Cervical Cancer Detection Pipeline with Synthetic Data

Sean Wade

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

David Wingate, Chair
Michael Jones
Christopher Archibald

Department of Computer Science

Brigham Young University

ABSTRACT

Cervical Cancer Detection Pipeline with Synthetic Data

Sean Wade
Department of Computer Science, BYU
Master of Science

Cervical cancer is one of the deadliest cancers amoung women worldwide. Every year there are over 250,000 deaths and 550,000 new diagnosis. These statistis are tragic because of how disproportionately they effect different populations, with low to middle income countries accounting for 85% of cervical cancer diagnosis.

One of the primary reasons for the inequality is the cost and availability of cancer screenings. Early diagnosis is extreamly effictive for treating cervical cancer. The goal of this project is to help in the development of low cost sensors and new algorithms to detect cervical cancer for poor areas. My contribution in this project is creating data infrastructure to prepare, augment, and train on the data. This is done through the implementation or several synthetic data generation algorithms, scripts for common pipeline tasks, and developing a python package for extending the pipelie.

Keywords: synthetic data, data augmentation, data pipeline, cervical cancer, medical imaging

Table of Contents

1 Introduction **1**

1.1 The Pap smear screening 1

1.2 Project Goals 2

2 Datasets **3**

2.1 SIPaKMeD 3

2.2 Herlev 4

2.3 Extensibility to future datasets 5

3 Synthetic Data Generation **6**

3.1 Tradidtional Data Augmentation 6

3.2 Inserting Cells 8

3.3 Conditional Generative Adversarial Networks 9

4 Conclusion **13**

Appendices **14**

A Cell Type Descriptions **15**

References **17**

Chapter 1

Introduction

Cervical cancer is one of the deadliest cancers amoung women worldwide. Every year there are over 250,000 deaths and 550,000 new diagnosis. Another tragic aspect of these statistics is how disproportionately they effect different populations, with low to middle income countries accounting for 85% of cervical cancer diagnosis.

One of the primary reasons for this inequality is the cost and availability of cancer screenings. But when caught early, cervical cancer can actually be fully cured. The pre-cancerous lesions of cervical cancer take almost a decade to convert into cancerous ones, leaving a longer than usual timeline for treatment. [1]

1.1 The Pap smear screening

The pap-smear screening was developed by Georges Papanicolaou. Using a small brush, a cytological sample is taken from the cervix and smeared onto a thin glass slide. To clarify the cells characteristics, the smear is stained using a special dye. This emphisizes the different components of the cells with specific colors, making it more clear in a microscope.[5] Each microscope slide contains up to 300,000 single cells with defferent orientations and overlap[5]. This has made automatic segmentation methods challenging.

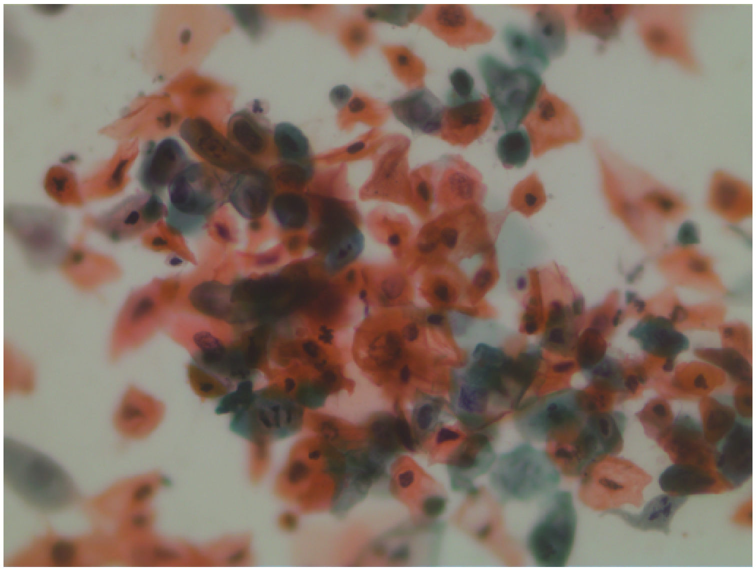


Figure 1.1: Example Pap smear slide

1.2 Project Goals

The goal of this project is to help in the development of low cost sensors and new algorithms to detect cervical cancer. Recent advancments in image segmentation using deep neural networks provides much room for improvment on current methods. The primary bottleneck for this specific application is the lack of labeled data. This can be attributed to two main factors: cells must be segmented by skilled pathologists, which is expensive and time consuming, and strict privacy laws arround medical data.

