**Introduction:**

Ulcerative Colitis (UC) is an autoimmune-related chronic inflammatory bowel disease (IBD) that affects the lining of the lower gastrointestinal (GI) tract. It is characterized by inflammation on the superficial layer of the GI tract that leads to erosions, ulcers, bleeding, and abnormal vascular patterns.

It is estimated that 3.1 million people have been diagnosed with IBD,[1] and the United States UC patient population is expected to incur lifetime total costs of $377 billion USD.[2] The key drivers of cost for IBD patients are treatment with drugs, emergency department use, and health care services.[3] UC incurs a large financial burden on patients and the US healthcare system.

UC has no cure, and treatment involves the lifelong use of anti-inflammatories, immunosuppressants, biologics, and/or surgery to remove the colon and rectum. The type of treatment will depend on the severity of the disease. The cost of treatment increases as severity increases.

Currently, the gold standard in diagnosing UC is through endoscopic imaging of the lower GI tract. Physicians conduct colonoscopies to collect images and biopsy tissue. Using their judgment, physicians will attempt to classify the presence of UC through the evaluation of these images.

Due to the differences in perceptions between physicians, it is not uncommon for IBD to be misdiagnosed. In a study conducted at the Asan Medical Center in Korea, 13.8% of the 2896 patients originally diagnosed with IBD had their diagnosis change from UC to CD, CD to UC, or IBD to non-IBD[4]

The correct classification of UC is key in the treatment of patients and proper healthcare resource utilization. With the steady decrease in gastroenterologists and differences in perception amongst gastroenterologists, there is a great need for a classifier for UC.

For this project, I used a pre-trained convolutional neural network (CNN) to perform a supervised machine learning task. The pre-trained model was trained on new data for the binary classification of UC and no-UC. The data was split into an 80%/10%/10% training-validation-test split. The model was used to predict the presence of UC in frames taken from endoscopes during colonoscopies. Evaluation of the model was based on accuracy and hyperparameters were tuned based on our objective function.

The specific CNN model that was used was ResNet50. This choice is based on the high accuracy seen in peer-reviewed papers on endoscopy classification tasks and its availability on PyTorch.

The goal of this project was to create a deep learning model that can aid gastroenterologists so they can make more accurate diagnoses in clinical practice. This will decrease the burden on healthcare spending and help healthcare systems achieve better patient outcomes.

**Problem Framing:** This was a binary classification task to detect the presence of UC in patients by looking at endoscopic images taken during colonoscopies.

**Data Framing:** The information available were images of visually normal and abnormal surfaces of the GI tract that are represented as pixel value intensities from a range of 0 to 255 in 3 color channels taken as frames from videos. The images were in a variety of different resolutions and the number of features was represented as the product of the height, width, and channel. Each image was paired with labels that correspond to pathology or normal GI. I used only images of the lower GI and their labels for my project.

**Objective Framing:** I used accuracy, precision, recall, and F1-score as my evaluation metrics in comparing CNN models based on their ability to classify endoscopic images of the low GI tract. Accuracy was used because it allowed me to compare the performance of our model to a baseline and their predictive ability for classifying UC in a metric that is easy to understand. Precision allowed for the assessment of how many positive predictions of UC were true positives. Recall gave us a picture of how many UC-positive images were correctly predicted. F1-scores combines both recall and precision into 1 metric. Hyperparameters, such as learning rate, were adjusted to maximize the accuracy of the models based on binary cross-entropy loss function.

**Data, Data Pipeline Plan, and Data Summary:**

The data was obtained from a private dataset created by the Computer Vision & Artificial Laboratory at Gdańsk University of Technology in Poland. The *ERS* dataset is available for research purposes and was attainable by contacting the team that created the dataset.

*ERS* contains around 6,000 and 115,000 precisely and imprecisely labeled frames from endoscopy videos, 3,600 and 22,600 precisely and imprecisely segmentation masks, and 1.23 million unlabeled frames from endoscopy videos. This data comes from 1520 videos taken from 1135 patients at the University of Gdańsk’s Clinic of Gastroenterology and Hepatology. The dataset was around 1~ terabyte in size after unzipping the files.

