Invasive Ductal Carcinoma Detection

Using Convolutional Neural Network

Barbara Payne
Bellevue University
DSC680-T302
Summer 2021

Abstract

Breast cancer is one of the most common types of cancer. It is expected within this year of 2021, that there will be 284,200 new cases (*Common Cancer Types*, 2021). There are many ways that breast cancer can be detected. A common way for breast cancer to be detected is done by the patient receiving a mammogram or an MRI. Both of these will take an image of the breast tissue. The images are then examined for any abnormalities. Machine learning can be a very useful tool in this case. More opportunities of detection and prevention come with the use of machine learning.

For my project, I have found a dataset that contains multiple images of breast tissue that contain invasive ductal carcinoma (IDC) and images of normal breast tissue. I plan on using a predictive model that will be able to detect which images contain IDC. This will be done by training the model to learn what these images look like for detection.

Introduction

Invasive Ductal Carcinoma (IDC)

Invasive Ductal Carcinoma (IDC) is the most common type of breast cancer – making up about 80%. The cancer invades the breast tissue by breaking through the milk duct wall. It can eventually spread to the lymph nodes and other areas of the body. Though it can happen in any age, about 2/3 of women aged 55 and older are diagnosed with IDC (*Invasive Ductal Carcinoma (IDC)*, 2021). Breast cancer can often show no symptoms, however there are some symptoms that can happen when you may have IDC. Some of these symptoms are lump in the breast, thickening of breast skin, rash or redness on the breast, swelling, new pain, dimpling around the nipple, discharge, lumps under the arm, or changes in appearance of the breast or the nipple.

The diagnosis of IDC can be done in multiple ways. This can be done by either a biopsy or through scanning. A digital mammography, ultrasound or MRI are just some of the ways the cancer can

be detected through scanning. Treatment for IDC is determined by the size and what stage the cancer is in. It could be any or any combination of the following: lumpectomy, mastectomy, breast reconstruction, radiation, chemotherapy, and hormonal therapy (Brown, 2017).

Detection of breast cancer is crucial. Typically, the early the cancer is found, the better. The earlier it is detected; treatments can be started earlier and the probability of beating the cancer goes up. There are four different ways to detect breast cancer. As mentioned, a biopsy can be performed. This is when tissue or fluids are removed from the breast and to be examined more thoroughly. A breast ultrasound can also be used to detect the cancer. This is done by using a machine that uses sound waves to make detailed pictures on inside the breast. The last two ways are to take images of inside the breast. A diagnostic mammogram is used when there are lumps present inside the breast or if the breast looks abnormal during a normal mammogram. This is a detailed x-ray of the breast. A magnetic resonance imaging (MRI) can also be taken of the breast. This will scan the breast and take an image of inside to link to a computer through magnets (*How is Breast Cancer Diagnosed?,* 2020). The images of the breast can tell a lot and it is important to be able to spot the cancer early on than later.

Image Classification and Deep Learning

Data science has a lot of benefits, especially when it comes to helping people. Deep learning recognizes patterns that can help distinguish between different images. For example, there could be two different images of animals – one a cat and one a dog. Deep learning can tell that they are both animals, but they are different by being trained to know what a cat looks like versus a dog. This can be very beneficial when it comes to health. There are many different types of health screenings, such as a screening for IDC, that can be used by deep learning to detect if a patient has IDC or not.

One-way deep learning is uses image classification is through the convolutional neural networks (CNN). It is the most commonly used method to analyze images and classify them. A benefit to using them is they require very little preprocessing. How it works is by extracting features from an image and

looks through many layers to detect what is being evaluated (Bonner, 2019). This is why it is one the best methods to use when classifying images.

Using deep learning for medical images is not new. There have been many studies to prove that it can be effective and some of what the flaws are. There has been one study where IDC could be detected if it was in an early stage or a later stage. Here they were able to gather images from The Cancer Genome Atlas and trained their model to differentiate between the different stages. They used methods of data mining, normalization, feature selection, training and evaluation (Roy, Kumar, Mittal, Gupta, 2020). Ultimately, they were successful in their work. This just goes to show that machine learning is very beneficial in our lives.

