02-251: Great Ideas in Computational Biology
Project Essay
Name: Shyam Sai

Breast Cancer: Diagnosis and Prognosis Using Keras and OpenCV

Shyam Sai

May 3rd, 2019

Abstract

Breast cancer is the second most common type of cancer and is one of the leading causes of global mortality. In addition, there is expected to be a 70% increase in breast cancer in the following years. It is for these reasons that building a tool to automatically diagnose breast cancer can be extremely useful. In this project, we build three computational tools in an effort towards fully automizing the diagnosis of breast cancer. We build a neural network using Keras for breast cancer diagnosis using a list of cellular properties from a breast mass biopsy with 76.79% accuracy. We also build a neural network using Keras for breast cancer prognosis using a list of cellular properties from a breast mass biopsy with 89.19% accuracy. Finally, we also build an image processing algorithm using OpenCV to accurately extract features from images of breast mass biopsies. In creating a solution to all three of these computational and scientific problems, we move one step closer towards the full automation of breast cancer diagnosis.

Introduction

Cancer is one of the leading causes of global mortality, accounting annually for more than 14 million deaths and costing more than $1.16 trillion in the United States alone in health care services and lost productivity (Cancer, 2017). Breast Cancer is the second most common form of Cancer, with 252 thousand new diagnoses of invasive cancer, 63 thousand new diagnoses of non-invasive cancer, and 40 thousand deaths annually in the United States. Overall, Breast Cancer mortality rates are on the decline. However, there will still be a 70% increase in the next few years (How Common Is Breast Cancer?, 2017). For this reason, developing a tool to automatically diagnose breast cancer is incredibly useful in today's world.

Background

Tools to diagnose breast cancer have been prevalent for many years. There are four commonly used ways of diagnosing breast cancer. The first, a mammogram, is essentially an X-ray of the breast. If an abnormality or other artifact is found on the mammogram of the breast, this could be indicative of breast cancer. According to a 1993 paper by Fletcher et. al, mammograms are able to correctly diagnose cancer between 68% to 79% of the time (Fletcher et. al, 1993). The second way, an ultrasound of the breast, uses sound waves to image the inside of the body. If lump-like structures are found, this could be indicative of breast cancer. The third way is breast magnetic resonance imaging or MRI. This process involves a magnet and radio waves to image the interior of the breast using a dye to detect abnormalities indicative of breast cancer. Finally, the fourth way of diagnosing breast cancer, and the one most pertinent to this paper, is a breast biopsy. A breast biopsy is the *only* definitive test for diagnosing breast cancer. Using a specialized needle-like device, a core of tissue is extracted from the breast and sent to a laboratory for analysis (Mayo Clinic, 2019). This laboratory analysis is exactly where technology can be used to improve the speed of diagnosis and to reduce human error in diagnosis. Instead of sending the core of tissue to a laboratory for analysis, it is feasible for a computer to process images of this tissue to automate the diagnosis process.

It was exactly this automation that motivated work at the University of Wisconsin Madison in 1991. A group of three researchers from two different academic fields collaborated in an

interdisciplinary research project to automate the process of diagnosing breast cancer using a biopsy. Dr. Olvi Mangasarian and Dr. Nick Street, both of the computer science department, collaborated with Dr. William Wolberg of the Human Oncology department. Together, they created a method to diagnose breast cancer using only a Fine Needle Aspiration, a type of biopsy that is less invasive. Using the Fine Needle Aspiration from the breast mass, the resulting material was stained on a microscope slide to highlight cell nuclei. This slide is then imaged using this microscope (Mangasarian et. al, 1995).

