# Parva Capsula – “swallow the future”

## Colorectal Cancer Diagnosis

*A scientifically proven venture in saving thousands of lives*


Sarvesh Prabhu

Senior, Lambert High School, Suwanee, GA

November 5th, 2022

IFoRE

POWERED BY SIGMA XI

November 3–6, 2022 | Alexandria, Virginia

Powered by

SIGMA XI

THE SCIENTIFIC RESEARCH HONOR SOCIETY

# Time is a virtue in winning the race with cancer

52,580 American lives were already lost in 2022; 1.4m worldwide

Colorectal cancer is a silent killer; but highly preventable if diagnosed earlier

To see the way forward, let's look back.

**Traditional Colonoscopy**, but people aren't motivated; why?

"We as the scientific community must converge with physicians to fulfill our social responsibility to save these lives."

IFoRE
POWERED BY SIGMA XI

# My solution is Parva Capsula

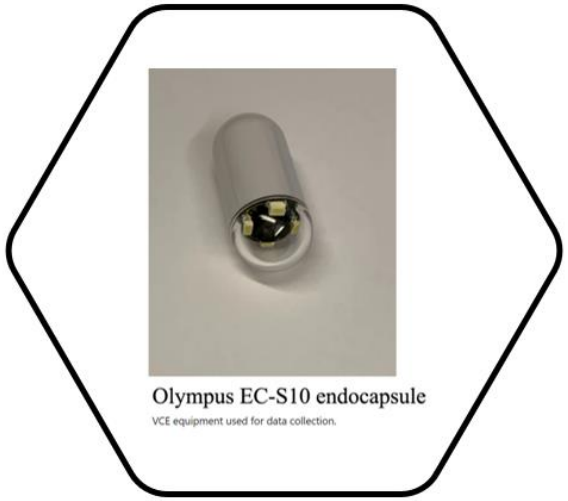
## A swallowable smart pill with a camera

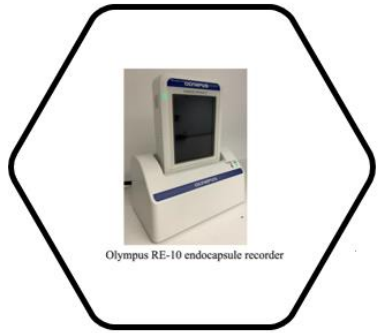
Imagine your colonoscopy is done while you are performing your regular daily duties.

- Olympus EC-S10 endocapsule
- Size of a vitamin tablet (26 mm x 13 mm size)

The actual devices used to capture the input dataset are shown on the right.

Patients of Bærum Hospital, Norway.



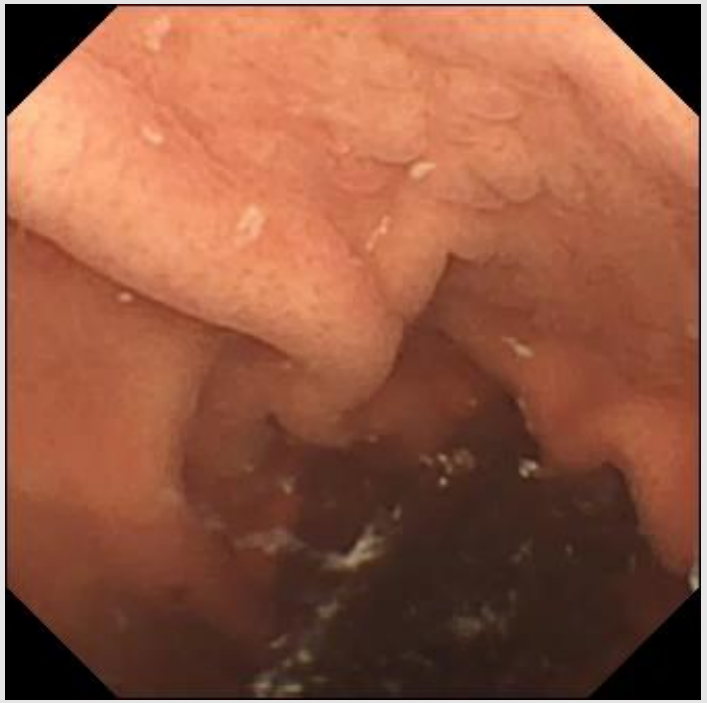




### Solution Rationale

- **Patient-centric**
- **Non-Invasive**
- **Hospitable Sustainable**
- **Scalable**
- **Profitable & VC-friendly**
- **GI Doc** views the recommendations in the comfort of their home
- **Hospital:** Precious space in the emergency is saved for other patients
- **Cost-effective** for Insurance carriers

