MACHINE LEARNING TECHNIQUES TO AID BREAST CANCER RECURRENCE PREDICTION

by

Madison Rose

A Signature Honors Project Presented to the

Honors College

East Carolina University

In Partial Fulfillment of the

Requirements for

Graduation with Honors

by

Madison Rose

Greenville, NC

May 2023

Approved by:

Dr. Nic Herndon

Dr. Venkat Gudivada

Abstract

Breast cancer is a leading cause of cancer death and one of the most common cancers among women. Treatment looks different for every patient due to a variety of factors. One factor that can change a patient's treatment plan is how aggressive their cancer is. Aggressive cancers are more likely to reoccur and require intense treatment options such as chemotherapy. Cancer aggression is currently measured by a recurrence score which can be determined by a pathologist viewing hematoxylin and eosin-stained slides (HE slides) from breast biopsies. Recurrence scores are an important factor considered by oncologists when crafting a treatment plan for their patients. However, these tests are costly and in high demand which limits patient access. In this work, applications of machine learning to the issue of breast cancer recurrence are discussed. The use of machine learning could greatly benefit recurrence prediction by aiding pathologists. The proposed method uses a three-step pipeline to accomplish this. Utilizing digitized HE slides, the pipeline will perform the following steps: data processing, clustering, and classification. Overall, this work aims to aid pathologists in recurrence score prediction and make recurrence score testing more accessible to breast cancer patients.

Background

Breast Cancer Recurrence

Women have a 1 in 8 chance of developing breast cancer and a 1 in 39 chance of dying from the disease in their lifetime. The American Cancer Society estimates 287,850 breast cancer diagnoses and 43,250 breast cancer deaths among women in the United States in 2022 (Giaquinto, Sung and Miller). The best forms of treatment vary among the different types of breast cancer. For some patients, hormone therapy alone is highly effective, and it is unnecessary to subject a patient to chemotherapy. Other cases are more severe and would benefit from a treatment plan that combines chemotherapy with hormone therapy. The Oncotype DX test makes it easier for oncologists to craft individual treatment plans for their patients. The Oncotype DX is used for patients with HER2-negative and ER-positive (estrogen receptor positive) breast cancer. The Oncotype DX test is a tumor profiling test which quantifies the 10-year risk for metastasis in patients as an integer number known as a recurrence score (Siow, De Boer and Lindeman). This score helps oncologists predict the likelihood that a patient will benefit from chemotherapy. The test is most effective for patients with tumors smaller than 5 cm that have not yet spread to the lymph nodes (Susan G. Komen).

An Oncotype DX score of 15 or lower means that the patient would likely not benefit from adding chemotherapy to their treatment plan. The side effects of chemotherapy outweigh the benefits that they would provide these patients with low recurrence scores. Therefore, a treatment plan with only hormone therapy is typically recommended for these patients. Patients with a recurrence score of 26 or higher are likely to benefit from a more aggressive treatment plan that utilizes both hormone therapy and chemotherapy. These patients are at a higher risk for metastasis. Patients with scores in the range 16 to 25 are patients who may see some benefit from the addition of chemotherapy to their treatment. However, it is unclear if the benefits they typically see from undergoing chemotherapy are from the chemotherapy itself or are from the ovarian suppression which is caused by chemotherapy. Oncologists will often recommend patients with a score in this range to combine either chemotherapy or an ovarian suppression treatment with their hormone therapy. These score range interpretations are for premenopausal women. These scores are interpreted differently for postmenopausal women (Susan G. Komen).

The use of adjuvant chemotherapy has been shown to be decreasing since the introduction of the Oncotype Dx test in 2004. However, this test is expensive which provides barriers to some patients (Siow, De Boer and Lindeman). One way to increase accessibility to these tests is to aid pathologists in the task of recurrence prediction. This can be achieved using machine learning. Machine learning and computer aided diagnosis is becoming more popular, and some studies have begun to try to apply this to breast cancer and recurrence prediction (Ha, Chang and Mutasa).