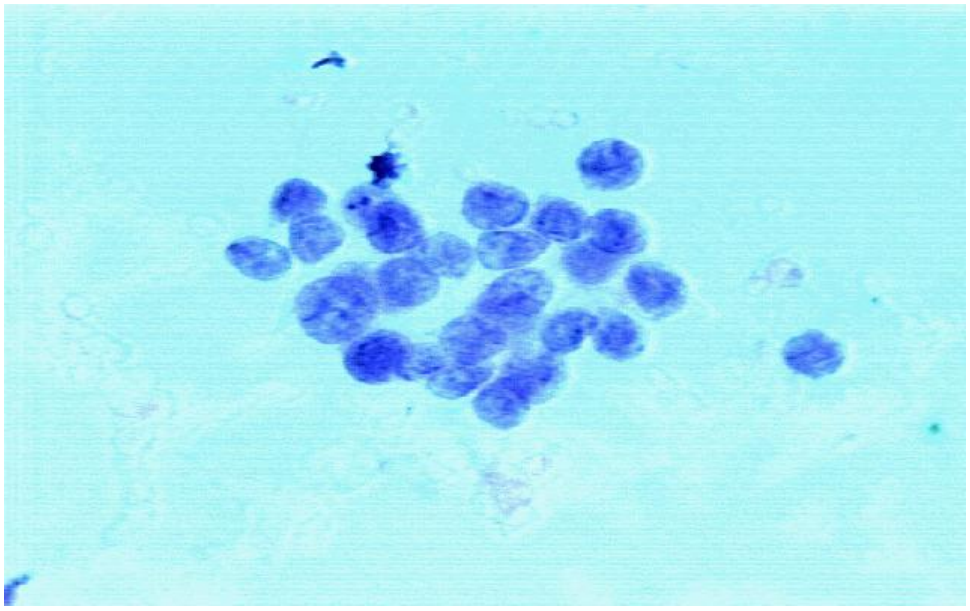

Figure 1: Example Image from Fine Needle Aspiration of Breast Mass

This is where semi-automation takes over in the diagnostic process. From this step, Mangasarian, Street, and Wolberg used Xcyt, an easy-to-use graphical computer program. A user uses Xcyt to trace around individual cell and cell nuclei, a process that takes a maximum of around five minutes. Xcyt then can infer data about an image. The features that Xcyt infers are area, radius, perimeter, symmetry, concavities, fractal dimensions, compactness, smoothness, and texture. Xcyt reports the mean value, extreme value, and standard error of each of these features (Mangasarian et. al, 1995).

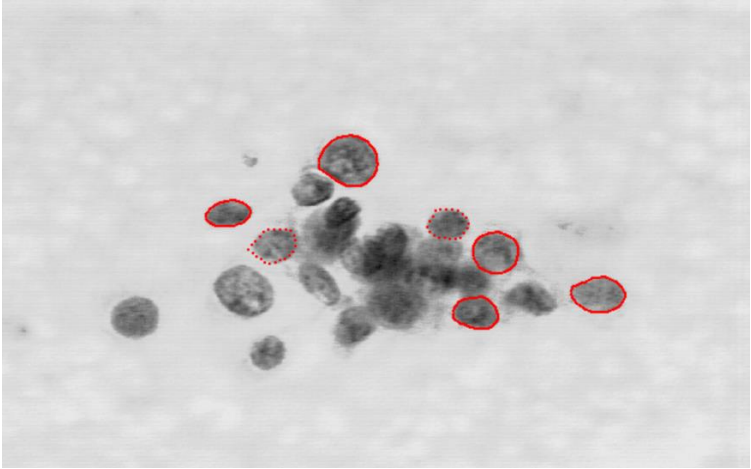

Figure 2: Xcyt used for image analysis

Using these features, along with 569 sample patients, the researchers created a diagnostic system based on linear programming. In solving the following linear program using the MINOS numerical optimization software, where *A* is a matrix of cancerous datapoints, *B* is a matrix of benign datapoints, and *w* is a weight vector, the researchers were able to create a diagnostic tool to classify malignant versus benign (Mangasarian et. al, 1995).

$$\begin{array}{lll} \underset{w,\gamma,y,z}{\text{minimize}} & \dfrac{e^T y}{m} + \dfrac{e^T z}{k} & \\ & Aw + y \geq e\gamma + e & \\ \text{subject to} & Bw - z \leq e\gamma - e & \\ & y, z \geq 0. & \end{array}$$

This linear program generates a separating plane separating benign points from cancerous points if one exists and creates a plane that minimizes the number of violations from a potential separating plane. The researchers found that the best possible separating plane is found when the extreme value of area, the extreme value of smoothness, and the mean value of texture are used as linear separators. They evaluated their resulting classifier using cross-validation to estimate the accuracy of their model at 97.5%. The researchers then implemented their model in the clinical practice of Dr. William Wolberg, where it has classified with 100% correctness since 1993 (Mangasarian et. al, 1995).

