



College of Design and Engineering

Department of Industrial Systems Engineering and Management

**Development of Deep Learning Models for Abnormality
Detection in Small Bowel Capsule Endoscopy Images**

Project Advisor:

A/P Cheung Wang Chi

Industrial Supervisor:

Dr. Sean Lam & Dr. Bochao Jiang

Executive Summary

Background and Problem Definition

Endoscopy is a procedure used to examine the interiors of various human organs. A traditional endoscopy involves inserting an endoscope, a long, thin, flexible tube with a camera at the end, into the body. However, endoscopes have limited capabilities and are not long enough to reach the entire digestive tract, in particular the small bowel. Furthermore, operators have to be well trained to use endoscopes without causing harm to the digestive system or stress to the patient. Capsule endoscopy provides a solution to overcome the limitations of traditional endoscopes. This procedure involves patients ingesting a small pill capsule equipped with a camera which captures images as it passes along the digestive tract. However, this procedure comes with its own set of challenges. Capsule Endoscopy (CE) takes about 50,000 pictures per patient, resulting in long and laborious reading times, risk of missing lesions and overall heavy workload for doctors to analyse, potentially discouraging them from using this procedure.

Purpose & Goal of Project

This project will develop deep learning models to accurately detect abnormalities in small bowel capsule endoscopy. These models will serve as a decision support system for doctors, partially automating the diagnosis process by providing a probability prediction for the presence of abnormalities in CE images. The model will also filter out images that cannot be used for diagnosis due to poor bowel prep, which results in large amounts of faecal matter clouding the CE image.

Methodology

The development of deep learning models involves the use of Artificial Intelligence (AI). AI has the capability to perform image recognition, which could be used to classify CE images. In detail, the project we propose is a Diagnosis Assistant Framework which utilises two different AI computer vision models. The first model is the bowel prep model (BPM) that would filter usable and unusable images, and the second model is the abnormality detection model (ADM) which filters the usable images to be photos with abnormalities and photos without abnormalities. Two data sets would be used for the project: the Kvasir-Capsule dataset and the SingHealth dataset. The data is cleaned and partitioned into “Usable” and “Unusable” groups, and the “Usable” group is further partitioned into a “Normal” and “Abnormal” group. The datasets are further split into training and testing sets to benchmark model performance. Different types of model architectures were used for CE image classification, and the results of the models are compared to select the best model.

Results

In this project we find that the use of Convolutional Neural Network (CNN) models with pre-trained layers yields acceptable levels of performance in both the BPM and ADM models. These models utilise transfer learning techniques that leverages on larger pre-trained models and repurposes them for a new task. This method takes advantage of much larger and complex CNNs while saving computational time needed to train them. The models tested are assessed by their F1-scores on the SingHealth validation dataset to strike a balance between precision and recall. Among all the models, the best BPM model is the VGG16 base w/ CNN (98.19%) while the best performing ADM model was the ADM DenseNet121 base w/ CNN (74.6%). When implemented into our Diagnosis Assistant Framework, we find that our models achieve acceptable levels of performance and there is potential for our models to yield immediate benefits for doctors. We recommend demonstrating a proof-of-concept tool to doctors in order

to garner further support while investing more time and expertise into improving the current models and framework.

Skills Acquired

Throughout this project, the team had the opportunity to apply machine learning methods in a real-world application. This includes using image augmentation techniques to pre-process image data, TensorFlow framework to build deep learning models, quality metrics to evaluate model performance, and the implementation of various drivers and packages needed to run models on Linux systems. The team also learnt the inner workings of deep learning convolutional neural networks and its advantages over other models in computer vision applications. Furthermore, project management soft skills were also put into practice such as Gantt Chart project scheduling, scoping of objectives and deliverables, and stakeholder management.

Acknowledgements

We would like to thank our project advisor, Prof Cheung Wang Chi, for his guidance and help throughout the project. Additionally, we would like to thank Dr Bok Shung Hwee for his continuous help in the coordination of different administrative matters in the module. We would also like to thank SingHealth for their guidance in the project; in particular, we would like to thank Dr Sean Lam for his guidance and feedback in the different aspects of data science in the medical research field, as well as Dr Bochao for her guidance in the field of gastroenterology and medical knowledge. We would also like to thank our CELC tutor, Ms Mahaletchumi Sivalingam, for her feedback and help in both our report writing and our presentation. Finally, we would like to thank the department of Industrial Systems Engineering and Management for the support given throughout this project.

Table of Contents

Executive Summary	I
Acknowledgements	III
❏	
2. Problem Definition and Drivers	2
2.1 Endoscopy Procedures and Present Disadvantages	2
2.2 Capsule Endoscopy	2
2.3 Challenges with Capsule Endoscopy Diagnosis	3
3. Approach to Automation in Medical Diagnosis	4
3.1 Artificial Intelligence in Medical Diagnosis	4
3.2 Benefits of AI in Diagnosis	4
3.3 Requirements of Proposed Solution	5
4. Methodology	6
4.1 Literature Review	6
4.1.1 Statistical Approaches to Computer Vision	6
4.1.2 Deep Learning Convolutional Neural Network Model	8
4.2 Diagnosis Assistant Framework	11
4.3 Data Acquisition	12
4.4 Data Preparation	13
4.4.1 Image Data Augmentation	14
4.4.2 Data Flow	15
4.5 Model Architecture	16
4.6 Model Training, Tuning, and Validation	18
5. Comparison of Models	20
6. Results & Benefits	22
6.1 Implementation of Diagnosis Assistant Framework	24
7. Recommendations	26
7.1 Data & Model Limitations	26
7.2 Future Works	27
7.2.1 Abnormality Classification Model	27
7.2.2 Detection of Contiguous Sets of Abnormal Images	28
7.2.3 Model Pipeline and Hyperparameter Tuning	29
9. Conclusion and Reflection	29

References	31
Appendix	35
A. Pathology Descriptions	35
B. BPM & ADM Model Performance	37

List of Figures

Figure 1: Examples of CE Recordings and Diagnosis	3
Figure 2: Unusable vs. Usable Images [13]	7
Figure 3: Conversion of Photos into Numerical RGB Values [14]	7
Figure 4: Accuracy Scores of Training and Testing Set Based on Depth	8
Figure 5: Convolution of an Input Data into an Output Feature Map [15]	10
Figure 6: General Structure of a Convolutional Neural Network [16]	10
Figure 7: Proposed Diagnosis Assistant Framework	11
Figure 8: Image Augmentation Methods for Resampling	14
Figure 9: Illustrated discrepancies in colour between datasets	15
Figure 10: Overall Data Preparation & Model Training Flowchart	16
Figure 11: Transfer Learning Model Architecture	17
Figure 12: Confusion Matrix for BPM and ADM models on Singhealth Validation Set	22
Figure 13: Lift and Gain Charts for ADM model	23
Figure 14: Interface to Upload CE Images	24
Figure 15: Results of Analysis from the BPM and ADM models	25
Figure 16: Threshold Slider and Summary of Probabilities Line Graph	26
Figure 17: Detection of Abnormalities in Sequences of Images	28

List of Tables

Table 1: Requirements and Benefits of Proposed Solution	6
Table 2: : Breakdown of Images and Labelled in Datasets	12
Table 3: Breakdown of Number of Images in each partition and split	16
Table 4: Summary of Top Performing Models based on F1-Score	20

1. Background of SingHealth

Singapore Health Services (SingHealth) is the largest organisation of healthcare institutions that offer healthcare that is accessible, inexpensive, and of high quality. It provides comprehensive, interdisciplinary, and integrated care with over 40 clinical specialisations, a network of acute hospitals, national specialty centres, polyclinics, and community hospitals. SingHealth collaborates with Duke-NUS Medical School as a component of the SingHealth Duke-NUS Academic Medical Centre to develop medical research and education to enhance patient care [1]. SingHealth has a common purpose, ‘Patients. At the Heart of All We Do’. Institutions under the purview of SingHealth range from primary to acute care, but the key principle of putting patients first remains; ensure patients are well-supported when they transition from one care environment to another [1].

As part of SingHealth’s efforts to develop medical research and improve patient care, our group worked closely with the Health Services Research Centre, together with Gastroenterology (GE) specialist Dr Bochoa. This venture aims to explore the use of artificial intelligence models as clinical decision support systems, serving as an assistant to clinical practices. Specifically, we are aiming to improve the process of diagnosing abnormalities in a patients’ small bowel capsule endoscopy images.

2. Problem Definition and Drivers

2.1 Endoscopy Procedures and Present Disadvantages

An endoscopy is a nonsurgical procedure to examine the digestive tract. Most endoscopy procedures use an instrument called an endoscope, a long, thin, flexible tube with a camera at the end. The endoscope provides high-resolution images of the throat, oesophagus, stomach, colon, and rectum [2]. Doctors can pass special surgical tools through the endoscope (to collect tissue samples or to remove a polyp) while examining the digestive tract of a patient [3].

However, there are multiple disadvantages of endoscopy that cannot currently be overcome. The traditional endoscope is limited to 600 cm, whereas the total length of the human digestive system is 900 cm which requires an additional rectal endoscopy to examine the whole system [4]. Tears may be caused during the process and lead to further inflection and bleeding. More negative effects such as fever, chest pain and difficulties in swallowing may occur among patients [5]. Long term training is required for operators to become skilful and observant, as well as avoid injuries to patients. Moreover, research has shown patients who undergo endoscopy experience are usually more stressed compared to receiving other operations [6]. Unnecessary general anaesthesia is applied to patients to avoid shameful feelings and embarrassment.

2.2 Capsule Endoscopy

Capsule endoscopy (CE) is a relatively new diagnostic tool that has become increasingly popular in recent years due to its many advantages over traditional endoscopy. It has a non-invasive procedure in which the possibility of causing additional injuries is prevented. Moreover, it provides a more comprehensive examination which has better visualisation and is not limited to its length compared to traditional endoscopes. [7] This technique involves

patients ingesting a small pill capsule equipped with a camera. The capsule captures images as it passes along the digestive tract, which are recorded using a device worn around the patient's waist. The CE procedure can transmit up to 50,000 to 60,000 photos of the digestive tract during the trip and has a higher diagnostic yield (71% vs 31%) compared to traditional process [8]. The CE procedure captures images of the pill capsule's journey throughout the patient's small bowel. These images are then viewed by doctors who diagnose the patient by spotting any abnormalities along the small bowel. We present a few examples of the CE images and abnormalities recorded using CE:



Figure 1: Examples of CE Recordings and Diagnosis

A comprehensive list of CE image pathologies can be found in Appendix A.

