Machine Learning Nanodegree: Capstone Project Proposal

By: Huong Ivy Nguyen

This project proposal is one of the Kaggle competitions that I found very interesting. The goal here is not about winning the competition but to use the knowledge and skills that I have gained throughout the Machine Learning nanodegree to tackle this Kaggle competition's problem. This competition can be found at https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening

Domain Background:

Cervical cancer is a type of cancer that has abnormal growth of cells in the cervical area that could potentially spread to the other parts of the body. During its early stage, no symptoms are usually detected. In the later stages, various unusual behaviors including vaginal bleeding or pelvic pain are observed. According to the National Institute of Health, cervical cancer is one of the most common cancers that affects the U.S. women nowadays – ranked 14th in frequency. The most recent data that was obtained by the NIH shows that the incidence rate for cervical cancer was 8.1 cases per 100,000 women per year during 2003-2007 in the United States. In 2010, there were approximately 12,200 women in the U.S. diagnosed with this type of cancer. The mortality rate is 2.4 deaths per 100,000 women per year. However, this type of cancer can be prevented effectively if it is caught at its pre-cancerous stage with using the appropriate treatments.

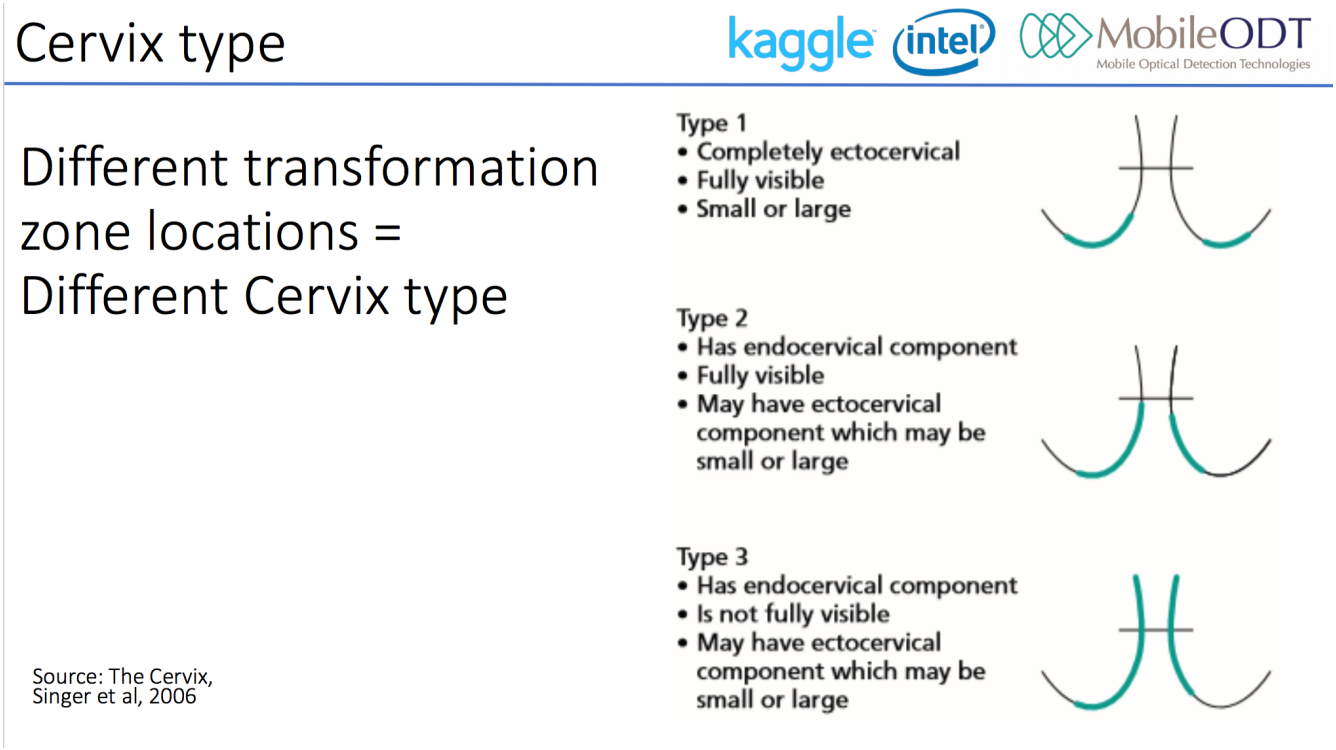
Before the 1940s, cervical cancer was one of the most tricky types of cancers to be treated in the United States due to the lack of medical diagnostic equipment and technology. In the 1950s, the introduction of the Papanicolaou (Pap) smear helped to decrease the number of deaths caused by this disease dramatically. However, this method does not always give the most accurate results since it is a screening test and not a diagnostic test. Moreover, the Pap smear test has been giving false negative up to 50% cases of cervical cancer, which proves the hunt and development of other methods to detect and diagnose cervical abnormal cells are dearly needed. Fortunately, imaging studies have been conducted to obtain better and more accurate results. Even more interesting, with the knowledge of machine learning and computer technology nowadays, these cervical images can be used to develop a high throughput algorithm to classify the type of cervical cancer in its pre-cancerous stage.

