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

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. 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.[1]

Cervical cancer accounts for 6.6% of cancer cases in the world. In 2018 alone, there were well over 570,000 cases. Even worse is the disproportionate way populations are effected. About 85% of deaths from cervical cancer came from lot to middle income countries. Fortunatly the high mortality rate from cervical cancer can be reduced by effective screening and early treatment.

Recent advances in neural networks can be applied to the early detection of cervical cancer. The primary bottleneck is the lack of labeled data and cost of generating it. This project will focus on new ways of generating synthetic cell data and using it to augment state of the art computer vision algorithms.

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 the Papanicolaou method. 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.

1.2 Project Goals

The Goal of this project is to develop a data pipeline and tools to take in raw pap-smear slide data and output1

-

Chapter 2

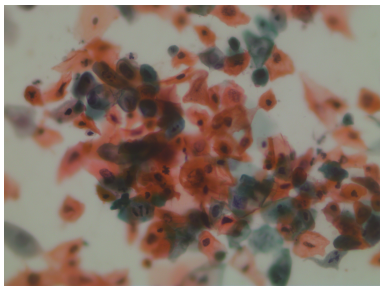
Datasets

Due to privacy with mediacl data and effort in labeling, there are only two open Pap smear datasets. For my project I use these two and make the pipeline generalizable to future datasets.

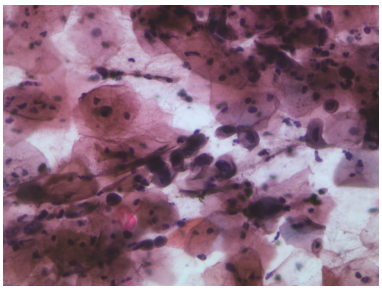
2.1 SIPaKMeD

The datasets used are constructed by cyto-technicians for classification purposes. These technicians use a microsope witha resolution of 0.201 µm / pixel to grab every cell.

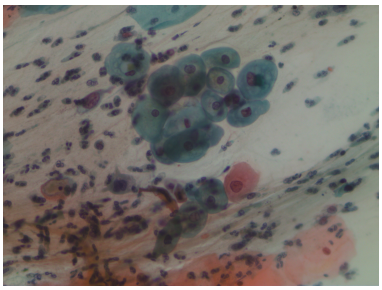
The SIPaKMeD dataset consists of 996 cluser cell images of Pap smear slides. From these there are 4049 indavidual cells that are segmented. These cells are then labeled in 5 categories: (a) Dyskeratotic, (b) Koilocytotic, (c) Metaplastic, (d) Parabasal and (e) Superficial-Intermediate.



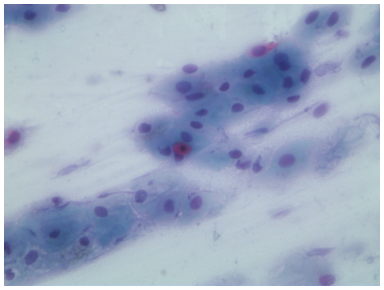
(a) Dysketarotic



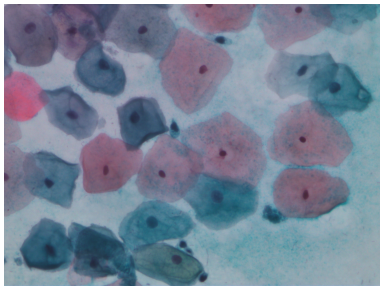
(b) Koilocytotic



(c) Metaplastic



(d) Parabasal



(e) Superficial-Intermediate

Figure 2.1: Cell Types

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

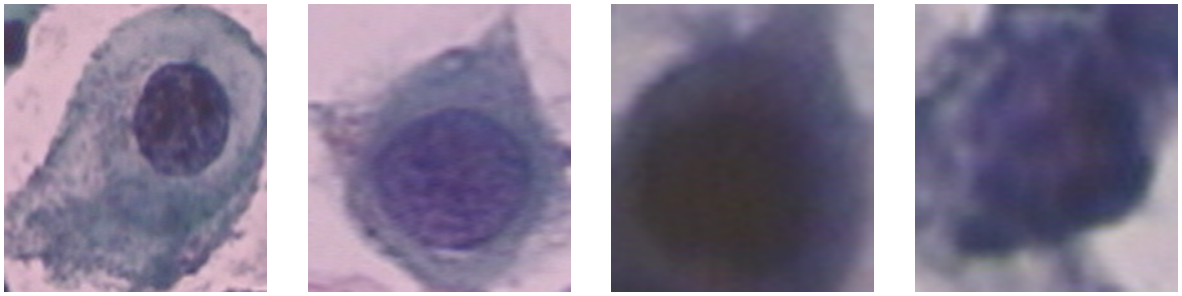
usually enlarged, eccentrically located, hyperchromatic and exhibit irregularity of the nuclear membrane contour.

