resection-margins       0.94      1.00      0.97        51
            esophagitis       0.98      0.78      0.87        51
           normal-cecum       0.98      1.00      0.99        51
         normal-pylorus       1.00      1.00      1.00        51
          normal-z-line       0.82      0.98      0.89        51
                 polyps       0.96      0.96      0.96        51
     ulcerative-colitis       0.98      0.98      0.98        51

               accuracy                           0.95       408
              macro avg       0.96      0.95      0.95       408
           weighted avg       0.96      0.95      0.95       408
```

Figure 4.11: Results of Classification Metrics: Accuracy, precision, recall and f1-score

To evaluate the efficacy of multi-class classification issues, the AUC-ROC curve is an essential tool. The degree of separability is a measure of the model's capacity to distinguish between classes. Values closer to 1 indicate greater performance, while a higher AUC indicates stronger predictive capability. Similarly, AUC is an important metric for assessing the performance of classification models since it indicates how well the model separates individuals with a condition from those without it [43].

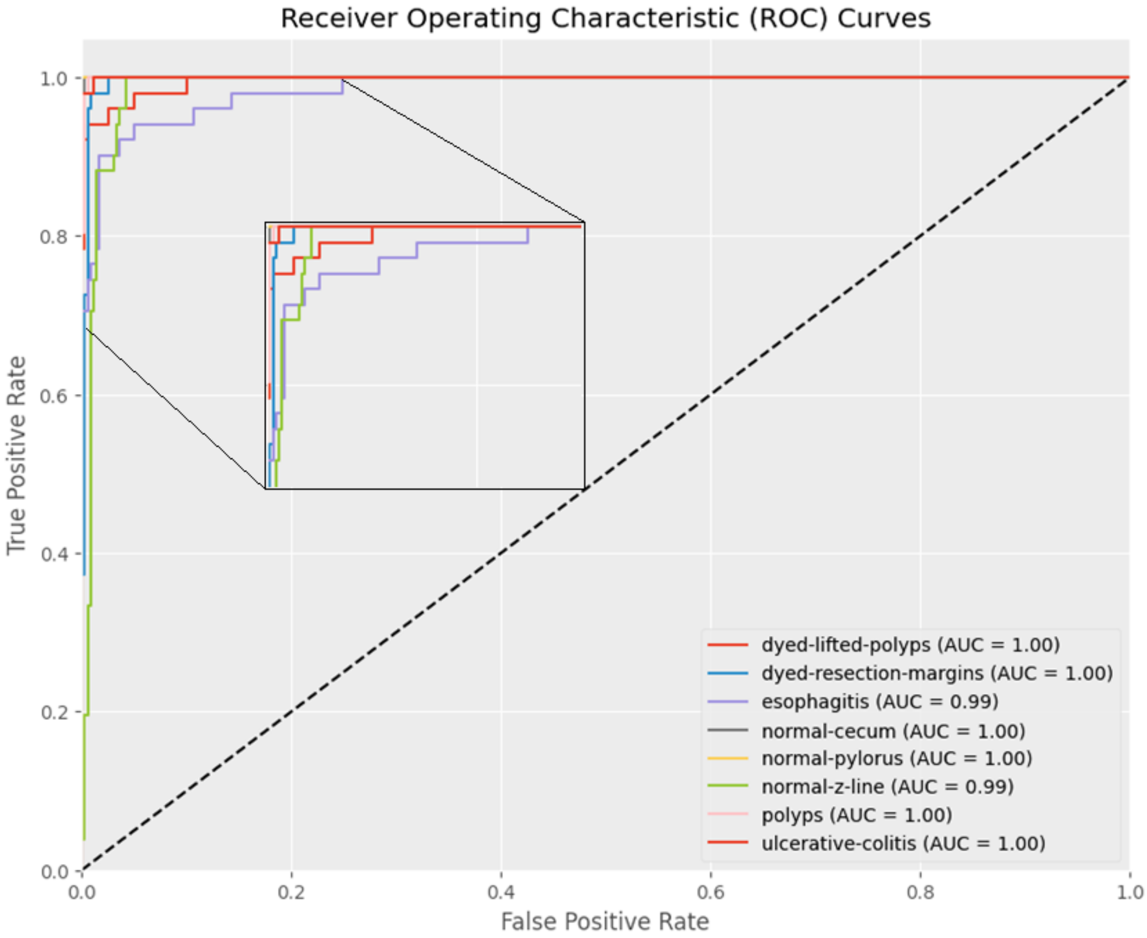


Figure 4.12: Class-wise AUC and ROC Curves

4.6 Results comparison to previous studies

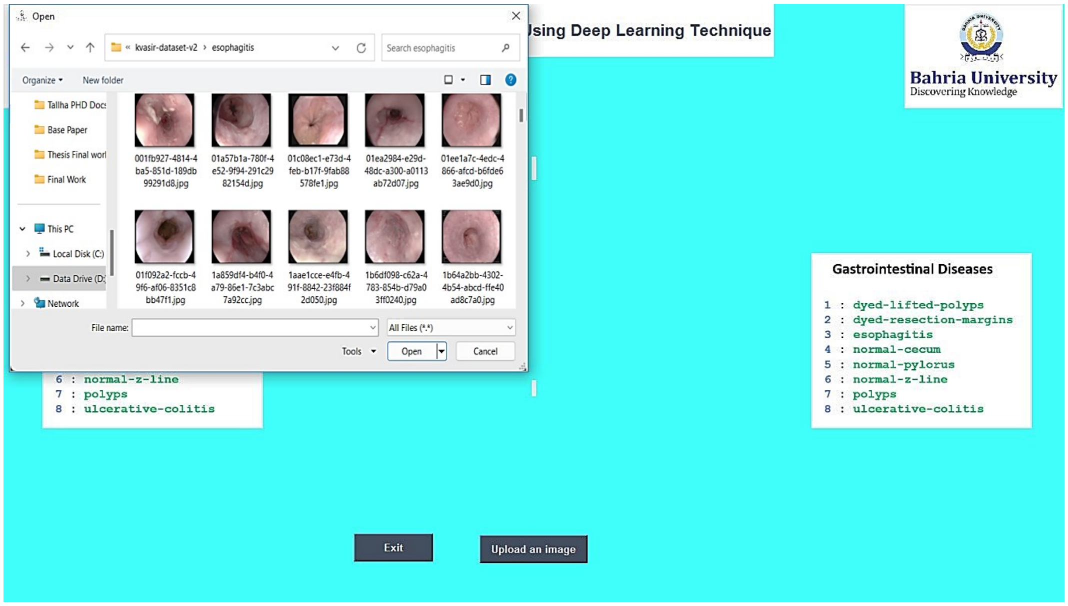
Study [33] achieved the accuracy of 94.80% utilizing pretrained VGG-16, the dataset contained 380 samples of gastroscopic images with 3 classes in total. Study [20] achieved validation accuracy of 93.46% through pretrained model ResNet-152 combined with Grad-Cam, Data augmentation applied and used KVASIR dataset with all 8 classes. Study [23] obtained dataset from local hospital and applied data augmentation, achieved 94.1% accuracy, precision of 91.1%, recall score was 92.4% with total dataset size of 5859 having 5 classes. The model incorporated was DLU-Net. Study [22] achieved promising results obtaining the accuracy of 96.40%, incorporating pretrained MobileNet-

V2 model. Combined Hyper-Kvasir and KVASIR-V2 dataset. Data augmentation applied and total sample size was kept as 4854 with 5 classes. The limitation of the study included the number of classes and chosen dataset size. **While the proposed study achieved promising results in classification of GI Images based on pretrained model Efficient Net V2B2 by achieving training accuracy of 97.03%, validation accuracy 94.03%, while testing accuracy of 95.34% and AUC score obtained as 99.75%, the dataset sample size was 7976 with 8 classes.**

Table 4.2: Performance comparison of proposed model with previous studies