Problem Statement:

Cervical cancer is easy to prevent if it is caught at its pre-cancerous stage. The most difficult part is how to deliver the most the effective treatment that is appropriate and responsive to each woman's physiology. A lot of women who have low income and at high risk for cervical cancer do not get the right treatments or screening due to the position of their cervix. This is very unfortunate since health providers are able to identify those who

are at high risk but could not offer the correct treatments to them. Even worse, wrong treatments can be high cost and even bring higher health risks. In addition, a treatment which works for one might not work for the others. Therefore, it is necessary to correctly identify the cervix type so cervical cancer patients will get the correct treatments.

The main goal of this project is to develop an algorithm using Convolutional Neural Networks (CNN) to correctly classify the cervix types based on the provided cervical images for each type; hence choose the suitable treatments for the patients. Different cervix types can be classified by identifying the transformation zone. This is because the location of the transformation zone is not always visible. The visibility level of this particular zone is related to the type of cervix the patient has. A summary of this relationship between the cervix type and the transformation zone location is described in the figure below.



Datasets and Inputs:

The datasets for this project can be found through the following link: https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data. The inputs of this project would be the cervical images of different types.

Solution Statement:

I plan to use deep learning technique, specifically Convolutional Neural Networks (CNN) to build the final solution for this problem. The main reason why I choose CNN over the

others is that CNN is more effective in determining the important features in a given image than most common computer vision techniques. The performance of CNN on the In addition, CNN is very good at filtering and finding specific pixel groupings that are important in the classification process.

Benchmark Model:

I plan to compare the results of the CNN to those obtained from an SVM algorithm. I am expecting that the CNN algorithm will have a much higher accuracy than the SVM algorithm. The primary reason for choosing SVM as a benchmark model is to compare the results between a wide-scoped general machine learning algorithm versus an image-oriented algorithm. If the above assumption regarding the performance of CNN is true, then we can conclude that we should give specific preference to CNN when it comes to image recognition and classification.

Evaluation Metrics:

I plan to use the loss and accuracy functions to evaluate the performance of my CNN model.

Project Design:

Step 1: Upload the datasets and inspect the image of cervix type. This step is to help me have a better visualization and understanding of each type in order to determine whether my CNN works appropriately.

Step 2: Obtain descriptive statistics of the dataset

- How many data points for each cervix type? Is one cervix type more common than the others? If so, would this behavior skew the algorithm at all?
- What are the max, min, and mean of the size the images? (These numbers can be used later for normalization purposes)
- How many types of images are in the dataset?

Step 3: Build the CNN model

- Split the dataset into training and testing set if not already done so
- Training stage:
  - Convert all images to 3-dimensional Numpy arrays
  - Convert labels to one-hot-encoding
  - Convert images to grayscale to reduce to 2 dimensions
  - Build layers for the CNN as previously done in Project 5 of the MLND

    - Train the dataset using Keras API
- Testing and classification stage:
    - Test the trained model against the testing dataset
    - Use the accuracy score and loss value to evaluate how well the performance is

Step 4: Build an SVM model for comparison

- Use the image preprocessing data points from step 3 to build an SVM algorithm using the SVC class of the SVM module in sklearn
- Classify the images using SVM and compare the performance with previous results obtained by CNN