In this project I addressed these challanges by building a image pipeline to solve the labeled data problem for Pap smear slides. To make this pipeline extensible to future datasets, I built the python package MediAug. By simply writing a small connector to structure the data, this library can take in images and augment them to an infite dataset. The augmentation methods range from simple standard practices, such as rotation, to complex methods like synthetic cell generation using generative adversarial neural networks.

Chapter 2

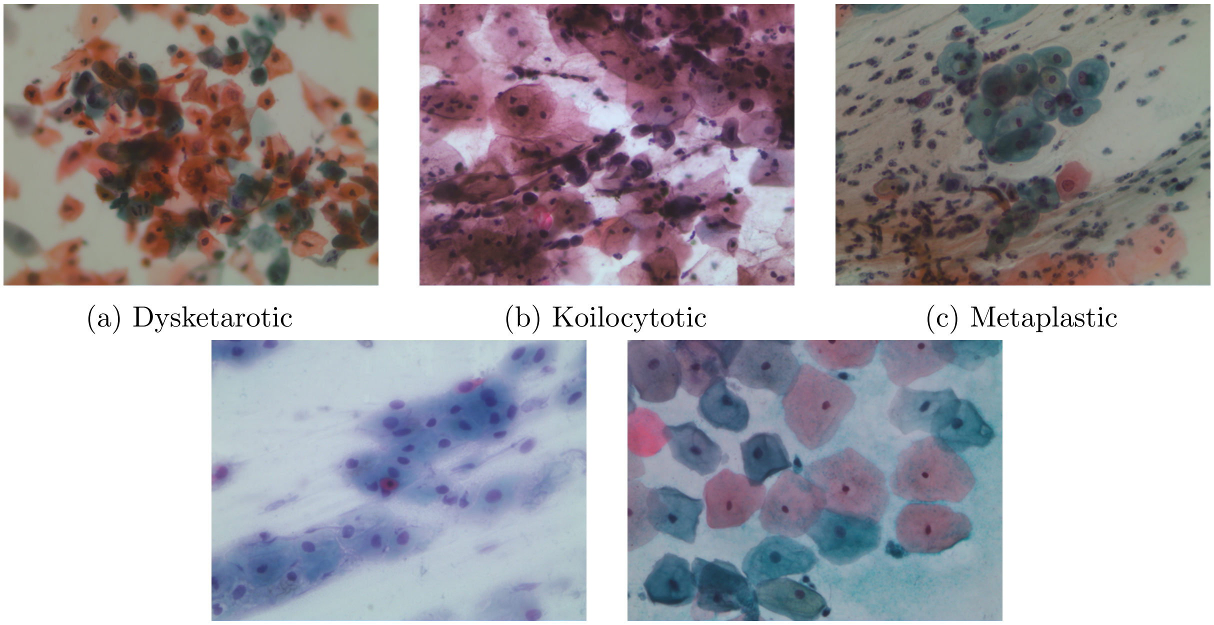
Datasets

Due to privacy with mediacl data and the effort required to label, there are only two open Pap smear datasets. For my project I used these two and made the pipeline generalizable to future datasets.

2.1 SIPaKMeD

The SIPaKMeD dataset consists of 996 cluser cell images of Pap smear slides. From these, there are 4049 indavidual cells that are segmented cyto-technicians. The resolution of these slides is 0.201 µm / pixel, with the final slide being 2048 × 1536. The cell segmentation is stored as a array of poygons.

These cells are grouped into 5 categories: (a) Dyskeratotic, (b) Koilocytotic, (c) Metaplastic, (d) Parabasal and (e) Superficial-Intermediate. Out of these categories, a-c are cancerous and d-e are normal. A full detailed description of each class is given in the appendix.



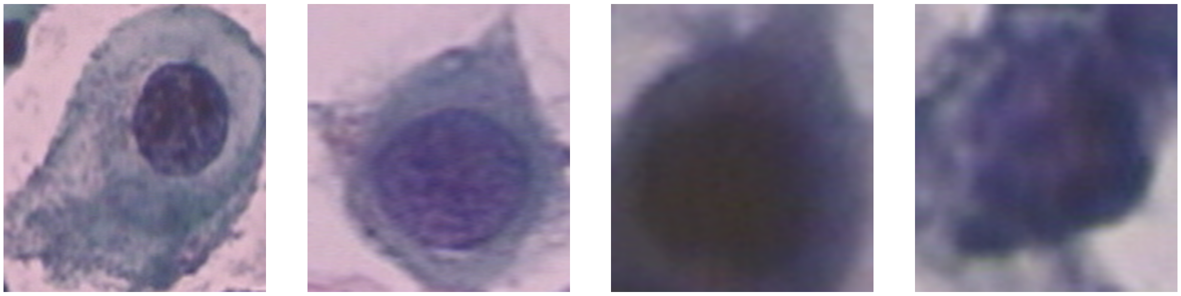
(a) Dysketarotic (b) Koilocytotic (c) Metaplastic