The dataset contained frames taken from videos of upper endoscopy and colonoscopy procedures in PNG format of various resolutions in RGB. Information about each frame is contained in a CSV file where they are observations given in rows. The columns contain information on *file, examination, sequence, frame, label,* and *mask*.

The dataset was separated into 5 categories: *gastro, colono, healthy, blood,* and *stool.* Each category has labels within them. *Colono* has 34 labels denoted by *c01 to c34* that represent different pathologies. There are 39590 labeled frames taken from 482 videos of patients. UC is represented in 3 labels: *c32, c33,* and *c34*. They compromise of 2765 labeled frames taken from videos of 68 patients. *Healthy* contained 20483 labeled images from 67 videos of patients. Under *healthy* was the feature label *h07* for colon, which contained 12031 labeled frames from 49 videos of patients.

I used only data from the *colono* category and *h07* for my project. Feature labels *c32, c33, and c34* were consolidated into a single label for *uc* with the value of 1. Feature Labels *c01 to c31 and h07* were combined and added into *no-uc* with the value of 0. *uc* was referred to as malignant and *no-uc* was referred to as benign, even though the terminology is not technically correct. This kept with the standard convention seen in binary classification tasks of medical images and increases clarity when working with the data.

Upon examining the first and last frames of randomly chosen videos, many of the frames contained no meaningful information. This is because the frames came from videos that include the insertion of the endoscope, blurry transitions to different parts of the GI tract, and removal of the scope. This information was not useful information for our model to learn.

To filter for frames that had meaningful information, I used segmentation masks as guidance for which images to include in my image dataset. Segmentation masks were annotations that partition the frames into different regions. This dataset gave us regions that would indicate a pathology. These segmentation masks were only in frames that were clear, focused, and where endoscopes reached the part of the colon gastroenterologists want to examine. I filtered for only these frames using the column *mask* from labels.csv and took the current frame and the two preceding and two proceeding frames and deleted any.

Data leakage was a potential problem because the frames came from videos of patients and could not be randomized solely on frames. To address this problem, the data was split by patients, which was indicated by the first four digits of each row in the column *file.* Two dictionaries were created for benign and malignant patients using the four digits as keys and the file paths as values.

The keys from the dictionaries were used to create a list called ‘patients’ that split the data into 80% train, 10% validation, and 10%, test sets called ‘X\_train’, ‘X\_val’, and ‘X\_test’, respectively. A list called ‘y’ was also made that contains 0’s and 1’s to denote benign or malignant and was split along with patients into ‘y\_train’, ‘y\_val’, and ‘y\_test’.

Using save\_split(), the images were resized, center cropped, converted to tensor, and normalized to fit the input requirements of ResNet. ResNet50 on PyTorch expected tensors to be a height and width of at least 224 and for the images to be loaded in a range from [0,1] and normalized using mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225].

save\_split() also split the frames into folders named X\_train, X\_val, and X\_test. The X\_train, X\_val, and X\_test contained 13739, 1123, and 2149 frames respectively.

X\_train, X\_val, and X\_test were loaded into DataLoader() to manage our data and avoid parallelization problems

In summary, I worked with 17011 labeled frames from 334 videos of patients. 3158 of the frames from 50 patients were labeled as the feature label *uc* as ‘1’ while the rest were

labeled as ‘0.’ The frames were examined and filtered to include only meaningful frames in my dataset. Training, validation, and test splits were performed at the patient level for the tabular data and then at the frame level for the frames. The images were then transformed to meet the ResNet50s input requirements and loaded into a dataloader.

A picture containing text, invertebrate, mollusk

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**Fig 1**: *0099/sample/frames/0001.png* from the *ERS* dataset and corresponding row from *image-labels.csv* visualized using Jupyter Notebook.