2.3 Challenges with Capsule Endoscopy Diagnosis

Throughout the diagnosis, CE takes about 50,000 pictures per patient. However, this creates laborious reading times for doctors who examine these pictures. Doctors have to spot images abnormalities and diagnose what the specific abnormality. According to observations of Dr Bochao, the Gastroenterology (GE) specialist from SingHealth. It takes 2-3 hours for junior doctors to detect these abnormalities carefully. Even more experienced senior doctors need 45 minutes to view all the images. The efficiency of image reading can be further affected by the doctor's fatigue and willingness. Hence, this brings the need for a model which can support doctors in identifying different abnormalities accurately and liberate them from repetitive work.

3. Approach to Automation in Medical Diagnosis

3.1 Artificial Intelligence in Medical Diagnosis

Artificial intelligence (AI) techniques, such as computer vision and deep learning, are able to process large volumes of various data types including imagery. Since AI is widely used in image classifications, it can help healthcare providers with disease diagnosis, patient risk detection, and drug subscriptions [9]. With increasing availability of data and rapid development in technology, AI-assisted decision support systems for endoscopy are becoming a possible solution to assisting GE specialists in automating the diagnosis of abnormalities in capsule images. Several newly developed AI-assisted colonoscopy applications have demonstrated impressive outcomes for the detection and categorization of colorectal polyps and adenomas. However, due to constraints in the design, validation, and testing of AI models under real-life clinical situations, their relevance for these applications in clinical practice has yet to be fully verified [10].

3.2 Benefits of AI in Diagnosis

The use of AI in medical diagnosis has incredible potential and benefits in the healthcare system. A study found that a sufficiently capable machine learning algorithm was able to analyse the root cause of patient illnesses, outscoring over 70% of doctors in a written test [11]. The potential of AI to outperform doctors could also mean that appropriate diagnosis by these AI models can be trusted and provided to patients in regions where expert doctors are not available [11].

Furthermore, these tools are not prone to human biases and other effects such as fatigue, which might impact a doctor's performance in diagnosing a patient's ailments. AI models could also help partially automate the diagnosis process, helping to diagnose obvious symptoms while

leaving more difficult diagnosis to doctors. One study published in the journal Nature Medicine found that an AI algorithm was able to accurately diagnose common childhood illnesses based on symptoms reported by parents, outperforming primary care physicians in certain scenarios [12]. This demonstrates the potential for AI models to assist in the diagnosis of common ailments, reducing the need for doctors to intervene in such cases. This further frees up precious time for doctors to focus on more complex cases, increasing the capacity for more patients in the healthcare system.

3.3 Requirements of Proposed Solution

To implement an AI solution for small bowel endoscopy diagnosis, we propose an AI-assistant computer vision model to partially automate the diagnosis process, serving as a decision support system for GE specialists. The overarching goal is to reduce the attention capacity and time required for GE specialists to diagnose all the images collected from a small bowel capsule endoscopy test. The proposed computer vision assistant will act alongside GE specialists, providing an independent diagnosis in confidence for the presence of abnormalities in each image from the small bowel. The computer vision assistant will be able to process a full set of images from the capsule endoscopy procedure, filter out images that cannot be used for diagnosis due to poor bowel prep, and identify images where an abnormality can be found. The specific details of each requirement for this framework are laid out in Table 1:

S/N	Requirement	Benefit
1	Ability to process up to a full set of capsule endoscopy images (50,000 images) in a reasonable amount of time.	Enable GE specialists to process images without having to individually read and diagnose each image. This will reduce the attention capacity and time needed to diagnose patient images.

<p>2 In instances of poor bowel prep, some images may be clouded by faecal matter and cannot be used for diagnosis. Model shall remove images that are deemed unusable for diagnosis.</p>	<p>Model should be able to filter out these images to reduce the number of images GE specialists will have to read, thereby reducing overall diagnosis time.</p>
<p>3 Model shall accurately detect and highlight when an abnormality is present in the image. This includes blood, polyps, ulcers, erosions, and other defined abnormalities.</p>	<p>Assist GE specialists in the reading of images so they will only need to diagnose what type of abnormality. This reduces the attention capacity needed for diagnosis, leaving the GE specialists with more capacity for more challenging diagnosis.</p>

Table 1: Requirements and Benefits of Proposed Solution

4. Methodology

This section outlines the literatures and methodologies used in computer vision applications. Based on our literature review findings, we will present our data preparation process, model specification, model training, and model selection process.

4.1 Literature Review

Before jumping into deep learning model solutions, our team took the time to evaluate past efforts with computer vision models. Given that this project is a new venture by SingHealth and the team, we reviewed and studied various approaches towards image classification with AI-based computer vision models.

4.1.1 Statistical Approaches to Computer Vision

In a preliminary study, our group built a baseline model based on a Random Forest Classifier. This is a statistical approach that utilises decision trees to make a prediction. Our baseline model attempts to fulfil the 1st and 2nd requirements to filter out unusable images due to poor bowel prep. To do this, the model should be able to identify images with large amounts of

yellow faecal matter and categorise them as “Unusable”. An example of a “Usable” and “Unusable” image from the CE process is illustrated in Figure 2:

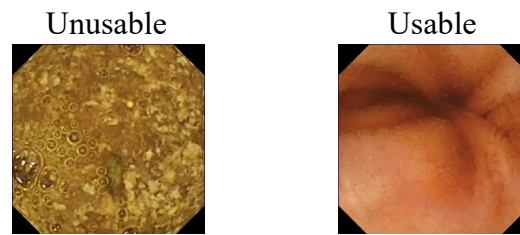


Figure 2: Unusable vs. Usable Images [13]

Data pre-processing of images for computer vision applications involves converting the photos into matrices of numerical data, which is done by representing every pixel in the photo as its Red-Green-Blue (RGB) values. As shown in Figure 3 below, a photo of 3 by 3 pixels can be converted to 3 matrices and 27 numbers ($3 \times 3 \times 3$). Similarly for the endoscopy photos, as their size is 336 by 336 pixels, each photo can be converted to a flatten array of $336 \times 336 \times 3 = 338,688$ values representing each pixel and colour.

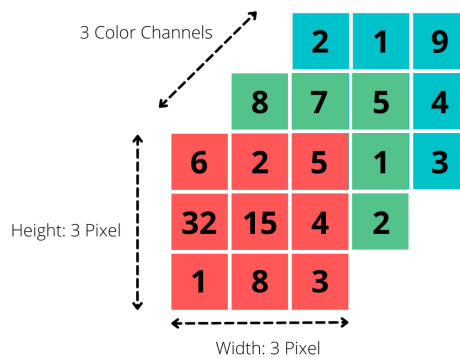


Figure 3: Conversion of Photos into Numerical RGB Values [14]

The values are then normalised so that their variance and averages are similar, ensuring that the data is not skewed. The random forest classifier would create multiple decision trees based on their ability to classify the training data accurately, and the depth of the trees can be controlled. The final classification of the test photos is based on an ensembled vote by all the decision trees. We expect the model to classify images based on the amount of yellow pixels found in each image (representative of faecal matter). This can be derived from the RGB values, from which the model will have to “learn” to identify yellow.

In the classification of the endoscopy images, many different forests with different depths of decision trees are made. Each of their results is analysed based on the 5-fold cross validation accuracy score. The optimal depth of the random forest balances complexity and generalizability. This prevents the model from overfitting to a dataset while ensuring that the model is complex enough to accurately classify images.

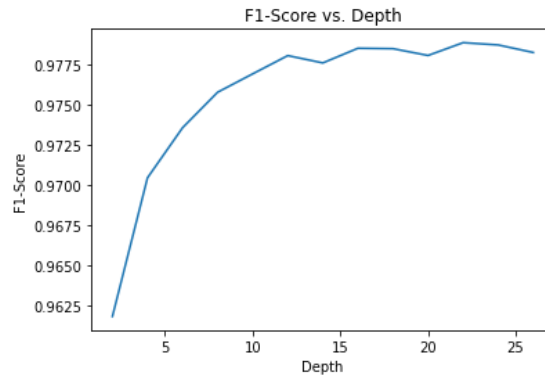


Figure 4: Accuracy Scores of Training and Testing Set Based on Depth

From our experiments with different depths, we find the optimal depth to be 12 for the highest cross-validation F1-score. The random forest generator can be generated relatively quickly, and the complexity can be easily tuned based on its depth. The results are also more interpretable. However, random forest models are not well optimised for high dimensional data and consume huge amounts of memory when implemented. To overcome this, images must be compressed to smaller pixel dimensions in the data pre-processing stage, leading to information loss along the way. In a balanced out-of-sample validation set, we find that it performs relatively poorly with an F1-score of 10.89%. As such, there is a need for a more complex and robust model with better image processing properties and memory optimization.

4.1.2 Deep Learning Convolutional Neural Network Model

Early classical approaches to computer vision models, such as the Random Forest model, relied on raw image pixel data to train and recognize images. While this approach may have been sufficient for simple image datasets, the model is not complex enough to capture the wide

variations in image data. For example, the position of the subject, lighting of the image, and the angle towards the subject of the image created large variations in the interpretability of the subject.

To overcome this variation, complex feature engineering techniques such as edge and contour detection can be employed to better highlight important features in each image. However, due to the complexity of image data, feature engineering is incredibly tedious and engineered features have a high probability of being biased towards the modeller and dataset. In order to overcome this, a more complex deep learning computer vision model is required to generate representative features of images without having to explicitly define and engineer them.

Deep learning neural network models present a feasible solution to generate and “learn” representative features from images. Unfortunately, this approach does not scale well due to the high dimensional nature of image data. As the number of pixels and colour channels increases in each image, the number of neural network model parameters would grow exponentially. This effect would result in an overly complex model that would likely overfit to the training data and have poor generalizability. In contrast, a Convolutional Neural Network (CNN) would be a more reasonable approach. These models mimic the way we process images with our eyes, dividing the image into sub-sections and analysing them individually. In the CNN, convolutions take subsets of the input image data, apply filters to them, and generate a new feature which contributes to the model’s feature map. The feature map highlights important attributes in the image that the model uses to process, interpret the subject, and classify the image.