Previous Studies	**Model and Dataset**	**Classes**	**Accuracy**	**Precision**	**Recall**	**F1-Measure**	**AUC**
Wengang Qiu et al. (2021) [23]	DLU-Net with Locally collected Dataset	5	94.1%	91.1%	92.4%	_	_
Ping Xiao et al. (2022) [33]	VGG-16 with Locally collected Dataset	3	94.80%	_	_	_	_
M Nouman Noor et.all(2023) [22]	MobileNet-V2 with Hyper-Kvasir and KVASIR-V2	5	96.40%	97.57%	93.02%	95.24%	_
Doniyorjon Mukhtorov et. all (2023) [20]	ResNet-152 with KVASIR-V2	8	93.46%	_	_	_	_
Proposed Study	**EfficientNet-V2B2 with KVASIR-V2**	**8**	**95.34%**	**96%**	**95.25%**	**95.25%**	**99.75%**

We also build application based on proposed model using Tkinter Python library for image classification purpose. Building this application involves saving the model file, attaching with the library offline by loading it on PyCharm. Tkinter Python library had the ability to create a graphical user interface (GUI) by providing a user-friendly experience for interacting with our medical image classification model. Tkinter further provides a range of widgets and tools for creating windows, buttons, labels, and other elements that can be used to design the application interface. This applications works by allowing us to upload image samples and then use the attached trained model file with the .h5 extension to classify the uploaded images. We loaded the trained model file into the program and utilized it for implication. This file contains the learnt weights and neural network architecture. Features like a file upload button added, a display area for the submitted sample, and an output section to show the results of the classification are added in the application's UI. Furthermore, Tkinter application offered the prospects for us to engage with the classification model. Below is attached the resulting images of application and its working for classification:

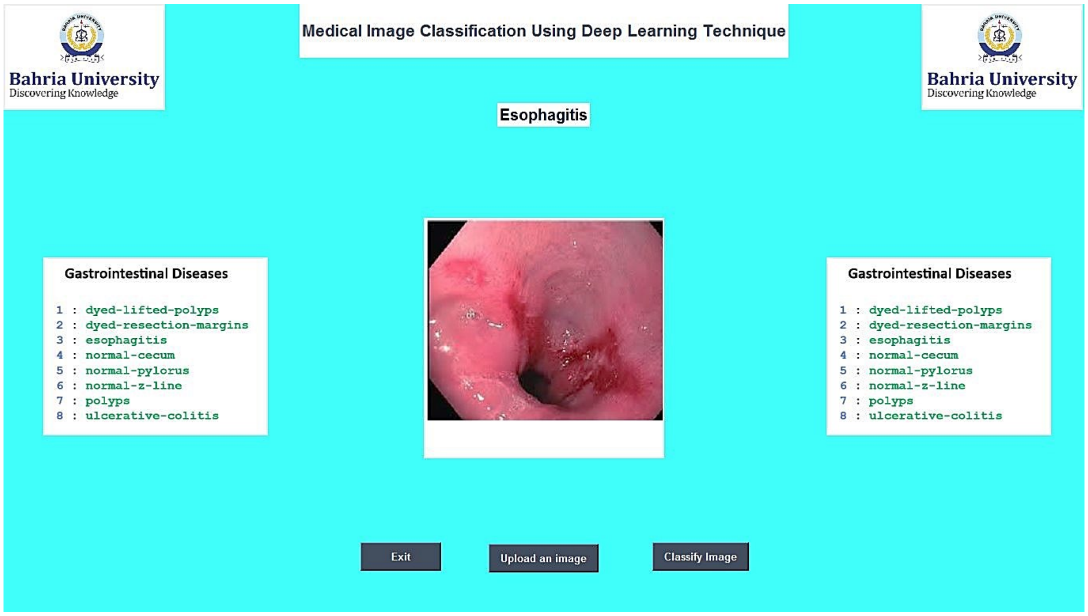

Open
« kvasir-dataset-v2 › esophagitis
Search esophagitis
Organize ▾
New folder
Tallha PHD Docs
Base Paper
Thesis Final worl
Final Work
This PC
Local Disk (C:)
Data Drive (D:
Network
001fb927-4814-4ba5-851d-189db99291d8.jpg
01a57b1a-780f-4e52-9f94-291c2982154d.jpg
01c08ec1-e73d-4feb-b17f-9fab88578fe1.jpg
01ea2984-e29d-48dc-a300-a0113ab72d07.jpg
01ee1a7c-4edc-4866-afcd-b6fde63ae9d0.jpg
01f092a2-fccb-49f6-af06-8351c8bb47f1.jpg
1a859df4-b4f0-4a79-86e1-7c3abc7a92cc.jpg
1aae1cce-e4fb-491f-8842-23f884f2d050.jpg
1b6df098-c62a-4783-854b-d79a03ff0240.jpg
1b64a2bb-4302-4b54-abcd-ffe40ad8c7a0.jpg
File name:
All Files (*.*)
Tools ▾
Open
Cancel
Jsing Deep Learning Technique
Bahria University
Discovering Knowledge
Exit
Upload an image

Gastrointestinal Diseases