Dataset

https://www.kaggle.com/paultimothymooney/breast-histopathology-images

The dataset that I chose can be found on Kaggle through the above link. As mentioned above, IDC is the most common form of breast cancer. This dataset was used to accurately identify and categorize the breast cancer subtypes. The dataset itself contains over 200,000 patches of over 150 images. There are more negative patches than there are positive at 198,738 to 78,786.

Methods

This dataset only contains images and no actual values. The images are divided into two categories normal (0) and IDC (1). With any data science project, exploratory data analysis (EDA) must be examined. Within this part of the project, I will need to explore the distribution of the data. From the initial look of the dataset, it is determined that there will be more normal images than images that contain IDC. I may want to reduce some of these images in the normal category so not to skew the results.

After EDA, I will want to train and create a CNN model. From some initial research on CNN models, I will be able to use the Tensorflow and Keras packages from previous course studies. Within this portion, I will want to examine the accuracy of the model to ensure that it is producing the most accurate results through the training.

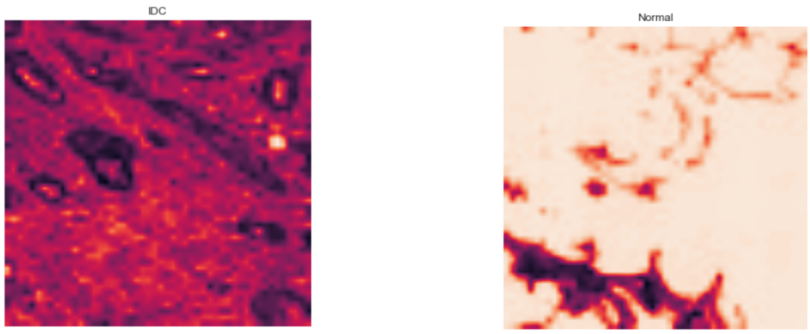
After training, I will test the model by having a set of images that were not used during training to see if it can accurately predict if the image contains IDC or not. This will then show if the project was a success or not.

Results

Exploratory Data Analysis

To start off, it would be important to know what an image containing IDC would look like versus a normal image. From the figure below, you can see that the IDC definitely has more color than the normal tissue.

Figure 1



Originally, the dataset provided the images by breaking down the images by the patient's images and what part of the images contained IDC and what were considered normal healthy tissue. I decided to create two folders for these images. One that would be used for training and one that would be used for testing. This way, I would not use the same images that would go into testing. I pulled the

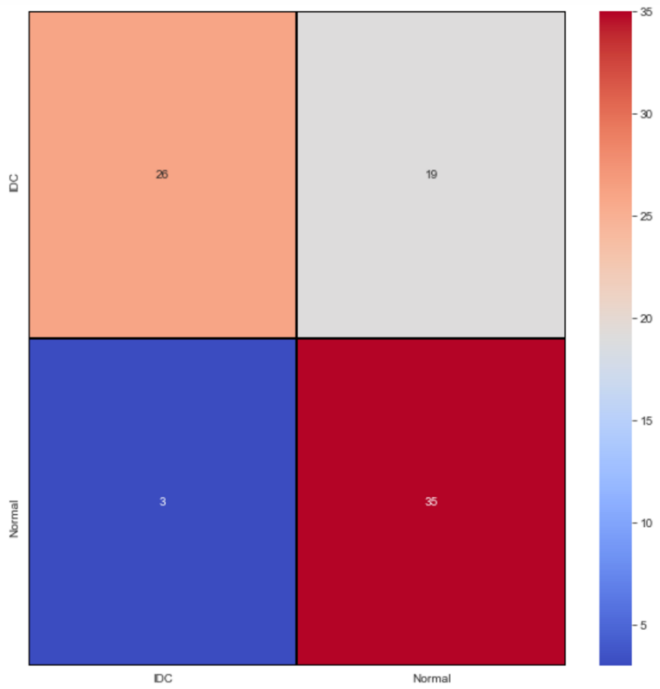
images at random across different patients. I had to take into account overfitting and had to think of a way to not include bias in the results.

