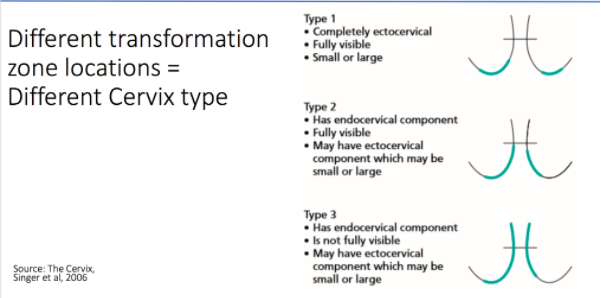
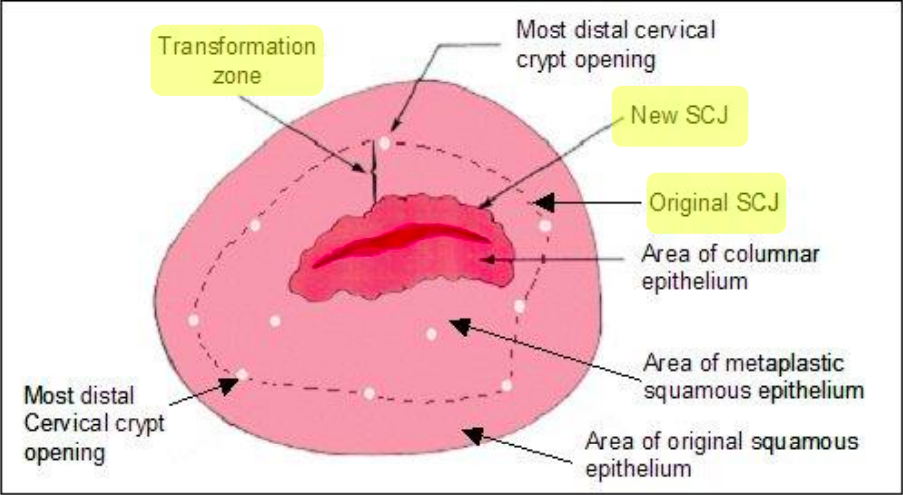
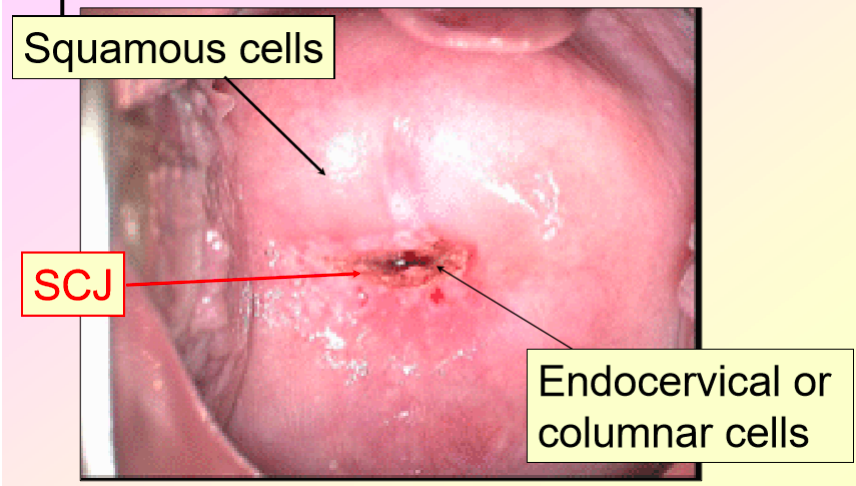
Developing an image-based features recognition algorithm to determine optimal treatment for cervical cancer in low and middle income countries

Keerat Singh

Cervical cancer has become a problem for most third world countries because of their inability to determine the appropriate method of treatment which can vary depending on patients’ physiological differences. Health providers don’t have the skills to identify which treatment will prevent the cancer; this is not only costly but it can even worsen the patient’s condition. Certain treatments that work for some women can obscure future cancer growth in others. By first knowing the type of a woman’s cervix, doctors should be able to treat the cancer accordingly. The goal of this research is to create an image classifier that will be able to determine the type of a woman’s cervix, and therefore the appropriate treatment.