In addition, the researchers also created a classifier for the prognosis of breast cancer reoccurrence. Based on 187 follow-up breast cancer patient cases where a patient either has

recurred breast cancer or a patient has come in for a checkup disease-free, the researchers created a time-based analysis to predict the likelihood of breast cancer recurrence. In solving the following linear program using the MINOS numerical optimization software, where *M* is a matrix of datapoints of cancer recurrence in a patient, *N* is a matrix of datapoints of disease-free patients, and *w* is a weight vector, the researchers were able to create a diagnostic tool to classify recurrence versus not (Mangasarian et. al, 1995).

$$\begin{array}{lrcl} \underset{w,\gamma,v,y,z}{\text{minimize}} & \frac{1}{m}e^T y + \frac{1}{k}e^T z + \frac{\delta}{m}e^T v & & \\ & -v \leq & Mw + \gamma e - t & \leq y \\ \text{subject to} & & -Nw - \gamma e + r & \leq z \\ & & v, y, z & \geq 0 \end{array}$$

This problem uses a linear programming-based technique called Recurrence Surface Approximation (RSA). This RSA technique led to a tool that was able to predict the disease-free survival of patients over several months, as shown in Figure 3 (Mangasarian et. al, 1995).

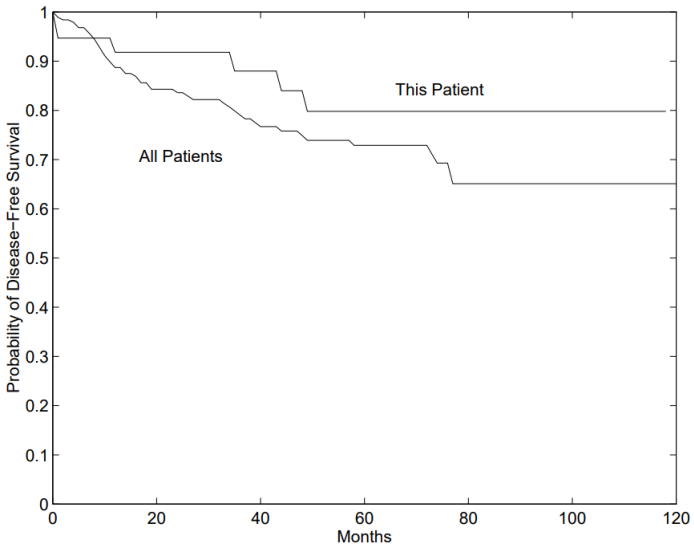

Figure 3: Probability of Disease-Free Survival vs Months

Though their research occurred over 25 years ago, the data and work of the Mangasarian, Street, and Wolberg is still pertinent. Creating a classifier for the diagnosis and prognosis of breast cancer is still important in saving time and saving lives today.

Scientific Problems

Three scientific, computational problems are presented to solve within this project.
Firstly, we aim to solve the problem of Breast Cancer Diagnosis. To achieve this, in contrast with the linear programming work by Mangasarian, Street, and Wolberg, we use a neural network.

Breast Cancer Diagnosis Problem:
Input: A series of data points, each consisting of a classification of benign or malignant, paired with a list of properties (as seen in Figure 4) about cells in an image of a breast mass tissue sample.
Output: A neural network that can classify a patient as benign or cancerous given a list of properties about cells in an image of a breast mass tissue sample, with accuracy better than 50%.

We also aim to be able to classify the prognosis of breast cancer recurrence. Prognosis of breast cancer is if breast cancer has returned to a patient given previous history of breast cancer. This classification will be similar to the Breast Cancer Diagnosis Problem in that we will examine a series of breast mass cell datapoints to determine if cancer recurrence has occurred or not. To achieve this, in contrast with the linear programming work by Mangasarian, Street, and Wolberg, we use a neural network.

Breast Cancer Prognosis Problem:
Input: A series of data points, each consisting of a classification of recurrent or disease-free, paired with a list of properties (as seen in Figure 4) about cells in an image of a breast mass tissue sample.
Output: A neural network that can classify a patient with a given history of breast cancer as recurrent or disease-free given a list of properties about cells in an image of a breast mass tissue sample, with accuracy better than 50%.

Finally, we also aim to remove the human from the process and fully automate the process of breast cancer diagnosis. To fulfill this goal, we must automate the process of extracting a list of

properties about cells in an image of a breast mass tissue sample that we can input into a neural network to diagnose breast cancer. This leads us to our third computational problem of extracting a list of properties of the cells in an image of a breast mass tissue sample.