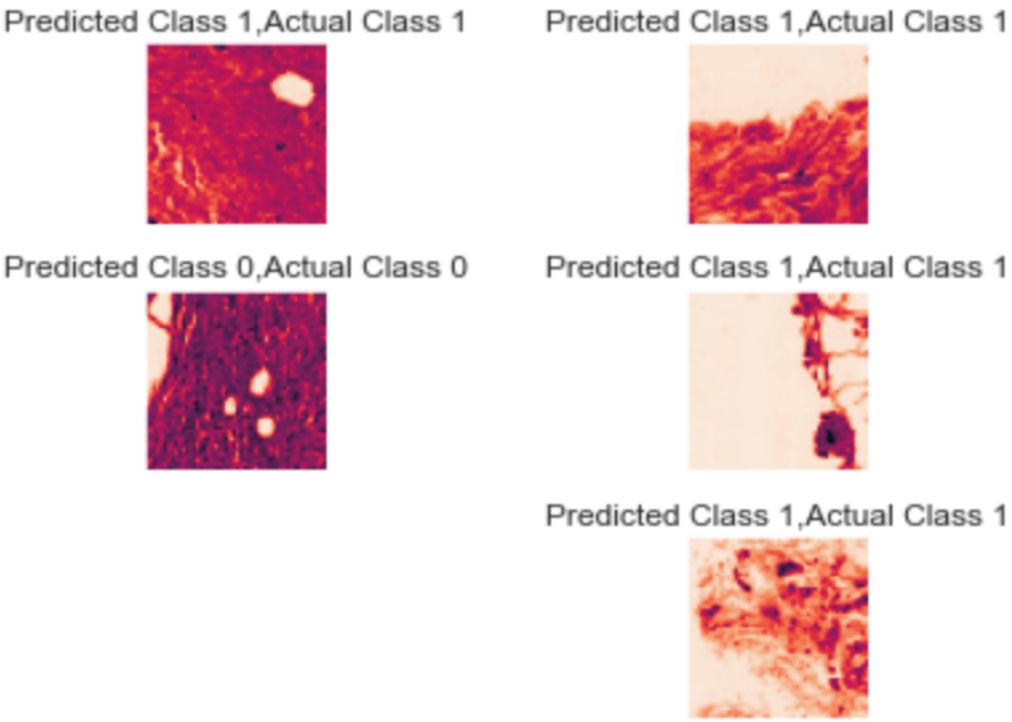
Model Creation

To start off, I thought of doing some data augmentation to help reduce some of the bias in the original data. This will also help with the quality of images and the dataset size. Data augmentation can also assist with overfitting and enhance the generalization of prediction in the model. Part of the CNN process is flattening of the image. The feature extraction is converted into a one-dimensional feature vector for its classifiers. Another method used in the creation of the model was implementing a learning rate variation. This is to determine how much the model should change based on the estimated error. I decided to go with a patience of two but kept the default learning rate of 0.01.