But there is a headwind – GI Docs don't trust the outcome; accuracy matters

IFoRE
POWERED BY SIGMA XI



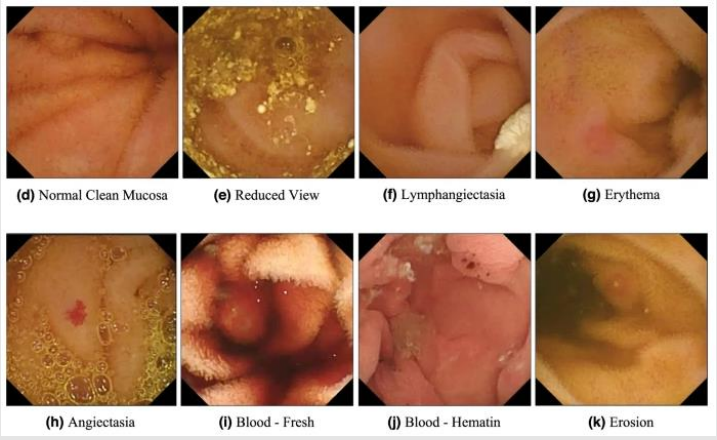
# Deep Learning & Random Forest Ensemble

Achieve the highest consistent accuracy in diagnosing colorectal cancer using medical imagery analysis.

Perform GI diagnosis in two ways in ECL: Deep Learning using the convolutional neural network & TensorFlow and Random Forest model.

Compare the model outcomes, hypertuning, and ensembling.

Conclude the research and publish results for the Healthcare, IFoRE, and Sigma Xi.




## HyperKvasir Dataset

- Data is collected from various patients of **Bærum Hospital, Norway**
- Partly labeled by experienced gastroenterologists
- **1m total images** (336 x 336 pixels), 374 videos at 6fps
- *4.74m images are also available but not used in the research*

# Preparing the models

Eliminating the noise in input images

Used a confusion matrix to plot the pixels and image resolution of all the 4.74m images

Picked the best 1m images of higher resolution for better prediction

**F1 score:** A measure of a test's accuracy by calculating the **harmonic mean of the precision and recall**

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

**Matthews correlation coefficient:** MCC considers true and false positives and negatives and is a balanced measure even if the classes are of very different sizes.

The $t_k$ is the number of times class k actually occurred, $p_k$ is the number of times class k was predicted, c is the total number of samples correctly predicted, and s is the total number of samples.

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}$$

# Solution rationale

## Model 1: Deep Learning using GNN bundle

- Feature extraction of biomarkers
- 78:22 ratio of training and test data
- Adams for training and passes the input to the binary cross entropy models through the kernel convolutions
- Adams optimizer descent algorithm for training and binary cross-entropy as the loss
- Tensor dimensions are [3 x 60 x 60]

```
compileDef :=    compile(optimizer=tf.keras.optimizers.Adam(),
                 loss=tf.keras.losses.binary_crossentropy,
                 metrics=[tf.keras.metrics.Accuracy(),
                          tf.keras.metrics.Precision(),
                          tf.keras.metrics.Recall()])
```

Kernel Convolution is measured by

$$(K * I)(i, j) = \sum_m \sum_n (I - m, j - n)K(m,n)$$

Precision 76-81%, Recall 86%

## Model 2: Random Forest learning tree

- Runs multiple decisions trees in a “forest pattern”
- The classification was based upon anomalies numerically, using number codes to identify and detect a cancerous biomarker through classification
- Input classification cross-verified over 24 trees over two variables per tree
- Also utilizes a 78:22 ratio of training and test data
- Max depth at 255 and the forest size at 10

