Anatomical Classification of the Gastrointestinal tract Using Ensemble Transfer Learning

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon

By
Fatemeh Sedighipour Chafjiri

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Electrical and Computer Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Abstract

Endoscopy is a procedure used to visualize disorders of the gastrointestinal (GI) lumen. GI disorders can occur without symptoms, which is why gastroenterologists often recommend routine examinations of the GI tract. It allows a doctor to directly visualize the inside of the GI tract and identify the cause of symptoms, reducing the need for exploratory surgery or other invasive procedures. It can also detect the early stages of GI disorders, such as cancer, enabling prompt treatment that can improve outcomes. Endoscopic examinations generate significant numbers of GI images. Because of this vast amount of endoscopic image data, relying solely on human interpretation can be problematic. Artificial intelligence is gaining popularity in clinical medicine. Assist in medical image analysis and early detection of diseases, help with personalized treatment planning by analyzing a patient's medical history and genomic data, and be used by surgical robots to improve precision and reduce invasiveness. It enables automated diagnosis, provides physicians with assistance, and may improve performance. One of the significant chyallenges is defining the specific anatomic locations of GI tract abnormalities. Clinicians can then determine appropriate treatment options, reducing the need for repetitive endoscopy. Due to the difficulty of collecting annotated data, very limited research has been conducted on the localization of anatomical locations by classification of endoscopy images. In this study, we present a classification of GI tract anatomical localization based on transfer learning and ensemble learning. Our approach involves the use of an autoencoder and the Xception model. The autoencoder was initially trained on thousands of unlabeled images, and the encoder then separated and used as a feature extractor. The Xception model was also used as a second model to extract features from the input images. The extracted feature vectors were then concatenated and fed into a Convolutional Neural Network for classification. This combination of models provides a powerful and versatile solution for image classification. By using the encoder as a feature extractor that can transfer the learned knowledge, it is possible to improve learning by allowing the model to focus on more relevant and useful data, which is extremely valuable when there are not enough appropriately labelled data. On the other hand, the Xception model provides additional feature extraction capabilities. Sometimes, one classifier is not enough in machine learning, as it depends on the problem we are trying to solve and the quality and quantity of data available. With ensemble learning, multiple learning networks can work together to create a stronger classifier. The final classification results are obtained by combining the information from both models through the CNN model. This approach demonstrates the potential for combining multiple models to improve the accuracy of image classification tasks in the medical domain. The HyperKvasir dataset is the main dataset used in this study. It contains 4,104 labelled and 99,417 unlabeled images taken at six different locations in the GI tract, including the cecum, ileum, pylorus, rectum, stomach, and Z line. After dataset preprocessing, which includes noise deduction and similarity removal, 871 labelled images remained for the purpose of this study. Our method was more accurate than state-of-the-art studies and had a higher F1 score while categorizing the input images into six different anatomical locations with less than a thousand labelled images. According to the results, feature extraction and ensemble learning increase accuracy by 5%, and a comparison with existing methods using the same dataset indicate improved performance and reduced cross-entropy loss. The proposed method can therefore be used in the classification of endoscopy images.

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Khan Wahid, without whom this research could not have been completed, nor could I have accomplished my master's studies. I am also grateful to Dr. Babyn for encouraging me and providing me with constructive comments that eased my journey. I am extremely grateful for my incredible parents and loving sister, whom I always have at my side, loving, supporting, and motivating me. Thanks to my lab mates who guided me through this process with their insight and knowledge. Thanks to the University of Saskatchewan for awarding me the Devolved Scholarship. I want to conclude by thanking my friends for all their unconditional support during this intense academic year.

I dedicate this thesis to my favourite people who never failed in supporting me unconditionally.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CDC	Compressed Domain Color
CE	Conventional Endoscopy
CNN	Nonvolutional Neural Network
CRB	Cramer-Rao Bounds
DHash	Differential Gradient Hash
DL	Deep Learning
DoA	Direction of Arrival
EGD	Esophagogastroduodenoscopy
GI	Gastrointestinal Tract
K-SVM	Kernel Support Vector Machine
ML	Machine Learning
MSE	Mean Squared Error
PLSA	Probabilistic Latent Semantic Analysis
PSNR	Peak signal-to-noise ratio
RFID	Radio Frequency Identification
ROC	Receiver Operating Characteristics
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
SIFT	Scale Invariant Feature Transform
TOA	Time of Arrival
WCE	Wireless Capsule Endoscopy
WE	Wireless Endoscopy

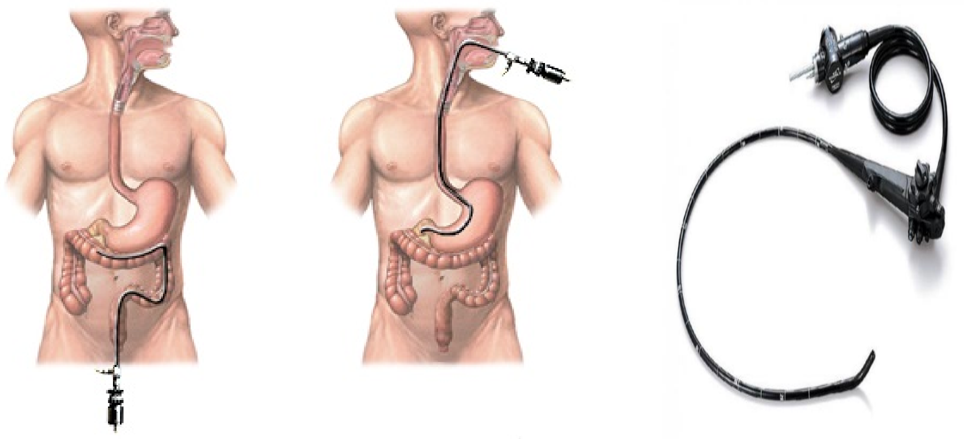
1 Introduction

In humans, gastrointestinal tract (GI) diseases are very common and cause significant healthcare concerns [1]. There are about 8 million deaths worldwide caused by gastrointestinal diseases [2]. In the GI tract, several severe disorders can occur without obvious symptoms. These include obscure GI bleeding [3, 4], ulcer infection [5], and benign and malignant tumours [6]. Consequently, gastroenterologists often recommend endoscopic examinations of the GI tract, especially in the elderly and high-risk people [7].

Today, most gastrointestinal diseases are diagnosed and examined by endoscopy [8, 9, 10]. There are three types of GI endoscopy: gastroscopy, colonoscopy, and wireless capsule endoscopy (WCE). In general, each of these methods is more beneficial for a specific section of the GI tract. Gastroscopy examines abnormalities of the upper GI tract, while colonoscopy is used to examine lesions in the lower GI tract. Conventional endoscopy [11] approaches (gastroscopy and colonoscopy) are also useful for routine GI screenings. A long, flexible tube is attached to a tiny camera that allows these techniques to visualize the digestive system [12]. However, the small intestine remains inaccessible [13]. Fig. 1.1 illustrates the conventional endocscopy.

There is the potential for capsule endoscopy (CE) to cause discomfort or harm patients; however, they can be used to perform real-time examinations and visualize many gastrointestinal tract disorders. GI inspections and scanning of inaccessible areas are possible with WCE, a non-invasive approach that can also detect lesions, which are characterized by abnormal changes or damage to the tissue in the small intestine between the upper and lower GI [1]. The capsule is designed to be swallowed by the patient and then travel throughout the GI tract. Images are continuously captured by the capsule's camera and transmitted to a data logger outside of the patient's body [14]. With WCE, pains and problems with sedation associated with conventional endoscopy may be eliminated [12]. The capsule endoscopy and

[1] https://medlineplus.gov/ency/presentations/100162_3.htm



(a)

(b)

Figure 1.1: a) Human GI system with the upper and lower endoscopy [1]b) Conventional endoscopy probe

how it functions is shown in Fig. 1.2.

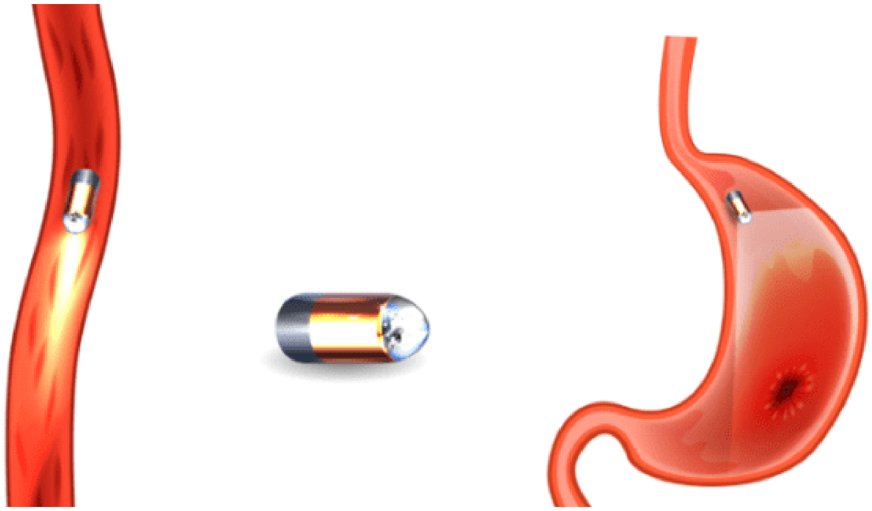


Figure 1.2: Wireless Capsule Endoscopy[2]

Endoscopic examinations that involve GI imaging produce a massive number of images, especially those involving WCE, which can produce between 50,000 and 120,000 images during a single procedure. Because of the vast amount of endoscopic image data, it is easy to make a misdiagnosis and decrease the level of diagnostic accuracy if we rely solely on our human ability to interpret the data. It would also be inefficient and time-consuming if we wanted to extract a few crucial images from a large number of endoscopic images

[2] https://www.gastrobh.com/contents/diagnostic-procedures/capsule-endoscopy-eso-pill

for physicians' consideration. As a result of these issues, computer-aided diagnosis (CAD) systems have been developed, which provide physicians with an objective reference when selecting, identifying, and classifying lesion images. By using these systems, lesion images can be selected, identified, and classified automatically, increasing the efficiency and accuracy of diagnosis in a less time-consuming manner [1].

There is a growing interest in artificial intelligence (AI) in clinical medicine, particularly in GI tract related problems. GI endoscopy can be made more efficient and effective by using artificial intelligence. Stress, fatigue, and limited experience can affect human brain performance in a negative way. AI will overcome the limitations and errors of human capability, provide higher speed, enhanced accuracy, and consistency, allowing endoscopic procedures to be performed more efficiently and with higher quality [15].

The term artificial intelligence refers to complex computer algorithms that are able to replicate some of the cognitive functions of the human brain, including the ability to learn and solve problems [16]. Their capability is based on what they receive as information and what they learn. Machine learning (ML) is the core principle behind this technology, which refers to the process of instructing computer algorithms to recognize patterns in data. Without being directly programmed, it allows them to learn and improve automatically [15]. By using ML, the computer learns how to understand previously unknown information and is able to produce reliable results when applied to new data, which were not included in its training dataset [17].

Deep learning (DL) is a derivative of machine learning that has emerged recently that is more powerful than ML. In Fig. 1.3 the relation between AI, ML, and DL has been demonstrated. DL extracts discriminative attributes automatically through an artificial neural network (ANN), which is typically a convolutional neural network (CNN) with many layers of nonlinear functions. [16, 18]. In computer-aided diagnosis systems, artificial intelligence, machine learning, and deep learning are increasingly being integrated to improve GI disease recognition and characterization [19]. These technologies are becoming more valuable every day as they have many applications in medicine.

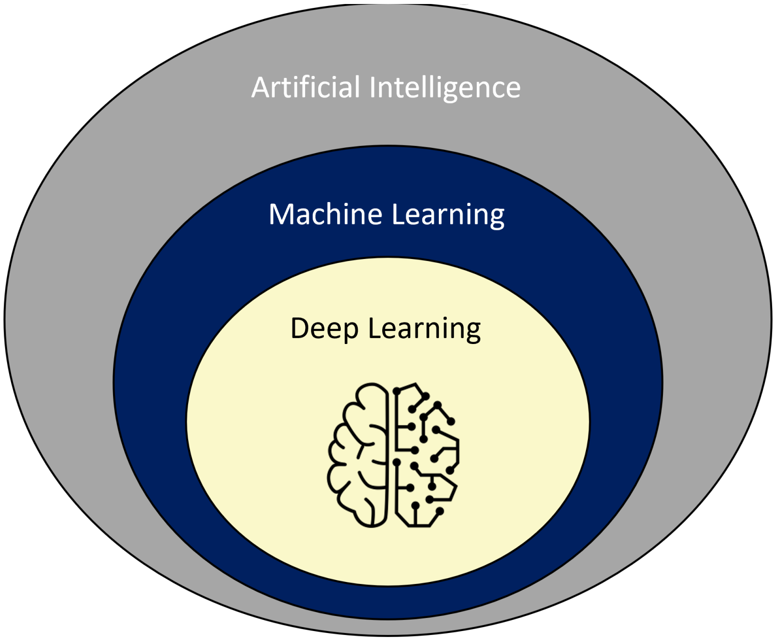


Figure 1.3: Relationship between AI, ML, and DL

1.1 Research Objectives and Questions

This study proposes a novel method of classifying the anatomical locations of the GI tract. In the medical field, the ability to detect the location of frames is a significant advantage. Doctors use the location of the abnormality to determine which treatment plan is most appropriate [20]. When performing surgery, it can be beneficial to have this information [21]. Defining the anatomical location within the GI tract effectively reduces the need for repetitive endoscopy procedures. It also allows the temporal tracking of abnormalities in the GI tract. Anatomic localization can be used to send capsules containing medications for drug delivery [6] and to automatically navigate an endoscope device to a specific location during an examination [22]. In addition, there are diseases that more typically occur in specific regions of the GI tract [7]. Dangerous bleeding, for example, usually originates in the stomach, small intestine, or duodenum. It may thus reduce the amount of time and human error associated with examinations if we provide locations along with frames, especially in high-risk anatomic areas.

In this research, a novel classification method is proposed to classify the anatomical locations of the GI tract. Among the earlier artificial intelligence studies in this field, the most notable flaw was the lack of properly labelled and available data. As a result, to address this problem, in our method, the first step is to take advantage of the vast amount of unlabeled data, which is publicly available. This is done through transfer learning. Afterwards, to improve the learning process, two networks extract features with different characteristics. The final decision is based on a combination of both of the feature vectors using ensemble learning.

Overal, the contribution of our work to the field of gastrointestinal image classification is significant for several reasons. The method has the advantage of providing high levels of accuracy in predictions, which are crucial when making clinical decisions using images. As a result, our model can reliably identify the image's location in the gastrointestinal tract, making it useful for medical professionals.

Secondly, our paper categorizes images into six locations, which is a challenging problem compared with previous works which categorized images into fewer categories. This work provides a more comprehensive understanding of the gastrointestinal tract, and is a significant advancement in medical image classification through its approach to this more complex problem.

Third, our method uses relatively few images while achieving high levels of performance. This is imperative due to the limited availability of labeled medical images, making it difficult to develop accurate models. The power of transfer learning and ensemble learning allows us to produce high levels of accuracy from a smaller number of images. This is because we leverage a large number of unlabeled data. This makes our method more accessible to medical professionals.

With these features in place, our work has made the following contributions:

- To develop a method for the anatomical classification of the GI tract.
- To develop a method to extract relevant knowledge from unlabelled data and use it to enhance the learning process in the classification task.
- To develop a method for enhancing prediction accuracy through ensemble learning, which integrates multiple learning models.
- To validate the classification method by measuring various metrics and to investigate its performance with respect to related state-of-the-art studies.

1.2 Thesis Organization

The structure of the thesis is as follows:

- **Chapter 1: Introduction**: The purpose of this chapter is to provide an overview of our research and to demonstrate its importance and contribution to the field.
- **Chapter 2: Related Work**: In this chapter, several studies on localization techniques, transfer learning, and ensemble learning are reviewed. Based on the shortcomings in the field, we have developed a method that aims to improve the accuracy of the task.
- **Chapter 3: Materials and Dataset Preparation**: This chapter describes the two main phases of data collection and preparation and the proposed anatomical classifier. First, the dataset and its preparation steps have been comprehensively discussed. Then, we explained the base concepts that are the building blocks of our method, including transfer learning through a trained decoder, the ensemble learning technique which we used to improve prediction and the Xception network architecture, which has been used as one of the main networks. Our deep learning approach was then developed using these concepts. As a final step, we review how the model was trained and how the results were evaluated.
- **Chapter 4: Results and Discussions**: Throughout this chapter, we present the results of our work regarding the anatomical classification of the gastrointestinal tract. We demonstrated that our method is highly accurate at classifying images into six different locations. As part of a comparison with state-of-the-art methods, we discussed the proposed method's performance using several metrics, and finally, to give an understanding of how our work compares to similar works, we provide a detailed comparison with similar works.
- **Chapter 5: Conclusion**: The conclusions presented in this chapter outline the key contributions of our proposed method as well as its advantages compared with previous research. Following this, we suggest further research and improvements based on our findings.

2 Related Works

Reviewing the literature on two principal subjects is necessary to show how our work contributes to current studies-the anatomic classification of the GI tract, and the use of transfer learning and ensemble learning in AI driven tasks. The first part introduces various methods of detecting or classifying anatomical locations that have been previously proposed. Among them, the methods we used for comparison have also been described. Then, in the second section, the advantages of using the main technique in this study were investigated in various tasks by reviewing the related papers.