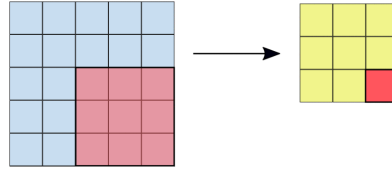


Figure 5: Convolution of an Input Data into an Output Feature Map [15]

The CNN model can be constructed using the Tensorflow Python package. In general, CNNs have a few key features and layers to process and classify images:

1. **Convolution Layers:** As described, this layer “learns” to extract import features from the input image data relevant to interpreting the subject of the image.
2. **Pooling Layers:** The convolved features are downsampled in the pooling layer to reduce the dimensions of the feature map, while preserving the most important feature information. This operation also saves computation and processing time.
3. **Fully Connected Layer:** These are the last layers of the CNN which perform classification based on the features extracted from the convolutions.

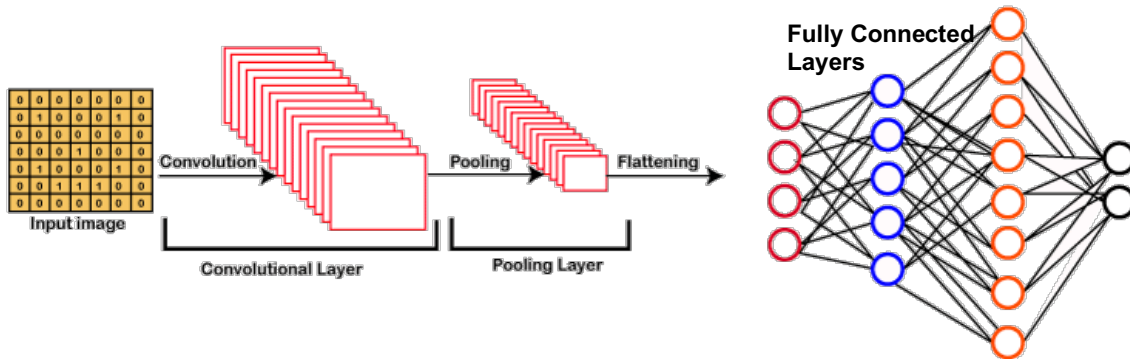


Figure 6: General Structure of a Convolutional Neural Network [16]

Typically, training a CNN model from scratch requires large amounts of labelled training data and computational time. Complex CNN models can take a few days to weeks to complete training [14]. Fortunately, pre-trained CNN models such as the VGG16[17] model can be used to reduce the amount of training required. These models have been previously trained on large datasets for large-scale image classification tasks, such the ImageNet dataset, with 1000 possible class categories [18]. We can utilise transfer learning techniques to leverage on the generic capabilities of these models, taking advantage of the previously trained weights and

repurposing them as feature extraction models for our small bowel CE images. This allows us to take advantage of much more complex models without having to train models from scratch, ultimately saving us computational time [19].

4.2 Diagnosis Assistant Framework

Based on the diagnosis requirements, we will build a small bowel capsule endoscopy diagnosis framework utilising Convolutional Neural Network (CNN) computer vision models for image classification. The project will have 2 focuses, identifying images that can be used for diagnosis and identifying abnormalities amongst the usable images. This will involve 2 binary CNN classification models, laid out in the following framework:

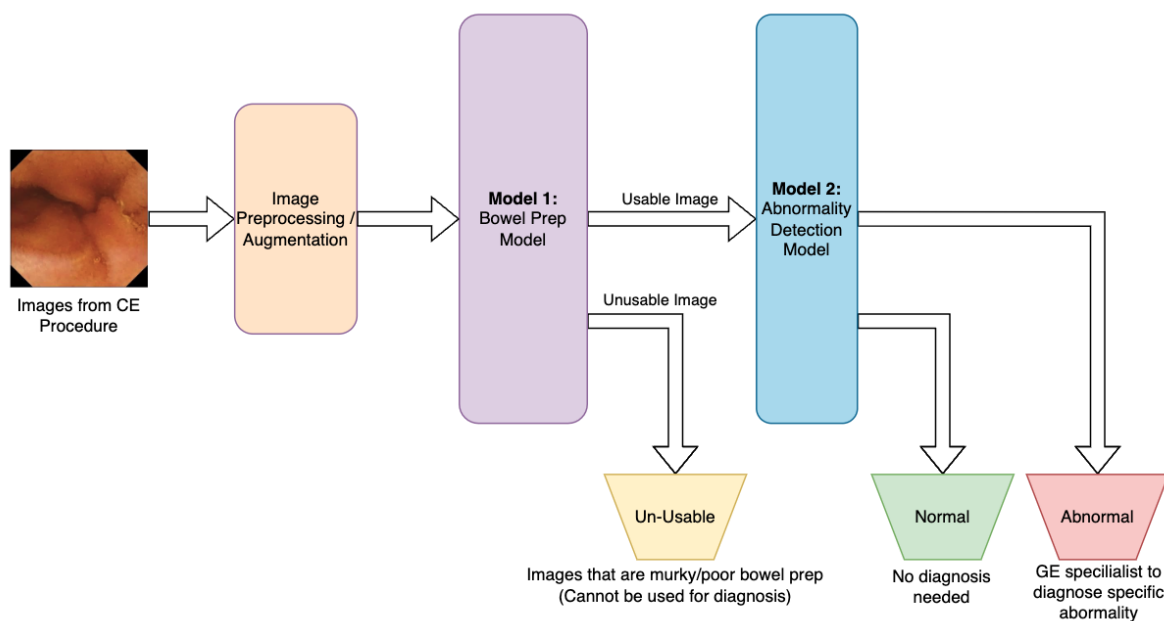


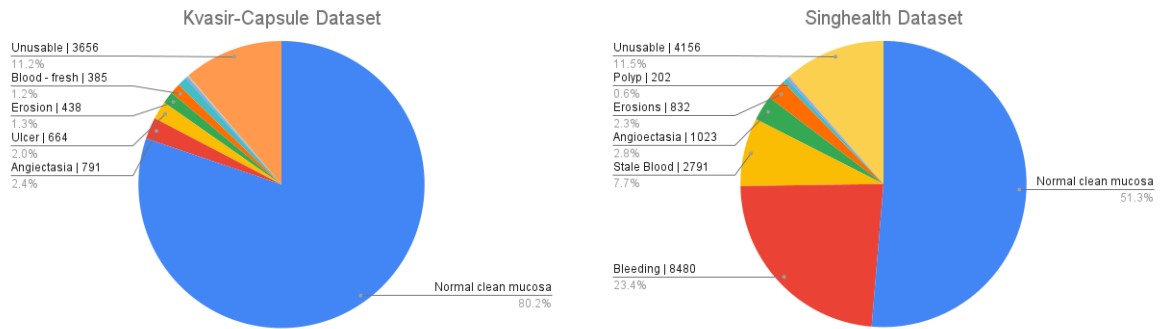
Figure 7: Proposed Diagnosis Assistant Framework

Every image from the capsule feed will pass through the 2 models and will be labelled based on the model’s classification outcome. Images that are deemed unusable by the Bowel Prep Model (BPM) will be filtered out, removing the need for the GE specialist to read those images. The usable images will then be used for abnormality detection, where the images with abnormalities will be highlighted by the Abnormality Detection Model (ADM). The GE specialist will then only have to read the images labelled as “Abnormal” to diagnose the

specific abnormality detected. Ultimately this reduces the amount of attention capacity and time needed to diagnose a full set of small bowel images for a patient.

4.3 Data Acquisition

Image data from past CE procedures will be utilised to train our CNN models. The images are acquired from two sources: the Kvasir-Capsule dataset [13] and SingHealth dataset with 32,524 and 36,171 images respectively. The Kvasir-Capsule dataset is an open-sourced dataset collected from examinations at a Norwegian hospital while the SingHealth dataset was collected from 29 patients in Singapore General Hospital. Each dataset was manually vetted by SingHealth’s GE specialist, Dr. Jiang Bochao, and labels were verified for accuracy. The breakdown of labelled images can be found in Table 2:



Label	Image Count	Label	Image Count
Normal clean mucosa	26087	Normal clean mucosa	18564
<i>Angiectasia</i>	791	<i>Angioectasia</i>	1023
<i>Blood - fresh</i>	385	<i>Bleeding</i>	8480
<i>Blood - hematin</i>	25	<i>Stale Blood</i>	2791
<i>Blood - stale clotted</i>	24	<i>Erosions</i>	832
<i>Erosion</i>	438	<i>Lymphangiectasia</i>	29
<i>Erythema</i>	82	<i>Polyp</i>	202
<i>Lymphangiectasia</i>	350	<i>Ulcers</i>	94
<i>Polyp</i>	22	Unusable	4156
<i>Ulcer</i>	664	Total: 36171	
Unusable	3656		
Total: 32524			

Table 2: Breakdown of Images and Labelled in Datasets

For the purposes of this study, all images that are not “Normal clean mucosa” (“Normal”) or “Unusable” will be considered labelled as “Abnormal”. Images from the Kvasir-Capsule dataset were extracted with size 336x336 pixels while images from the SingHealth dataset were 400x400 pixels. Both datasets are in JPEG format. Since images from the Kvasir-Capsule dataset are open-sourced, images can be uploaded and processed online. In contrast, the SingHealth dataset images are governed by SingHealth’s privacy policy and can only be accessed locally (offline) on SingHealth’s workstations.

4.4 Data Preparation

With the image labels manually verified, both datasets were partitioned into training data and validation data. Images in each dataset are partitioned into “Usable” vs. “Unusable” labels and “Normal” vs. “Abnormal” labels to train and validate the BPM and ADM models respectively. In addition to partitioning by label, each dataset is split into training and testing sets, where the same patients are kept within each training or testing set. This is to prevent leakage on a patient level across the training and testing data where models might classify based on similar looking bowel structure instead of actual abnormalities in the image features. While patient level data is available on the SingHealth dataset, it had to be inferred on the Kvasir-Capsule dataset. We assume patient source by utilising the image file names in the Kvasir-Capsule dataset, where sets of images with contiguous file names were assumed to be source from a unique patient. The Kvasir-Capsule dataset is split into 80% for training and 20% for testing for both “Usable” vs. “Unusable” and “Normal” vs. “Abnormal” labels. We also ensure that both the training and testing set for “Normal” vs. “Abnormal” both contain all the abnormalities listed in Table 2. The SingHealth Dataset is similarly split 50% for tuning and 50% for validation for both “Usable” vs. “Unusable” and “Normal” vs. “Abnormal” labels.