A picture containing diagram

Description automatically generated

Graphical user interface, application

Description automatically generated

**Fig 2**: *0006/seq\_01/frames/000180.png* from the *ERS* dataset and observations from *image-labels.csv filtered for label = ‘h07’* visualized using Jupyter Notebook.

**Research Papers:**

1. Transfer learning for medical image classification: a literature review: <https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-022-00793-7>
2. Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6537821/>
3. Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images:

<https://www.nature.com/articles/s41598-022-06726-2>

1. Ulcerative Colitis:

<https://www.nature.com/articles/s41572-020-0205-x>

1. Colonoscopic evaluation in ulcerative colitis: <https://academic.oup.com/gastro/article/2/3/161/2910194>
2. Dermatologist-level classification of skin cancer with deep neural networks: <https://www.nature.com/articles/nature21056>

***Transfer learning for medical image classification: a literature review***

This paper offered clarity on different models and transfer-learning (TL) types used in peer-reviewed papers that utilize TL for medical image classification tasks. 121 papers from 2016-2020 from PubMed and Web of Science were identified that met the authors’ inclusion criteria. The papers were mainly assessed for the type of models, model performance, and TL type. The most popular models and transfer-learning types were counted. It was found that Inception, AlexNet, VGG, and ResNet were some of the most popular models in their sample of papers. When it came to TL approaches, feature extractor was the most widely used amongst the other forms (fine-tuning from scratch, feature extractor hybrid, and fine-tuning.

The paper did a good job of providing information that can be used as guidance for projects utilizing deep learning in medical image classification. The authors present data on the accuracy of popular pretrained models and TL type based on the types of images, which can be used in model selection and training style for a specific case. The authors were very selective in their choices of paper, though I believe they would have been better served analyzing more than 2 web databases in their selection process. Despite this, the paper captures the landscape of TL in medical image classification well.

***Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images***

This paper investigated deep learning models in the evaluation of endoscopy images in the classification and grading of UC. For the data, the authors utilized a publicly available multi-class dataset called *HypeKvasir* that contains thousands of labeled endoscopic images with a thousand labels falling under UC. The authors had two training objectives: diagnosis of disease and grading of UC severity. Both problems were framed as binary classification tasks. The authors then chose four models based on the most popular models used in image classification. The authors chose to retrain all the layers in each model. The hyperparameters were tuned and the models were evaluated based on their accuracy, recall, precision, and F1-scores. All models were compared to each other and a baseline model with no skill. In this case, their skillless model was a model that predicted based on ‘1’ majority class and ‘0’ otherwise. The authors showed that all models were more accurate in binary classification tasks than the skillless model, with ResNet121 achieving the highest accuracy.

This paper shows the viability of using deep learning in the context of classification tasks for UC. The field is very new, with the earliest papers dating back to 2019 when conducting a database search on PubMed and Web of Science. In looking at the data used, it was very limited in scope, with only a couple of labels for lower GI pathologies. They also included data from upper GI, which has its own distinct set of specific diseases. The material and methods used by the author are the most common and widely used methods in the field of medical image classification. It is promising that the authors were able to achieve an accuracy of 87.50% in their best model, though much better accuracy needs to be achieved since stakes are higher in the medical field. Nonetheless, this paper establishes a strong foothold for the use of CNN models in this domain.

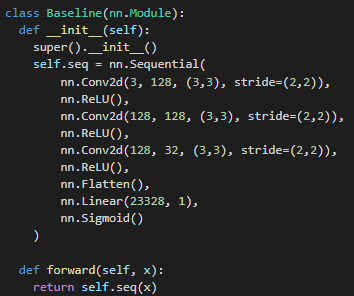
***Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis***

In this study, the authors aimed to assess the viability of deep learning models in grading the endoscopic severity of UC compared to an experienced human. The authors collected electronic health records from a healthcare system connected to the University of Michigan and filtered for UC patients that met their specific inclusion and exclusion criteria. Images from included patients were retrieved and a dataset was created. The study used two blinded independent experienced physician reviewers that would grade images in a binary classification task of remission versus moderate-to-severe UC. A third reviewer will then act as a judge of disagreement in classification between the two reviewers with the consensus acting as a ground truth for the dataset. A CNN model using pretrained Inception V3 was then created and retrained using this dataset. The model and the two human reviewers’ predictions were then compared. Using the Cohen kappa coefficient, it was found that the model had similar performance to the two individual reviewers. The same similarity was also seen in the accuracy of disease severity scoring.