2.1 Endoscopic localization

Endoscopic localization methods can vary based upon the techniques used which can be radio frequency [23, 24], magnetic [25, 12], image processing [26, 27, 28, 29, 30, 31, 32, 13, 33], etc. Depending on the type of output they have, they can also vary. For example, the output of [29, 30, 31, 32, 13] are predicted locations while [26, 27, 28] have estimated the median error.

For example WCE takes pictures and sends them by using RF signals. The capsule already contains all the necessary equipment for RF-based localization methods, and this fact makes RF-based localization worth exploring for researchers. Among the techniques investigated are Time of Arrival (TOA), Direction of Arrival (DoA), Received Signal Strength Indicator (RSSI), and Radio Frequency Identification (RFID)-based methods. In [23] Cramer-Rao bounds (CRB) are derived for location estimators using measured received signal strength (RSS). They calculate bounds on location estimators in three digestive organs: the stomach, small intestine, and large intestine utilizing a 3-D human body model conducted with an application of full-wave simulation software and a log-normal model for RSS propagation. The authors analyze the factors that affect localization accuracy. The number of pills cooperating, the topology of the external sensor array and the random variation in the transmitting power of nodes are all included in this analysis. [24] introduces a localization technique for tracking capsules using RF that integrates directional-of-arrival, time-of-arrival, and inertial measurement unit measurements using the Kalman filter. A signal is transmitted from an emitter to the WCE, and a response is sent back from the WCE to a set of antenna arrays. A measurement is then taken of the round-trip TOA and the DOA of the signals transmitted from the WCE to the antenna arrays. An error of up to 10 mm is associated with their method.

Magnetic positioning technology is another method for capsule localization which detects the change of the magnetic intensity generated by a magnet attached to the capsule [34]. Magnetization fields are less affected by the unstructured environment of the GI than RF signals [12]. To reduce drifting caused by geomagnetic noise, a localization system with noise cancellation is proposed in [25]. Therefore, the initial guess will be obtained in a simple and effective way. With the proposed system, positioning and orientation errors are lower than those seen in prior works with the same configuration. As a result, the localization system can be worn while the patient moves around. Furthermore, the proposed algorithm makes positioning errors more consistent within the localization region. According to the proposed algorithm, the average positioning error obtained from 16 digital magnetic sensors in a volume of 380 mm by 270 mm by 240 mm is around 10 mm, and the average orientation error is around 12 mm. The authors present a hybrid method called MagnetOFuse in [12] to localize the capsule within the human gastrointestinal tract. A magnet and four cameras are mounted on the side walls of the capsule. With nine three-axis Hall effect sensors, magnetic localization determines the capsule's global location. They have implemented low-resolution monochromatic cameras along the capsule's side to measure the capsule's movement and increase tracking accuracy. It is shown by the results that the proposed hybrid method can compensate for the relative movement of GI tracts and has a 3.5 mm position error on average.

There have been a number of studies on localization through image processing techniques. This method does not require any extra equipment, so it would be a suitable way to estimate locations. Vu et al. [26] present an efficient method of segmenting the digestive organs using video capsule endoscopy. In this method, the colors of digestive organs are analyzed based on their unique characteristics. As the first step, they present a color model containing the color components of GI walls and non-wall areas. Each color component is analyzed according to the distribution along the time dimension of the wall regions extracted from images. It is used to learn which colors are dominant enough to discriminate between digestive organs. In order to detect the boundary of two adjacent regions, the strongest candidates are combined to create a representative signal. A method for automatically discriminating between esophageal, stomach, small intestine, and colon tissue is described in Mackiwicz et al. [26] work. In their study, they demonstrate that adding texture and motion features to their classifiers improves performance. To process the WCE image, they first divide it into sub-images and reject sub-images that lack clear tissue definition. They create a feature vector based on color, texture, and motion information from the entire image and valid sub-images. The hue and saturation histograms are used to generate color features, which are then compressed by an adapted discrete cosine transform and principal component analysis. Local binary patterns are used to create a second feature that combines color and texture information. A hidden Markov model is employed to segment videos into meaningful sections based on support vectors or multivariate Gaussian classifiers.

An analysis of the movement of the endoscopy capsule was presented by Bao et al. [33]. A Kernel Support Vector Machine (K-SVM) is used to segment endoscopic images into sub-regions and classify them. They used a quantized feature vector, which has a naturally resistant characteristic to noise, to better represent the image. Furthermore, the kernel function transforms low-dimensional feature vectors into higher-dimensional hyperplanes for non-linear decision-making. The results of the experiment indicate an accuracy of 93% for speed estimation and a localization error of 2.49 cm. Cunha et al. [28] presented a Bayesian and support vector machine approach and compared to segment the gastrointestinal tract into its four major topographic areas. Consequently, clinically relevant gastric and intestinal sections and transit times can be automatically determined. This can reduce the time it takes to annotate exams. Based on color change pattern analysis, Lee et al. [29] propose a technique for segmenting WCE videos into anatomic parts using the energy of muscular contractions. Basic to the concept is that different sections of the digestive system have different patterns of intestinal contractions that can be used as features. WCE video contractions are first characterized in the frequency domain using the energy function. The WCE video is then segmented into events using the high frequency content function. Either an entry has been detected in the next organ or an unusual event has been detected at the event boundary. Segmented events are classified into higher level events that represent GI locations. Researchers report an event detection algorithm that has a recall and precision of 76% and 51%, respectively. Compressed domain color (CDC) descriptors are compared in Marques et al. [30] with traditional full decoded images, to determine the anatomical classification of wireless capsule endoscopy images. According to the results based on 26469 images divided into stomach, small intestine, and large intestine, the difference in classification accuracy is less than 1%. Moreover, these findings indicate that errors are mostly located near zone transitions, which can be addressed with other visual descriptors like shape and motion. The authors conclude that they can use CDC when it comes to this type of classification while sacrificing a small amount of accuracy.

Shen et al. [31] proposed an unsupervised learning approach that employs Scale Invariant Feature Transform (SIFT) for extracting local image features and probabilistic latent semantic analysis (PLSA) for clustering data. Using the GoogleNet architecture, Takiyama et al. [32] designed a CNN-based diagnostic program. They trained their model using an esophagogastroduodenoscopy (EGD) dataset containing 27,335 images from four main anatomical locations of the larynx, the esophagus, the stomach, and the duodenum, as well as three subsequent subclassifications for stomach images, including upper, middle, and lower regions. They have achieved overall accuracy of 97% with area under the curve (AUC) of the receiver operating characteristics (ROC) 1.00 for larynx and esophagus images, and 0.99 for stomach and duodenum images and 0.99 for the upper, middle, and lower stomach. Using 9,995 colonoscopy images, Satio et al. [13] trained a CNN, then tested its performance on 5,121 colonoscopy images that include seven different locations: the terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, the anus, and indistinguishable parts. Their system is based on the GoogLeNet deep CNN without any modifications. An analysis of the concordance between diagnoses made by endoscopists and the diagnoses made by CNN was conducted. A primary objective of the study was to measure the sensitivity and specificity of CNN for identifying anatomical categories in colonoscopy images. By calculating the area under the curve for each location, they assessed the performance of the constructed CNN. 0.979 was obtained for the terminal ileum, 0.940 for the cecum, 0.875 for ascending colon to transverse colon, 0.846 for descending colon to sigmoid colon, 0.835 for the rectum, and 0.992 for anus. According to the CNN system, 66.6% of images were correctly classified during the tests.

Table 2.1: Comparison of the previous works on the localization of the GI tract

Ref	**ML (DL)**	**Method**	**Number of anatomical locations**	**Anatomic locations included**	**Performance metrics and results**
[29]	No (No)	Variation in HSV intensity in subsequent frames using event correlation	4	Esophagus, stomach (entering stomach), small intestinal (entering duodenal and ileum), and colon	Recall: 76%; Precision: 51%; F1-score:61%
[33]	No (No)	Feature Points Matchingfor capsule speed estimation	-	Speed estimation accuracy and location error	93% accuracy for speed estimation and 2.49 cm for localization error
[26]	No (No)	PCA and customized thresholding approach with color features	2	Median error in frame number prediction for detecting pylorus; ileocecal valve	Pylorus:105; ileocecal valve: 319(frames)
[23]	No (No)	Using RSS, DoAor ToA	-	average RMSE for predicting capsule location	100 mm RMSE with 10 sensors on body surface
[25]	No (No)	Adding small magnet in capsule	-	Capsule inside a volume of 380 mm by 270 mm by 240 mm covered by 16 digital magnetic sensors	10 mm RMSE error
[27]	Yes (No)	multivariate Gaussian classifiers with color, texture,motion features	3	Median error in frame number prediction for detecting esophagogastric junction; pylorus; ileocecal valve	Esophagogastric junction: 8; pylorus: 91; ileocecal valve: 285 (frames)

Table 2.1: Comparison of the previous works on the localization of the GI tract

Ref	**ML (DL)**	**Method**	**Number of anatomical locations**	**Anatomic locations included**	**Performance metrics and results**
[28]	Yes (No)	SVM with color and texture features	3	Median error in frame number prediction for detecting esophagogastric junction; pylorus; ileocecal valve	esophagogastric junction: 2; pylorus: 287; ileocecal valve: 1057 (frames)
[30]	Yes (No)	SVM with color features	3	Stomach, small intestine, and large intestine	85.2 % (overall accuracy)
[31]	Yes (No)	SIFT features matched using random sample consensus and tracked using Kanade-Lucas-Tomasitracker	-	Robotic-assisted setup provided for evaluation	2.70 ± 1.62 cm localization error
[32]	Yes (Yes)	CNN	6	Larynx, esophagus, stomach (upper, medium, lower), duodenum	AUC: 100% for larynx and esophagus 99% for stomach and duodenum Accuracy: 97%
[13]	Yes (Yes)	CNN	7	The terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, the anus, and indistinguishable parts	AUC: 97% for the terminal ileum; 94% for the cecum; 87% for ascending colon to transverse colon; 84% for descending colon to sigmoid colon; 83% for the rectum; 99% for the anus. Accuracy: 66%

2.2 Transfer Learning and Ensemble learning

A recent method of machine learning is transfer learning. This method involves reusing a model created for one task as a basis for another [35]. It can be difficult or expensive to collect training data in some cases. Hence, it may be necessary to train high-performance learners using easy-to-obtain data like unlabeled images from different domains [36]. In transfer learning, knowledge is extracted from a task in the form of parameters, features, samples, instances, etc., which are then applied to a new task. Transfer learning can be applied to supervised learning, unsupervised learning, and reinforcement learning. In unsupervised transfer learning, which is usually a form of clustering or dimensionality reduction [35], knowledge is used in order to improve the learning of the target predictive function. During training, no labelled data is present in either the source or target domains [37].

Several studies and applications have demonstrated the benefits of transfer learning in the gastrointestinal tract. A major benefit of transfer learning in the area of the gastrointestinal tract is the ability to improve accuracy and performance during diagnostic procedures. An example of transfer learning in practice is the use of pre-trained deep learning models that are fine-tuned based on smaller datasets of endoscopic images. In comparison with training from scratch, this approach improves classification accuracy significantly. Transfer learning also reduces computational costs and time in the field of the gastrointestinal tract. By utilizing pre-trained models, the low-level features in the data are already known, so they can be fine-tuned for the new task rather than having to learn them all over again. Moreover, this process transfers knowledge, helping to improve accuracy and performance without requiring extensive training data.

Sometimes, depending on the problem we are trying to solve and the quality and quantity of the available data, one classifier is not enough in machine learning. Using only one classifier can result in errors or shortcomings because it can only reflect one aspect. The classification can be improved if multiple learners work together. A stronger classifier can be created using ensemble learning [37]. In the field of the gastrointestinal tract, ensemble learning has been widely used to predict various outcomes, such as the risk of disease, the severity of disease, and the response to therapy. Ensemble learning has been shown to consistently improve the accuracy of predictions compared to individual models. It can also increase the stability of the final prediction by reducing the variance of individual models. This can be especially useful in the field of the gastrointestinal tract, where the prediction of disease outcomes is often complex and subject to a high degree of variability. Another benefit of ensemble learning is to improve the robustness of the predictions of outliers by combining the predictions of multiple models that have different sensitivities to outliers.

The aforementioned benefits make transfer learning and ensemble learning attractive approaches for researchers and practitioners working in the field of the gastrointestinal tract. Our anatomical classifier for the GI tract was developed using transfer learning and ensemble learning. These two methods of machine learning and their benefits for related tasks are reviewed in the following paragraphs.

Using the K-Vasir capsule endoscopy data set, [38] evaluated how transfer learning affects gastrointestinal disease classification in wireless capsule endoscopy images. Fine-tuning the pre-trained ResNet50 is the basis of the proposed method. The pre-trained model ResNet50 is fine-tuned via transfer learning in order to extract deep features from WCE images. With the modified deep convolution neural network ResNet50, deep features from HSV images are extracted and then fed into a softmax classifier for gastrointestinal disease detection. As a result of their approach, high accuracy was observed when classifying wireless capsule endoscopy images into either ulcer, polyp, and normal.

A convolutional neural network and transfer learning type fine-tuning are used in [39] to aid medical diagnostic processes for diseases and anomalies related to the gastrointestinal tract. The proposed method mainly relies on transfer learning via fine-tuning the VGG16 convolutional neural network, which has been trained with ImageNet datasets. It was possible to achieve high-level representations of endoscopy images by learning from natural images and transferring that knowledge to the medical field. The Kvasir dataset was used to evaluate the proposed strategy and achieved 94.6% accuracy. As a result, machine learning and transfer learning remain promising alternatives for supporting fast and accurate medical decisions.

Using deep learning techniques, Qiaosen et al. [40] seek to demonstrate the potential for detecting GI diseases. Using WCE image data, this study confirms that their method which is based on convolutional neural networks can detect GI diseases using integrated frameworks for convolutional neural networks. In this study, transfer learning techniques are used, which have been shown to enhance the learning abilities of models. An ensemble learning technique is also used in this technique, which improves the performance of single classifiers. Three convolutional neural networks serve as the underlying backbone for predicting images, which are also known as independent base classifiers. Using pre-trained weights, the backbone networks use transfer learning to accelerate model convergence. After training the model on a broader training set, it is migrated, applied to a smaller, more specific model, and modified as necessary. The classification results are then used to feed into an integrated classifier containing a number of fully connected layers in order to obtain the final prediction. The process is an example of ensemble learning, which improves the prediction performance of a single model by training multiple models and combining their predictions. In general, the prediction performance of a multiclassifier system will be better than that of an independent classifier system. This is because the error of the independent classifier is likely to be compensated for by the other classifiers. A 94.9 percent accuracy rate is achieved with the proposed methods on the test dataset.

Ghosh et al. [41] presents a computer-aided diagnostic tool for the automated analysis of CE images. In this tool, abnormalities of the small intestine are detected, particularly bleeding. The tool is based on a CNN deep learning framework, which uses a pre-trained AlexNet neural network to train a transfer learning CNN to identify bleeding and non-bleeding CE images. The bleeding zones within the bleeding images are also identified using a deep learning-based semantic segmentation technique, which leverages the SegNet deep neural network. The performance of the proposed tool was evaluated using two publicly available clinical datasets. Compared to manual inspection and annotation of CE images by a physician, the proposed framework reduces annotation time and human labour costs, increases detection accuracy, and provides additional benefits such as bleeding zone delineation.

Almanifi et al. [42] used ensemble methods to aid in diagnosis by using three pre-trained models. The models were trained on an 8000-image Kvasir dataset, which consisted of eight classes of different parts of the GI tract along with different diseases. ResNet50, MobileNetV2, and Xception, three popular pre-trained CNNs were used. Compared to other studies that utilized the same dataset, the ensemble method significantly increased the prediction accuracy of all three models. It demonstrates how the ensemble method has great potential in medical machine learning. Based on the model's performance, 99.2% accuracy, 0.9977 AUC, and 99.29% F1 score were achieved.

Nadeem et al. [43] present a model for classifying gastrointestinal abnormalities using endoscopic images. A dataset was provided by the MediaEval Benchmarking Initiative for Multimedia Evaluation for the purposes of training and testing. Machine learning, ensemble learning, and multimedia content analysis were used for classification in the study. Considering the composite nature of features, they train separate models using logistic regression and kernel discriminant analysis using spectral regression for each one. Predictions were then made using the ensemble technique. Based on the ensemble method, logistic regression was found to give the best results using six different features, including Local Binary Patterns and Haralick textures. On testing data, it achieved an accuracy of 94% with an F1-score of 0.76 and an MCC of 0.73.

Vieira et al. [44] discuss a two-step procedure for automatically detecting tumours in WCE based on the region of interest selection and classification. Automatic segmentation based on a Gaussian Mixture Model is used in the first step to separate abnormal from normal tissue. As a second step, they propose an ensemble system with a partition of the training data and a new training scheme for the ensemble system. This paper contributes to the ensemble-based classification module, which guarantees no significant loss of diversity during incremental training by using data partitioning and ensemble structures. Results from experiments demonstrate the superiority of the proposed algorithm over state-of-the-art methods in three main areas: feature extraction, tissue separation, and classification. It was found that the proposed feature set and classification module improved accuracy by 1.7% and 1.2%, respectively.

Khan's research [45] aims primarily to detect diseases and abnormalities in the Gastrointestinal Tract with the use of multimedia data. A traditional colonoscopy is used to collect this data. To extract information from visual data, multimedia content analysis techniques have been applied, as well as machine learning techniques for classification. In this study, the VGG 19 model is used to extract plentiful visual concepts by using pre-training data from the ImageNet challenge and retraining of the last 2 layers with medical images. Based on the features that have been extracted, logistic regression, random forest, and extremely random tree classifiers are trained. Pre-computed texture features and VGG features, the features extracted from VGG19 pre-trained models, were included in the feature set. As a result of ensemble training, the final model utilizes weighted majority voting among all independent models. Using majority voting of logistic regression, random forest, and extra trees classifiers, 98% accuracy, 0.76 F1, and 0.75 MCC are achieved on testing data.