Feature Analysis of Breast Cancer Mass Images Problem:
Input: An image of stained cells from a Fine Needle Aspiration of a breast mass.
Output: A list of properties about the cells in the image, picked from a list of possible inputs to our neural network (as seen in Figure 4). At a minimum, the radius, area, and perimeter of cells in an image.

If solving of all three of these problems is successful to at least a minimal degree, it is possible to now completely automate the process of diagnosis and prognosis of breast cancer from biopsy to final classification of benign vs. malignant.

Data

In designing a neural network for both the diagnosis and prognosis of breast cancer, it is important to specify what data is accessible for training and testing. The data taken for this project is taken from the Wisconsin Breast Cancer Dataset, from work done in 1993 by Mangasarian, Street, and Wolberg at the University of Wisconsin at Madison. Two datasets from that study are used in order to train the neural networks.

The first dataset from the University of Wisconsin study is that for the diagnosis of breast cancer. The dataset includes data from breast mass biopsies of 569 patients. Each breast mass biopsy is imaged and analyzed using the software Xcyt, leading to a dataset where each datapoint holds 11 properties about the cells in the breast mass tissue sample. These properties are listed in Figure 4.

The second dataset from the University of Wisconsin study is that for the prognosis of breast cancer recurrence. The dataset includes data from breast mass biopsies of 187 patients. Each breast mass biopsy is imaged and analyzed using the software Xcyt, leading to a dataset where

each datapoint holds 14 properties about the cells in the breast mass tissue sample. These properties are listed in Figure 4.

Diagnosis Dataset	Prognosis Dataset
- Benign vs. Malignant - Radius - Texture - Perimeter - Area - Smoothness - Compactness - Concavity - Concave Points - Symmetry - Coastline Approximation	- Recurrent vs. Disease-free - Radius - Texture - Perimeter - Area - Smoothness - Compactness - Concavity - Concave Points - Symmetry - Coastline Approximation - Tumor Size - Lymph Node Status - Time since last known disease-free living

Figure 4: Properties from University of Wisconsin Study

Each property listed in Figure 4 is represented by three distinct values: the mean value over all cells, the extreme value over all cells, and the standard error over all cells. This goes for all properties except “Benign vs. Malignant,” “Recurrent vs. Disease-free,” “Tumor Size,” “Lymph Node Status,” and “Time.”

In addition, for use in developing an image processing algorithm to extract features from the images of the cells from the breast mass biopsies, the Wisconsin Breast Cancer Dataset provides several images of stained imaged cells from a breast mass biopsy, similar to those seen in Figure 1 and 6. These images will be invaluable in creating and testing an image processing algorithm.

Methods and Techniques

To build the neural networks for breast cancer diagnosis and prognosis, Keras was used. Keras is a neural network interface that runs on top of TensorFlow and is written in Python. It is incredibly useful for fast prototyping of neural networks, creating an easy-to-use but highly functional system that can quickly create high-performance neural networks.

Within Keras, there are two models of building neural networks. The first, Sequential, allows for easy and quick building of a layer-by-layer neural network. The second, Functional, is used in more complex purposes, such as use with acyclic graphs and multi-output models. Because this project requires a neural network with a singular, straightforward output, the Sequential model was chosen. This allows us to select layers to be used in our neural network and to simply and easily add them to our model.

The first neural network to be built is to solve the **Breast Cancer Diagnosis Problem**. This neural network would have to accept an input of 10 datapoints (the mean value was used for each property) and would have to output a classification of either benign or cancerous. This neural network had 569 datapoints at its disposal from both benign and cancerous breast masses, with 357 benign datapoints and 212 malignant.

The neural network designed for this dataset took in input with dimensions of 10, as 10 datapoints were being used in each training and test example. The kernel weights were initialized using a random, uniform initializer. The layers within the neural network were simply layers of Dense layers alternated with Dropout layers. Dense layers are densely connected neural network layers, while dropout layers set a random selection of input to 0, essentially randomly removing certain connections from the neural network to prevent overfitting. Finally, the optimizer used was stochastic gradient descent, a randomized algorithm with a batch size of 1 used to minimize the loss by finding the best combinations of weights and bias. The loss function for this neural network was defined as the mean squared error. This neural network trained over 100 epochs.

