**Invasive Ductal Carcinoma Detection Model**

**Business Understanding**

Every year in the United States, there are around 240,000 cases of breast cancer in women and around 2,100 cases in men with about 42,000 women dying and 500 men dying.[[1]](#footnote-0) Yearly, the numbers are rising and less people are getting tested due to costs, time commitment, invasiveness, and lack of education on this subject. Currently, the most common methods for detection include mammography and a magnetic resonance image (MRI)[[2]](#footnote-1) with the most common type of breast cancer being Invasive Ductal Carcinoma (IDC).[[3]](#footnote-2) It is recommended that women between the ages of 50-74 undergo these tests every 2 years.[[4]](#footnote-3) A study showed that the accuracy of a mammogram depends on breast density and age.[[5]](#footnote-4) Along with the accuracy, there is also the high risk of radiation that comes with a mammography.

Through our model, we are classifying if a person has breast cancer or not through a histological image. A histological image is where a sample of a person’s tissue is taken and observed under a microscope. Our model allows for the histological image to be passed through our model, instead of having a professional go through it. Once the image is processed through our model, it also highlights the part where the potential tumor is in the breast. Hence, this is a stepping stone towards diving more in depth of breast cancer and having early preventative measures. This could be integrated in men’s annual health check-up appointments as well. This is just a time efficient and cost efficient way of checking and it also allows the accuracy to increase.

Through our model we could help catch breast cancer at earlier stages and more proactively in an efficient manner with less billable hours to a doctor. From a business perspective, this model would be profitable for the healthcare industry. Hence, we would target hospitals that focus on cancer.

**Data Understanding**

The industry our project belongs to is the healthcare industry. We received the data from GitHub as breast histopathology images.[[6]](#footnote-5) The dataset, consisting of 279 whole mount slide images of breast cancer specimens, was processed and cleaned. Image patches of size 50 x 50 were extracted, resulting in 277,524 patches. Each patch was labeled based on the presence (78,786 patches) or absence (198,738 patches) of the most common form of breast cancer, IDC.

The dataset was uploaded on May 16, 2019 and has not been updated since then. There are many potential biases with this dataset, including data accuracy, bias based on gender, ethical considerations, and outdated data. Despite this, we are proceeding to use this data due to the fact that there was limited data availability for this along with the fact that the dataset had already gone through previous validation. This made it easier for us to understand and narrow down what we were looking for in our model and to spend more time improving our model.

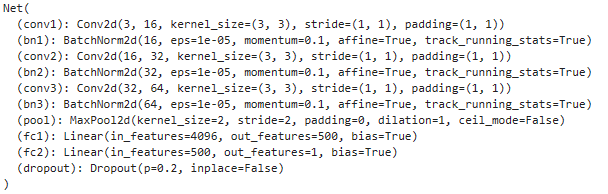
**Data Preparation**

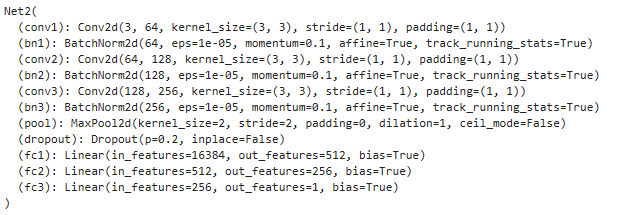
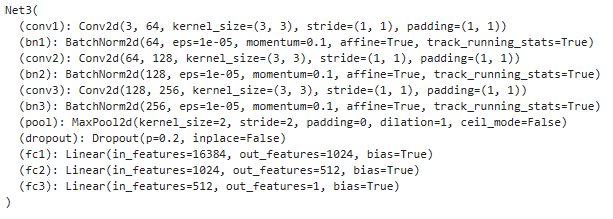
Recognizing the vast dataset, we initiated with a curated subset of 170 patients, ensuring computational efficiency while still encompassing around 180,000 images from the original 279 patients. The initial biopsy image collection underwent a transformation, resizing from 50x50 to a standardized 64x64 for optimal deep learning computation.