```
1 : dyed-lifted-polyps
2 : dyed-resection-margins
3 : esophagitis
4 : normal-cecum
5 : normal-pylorus
6 : normal-z-line
7 : polyps
8 : ulcerative-colitis
```

```
6 : normal-z-line
7 : polyps
8 : ulcerative-colitis
```

Figure 4.13: Application with graphical user interface for proposed model



Medical Image Classification Using Deep Learning Technique

Esophagitis

Gastrointestinal Diseases

```
1 : dyed-lifted-polyps
2 : dyed-resection-margins
3 : esophagitis
4 : normal-cecum
5 : normal-pylorus
6 : normal-z-line
7 : polyps
8 : ulcerative-colitis
```

Gastrointestinal Diseases

```
1 : dyed-lifted-polyps
2 : dyed-resection-margins
3 : esophagitis
4 : normal-cecum
5 : normal-pylorus
6 : normal-z-line
7 : polyps
8 : ulcerative-colitis
```

Figure 4.14: Working of the Application with GUI for proposed model

4.7 Summary of the Chapter

Data analysis, results and findings of the study are covered in Chapter 4, with a particular emphasis on the KVASIR-V2 dataset produced by Norway's Vestre Viken Health Trust. Comparative analysis of results by previous studies discussed. At the very beginning, the chapter provides a descriptive and demographic study of the dataset, drawing attention to the fact that it is diverse in terms of the ages, genders, and ethnic backgrounds of the patients and that it also preserves privacy by not including certain personal information. Preprocessing and dataset visualization done highlighting manual quality-checks, sample balancing, and removal of image samples containing either extreme noise, high distortions, low quality or seems to have low impact to contribute for feature extraction. After introducing the dataset, this chapter moves on to validity and reliability analysis by testing three pretrained models with different optimizers: EfficientNetV2B2, AlexNet, and VGG16. Exploring and fine-tuning the model EfficientNetV2B2 through experimental and factor analysis, we adjusted the hyperparameters and evaluated the model's performance. To wrap off the chapter, we test our hypotheses and find that transfer learning works well for GI image classification. We also show you all the details of our results, including ROC curves, F1-score metrics, recall, accuracy, and precision. Furthermore, we gone through the creation of a Tkinter-based sample categorization application, which enhances user involvement with the suggested model. Results show promise for medical picture categorization using deep learning models, especially EfficientNetV2B2 and SGD optimizer.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Discussion

The purpose of this research was to evaluate the potential of deep learning in healthcare, with a focus on the classification of gastrointestinal (GI) images using the preprocessed 7976 samples from the KVASIR V2 dataset, which has 8 classes. Limitations in the dataset and the amount of samples were common in earlier investigations [20], [22], [33]. At the outset, three convolutional neural network (CNN) models were assessed in this research, two were pretrained on ImageNet Dataset EfficientNetV2B2, and VGG-16, and third model included AlexNet. The following measures were used for evaluation: accuracy, precision, recall, f1-measure, confusion matrix, and category accuracies. The EfficientNetV2B2 model was chosen for additional fine-tuning because it showed greater performance according to these criteria. outcomes from training, validation, and testing using the EfficientNetV2B2 model with the SGD optimizer were 97.03%, 94.03%, and 95.34%, respectively, indicating positive outcomes for the study. In comparison to AlexNet, the use of transfer learning showed substantial advantages, especially in terms of reduced training time and processing resources. Training efficiency and general performance in GI medical image classification were both enhanced by the use of transfer learning utilizing pretrained models.

5.2 Implications of the study

This study investigated the transfer learning procedure and domain adaptation methods to adapt models to new datasets, focusing on interpretability and explain-ability of model predictions. The implications of this study on GI (Gastrointestinal) image classification are far-reaching and have the potential to significantly impact the field of technology and medical. Advanced image classification techniques can lead to improved diagnostic accuracy, aiding in early detection and treatment of medical conditions. By automating image classification, the workload of medical professionals can be reduced. Additionally, the detailed information provided by accurate image classification can contribute to the development of personalized treatment plans. Furthermore, it can support medical research endeavors and accelerate the advancement of new medical technologies. Since the use of AI and Deep Learning in medical image classification becomes more prevalent, there will be a need for updated regulations and ethical guidelines to ensure patient privacy, data security, and the responsible use of technology in healthcare settings.

5.2.1 Practical implications