The second neural network to be built is to solve the **Breast Cancer Prognosis Problem**. This neural network would have to accept an input of 13 datapoints (the mean value was used for each property except tumor size, lymph nodes, and time) and would have to output a classification of either recurrent or disease-free. This neural network had 198 datapoints at its disposal from both recurrent and disease-free breast masses, with 47 recurrent datapoints and 151 disease-free datapoints.

The neural network for prognosis looked very similar to the neural network used for diagnosis, as the input data is very similar. The kernel initializer was also random and uniform, and the optimizer was also stochastic gradient descent. The layers within the neural network were simply layers of Dense layers alternated with Dropout layers, as before. The loss function for this neural network was defined as the mean squared error. This neural network also trained over 100 epochs.

Both neural networks for diagnosis and prognosis look like the one visualized in Figure 5.

Layer (type)	Output Shape
dense_1 (Dense)	(None, 1000)
dropout_1 (Dropout)	(None, 1000)
dense_2 (Dense)	(None, 500)
dropout_2 (Dropout)	(None, 500)
dense_3 (Dense)	(None, 1)

Figure 5: Visualization of Neural Network Structure

The final problem to solve within this project is the **Feature Analysis of Breast Cancer Mass Images Problem**. To solve this problem, the open source library OpenCV was used. OpenCV is a useful image processing tool that can be used within Python to use its computer visions capabilities to isolate and extract features of the cells within images. Images of cells from breast

mass biopsy tissue were used to create and test the image processing algorithm. The image processing algorithm developed was a 5-step process:

1. Convert the color image into a grayscale image.
2. Binarize the image
3. Apply noise removal onto the image
4. Apply a connected components analysis onto the image to isolate components
5. Determine final cell boundaries; calculate radius, area, and perimeter using boundaries.

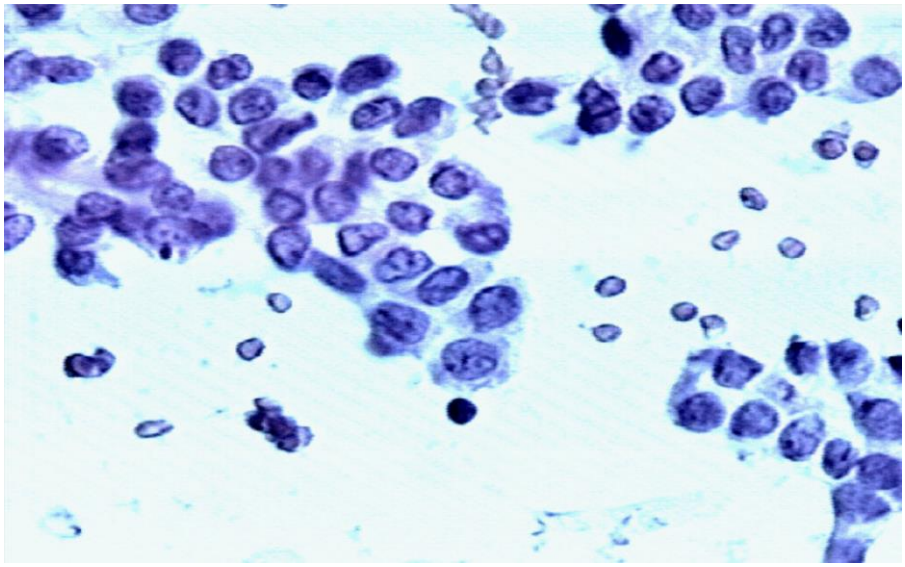

Figure 6: Unprocessed Image of Cells from Breast Biopsy

Because the breast cancer cell images (Figure 6) start out as bright blue images, it is important to first use OpenCV to convert the image into a grayscale image. This involves scaling the brightness of all the points so that the darkest point is skewed towards black and the lightest point is skewed towards white. Converting the image to grayscale facilitates the rest of the image processing algorithm.

Once the image has been converted to grayscale, we use OpenCV to binarize the image. This involves taking a threshold value and converting all pixels above that value to white and converting all pixels below that value to black. Binarization is useful, as it often makes it so an image's background is turned to black while the foreground, the cells, are white, making it easier to isolate important aspects of a picture. This is illustrated in Figure 7.