This study shows that deep learning models can make perform on par with domain experts in a specific binary classification task in this field. This is powerful knowledge as future studies using CNN models can use this knowledge as a baseline and iterate. A concern I have is that the authors did not explain the rationale for model selection, which could be explored to obtain better results for deep learning. The experiment design also would have been better with more than two reviewers and a group of judges. Nonetheless, this paper was a good first step.

**Baseline:**

The baseline model that was used was an untrained CNN. This model was created as a point of comparison for a novel dataset. The model contained a section for feature extraction and classification. Feature extraction was performed using 3 convolution layers, each followed by an activation function called Rectified Linear Unit (ReLU) that is used for its efficiency and reduction in likelihood of vanishing gradients in CNN models. After feature extraction, a fully connected layer performed the classification task using a linear layer and a sigmoid activation function to make the output binary.



**Fig 3.** *Baseline CNN*

**Model, Results, and Discussion:**

ResNet50 pretrained on ImageNet with a fully connected layer was used as our model of interest. The feature extraction section is separated into 5 stages. Each stage contains convolution layers and identity blocks. The fully connected layer performs classification with a sigmoid activation function.

The strength of ResNet is the utilization of skip connections that allow for the “jumping” over of layers. This solves the vanishing gradient problem that occurs with neural networks with many layers where gradients decrease to zero.

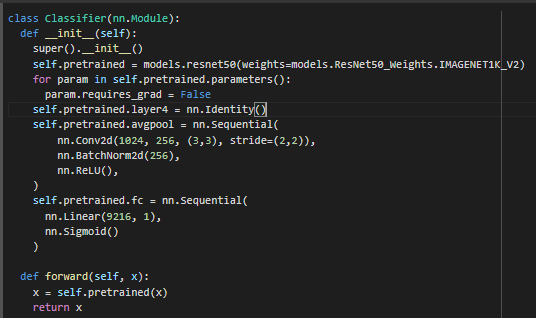
Both a baseline and ResNet were trained using batch sizes of 32, 256, and 256 for train, validation, and test sets. Larger batch sizes were used for validation and testing since they did not affect how well training converges. Both models were run for 100 epochs, using Adam optimizer to optimize our binary cross entropy loss. The parameter weight\_decay was set to 0.1 for our baseline model and 0.2 for ResNet. This was done as both models were experiencing issues with overfitting and 0.2 closed the gap in train and validation error for ResNet. The best model for baseline and ResNet was selected based on the model with the best validation loss. The best baseline and best ResNet were then tested on the test set and evaluation metrics were calculated.

The results were as follows:

|  | ***Baseline*** | ***ResNet50*** |
| --- | --- | --- |
| Accuracy | 0.52 | 0.72 |
| AUROC | 0.63 | 0.93 |
| F1 | 0.27 | 0.62 |
| Precision | 0.29 | 0.98 |
| Recall | 0.25 | 0.45 |

The results show that ResNet50 performed better in all evaluation metrics than the baseline model. This is to be expected since a pretrained ResNet transfer knowledge gained from looking at images on ImageNet (source task) was applied to the target task of looking at images of UC. ResNet50 was also much more complex than the baseline, which only helps it in a difficult classification task.

ResNet50 performed well overall with accuracy scores in some metrics, like AUROC, being on par with scores seen in CNN models used in similar binary classification tasks.