The practical implicatioons of this study included to study Deep Learning Models to recognize the Gastrointestinal (GI) medical images patterns and features with promising accuracy that task may be very difficult for medical practitioners due to time constraints and performing comparison with large datasets. The use of deep learning algorithms to classify medical images has practical implications that might revolutionize the healthcare system. Early disease diagnosis, with the help of the EfficientNet V2B2 model with the examined hyperparameters, can lead to better patient outcomes and faster interventions, according to the proposed study. Misdiagnosis is a major problem in healthcare, and the Deep Learning model's ability to accurately analyze medical images, especially gastrointestinal (GI) imaging, can help fix that. Improved workflow efficiency, the proposed model study has the capability to automate picture processing, which means faster decision-making, and better patient throughput. The proposed research could also aid in the development of telemedicine by constructing remote diagnoses utilizing this study, which would increase people's access to healthcare. This study achieved a significant results following Efficient Net V2B2 with SGD optimizer achieved promising results as training accuracy of 97.03%, validation accuracy 94.03% , while testing accuracy of 95.34%. Further an application interface with Tkinter library have been built to test and interact with classification process. The application provided graphical interface and simple to use.

5.3 Limitations of the study

This study utilized KVASIR V2 dataset with all of the 8 classes to classify. Further achieved the testing accuracy of 95.34% on unseen data and utilizing transfer learning. In aspects of the limitations we suggest further to test more optimizers like AdaDelta, RMSProp and study further Deep Learning technologies to enhance the validation and testing accuracy. There is also need to implement Ensemble Learning approach by fitting and training two or more models on same dataset and combining the predictions of the models to further test the accuracy.

5.4 Future research directions

It is recommended as a course of action for future studies that exploring further advanced deep learning architectures tailored to handle the complexity of gastrointestinal images, incorporating multimodal approaches to integrate information from different imaging modalities. Training the dataset on KVASIR V2 and testing accuracy on other open source datasets like Gastrovision. The open source Gasrovision dataset have more number of classes but is very limited in number of samples. Further it is suggested to adapt Ensembel Learning approach. This method allows for the combination of predictions from multiple models that have been fitted to the same data. Theoretically, ensemble learning has the potential to outperform individual models. Studying transfer learning at further level required to better and domain adaptation methods to adapt models to new datasets, focusing on interpretability and explainability of model predictions, integrating deep learning models into clinical practice, addressing robustness and generalization challenges, and validating the impact of these models on clinical decision-making and patient outcomes. These research directions aim to advance the field and contribute to clinically impactful and robust solutions for gastrointestinal disease diagnosis and monitoring.

5.5 Summary of Chapter 5 / Conclusion

The AI and Deep Learning evolved into almost every field of life, the maximum limit of technology is being tested and we are going exponentially in the field. Similarly this study was intended to reduce human effort by utilizing Deep Learning technologies in the field of medical by classifying Gastrointestinal (GI) images from an open source Dataset KVASIR V2 with all 8 classes and after preprocessing we utilized 7976 number of samples. In most of the previous studies dataset was found as the main limitation as limited number of samples and classes were used. We carried out initial results of 3 CNN based models out of which two models EfficientNetV2B2 and VGG-16 were pretrained on ImageNet Dataset while the third model was AlexNet. Evaluation metrics reported in this study include accuracy, precision, recall, f1-measure including categorical accuracies and Confusion matrix. Based on above metrics we selected the best model EfficientNetV2B2 for further fine tuning. **The study achieved promising results in classification of GI Images based on pretrained model Efficient Net V2B2 with SGD optimizer by achieving training accuracy of 97.03%, validation accuracy 94.03%, while testing accuracy of 95.34%.** Transfer learning technique was tested by utilizing above two pre-trained models as a foundation, transfer learning proved to be great for substantial reductions in training time and processing resources beside AlexNet. Further utilizing Transfer learning leaded to better training and improved performance in classification of Gastrointestinal (GI) medical images. Lastly an application with GUI interface was built using Tkinter python library to better interact with image classification process.

ANNEXURES

Implementation Code of Final Model “EfficientNetV2B2” For Medical Images Classification Using Deep Learning Technique

```
pip install split-folders

from google.colab import drive
drive.mount('/content/drive')


import splitfolders
input_folder='/content/drive/MyDrive/Colab Notebooks/Thesis/Rev1-kvasir-dataset-v2'
splitfolders.ratio(input_folder, output='dataset', seed=42, ratio=(.85,.10,.05), group_prefix=None)