(d) Parabasal (e) Superficial-Intermediate

Figure 2.1: Cell Types

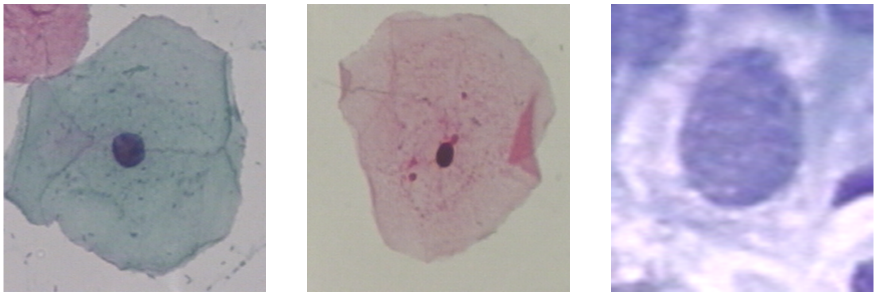
2.2 Herlev

The Herlev dataset is comprised of 917 isolated, single cell images. These are distributed unequally between seven different classes of cells. Superficial squamous, intermediate squamous, columnar, mild dysplasia, moderate dysplasia, severe dysplasia and carcinoma in situ.[4]



Light Dysplastic Moderate Dysplastic Severe Dysplastic Carcinoma in Situ

Figure 2.2: Abnormal Cells



Intermediate Superficiel Columnar

Figure 2.3: Normal Cells

2.3 Extensibility to future datasets

Since we are developing a low cost sensor, all these tools I made will fit the new data into the pipline. This is done by simply extending the DatasetConnector class to put the data in the correct format for the Dataset class. The correct format is simply:

```
newfolder
├─ class1
│  ├─ img1.png
│  ├─ img2.png
│  └─ ...
├─ class2
│  ├─ img1.png
│  ├─ img2.png
│  └─ ...
└─ ...
```

More details are given in the MediAug documentation.

Chapter 3

Synthetic Data Generation

Augmented image datasets have become invaluable to current state of the art computer vision algorithms. Data hungry neural networks require masize amounts of data to learn the match patterns in the distribution of images. Even in situations where there are plenty of data available it is often not labeled and synthetic data is still neccesary.

Another important benifit of augmenting image data is that it serves as a very effective means of regularization. Regularization can be defined as a modication we make to a learning algorithm that is intended to reduce generalization error, but not training error. With a large neural networ and small number of samples, it is very easy for a model to have high varience by memorizing the data. Continually perturbing, rotating, and zooming prevent this by lowering the training accuray.

3.1 Tradidtional Data Augmentation

Traditional data augmentation relies on making small perterbations that do not comprimise of the symantic content of the image. Depending on the domain of the data, several different types of data augmentaion can be used. This could be randomly fliping and rotating the image. Others include random color altertions, adding noise, random zooming, ect. When choosing what opperations to perform it is important to consider if the resulting image still falls in the distribution of ones images.

For Pap smear slides and cervical cells there are many ways to generate new data. Cell images are rotationally invarient and scale invarient, so we can randomly rotate and zoom in to differnt parts of the slide. In addition this data is well suited to elastic deformations and altering the colors.

The MediAug package proviedes a simple API to apply traditional image augmentation operations. The Augmentor class takes in a Dataset and outputs augmented images. This class can be used as either a generator function, or it can write a out a new dataset. To choose which operations the Augmentor does, you can add a step to the pipeline called an Operation. These inlude methods such as flipping, zooming, random cropping, and elastic distortion. An Operation also has a probability for happening that can be set. Bellow are several examples of these simple augmentaions on a slide.