4.4.1 Image Data Augmentation

For the Kvasir-Capsule dataset, we find an imbalance in class labels, where class labels are skewed towards “Usable” and “Normal” for both partitions. The number of images in these majority classes outnumber their respective minority classes with a ratio of 10:1 for both partitions. Utilising this imbalance datasets to train the BPM and ADM models would result in models biased towards the majority classes [20], resulting in models with low recall. To overcome this, we utilise image augmentation to resample images in the minority class, thereby creating new synthetic data to train both BPM and ADM models. Augmentation methods implemented include horizontal flip, vertical flip, random rotations, and random colour jitters utilising the Albumentations [21] Python package. Using a randomised combination of these methods, we create “new” unseen image data that the BPM and ADM models can utilise to learn the features required for their respective classification tasks. Figure 8 shows examples of “new” synthetic image data generated using the predefined augmentation methods:

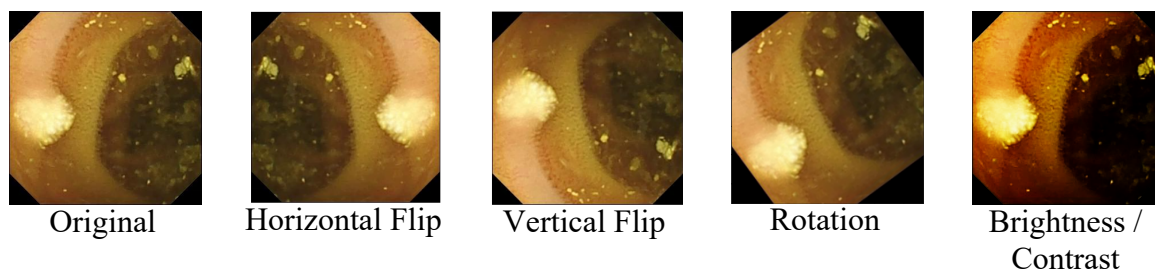


Figure 8: Image Augmentation Methods for Resampling

With resampling, we can generate new images for minority class labels. For the “Usable” vs. “Unusable” partition, we resample the minority “Unusable” images to create a new sample with a ratio of 2:1 images for “Usable”: “Unusable”. Similarly, for the “Normal” vs. “Abnormal”, we resample each type of abnormality to create a new sample with a ratio of 3:1 images for “Normal”: “Abnormal”.

4.4.2 Data Flow

The overall data preparation flow can be found in Figure 10: Overall Data Preparation & Model Training Flowchart. The BPM and ADM models are first trained on the Kvasir-Capsule Dataset. The suitable models are then transferred to the SingHealth workstation offline for additional fine-tuning on the SingHealth dataset. Fine-tuning is required since the SingHealth dataset images have slight discrepancies in colour as compared to the Kvasir-Capsule dataset. These discrepancies arise from the different types of capsule cameras used for the CE procedure. The Kvasir-Capsule dataset utilises the Olympus EC-S10 endocapsule while the SingHealth Dataset utilises the PillCam SB 3 capsule endoscopy system [22]. We find that images from the SingHealth dataset tend to have deeper blacks and less saturation as compared to the Kvasir-Capsule dataset. Example of the discrepancies are illustrated in Figure 9:

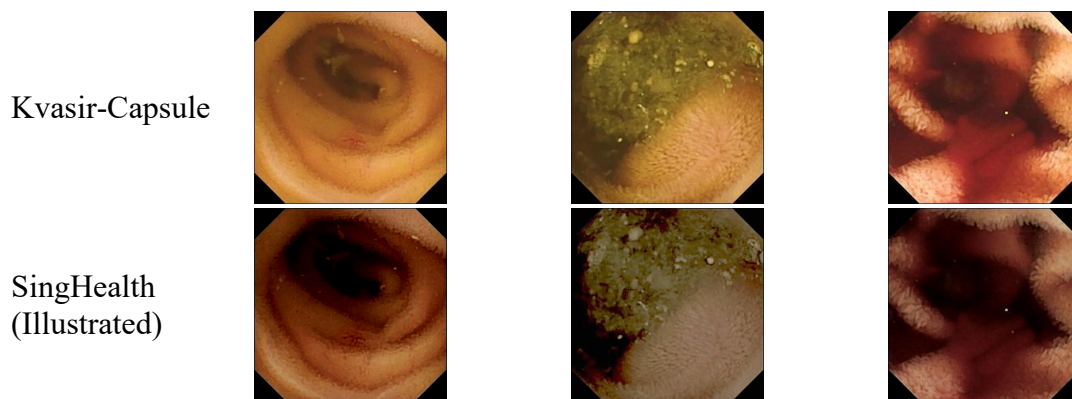


Figure 9: Illustrated discrepancies in colour between datasets

We will account for the differences in CE image colour by fine-tuning our model weights on the SingHealth dataset. The overall data preparation and model training flow can be found in Figure 10:

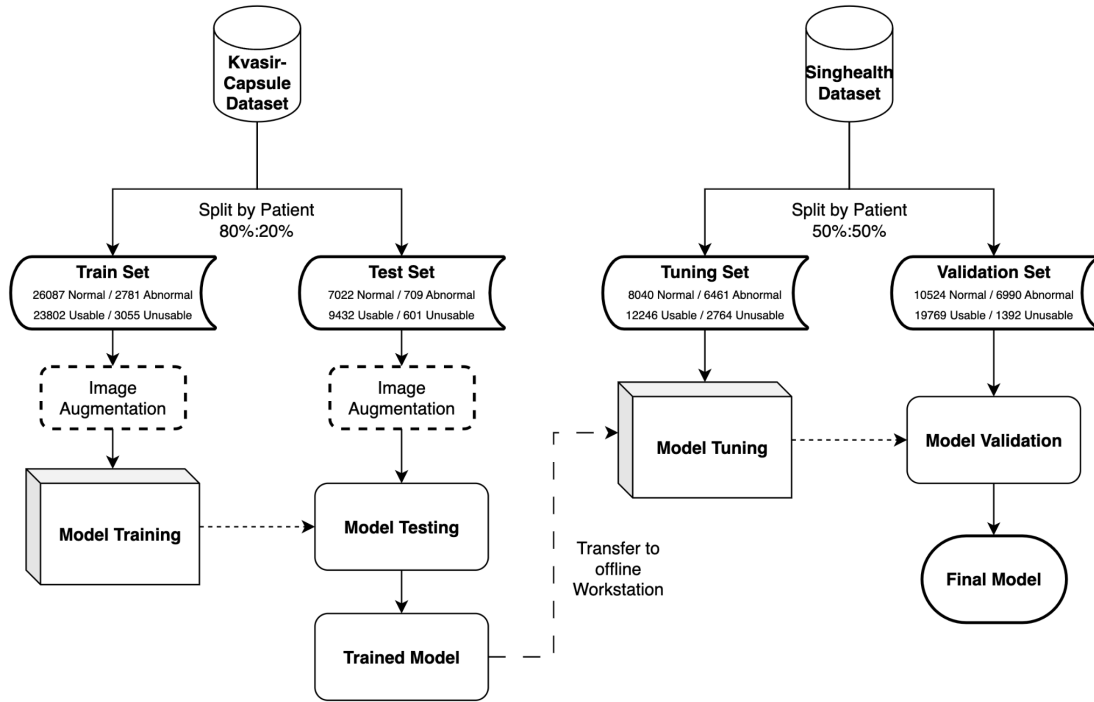


Figure 10: Overall Data Preparation & Model Training Flowchart

The final breakdown of the number of images in each data partition and split are summarised in Table 3:

	Kvasir-Capsule Dataset		SingHealth Dataset	
Data Partition	Training Set (~80%)	Testing Set (~20%)	Tuning Set (~50%)	Validation Set (~50%)
“Usable” vs. “Unusable”	23802 Usable 3055 Unusable	9432 Usable 601 Unusable	12246 Usable 2764 Unusable	19769 Usable 1392 Unusable
“Normal” vs. “Abnormal”	26087 Normal 2781 Abnormal	7022 Normal 709 Abnormal	8040 Normal 6461 Abnormal	10524 Normal 6990 Abnormal

Table 3: Breakdown of Number of Images in each partition and split

4.5 Model Architecture

In this study we explore various image classification models to determine the best model that is fit-for-purpose. For the BPM model, we experiment with Random Forest[23], Deep Learning CNN[24] models, and pre-trained CNN models VGG16[25], ResNet50[26], Xception[27], DenseNet121[28] as base models. For deep learning models, we incorporate image pre-processing layers to standardise the input image. Details of each model type are as follows:

1. Random Forest [23] Classifier: Image array with 3 colour channels (RGB) is flattened to be used as a vector input for the classifier.
2. CNN [24] Model: 3 convolutional layers with max pooling and 3 final classification layers with dropout regularisation.
3. Transfer Learning base w/ CNN Models: Pre-trained layers are used as a base model with the top layer removed. Weights for the base model will not be updated during training. Output from the pre-trained layer is passed into 2 convolutional layers with max pooling followed by 5 final classification layers.

For deep learning models utilising transfer learning, the features extracted from the pre-trained base model are passed on to additional trainable convolutional layers to further extract CE diagnosis-specific features that would contribute to more accurate classification. This leverages on the feature extraction capabilities of the pre-trained models while avoiding the need to train these complex models from scratch. The architecture of our transfer learning models is summarised in Figure 11:

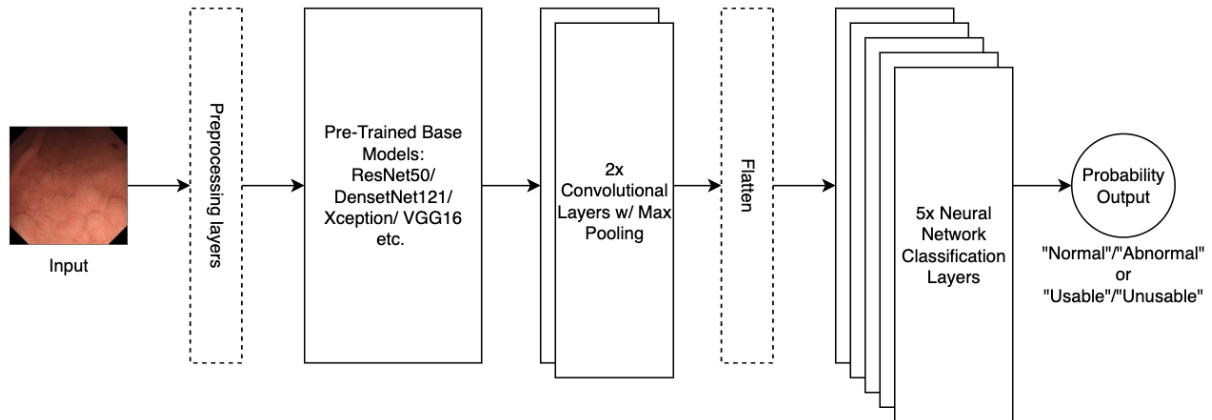


Figure 11: Transfer Learning Model Architecture

From the BPM models, we find that deep learning models with pre-trained CNN base models generally perform better for classification tasks and generalise better to the SingHealth dataset. We therefore rely on similar model architectures with pre-trained base layers for the ADM

models. Similarly, this includes utilising VGG16[25], ResNet50[26], Xception[27], and DenseNet121[28] as base layers and additional convolutional layers with max pooling. Both BPM and ADM deep learning models utilise Stochastic Gradient Descent (SGD) optimizers for better generalised performance [29]. Prediction quality metrics are also monitored including Precision, Recall, F1-Score, Area Under Precision at Recall Curve (AU-PRC) and Area under Receiver Operating Characteristic (AU-ROC). These quality metrics will be used to evaluate the performance of each model during training and validation.