import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns  #data visualization library based on matplotlib
%matplotlib inline
 #inline allows us to quickly visualize data without opening them in separate window or saving
them as image
import cv2 # to solve computer vision problems
import os
os.environ["TF_CPP_MIN_LOG_LEVEL"] = "2" # os.environ: maps the user's environmental
variables while TP_CPP_MIN_LOG_LEVEL is used to disable or control log about memory and
gpu resources allocation error or messages.
# we can also adjust the verbosity by changing the value of TF_CPP_MIN_LOG_LEVEL:
# 0 = all messages are logged (default behavior)
# 1 = INFO messages are not printed
# 2 = INFO and WARNING messages are not printed
# 3 = INFO, WARNING, and ERROR messages are not printed
import warnings
warnings.filterwarnings('ignore')
from sklearn.metrics import confusion_matrix, classification_report
import tensorflow as tf
from tensorflow.keras.applications import EfficientNetV2B2
from tensorflow.keras.layers import Activation, BatchNormalization, Conv2D, Dense, Dropout,
Flatten, MaxPooling2D
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.optimizers import SGD #Adam is a gradient based optimization algo
which relies on adaptive estimates of lower-order moments
from tensorflow.keras.losses import CategoricalCrossentropy #cross entropy loss is a metric to
measure how well a classification model performs
from tensorflow.keras.regularizers import l2 #L2 is a regularizer that applies a L2 regularization
penalty
from tensorflow.keras.callbacks import ReduceLROnPlateau, EarlyStopping
```

```
train_dataset_path = '/content/dataset/train'
validation_dataset_path = '/content/dataset/val'


IMG_WIDTH = 224
IMG_HEIGHT = 224
BATCH_SIZE = 32


train_datagen = ImageDataGenerator(rescale=1.0/255,
                                   zoom_range=0.2,
                                   width_shift_range=0.2,
                                   height_shift_range=0.2,
                                   fill_mode='nearest')
train_generator = train_datagen.flow_from_directory(train_dataset_path,
                                                    target_size=(IMG_WIDTH, IMG_HEIGHT),
                                                    batch_size=BATCH_SIZE,
                                                    class_mode='categorical',
                                                    shuffle=True)


validation_datagen = ImageDataGenerator(rescale=1.0/255)
validation_generator = validation_datagen.flow_from_directory(validation_dataset_path,
                                                              target_size=(IMG_WIDTH, IMG_HEIGHT),
                                                              batch_size=BATCH_SIZE,
                                                              class_mode='categorical',
                                                              shuffle=False)


import cv2


labels = {value: key for key, value in train_generator.class_indices.items()}

print("Label Mappings for classes present in the training and validation datasets\n")
for key, value in labels.items():
    print(f"{key} : {value}")


fig, ax = plt.subplots(nrows=2, ncols=5, figsize=(15, 12))
idx = 0

for i in range(2):
    for j in range(5):
        label = labels[np.argmax(train_generator[0][1][idx])]
        ax[i, j].set_title(f"{label}")
        ax[i, j].imshow(train_generator[0][0][idx][:, :, :])
        ax[i, j].axis("off")
        idx += 1
```

```
plt.tight_layout()
plt.suptitle("Sample Training Images", fontsize=21)
plt.show()


from keras import Input
from tensorflow.keras import layers , models
from tensorflow.keras import Model


pre_trained_model = EfficientNetV2B2(include_top = True,
                                     include_preprocessing=False,
                                     pooling=None,
                                     weights = 'imagenet',
                                     input_tensor=Input(shape=(224, 224, 3)))
last_layer = pre_trained_model.get_layer('top_dropout')
print('last layer output shape: ', last_layer.output_shape)
last_output = last_layer.output
x = layers.Dense(8, activation='softmax')(last_output)
model = Model(pre_trained_model.input, x)
model.summary()


# Visualizing the model
tf.keras.utils.plot_model(model, show_shapes=True, show_layer_names=True)


# Compile and fit the model
model.compile(optimizer='SGD', loss='categorical_crossentropy', metrics='accuracy')
hist = model.fit(train_generator,
        epochs=45,
        verbose=1, #Verbose=1 means show both progress bar and one line per epoch,
verbose=0 means silent, verbose=2 means one line per epoch i.e. epoch no./total no. of epochs
        validation_data=validation_generator,
        steps_per_epoch = 4956//32, # 4956/32=154, step per epoch argument is used to run
training only on a specific no. of batches from the dataset, it specifies
        # how many training steps the model should run using this Dataset before moving on
to the next epoch.
        # The steps per epoch denote no. of batches to selcted for one epoch, if 154 steps are
selected then the network will train for 154 batches to complete one epoch
        validation_steps = 712//32,
        )


# Plot of the Accuracy and Error combined
h = hist.history
plt.style.use('ggplot')
plt.figure(figsize=(10, 5))
plt.plot(h['loss'], c='red', label='Training Loss')
plt.plot(h['val_loss'], c='red', linestyle='--', label='Validation Loss')
plt.plot(h['accuracy'], c='blue', label='Training Accuracy')
plt.plot(h['val_accuracy'], c='blue', linestyle='--', label='Validation Accuracy')
plt.xlabel("Number of Epochs")
plt.legend(loc='best')
plt.show()


test_dataset_path = '/content/dataset/test'


test_datagen = ImageDataGenerator(rescale=1.0/255)  # the rescale=1./255 will convert the
pixels in range [0,255] to range [0,1]. This process is also called Normalizing the input. Scaling
every images to the same range [0,1] will make images contributes more evenly to the total loss
test_generator = test_datagen.flow_from_directory(test_dataset_path,
                                                   target_size=(IMG_WIDTH, IMG_HEIGHT),
                                                   batch_size=BATCH_SIZE,
                                                   class_mode='categorical',
                                                   shuffle=False)