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]

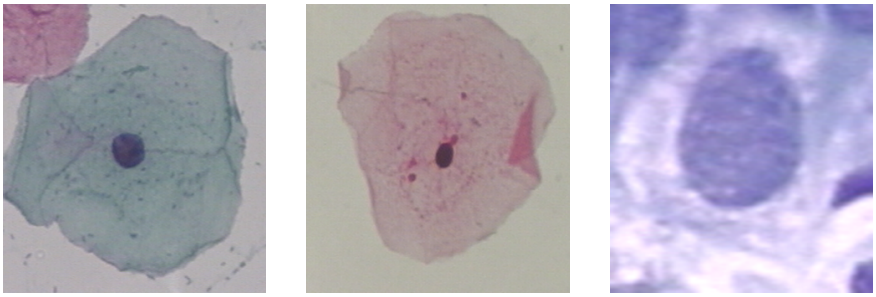
2.2 Herlev

The Herlev dataset is comprised of 917 isolated single cell images. These are distributed unequally between seven classes of cells. Superficial squamous epithelia, intermediate squamous epithelia, columnar epithelial, mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, severe squamous non-keratinizing dysplasia and squamous cell carcinoma in situ intermediate.[3]



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. More details are given in the attached API.

Chapter 3

Background

3.1 General Data Augmentation

Data hungry neural networks make us need data augmentation. Depending on the domain of the data, severl different types of data augmentaion can be used. Very tradiional is fliping and rotating the data. Otheres include random color altertions, adding noise, random zooming, ect.

Current automatic screening scystems that incorperate pap smear have a very similar workflow: cell segmentation, cytoplasm and nuclei segmentation, feature extraction, and cell classification. This is very limited and very sensitive to cell overlap.

Whole slide classification (no segmentation) [4]

Chapter 4

Synthetic Data Generation

There are lots of different manaul augmentations I do to the images...