The three cervix types are indicated below.

In essence, all the image classifier will do is it will analyze the numerical properties of different categorical image features and determine what matches the pattern for each type. What is being analyzed and how the analysis occurs are what differentiate solutions from each other.



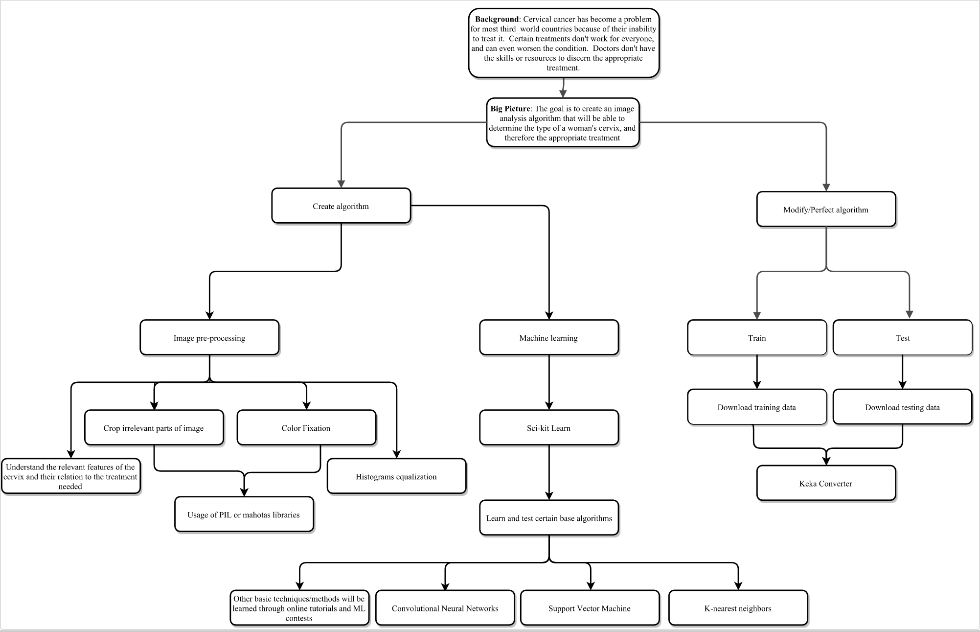
The two most important areas for distinguishing between cervixes are the squamo-columnar junction (SCJ) and the transformation zone. The SCJ is the region in the uterine cervix in which the squamous lining of the vagina is replaced by the columnar epithelium.  This marker defines the lesion in the tissue. The transformation zone is the area between the original SCJ and the new SCJ. Most cervical cancers begin in the cells of the transformation zones.

The most traditional image classification algorithms look for clusters among training data, categorize each cluster based on the associated factors, then match incoming data to those clusters. There are, although, newer algorithms and methods being used in similar studies that could be useful for this research. These not only highlight the algorithms being used, but also certain techniques being used for the preprocessing. The most similar study was creating an automatic CT liver image classification system that would be used to detect two types of liver classification (Chen et. al, 1998). It used a boundary detection and extraction approach for the preprocessing. This is essentially a cropping method that first transforms the entire image into a 16 x 16 feature map of binary values based on whether a section of the map is part of the liver liner (the important feature in the study). Then using a Catmull-Rom algorithm, the detected liver boundary was interpolated back into the image. This is probable possibility, as the cervix contains a transformation zone boundary and an SCJ location that are critical to classification. Once this “cleaning” of the data was completed, they then identified the distinguishing features of the two types of liver cancer. Using an NFB feature curve, they gave numerical values to qualitative data. This seems to be a popular approach in quantifying feature-based data points. I will need to use this feature curve in that both my factors are categorical. They will not be identified on a binary scale rather a spectrum from 1 to 100. They first tried a modified probabilistic neural network with a Kohonen self-organization algorithm to train the data on. This is a statistical approach that yielded unsatisfactory results. It seems that this trainer could have been improved with more data, but it is unclear the true source of error. They found much more success in Specht’s probabilistic neural network. One advantage to this approach is that it increases the learning capabality of the system, which works well with my abundance in training data. It also minimizes the misclassification of false negatives and false positives which is extremely important for any medical classification system. This yielded much better results, meaning I will definitely consider this when training.