# Predict the accuracy for the test set
model.evaluate(test_generator)


predictions = model.predict(test_generator)


fig, ax = plt.subplots(nrows=2, ncols=5, figsize=(12, 10))
idx = 0

for i in range(2):
    for j in range(5):
        predicted_label = labels[np.argmax(predictions[idx])]
        ax[i, j].set_title(f"{predicted_label}")
        ax[i, j].imshow(test_generator[0][0][idx])
        ax[i, j].axis("off")
        idx += 1

plt.tight_layout()
plt.suptitle("Test Dataset Predictions", fontsize=20)
plt.show()


from keras.preprocessing import image


# Testing the Model
test_image_path = '/content/drive/MyDrive/PolypsTest.jpg'
def generate_predictions(test_image_path, actual_label):
    # 1. Load and preprocess the image
    test_img = image.load_img(test_image_path, target_size=(224, 224))
    test_img_arr = image.img_to_array(test_img)/255.0
    test_img_input = test_img_arr.reshape((1, test_img_arr.shape[0], test_img_arr.shape[1],
test_img_arr.shape[2]))
    # 2. Make Predictions
    predicted_label = np.argmax(model.predict(test_img_input))
    predicted_disease = labels[predicted_label]
    plt.figure(figsize=(4, 4))
    plt.imshow(test_img_arr)
    plt.title("Predicted Label: {}, Actual Label: {}".format(predicted_disease, actual_label))
    plt.grid()
    plt.axis('off')
    plt.show()
# call the function
generate_predictions(test_image_path, actual_label='polyps')


y_pred = np.argmax(predictions, axis=1)
y_true = test_generator.classes



from sklearn.metrics import confusion_matrix

import sklearn.metrics as metrics


cm = confusion_matrix(y_true, y_pred, normalize=None)

print("Confusion matrix for the test set:")
print(cm)


cf_mtx = confusion_matrix(y_true, y_pred)

group_counts = ["{0:0.0f}".format(value) for value in cf_mtx.flatten()]
group_percentages = ["{0:.2%}".format(value) for value in cf_mtx.flatten()/np.sum(cf_mtx)]
box_labels = [f"{v1}\n({v2})" for v1, v2 in zip(group_counts, group_percentages)]
box_labels = np.asarray(box_labels).reshape(8, 8)

plt.figure(figsize = (12, 10))
sns.heatmap(cf_mtx, xticklabels=labels.values(), yticklabels=labels.values(),
           cmap="YlGnBu", fmt="", annot=box_labels)
plt.xlabel('Predicted Classes')
plt.ylabel('True Classes')
plt.show()


diseases= ['dyed-lifted-polyps', 'dyed-resection-margins', 'esophagitis', 'normal-cecum', 'normal-
pylorus', 'normal-z-line','polyps','ulcerative-colitis']
```

```
classification_metrics = metrics.classification_report(y_true, y_pred, target_names=diseases )
print(classification_metrics)


#Plotting aoc and roc curve

from sklearn.metrics import roc_curve, auc, roc_auc_score
import matplotlib.pyplot as plt

# Calculate ROC and AUC
fpr = {}
tpr = {}
roc_auc = {}

for i in range(len(diseases)):
   fpr[i], tpr[i], _ = roc_curve(test_generator.classes == i, predictions[:, i])
   roc_auc[i] = auc(fpr[i], tpr[i])

# Plot ROC curves
fig, ax = plt.subplots(figsize=(10, 8))
for i in range(len(diseases)):
   ax.plot(fpr[i], tpr[i], label=diseases[i] + ' (AUC = %0.2f)' % roc_auc[i])

ax.plot([0, 1], [0, 1], 'k--')  # Random classifier line
ax.set_xlim([0.0, 1.0])
ax.set_ylim([0.0, 1.05])
ax.set_xlabel('False Positive Rate')
ax.set_ylabel('True Positive Rate')
ax.set_title('Receiver Operating Characteristic (ROC) Curves')
ax.legend(loc='lower right')
plt.show()

# Plotting AUC scores
fig, ax = plt.subplots(figsize=(10, 8))
ax.bar(diseases, [roc_auc[i] for i in range(len(diseases))])
ax.set_xlabel('Disease')
plt.xticks(rotation=45, ha='right')
ax.set_ylabel('AUC Score')
ax.set_title('Area Under the Curve (AUC) Scores')
plt.show()


from tensorflow.keras.models import load_model
model.save('/content/drive/MyDrive/Colab Notebooks/Thesis/full final effnet with sgd 48epoch
224dpi rev1.hdf5')