Rezaei et al. [46] aim to predict the probability of gastric cancer occurrence and deaths associated with it. Their ensemble method combines multiple machine learning methods to achieve this goal, such as logistic regression, random forest, gradient-boosted decision trees, and deep neural networks. Using ensemble methods, they aim to minimize prediction errors for a large number of patients' features. Their results demonstrated the superiority of the ensemble method in predicting gastric cancer and deaths associated with it compared with other machine learning-based methods. It has been proven again that ensemble learning is feasible and effective in the gastrointestinal field.

The literature review concludes that gastrointestinal tract location classification is important for medical purposes. It has highlighted the range of methodologies employed, their level of achievement, and the advantages of ensemble and transfer learning. There has been evidence that applying transfer learning and ensemble learning to this field improves the performance of models. On the basis of these findings, we will present a novel method of predicting gastrointestinal anatomical locations using transfer learning and ensemble learning in the next chapter. In contrast to previous methods, we propose a method that overcomes challenges faced in previously classified GI tract locations, including limited labelled images and the use of private datasets, low prediction rates, and having fewer categories. We achieved high accuracy while classifying six locations by using transfer learning and ensemble learning techniques, surpassing the number of categories in most related works' classifications with much less labelled data. As a result of the application of these techniques, a more efficient and accurate method has been developed, providing significant benefits to the field. With further research and development, these approaches can be refined to improve localization accuracy and enhance the effectiveness of medical diagnosis.

3 Materials and Dataset Prepration

This study proposes a method to classify gastrointestinal tract locations using ensemble transfer learning. To improve prediction performance, we have taken advantage of thousands of unlabelled data and an encoder for the purpose of transfer learning. Therefore, the encoder extracts the most relevant features and the final result is based on two different models using ensemble learning. This project was developed with Google Colab Pro, Python 3, and the TensorFlow 2.11.0 framework, along with the Keras 2.11.0 high-level API. Using a GPU provided by Google Colab, deep learning models were trained.

3.1 Dataset

This study uses the HyperKvasir dataset [47] as its main dataset. In this database, 10,662 labelled images have been stored using the JPEG format that has been classified by their location within the GI tract and the type of finding identified. The dataset also contains an additional 99,417 unlabeled images. Of the 10,662 labelled images, 4,104 have been labelled with their associated locations, while the remainder has been labelled with their respective diseases. For the purposes of this study, the 4,104 labelled images and the unlabeled portion of the dataset were utilized. Additionally, the dataset contains videos from the GI tract, although they were not used in this study. During endoscopic procedures, landmarks are characteristics of the GI tract that help with orientation. Additionally, they serve to confirm a thorough examination. Lower GI tract landmarks can be identified in the terminal ileum, colon, and rectum. Upper GI tract landmarks include the esophagus, stomach, and duodenum. These locations are demonstrated in Fig. 3.1 as well as a few image samples from the Kvasir dataset in Fig. 3.2.

The images and videos were collected using standard endoscopy equipment manufactured by Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany) at the Department of Gastroenterology, Bærum Hospital, Vestre Viken Hospital Trust, Norway

[47]. The dataset documentation indicates that the study took appropriate steps to ensure the ethical use of the data. Specifically, the data was fully anonymized and approved by the Privacy Data Protection Authority.

In the Table 3.1, further description of the dataset used in this study is shown. Our study used a well-known dataset that has been widely used in previous studies. We acknowledge, however, that the dataset contains some noise and similar images, which we removed by carefully reviewing and removing duplicates. A number of validation measures were conducted to ensure the dataset's quality and validity. Detailed information about the pre-processing steps used to get better quality data will be presented in the following sections.

Table 3.1: Description of the labelled dataset used in this study

Position	Number of images before preprocessing	Number of images after preprocessing
Cecum	1009	164
Ileum	9	52
Pylorus	999	234
Rectum	391	87
Stomach	764	172
Z line	932	162
Total	4104	871

Overall, the dataset used in this work contains 4,104 labelled and 99,417 unlabeled images taken at six different locations in the GI tract, including the cecum, ileum, pylorus, rectum, stomach, and Z line which is the anatomical junction between the squamous epithelium of the esophagus and the columnar epithelium of the stomach. The table 3.1 shows that the Ileum class has fewer images than the other locations. A few other images were collected from this location and added to the dataset to balance this number. In the second dataset [48], images and videos were captured by endoscopy cameras at 10 different anatomical locations within the GI region of patients. Approximately 80-110 distinct patients were used to produce the image dataset, which consists of both CE and WCE frames. As part of another study [49] we conducted on the anatomical classification of the GI tract, the same dataset was used with a few-shot learning technique. This machine-learning technique involves training models to identify new classes of objects based on very few data examples. Few-shot learning algorithms can learn from just a few examples of each new class, unlike traditional supervised learning methods, which require large amounts of labelled data. Several applications of few-shot learning have shown promising results, including computer vision, natural language processing, and robotics, and it is a rapidly evolving field.

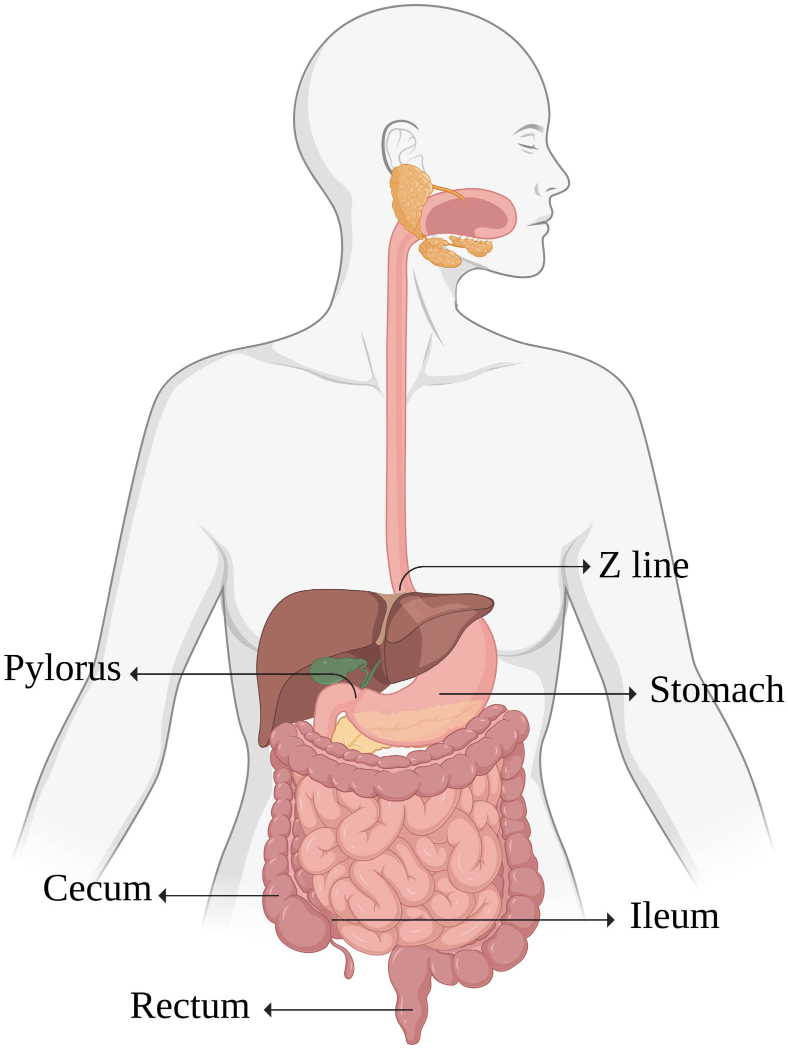


Figure 3.1: Anatomical locations covered in the Kvasir dataset and this study

It is possible for images to contain noise as well as useful information. Existing noise in an image reduces clarity and introduces misleading information. These artifacts make classification more complex [50]. An ideal denoising result would be an image that preserves only essential information. Some images in the main dataset contain noise, including black corners, green boxes and texts. These items will be considered inputs by the network while they should not have any effect on the predictions. Mispredictions will be more likely if they only appear in some categories as well. If we have a specific artifact in just one of

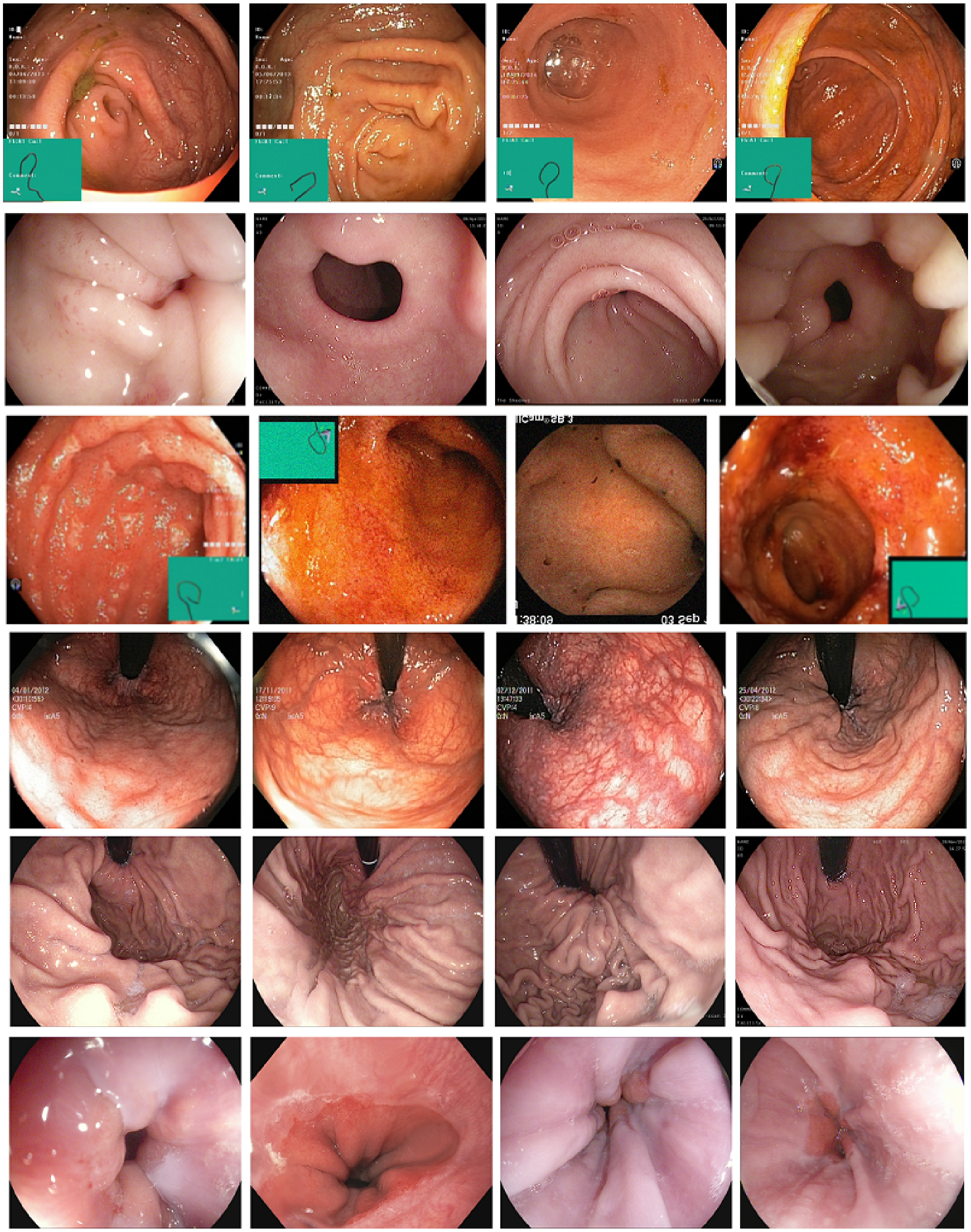


Figure 3.2: Several image samples from the Kvasir labelled dataset. From the top row to the bottom, images of the cecum, ileum, pylorus, rectum, stomach, and z-line have been illustrated, respectively.

the categories, it will produce features that the model relies on rather than the real useful features from the textures, colors, etc. This has been prevented here by cropping the images and removing all the artifacts.

Images were resized to 128×128 pixels to have the same size. A high degree of similarity was observed between images on the dataset, according to experiments. Similar images in a dataset introduce bias into the dataset, increasing the likelihood of deep neural networks learning specific patterns based on those images. This prevents the model from generalizing to new images from other datasets. Using the differential gradient hash (D-hash) Algorithm [51], similar images have been detected and removed from the dataset. D-hash is a hashing algorithm that compares the difference in gradient between adjacent pixels, generating a 64-bit image signature (a matrix with eight rows and eight columns). It is resilient to the aspect ratio, image size, brightness, and contrast differences and can be used to compare images with each other for similarity. By comparing two peak signal-to-noise ratio (PSNR) values of two distinct images from the dataset, a threshold was established, and images that showed similarity beyond that threshold were removed.

The amount of labelled data available for specialized image and video classification tasks is often insufficient. This is especially true in the medical industry, where privacy concerns restrict access to data. The goal of image augmentation is to increase the size and diversity of the training data by applying various transformations to the original images. It is another way that model overfitting can be reduced by increasing the amount of training data with information exclusively from the training data. It is not a new field, and a variety of data augmentation techniques have been used to solve specific problems. Here, we used some traditional image augmentation methods [52]. Image augmentation techniques that we utilized in this work include rotating, flipping, zooming in, and zooming out. Rotating involves rotating the image by a certain degree to increase the robustness of the model to rotations. Flipping involves flipping the image horizontally or vertically. Zoom in and zoom out involve changing the scale of the image, either by enlarging or reducing its size, and shading with a hue changes the colour intensity of the image. These techniques can be applied individually or in combination to produce a diverse set of augmented images. These augmentation techniques help to prevent overfitting and improve the generalization performance of the model on unseen data.

After dataset preprocessing, which includes noise deduction, similarity removal, and image augmentation, 871 labelled images remained for the purpose of this study.

3.2 Proposed Pipeline

A brief description of basic concepts is provided in this section, followed by an explanation of the proposed method based on the information provided.

3.2.1 Transfer Learning: using an encoder trained on unlabelled GI tract images

Data representation or features heavily influence the performance of machine learning methods [53]. As an input to a supervised predictor, a good representation is very useful and can improve predictions. While it is undeniable that the efficacy of a representation learned from data is heavily dependent on the task for which it is to be used, some representations have properties that can be applied simultaneously to many real-world problems [54]. Since there are not enough labelled images, learning useful representations from a vast amount of unlabeled data without supervision can be helpful for solving artificial intelligence problems, including classification problems. This combination of learning can be achieved through transfer learning. By transferring knowledge across tasks, a learning algorithm can exploit commonalities between them and share statistical strength across them. This makes it easier to extract useful information when building classifiers or other predictors in a supervised manner.

An autoencoder is an unsupervised learning algorithm that employs backpropagation for training. Its schematic is shown in Fig. 3.3. For a given network configuration, an autoencoder learns efficient embeddings of unlabeled data. In an autoencoder, two parts are involved, an encoder and a decoder. Using an encoder, we compress data from a higher dimension into a lower dimension referred to as latent space and conversely, the decoder converts data from a lower dimension to a higher dimension again. Using the decoder, we are able to ensure that the latent space captures most of the beneficial information from the dataset. This is because the decoder tries to output the information that was originally input to the encoder as accurately as possible.

Additionally, autoencoders can be used to improve the quality of data. Data quality has a significant impact on the outcomes of data-driven approaches. There is a strong

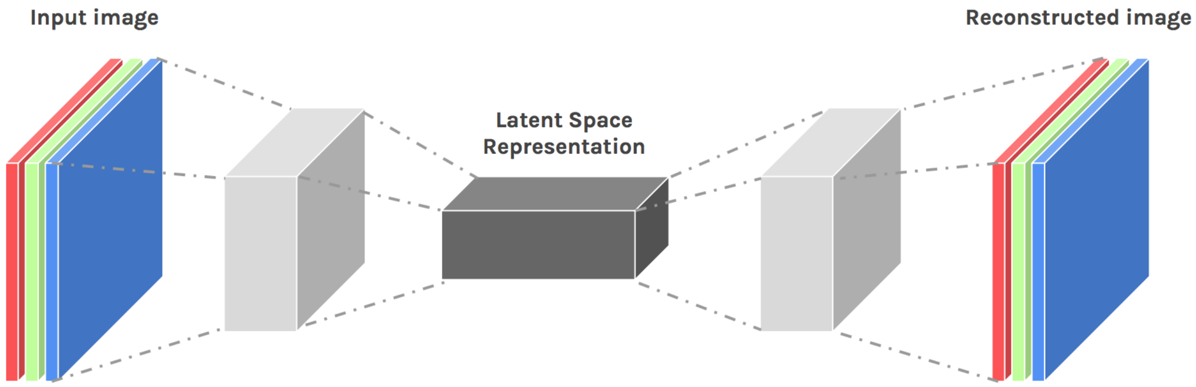


Figure 3.3: The block diagram of an auto encoder

correlation between better features and better results. Because the data contains unnecessary and redundant information, data quality can't always be guaranteed. Information that is not useful will not only decrease detection accuracy but also increase processing time. As long as it eliminates useless information and keeps as much influential information as possible, dimension reduction may improve classification accuracy [55].