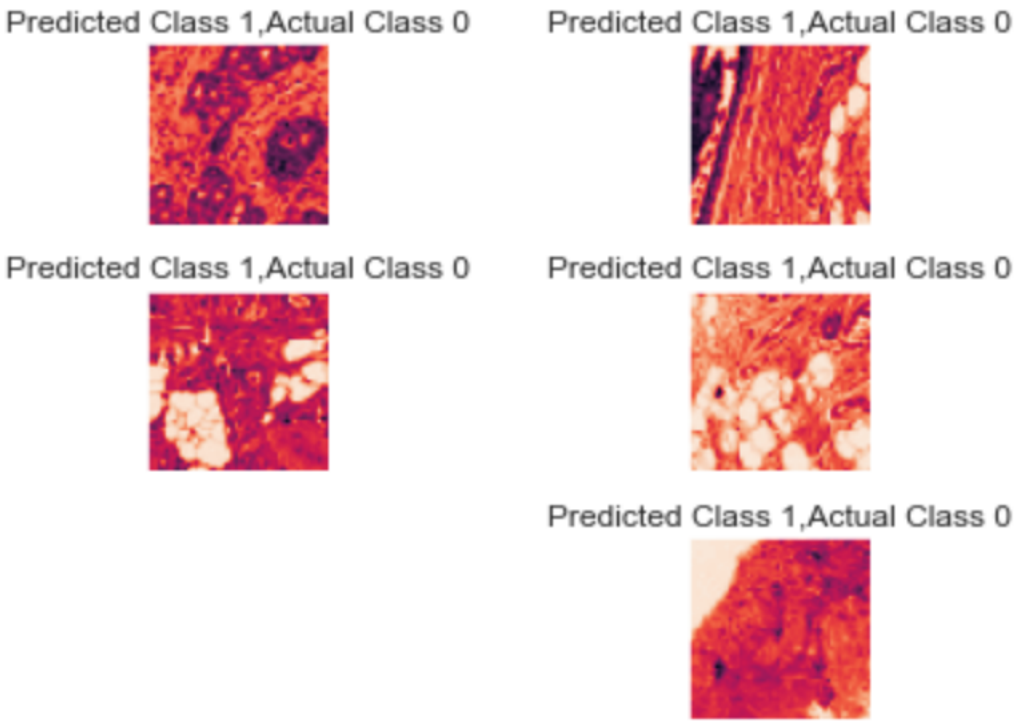
Initial Results

From my initial results, I had an accuracy of 73.49%. This was a little lower than what I was hoping for. The training loss was .5769, which was pretty high. At this point, I may need to examine the data augmentation that I am using and make some adjustments. I may be zooming too much into the image itself as some of the images are very similar in color in the classification of normal and IDC. I had done a confusion matrix to help identify how many true positives, true negatives, false positives, and false negatives.





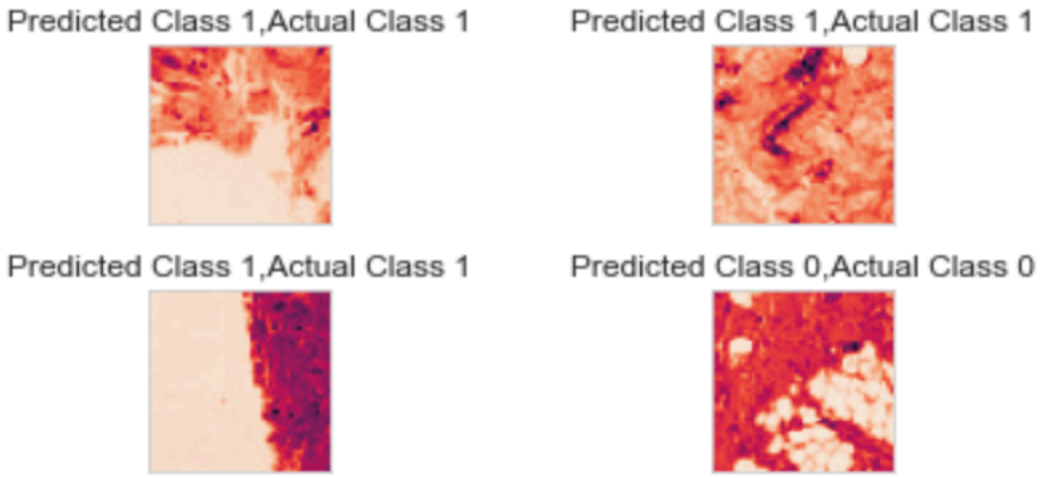
The above image shows some of the accurately predicted images. Below are a sample of the images that were inaccurately predicted.