4.6 Model Training, Tuning, and Validation

To train the BPM and ADM models, we rely on a variety of GPU-enabled cloud computing resources to accelerate the training process [30]. Since these cloud computing platforms require data to be uploaded, we can only perform training for the Kvasir-Capsule dataset. As these resources are more readily accessible and have better GPU performance, we primarily rely on them to train the BPM and ADM model weights. The specifications for the cloud computing platforms used in this study are as follows:

- Kaggle Notebook: 2x Nvidia Tesla T4 GPUs (16GB RAM, 65 TFLOPS FP32), 2 Intel Xeon Haswell CPU cores, 13GB RAM [31]
- Google Colab: Nvidia Tesla T4 GPU (16GB RAM, 65 TFLOPS FP32), Intel Xeon Haswell CPU, 12 GB RAM [32]

In comparison, the workstation storing the SingHealth dataset has limited GPU capabilities. We therefore avoid training models from scratch on this workstation and transfer our model architecture and trained weights for the ADM and BPM models to the workstation only for fine-tuning. The specification of the workstation are as follows:

- SingHealth Workstation: Nvidia Quadro P2200 GPU (5GB RAM, 3.8 TFLOPS FP32), Intel Xeon Gold 5218, 128GB RAM running Ubuntu 18.04 LTS

As a barebones offline Linux-based system, our team had to set up the Python-based workflow and GPU driver capabilities through manual installer packages to run our ADM and BPM models. The SingHealth Workstation is configured to run TensorFlow 2.11.0 on Python 3.9.5 with cuDNN 8.1 and CUDA toolkit 11.2 on the Nvidia GPU.

For the Kvasir-Capsule dataset, we previously resampled images within the minority “Abnormal” classes to make up for the imbalance. To further enhance the ADM models’ robustness against the data imbalance, we utilise class training weights for each deep learning model with heavier weightage being placed on the minority “Abnormal” class. Both the BPM and ADM deep learning models employ a decaying learning rate alongside the SGD optimizer to increase the probability of convergence on a global minimum. Additionally, this method is found to help networks learn more complex and nuanced patterns, such as the subtler abnormalities found in the CE images [33]. The quality metrics previously mentioned are monitored on the training and validation sets during training, with early stopping when there is marginal improvement on the model’s AU-ROC performance.

For the Kvasir-Capsule dataset, the data is passed through the model 100 times (with early stopping), where the model weights are updated with each pass. In comparison, the SingHealth data is passed through 5 times on the SingHealth workstation to fine-tune the weights for the CE images in that set. With this training flow, we can leverage on the stronger cloud computing resources to train our model weights from scratch and make minor adjustments later on the SingHealth dataset to account for the differences in CE image colour and hardware.

5. Comparison of Models

We evaluate the predictive performance of each model on the Kvasir-Capsule test set and the SingHealth validation set after tuning. Each model’s binary classification threshold is optimised to maximise F1-score [34] and the AU-PRC performance is recorded as it is robust against data imbalance [35]. We list the performance measures for each model type on both datasets and classification tasks in Table 4:

Kvasir-Capsule Test Dataset				
Model Name	Precision	Recall	F1-Score	AU-PRC
BPM Random Forest	0.9645	0.9690	0.9668	0.9890
BPM CNN	0.9925	0.4932	0.6590	0.9803
BPM VGG16 base w/ CNN	0.9538	0.9723	0.9630	0.9942
ADM Xception base	0.6853	0.5574	0.6148	0.6841
ADM Xception base w/ CNN	0.6322	0.6292	0.6307	0.6966
ADM DenseNet121 base w/ CNN	0.8071	0.5957	0.6855	0.7266

SingHealth Validation Dataset				
Model Name	Precision	Recall	F1-Score	AU-PRC
BPM Random Forest (no tuning)	0.9840	0.0576	0.1089	0.9207
BPM CNN	0.9879	0.0281	0.0546	0.8637
BPM VGG16 base w/ CNN	0.9846	0.9791	0.9819	0.9984
ADM Xception base	0.6703	0.7369	0.7030	0.7983
ADM Xception base w/ CNN	0.6923	0.7834	0.7351	0.8266
ADM DenseNet121 base w/ CNN	0.7315	0.7611	0.7460	0.8415

Table 4: Summary of Top Performing Models based on F1-Score

A complete list of all models tested can be found in Appendix B. Precision measures the proportion of true positive cases picked up by the model, given by $Precision = \frac{TP}{TP + FP}$ where TP is the number of true positives and FP is the number of false positives. Recall measures the sensitivity of the model, given by $Recall = \frac{TP}{TP + FN}$ where FN is the number of false negatives. Ideally, we seek a model which balances out on precision and recall, where the model is able to pick up as many “Abnormal” and “Usable” cases as possible while minimising the false positive and false negative rates. In practice, this allows GE specialists to minimise time wasted

on reading images that are normal and unusable (high precision) and minimises the chance of missing images that are usable and have abnormalities (high recall). Models are therefore benchmarked based on F1-Score, a measure that takes a balance between precision and recall

given by $F1 = \frac{precision \cdot recall}{precision + recall}$.

For the BPM models, we observe that the random forest model yields generally good F1-scores on the Kvasir-Capsule dataset. However, when tested on the SingHealth dataset, we find that the model performance severely deteriorates. This signals that the model is not able to account for the subtle differences in colour between the datasets and cannot be generalised as well when compared to the “BPM CNN w/ VGG16” model. We also observe that the “BPM CNN” model without any transfer learning base models performs poorly on both datasets. This suggests that more training is needed since training CNN models from scratch often takes copious amounts of time and data [36].

For ADM models, we find that the models perform better on the SingHealth dataset as compared to the Kvasir-Capsule dataset. This may seem counterintuitive since only minor tuning and updates were done on the SingHealth dataset. However, we can account for this increase in performance based on the specific abnormalities found in each dataset. From Table 3 we find that the SingHealth dataset has more cases of bleeding abnormalities. These abnormalities are more conspicuous and easier to identify due to significant colour differences in CE images as seen in Figure 1. In contrast, the Kvasir-Capsule dataset has more cases of Angioectasia and Ulcers, which can be subtler and harder to identify. A description of CE image pathologies can be found in Appendix A.

6. Results & Benefits

Since models will eventually be implemented on SingHealth-specific CE images, we select models based on the best F1-score on the SingHealth validation dataset. The best performing BPM model was the BPM VGG16 base w/ CNN with an F1-score of 98.19% and an AU-PRC of 99.84% while the best performing ADM model was the ADM DenseNet121 base w/ CNN with an F1-score of 74.6% and an AU-PRC of 84.51%. The confusion matrix for both models' performance on the SingHealth validation set can be found in Figure 12:

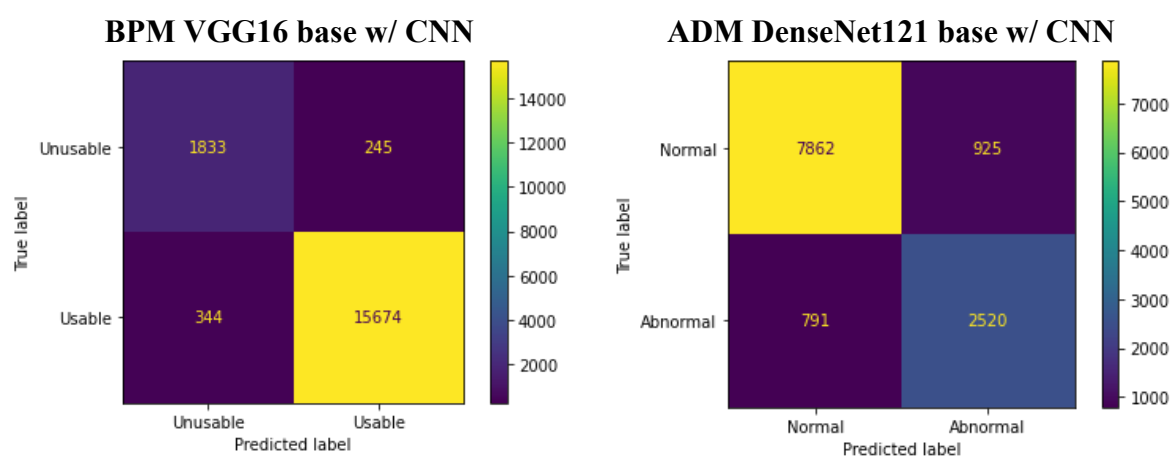


Figure 12: Confusion Matrix for BPM and ADM models on SingHealth Validation Set

The BPM model has an inference time of 166 seconds on the set of 18,096 images (109 images/sec) while the ADM model has an inference time of 86 seconds on the set of 12,098 images (140 images/sec). This performance is much faster than a completely manual diagnosis from GE specialists.

The VGG16 model is an iteration of the AlexNet model introduced in 2014. The architecture utilises 3x3 convolutional filters with an increased number of layers to further enable the model to generalise [25]. By stacking these layers, more information can be captured without increasing the number of parameters. In comparison, the DenseNet121 model incorporates dense connections to concatenate the feature maps of all layers in the model. Each layer

learns the residuals from the previous, allowing early layers to capture low-level features and future layers to capture high level features from complex images [37]. We hypothesise that the VGG16 base layer was not able to perform as well as the DenseNet base for the ADM model due to the larger filter size. The more nuanced features in the abnormalities are lost when passed through the 3x3 convolutional filters and the features extracted by the VGG16 model are not as precise as those extracted by the DenseNet121 model.

