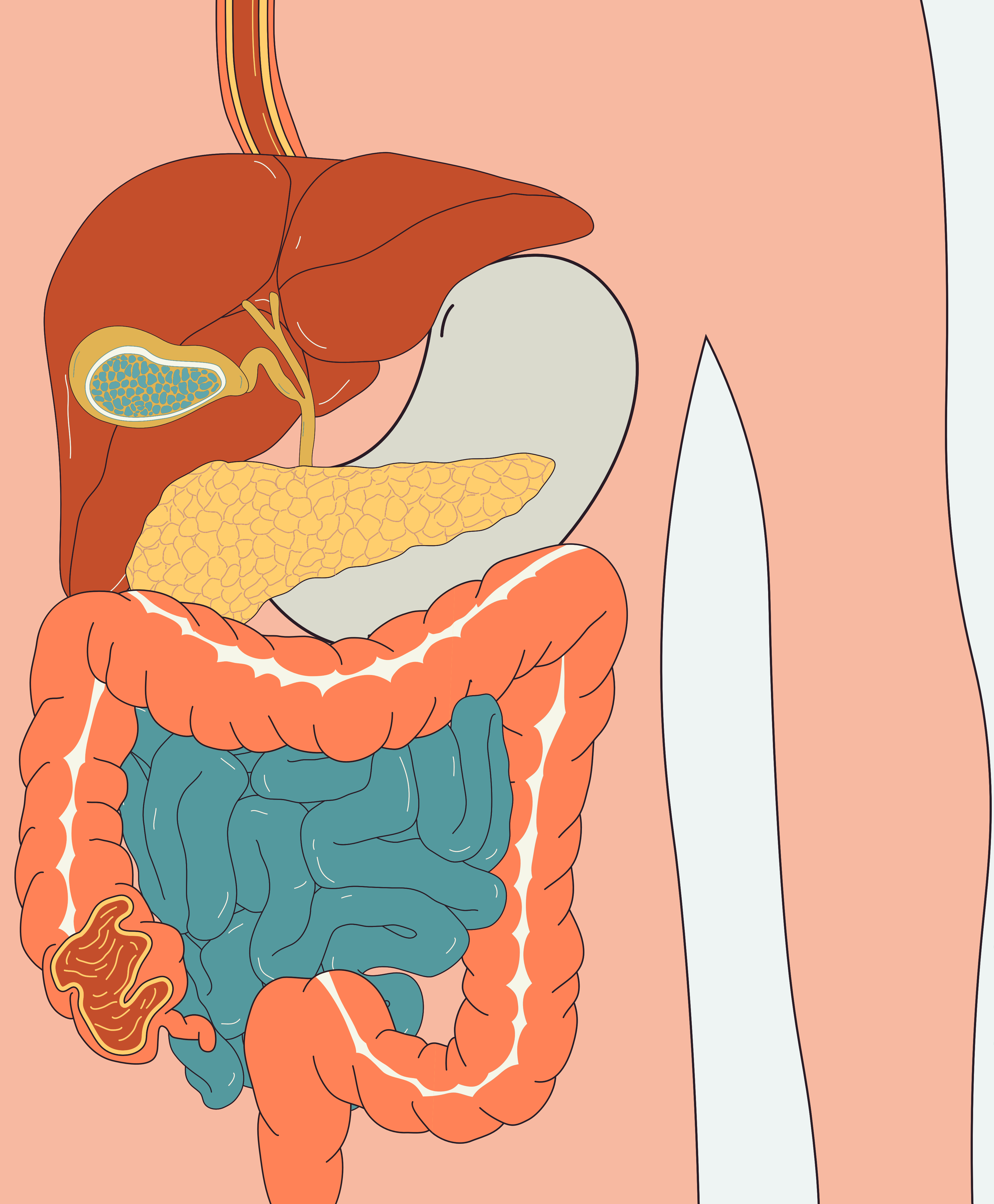
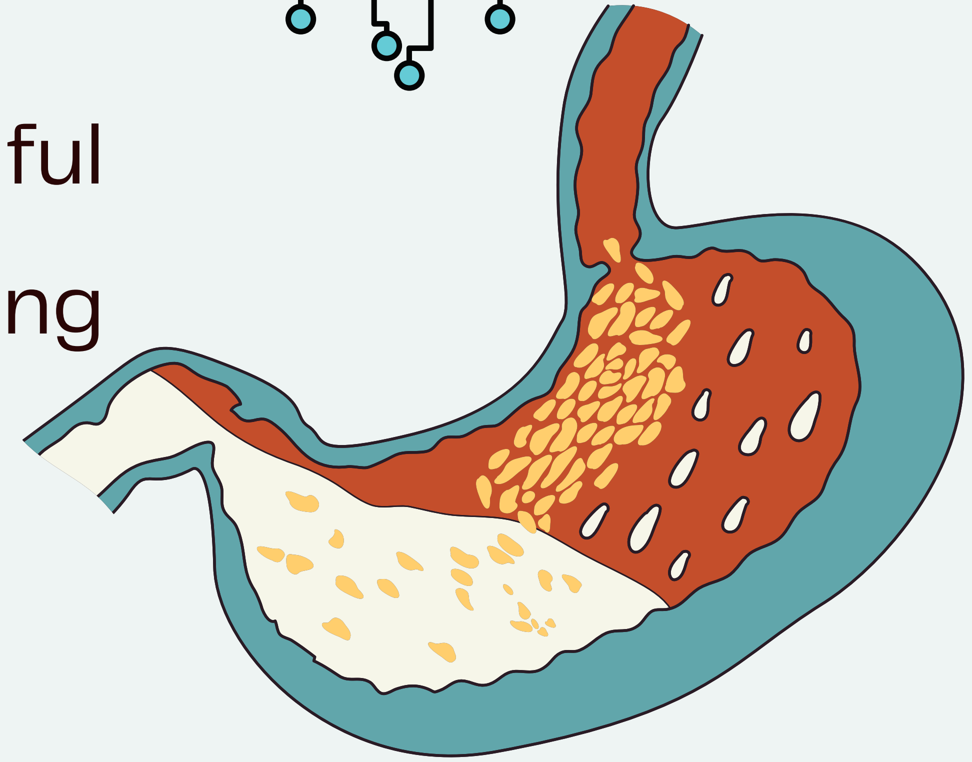