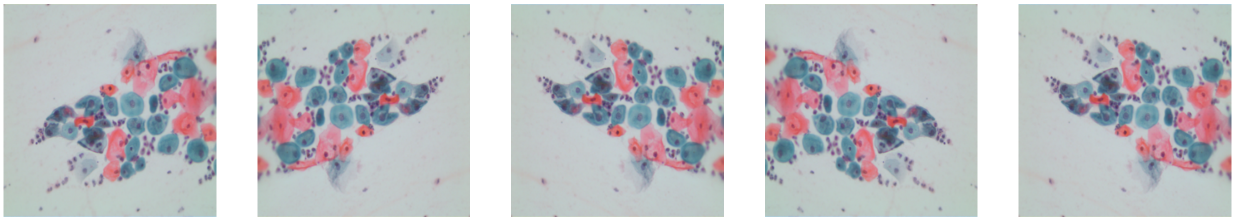


Figure 3.1: Horizantal and vertical flipping

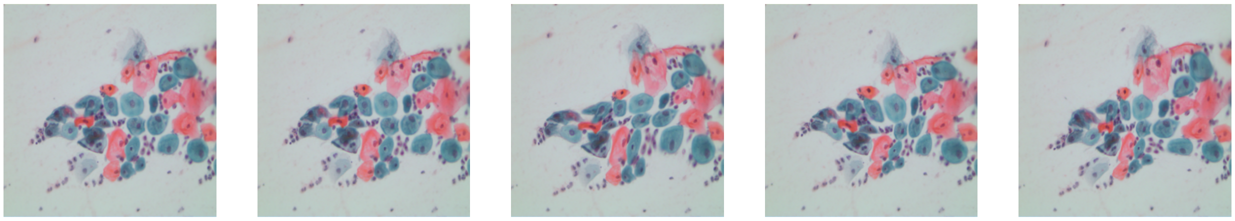


Figure 3.2: Elastic Deformation

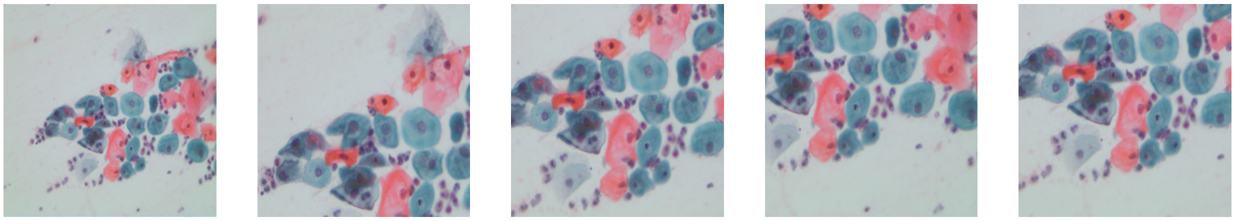


Figure 3.3: Croppin

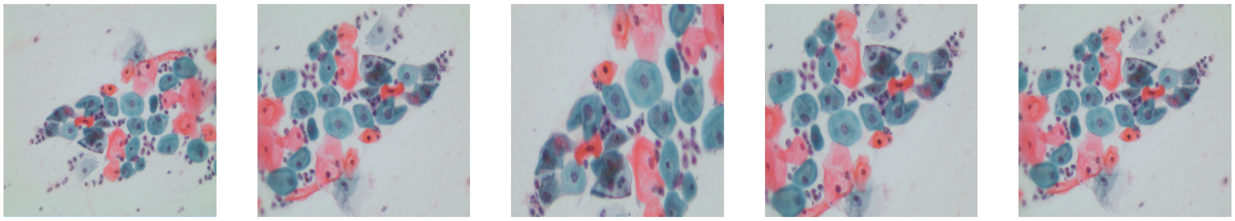


Figure 3.4: Rotation

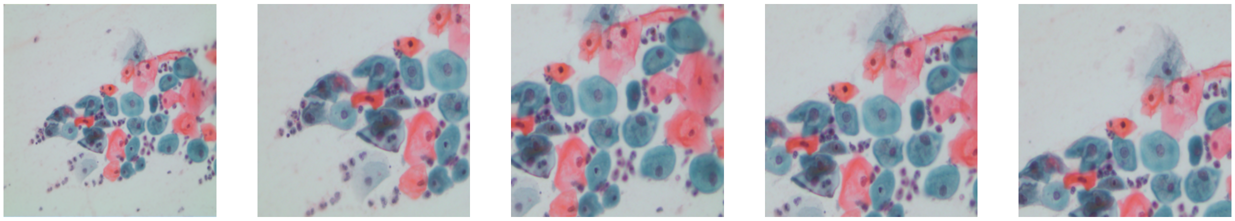


Figure 3.5: Zoom

3.2 Inserting Cells

The primary use case this pipeline is built for is segmenting cancerous cells on a slide. Both availabld datasets are not sufficent to solve this task. The SIPaKMeD dataset has some of the cells labeled, but not all. This results in negative signals for correct segmentations while training.