The entire encoder-decoder architecture is continuously trained on the loss function in order to ensure that the input is reconstructed at the output. The loss function is therefore defined as the mean square error between the encoder input and the decoder output. It is intended to use a very low-dimensional latent space so that maximum compression can be achieved at the same time as minimizing errors. However, there are no limits on the values of the latent space, but information will be lost if its dimensionality is reduced beyond a certain point. Any dimension can be used in the latent space as long as the decoder function can reconstruct the input.

The encoder section of an auto-encoder can be split from the encoder-decoder architecture and used as a feature extractor in another network to apply transfer learning. A more discriminating feature is obtained by an encoder, they can have the clearest boundary among the categories, and improve the process of classification [56]. Fig. 3.4 shows how the knowledge can be transferred from a trained encoder to another network. The summary of the designed autoencoder in this study and its layers has also been illustrated in table. 3.2

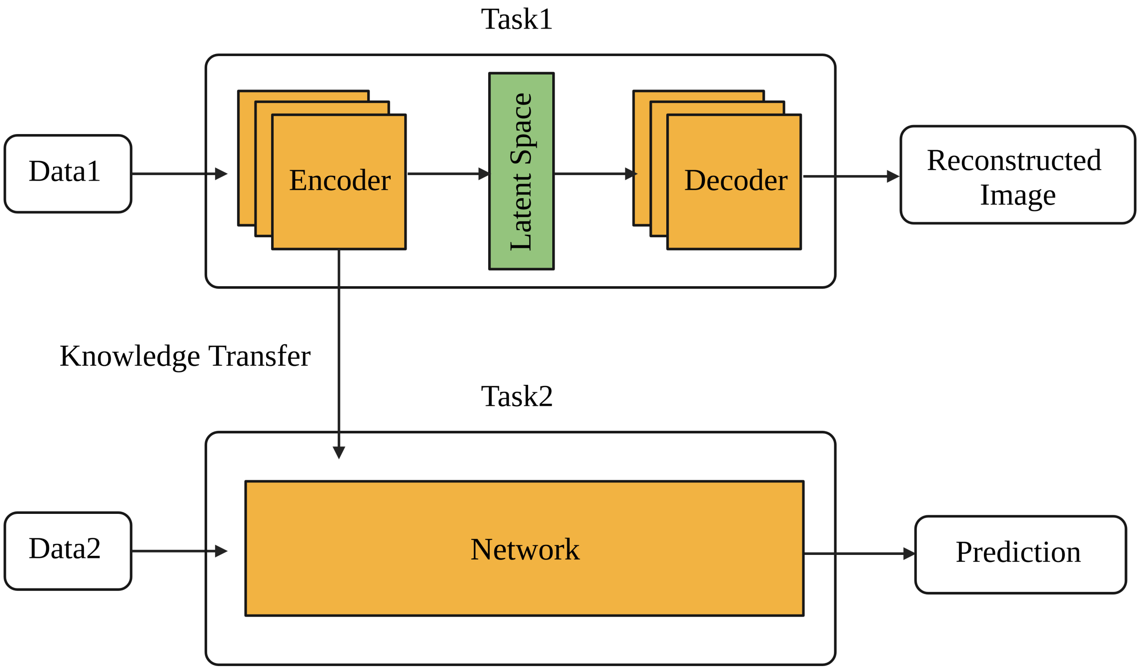


Figure 3.4: Transferring knowledge from the trained auto-encoder to another task

Table 3.2: A summary of the layers for the designed autoencoder. The number of trainable parameters are 134,724,995 and the number of non-trainable parameters are 0.

Layer (type)	**Output Shape**	**Param #**
img_in (InputLayer)	[(None, 128, 128, 3)]	0
conv2d (Conv2D)	(None, 128, 128, 128)	3584
max_pooling2d (MaxPooling2D)	(None, 64, 64, 128)	0
conv2d_1 (Conv2D)	(None, 64, 64, 128)	147584
max_pooling2d_1 (MaxPooling2D)	(None, 32, 32, 128)	0
conv2d_2 (Conv2D)	(None, 32, 32, 64)	73792
max_pooling2d_2 (MaxPooling2D)	(None, 16, 16, 64)	0
flatten (Flatten)	(None, 16384)	0
dense (Dense)	(None, 4096)	67112960
dense_1 (Dense)	(None, 16384)	67125248
reshape (Reshape)	(None, 16, 16, 64)	0
conv2d_3 (Conv2D)	(None, 16, 16, 64)	36928
up_sampling2d (UpSampling2D)	(None, 32, 32, 64)	0
conv2d_4 (Conv2D)	(None, 32, 32, 128)	73856
up_sampling2d_1 (UpSampling2D)	(None, 64, 64, 128)	0
conv2d_5 (Conv2D)	(None, 64, 64, 128)	147584
up_sampling2d_2 (UpSampling2D)	(None, 128, 128, 128)	0
conv2d_6 (Conv2D)	(None, 128, 128, 3)	3459

3.2.2 Xception: the second network used for extracting features

In 2017, Chollet proposed a convolutional neural network architecture based entirely on depthwise separable convolution layers called Xception [57]. In the Xception architecture, 36 convolutional layers are used to extract features from images, followed by a logistic regression layer for the purpose of image classification. The 36 convolutional layers can be divided into 14 modules, with linear residual connections around them, except for the first and last modules. First, the data passes through the entry flow, then it passes through the middle flow, which is repeated eight times, and then it passes through the exit flow. Batch normalization follows all Convolution and Separable Convolution layers.

As an alternative to classical convolutions, depthwise separable convolutions are more efficient in terms of computation time. First, there is a depthwise convolution layer that filters the input, and second a 1×1 convolution layer named point-wise convolution layer that creates new features by combining these filtered values. Combining depthwise and pointwise convolutions creates a ”depthwise separable” convolution block. Convolution blocks with depth wise separables perform the same function as traditional convolution blocks, but they are much faster.

On our dataset, Xception performed the best in terms of accuracy after training and evaluating several CNN models. ResNet-50 [58] had results similar to Xception among the other CNN models. Nevertheless, we decided to use Xception in our proposed method because it has a number of advantages. A major benefit is that it has a faster training time, so we could experiment more quickly. Furthermore, Xception converges faster, requiring fewer epochs to achieve high accuracy. In situations where computational resources are limited or where training time is important, this can be beneficial. We believe that Xception is the right choice for our proposed method because of its superior performance and faster training time.

The depthwise separable convolution block has been shown in figure 3.5

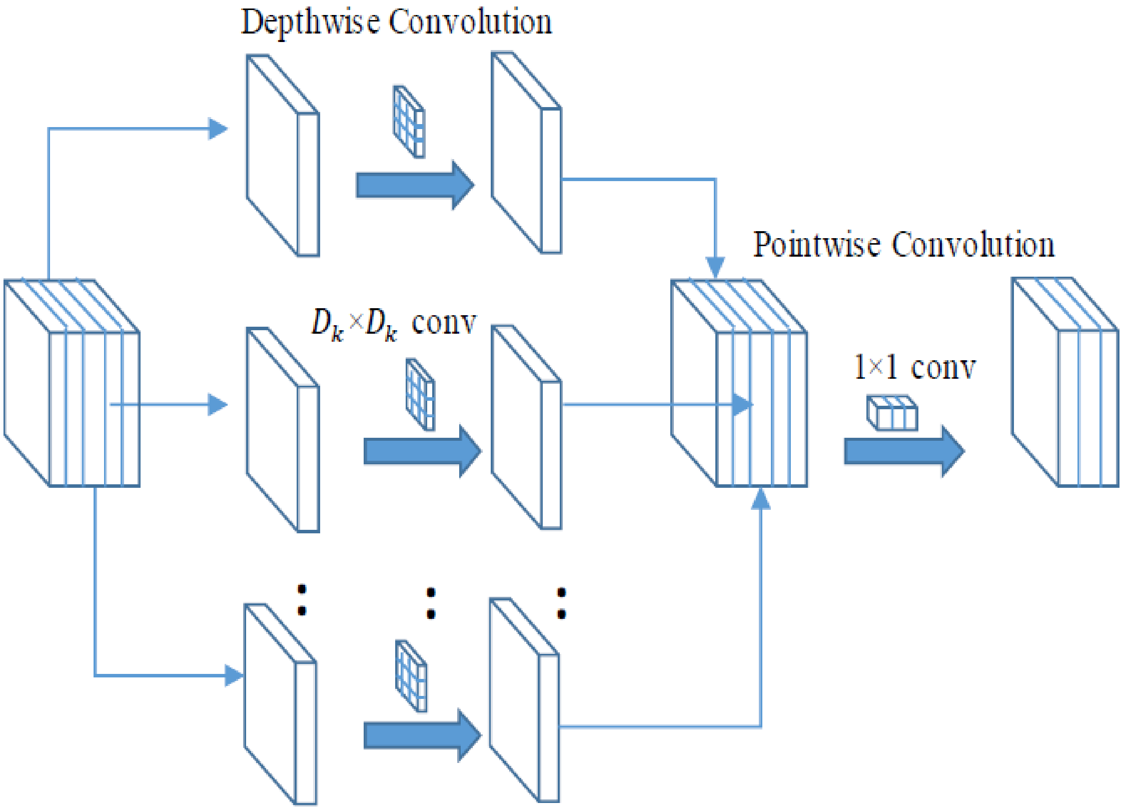


Figure 3.5: Depth wise Separable Convolution Block [1]

[1] [7]

The Xception architecture is illustrated in figure 3.6.

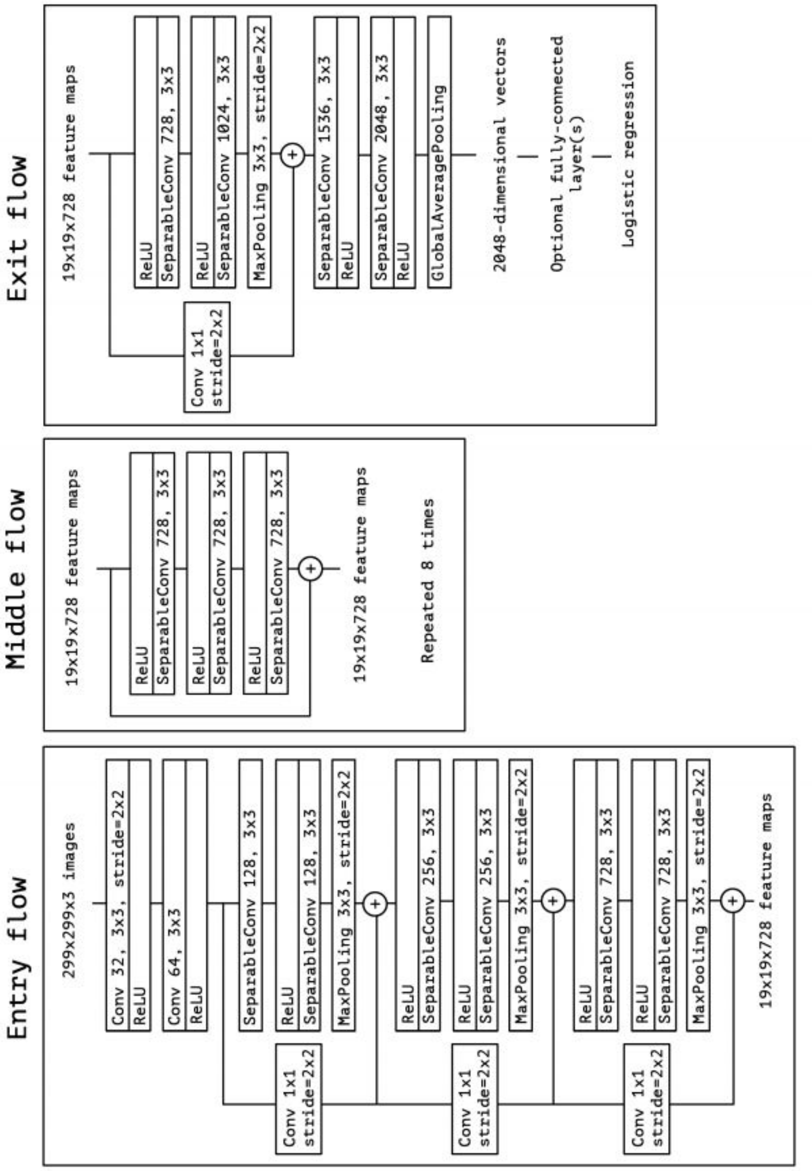


Figure 3.6: Xception architecture [2]

[2] [57]

3.2.3 Ensemble Learning: predicting by multiple learners working together

Ensemble learning has drawn increasing attention over the past decade, and researchers have made significant contributions to the field. An ensemble learning approach efficiently integrates various machine learning algorithms into a unified framework. Hence, the complementary information provided by each algorithm is effectively utilized to improve overall performance [59].

In ensemble methods, multiple learners work together to solve a problem. Instead of constructing a single learner from training data, ensemble methods aim at creating a collection of learners and then combining them. Ensembles often generalize better than base learners. Besides ensemble learning, committee-based learning or multiple classifier systems learning are also terms used to describe ensemble learning [60]. Fig. 3.7 illustrates a typical ensemble architecture.

We call these individual learners weak learners. Weak learners are models that can be used to obtain a meta-model. In this ensemble learning architecture, inputs are passed to each weak learner and each of them extracts feature vectors. We can use the combined features to build a more detailed feature to be fed into our last classifier. Features can be extracted differently based on weak learners, and this is one of the main reasons for improved prediction.

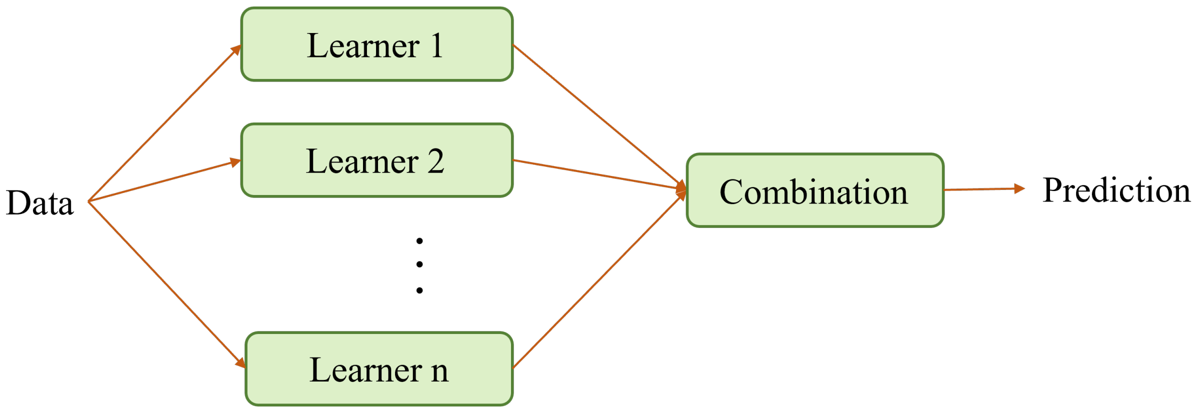


Figure 3.7: Ensemble Architecture

3.2.4 Proposed Method

An introduction to the proposed method is presented here, followed by detailed discussions of each module. To make the image ready for the system, a pre-processing stage has been performed. The resulting images are then fed into the model so that the anatomical locations of the GI tract can be determined. In the following, we describe how the classification methodology works.

An image classification problem can be solved using the convolutional neural network, one of the deep learning frameworks. In order to extract useful features from an image, convolution is employed. To overcome the problem of the small number of samples in the dataset, the proposed architecture uses the transfer learning paradigm for feature extraction. In the feature extraction block, two pre-trained models are utilized: encoder and Xception. The encoder is not trainable, and both models' top layers are omitted. Xception is a powerful convolutional neural network trained on the ImageNet dataset [61] There are a limited number of images labeled with GI tracts anatomical locations, and expanding this dataset will take specialists a lot of time and effort. There are, however, a significant number of images that are unlabeled. A new approach is proposed to take advantage of this extensive dataset. With an unsupervised method, a model is trained on an unlabeled dataset. To detect locations, trained weights are then used in the main model. With this approach and using transfer learning, we can utilize both labelled and unlabeled data. As part of this work, an auto-encoder is developed and trained to encode and reconstruct the image from the encoded data. The auto-encoder is composed of two networks, the encoder and the decoder. After training the auto-encoder, the encoder has been separated from the auto-encoder and has been used as a transfer learning tool. An encoder can learn the most informative GI tract image features since all the image information should be preserved during the encoding process. We use this encoder in our model as a pre-trained network for feature extraction.