4.1 Tradidtional Data Augmentation

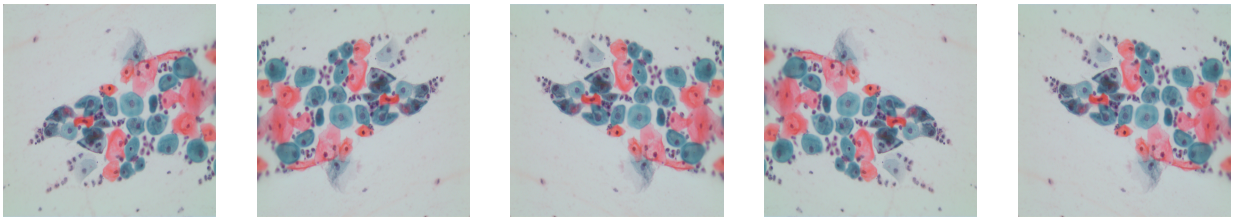


Figure 4.1: Horizantal and vertical flipping

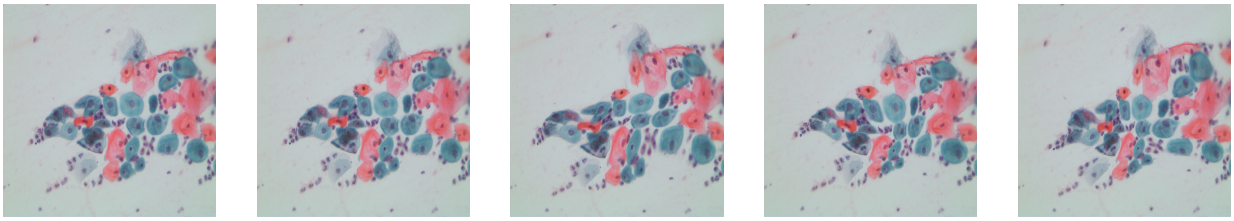


Figure 4.2: Elastic Deformation

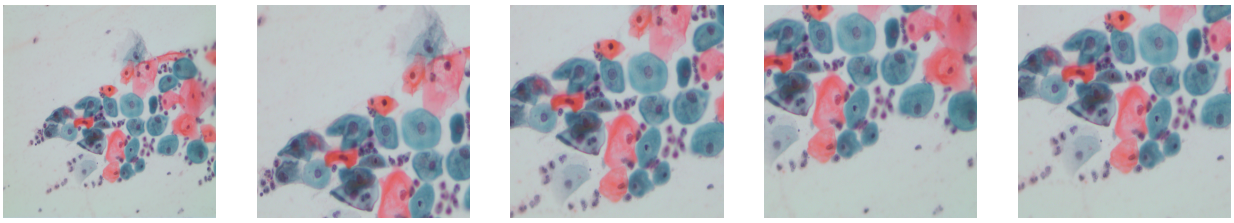


Figure 4.3: Croppin

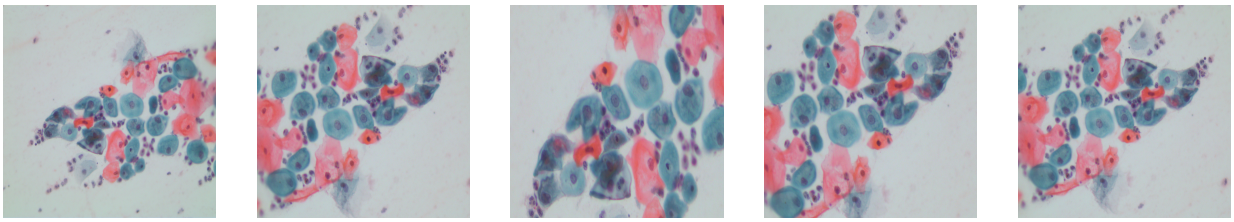


Figure 4.4: Rotation

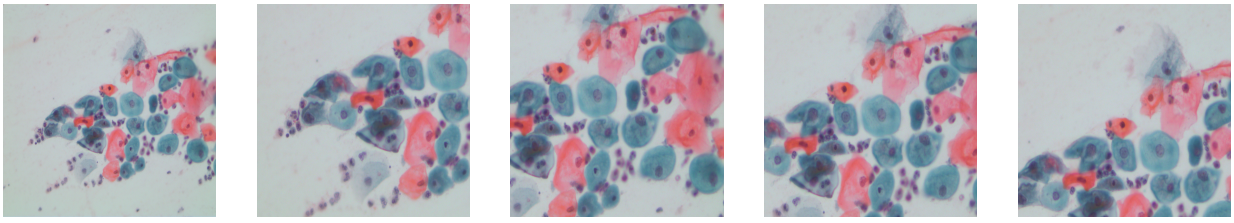


Figure 4.5: Zoom

4.2 Inserting Cells

This is where I randomly crop and segment labeled cells in other images. I place them onto the image. To help it blend in more I take the hard segmentation boundy and dialate the mask. I then do a Guassian filter to blur the opacity so it is not just a hard border on the inserted images.

current challanges:

- different size cells from different samples

- dosnt fit into the context when put randomly
- there are different colors of die that are not consistant accross classes (color correct somehow?)
- placing above other cells

4.3 Conditional Generative Adversarial Networks

Generative adversarial networks, GANs, are generative models that learn a mapping from a random noise vector z to output image y, $G : z \rightarrow y$. A traditional GAN can be broken down into 2 networks, the generator and the discriminator. The generator is G described above. The goal of the discriminator is to take in a generated and real image and classify which is real.

Conditional GANs differ from this learning a mapping from an observed image, x, and a random noise vector z. This gives $G : \{x, z\} \rightarrow y$. Mathematically the objective funcitn of a conditional GAN can be expressed as