Other less relevant, in terms of purpose, studies highlighted a few more interesting ideas that I will try to implement, one of those being histogram equalizations. Histogram equalizations were used in a face recognition classifier outlined in a study (Lawrence, Giles, Tsoi, & Back, 1997). The technique increases color contrast after cropping and detecting valuable features of the image. This could be extremely useful for my study just because the cervix images come with very little difference in color. By increasing the difference in color shades or tones between important features, it would be much easier to later detect in the system. Another interesting idea is the use of gray level thresholding (Saeed & Werdoni, 2014). This preprocessing technique is mainly used for thin boundary lines, as the algorithm searches all individual pixels of an image and classifies them into either object or background pixels. This could be useful if the 16 x 16 feature map doesn’t isolate the parts well enough.

All of the images that will be used to train the algorithm are coming from the Kaggle database. The data files are available through .txt and .7z files, meaning that I will need to download a .7z file converter to access parts of it. The cervices within the images are all considered non-cancerous, but they all showcase the individual transformation zone locations, which ultimately determine the different cervix types. Once the data is fully accessible, I will need to create the first part of the software that will process the images. I have extensive experience with the Python language, but I will need to learn how to use certain libraries that help in analyzing images. These libraries primarily include OpenCV, NumPy, SciPy, and Pillow. The main library, OpenCV, has a full online walkthrough of all of the library’s features that I plan on working through before proceeding. This tutorial contains OpenCV lessons on the image processing functions, image input and output, the 2D features framework, deep neural networks, and the machine learning classes of the library. While the tutorial does contain brief introductions to the deep learning aspects of it, it will mainly serve as an introduction to the library itself. To learn the image classification skills needed to create the algorithm, I will take the online CV-Tricks tutorials that explain how to use different pre-existing algorithms, such as convolutional neural networks, a support vector machine, and K nearest neighbors. In addition, Kaggle has multiple machine learning sample projects that serve as useful introductions. I feel that it is important to get a basic understanding of the main part of my project before delving into the other subsections of it. Therefore, the first thing I will do is going to be to complete two of the introductory problems offered by Kaggle (Titanic and Facebook Recruiting). These problems will run parallel to the image processing. Therefore, I will also be coding some of the preprocessing functions into Jupyter notebook with the training datasets at the same time. By learning the basics of machine learning at first, I will have a better understanding of how the initial image pre-processing will fit in with the rest of the project.

Below is a diagram of each objective moving forward in the project. It is essentially a map of everything mentioned in the above paragraph with some minor additions.



My end goal with this project is to create a fully-functioning, easily accessible application that will take in an image of a cervix and output its type based on its transformation zone. Finishing the project with an app is the optimal scenario because that is how my research could most help those countries in need. I have pre-existing experience with Java and Swift, which are the languages I plan to use if Python isn’t sufficient. This cheap and reliable technology can become an invaluable asset in the cervical cancer toolkit for healthcare providers in less developed regions of the world.

References

Chen, E., Chung, P., Chen, C., Tsai, H., & Chang, C. (1998). An automatic diagnostic system for CT liver image classification. *IEEE Transactions on Biomedical Engineering,* *45*(6), 783-794. doi:10.1109/10.678613

Lawrence, S., Giles, C., Tsoi, A. C., & Back, A. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks,* *8*(1), 98-113. doi:10.1109/72.554195

Saeed, K., & Werdoni, M. (2014). A New Approach for Hand-Palm Recognition. *Enhanced Methods in Computer Security, Biometric and Artificial Intelligence Systems,* 185-194. doi:10.1007/0-387-23484-5\_18