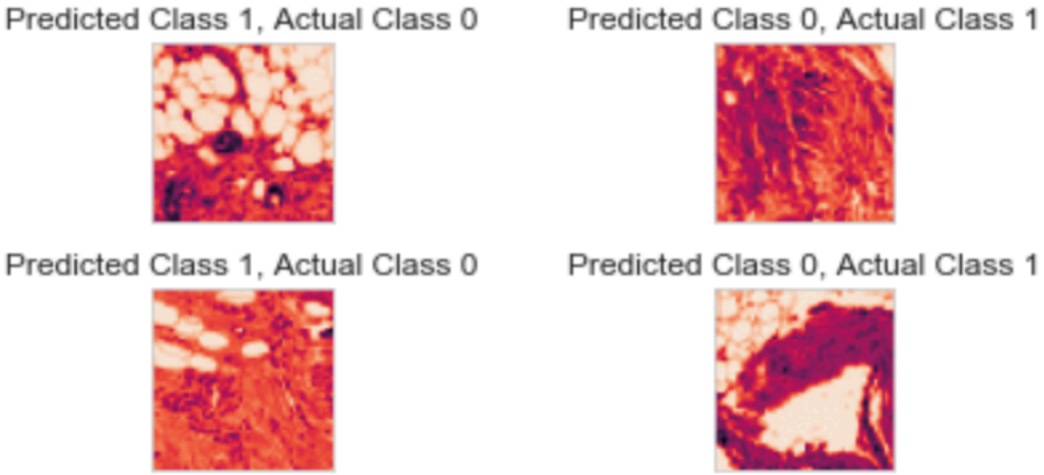
Class 1 is considered “normal” where 0 is considered “IDC”. Each of these images are heavy in the red/purple colors. One would think that these could be IDC, but they are actually normal. This is why I believe I may need to reevaluate how much I am zoomed into the image and look at a broader image. My only concern would be how much this would take to process.

Final Results

I adjusted how many images were being examined, but I did not adjust the zooming. My accuracy did improve, and the overall accuracy became 86.82%. However, my loss also increased, though slightly. The final accuracy was 60%. I had high accuracy and high loss, meaning that when the errors were present, they were big. Below is now the sample of accurately predicted images.



From this small sample, it looks like that it can predict a healthy/normal breast tissue image versus one that contains IDC. Below are the inaccurate prediction sample. It may be that I am zooming in too much, even though it is at 10%. Some of the normal/healthy can look similar to an image that contains IDC.

