INTRODUCTION

As the decision situations become increasingly more complex, advanced analytical techniques are gaining popularity in addressing complex classification type decision problems in many fields including healthcare and medicine. Because of the rapid increase in the collection and storage of large quantities of data (facilitated by improving software and hardware capabilities coupled with increasingly lower cost of acquiring and using them), data and model driven decision making is becoming a mainstream practice in every field imaginable (from art to business, medicine to science). One area where faster and better decisions could make a significant difference is in healthcare/medicine. This data rich field can undoubtedly use what modern day decision analytics has to offer. In this study, we used various Machine Learning methods to address some classification type decision problems, namely prediction of chronic kidney disease, using general health check up parameters like blood pressure, etc; and, Tuberculosis, using X-Ray images of the patient.

Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Our kidneys filter wastes and excess fluids from our blood, which are then excreted in our urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in the body.

Tuberculosis is a disease that affects many people in developing countries. While treatment is possible, it requires an accurate diagnosis first. In these countries there are, in many cases, available X-ray machines (through low-cost projects and donations), but often the radiological expertise is missing for accurately assessing the images. An algorithm that could perform this task quickly and cheaply could drastically improve the ability to diagnose and ultimately treat the disease.

In more developed countries, X-ray radiography is often used for screening new arrivals and determining eligibility for a work-permit based on their health. The task of manually examining images is time consuming and an algorithm could increase efficiency, improve performance and ultimately reduce cost of this screening.

There are 2 different kinds of datasets, for covering various sectors of data collection used in healthcare for decision making. The first dataset was taken from the UCI machine learning repository, for the detection of Chronic Kidney Disease (CKD), which has 25 features (eg. red blood cell count, white blood cell count, etc). The target is the 'classification' of the patients as suffering from CKD or not, and prediction of the disease in other patients. The second dataset is obtained from The standard digital image database for Tuberculosis, created by the National Library of Medicine, Maryland, USA in collaboration with Shenzhen No.3 People’s Hospital, Guangdong Medical College, Shenzhen, China, and The tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. It focuses on training the algorithm for detection of Tuberculosis in patients, based on their chest X-ray images.

To start with some of the very best methods, this thesis explores the implementation of Histogram of Oriented Gradients along with Support Vector Machines/ K-nearest neighbor, and neural network based Convolutional Neural Networks (CNNs). These methods have been propitiously utilized for disparate image problems and obtaining breakthrough performances in image detection, segmentation, classification, and other related challenges. HOG is used as feature descriptor for images which is a representation of an image or an image patch that simplifies the image by extracting useful information and throwing away extraneous information. In the HOG feature descriptor, the distribution (histograms) of directions of gradients ( oriented gradients ) are used as features which when fed into an image classification algorithm like SVM or KNN gives good results.

On the other hand, CNNs learn features and train the classifier at the same time in a supervised manner. End-to-end feature extraction and classification can thus be concurrently performed, which has made CNN the most favored choice for image recognition problems. However, CNN with shallow layers do not have enough discriminative power, while deep CNN are computationally expensive to train and can be easily over-fitted. Thus, we need proper parameter training. 3

This study is a testament to the improving capabilities of analytic techniques in support of better decision making, especially in situations constraint by limited resources