In employing a meticulous validation strategy, we partitioned the dataset into a 70% training set, a 15% validation set, and a 15% test set. The preprocessing pipeline, encompassing resizing, normalization, and tensor conversion, assumes paramount significance. Firstly, resizing guarantees uniformity, a prerequisite for neural networks to consistently process data. Secondly, transforming images into tensors and normalizing pixel values optimizes the data format, enhancing the model's learning efficiency

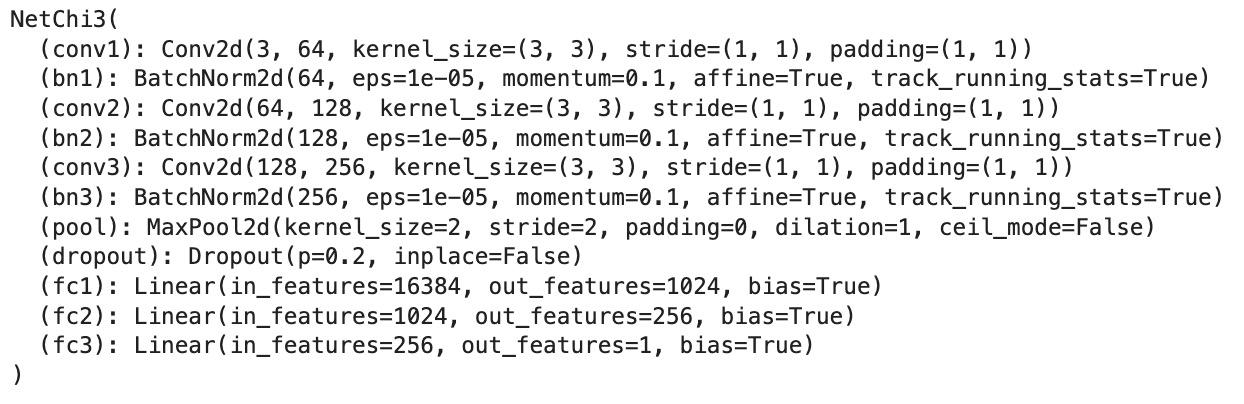
Far from routine, these preprocessing steps are pivotal. They are instrumental in ensuring the effectiveness of our neural network training, leading to heightened model accuracy and reliability. Ultimately, this precision translates into superior performance for our machine learning systems, harmonizing seamlessly with our overarching goal of cultivating robust and efficient solutions for our projects.

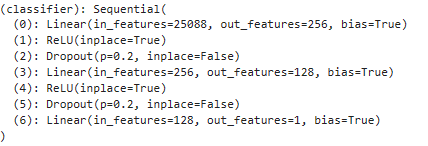
**Modeling**

Based upon the nature of the obtained data, we chose the ***Convolutional Neural* *Network* *(CNN)*** to achieve the extraction of important features from visual images and the accurate detection of the presence of IDC. To find the most effective model, we decide to incorporate and fine tune the following two options: 1) a self-constructed CNN 2) a VGG16 CNN with fully-connected layers adjusted. As shown below, 5 architectures are employed in the later investigation: Net, Net2, Net3, Net3Chi, and adjusted VGG16, as the result of fine tuning.

***Net*** has three convolutional layers, each followed by a batch normalization layer, a max pooling layer, and a Relu activation layer. The output then has a shape with 64 filters and 8x8 features. The flatten data is then passed through two linear layers (***64x8x8 -> 4096 -> 500 -> 1***). 

***Net2****,* ***Net3 and NetChi3*** are developed from Net’s outline. All of them transform images to a shape with 256 filters and 8x8 features.

Moreover, ***Net2*** has three linear layers with feature size progressing from **256x8x8 to 512 to 256 to 1**.The linear layers progress the feature size from **256x8x8 to 1024, to 512, then to 1.** 

***NetChi3*** is similar to Net3, but the second and third fully connected layers convert the feature size from 1024 to 256 then to 1.