model.save('/content/drive/MyDrive/Colab Notebooks/Thesis/full final effnet with sgd 48epoch
224dpi rev1.h5')
```

REFERENCES

[1] Al-mekhlaf, Z. G., Senan, E. M., Alshudukhi, J. S., & Mohammed, B. A. (2023). *Hybrid Techniques for Diagnosing Endoscopy Images for Early Detection of Gastrointestinal Disease Based on Fusion Features*. *2023*.

[2] American Cancer Society. (n.d.). What are the risk factors for stomach cancer? What Are the Risk Factors for Stomach Cancer? Retrieved May 2, 2023, from *https://www.cancer.org/cancer/stomach-cancer/causes-risks-prevention/risk*-factors.html

[3] Choi, J., Shin, K., Jung, J., Bae, H. J., Kim, D. H., Byeon, J. S., & Kim, N. (2020). Convolutional neural network technology in endoscopic imaging: Artificial intelligence for endoscopy. *Clinical Endoscopy*, *53*(2), 117–126. https://doi.org/10.5946/ce.2020.054

[4] Daniyal, M., Ahmad, S., Ahmad, M., Asif, H. M., Akram, M., Rehman, S. U., & Sultana, S. (2015). Risk factors and epidemiology of gastric cancer in Pakistan. Asian Pacific Journal of Cancer Prevention, 16(12), 4821–4824. https://doi.org/10.7314/apjcp.2015.16.12.4821

[5] Hirasawa, T., Ikenoyama, Y., Ishioka, M., Namikawa, K., Horiuchi, Y., Nakashima, H., & Fujisaki, J. (2021). Current status and future perspective of artificial intelligence applications in endoscopic diagnosis and management of gastric cancer. *Digestive Endoscopy*, *33*(2), 263–272. https://doi.org/10.1111/den.13890

[6] Hmoud Al-Adhaileh, M., Mohammed Senan, E., Alsaade, F. W., Aldhyani, T. H. H., Alsharif, N., Abdullah Alqarni, A., Uddin, M. I., Alzahrani, M. Y., Alzain, E. D., & Jadhav, M. E. (2021). Deep Learning Algorithms for Detection and Classification of Gastrointestinal Diseases. *Complexity*, *2021*. https://doi.org/10.1155/2021/6170416

[7] Idrees, R., Fatima, S., Abdul-Ghafar, J., Raheem, A., & Ahmad, Z. (2018). Cancer prevalence in Pakistan: Meta-analysis of various published studies to determine variation in cancer figures resulting from marked population heterogeneity in different parts of the country. *World Journal of Surgical Oncology, 16*(1). https://doi.org/10.1186/s12957-018-1429-z

[8] S, R., V.A, S., G.K, K., C.S, H., P, D., G.E, C., & V, S. (2023). GASTROEFFNETV1-CNN based automated detection of gastrointestinal abnormalities from capsule endoscopy images. https://doi.org/10.21203/rs.3.rs-2588671/v1

[9] Islam, M. M., Poly, T. N., Walther, B. A., Lin, M. C., & Li, Y. C. (2021). Artificial intelligence in gastric cancer: Identifying gastric cancer using endoscopic images with convolutional neural network. *Cancers*, *13*(21). https://doi.org/10.3390/cancers13215253

[10] Kailin, J., Xiaotao, J., Jinglin, P., Yi, W., Yuanchen, H., Senhui, W., Shaoyang, L., Kechao, N., Zhihua, Z., Shuling, J., Peng, L., Peiwu, L., & Fengbin, L. (2021). Current Evidence and Future Perspective of Accuracy of Artificial Intelligence Application for Early Gastric Cancer Diagnosis With Endoscopy: A Systematic and Meta-Analysis. *Frontiers in Medicine*, *8*(March), 1–11. https://doi.org/10.3389/fmed.2021.629080

[11] Kaul, D., Raju, H., & Tripathy, B. K. (2022). Deep Learning in Healthcare. In *Studies in Big Data* (Vol. 91). https://doi.org/10.1007/978-3-030-75855-4_6

[12] Lee, J. H., Kim, Y. J., Kim, Y. W., Park, S., Choi, Y. i., Kim, Y. J., Park, D. K., Kim, K. G., & Chung, J. W. (2019). Spotting malignancies from gastric endoscopic images using deep learning. *Surgical Endoscopy*, *33*(11), 3790–3797. https://doi.org/10.1007/s00464-019-06677-2

[13] Lee, S. A., Cho, H. C., & Cho, H. C. (2021). A Novel Approach for Increased Convolutional Neural Network Performance in Gastric-Cancer Classification Using Endoscopic Images. *IEEE Access*, *9*, 51847–51854. https://doi.org/10.1109/ACCESS.2021.3069747

[14] Li, L., Chen, M., Zhou, Y., Wang, J., & Wang, D. (2020). Research of Deep Learning on Gastric Cancer Diagnosis. *2020 Cross Strait Radio Science and Wireless Technology Conference, CSRSWTC 2020 - Proceedings*, 19–21. https://doi.org/10.1109/CSRSWTC50769.2020.9372583

[15] Luo, Q., Yang, H., & Hu, B. (2023). Application of artificial intelligence in the endoscopic diagnosis of early gastric cancer, atrophic gastritis, and Helicobacter pylori infection. *Journal of Digestive Diseases*, *August 2022*, 666–674. https://doi.org/10.1111/1751-2980.13154

[16] Ma, L., Su, X., Ma, L., Gao, X., & Sun, M. (2023). Deep learning for classification and localization of early gastric cancer in endoscopic images. *Biomedical Signal Processing and Control*, *79*(P2), 104200. https://doi.org/10.1016/j.bspc.2022.104200

[17] Malviya, A., Sengar, N., Dutta, M. K., Burget, R., & Myska, V. (2022). Deep Learning Based Gastro Intestinal Disease Analysis Using Wireless Capsule Endoscopy Images. *2022 45th International Conference on Telecommunications and Signal Processing, TSP 2022*, 221–225. https://doi.org/10.1109/TSP55681.2022.9851383

[18] Mohammad, F., & Al-Razgan, M. (2022). Deep Feature Fusion and Optimization-Based Approach for Stomach Disease Classification. *Sensors*, *22*(7), 1–17. https://doi.org/10.3390/s22072801

[19] Mohapatra, S., Kumar Pati, G., Mishra, M., & Swarnkar, T. (2023). Gastrointestinal abnormality detection and classification using empirical wavelet transform and deep convolutional neural network from endoscopic images. *Ain Shams Engineering Journal*, *14*(4), 101942. https://doi.org/10.1016/j.asej.2022.101942