**Fig 4.** ResNet50

**Link to Folder Containing Data and Code:** <https://drive.google.com/drive/folders/15_IIC7WzyY01HOifO84-kpe3Uh7anmcp?usp=sharing>

**Applied Setting:**

This model will aid gastroenterologists in classifying images of patients’ colonoscopies. This is important as human perception varies amongst physicians and evaluation of endoscopic images is crucial in the diagnosis of UC. Getting a second opinion from a gastroenterologist is time-consuming and costly. Patients often present with worse health outcomes from misdiagnosis of UC due to the use of improper treatment. This model can be utilized to support gastroenterologists in making diagnostic decisions in clinical practice.

Will a gastroenterologist, with the aid of a deep learning model, be able to correctly classify the presence of Ulcerative Colitis in patients, differently on average than a gastroenterologist with no aid?

* Null Hypothesis: DL model + Gastroenterologist Accuracy = Gastroenterologist Accuracy
* Alternative Hypothesis: DL model + Gastroenterologist Accuracy != Gastroenterologist Accuracy
* Outcome of Interest: Accuracy Rate in the classification of UC
* Target Population: The population will focus on gastroenterologists that evaluate endoscopic images and diagnosis lower GI pathologies.
* Experiment Conditions:
  + Treatment group will be comprised of gastroenterologists that will classify images taken from colonoscopies of patients with no previous diagnosis of lower GI disease with the aid of a deep learning model.
  + Control group with be comprised of gastroenterologists that will classify the same group of patients with no aid from the DL model.
  + A panel of expert gastroenterologists will determine the accuracy of classification for both the treatment and control groups.
* Unit of randomization: Gastroenterologist.
  + They will need to be randomized because gastroenterologists have various levels of exposure to certain diseases, different levels of expertise, and different levels of experience. We need to account for years of practice as a gastroenterologist: a gastroenterologist that has practiced for more years would have more experience in the field and would be more likely to have attended more GI-specific conferences. We would also like to know the subspecialty of our gastroenterologists as a gastroenterologist subspecializing in upper GI diseases would not have the same level of domain knowledge as one that subspecializes in lower GI diseases. The number of colonoscopies performed and the number of UC patients a gastroenterologist has worked with also play a factor in the level of experience a gastroenterologist has, where there is a positive correlation between experience and exposure.

**The Good, The Bad, and The Ugly: Lessons Learned**

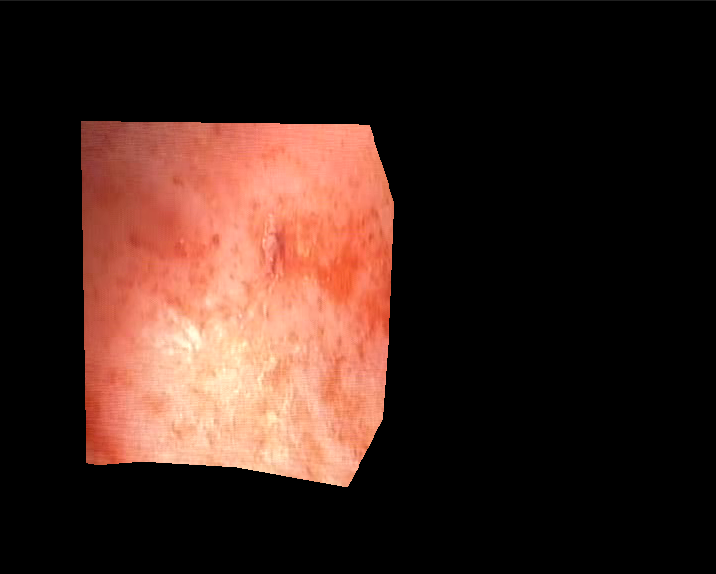
The dataset was extremely tough to work with due to my lack of experience with working with this type of data. There were many things to consider when working with frames from videos. The original dataset was extremely large, being around 1 TB in size. I was unable to work with the data on Google Colab, my personal computer, and Rutgers’s Amarel. I utilized a friend’s computer, which had a 4 TB solid-state drive, to filter through the original dataset to create a dataset with my frames of interest. This issue could have possibly been solved if I knew a way to work with the data in batches.