**Adjust *VGG16*** inherits the default VGG16 architecture beyond the linear layers, while the features of linear layers are reconstructed to be 25088 to 256, 256 to 128, and 128 to 1.

As the architecture differs in the content of complexity from each other, we can track and evaluate the performance of each architecture with other hyperparameters held constant during the training stage and pick the most desirable architecture for further fine tuning. Net has the simplest architecture, hence used in building the baseline model, and the other 4 architectures are more complex with the potential to capture the data complexity better but also with the inherent risk of overfitting, which may require additional computation and fine tuning steps.

**Implementation**

Via trying out different values for certain hyperparameters in the training process to fine tune the baseline model for gaining better performance on the validation dataset, We built 11 models in sequence. As our investigation interest is a binary classification problem, All models use BCEWithLogitsLoss as the loss function and Adam as the optimizer. The main hyperparameters of our models are listed as follows:

| **Model** | **Batch**  **Size** | **Pos\_**  **weight\*** | **Learning**  **Rate** | **Architecture** | **Epochs** | **Test**  **Accuracy** | **Test**  **F1 Score** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Baseline\_nn** | 512 | 4.0 | 0.007 | Net | 50 | 0.847781 | 0.786033 |
| **Model\_vgg** | 128 | 4.0 | 0.007 | vgg16 | 25 | 0.804534 | 0.737811 |
| **Model2** | 512 | 4.0 | 0.008 | Net2 | 30 | 0.829452 | 0.769279 |
| **Model3** | 512 | 4.0 | 0.008 | Net3 | 35 | 0.833637 | 0.773229 |
| **Model4** | 128 | 4.0 | 0.008 | Net3 | 35 | 0.842240 | 0.77432 |
| **Model5** | 128 | 4.0 | 0.003 | Net3 | 50 | 0.847510 | 0.789876 |
| **Model6** | 128 | 4.0 | 0.002 | Net | 50 | 0.856927 | 0.798801 |
| **Model7** | 128 | 4.0 | 0.002 | Net2 | 40 | 0.865065 | 0.803743 |
| **Model8** | 128 | 3.0 | 0.002 | Net2 | 40 | 0.861810 | 0.802481 |
| **Chi3** | 512 | 4.0 | 0.007 | NetChi3 | 50 | 0.832435 | 0.784423 |
| **Chi4** | 512 | 4.0 | 0.003 | NetChi3 | 30 | 0.817466 | 0.773816 |

\*Pos\_weight: optional hyperparameter of BCEWithLogitsLoss that places weight on positive classification, helpful to mitigate the imbalance of the number of data in each category.

**Step 1: Build Baseline Model: baseline\_nn**

The choice of using Net as a baseline model is driven by its balanced structure, which is complex enough to capture hierarchical features from image data, yet not too deep to require excessive computational resources, fitting for projects with moderate-sized datasets and computational constraints. The performance of the baseline model will also be a comparable metric for hyperparameter tuning.

**Step 2: Utilizing the Pre-Trained VGG16 Model**

To enhance the performance of important feature extraction from the visual images, we employed a pre-trained VGG16 model. This approach leverages the robust feature representations learned from diverse datasets in VGG16. However, due to the great complexity of the VGG model and our computation hardware constraint, we decided not to further fine tune the VGG model in consideration of time and cost efficiency.

**Step 3: Fine-Tuning Models**

The adjustable hyperparameters mainly concentrate in three parts: the model architecture, the optimizer, the loss function, and the batch size of the data. Model architecture directly influences the complexity of the model, which plays a great role in reaching a balance between unfitting and overfitting. The learning rate applied to the optimizer is the most relevant hyperparameter that could impact the model training and the validation performance. Furthermore, to fix the imbalance of class distribution, we also raise the training weight on the positive cases in the loss function. Last but not the least, the effect of the batch size of the data can not be neglected, as the larger the batch size is, the faster the model is trained, while the trade-off is a slower convergence. We changed the value of ONE hyperparameter at a time to investigate the direction and the magnitude of the improvement of the evaluation metrics in order to refine the model in the next model design.