Random Forest Classification is measured by

$$\text{Gini} = 1 - \Sigma^{c}_{i=1} (P_i)^2$$

Precision 92.3%, Recall 96%

IFoRE
POWERED BY SIGMA XI

# Experimentation and final algorithm

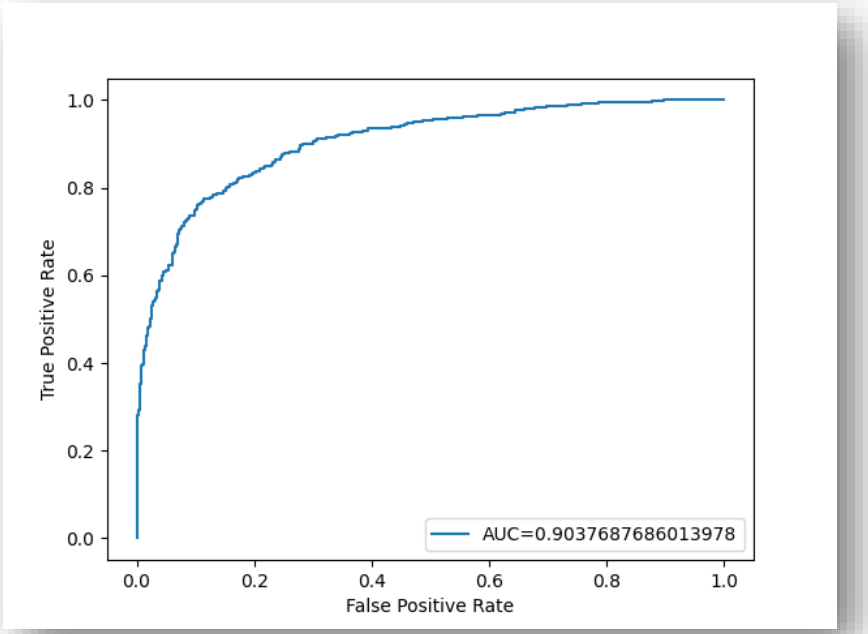
The experimentation contains three parts.

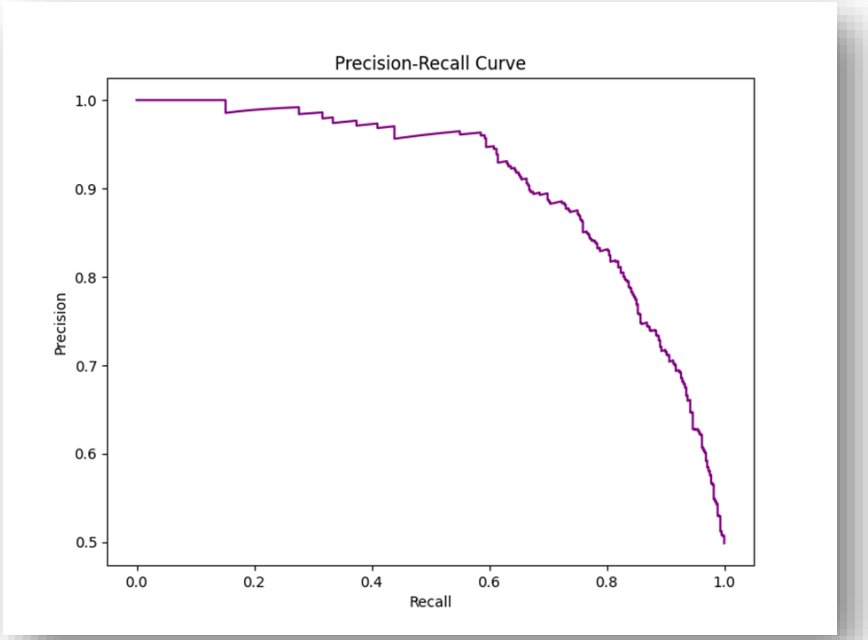
1) Execute the GNN model inference 1,000 times for every raw image acquired by the smart pill with optimized parameters for every run. Accuracy is 80.6%.
2) Feed the GNN biomarkers and pre-defined labels to the Random Forest model, attaining the most confident outcome by counting the most votes for the conclusion. Accuracy improved to 92.3%.
3) Further improved the accuracy by broadening the input to not just one image but a collection of images.

    So, I traveled 3 seconds before and 3 seconds after for every image to get a cohesive view, angle, and lighting conditions. The smart pill I used yielded six frames per second, allowing 36 images to explore.

    For each of the 36 images, I used GNN to retrieve the biomarkers and fed the $\emptyset^{n}(x) = T(\emptyset^{n-1}(x))$ input vectors to Random Forest, boosting the precision with a **99.8%** accuracy rate.

# ROC & Precision-Recall Curve




## ROC Curve of Random Forest

The area under the curve is .903 with Random Forest, which is better;

