Colon Disease Classification Using Machine Learning Techniques

CSCI 5622
Mitch Fulton, Dongyao Wang, Sushma Colanukudhuru, Chris Pillion

Intestinal diseases currently affect between 60 and 70 million people worldwide. While some of the diseases within the colon are not fatal, many are. Colorectal cancer, for example, is the third most common kind of cancer worldwide, causing 50,000 deaths in 2016 alone. When detected early, treatment provides a 94% chance of survival. With only two-thirds adults being checked for gastrointestinal disease regularly, the detection of ailments from a single colonoscopy has an increased importance. Because the detection of intestinal diseases is almost purely visual, the problem is difficult to address with traditional computational methods. However, with the increased prevalence and maturity of machine learning methods in recent years, a visual approach could be realized.

Diseases within the body manifest themselves differently for each person they affect. However, there are patterns that particular diseases display, and these patterns serve as identifying marks for each disease. It is this uniformity of patterns, combined with the nonuniformity of the manifestation, that caused our group to consider machine learning for this problem. Our current dataset consists of around 300 classified images, each containing one of seven different types of colon disease. This dataset has been provided to us by the Advanced Medical Technologies Lab (AMTL), which is under the Mechanical Engineering department at the University of Colorado. In addition to the still images, several videos were also included in the dataset. While the dataset size is not necessarily ideal to guarantee a robust model off a small training set, we believe it will set up a sufficient baseline to help further research this idea in the future. Our group is looking to expand this dataset in whatever way possible, but will have limited time and resources to do so.

To approach this machine learning problem, we have to first consider what limitations there are within the data. Because the dataset is very small, we must consider doing initial image manipulation to artificially increase the size of the dataset. This includes but is not limited to: scaling, rotation, warping or shearing, and filtering the colors within the images and pulling single frames from the videos. Once the larger dataset has been created the type of machine learning must be chosen. For this particular size of dataset, the algorithms that could learn an acceptable model are limited. Of all the models, we determined that the two implementation possibilities for our use case are Bayesian Learning and Support Vector Machines (SVM). Because the training set is comprised of images, the input has a very high dimension. Of our two possible models that can accurately model such a small dataset, SVMs produce a more accurate prediction with high-dimensional inputs, so this approach will likely be used by our group. In addition to this, there have been other approaches to image and disease classification attempted in the past. A further literature review will be conducted to generate more approaches to our problem.