To quantify the benefits of utilising the ADM model in the CE diagnosis process, we can utilise lift and gain analysis [38]. The predicted probability of an abnormality for images in the validation set is sorted in descending order (1 being abnormal and 0 being normal) and split into 20 deciles of 393 images each. Gain represents the percentage of abnormalities captured in each decile, where the first deciles are images with the highest probability of an abnormality. From the Gain chart, we can observe that 43.2% of abnormalities are captured in the first decile. This represents a lift value of 8.64, where lift represents the ratio of observing an abnormality in the decile with the model as compared to observing an abnormality without the model. In practice, this means that the GE specialist is 8.64 times more likely to observe an abnormality if he or she only reads images in the first decile. The lift and gain charts for all deciles can be found in Figure 13.

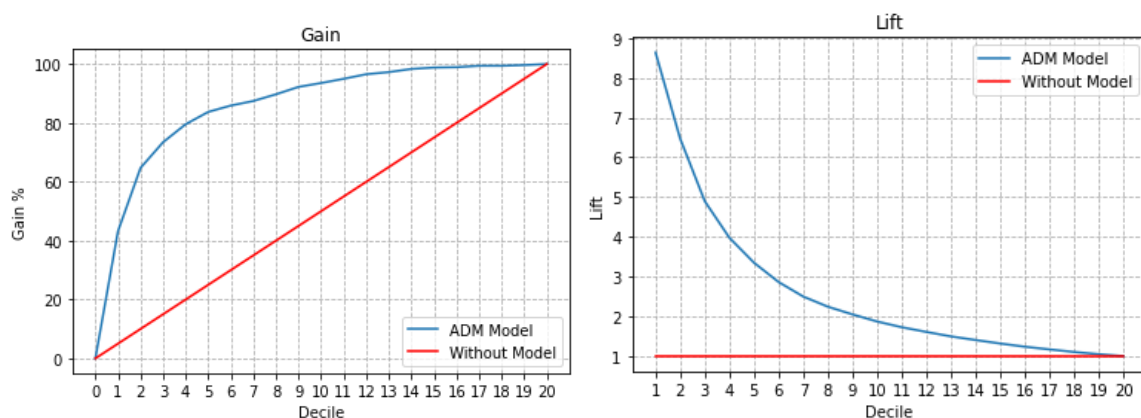


Figure 13: Lift and Gain Charts for ADM model

The current results demonstrate that both the BPM and ADM computer vision models based on deep learning CNNs are able to achieve acceptable levels of performance across the quality metrics. Integrating the models into the proposed framework in Figure 7, we find that the selected models are able to meet the requirements set out in Table 1 and achieve the desired benefits. The concept will be able to serve as a decision support system to partially automate the CE diagnosis. If implemented into practice, the CE image diagnosis assistant framework will ultimately benefit GE specialists, reducing the attention capacity and time required for diagnosis.

6.1 Implementation of Diagnosis Assistant Framework

To implement our models into a user-friendly system, our team utilised the Streamlit app framework to package our models into a web-based tool as a proof-of-concept. This allows doctors and GE specialists without technical expertise to easily utilise our models and experiment using them in the CE diagnosis process. Users can experience how the models and framework developed might ultimately improve clinical performance. In this tool, users can upload images recorded from the CE procedure. Users can then click the “Analyse Images” button to pass all the uploaded images through our diagnosis framework. Figure 14 illustrates how users can upload CE images to the Steamlit web tool.

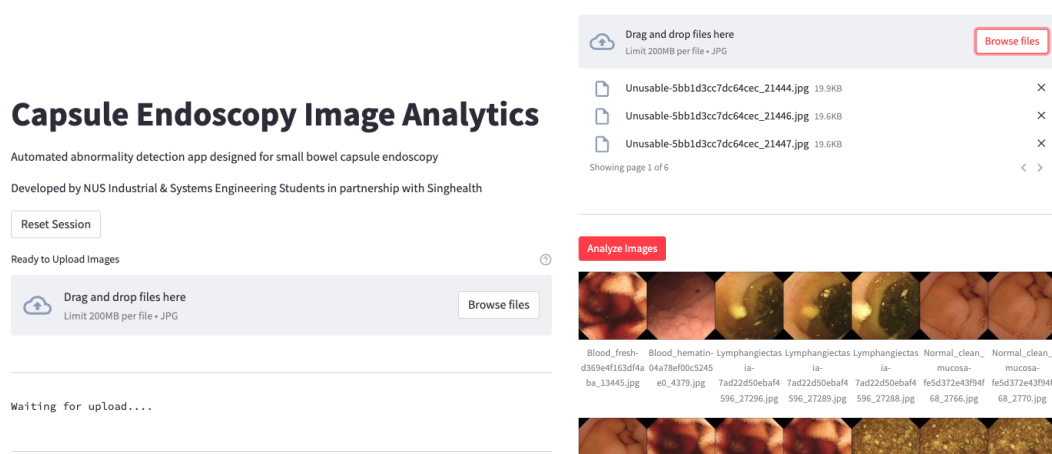


Figure 14: Interface to Upload CE Images

Once the analysed button is clicked, uploaded images are passed into our BPM and ADM models in the back end, where both models will classify each image based on usability and probability of an abnormality. “Usable” and “Unusable” images are divided into 2 tabs, each showing the probability score from the ADM model. Since the “Unusable” images are clouded by faecal matter, the ADM model output may be unreliable as compared to the “Usable” images. Images are highlighted in red, yellow, and green based on a user-defined threshold of abnormal, uncertain, and normal respectively. This allows the user to focus on the images highlighted in red, where the ADM model finds a high probability of an abnormality. Users can also choose to further read images in the uncertain yellow category to be more comprehensive in the diagnosis. The output in the tool is illustrated in Figure 15:

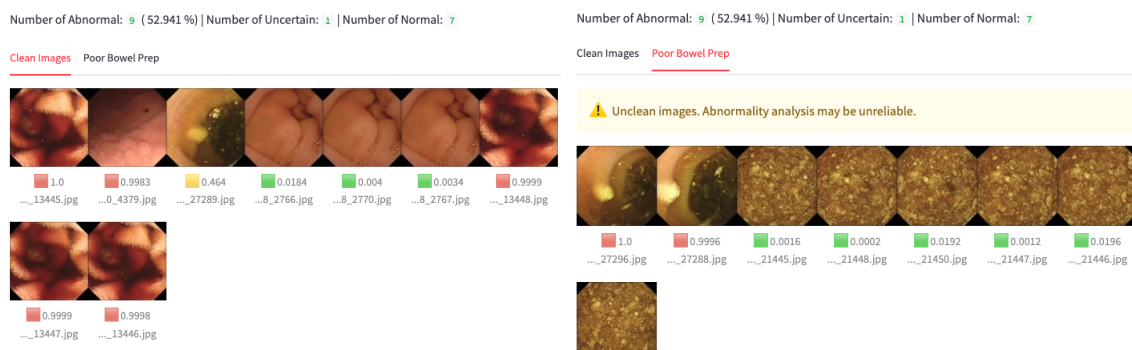


Figure 15: Results of Analysis from the BPM and ADM models

Users can set the threshold for abnormal, uncertain, and normal with a slider above the results. As shown in Figure 16, this allows the user to tune the classification and labelling threshold based on the model and sensitivity they desire. Furthermore, users can also view the ADM output of all the images in a line graph. If images are uploaded sequentially, we expect an abnormality to be present in contiguous sets of images. Users can therefore infer which frames are likely to contain an abnormality where there are continuously high probabilities of abnormalities.

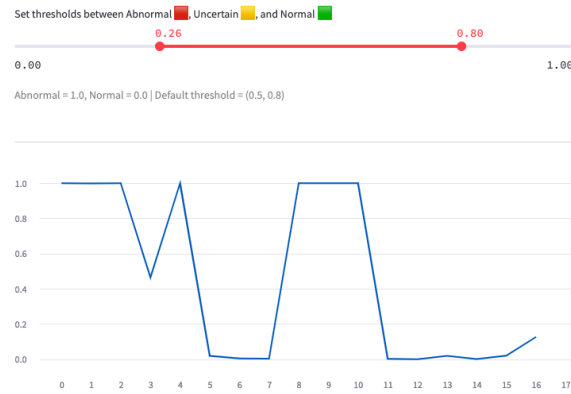


Figure 16: Threshold Slider and Summary of Probabilities Line Graph

With this proof-of-concept web tool, we hope to demonstrate the benefits of implementing and adopting the diagnosis assistant framework into the CE diagnosis procedure. With these benefits, we hope to garner support for further development of the tool and models.

7. Recommendations

While the current BPM and ADM models yield promising results, we acknowledge the limitations of the model architecture, training, and validation process. Future improvements can also be made to the diagnosis assistant pipeline to improve the reliability of the results.

7.1 Data & Model Limitations

In this project, we utilise 2 different sources of image data for model training and tuning. We find that images in the 2 datasets have variation in colour and contrast, where tuning was needed on the SingHealth dataset to reach acceptable levels of performance. We propose the 2 ADM and BPM models under the assumption that future CE images will be sourced from the same hardware and will have similar colour and contrast as the current data available in the SingHealth dataset.

In addition, our initial data preparation split both datasets into training and validation sets based on the patients. This resulted in 2 sets of testing and validation data for each ADM and BPM

model. Given more data and computational time, we would utilise cross-validation on more data splits (folds) in order to benchmark the predictive performance of each model across the full dataset. This process would involve partitioning the full dataset into multiple folds where each fold is used for testing and the others are used for training. This is a computationally intensive process that would involve training and testing a model for each fold of the data. However, this approach prevents overfitting and represents more generalised performance metrics, reducing the chance of selecting models that are biased towards a small subset of available data [39].

7.2 Future Works

Given the project timeline and resource constraints, our team was not able to continue the development and improvement of our models and Diagnosis Assistant Framework. In this section, we propose the next steps so future teams can leverage on the work we have done to further improve our proposed solutions and unlock more potential to assist GE specialists in the CE procedure.

7.2.1 Abnormality Classification Model

In this project, we implement an ADM binary classification model to detect specific abnormalities in CE images. The output of this model is a probability prediction of the presence of an abnormality within the image but does not specify the type of abnormality. This means that GE specialists will still need to intervene to diagnose the specific abnormality within each image. To better support the CE diagnosis procedure, additional models can be implemented to detect the specific type of abnormality within each abnormal image. This approach can range from building a single multiclass classification model to building a detection model for each type of abnormality. The ability to diagnose the specific type of abnormality serves to further lighten the workload and scrutiny that GE specialists must invest into each patient's CE images.