At this point, the two models are working together and extracting features from the data. Depending on the model, features can be extracted differently, and a concatenation of all extracted features is used in the proposed model. Afterward, a fully connected layer with 64 neurons is added after the final convolution layer has been flattened. Following this, we transform these 64 neurons into 6 final neurons, each representing a different location using softmax activation. A softmax layer produces a probabilistic output for each class. Based on the value associated with the higher probability class, a prediction is also derived. Fig. 3.8 shows the proposed pipeline, including the preprocessing section as well. The different steps of the proposed method have also been listed in Algorithm. 1

Algorithm 1 Proposed Anatomical Classifier

```
SETUP
  Tensorflow and the required libraries initialization
  Load the datasets
    $unlabeled\_dataset, labeled\_dataset$

METHOD
  Train the Autoencoder on the large unlabeled dataset
    $autoencoder = train\_autoencoder(unlabeled\_dataset)$
  Separate the encoder part of the Autoencoder
    $encoder = extract\_encoder(autoencoder)$
  Use the encoder part as a feature extractor
    $autoencoder\_features = extract\_features\_from\_encoder(labeled\_dataset)$
  Use Xception model as a feature extractor
    $xception_features = extract\_features\_from\_xception(labeled\_dataset)$
  Concatenate the feature vectors obtained from the encoder and Xception
    $final\_features = concatenate\_features(encoder\_features, xception\_features)$
  Feed the final feature vector to a CNN model and classify images
    $CNNmodel.classify(final\_features)$
```

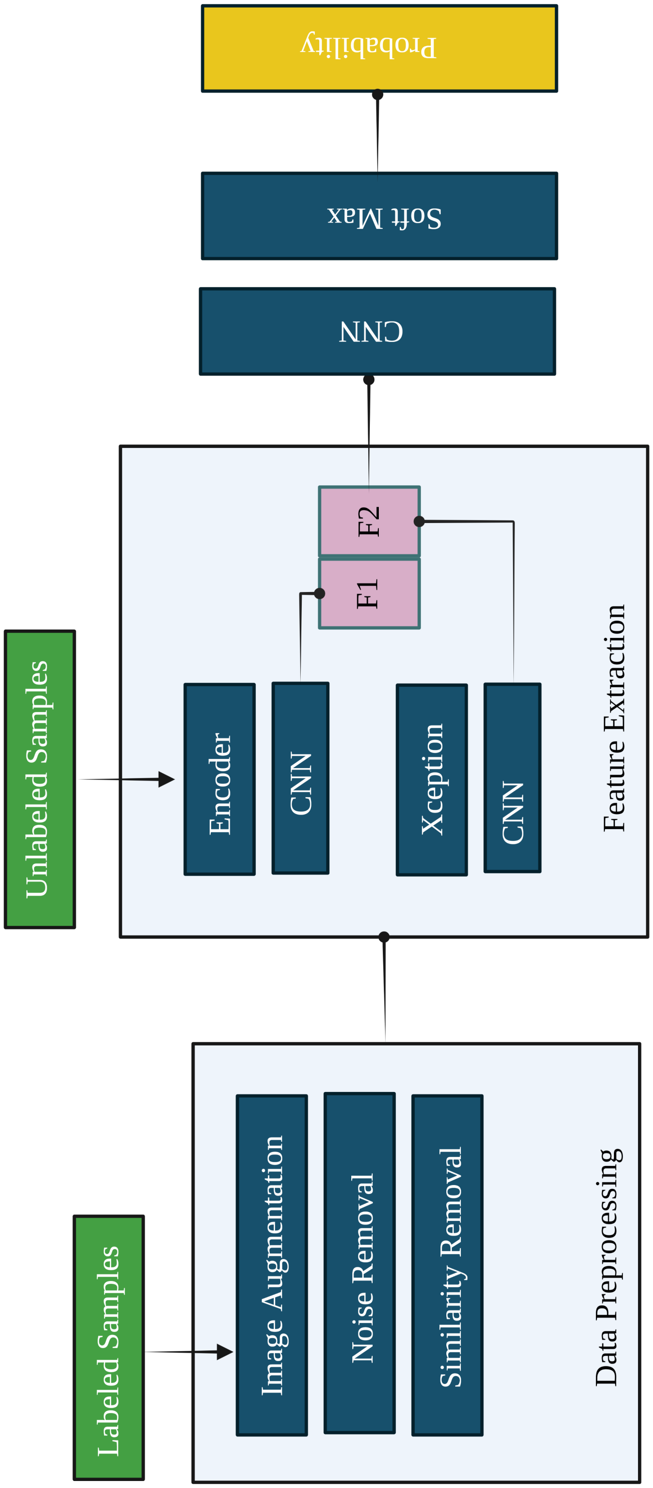


Figure 3.8: The proposed pipeline

3.3 Model Training and Performance Evaluation

In Machine Learning, the loss function represents how accurately your ML model can predict the expected outcome. The loss value is high if our model did not perform well while the low loss value indicates that our model performed well. In order to train an accurate model, it is crucial to select the right loss function. Depending on the loss function, your model may learn differently as each loss function has a specific property. In this study, for training the autoencoder, Mean Squared Error (MSE) has been used. MSE is calculated by squaring the difference between your model's predictions and the ground truth, which in this case should be the same image, square it and then average it over the entire dataset. It is formally defined by 3.1:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \mathrm{\hat{y}_j})^2 \tag{3.1}$$

Where N is the number of samples.

The Xception model was pre-trained on the ImageNet dataset. For its training process, we used the sparse categorical cross entropy loss function. Based on how far the prediction is from the actual expected value, a loss is calculated for each probability of the predicted class. Due to the logarithmic nature of the cost, large differences lead to large costs, while small differences tend to receive small costs. When a model is perfect, it has zero sparse categorical cross-entropy loss. Its equation is provided in 3.2:

$$\mathrm{L_{SCCE}} = -\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{t_i}\mathrm{log}(\mathrm{p_i}) \tag{3.2}$$

Where N is the number of samples, t_i is the true label of the i^{th} sample, and p_i is the predicted label of the i^{th} sample.

Minimizing the cost function is important because it describes the difference between the true value and the predicted value. An optimization technique minimizes the cost function by adjusting parameters. The optimization technique used here is the Adam optimization method [62]. Several deep learning applications in computer vision and natural language processing are now using the Adam optimization algorithm as an extension to stochastic gradient descent. In stochastic gradient descent, the learning rate remains constant for all weight updates. In contrast, Adam calculates adaptive learning rates based on estimates of the first and second moments of gradients for different parameters. The Adam optimizer is initially configured with a learning rate of 0.001 and a decay rate of 1e-6. A learning rate determines the step size during each iteration of the optimization process, while a decay rate determines how fast the learning rate should decay over time. Depending on the particular problem and data set being trained, the learning rate and decay rate will vary. To find the optimal values, these values have been tuned through experimentation.

We evaluated the proposed network with the 10-fold stratified cross-validation method using the stratified K-fold [63]. For the multiclass problem, along with the accuracy, the precision, recall, F1-score and the AUC are reported. AUC is a performance measurement to check and visualize multi-class classification, and additionally specifies the degree of separability. The higher the AUC, the better the model is at distinguishing between classes.

The accuracy of a model is not the only metric that can be used to evaluate its performance. Other metrics are also included in the evaluation process to explain the classification results. In the following paragraphs, they will be discussed.

Precision is defined in 3.3:

$$\frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{3.3}$$

In other words, stated to determine how accurate and precise your model is when compared with the predicted positives. The number of actual positives is stated. If there is a high cost associated with false positives, precision can be an effective and useful measure.

On the other hand, when we are more concerned with false negatives, recall can be used to select the most appropriate model. Using recall, we calculate how many of the Actual Positives in our model were correctly labelled as Positive. The formula is illustrated below in 3.4:

$$\frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{3.4}$$

In the context of medicine, recall can be an important metric, when the aim is to identify as many real positive cases as possible [64]. This means that if a sick patient is identified as negative (false negative) and does not receive treatment, the cost is much higher.

F1 which is a function of Precision and Recall is described in 3.5:

$$\frac{2\,(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3.5}$$

When comparing two different classifiers, the F1 score can be used as the most appropriate measure, since it gives a balance between recall and precision by taking their harmonic mean. Furthermore, it helps with comparisons when there is an uneven distribution of classes. If the problem is multi-class classification, the F1-score for the entire model can be calculated by taking the arithmetic mean of the F1-scores of all the classes.

4 Results and Discussion

This section discusses the performance of the proposed pipeline, which includes the data preprocessing steps and the anatomical classifier of the GI tract using ensemble learning. A numerical and visual representation of the results illustrates the contribution of the feature extractor network to the analysis. Also, a comparison between the proposed method and other studies utilizing different types of classifiers is made to see whether the proposed method is more accurate.

4.1 Experimental Results

First, we discuss the preprocessing step used to prepare the dataset, and then we analyze the network's performance.

4.1.1 Data preprocessing

The collected dataset consists of images taken at six different locations in the GI tract. In order to improve the performance of the network, image samples were processed for noise removal. A few samples have been shown in Fig. 4.1.

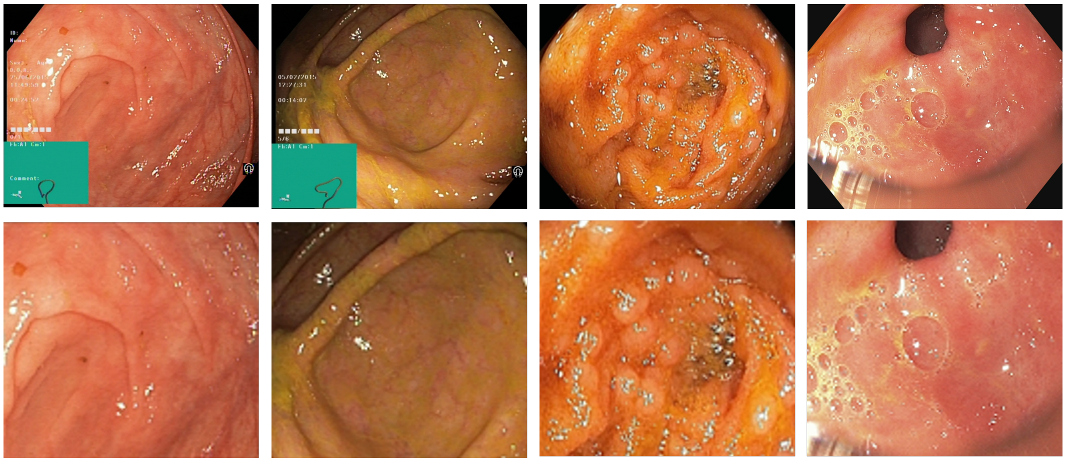
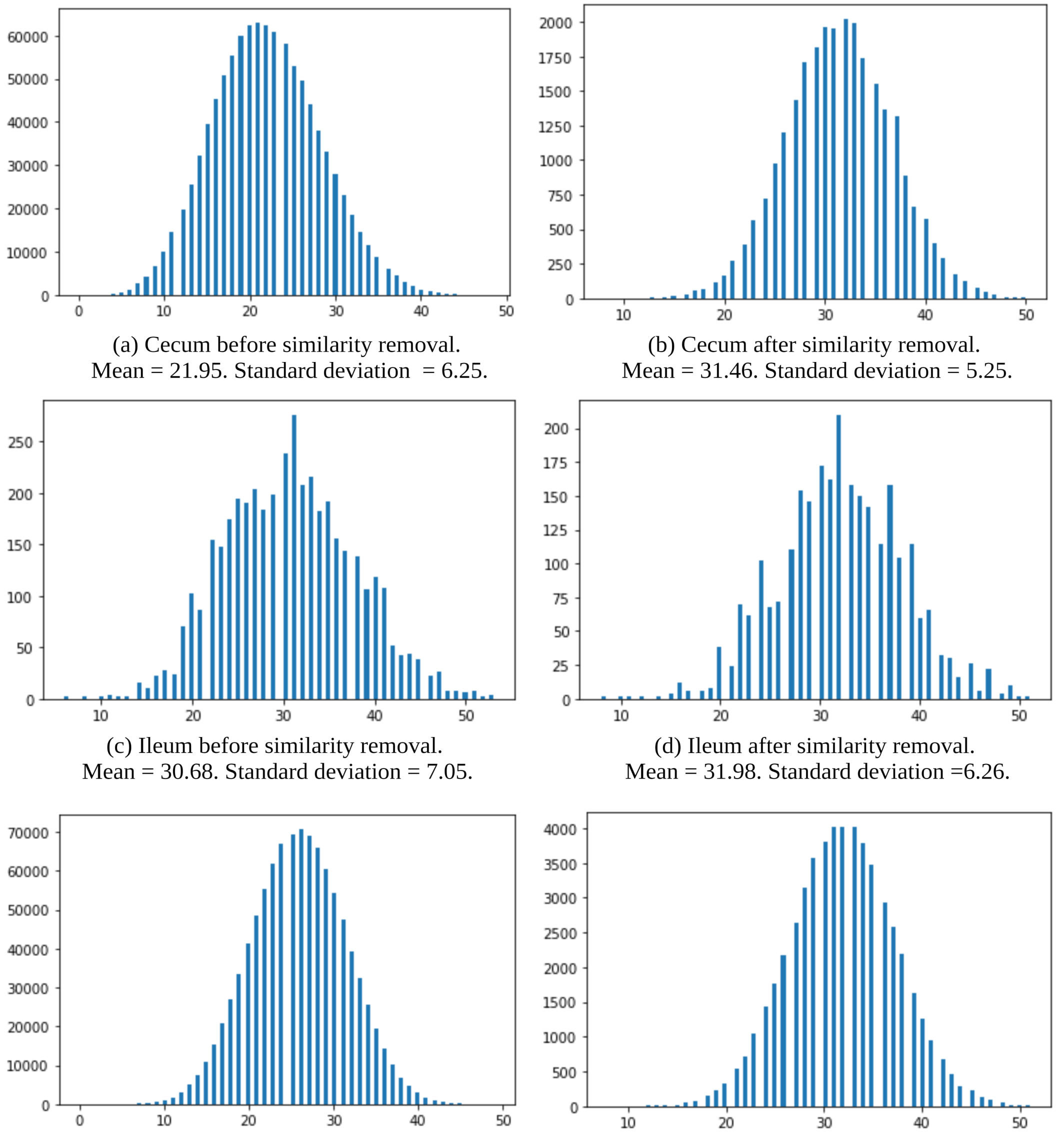


Figure 4.1: A few image samples before and after pre-processing steps. The top row illustrates images with some existing noises that may mislead the network's training process.

In addition, there was a high level of similarity between images within each category. As a result, some images could receive more weight in the training step, or the network might perform incorrectly when it comes to validation. In order to address this issue, image hashing was used to detect and remove similar image samples from the dataset. After removing similarities, we compared the original dataset with the resulting data samples. We have developed a differential gradient hash to detect similar images. Based on the distance (difference) of each hash with the rest of the image samples of the same category (location), an array has been created. The lower the value of the differences, the more similar the images are. As a result, a dataset with less similarity has a smaller distribution around the lower values. The distribution of this array for each location is demonstrated in Fig. 4.3 with the mean and standard deviation of the values. Once similar images are removed, the mean of the hash differences between image samples increases.



(a) Cecum before similarity removal.
Mean = 21.95. Standard deviation = 6.25.

(b) Cecum after similarity removal.
Mean = 31.46. Standard deviation = 5.25.

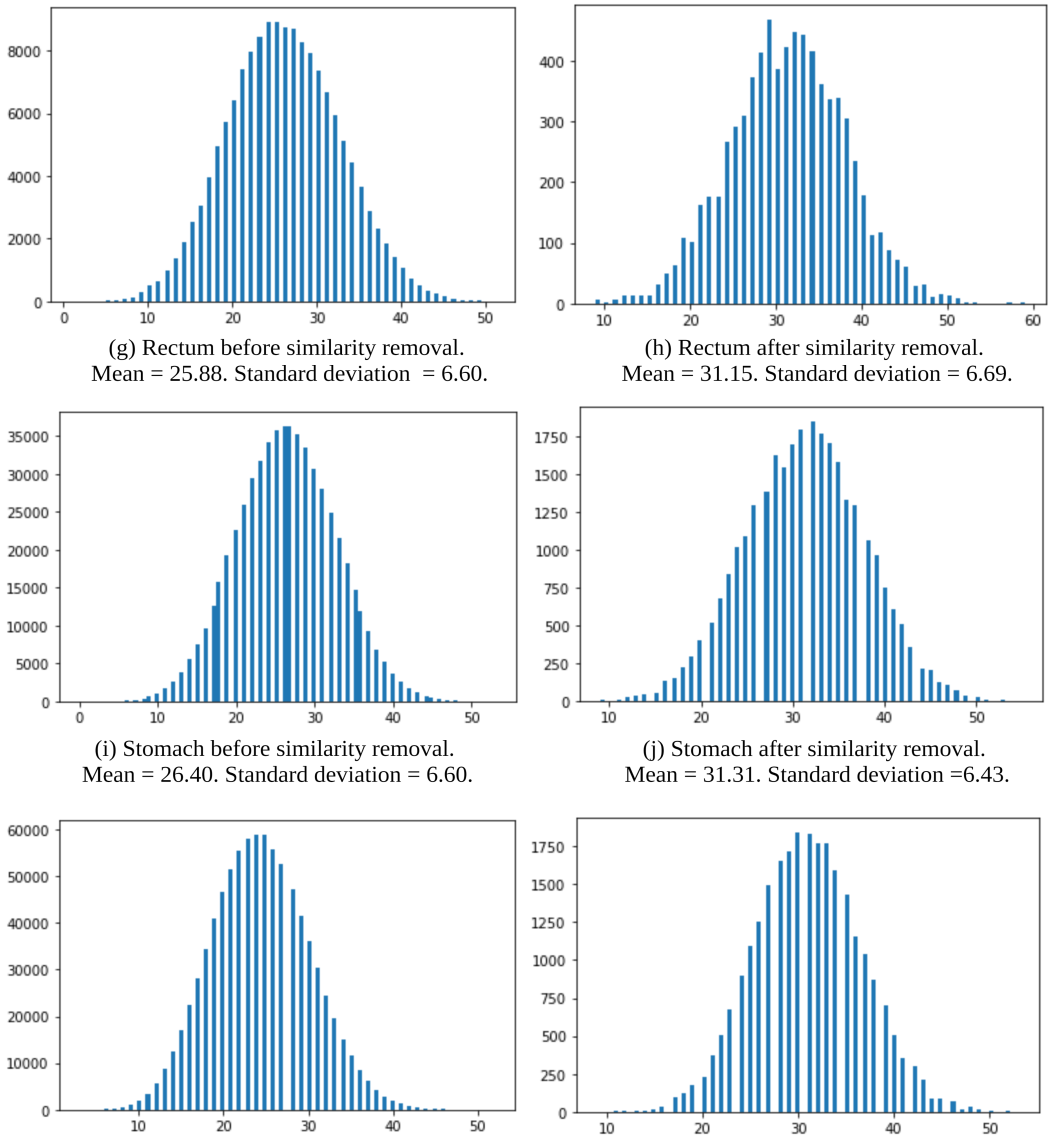
(c) Ileum before similarity removal.
Mean = 30.68. Standard deviation = 7.05.