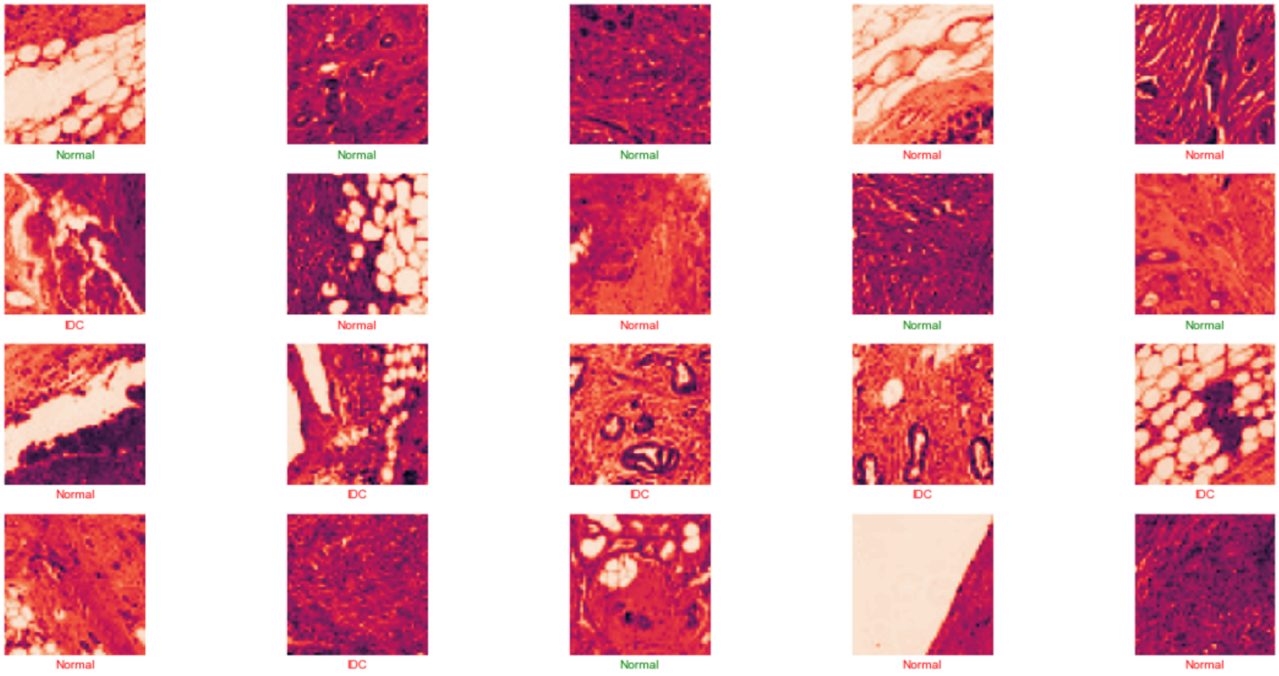


Conclusion

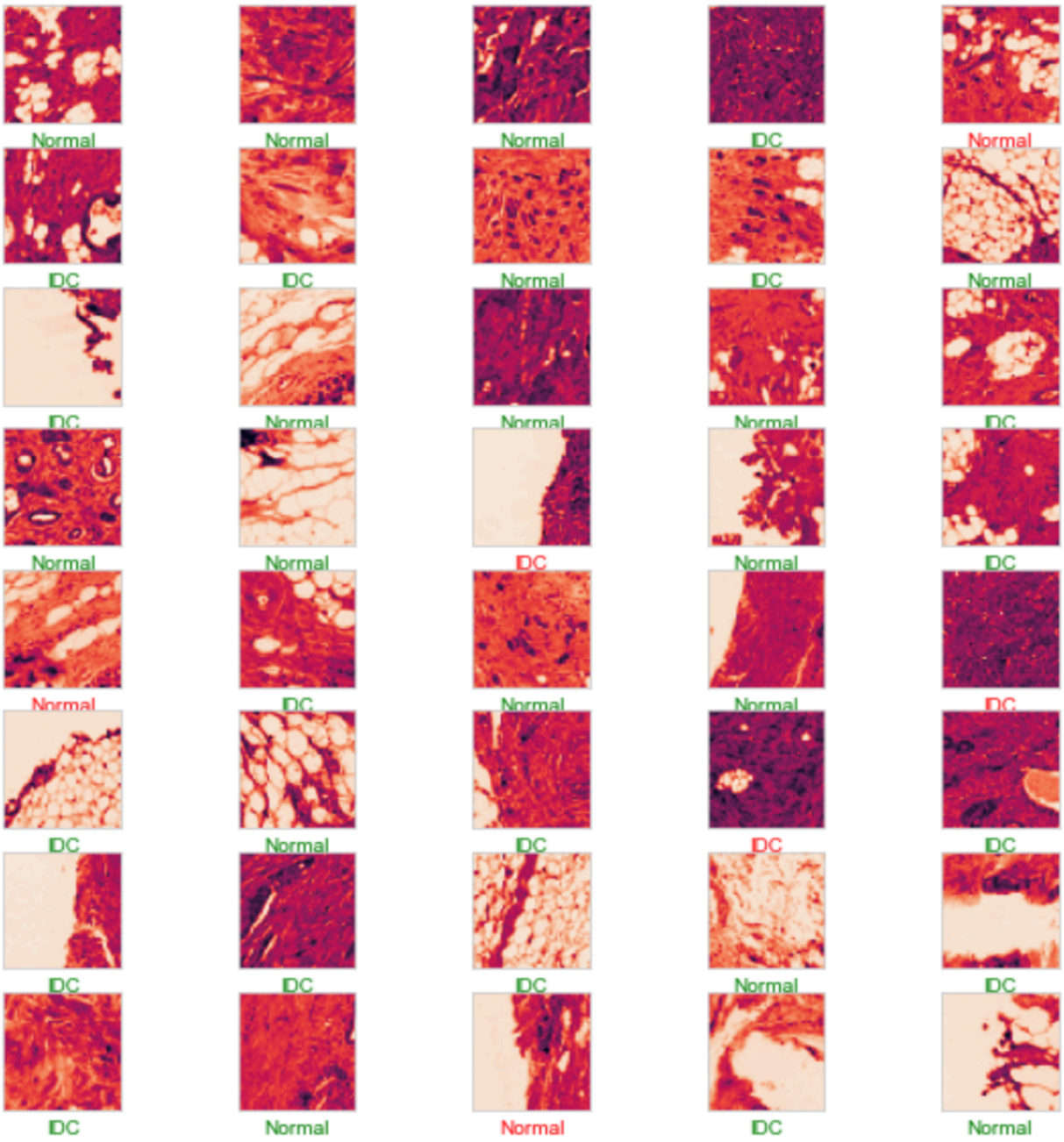
This has been very eye opening and a great learning opportunity on image detection and prediction. This project also proves how important it is to have extremely accurate results. Though my accuracy improved from the initial results, I still believe that can be better. More tweaking needs to be done to the overall project. I need to take a look more into my dataset and the data augmentation that I