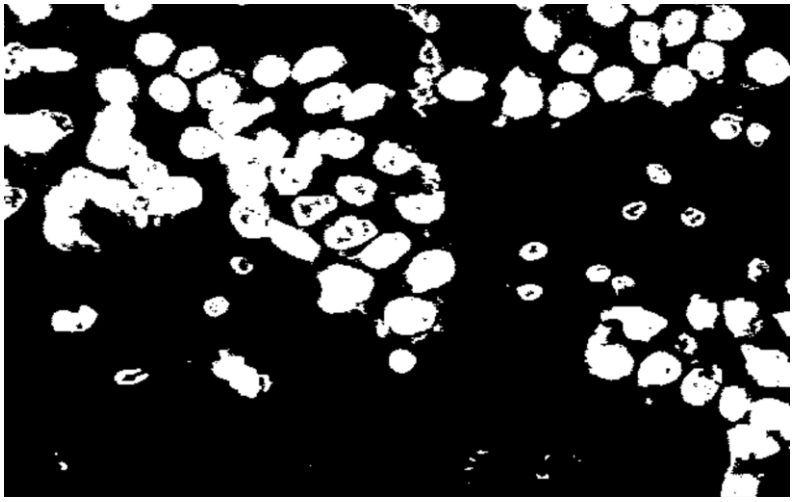

Figure 7: Image after Binarization

One problem with binarization is that it always does not highlight just the important parts of the image. For example, in Figure 7, there are small white artifacts near the bottom of the image that are not cells, and thus should not be highlighted as such. In order to remove small artifacts and further develop the white highlighting of the cell boundaries, we apply OpenCVs noise removal to the image. This process removes small artifacts from an image while further clumping large artifacts, differing them from their background, as seen in Figure 8.

Figure 8: Image with Noise Removal Applied

Once the small artifacts are removed from the image, we can begin to see the outlines of cells within the image. From here, we can implement an OpenCV function called Connected Components. An incredibly simplified explanation is that Connected Components applies an

algorithm that looks at each pixel within the image and compares it to its neighboring pixels. If the pixels match, the group of pixels is treated as a connected component. In applying this algorithm to the entire image, we can extract connected components from the image, as illustrated in Figure 9.

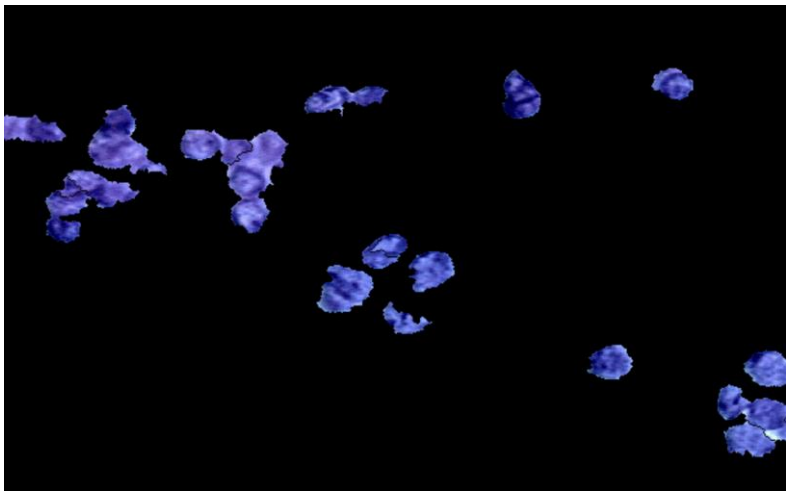

Figure 9: Connected Components within the Image

Notice that the connected components in this image highlight the cells within the image. This is exactly the result we wanted and have now isolated the cells within the image. Now, we can extract the borders of these cells to extract data about the radius, area, and perimeter of these cells, as seen by the red borders in Figure 10.