The main challenge of the implementation is the computation constraint. Due to the deficit of computational resources and the considerable size of the raw data, the training time of any of the models is over 90 minutes with each training epoch taking 5 minutes on average. In addition to the limited interpretability of CNN, therefore, any single change to the hyperparameters needs to be very cautious and careful. Moreover, for faster computation, we cannot incorporate the data of all 279 patients, so the model we develop can be potentially biased, though we’ve deployed some generalization techniques like dropout. We believe that the model can be further improved with greater resources on hand in the future.

In summary, each step of our modeling strategy was guided by the goals of improving the model's predictive accuracy for IDC in breast cancer specimens and ensuring its relevance to our specific dataset. The objective was to develop a model that not only exhibits high accuracy and low loss but also effectively generalizes to new, unseen data, contributing significantly to the early detection and diagnosis of breast cancer.

**Evaluation**

When training our neural network, we prioritize minimizing validation loss to ensure the model's generalizability and prevent overfitting. After achieving a satisfactory level of validation loss, we evaluate the model on a separate test dataset. This crucial step assesses real-world performance using two key metrics: accuracy and F1-score. These metrics provide a balanced understanding of precision and recall, guiding our efforts to fine tune the model for optimal practical performance.

Model7 stands out as the top-performing model, showcasing a test accuracy of 0.865065 and a test F1 score of 0.803743. Impressively, it achieves a substantial lift of 16.96% from the null model, where accuracy is 0.6954 when all predictions are set to 0. This surpasses the baseline model by 1.73% in accuracy and 0.018 in F1 score.

| **Model** | **Batch**  **Size** | **Pos\_**  **weight\*** | **Learning**  **Rate** | **Network\*** | **Epochs** | **Test**  **Accuracy** | **Test**  **F1 Score\*** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model6** | 128 | 4.0 | 0.002 | Net | 50 | **0.856927** | **0.798801** |
| **Model7** | 128 | 4.0 | 0.002 | Net2 | 40 | **0.865065** | **0.803743** |
| **Model8** | 128 | 3.0 | 0.002 | Net2 | 40 | **0.861810** | **0.802481** |

\*F1 Score: a metric ranging from 0-1. Higher the F1 score, higher both the recall and precision.

**Deployment**

We plan to go about the deployment of our IDC detection model in phases. Phase 1 will consist of the development of a User Interface Development, Phase 2 will consist of an Emphasis on Model Interpretability, Phase 3 will consist of Regulatory Compliance, Phase 4 will consist of Marketing, and lastly, Phase 5 will consist of Implementation and/or Integration in hospitals.

Phase 1 is the User Interface Development. We plan to develop an interface in order for healthcare professionals to interact with our model. This interface will only be accessible to our clients with different portals for each hospital and to ensure maximum data security, each professional will have their own login information and accessibility depending on their role.

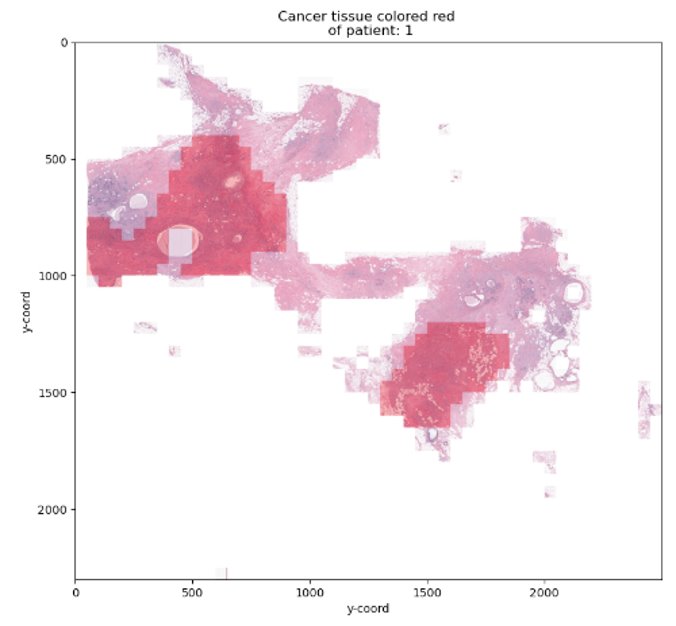
Phase 2 is the Model Interpretability. We want to enhance the interpretability of the predictions that our model will give. This is in order to ensure that a healthcare professional is able to understand what the model has done and how it has gone about analyzing the image. This allows for better decision making as well.