tried. I believe some adjustments can be made there to help improve the accuracy and lower the loss of the model.

Below are 20 images from my initial data that show how many it predicted correctly versus wrong. The text that is red was inaccurately predicted, where the green is accurately predicted. Out of these 20 images, only 6 were accurately predicted. Again, I truly believe this has to do more with the overall color of the image and how much I have zoomed in.



Now, here is 40 images that show the accurate versus inaccurate results of the model. I believe I zoomed in too much when training and that is throwing off the end results and why the loss of the model is so great.


Normal
Normal
Normal
IDC
Normal
IDC
IDC
Normal
IDC
Normal
IDC
Normal
Normal
Normal
IDC
Normal
Normal
IDC
Normal
Normal
IDC
Normal
IDC
Normal
Normal
IDC
IDC
Normal
IDC
IDC
IDC
IDC
IDC
Normal
IDC
IDC
IDC
Normal
IDC
Normal
Normal
IDC
Normal

References

Bonner, A. (2019, June 1). *The Complete Beginner's Guide to Deep Learning: Convolutional Neural Networks*. Medium. https://towardsdatascience.com/wtf-is-image-classification-8e78a8235acb.

Brown, K. (2017, November 3). *Invasive Ductal Carcinoma (IDC) Breast Cancer: Johns Hopkins Breast Center*. Invasive Ductal Carcinoma (IDC). https://www.hopkinsmedicine.org/breast_center/breast_cancers_other_conditions/invasive_ductal_carcinoma.html.

Centers for Disease Control and Prevention. (2020, September 14). *How Is Breast Cancer Diagnosed?* Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm.

Common Cancer Types. National Cancer Institute. (2021, April 22). https://www.cancer.gov/types/common-cancers.

Invasive Ductal Carcinoma: Diagnosis, Treatment, and More. Invasive Ductal Carcinoma (IDC). (2020, January 21). https://www.breastcancer.org/symptoms/types/idc.

Roy, S., Kumar, R., Mittal, V., & Gupta, D. (2020, March 5). *Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning*. Nature News. https://www.nature.com/articles/s41598-020-60740-w.

Ten Questions

1. What made you choose this topic?
2. Why did you choose the methods you chose?
3. What were you hoping to gain from this project?
4. What went wrong in the project?
5. What went right in the project?
6. Is there anything you would have done differently?
7. What were your initial predictions?
8. What were challenges that you faced?
9. How did you overcome those challenges?
10. What you know now, what would you tell your past self when starting the project?