MediAug is able to address this problem by inserting known cells and masks onto slides. Since we know the ground truth for a health slide is no segmentation, we can then add random cancerous cells with their corresponding segmentation to the slide. This is a weakly supervised data augmentation method that can gives us the dataset we want.

When a Dataset is created, you can specify which classes of slides to use and what classes of cells. There are then a variety of hyperparameters such as range of the amount of cells to add, rotation, scale, ect. To help the cells blend in once inserted I dialate the cell mask and then do a guassian blur kernel. I then use the result as an alpha mask to blend it in so the edges are not as sharp. These parameters need to be adjusted to get the best resuts.

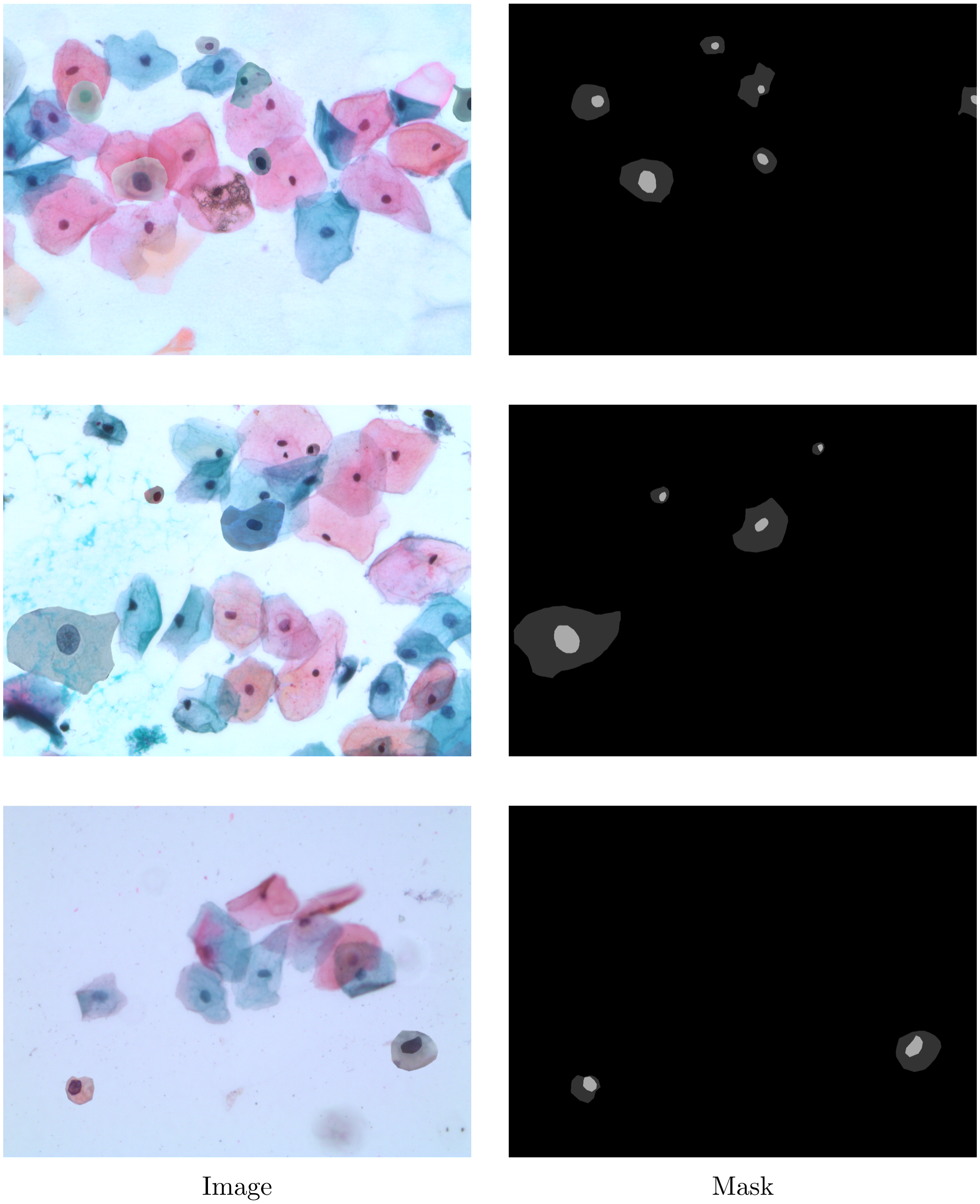


Figure 3.6: Inserted cell examples

3.3 Conditional Generative Adversarial Networks

Another major contribution to the package is using generative adversarial networks (GANs) to generate completly new images and their corresponding segmentation masks. To do this I implemented the Pix2Pix conditional GAN paper [3]. The high level idea is given an input image and random vector, the network will produce and output image that is drawn from the data distribution condtional on the input image. For my case the input was the mask of a cell and the output was the image of the cell. With this GAN trained I can generate new cwll images and know their segmentation map.