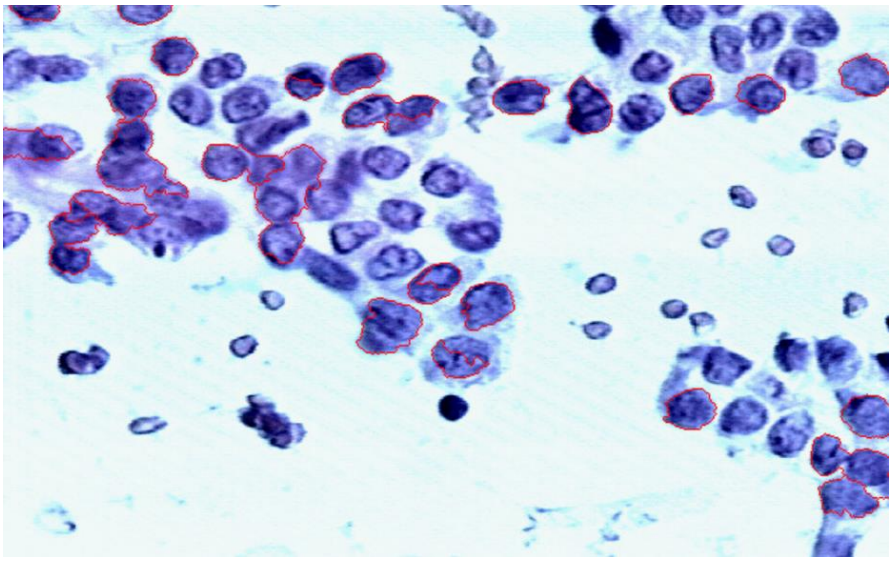

Figure 10: Cell Borders Highlighted on the Image

This algorithm, in a short 5-step process, is successfully able to take an image of cells from breast cancer biopsy tissue and isolate the cells, deriving cell boundaries around many of the cells within the image. These cell boundaries can then be used to successfully calculate features from the image, like the radius, area, and perimeter of the cells.

Results

Many computational problems were attempted within this project. Each had varying degrees of success, and the results of the training and development of these neural networks and image processing algorithm can be determined through calculating each of their accuracies.

Firstly, the neural network to solve the **Breast Cancer Diagnosis Problem**. This neural network was trained for 100 epochs, on 357 benign datapoints and 212 malignant datapoints, with 455 Training datapoints and 114 test datapoints. It finished training with an accuracy of 76.79%. This is significantly better than 50%, the accuracy of a model that chose benign or cancerous randomly. This accuracy is also within the accuracy of mammogram diagnoses in 1993, as it falls between 68% and 79%. This preliminary result leads to the belief that using a neural network to diagnose breast cancer is more than feasible. However, this is still a long call from the 97.5% accuracy of Mangasarian, Street, and Wolberg.

Secondly, the neural network to solve the **Breast Cancer Prognosis Problem**. This neural network was trained for 100 epochs, on 151 disease-free datapoints and 47 recurrent datapoints, with 158 Training datapoints and 40 test datapoints. It finished training with an accuracy of 89.19%. This is significantly better than 50%, the accuracy of a model that chose benign or cancerous randomly. This accuracy is also incredibly close to 100%, the accuracy of a perfect model. This accuracy is surprisingly high, and this result brings hope that a classifier for the prognosis of breast cancer is strongly feasible.

Finally, we can determine the accuracy of the image processing algorithm developed for the **Feature Analysis of Breast Cancer Mass Images Problem**. When the image processing

algorithm is run on the image in Figure 6, the following average cell radius, average cell perimeter, and average cell area are inferred from the image.

Image Processing Algorithm	Mangasarian et. al Dataset Averages
Average Radius: 18.48437 pixels Average Perimeter: 85.927037446 pixels Average Area: 812.625 pixels squared	Average Radius: 14.12729 pixels Average Perimeter: 91.96903 pixels Average Area: 654.889 pixels squared

Each of these properties is displayed next to the average for each property within the Mangasarian, Street, and Wolberg dataset, data that was used in serious scientific work at the highest level. The averages from the algorithm are incredibly comparable with the averages from the University of Wisconsin Madison dataset. The average radius, as calculated by the image processing algorithm, has only a 26.721% difference from the scientific dataset. The average perimeter, as calculated by the image processing algorithm, has only a 6.793% difference from the scientific dataset. The average area, as calculated by the image processing algorithm, has only a 21.497% difference from the scientific dataset. These percent differences are not very large and show that the image processing algorithm developed in this project has real feasibility in extract cellular properties from a breast cancer biopsy image.

These results show that there is real feasibility in completely automizing the breast cancer diagnosis process, in using the image processing algorithm developed to extract properties from a breast cancer biopsy image for use in passing on to a breast cancer diagnosis neural network.

Future Work

Multiple possibilities exist for future work within this project. The first option for future work is producing a percentage of error for how accurate the image processing algorithm is in extracting the radius, area, and perimeter of cells from cell images. This can be calculated by comparing results from the algorithm to results from Xcyt, or by manually calculating these properties for a multitude of images and seeing if the results from the image processing algorithm match. Taking

this future step could further validate the correctness of the image processing algorithm, possibly even validating it for use in practical applications in cellular imaging outside of this project.