SCALABLE MEDICAL IMAGE ANALYTICS WITH DASK AND PYSPARK: A CASE STUDY ON THE KVASIR DATASET

GROUP NAME	MEDAI VISIONARIES
GROUP MEMBERS	ABDUL MOIZ KHAN SHUMAILA JAVED

PROJECT OVERVIEW

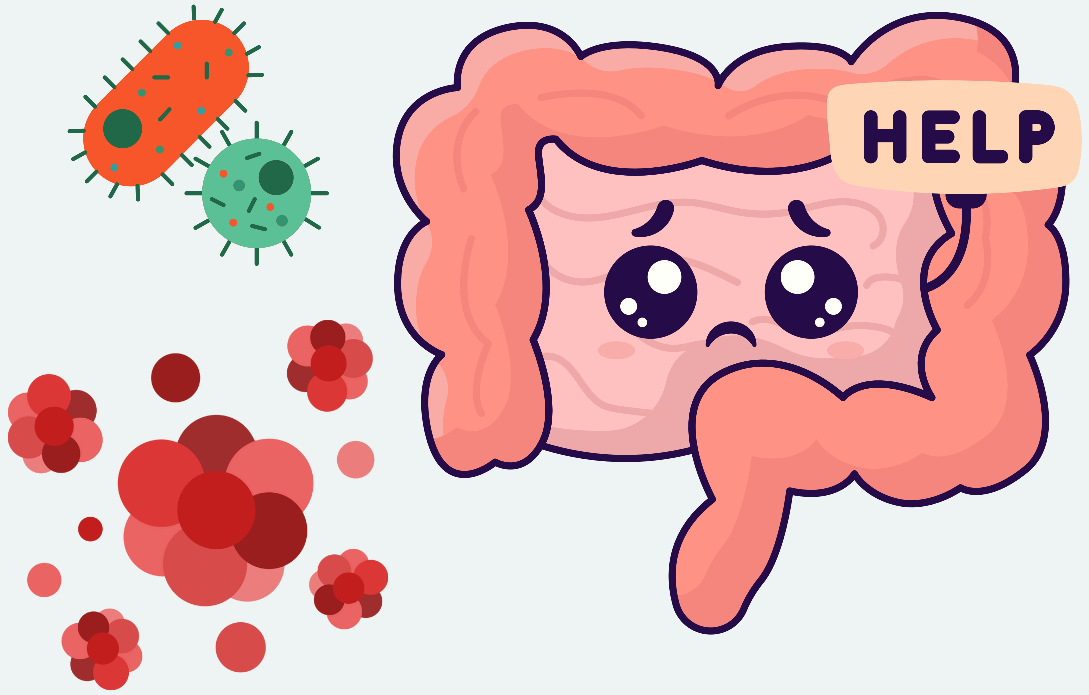
- A scalable, end-to-end data analytics and machine learning pipeline using the **Kvasir** dataset, a comprehensive collection of endoscopic images of the gastrointestinal (GI) tract.
- Focusing on transforming raw medical images into meaningful insights through distributed data processing, deep learning based disease detection, and interactive visual analytics.
- Dask/Spark for large-scale image analysis
- Power BI for dynamic visualization
- Develop predictive models capable of detecting abnormalities such as polyps, which are early indicators of colorectal cancer.