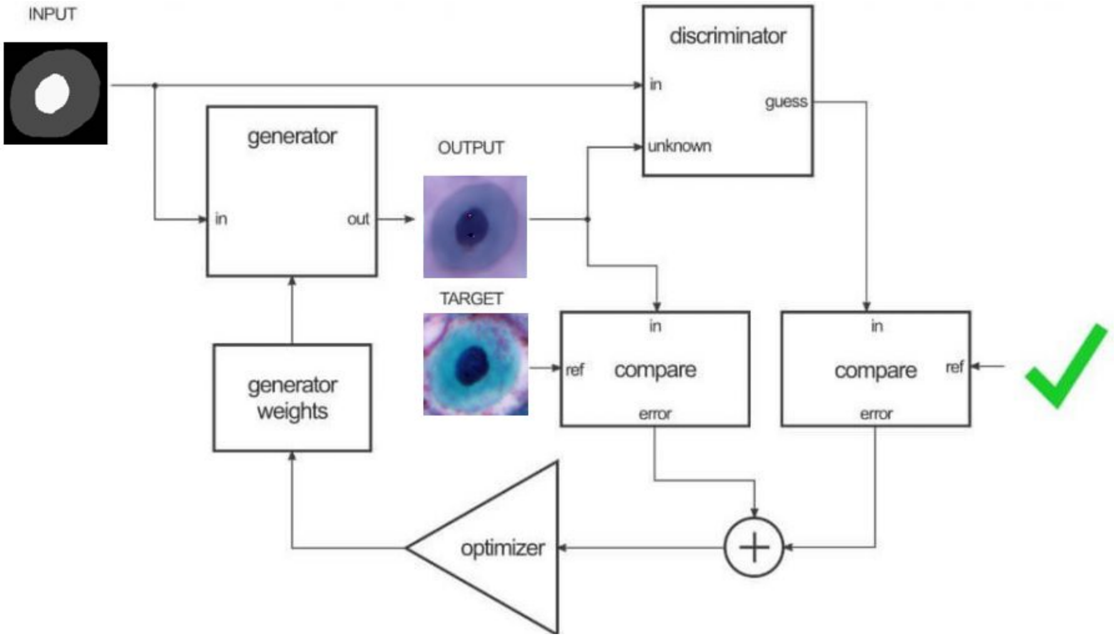


Figure 3.7: Pix2Pix Archetecture [2]

More specifically we are learning a mapping from an observed image, x, and a random noise vector z. This gives $G : \{x, z\} \rightarrow y$. The noise vector is important because otherwise this would just be deterministic. By picking z we are randomly sampling from a distibution of cells with this shape. As illustraited above the archetecture has two networks, the generator G and descriminator D. The generator produces y and the descriminiator takes in a real y and generated y D and gives an estimate of the probability that y is real

Mathematically the objective function of a conditional GAN can be expressed as:

$$\mathcal{L}_{cGAN} = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]$$

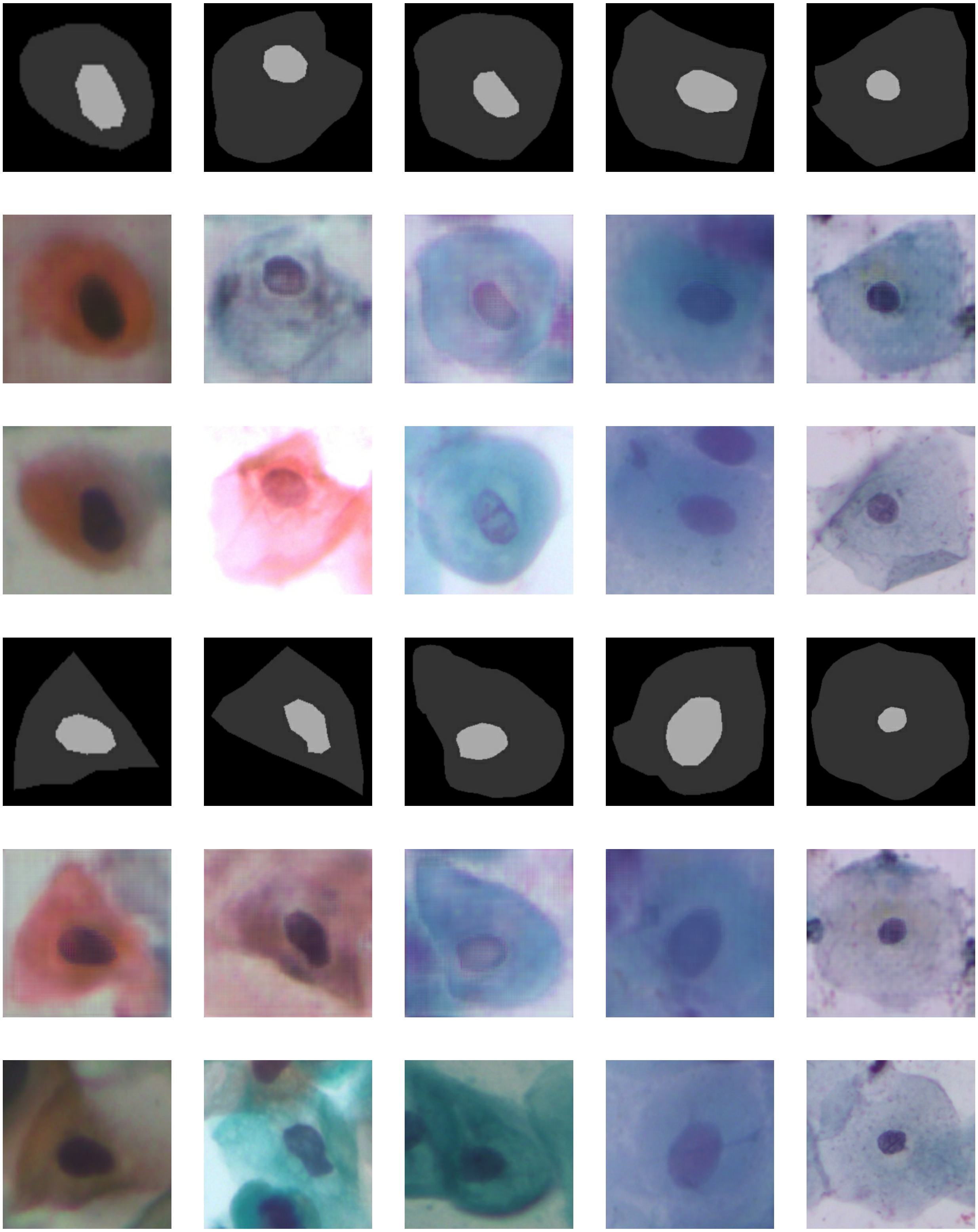


Figure 3.8: Cell generation from input mask

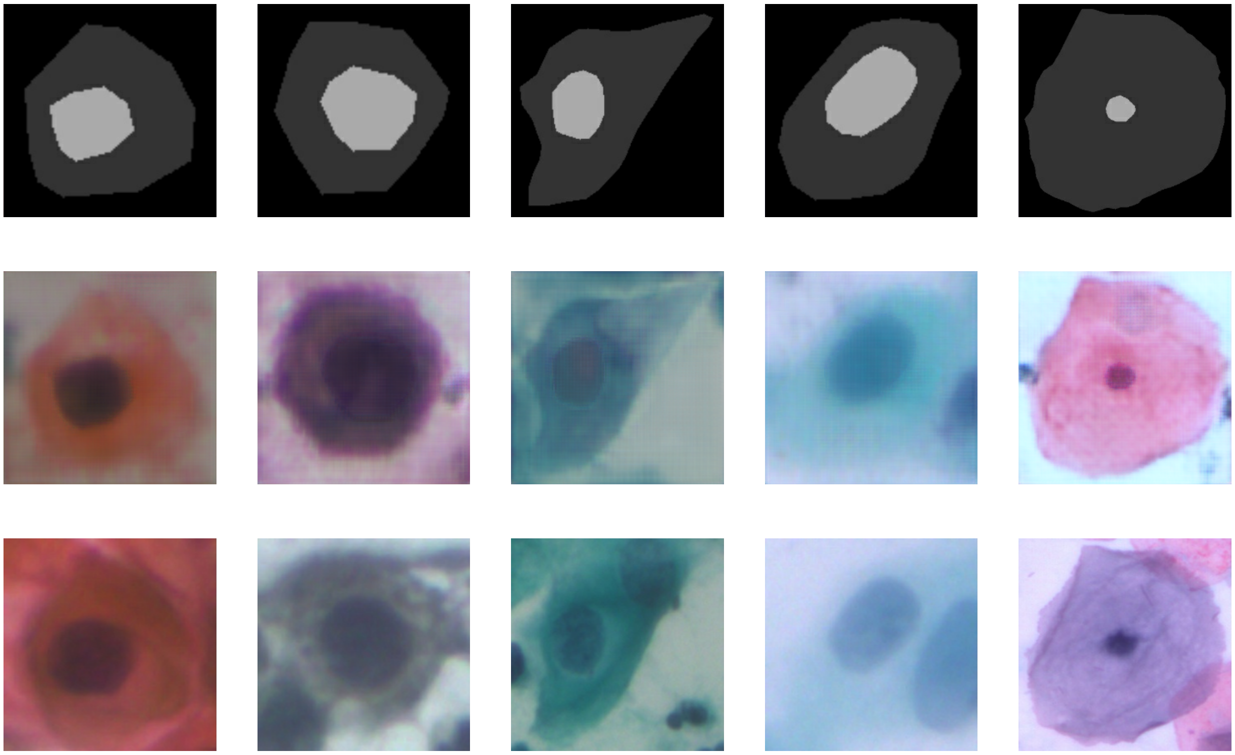


Figure 3.9: Cell generation from input mask

The generated synthetic cells above were trained with asegmnetation mask and cell not in the training data. This similarity between the generated and real shows how well the GAN works. To additionally verify the value of the synthetic data, a simple convolutional neural network was trained on only synthetic images and told to classify the cell. Once trained the networks accuracy was compared using sythetic test data and real test data, and both were within 1%.

Chapter 4

Conclusion

Going forward in cervical cancer research, the pipline I created will be very useful. The MediAug package provides tools to create weekly supervised training sets that are infinite in size. The APi to do this allows the user to tune rotation, position, scale, and alpha. To further stretch the limited value of the limited data, scripts are provided to train a condtional GAN to create completly new cells.

Appendices

Appendix A

Cell Type Descriptions

Superficial-Intermediate cells constitute the majority of the cells found in a Pap test. Usually they are flat with round, oval or polygonal shape cytoplasm stains mostly eosinophilic or cyanophilic. They contain a central pycnotic nucleus. They have well defined, large polygonal cytoplasm and easily recognized nuclear limits (small pycnotic in the superficial and vesicular nuclei in intermediate cells). These type of cells show the characteristics morphological changes (koilocytic atypia) due to more severe lessions.

Parabasal cells are immature squamous cells and they are the smallest epithelial cells seen on a typical vaginal smear. The cytoplasm is generally cyanophilic and they usually contain a large vesicular nucleus. It must be noted that parabasal cells have similar morphological characteristic with the cells identified as metaplastic cells and it is difficult to be distinguished from them.