7.2.2 Detection of Contiguous Sets of Abnormal Images

The CE images are captured and recorded sequentially as the capsule passes through the patient's digestive tract and images are recorded with file names in sequential order. Abnormalities found using CE will therefore also be recorded in sets of sequential, contiguous frames. We can therefore highlight sets of contiguous abnormal images with our ADM model and overcome any inaccuracies by using a majority vote algorithm for images within the search window. With this approach, we can search and identify sequences of images that capture the same abnormality, thereby overcoming minor inaccuracies in the ADM model. An example of this approach can be illustrated in Figure 17:

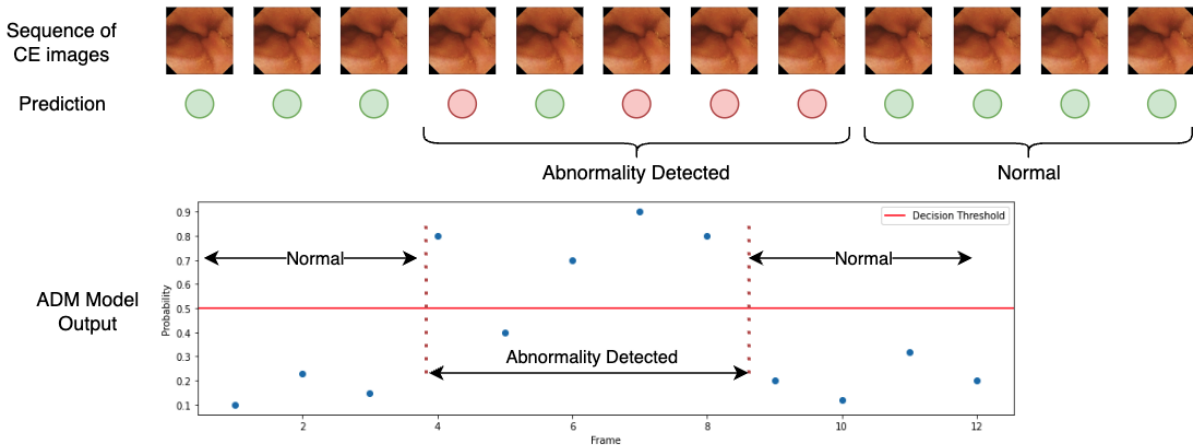


Figure 17: Detection of Abnormalities in Sequences of Images

To implement this approach, experiments can be designed to test the search window size, majority vote threshold, and ADM model prediction threshold to determine the optimal parameters that maximise the performance in finding sequences of abnormal images.

Temporal filtering techniques can also be employed to overcome any false negatives and false positives predicted by the ADM model [40], thereby improving the overall reliability of the diagnosis assistant framework.

7.2.3 Model Pipeline and Hyperparameter Tuning

In this project, we formulate the architecture and layer specifications of our CNNs based on common practices found during our literature review. However, there is more room to experiment with different CNN architectures and hyperparameters such as the number of hidden layers, regularisation layers, model learning rate, and the activation functions used. We recommend building a pipeline to package the data preparation, model hyperparameter specification, model training, and model testing into a more efficient process for cross validation. This allows more experiments to be conducted on optimising the model architecture and hyperparameters of the model layers, unlocking greater performance and potential from CNN models [41].

9. Conclusion and Reflection

In this project, we demonstrate the potential for an AI computer vision model to serve as a decision support system and diagnosis assistant for GE specialists in the CE diagnosis process. The diagnosis assistant framework we propose breaks down the diagnosis process into 2 stages: filtering out usable images and labelling images with abnormalities. These 2 stages are fulfilled by the Bowel Prep Model (BPM) and Abnormality Detection Model (ADM) respectively. Both models are trained and tuned on the Kvasir-Capsule and SingHealth dataset respectively. Evaluating the quality metrics of both models on a hold-out validation set, we find that they achieve acceptable levels of performance and are able to serve as a proof-of-concept for future development.

This project presented our team with the opportunity to apply machine learning in a real-world application. We learnt the basics of computer vision models and algorithms, including image augmentation, resampling techniques, and classical computer vision models which served as

the foundation for us to adopt more advanced deep learning methods. Throughout the literature review process, we discovered the inner workings of Convolutional Neural Networks and transfer learning techniques, eventually implementing them into our models.

In addition to academics, our team overcame many practical challenges such as setting up the SingHealth workstation remotely offline. This was a tedious process of debugging Python installation packages for each framework used, installing them manually (as compared to simply using “pip install” commands) and ensuring that the deep learning Nvidia cuDNN and CUDA driver was harmonising with our Python environment. Furthermore, our team had the chance to practise project management skills, including Gantt Chart scheduling, scoping of objectives and deliverables, and stakeholder management.

References

- [1] “Singapore Health Services - Singapore hospitals and doctors,” *SingHealth*. [Online]. Available: <http://www.singhealth.com.sg/>. [Accessed: 26-Oct-2022].
- [2] “Types of endoscopy,” *Cancer.Net*, 24-Apr-2020. [Online]. Available: <https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/types-endoscopy>. [Accessed: 27-Oct-2022].
- [3] B. Dooley, “Endoscopy vs colonoscopy - what's the difference?,” *Gastroenterology Consultants of San Antonio*, 26-Oct-2021. [Online]. Available: <https://www.gastroconsa.com/endoscopy-vs-colonoscopy-whats-the-difference/>. [Accessed: 26-Oct-2022].
- [4] “Upper Endoscopy,” Mayo Clinic, 26-Aug-2022. [Online]. Available: <https://www.mayoclinic.org/tests-procedures/endoscopy/about/pac-20395197>. [Accessed: 27-Oct-2022].
- [5] S. H. Kim and H. J. Chun, "Capsule Endoscopy: Pitfalls and approaches to overcome," in *Diagnostics*, vol. 11, no. 10, pp. 1765, Oct. 2021. [Online]. Available: <https://doi.org/10.3390/diagnostics11101765>. [Accessed: Mar. 17, 2023].
- [6] C. Yang, V. Sriranjani, A. M. Abou-Setta, W. Poluha, J. R. Walker, and H. Singh, “Anxiety associated with colonoscopy and FLEXIBLE SIGMOIDOSCOPY: A systematic review,” *The American journal of gastroenterology*, Dec-2018. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768596/>. [Accessed: 27-Oct-2022].
- [7] “Wireless Capsule Endoscopy for Gastrointestinal Imaging and the Patency Capsule,” *Healthy Blue Provider*, 06-Jul-2022. [Online]. Available: https://provider.healthybluesc.com/dam/medpolicies/healthybluesc/active/guidelines/glpw_d080197.html. [Accessed: 21-Mar-2023].

- [8] V. L. Leva and R.-R. Alberto, “Effectiveness and safety of capsule endoscopy in the diagnosis of small bowel diseases,” *Journal of clinical gastroenterology*. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/18277887/>. [Accessed: 27-Oct-2022].
- [9] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, “Artificial Intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda,” *Journal of ambient intelligence and humanized computing*, 13-Jan-2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8754556/#:~:text=Artificial%20intelligence%20can%20assist%20providers,discovery%2C%20and%20patient%20risk%20identification>. [Accessed: 30-Oct-2022].
- [10] M. Taghiakbari, Y. Mori, and D. von Renteln, “Artificial Intelligence-Assisted Colonoscopy: A review of current state of Practice and Research,” *World journal of gastroenterology*, 21-Dec-2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8704267/>. [Accessed: 30-Oct-2022].
- [11] J. Collingwood, “Artificial Intelligence in medical diagnosis,” *Southern Medical Association*, 07-Oct-2021. [Online]. Available: <https://sma.org/ai-in-medical-diagnosis/>. [Accessed: 30-Oct-2022].
- [12] R. A. N. D. Y. GLICK, “Artificial Intelligence in medical diagnosis,” *Southern Medical Association*, 07-Oct-2021. [Online]. Available: <https://sma.org/ai-in-medical-diagnosis/>. [Accessed: 21-Mar-2023].
- [13] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler, and P. Halvorsen, “Kvasir-capsule, a video capsule Endoscopy

- Dataset,” Scientific data, 27-May-2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8160146/>. [Accessed: 30-Oct-2022].
- [14] N. Lang, “Using convolutional neural network for Image Classification,” Towards Data Science, 05-Dec-2021. [Online]. Available: <https://towardsdatascience.com/using-convolutional-neural-network-for-image-classification-5997bfd0ede4>[Accessed: 30-Oct-2022].
- [15] “ML Practicum: Image Classification” Google, 18-Jul-2022. [Online]. Available: <https://developers.google.com/machine-learning/practica/image-classification/convolutional-neural-networks>. [Accessed: 30-Oct-2022].
- [16] “Convolutional Neural Network - Javatpoint,” www.javatpoint.com. [Online]. Available: <https://www.javatpoint.com/keras-convolutional-neural-network>. [Accessed: 30-Oct-2022].
- [17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," arXiv preprint arXiv:1409.1556, 2014. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6795963> [Accessed: Mar. 15, 2023].
- [18] . Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. [Online]. Available: <https://ieeexplore.ieee.org/document/5206848> [Accessed: Mar. 15, 2023].
- [19] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler, and P. Halvorsen, “Kvasir-capsule, a video capsule Endoscopy Dataset,” *Nature News*, 27-May-2021. [Online]. Available: <https://www.nature.com/articles/s41597-021-00920-z>. [Accessed: 21-Mar-2023].

- [20] M. Sharma, "5 Techniques to Handle Imbalanced Data for a Classification Problem," Analytics Vidhya, Jun. 2021. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>. [Accessed: Mar. 15, 2023].
- [21] B. Buslaev, A. Parinov, E. Khvedchenya, and V. Ivashkin, "Albumentations Documentation," 2021. [Online]. Available: <https://albumentations.ai/docs/>. [Accessed: Mar. 15, 2023].
- [22] Medtronic, "PillCam™ SB 3 Capsule Endoscopy System," Medtronic, [Online]. Available: <https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-sb3-system.html>. [Accessed: Mar. 15, 2023].
- [23] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. [Online]. Available: <https://link.springer.com/article/10.1023/a:1010933404324>. [Accessed: Mar. 15, 2023].
- [24] B. Lee and W. Lee, "Convolutional neural networks: an overview and application in radiology" Insights into Imaging, vol. 9, no. 1, pp. 1-18, 2018. [Online]. Available: <https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9#:~:text=CNN%20is%20a%20type%20of,%2D%20to%20high%2Dlevel%20patterns>. [Accessed: Mar. 15, 2023].
- [25] S. Hochreiter and J. Schmidhuber, "Very Deep Convolutional Networks for Large-Scale Image Recognition" arXiv preprint arXiv:1409.1556, 2014. [Online]. Available: <https://arxiv.org/abs/1409.1556>. [Accessed: Mar. 15, 2023].
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778. [Online]. Available: <https://arxiv.org/abs/1512.03385>. [Accessed: Mar. 15, 2023].