DATASET DESCRIPTION

- Name: Kvasir Dataset (Version 2 from Kaggle)
- Size: Approximately 2–5 GB (depending on extracted image metadata and augmentations)
- Type: Image dataset containing labelled images from endoscopic examinations
- Classes include:

1. Dyed-lifted polyps
2. Normal z-line
3. Polyps
4. Ulcerative colitis
5. Esophagitis
6. Dyed-resection margins
7. Normal pylorus, Cecum, etc.



The dataset also includes annotation masks, which can be analyzed for segmentation or used to extract pixel-level features such as area, shape, or color intensity.

KEY QUESTIONS

1.Exploratory Questions:

- What is the distribution of gastrointestinal conditions in the dataset (by class)?
- Are there any imbalances or biases across categories?
- What are the average image dimensions and color intensity patterns per class?
- How does dataset quality vary (missing labels, corrupted images, etc.)?

2.Analytical Questions:

- Can we identify patterns in color histograms or texture features that help distinguish between healthy and abnormal tissues?
- Can we use statistical inference to validate if certain image features significantly correlate with polyp presence?
- What are the most informative image features (brightness, contrast, saturation, edge density) for detecting abnormalities?

3.Predictive Questions (Stretch Goal):

- How does model performance change when distributed training is applied using Dask or PySpark MLlib?

PROJECT STAGES

01

DATA PREPARATION

Load and validate the Kvasir dataset using Dask or PySpark; handle missing or corrupted data, resize images, and extract metadata such as dimensions, brightness, and class labels.

02

FEATURE ENGINEERING

Perform distributed EDA to explore class distribution, pixel intensity variations, average RGB histograms, and correlations between image features.

03

VISUALISATION AND DASHBOARD

Build interactive Power BI dashboards visualizing dataset composition, polyp detection trends, class frequency, and feature relationships.

04

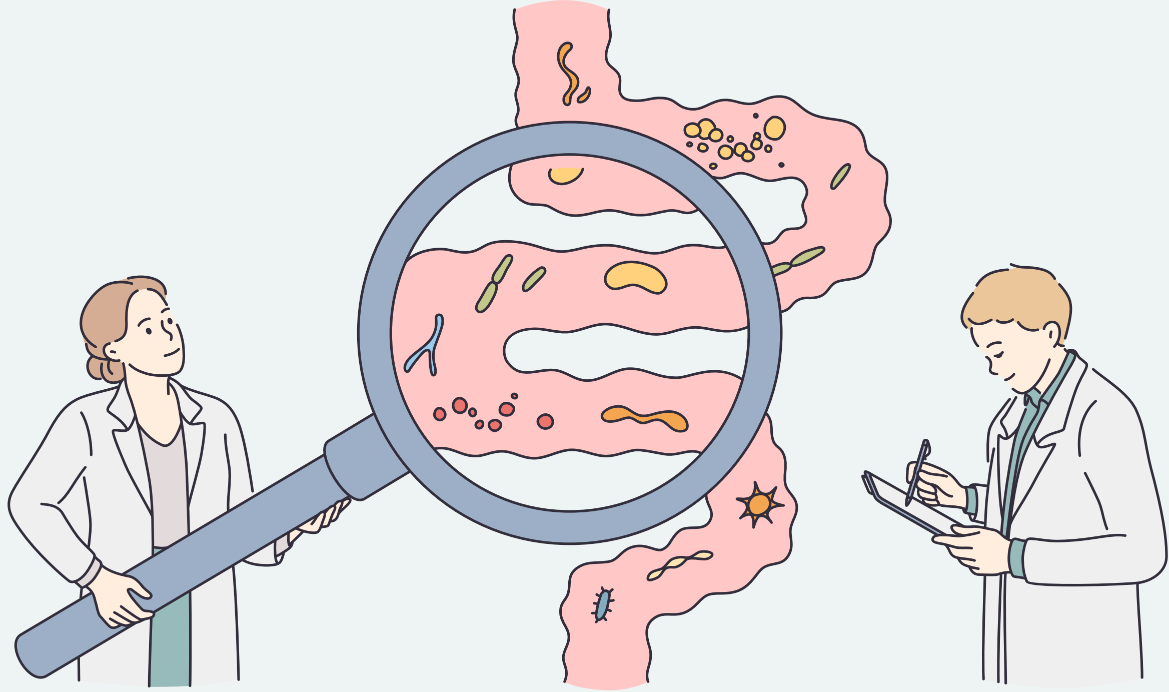
ML TRAINING

Train convolutional neural network (CNN) models to classify image categories; evaluate accuracy, precision, recall, and confusion matrix.

05

FINAL REPORT

Summarize insights, Power BI dashboard findings, and predictive model performance through a comprehensive narrative.



EXPECTED OUTCOME

By the end of this project, we expect to achieve:

- A clear understanding of the distribution and quality of the Kvasir dataset.
- Insights into visual patterns and statistical characteristics of polyps and related GI diseases.
- Experience with big data frameworks (Dask/PySpark) for large-scale medical image analysis.
- A well-designed Power BI dashboard that effectively communicates findings to a non-technical audience.