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

4.3.1 Pix2Pix

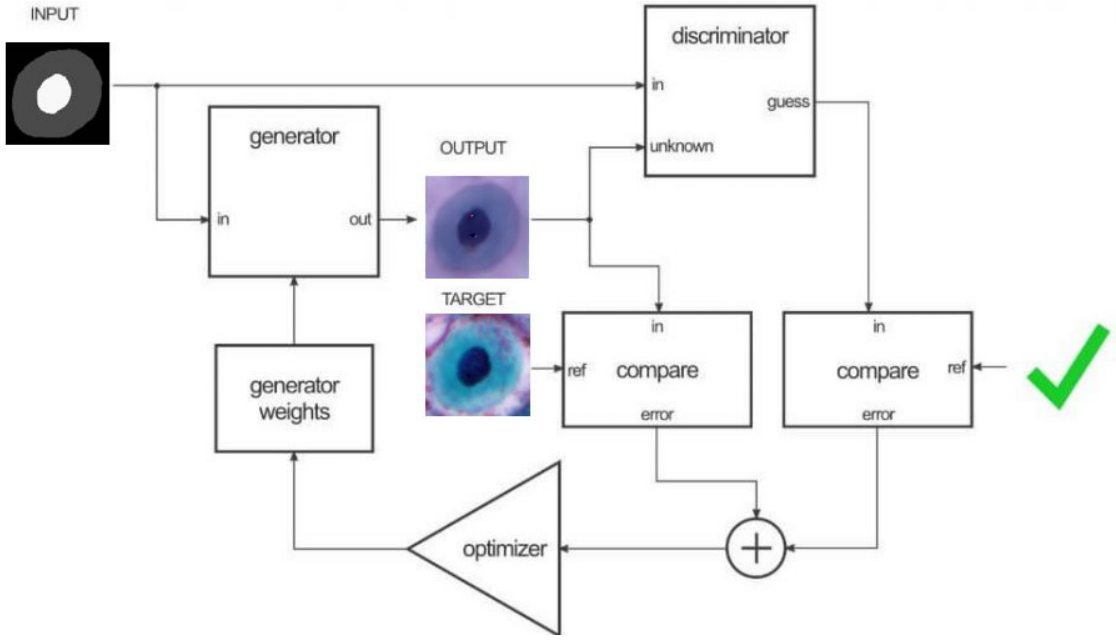
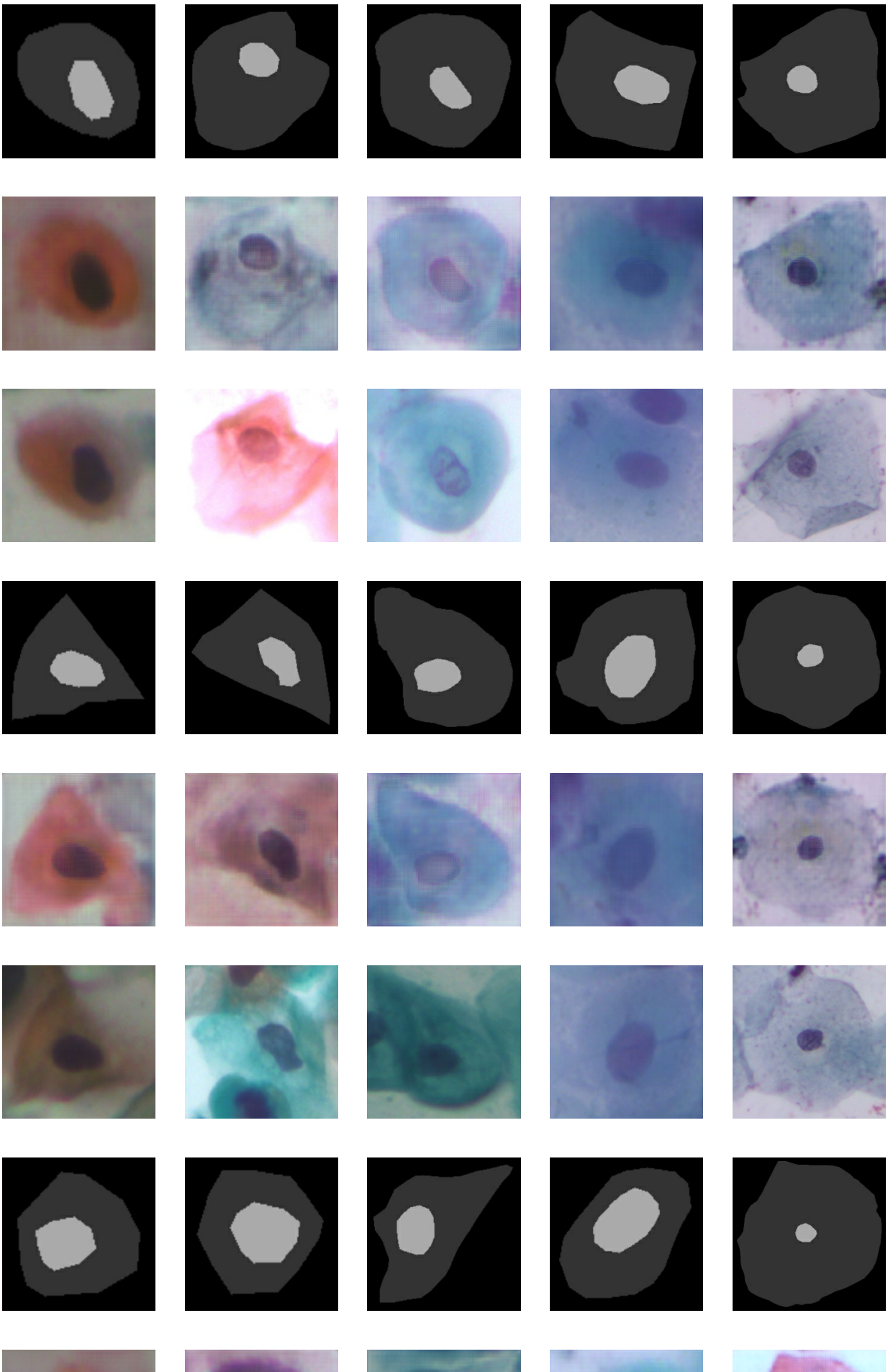
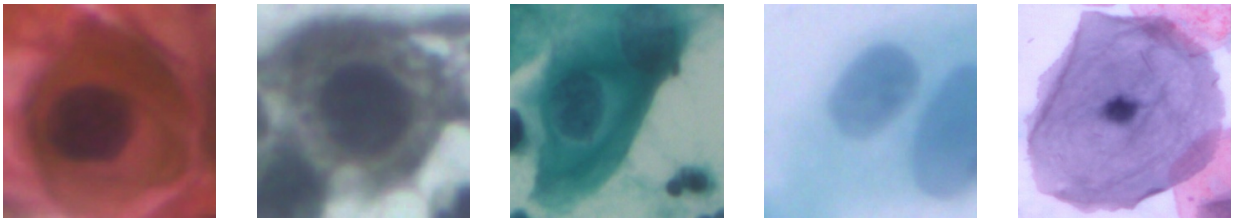