Koilocytotic cells correspond most commonly in mature squamous cells (intermediate and superficial) and some times in metaplastic type koilocytotic cells. They appear most often cyanophilic, very lightly stained and they are characterized by a large perinuclear cavity. The periphery of the cytoplasm is very dense stained. The nuclei of koilocytes are usually enlarged, eccentrically located, hyperchromatic and exhibit irregularity of the nuclear membrane contour.

Dysketarotic cells are squamous cells which undergone premature abnormal keratinization within individual cells or more often in three-dimensional clusters. They exhibit a brilliant orangeophilic cytoplasm. They are characterized by the presence of vesicular nuclei, identical to the nuclei of koilcytotic cells. In many cases there are binucleated and/or multinucleated cells.

Metaplastic Cells are small or large parabasal-type cells with prominent cellular borders, often exhibiting eccentric nuclei and sometimes containing a large intracellular vacuole. The staining in the center portion is usually light brown and it often differs from that in the marginal portion. Also, there is essentially a darker-stained cytoplasm and they exhibit great uniformity of size and shape compared to the parabasal cells, as their characteristic is the well defined, almost round shape of cytoplasm.[6]

References

[1] A. sreedevi, r. javed, and a. dinesh, epidemiology of cervical cancer with special focus on india, int j womens health, vol. 7, pp. 40514, 2015.

[2] Pix2pix - image-to-image translation neural network, Nov 2018. URL `https://neurohive.io/en/popular-networks/pix2pix-image-to-image-translation/`.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[4] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005*, pages 1–9. NiSIS, 2005.

[5] Jonas Norup. Classification of pap-smear data by tranduction neuro-fuzzy methods. 2005.

[6] Marina Plissiti, P. Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, O. Krikoni, and Antonia Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. pages 3144–3148, 10 2018. doi: 10.1109/ICIP.2018.8451588.