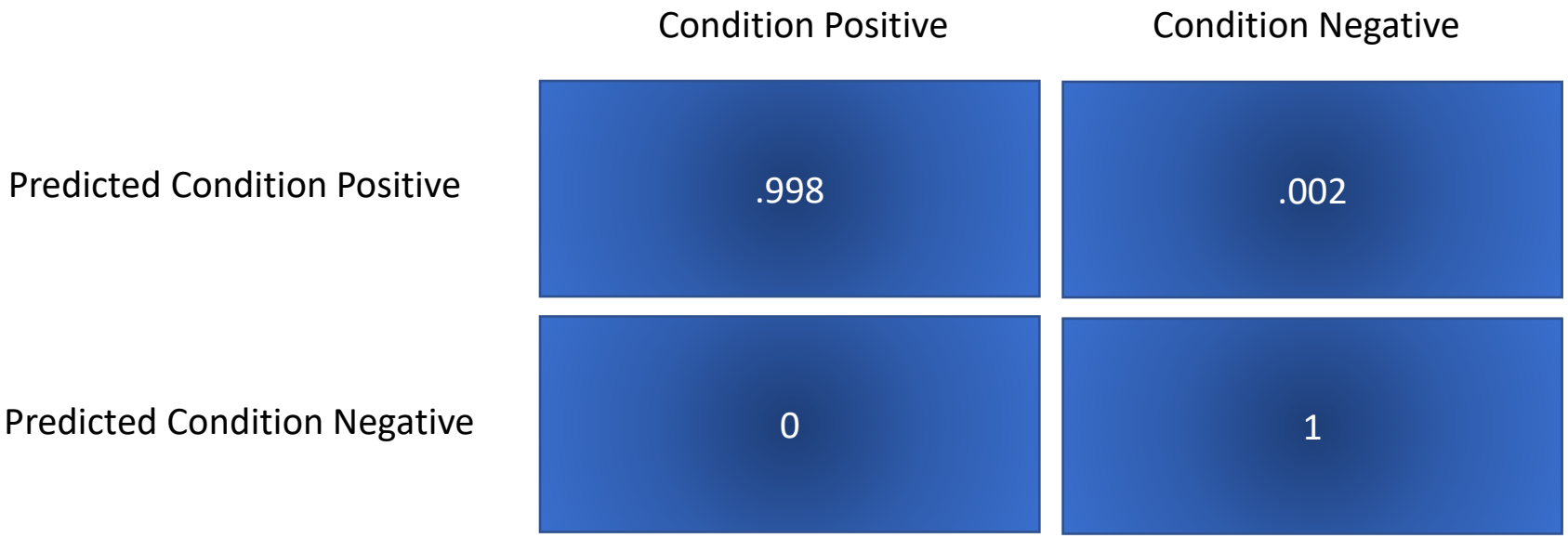
I desire better accuracy for physician recommendations for patient diagnosis in pragmatic applications.

## Precision-Recall Curve of the ensembled model

The precision-recall curve generated from the ensembled model confirms the highest precision achieved with 99.8% accuracy and 100% recall.

Lastly, I wanted to view the results in a physician-decipherable method as a confusion matrix.

# Confusion matrix of the ensembled model- “physician approachable”



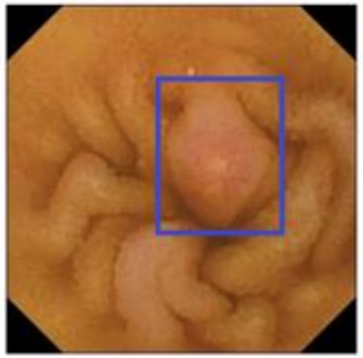
# The Verdict

The research concludes that ensembling the GNN and multi-layered Random Forest yields the highest accuracy and consistent results.

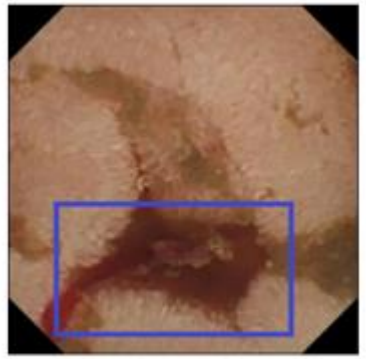
The ensembling yields 99.8% precision and 100% recall, enabling the practitioner to make an informed treatment decision using a non-invasive smart pill for the patient in need of urgent care.

Research Artifacts & Code in GIT Hub
SarveshPrabhu90/GI-Imagery-Analysis-Models

**Diagnosis:** Ulcer



Erosion



Digestive tract bleeding

# Next Steps

Furthering myself as an entrepreneur and making Parva a reality.

Partnering with AGA leadership, SonarMD, and Dr. Kosinski, Larry, effective December 19, 2022.

Partner with Illinois Gastroenterology Group (IGG) to field test the solution with GI Docs.

# Thank you!

and to saving tomorrow's lives today!

Research Artifacts & Code in GIT Hub
SarveshPrabhu90/GI-Imagery-Analysis-Models


## Sarvesh Prabhu

Senior, Lambert High School, Suwanee, GA
November 5th, 2022


IFoRE
POWERED BY SIGMA XI
November 3–6, 2022 | Alexandria, Virginia

Powered by
SIGMA XI
THE SCIENTIFIC RESEARCH HONOR SOCIETY