Another issue I had working with frames occurred when looking through the images and trying to solve the problem of data leakage. The problems and solutions were mentioned in **Data, Data Pipeline Plans, and Data Summary.** In doing so, I reduced the original amount of data from 51621 frames and 531 videos of patients to 17011 frames and 334 patients in my final dataset. This left me with less data than I have anticipated and most likely impacted my model.

The training, validation, and test split were done based on patients (not videos of patients), which created an uneven amount of frames going into each data split since some patients had more frames than others. This made the data difficult to stratify, leading to class imbalances. Evaluation metrics, like accuracy, become misleading due to class imbalances. Accuracy was kept and F1-score was added as an additional evaluation metric. If time permitted, I would have tested oversampling UC and undersampling non-UC frames.

The original dataset included segmentation masks that could be used to improve the performance of the model by giving. Masks were images that give information on areas of interest in a frame. New mask images were generated by turning a frame and its corresponding mask, if it existed, into an array and finding the product. The product would then be converted back to a PNG file and saved in the *\*\_masked* folders. The masks added complexity that I have yet to have the skills to master and were excluded from the final product.

Even with all the difficulties and setbacks, I enjoyed working on this project. From applying new preprocessing techniques and deep learning concepts to learning more about the disease that ails me. If I had more time, I would have done a lot of things differently, like using cross-validation for model selection. In all, I found the project an overall success considering having no baseline knowledge of deep learning at the beginning of the project.



**Fig 5**: Left: *0075-seq\_01-frames\_000083.png* Right: *0075-seq\_01-frames\_000083\_masked.png*

Bottom: original *0075-seq\_01-frames\_000083\_masked.png (recreated with paint since Amarel was down during the writing of this proposal)*

References:

1. Xu F, Dahlhamer JM, Zammitti EP, Wheaton AG, Croft JB. Health-Risk Behaviors and Chronic Conditions Among Adults with Inflammatory Bowel Disease - United States, 2015 and 2016. MMWR Morb Mortal Wkly Rep. 2018 Feb 16;67(6):190-195. doi: 10.15585/mmwr.mm6706a4. PMID: 29447146; PMCID: PMC5815485.
2. Lichtenstein GR, Shahabi A, Seabury SA, Lakdawalla DN, Espinosa OD, Green S, Brauer M, Baldassano RN. Lifetime Economic Burden of Crohn's Disease and Ulcerative Colitis by Age at Diagnosis. Clin Gastroenterol Hepatol. 2020 Apr;18(4):889-897.e10. doi: 10.1016/j.cgh.2019.07.022. Epub 2019 Jul 18. PMID: 31326606.
3. Park KT, Ehrlich OG, Allen JI, Meadows P, Szigethy EM, Henrichsen K, Kim SC, Lawton RC, Murphy SM, Regueiro M, Rubin DT, Engel-Nitz NM, Heller CA. Corrigendum to The Cost of Inflammatory Bowel Disease: An Initiative From the Crohn's & Colitis Foundation. Inflamm Bowel Dis. 2020 Jun 18;26(7):1118. doi: 10.1093/ibd/izaa081. Erratum for: Inflamm Bowel Dis. 2020 Jan 1;26(1):1-10. PMID: 32374862; PMCID: PMC7301401.
4. Lee HS, Choe J, Lee HJ, Hwang SW, Park SH, Yang DH, Kim KJ, Ye BD, Byeon JS, Myung SJ, Yoon YS, Yu CS, Kim JH, Yang SK. Change in the diagnosis of inflammatory bowel disease: a hospital-based cohort study from Korea. Intest Res. 2016 Jul;14(3):258-63. doi: 10.5217/ir.2016.14.3.258. Epub 2016 Jun 27. PMID: 27433148; PMCID: PMC4945530.