Colorectal Cancer diagnosis: A comparative study between Neural Networks vs. Random Forest Deep learning for business use cases

June 20, 2022

Abstract

The advancement of AI & ML in the medical field, particularly cancer diagnosis, has long been held back on the basis of accuracy limitations and a lack of trust from the practitioners. Since the diagnosis changes the course of action taken on a patient, any error, whether a false positive or negative, leads to loss of life or potentially unnecessary treatment; As a result, ML has not played the role of a primary predictor, acting as an optional aid.

The scope of my research is to get to a consistently accurate diagnosis, possible by highlighting the areas of interest to the Physician (whether it’s a polyp, ulcer, etc.), allowing faster conclusion several hours faster than a traditional procedure while being non-invasive. The accuracy is realized by comparing the two dominant ML models: Neural Networks and Random Forest. These two models will be created using two of the expansive HPCC bundles: the Generalized Neural Network, and the Learning Trees.

Both models are fed images of various colorectal cancer biomarkers, including but not limited to Polyps, Ulcers, and Colon Erosions, and photos of a standard, functional colon pathway. The training and test dataset represents various upper and lower GI tract use cases. The research utilized up to 110,000 images and 890,000 images for rest purposes. For better accuracy, the research utilized the images from both traditional scope and Smart Pill.

The first model, the generalized neural network (as pertained to the bundle), utilizes sparse cross-entropy to classify and detect multiple biomarkers within the training dataset. By identifying the biomarkers, the model makes the decision as to whether the patient has colorectal cancer; with a lack of a marker showing no cancer.

The second model, the Random Forest, will use a repeated binary cross-entropy pattern using a multitude of splits, redistributing these results using SORT, and further analyzing in a sequential order.

Both models have been compared with the same metrics to fairly determine the superior model. Direct accuracy will be measured using precision and recall as well as the F-1 score. The results will also be verified using the log-loss and visualized using a confusion matrix. This way, the two popular methods of image classification will be able to be reliably compared from a value and visual standpoint.

I am ecstatic to offer the conclusion of my research to the Healthcare community and push the envelope on the accuracy of ML models in medical diagnosis and a plethora of future imagery use cases. I am grateful to have had the opportunity to work with the industry's superior HPCC Platform, my awesome mentor Mr. Bob Foreman, and to have been able to utilize its incredible supercomputing and expansive libraries.