- [27] A. Radford, L. Metz, and S. Chintala, "Xception: Deep Learning with Depthwise Separable Convolutions" in International Conference on Learning Representations, 2016. [Online]. Available: <https://arxiv.org/abs/1610.02357>. [Accessed: Mar. 15, 2023].
- [28] D. P. Kingma and J. Ba, "Densely Connected Convolutional Networks" arXiv preprint arXiv:1608.06993, 2014. [Online]. Available: <https://arxiv.org/abs/1608.06993>. [Accessed: Mar. 15, 2023].
- [29] L. Cheng, Q. Liu, J. Liu, and Y. Huang, "A Comprehensive Survey on Hyperparameter Optimization for Deep Learning," in Proceedings of the 4th Workshop on Optimization for Machine Learning, 2021, pp. 1-15. [Online]. Available: <https://opt-ml.org/papers/2021/paper53.pdf>. [Accessed: Mar. 15, 2023].
- [30] Google Cloud, "Using GPUs on AI Platform Training," Google Cloud, [Online]. Available: <https://cloud.google.com/ai-platform/training/docs/using-gpus>. [Accessed: Mar. 15, 2023].
- [31] Kaggle. 2021 Kaggle Machine Learning & Data Science Survey. [Online]. Available: <https://kaggle.com/competitions/kaggle-survey-2021> [Accessed: Mar. 15, 2023].
- [32] A. Kazemnejad, "How to Do Deep Learning Research with Absolutely No GPUs - Part 2," [Online]. Available: https://kazemnejad.com/blog/how_to_do_deep_learning_research_with_absolutely_no_gpus_part_2/. [Accessed: Mar. 15, 2023].
- [33] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," in International Conference on Learning Representations, 2019, pp. 1-16. [Online]. Available: <https://arxiv.org/abs/1809.11096>. [Accessed: Mar. 15, 2023].

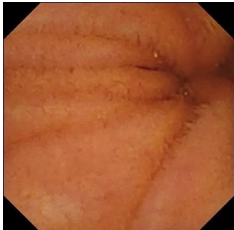
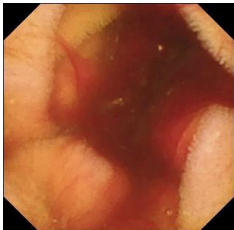
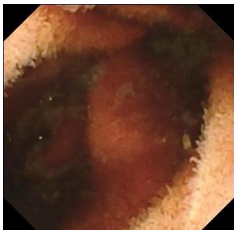

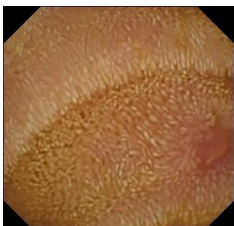
- [34] F. Pedregosa et al., Eds., "Python Machine Learning," Springer, 2nd ed., 2017. [Online]. Available: <https://link.springer.com/book/10.1007/978-3-319-98074-4>. [Accessed: Mar. 15, 2023].
- [35] J. M. Tingle et al., "An empirical evaluation of sampling methods for the classification of imbalanced data" *British Journal of Anaesthesia*, vol. 126, no. 3, pp. e68-e72, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/35901023/>. [Accessed: Mar. 15, 2023].
- [36] Google Developers, "Leveraging Pretrained Models for Image Classification," Google Developers, [Online]. Available: <https://developers.google.com/machine-learning/practica/image-classification/leveraging-pretrained-models>. [Accessed: Mar. 15, 2023].
- [37] A. Savakis, "<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9137242>," *IEEE Xplore* Full-text PDF: 09-Jul-2020. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9137242>. [Accessed: 21-Mar-2023].
- [38] M. Afzal, "Model Benefit Evaluation with Lift and Gain Analysis," *Towards Data Science*, [Online]. Available: <https://towardsdatascience.com/model-benefit-evaluation-with-lift-and-gain-analysis-4b69f9288ab3>. [Accessed: Mar. 15, 2023].
- [39] A. Sharma, "Cross-Validation in Machine Learning," *Towards Data Science*, [Online]. Available: <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>. [Accessed: Mar. 15, 2023].
- [40] [1] L. H. Abdullatif et al., "AI-assisted diagnosis and treatment of prostate cancer: An update," *Expert Review of Anticancer Therapy*, vol. 21, no. 11, pp. 1115-1122, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/36010210/>. [Accessed: Mar. 15, 2023].

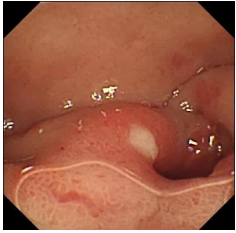

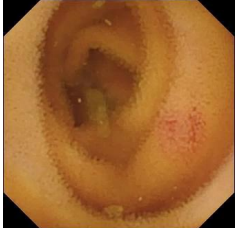
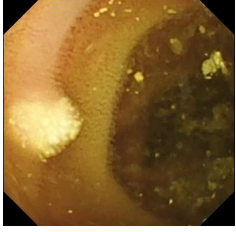
- [41] M. Khan, M. H. Malik, and S. J. Lee, "Deep Learning for Handwritten Text Recognition: A Comprehensive Review," in *Advanced Computing and Big Data Processing*, J. Abraham et al., Eds. Cham: Springer International Publishing, 2021, pp. 49-69. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-72711-6_4. [Accessed: Mar. 15, 2023].

Appendix

A. Pathology Descriptions

Pathology description provided by Dr. Jiang Bochao from SingHealth.

Label	Description	Sample Picture
Normal	Smooth, uniform, healthy-looking with occasional folds, uniform short finger-like projections from the mucosa known as “villi”	
Fresh Blood	Presence of bright red liquid stains on mucosa	
Stale Blood/ Clotted blood	Presence of dark maroon coloured liquid stains on mucosa	
Erythema	Limited patches of pinkish red mucosa with surrounding normal mucosa, there should be no disruption to the surface integrity (i.e. villi should still be present)	
Erosion	Limited patch of pinkish red mucosa with surrounding normal mucosa, but the surface integrity may be disrupted, villi may not be seen, in some cases there may be a small pin-point white spot	

Ulcer	Pinkish red mucosa with surrounding normal mucosa, and a prominent white spot in the middle; the centre may sometimes appear slightly excavated	
Polyp	Protruded mass, may be of slightly irregular edges. Colour of mucosa appears similar to the surrounding mucosa, and surface integrity should be preserved (eg villi should be seen)	
Angioectasia	Reddish streaks suggestive of prominent and very superficial blood vessels just beneath the mucosa	
Lymphangiectasia	Whitish, chalk-like, usually rounded edges; may be flat or protruding (different from polyp, which has normal overlying mucosa). There will be no villi seen on the areas of lymphangiectasia	

B. BPM & ADM Model Performance

Bowel Prep Model Name	SingHealth Dataset						Kvasir-Capsule Dataset					
	AU-ROC	Precision	Recall	Optimal F1 Threshold	F1-Score	AU-PRC	AU-ROC	Precision	Recall	Optimal F1 Threshold	F1-Score	AU-PRC
BPM Random Forest (max_depth=12)	0.5934	0.9840	0.0576	0.500	0.1089	0.9207	0.9798	0.9645	0.9690	0.420	0.9668	0.9890
BPM CNN	0.3811	0.9879	0.0281	0.500	0.0546	0.8637	0.9582	0.9925	0.4932	0.050	0.6590	0.9803
BPM VGG16 base	0.7760	0.9886	0.1569	0.500	0.2709	0.9608	0.9534	0.9210	0.9826	0.600	0.9508	0.9804
BPM VGG16 base w/ CNN	0.7261	0.9665	0.3155	0.500	0.4758	0.9505	0.9865	0.9538	0.9723	0.170	0.9630	0.9942
BPM Xception base w/ CNN	0.7674	0.9253	0.7896	0.500	0.8522	0.9647	0.9759	0.9423	0.9626	0.700	0.9523	0.9884
After Tuning												
BPM VGG16 base	0.7789	0.9285	0.9092	0.500	0.9187	0.9654						
BPM VGG16 base w/ CNN	0.9889	0.9846	0.9791	0.25	0.9819	0.9984						
BPM Xception base w/ CNN	0.9316	0.9390	0.9576	0.2	0.9482	0.9904						

Abnormal Detection Model Name	SingHealth Dataset						Kvasir-Capsule Dataset					
	AU-ROC	Precision	Recall	Optimal F1 Threshold	F1-Score	AU-PRC	AU-ROC	Precision	Recall	Optimal F1 Threshold	F1-Score	AU-PRC
ADM IncepRes base	0.8377	0.5131	0.8562	0.995	0.6416	0.7542	0.8846	0.7588	0.5646	0.625	0.6474	0.6987
ADM VGG16 base	0.7731	0.3462	0.9851	0.995	0.5123	0.6524	0.7747	0.5604	0.4438	0.320	0.4953	0.5094
ADM Xception base	0.8174	0.6269	0.7014	0.940	0.6621	0.7292	0.8947	0.6853	0.5574	0.620	0.61482	0.6841
ADM ResNet50 base	0.8123	0.3850	0.9495	0.995	0.5479	0.7425	0.8550	0.7428	0.5215	0.750	0.6128	0.6569
ADM DenseNet121 base	0.9386	0.4566	0.9982	0.995	0.6266	0.9306	0.8784	0.6499	0.5239	0.655	0.5801	0.6381
ADM Densenet121 base w/ CNN	0.8507	0.5854	0.8787	0.940	0.7027	0.8493	0.9090	0.8071	0.5957	0.390	0.6855	0.7266
ADM Xception base w/ CNN	0.7532	0.6084	0.6818	0.825	0.6430	0.7429	0.9135	0.6322	0.6292	0.56	0.6307	0.6966
After Tuning												
ADM Xception base	0.8901	0.6703	0.7369	0.545	0.7030	0.7983						
ADM Xception base w/ CNN	0.9113	0.6923	0.7834	0.560	0.7351	0.8266						
ADM Densenet121 base w/ CNN	0.9070	0.7315	0.7611	0.305	0.7460	0.8415						

BLANK