Capestone proposal: Cervical Cancer Screening

Domain Background

Since the inception of Machine Learning people have been applying it across various disciplines. One such domain where I think its application is making life changing difference is Medical Field. Some of the application of ML in Medical imaging are Diabetic retinopathy, Blood Flow Quantification, Tumor Detection etc. One of the problems in which I have liking is Cervical Cancer.

Cervical cancer is a cancer arising from the cervix. It is due to the abnormal growth of cells that can invade or spread to other parts of the body. Early on, typically no symptoms are seen. Later symptoms may include abnormal vaginal bleeding, pelvic pain, or pain during sexual intercourse.

Worldwide, cervical cancer is both the fourth-most common cause of cancer and the fourth-most common cause of death from cancer in women. In 2012, an estimated 528,000 cases of cervical cancer occurred, with 266,000 deaths. This is about 8% of the total cases and total deaths from cancer. About 70% of cervical cancers occur in developing countries. In low-income countries, it is one of the most common causes of cancer death. In developed countries, the widespread use of cervical screening programs has dramatically reduced rates of cervical cancer.

Source: *Cervical cancer - https://en.wikipedia.org*

Other references: *Denny L (2012) Cervical cancer: Prevention and T/t . Discover Med 14: 125-131, Ginsburg OM (2013) Breast and cervical cancer control in low and middle-income countries: Human rights meet sound health policy 1: e35-e41.*

Problem Statement

Cervical cancer is so easy to prevent if caught in its pre-cancerous stage. However, due in part to lacking expertise in the field, one of the greatest challenges of these cervical cancer screen and treat programs is determining the appropriate method of treatment which can vary depending on patients' physiological differences. Especially in rural parts of the world, many women at high risk for cervical cancer are receiving treatment that will not work for them due to the position of their cervix. This is a tragedy: health providers can identify high risk patients

but may not have the skills to reliably discern which treatment which will prevent cancer in these women. Even worse, applying the wrong treatment has a high cost. A treatment which works effectively for one woman may obscure future cancerous growth in another woman, greatly increasing health risks. The solution to this problem I think is to provide the Health Care providers with a system which will determine Cervix type in real time, so that they can easily figure out patient's treatment eligibility based on it.

Datasets and Inputs

Figure 1: Type 1

Figure 2: Type 2

Figure 3: Type 3

To determine the type of treatment fit for a Cervix cancer patient first we need to determine the type of cervix a patient has. Keeping that in mind data has been collected and has been divided into 3 types.

The dataset consists of collection of images of each type of cervix. These different types of cervix in data set are all considered normal (not cancerous), but since the transformation zones aren't always visible, some of the patients require further testing while some don't. This decision is very important for the healthcare provider and critical for the patient. Identifying the transformation zones is not an easy task for the healthcare providers, therefore, an algorithm-aided decision will significantly improve the quality and efficiency of cervical cancer screening for these patients.

Cervix type

Different transformation zone locations = Different Cervix type

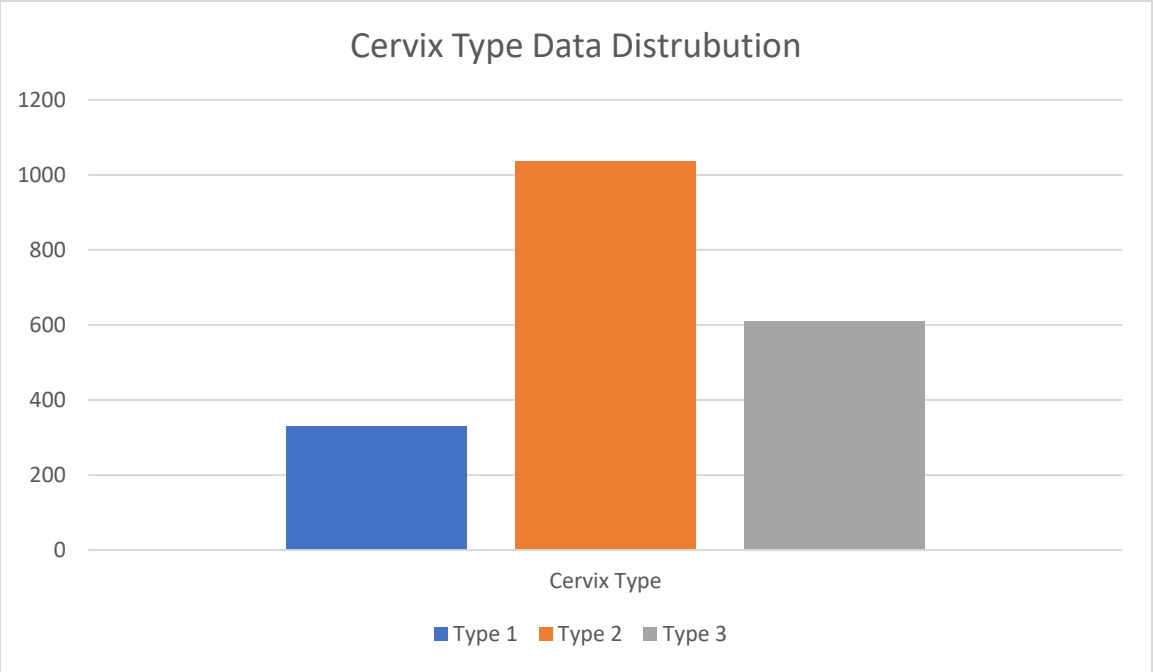
Type 1
- Completely ectocervical
- Fully visible
- Small or large