(d) Ileum after similarity removal.
Mean = 31.98. Standard deviation =6.26.

(e) Pylorus before similarity remova.
Mean = 25.91. Standard deviation = 5.58.

(f) Pylorus after similarity removal.
Mean = 31.87. Standard deviation = 5.35

Figure 4.2: The distribution plot of the hash distances of each location image samples. From (a) to (f), 3 locations of the cecum, ileum, and pylorus, before and after similarity removal, have been illustrated. Their mean and standard deviation have also been reported.



(g) Rectum before similarity removal.
Mean = 25.88. Standard deviation = 6.60.

(h) Rectum after similarity removal.
Mean = 31.15. Standard deviation = 6.69.

(i) Stomach before similarity removal.
Mean = 26.40. Standard deviation = 6.60.

(j) Stomach after similarity removal.
Mean = 31.31. Standard deviation =6.43.

(k) Z line before similarity remova.
Mean = 24.39. Standard deviation = 6.60.

(l) Z line after similarity removal.
Mean = 30.99. Standard deviation = 5.58

Figure 4.3: The distribution plot of the hash distances of each location image samples. From (g) to (l), 3 locations of the rectum, stomach, and z-line before and after similarity removal has been illustrated. Their mean and standard deviation have also been reported.

4.1.2 The proposed network performance

In order to classify GI tract locations anatomically, an image dataset containing 871 images was used in which 164, 52, 234, 87, 172, and 162 samples are available for the cecum, ileum, pylorus, rectum, stomach, and Z-line locations, respectively. A total of 80% of each category is used for training, while 20% is excluded as a test set. In order to train CNN models, fixed-size images must always be used. Therefore, in order to demonstrate the performance of the model on variant input data, the GI tract images have been resized into 128 x 128 × 3. The learning rate is adjusted to 0.0001. After the network has been trained, its performance is evaluated and presented in Table. 4.1

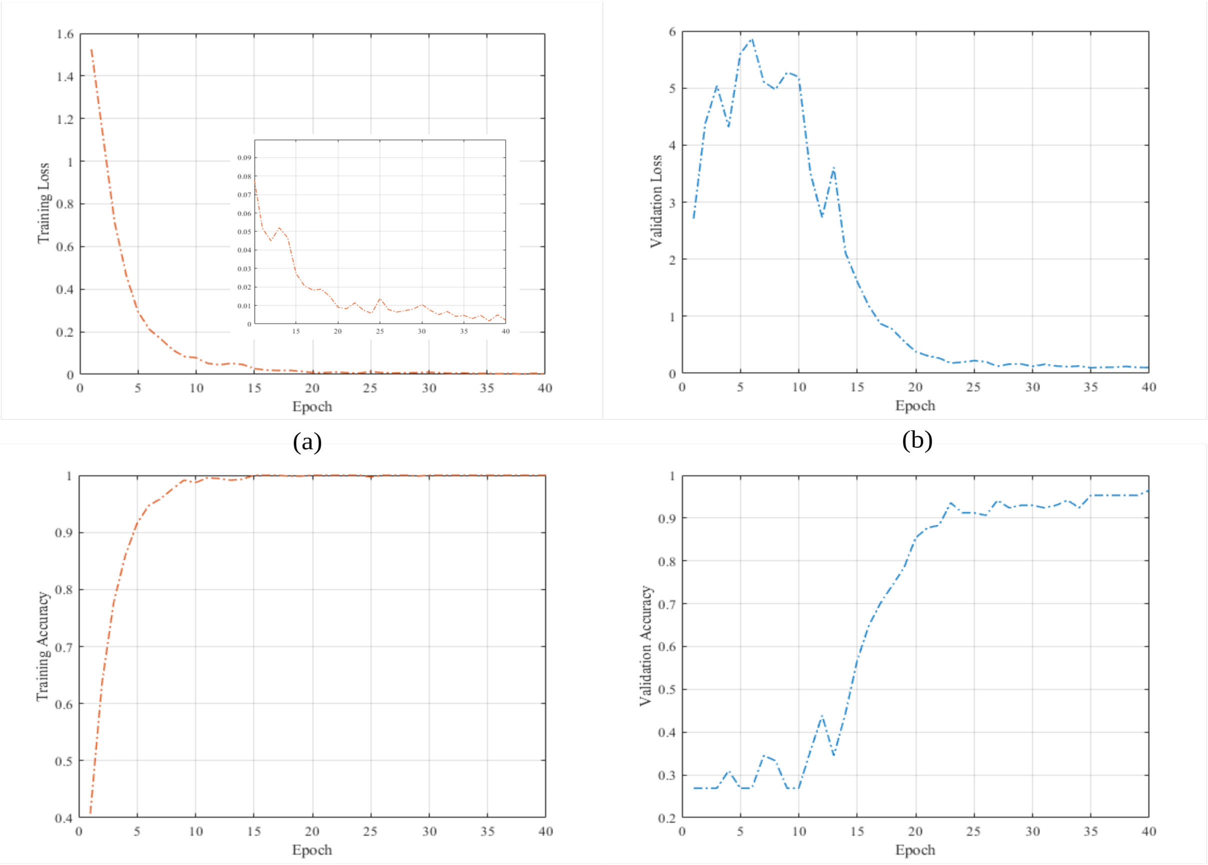
Table 4.1: Statistical performance of the proposed method

Category	Accuracy	Precision	Recall	F1-score
Cecum	0.97	0.91	0.97	0.94
Ileum	0.80	0.89	0.80	0.84
Pylorus	0.98	0.98	0.98	0.98
Rectum	0.94	1.0	0.94	0.97
Stomach	1.0	1.0	1.0	1.0
Z line	0.97	0.97	0.97	0.97
Total	0.97	0.97	0.97	0.97

An overall accuracy of 97% was achieved by the proposed method. A summary of each category's results is also provided.

Other metrics have also been calculated and recorded in Table 4.1 along with accuracy. Stomach has an accuracy, precision, recall, and F1-score of 1.0. Pylorus and Z line have the next highest performances with 0.98 and 0.97 for all of the measures. Accuracy, precision, recall, and F1-score for the cecum are 0.97, 0.91, 0.97, 0.94, respectively. The performance of the rectum and illeum is 0.94, 1.0, 0.94, 0.97, and 0.8, 0.89, 0.8, 0.84 for accuracy, precision, recall, and F1-score.

The learning and validation curves are illustrated in Fig. 4.4 which shows the network took 40 epochs to achieve 97% accuracy. To stop the learning process, early stopping has been used.



(a)

(b)

(c)

(d)

Figure 4.4: (a) and (b) show the loss curve, and (c) and (d) demonstrate the accuracy curves of the learning and validation process for the proposed method. To better illustrate the improvements in learning during the past epochs, a magnified view of the learning loss curve has also been presented.

For the purpose of checking or visualizing the performance of a multiclass classification problem, we use the AUC of the ROC, an effective metric for evaluating classification models' performance. The ROC - AUC measures the discriminative ability of classification models. ROC graphs were first developed for radar signal detection. In [65], it was first suggested that ROC analysis be applied to medical research.

AUC is a measure of how well the model distinguishes between categories, and the higher the AUC, the more accurate the model is at predicting classes. Considering that the goal of this study is to classify the anatomical locations of the GI tract, the higher the AUC, the better the model can distinguish between them.

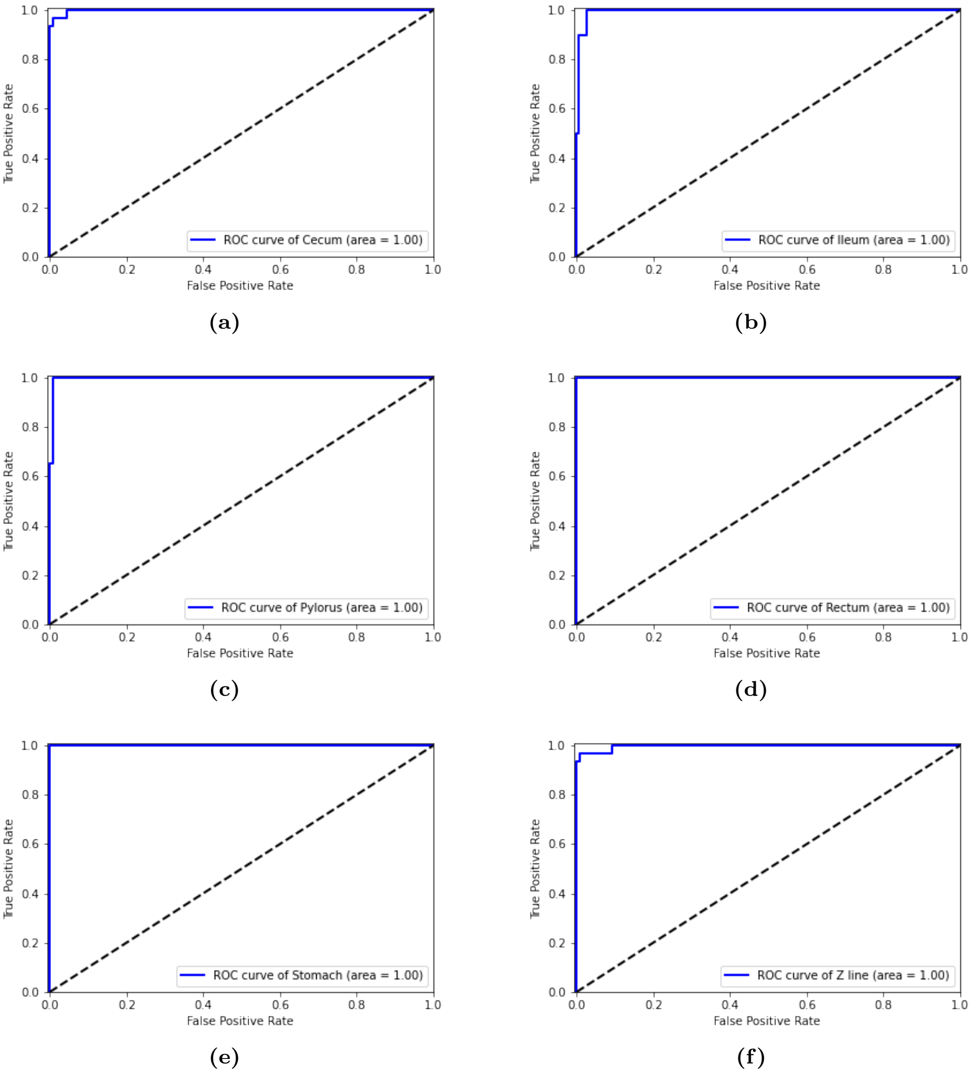
Plotting ROC curves for each category in a multiclass model can be accomplished using the One vs ALL methodology. As shown in Fig. 4.5, there are 6 ROC curves related to 6 different anatomical locations of the GI tract. An optimal discriminatory model would have a ROC curve that starts from point (0,0) and goes to point (0,1), and then from this point to point (1,1). From the figure, it is evident that the ROC curve of each location has the optimal plot curve with an AUC of 1, meaning it has the best measure of separability.

This network was also trained on the same dataset without using ensemble learning to evaluate the effectiveness of using the trained encoder on 99,417 unlabeled images as a feature extractor. The results in Table 4.2 demonstrate that using feature extraction enhances prediction by 5% when compared with no feature extraction.

Table 4.2: The effect of using trained encoder as the feature extraction

Method	Accuracy	Loss
Xception network without feature extraction	92%	0.24
Xception network with feature extraction using the trained encoder	97%	0.1

Confusion matrix is another performance measurement for classification problems. It is highly beneficial to create a confusion matrix in order to provide the needed data to calculate quantify metrics including Recall, Precision, Specificity, Accuracy, AUC-ROC curves, etc. Table 4.3 shows the confusion matrix of the proposed model and the network with no feature extraction. By comparing these two confusion matrixes we can estimate the effectiveness of using the encoder for feature extraction for each individual anatomical location. According to the table, we achieved 13, 7, 18, 3, 4 percent improvement in accuracy for the locations



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 4.5: The proposed network recognized the anatomical location of gastro-intestinal endoscopy images very accurately with area under the curve (AUC) values of 1.00 for all the locations (cecum, ileum, pylorus, rectum, stomach, z line).

of the ileum, pylorus, rectum, stomach, and Z line respectively. The highest improvements are 18 and 13 percent related to the locations of the rectum and ileum. These two locations contained fewer image samples than other locations. This indicates that, in classification problems with less labeled images, feature extraction and ensemble learning improve results efficiently.

Table 4.3: Comparison of the proposed method's confusion matrix with the network without feature extraction and ensemble learning. Top table is for the proposed method and the bottom table illustrates the result for the network without ensemble learning.

Predicted label

True label

Locations	Cecum	Ileum	Pylorus	Rectum	Stomach	Z line
Cecum	**0.97**	0.03	0	0	0	0
Ileum	0.2	**0.8**	0	0	0	0
Pylorus	0	0	**0.98**	0	0	0.02
Rectum	0.06	0	0	**0.94**	0	0
Stomach	0	0	0	0	**1**	0
Z line	0	0	0.03	0	0	**0.97**

Predicted label

True label

Locations	Cecum	Ileum	Pylorus	Rectum	Stomach	Z line
Cecum	**0.97**	0	0.03	0	0	0
Ileum	0.33	**0.67**	0	0	0	0
Pylorus	0.02	0	**0.91**	0	0.02	0.04
Rectum	0.12	0.06	0	**0.76**	0	0.59
Stomach	0.02	0	0	0	**0.97**	0
Z line	0	0	0.07	0	0	**0.93**

In deep learning, interpretability is an influential concept. For intelligent systems to be integrated meaningfully into our everyday lives, transparent models must be built that can explain their reasoning toward the prediction process they have. It is usually difficult to balance accuracy with interpretability. Deep models employ complicated non-interpretable modules that deliver superior performance. Interpreting these models is difficult due to their complexity. However, to be reliable, AI systems need to be understood and have clear justifications for their decision-making process. To have a further look at how the proposed model detects the anatomical locations of the GI tract, a heatmap map from the CNN model is drawn. A heatmap map from the CNN model is drawn to further investigate how the proposed model detects the anatomical locations of the GI tract. Heatmap is also known as class activation map, resulting from gradients of any particular concept flowing through the last convolutional layer. This is done to identify the most crucial regions in the image for the predicting procedure [66].

By understanding which components of an image belong to a particular class, we can more efficiently localize images. Fig. 4.6 shows the above-mentioned map.

As can be seen from the heatmap, the colours represent different confidence scores that GradCAM has calculated through the computation of gradients. The red colour signifies the regions having the highest confidence in classifying the related image into the correct location. In contrast, the blue region indicates the less effective regions for prediction. This technique can also indicate that the proposed model does not make predictions based on noise or external objects as they have no impact on the decision-making process.

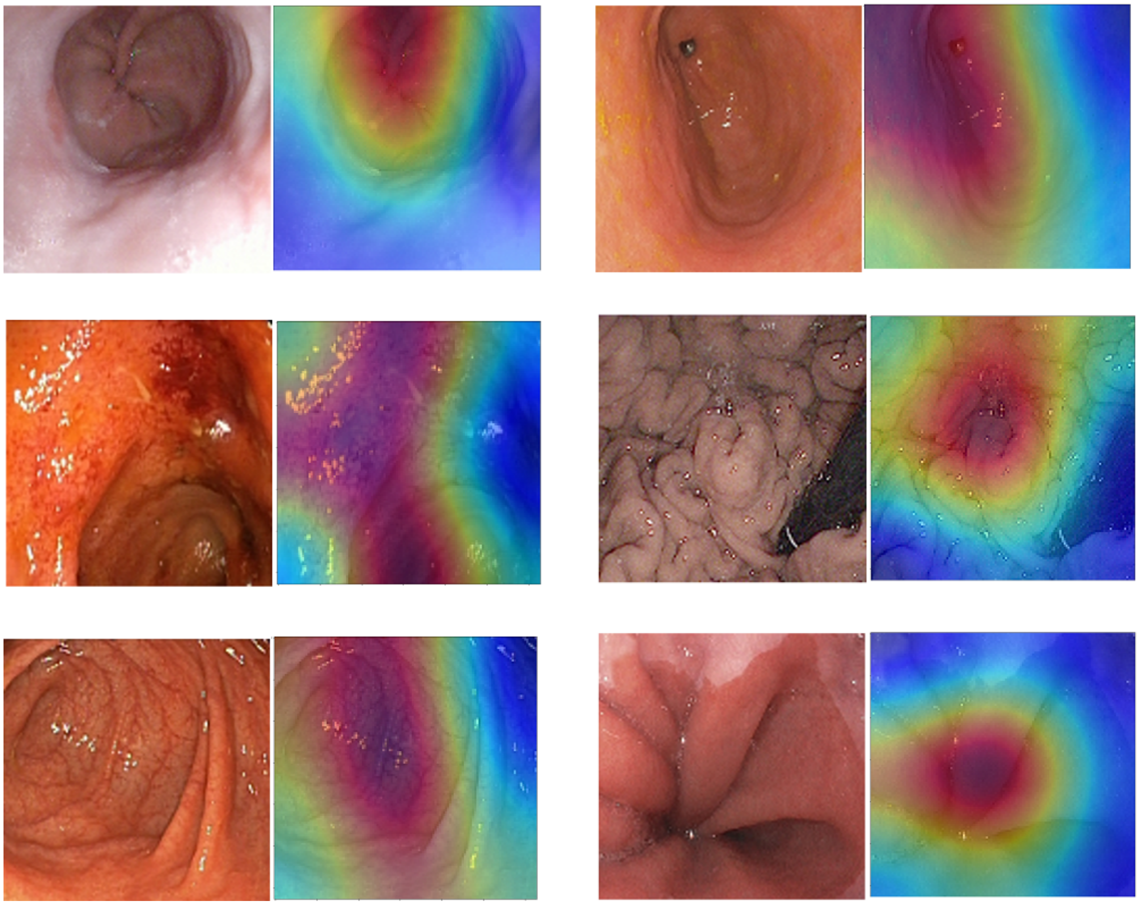


Figure 4.6: Visual explanation of decisions to highlight the important regions in the input images

In most medical-related problems, the main challenge is the lack of enough labelled images. This challenge can be overcome in two ways: collecting more samples for training the model or designing architectures that can learn from fewer samples and predict accurately based on them. It is not always feasible to collect more labelled data in this field; however, there can be a great deal of unlabelled data that the network can take advantage of. An encoder has been trained on unlabelled images and used as a feature extractor in the network, and a total number of 840 labelled images from 6 different locations, cecum, ileum, pylorus, rectum, stomach, and Z line have been used for the training process.