Phase 3 is Regulatory Compliance with Data Privacy and Ethical Considerations. To avoid any data privacy issues, we plan to encrypt patient information through the use of a Patient ID. The model will only have access to the Patient ID while the healthcare professional will have access to both the Patient ID and patient name in order to match them.

Phase 4 is Marketing. Our model will go through the pilot phase for the first month in order to market itself to our clients. It will be a subscription system where the first month is free in order for hospitals to get a chance to try it out and see if they are interested in it. Afterwards, it will be a yearly subscription to access the interface and to use the model. We will also market through outreach: patient advocacy groups, talking to healthcare professionals, attending cancer societies.

Phase 5 is Implementation and/or Integration. Implementation would include the pilot system and then through that, gaining feedback from healthcare professionals and improving our model based on their feedback. We would also gain the trust of the professionals in this manner and they will be more likely to give our model a chance as we will be working with them and taking their suggestions. Integration wise, we could integrate our model in pre-existing systems. We could collaborate with hospital IT departments to integrate the model and ensure compatibility with electronic health records and other tools. Implementation and integration includes educating healthcare professionals and patients. To professionals, we need to provide comprehensive training on how to use the model effectively and efficiently. This includes understanding the model’s predictions (Phase 2 - Model Interpretability), integrating it into their workflow on a daily basis, and interpreting results accordingly. We also want to educate patients in understanding the benefits of the model and would address any concerns they might have regarding the use of AI in healthcare. Through all of this education, we want to emphasize to both patients and professionals the importance of early detection.

To add onto Phase 5, we will also include an image of the breast tissue slice of a patient with the cancerous tissue areas colored red. An example of this can be seen in the image below. For the time being, this will be included in the pilot, but hospitals will have to pay extra for this feature to be included. This allows for time saved by the healthcare professional and costs to be saved by the patient.



One potential drawback is that our dataset is 4 years old and hasn’t been updated since then, hence, there could be more detection approaches or recent breakthroughs in catching this type of breast cancer that our model has not accounted for. There could also be more types of classifiers of this type of breast cancer that have not been taken into consideration. To keep improving our model and give the most accurate results to our clients, we plan to look at the gender of the person and segregate our model based on that in the future along with updating our dataset. Since the gender is not classified either, it might reduce the accuracy of our current model. Hence, this is something we would have to consider immediately when marketing this to health care professionals.

**Appendix**

Group Contributions -

Saloni Agarwal: Business Understanding, Data Understanding, Deployment

Chi Chen: Data Preparation, Modeling, Evaluation

Peihan Liu: Modeling, Implementation, Evaluation

Tanya Ravi: Data Preparation, Presentation

1. <https://www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United,What%20Is%20Breast%20Cancer%3F> [↑](#footnote-ref-0)
2. <https://www.cancer.gov/types/breast/patient/breast-screening-pdq#:~:text=Mammography%20is%20the%20most%20common,use%20depends%20on%20certain%20factors> [↑](#footnote-ref-1)
3. <https://www.breastcancer.org/types/invasive-ductal-carcinoma> [↑](#footnote-ref-2)
4. <https://www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United,What%20Is%20Breast%20Cancer%3F> [↑](#footnote-ref-3)
5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5638217/> [↑](#footnote-ref-4)
6. https://github.com/sayakpaul/Breast-Cancer-Detection-using-Deep-Learning [↑](#footnote-ref-5)