[20] Mukhtorov, D., Rakhmonova, M., Muksimova, S., & Cho, Y.-I. (2023). Endoscopic Image Classification Based on Explainable Deep Learning. *Sensors*, *23*(6), 3176. https://doi.org/10.3390/s23063176

[21] Niu, P. H., Zhao, L. L., Wu, H. L., Zhao, D. B., & Chen, Y. T. (2020). Artificial intelligence in gastric cancer: Application and future perspectives. *World Journal of Gastroenterology*, *26*(36), 5408–5419. https://doi.org/10.3748/wjg.v26.i36.5408

[22] Noor, M. N., Nazir, M., Khan, S. A., & Song, O. (2023). *Pretrained Deep Convolutional Neural Network*.

[23] Qiu, W., Xie, J., Shen, Y., Xu, J., & Liang, J. (2022). Endoscopic image recognition method of gastric cancer based on deep learning model. *Expert Systems*, *39*(3), 1–8. https://doi.org/10.1111/exsy.12758

[24] Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., & Halvorsen, P. (2017). Kvasir. Artifact Digital Object Group. https://doi.org/10.1145/3193289

[25] Ramamurthy, K., George, T. T., Shah, Y., & Sasidhar, P. (2022). A Novel Multi-Feature Fusion Method for Classification of Gastrointestinal Diseases Using Endoscopy Images. *Diagnostics*, *12*(10). https://doi.org/10.3390/diagnostics12102316

[26] Rana, M., & Bhushan, M. (2022). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-022-14305-w

[27] Sedighipour Chafjiri, F., Mohebbian, M. R., Wahid, K. A., & Babyn, P. (2023). Classification of endoscopic image and video frames using distance metric-based learning with interpolated latent features. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-14982-1

[28] Stomach cancer - diagnosis. *Cancer.Net. (2022, September 7). Retrieved May 1, 2023, from https://www.cancer.net/cancer-types/stomach-cancer/diagnosis*

[29] Shibata, T., Teramoto, A., Yamada, H., Ohmiya, N., Saito, K., & Fujita, H. (2020). Automated detection and segmentation of early gastric cancer from endoscopic images using mask R-CNN. *Applied Sciences (Switzerland)*, *10*(11). https://doi.org/10.3390/app10113842

[30] Sivari, E., Bostanci, E., Guzel, M. S., Acici, K., Asuroglu, T., & Ercelebi Ayyildiz, T. (2023). A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models. *Diagnostics*, *13*(4). https://doi.org/10.3390/diagnostics13040720

[31] Tang, S., Yu, X., Cheang, C. F., Liang, Y., Zhao, P., Yu, H. H., & Choi, I. C. (2023). Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. *Computers in Biology and Medicine*, *157*(November 2022), 106723. https://doi.org/10.1016/j.compbiomed.2023.106723

[32] Teramoto, A., Shibata, T., Yamada, H., Hirooka, Y., Saito, K., & Fujita, H. (2022). Detection and Characterization of Gastric Cancer Using Cascade Deep Learning Model in Endoscopic Images. *Diagnostics*, *12*(8), 1–12. https://doi.org/10.3390/diagnostics12081996

[33] Xiao, P., Pan, Y., Cai, F., Tu, H., Liu, J., Yang, X., Liang, H., Zou, X., Yang, L., Duan, J., Xv, L., Feng, L., Liu, Z., Qian, Y., Meng, Y., Du, J., Mei, X., Lou, T., Yin, X., & Tan, Z. (2022). A deep learning based framework for the classification of multi- class capsule gastroscope image in gastroenterologic diagnosis. *Frontiers in Physiology*, *13*(November), 1–9. https://doi.org/10.3389/fphys.2022.1060591

[34] Yue, G., Wei, P., Liu, Y., Luo, Y., Du, J., & Wang, T. (2023). *Automated Endoscopic Image Classification via Deep Neural Network With Class Imbalance Loss*. *72*.

[35] Zhao, X. (2022). Research and application of deep learning in image recognition. *Journal of Physics: Conference Series*, *2425*(1), 994–999. https://doi.org/10.1088/1742-6596/2425/1/012047

[36] Zhao, Y., Hu, B., Wang, Y., Yin, X., Jiang, Y., & Zhu, X. (2022). Identification of gastric cancer with convolutional neural networks: a systematic review. *Multimedia Tools and Applications*, *81*(8), 11717–11736. https://doi.org/10.1007/s11042-022-12258-8

[37] N. Shahriar, "What is Convolutional Neural Network-CNN (deep learning)," Medium, https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5 (accessed Feb. 21, 2024).

[38] N. Micheletti *et al.*, *Generative AI mitigates representation bias using synthetic health data*, Sep. 2023. doi:10.1101/2023.09.26.23296163

[39] X. Du, Y. Sun, Y. Song, H. Sun, and L. Yang, "A comparative study of different CNN models and transfer learning effect for underwater object classification in side-scan sonar images," *Remote Sensing*, vol. 15, no. 3, p. 593, Jan. 2023. doi:10.3390/rs15030593

[40] R. Poojary and A. Pai, "Comparative study of model optimization techniques in fine-tuned CNN Models," *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, Nov. 2019. doi:10.1109/icecta48151.2019.8959681

[41] A. Debnath *et al.*, "A smartphone-based detection system for tomato leaf disease using EFFICIENTNETV2B2 and its explainability with Artificial Intelligence (AI)," *Sensors*, vol. 23, no. 21, p. 8685, Oct. 2023. doi:10.3390/s23218685

[42] H. N. B, "Confusion matrix, accuracy, precision, recall, F1 score," Medium, https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd (accessed Jan. 22, 2024).

[43] S. Abd ElGhany, M. Ramadan Ibraheem, M. Alruwaili, and M. Elmogy, "Diagnosis of various skin cancer lesions based on fine-tuned Resnet50 Deep Network," *Computers, Materials &amp; Continua*, vol. 68, no. 1, pp. 117–135, 2021. doi:10.32604/cmc.2021.016102