Type 2
- Has endocervical component
- Fully visible
- May have ectocervical component which may be small or large

Type 3
- Has endocervical component
- Is not fully visible
- May have ectocervical component which may be small or large

Source: The Cervix, Singer et al, 2006

Data Set Distribution



- The above count is of raw data, it may vary after filtering.
- Total samples in the dataset 1975.
- Images have dimension varying from 2000x3000 pixel to 3000x4000 pixel (approx.).

- 80% of the data will be used for training the network and rest will be used for testing.

Data set Source: Cervical cancer - https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data

Solution Statement

The health care providers are facing problems in determining the Cervix type. To assist them with this task a system can be developed which given an image of cervix will determine its type in the real time. Using this information, they will have some ease in determining the cancer treatment fit for the patient.

Benchmark model

Current benchmark for the give problem statement is **0.70**, which is a multiclass loss value basically Categorical Cross entropy Loss, achieved by a team named "***Towards Empirically Stable Training***". They have made use of R-CNN models with VGG-16 feature extractors to achieve this feat.

Evaluation metrics

Will make use of ***Categorical Cross entropy Loss*** metric to evaluate performance of the model.

Categorical Cross entropy Loss = $-\sum_{c=1}^{M} y_{o,c} * \log(p_{o,c})$

where ***M*** is number of classes (dog, cat, fish), ***log*** is the natural log, ***y*** is binary indicator (0 or 1) if class label ***c*** is the correct classification for observation ***o*** and ***p*** is predicted probability observation ***o*** is of class ***c***.

Project Design

The Problem stated in the second section in this proposal is clearly an image classification problem. Considering the high accuracy and performance of Deep neural networks in image processing, I will be making use of Deep learning to achieve the result.

Step one in approaching the solution will be basic analysis and processing data. In this step outliers will be detected and removed. In this case images which doesn't have clear view of cervix will be considered as outliers and will be removed.

Going futher, I will be making use of Keras with TensorFlow backend, as a support framework for creating deep networks. To make data compatible with the keras network they will have to converted to tensors which will require some more pre-processing. Biggest challenge for me in

this project, other than coming up with a network architecture will be to determine the proper size of the input images to the network (*Dimensionality reduction problem*). The images present in the dataset are high quality ranging from 2K to 4K. If I will attempt to directly train the images in 2K or 4K, I will end up dealing with millions of parameters to trains. Futher more, I don't think I have luxury to convert images to greyscale as RGB data may be crucial in training the network, though an attempt will be made to see the difference in the accuracy of both the approaches.

Next, I will work on network architecture. One of the potential architectures that I will be trying is:

Layer	Filter size	N. of filters	Stride	Padding
Conv-1_1	5	32	2	2
Conv-1_2	3	32	1	1
Pool-1	2		2	
Conv-2_1	3	64	1	1
Conv-2_2	3	64	1	1
Pool-2	2		2	
Conv-3_1	3	96	1	1
Conv-3_2	3	96	1	1
Pool-3	2		2	
Conv-4_1	3	128	1	1
Conv-4_2	3	128	1	1
Pool_4	2		2	
Conv-Score	1	5	1	

I will also be trying Transfer learning with Resnet architecture.

Finally, I will conclude the project with comparing my results with the benchmark model.