The second option for future work is to improve the classifying accuracy of the two neural networks. This can be achieved in a multitude of ways.

The first opportunity to improve the accuracy of the two neural networks for diagnosis and prognosis is through transfer learning. Transfer learning is a technique where a model that is previously trained to classify one task is repurposed to classify another, separate task. If a related predictive problem's model is found, perhaps, a model that also classifies using cellular data, it may be possible to reuse and tune that model for the current problem of breast cancer diagnosis to improve the accuracy of diagnosis.

A second route for improvement of the two neural networks is by using a validation set. In training our neural networks, we only used a training set and a test set. If a validation set was to be specified, this could help us understand if either of the neural networks was overfitting or underfitting. If the validation loss was higher than training loss than we could determine that the neural networks might be overfitting. This would allow us to correct for that, and to improve the accuracy for both classifiers.

Another route for improvement of the two neural networks for diagnosis and prognosis is dependent on the accuracy and correctness of the image processing algorithm developed in this project. If the image processing algorithm is shown to have a high probability of correctness, the image processing algorithm can be used on larger datasets of breast cancer breast mass biopsy cell images, creating more data for the neural networks to train on. A larger set of training data for each neural network will most certainly improve the accuracy of both networks.

Finally, an end goal for this project is to develop a web application where a user can input an image of a breast cancer biopsy tissue sample, and the website would output a breast cancer diagnosis. This would require the website to run the image processing algorithm on the image and pass the resulting feature data to a neural network to classify the image as being of a benign

tumor or malignant. This tool could have potential viable implications in locations and places without access to a specialized doctor that is able to diagnose breast cancer, instead using a computer for the job. This can further help people around the world diagnose and catch breast cancer quicker, potentially saving thousands of lives.

Conclusions

This project successfully created two neural network classifiers to both diagnose breast cancer and to classify the prognosis of breast cancer, with 76.79% and 89.19% accuracy, respectfully. In addition, this project successfully developed an image processing algorithm to estimate the radius, area, and perimeter of cells in a stained image of a breast cancer biopsy, falling well within the averages of published data.

These two accomplishments together point towards fully autonomous diagnosing of breast cancer. The image processing algorithm and diagnosis neural network make it so that if a breast mass biopsy is taken, with a little microscopy work, an accurate diagnosis can be made using no human involvement past the biopsy. And, with the advent of lab robots, it may soon be possible to automate the process of acquiring a Fine Needle Aspiration from a breast mass. The lab robots may also automate the process of staining and imaging the cells. This development would fully automate the diagnosis of breast cancer, making it so that places without specialized doctors can have fully automized clinics for breast cancer diagnosis. This final goal can potentially save thousands of lives in rural and impoverished areas.

This project, proposing two accurate neural network classifiers and an image processing algorithm, is only the first step in the final goal of fully automating the diagnosis of breast cancer. Hopefully, with the algorithms and analysis put forth within this project, further work can be done towards the automation of cancer diagnosis.

Acknowledgments

I'd like to thank Professors Phillip Compeau and Carl Kingsford in putting together this unique and interesting class, and for assigning this incredibly fun final project. I'd also like to thank my TAs Wendy and Hongyu and would like to thank everyone else in the class with me. I took three technical classes this semester, and 02-251 was by far the most interesting and fun, all capped off by this brilliantly designed final project assignment. This class was a real treat to be in.

I'd also like to thank the late Dr. Vandana Chinwalla, the high school teacher that showed me the world of computational biology. She passed away to breast cancer in 2018.

Works Cited

Cancer. (2017). http://www.who.int/mediacentre/factsheets/fs297/en/

Fletcher, Black, Harris, Rimer, Shapiro. Report on the International Workshop on Screening for Breast Cancer. (Periodical Report). (1993, October 25). Cancer Researcher Weekly.

How Common Is Breast Cancer? (2017). https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

Mangasarian, Olvi L., Street, W. Nick, and Wolberg, William H. "Breast Cancer Diagnosis and Prognosis via Linear Programming." Operations Research 43.4 (1995): 570–577. Web.

Mayo Clinic. "Breast Cancer." Mayo Clinic, Mayo Foundation for Medical Education and Research, 10 Jan. 2019, www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.