In Table 4.4, a comparison of the proposed method's results with those from existing studies is presented. These studies performed the same tasks as the proposed method, classification and labelling of GI tract anatomic locations.

In summary, Lee et al. [29] implemented a system for detecting 5 locations of the esophagus, stomach, duodenal, ileum and colon based on colour changes observed in consecutive video frames and achieved a 61% F1-score; no machine learning or deep learning approaches were utilized in their work. The proposed approach outperforms their result by 36% concerning F1-score.

Marques et al. [30] used colour features and SVM for the stomach, small intestine, and large intestine (3 locations) to classify WCE frames, achieving an overall accuracy of 85.2%. Shen et al. [31] used SIFT local feature extraction on WCE images and unsupervised learning based on clustering for localization of the stomach, small intestine, and large intestine (3 locations). They achieved an overall accuracy of 97.6%.

In their work, the accuracy is 0.6% higher than ours, but the number of locations classified is only three. The problem becomes more complex when there are more classes. In other fields, such as anomaly detection, expanding the number of classes is also investigated. A study conducted by Mohammed et al. [67] demonstrated that adding classes caused the problem to become more complex and that performance decreased. On the other hand, a method's efficiency increases as it predicts more locations, which makes it more precise when defining locations.

Saito et al. [13] achieved an overall accuracy of 66% when applying a CNN to standard colonoscopy images, including the terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, and the anus (6 locations). Takiyama et al. [32] were pioneers who used standard endoscopy images for training a CNN to classify input images as either the larynx, esophagus, stomach (upper, medium, and lower part) or duodenum (6 locations). They achieved 97% accuracy with an AUC more than 99%. While they have the same accuracy as our work, in our proposed method, the number of images used for training is significantly lower than that employed in theirs. They utilized 27,335 image samples for the training process while our network achieved an accuracy of 97% with 840 labelled image samples.

Another experiment has been conducted in order to compare several methods including the base methodology of some studies presented in Table 4.4 with the proposed method based on feature extraction and ensemble learning. According to this table, [32] and [13] employed CNN, and [31] employed unsupervised data clustering with SIFT. The same dataset has been applied to all of them and the results are illustrated in Fig. 4.7. In this experiment, the proposed method has been compared with the Xception network which is the main architecture used in this study, Resnet-50, VGG-15, simple CNN, and SVM with Scale Invariant Feature Transform (SIFT) features. By using the processed dataset, which includes 840 labelled images from HyperKvasir, the proposed method outperforms the other models in all categories.

While we were unable to make a direct comparison of our proposed method's complexity with the state-of-the-art due to the lack of their source codes and materials, we can report the effectiveness of our approach in terms of its complexity. Specifically, our method achieved a mean validation accuracy of 0.97 with a total of 40 epochs. It was trained on Google Colab Pro with its provided GPU in approximately 85 seconds. A model's training time is directly proportionate to the computational resources required to train it, thereby the proposed method is considered computationally efficient. In addition, our method offers a number of unique advantages, such as requiring fewer labeled data while categorizing images into six different locations.

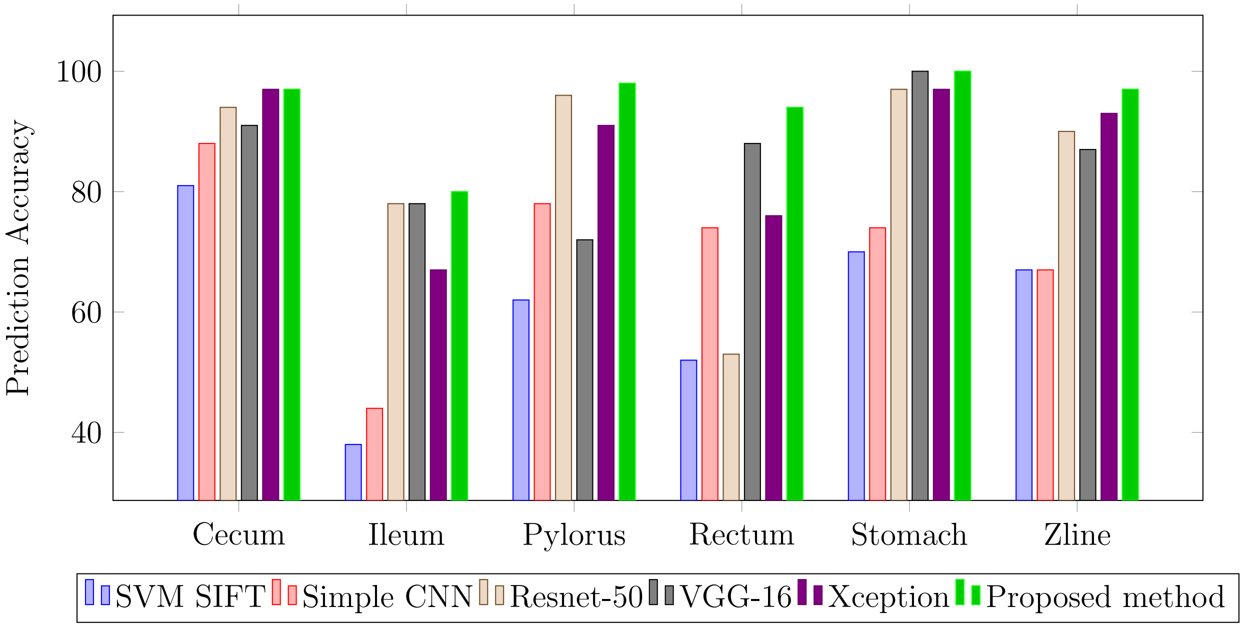


Figure 4.7: Different methods' performance on the same dataset

Table 4.4: Comparison between the proposed method and other research works

Study	Method	Number of locations	Locations	Training size	Test size	Result
[32]	Convolutional Neural Network	6	Larynx, Esophagus, Stomach (Upper, Medium, Lower), Duodenum	27335	13048	97%
[13]	Convolutional Neural Network	6	the terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, the anus	4100	1025	66%
[29]	Variation in HSV intensity in subsequent frames using event correlation	5	esophagus, stomach (entering stomach), small intestinal (entering duodenal and ileum), and colon	N/A	External Dataset (10 videos, number of frames is NA)	Recall: 76%; Precision: 51%; 61% F1-score
[31]	The probabilistic latent semantic analysis model for unsupervised data clustering with Scale Invariant Feature Transform (SIFT) features	3	stomach, small intestine, and large intestine	50000	10-fold cross validation	97.6%
[30]	SVM with color features	3	stomach, small intestine, and large intestine	26469	10-fold cross validation	85%
[49]	Attention-based SNN with Manifold mixup	10 for CE 7 for WCE	Esophagus, Cardia, Angularis, Pylorus, Duodenum, Jejunum, Ileum, Colon, Rectum, Anus	78 CE 27 WCE	External Dataset (2570 CE, 1825 WCE)	CE: 93% Accuracy WCE: 86% F1-score
Proposed method	Convolutional Neural Network	6	Cecum, Ileum, Rectum, Pylorus, Stomach, Z-line	700	171	97%

5 Conclusion and Future Works

5.1 Conclusion

A novel classification method was developed in this study to classify the anatomical locations of endoscopic images of the GI tract. In order to accurately follow up on an GI tract abnormality, treat it, or perform surgery, the accurate location of any abnormality must be precisely determined. Computer-aided intelligence systems can help better navigate the challenges associated with the anatomical classification of GI tract images. To address this challenge, having lots of well-labelled data is typically essential, but the quality and quantity of labelled medical images, especially endoscopic images, are not yet sufficient. A major objective of our method was to overcome this lack of properly labeled and available data. As a result, we employed two different machine learning methods to accomplish this goal. The first step in our method is to use transfer learning and take advantage of the vast amount of unlabeled data that is publicly available. Following that, two networks extract features with different characteristics to improve the learning process. With ensemble learning, both feature vectors are considered in the final decision.

In the first part of our solution an auto-encoder is trained on 99,417 GI tract unlabeled images from the Kvasir dataset to learn efficient embeddings of GI tract unlabeled data. The encoder section of this auto-encoder is then separated from the encoder-decoder architecture and used as a feature extractor in the main network to transfer the extracted knowledge. As all the information in the GI tract image should be preserved during encoding, this knowledge includes the most informative GI tract image features. By employing this technique, a more discriminating feature can then be obtained. For the second part of our solution, we integrate two machine learning models into a unified framework to improve the final predictions. This prediction is based on a concatenation of the features extracted from both models, the trained encoder and the Xception network. Preprocessed and filtered Kvasir labelled images from 6 different anatomical locations of the gastrointestinal tract are included in the data for this part's training and testing procedure.

The suggested technique yields an F1-score, AUC, and overall accuracy of 97%, 100%, and 97% respectively. To evaluate the effectiveness of these machine learning techniques, the network was also trained on the same dataset without transferring knowledge and ensemble learning. It is demonstrated that the use of feature extraction and ensemble learning improves prediction by 5% over the absence of these techniques. According to the analysis of each category separately, there is a significant improvement of 18 and 13 percent in the location of the rectum and ileum. Compared to other locations, these two had fewer image samples. The results indicate that feature extraction and ensemble learning are efficient approaches to improving results in classification problems with less labeled images.

Various studies have been carried out to demonstrate the significance of the proposed method. A detailed comparison shows our work's advantage over previous works. Research works can be analyzed through a variety of metrics including accuracy, number of locations, and number of labeled images. While having fewer than a thousand labeled images, our method achieved higher accuracy than other prior studies. A further improvement over most previous approaches is that our method is able to categorize more anatomical locations than previous studies. It is important to note that as more classes are added, the problem becomes more complex, making classification more difficult.

5.2 Future Works

The proposed method of predicting gastrointestinal tract locations using transfer learning and ensemble learning has shown promising results in overcoming challenges faced by previous methods. There are, however, some avenues for further research in this field. These future works could significantly contribute to the advancement of medical science by providing more accurate and efficient solutions.

There is a potential direction for future work in the field of gastrointestinal image classification that involves addressing the limitations of labeled data. The challenge arises from the fact that most datasets in this field are private and there are not enough appropriately labeled images to build effective machine learning models. The development of medical technology, such as capsule endoscopy, offers opportunities for the collection of large volumes of images that can be annotated by medical professionals. Therefore, we are able to not only propose novel architectures that are more flexible and effective for classifying images into multiple categories, but also demonstrate their generalizability on a variety of datasets. The future of research in this area could therefore focus on developing new techniques for efficiently collecting and annotating large amounts of image data generated by medical devices such as capsule endoscopy.

The incorporation of additional medical data, imaging modalities, and patient features into image classification is another promising area of future research. In order to guide the classification process, we could use patient history data, such as medical conditions, medications, and lifestyle factors. Furthermore, other medical imaging modalities, such as MRI or CT scans, may provide complementary information that can enhance our model's accuracy. Our model could also be improved by incorporating extracted patient data, such as age, gender, and genetic information. The development of an algorithm that receives all of these data sources and makes predictions based on them could be one avenue for future research in this area. Especially in complex cases, such as patients with multiple comorbidities, this approach could be useful for improving the accuracy and interpretability of our model. Rather than a model ensemble between different architectures, we could also use a model ensemble between different data sources.

Other Publications

1. F. S. Chafjiri, M. R. Mohebbian, K. Wahid and P. Babyn, ”Classification of Endoscopic Image and Video Frames using Distance Metric-Based Learning with Interpolated Latent Features,” *Multimedia Tools and Applications*, 2022. (Under review)

2. M. R. Mohebbian, F. S. Chafjiri, S. S. Vedaei and K. Wahid, ”Efficient Color Transformation for Bayer CFA Compression with FPGA implementation,” *Multimedia Tools and Applications*, 2022. (Under review)

3. K. M. M. Rahman, S. K. Mohammed, S. S. Vedaei, M. R. Mohebbian, F. S. Chafjiri and K.Wahid, “A low complexity lossless Bayer CFA image compression,” *Signal, Image and Video Processing*, vol. 15, pp. 1767-1775, 2021, doi: 10.1007/s11760-021-01921-6.

Bibliography

[1] F. Bianchi, A. Masaracchia, E. Shojaei Barjuei, A. Menciassi, A. Arezzo, A. Koulaouzidis, D. Stoyanov, P. Dario, and G. Ciuti, "Localization strategies for robotic endoscopic capsules: a review," *Expert Review of Medical Devices*, vol. 16, pp. 1–23, 05 2019.

[2] D. Schwartz, M. Wiersema, K. Dudiak, J. Fletcher, J. Clain, W. Tremaine, A. Zinsmeister, I. Norton, L. Boardman, R. Devine, B. Wolff, T. Young-Fadok, N. Diehl, J. Pemberton, and W. Sandborn, "A comparison of endoscopic ultrasound, magnetic resonance imaging, and exam under anesthesia for evaluation of crohn's perianal fistulas," *Gastroenterology*, vol. 121, pp. 1064–72, 12 2001.

[3] G. Ciuti, A. Menciassi, and P. Dario, "Capsule endoscopy: From current achievements to open challenges," *IEEE reviews in biomedical engineering*, vol. 4, pp. 59–72, 01 2011.

[4] Y. Zheng, L. Hawkins, J. Wolff, O. Goloubeva, and E. Goldberg, "Detection of lesions during capsule endoscopy: Physician performance is disappointing," *The American journal of gastroenterology*, vol. 107, pp. 554–60, 01 2012.

[5] M. Turkoz, S. Kim, Y. Son, M. Jeong, and E. Elsayed, "Generalized support vector data description for anomaly detection," *Pattern Recognition*, vol. 100, p. 107119, 11 2019.

[6] P. Pedersen, D. Bar-Shalom, S. Baldursdottir, P. Vilmann, and A. Müllertz, "Feasibility of capsule endoscopy for direct imaging of drug delivery systems in the fasted upper-gastrointestinal tract," *Pharmaceutical research*, vol. 31, 02 2014.

[7] N. Hosoe, Y. Hayashi, and H. Ogata, "Colon capsule endoscopy for inflammatory bowel disease," *Clinical Endoscopy*, vol. 53, 01 2020.

[8] F. Bianchi, G. Ciuti, A. Koulaouzidis, A. Arezzo, D. Stoyanov, S. Schostek, C. Oddo, A. Menciassi, and P. Dario, "An innovative robotic platform for magnetically-driven painless colonoscopy," *Annals of Translational Medicine*, vol. 5, pp. 421–421, 11 2017.

[9] M. J. A. Holzheimer, René G, *Surgical Treatment: Evidence-Based and Problem-Oriented*, 2001.

[10] A. H. Khan, M. H. A. Sohag, S. S. Vedaei, M. R. Mohebbian, and K. A. Wahid, "Automatic detection of intestinal bleeding using an optical sensor for wireless capsule endoscopy," pp. 4345–4348, 2020.

[11] Z. Liu, H. Li, K. Yu, S.-H. Xie, A. King, Q. Ai, W. Chen, X. Chen, Z. Lu, L. Tang, L. Wang, C. Xie, W. Ling, Y. Lu, Q. Huang, A. Coghill, C. Fakhry, R. Pfeiffer, Y. Zeng, and A. Hildesheim, "Comparison of new magnetic resonance imaging grading system with conventional endoscopy for the early detection of nasopharyngeal carcinoma," *Cancer*, vol. 127, 07 2021.

[12] S. S. Vedaei and K. A. Wahid, "Magnetofuse: A hybrid tracking algorithm for wireless capsule endoscopy within the gi track," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[13] H. Saito, T. Tanimoto, T. Ozawa, S. Ishihara, M. Fujishiro, S. Shichijo, D. Hirasawa, T. Matsuda, Y. Endo, and T. Tada, "Automatic anatomical classification of colonoscopic images using deep convolutional neural networks," *Gastroenterology Report*, vol. 9, 12 2020.

[14] S. Vedaei and K. Wahid, "A localization method for wireless capsule endoscopy using side wall cameras and imu sensor," *Scientific Reports*, vol. 11, 05 2021.