Figure 4.6: Pix2Pix Archetecture [2]





Chapter 5

Segmentation Model

5.1 Convolutional Neural Network

5.2 Unet Segmentaion

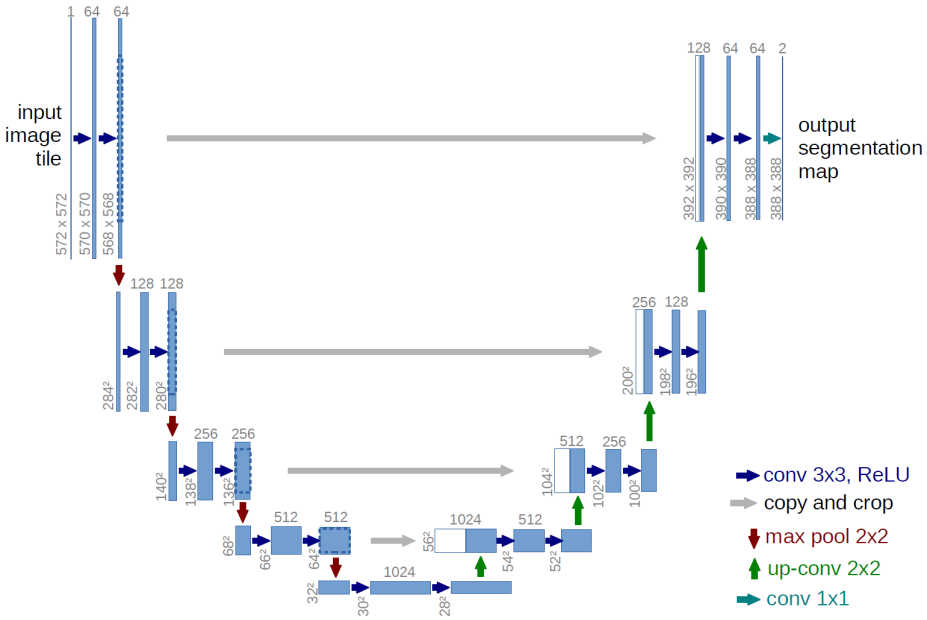


Figure 5.1: Unet Archetecture

Chapter 6

Experiments and Results

Trainging a cnn model on synthetic. giving it real data to predict gives accuracy of .2409. The real one is .2589

Cell 5 part classification: CNN

loss: 0.0116

acc: 0.9953

val loss 0.2423

val acc: 0.9513

Chapter 7

Future Work

References

[1] World health organization. cervical cancer, 2018. [online; accessed 12-december-2018].

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

[3] 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.

[4] Kranthi Kiran GV and G Meghana Reddy. Automatic classification of whole slide pap smear images using cnn with pca based feature interpretation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[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.