[15] A. el Hajjar and J.-F. Rey, "Artificial intelligence in gastrointestinal endoscopy: General overview," *Chinese Medical Journal*, vol. 133, p. 1, 01 2020.

[16] C. Berre, W. Sandborn, S. Aridhi, M.-D. Devignes, L. Fournier, M. Smail, S. Danese, and L. Peyrin-Biroulet, "Application of artificial intelligence to gastroenterology and hepatology," *Gastroenterology*, vol. 158, 10 2019.

[17] A. Ebigbo, C. Palm, A. Probst, R. Mendel, J. Manzeneder, F. Prinz, L. Souza Jr, J. Papa, P. Siersema, and H. Messmann, "A technical review of artificial intelligence as applied to gastrointestinal endoscopy: clarifying the terminology," *Endoscopy International Open*, vol. 07, pp. E1616–E1623, 12 2019.

[18] M. K. Sana, Z. Hussain, M. Maqsood, and P. Shah, "Artificial intelligence in celiac disease," *Computers in biology and medicine*, vol. 125, p. 103996, 09 2020.

[19] P. Visaggi, N. De Bortoli, B. Barberio, V. Savarino, R. Oleas, E. Rosi, S. Marchi, M. Ribolsi, and V. Savarino, "Artificial intelligence in the diagnosis of upper gastrointestinal diseases," *Journal of Clinical Gastroenterology*, vol. Publish Ahead of Print, 11 2021.

[20] A. Eshkevari and S. Sadough, "An improved method for localization of wireless capsule endoscope using direct position determination," *IEEE Access*, vol. PP, pp. 1–1, 11 2021.

[21] Y. Wang, S. Yoo, J.-M. Braun, and E. Nadimi, "A locally-processed light-weight deep neural network for detecting colorectal polyps in wireless capsule endoscopes," *Journal of Real-Time Image Processing*, vol. 18, 08 2021.

[22] N. Stap, F. Van der Heijden, and I. Broeders, "Towards automated visual flexible endoscope navigation," *Surgical endoscopy*, vol. 27, 05 2013.

[23] Y. Ye, P. Swar, K. Pahlavan, and K. Ghaboosi, "Accuracy of rss-based rf localization in multi-capsule endoscopy," *International Journal of Wireless Information Networks*, vol. 19, 09 2012.

[24] A. R. Nafchi, S. T. Goh, and S. A. R. Zekavat, "Circular arrays and inertial measurement unit for doa/toa/tdoa-based endoscopy capsule localization: Performance and complexity investigation," *IEEE Sensors Journal*, vol. 14, no. 11, pp. 3791–3799, 2014.

[25] G. Shao, Y. Tang, L. Tang, Q. Dai, and Y.-X. Guo, "A novel passive magnetic localization wearable system for wireless capsule endoscopy," *IEEE Sensors Journal*, vol. 19, no. 9, pp. 3462–3472, 2019.

[26] H. Vu, Y. Yagi, T. Echigo, M. Shiba, K. Higuchi, T. Arakawa, and K. Yagi, "Color analysis for segmenting digestive organs in vce," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2468–2471.

[27] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy color video segmentation," *Medical Imaging, IEEE Transactions on*, vol. 27(12), pp. 1769 – 1781, 01 2009.

[28] J. P. S. Cunha, M. Coimbra, P. Campos, and J. M. Soares, "Automated topographic segmentation and transit time estimation in endoscopic capsule exams," *IEEE Transactions on Medical Imaging*, vol. 27, no. 1, pp. 19–27, 2008.

[29] J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang, "Automatic classification of digestive organs in wireless capsule endoscopy videos," in *Proceedings of the 2007 ACM Symposium on Applied Computing*, ser. SAC '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 1041–1045. [Online]. Available: https://doi.org/10.1145/1244002.1244230

[30] N. Marques, E. Dias, J. P. S. Cunha, and M. Coimbra, "Compressed domain topographic classification for capsule endoscopy," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 6631–6634.

[31] Y. Shen, P. Guturu, and B. P. Buckles, "Wireless capsule endoscopy video segmentation using an unsupervised learning approach based on probabilistic latent semantic analysis with scale invariant features," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 1, pp. 98–105, 2012.

[32] T. Hirotoshi, O. Tsuyoshi, I. Soichiro, F. Mitsuhiro, S. Satoki, N. Shuhei, M. Motoi, and T. Tomohiro, "Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks," *Scientific Reports*, vol. 8, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-25842-6

[33] G. Bao and K. Pahlavai, "Motion estimation of the endoscopy capsule using region-based kernel svm classifier," in *IEEE International Conference on Electro-Information Technology , EIT 2013*, 2013, pp. 1–5.

[34] Z. Zheng, X. He, and C. Hu, "Magnetic localization and orientation of the capsule endoscope based on a random complex algorithm," *Medical Devices: Evidence and Research*, vol. 8, p. 175, 04 2015.

[35] S. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, pp. 1345 – 1359, 11 2010.

[36] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, 05 2016.

[37] W. Yan, *Computational Methods for Deep Learning: Theoretic, Practice and Applications*, 01 2021.

[38] Y. Oukdach, Z. Kerkaou, M. El Ansari, L. Koutti, and A. El Ouafdi, “Gastrointestinal diseases classification based on deep learning and transfer learning mechanism,” pp. 1–6, 10 2022.

[39] J. Escobar, N. Gomez, K. Sanchez, and H. Arguello, “Transfer learning with convolutional neural network for gastrointestinal diseases detection using endoscopic images,” pp. 1–6, 08 2020.

[40] Q. Su, F. Wang, D. Chen, G. Chen, C. Li, and L. Wei, “Deep convolutional neural networks with ensemble learning and transfer learning for automated detection of gastrointestinal diseases,” *Computers in Biology and Medicine*, vol. 150, p. 106054, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522007673

[41] T. Ghosh and J. Chakareski, “Deep transfer learning for automated intestinal bleeding detection in capsule endoscopy imaging,” *Journal of Digital Imaging*, vol. 34, 03 2021.

[42] O. R. A. Almanifi, M. A. M. Razman, I. M. Khairuddin, M. A. Abdullah, and A. P. A. Majeed, “Automated gastrointestinal tract classification via deep learning and the ensemble method,” in *2021 21st International Conference on Control, Automation and Systems (ICCAS)*, 2021, pp. 602–606.

[43] S. Nadeem, M. A. Tahir, S. S. A. Naqvi, and M. Zaid, “Ensemble of texture and deep learning features for finding abnormalities in the gastro-intestinal tract,” in *Computational Collective Intelligence*, N. T. Nguyen, E. Pimenidis, Z. Khan, and B. Trawiński, Eds. Cham: Springer International Publishing, 2018, pp. 469–478.

[44] P. Vieira, N. Freitas, J. Valente, I. Vaz, C. Rolanda, and C. Lima, “Automatic detection of small bowel tumors in wireless capsule endoscopy images using ensemble learning,” *Medical Physics*, vol. 47, 07 2019.

[45] Z. Khan, “Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract,” 10 2018.

[46] H. Rezaei, A. Amjadian, M. Sebt, R. Askari, and A. Gharaei, “An ensemble method of the machine learning to prognosticate the gastric cancer,” *Annals of Operations Research*, 09 2022.

[47] H. Borgli, V. Thambawita, P. Smedsrud, S. Hicks, D. Jha, S. Eskeland, K. Randel, K. Pogorelov, M. Lux, D. T. Dang Nguyen, D. Johansen, C. Griwodz, H. Stensland, E. Garcia Ceja, P. Schmidt, H. Hammer, M. Riegler, P. Halvorsen, and T. de Lange, “Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific Data*, vol. 7, 08 2020.

[48] “The gastrointestinal image site, gastrolab,” http://www.gastrolab.net/, accessed November 17, 2020.

[49] F. Sedighipour Chafjiri, M. R. Mohebbian, K. A. Wahid, and P. Babyn, “Classification of endoscopic image and video frames using distance metric-based learning with interpolated latent features,” 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-14982-1

[50] A. Sultana, I. Dumitrache, M. Vocurek, and M. Ciuc, “Removal of artifacts from dermatoscopic images,” *IEEE International Conference on Communications*, pp. 1–4, 05 2014.

[51] Y. Malkov and D. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 03 2016.

[52] E. Bjerrum, “Smiles enumeration as data augmentation for neural network modeling of molecules,” 03 2017.

[53] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 1798–1828, 08 2013.

[54] M. Tschannen, O. Bachem, and M. Lucic, “Recent advances in autoencoder-based representation learning,” in *Proc. Workshop Bayesian Deep Learn.* NeurIPS, 2018.

[55] R. Yao, C. Liu, L. Zhang, and P. Peng, "Unsupervised anomaly detection using variational auto-encoder based feature extraction," in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2019, pp. 1–7.

[56] C. Dong, T. Xue, and C. Wang, "The feature representation ability of variational autoencoder," 06 2018, pp. 680–684.

[57] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," pp. 1800–1807, 07 2017.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," vol. 7, 12 2015.

[59] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, 08 2019.

[60] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 06 2012, vol. 14.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: a large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 06 2009.

[62] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, 01 2012.

[64] D. M. W. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.16061

[65] P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-dependent roc curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000. [Online]. Available: http://www.jstor.org/stable/2676971

[66] R. Rs, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” 10 2017, pp. 618–626.

[67] A. Mohammed, I. Farup, M. Pedersen, S. Yildirim Yayilgan, and Hovde, “Ps-devcem: Pathology-sensitive deep learning model for video capsule endoscopy based on weakly labeled data,” *Computer Vision and Image Understanding*, vol. 201, p. 103062, 08 2020.

Appendix A

Wireless Capsule Endoscopy

Both traditional endoscopy and wireless capsule endoscopy may be enhanced with the proposed image classifier. The interpretation of images captured by wireless capsule endoscopy can, however, be challenging due to noise and artifacts, which make it difficult to accurately categorize images into different anatomical locations or identify specific features. As it allows for non-invasive visualization of the entire digestive tract, wireless capsule endoscopy is a promising technology for the diagnosis and monitoring of various gastrointestinal disorders. Given the importance of capsule endoscopy, using AI solutions to enhance its capabilities is one of the key components of this study. Fig. A.1 shows a few images taken by capsule endoscopy.

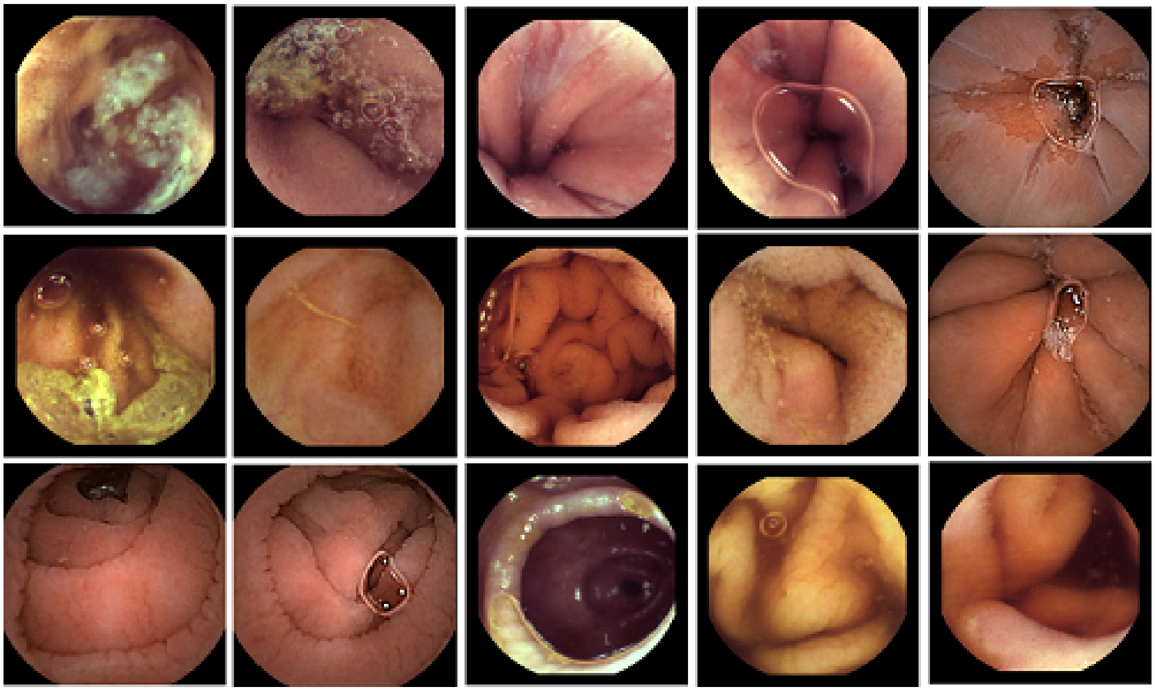
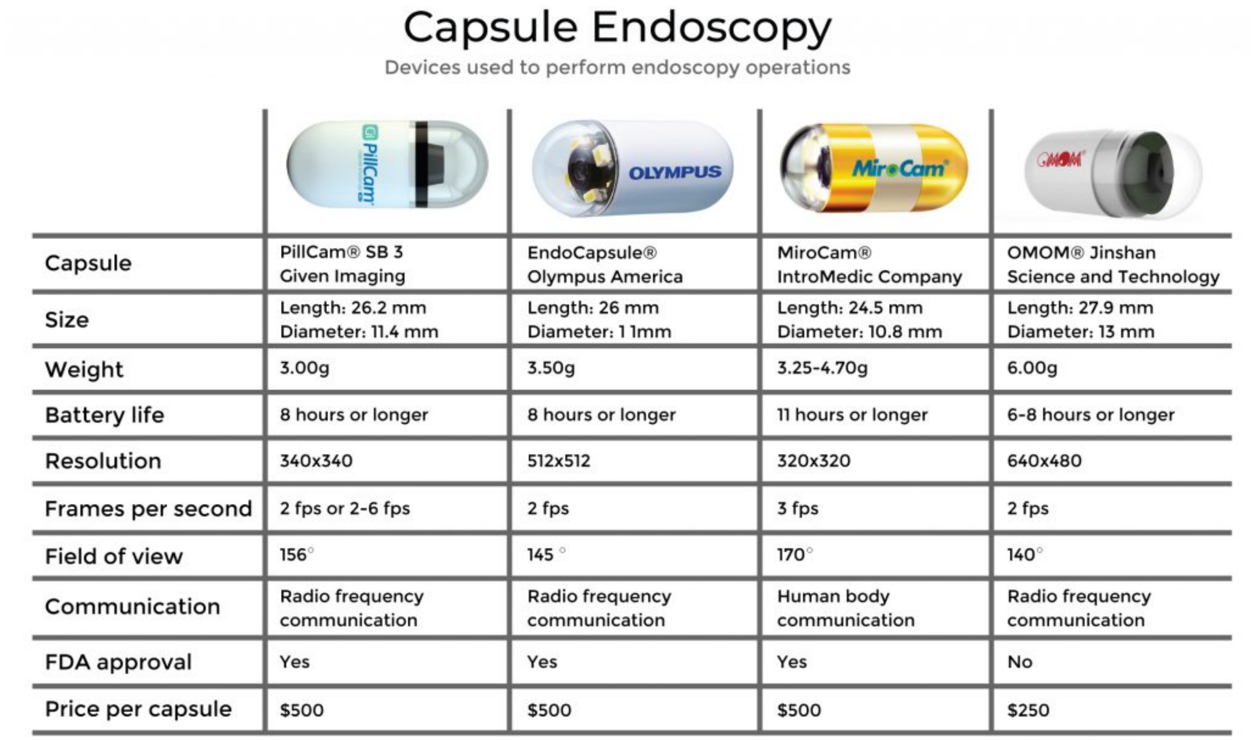


Figure A.1: A few samples captured by capsule endoscopy

The use of an image classifier in wireless capsule endoscopy can assist in interpreting videos and images captured by the capsule. Typically, capsules are equipped with cameras that capture images and videos as they travel through the digestive system. These images and videos can then be used to detect and analyze various gastrointestinal conditions. There are a variety of types of endoscopic capsules available on the market, as shown in Fig. A.2.



Capsule Endoscopy

Devices used to perform endoscopy operations

Capsule	PillCam® SB 3 Given Imaging	EndoCapsule® Olympus America	MiroCam® IntroMedic Company	OMOM® Jinshan Science and Technology
Size	Length: 26.2 mm Diameter: 11.4 mm	Length: 26 mm Diameter: 1 1mm	Length: 24.5 mm Diameter: 10.8 mm	Length: 27.9 mm Diameter: 13 mm
Weight	3.00g	3.50g	3.25-4.70g	6.00g
Battery life	8 hours or longer	8 hours or longer	11 hours or longer	6-8 hours or longer
Resolution	340x340	512x512	320x320	640x480
Frames per second	2 fps or 2-6 fps	2 fps	3 fps	2 fps
Field of view	156°	145 °	170°	140°
Communication	Radio frequency communication	Radio frequency communication	Human body communication	Radio frequency communication
FDA approval	Yes	Yes	Yes	No
Price per capsule	$500	$500	$500	$250

Figure A.2: Different types of endoscopic capsules on the market and their information [1]

Real-time analysis of wireless capsule endoscope images using image classifiers involves the application of machine learning models to the real-time images and videos captured by the capsule endoscope while they are being transmitted to a device. Having immediate feedback during a procedure can allow the clinician to adjust the intervention in real-time based on the analysis of the images. Detecting bleeding or other urgent abnormalities through real-time image analysis can be especially useful in situations where immediate action is needed.

[1] https://igniteoutsourcing.com/healthcare/capsule-endoscopy-technology/