Data Processing

Data processing is an important step in machine learning. The main purpose of a data processing step is to clean data to allow for better analysis. This may consist of discarding irrelevant data, cleaning inaccurate data or splitting data. In regard to computer vision tasks in machine learning, discarding irrelevant data and splitting data are used often (García, Ramírez-Gallego and Luengo). It is important when training classification models to only include relevant and accurate data as to not negatively influence the model's decision making. Splitting data becomes important due to large image sizes. Most modern neural networks require small image inputs, such as images of size 256 x 256 pixels (Simonyan and Zisserman). However, most modern-day images are much larger and thus must be compressed or split to accurately and efficiently train a machine learning model. There are many different data processing techniques that can be used. For example, for smaller datasets, manual data processing may be most effective. Larger datasets may benefit from scripts to clean data. Other machine learning algorithms can also be applied to data processing. Effective data processing can speed up processing time and increase model accuracy (García, Ramírez-Gallego and Luengo).

Clustering

Clustering is a machine learning approach that looks to group data based on similarities. This approach is a form of unsupervised learning in which a model is provided data with no labels. This can allow the model freedom in deciding its groupings. Accuracy is determined by evaluating the output groups to see if there is a clear label and how well the data in the group matches the label (Janiesch, Zschech and Heinrich). Clustering can also be used in data processing. This approach could implement clustering to break data into groups of relevant and irrelevant data and only keep the relevant data group in the dataset. Many types of clustering algorithms are available today and the best algorithm to use is dependent on the type of data and the desired groupings (Zhou).

Classification

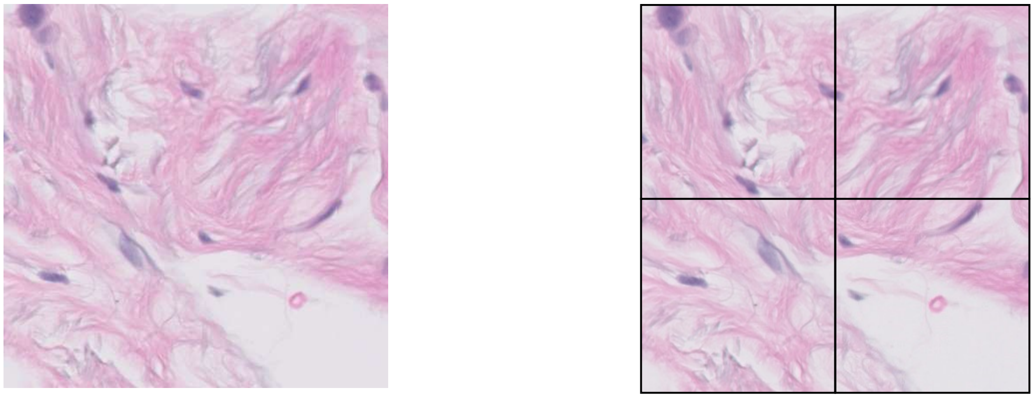
Classification is a popular machine learning approach that aims to predict an output label given input data. Unlike clustering, classification models generally utilize supervised learning. This means that the model is given both the data and the desired labels. This allows the model to recognize patterns. Models will adjust their internal configuration during training based on the data they see. After training, the model is shown unseen data and is asked to predict the output label. The accuracy is determined by how often the model predicted the correct label. There is a wide variety of classification algorithms that specialize in classifying different types of data (Zhang, Lipton and Li).

Methods

In this work, a three-step machine learning pipeline for breast cancer recurrence prediction is laid out. The steps include data processing, clustering, and classification. Analysis from the data processing step is also included. Resources from Amazon Web Services were utilized for the data processing and will be used in future implementation of the machine learning pipeline.

Data Processing

The dataset used here is a custom dataset that consists of 418 anonymized breast biopsy whole slide images along with a pathologist labeled recurrence score. This dataset was provided by Dr. Joseph Geradts from the Brody School of Medicine. These images were hosted in a pathology tool, Philips Digital Pathology, which allowed for viewing the images as well as the ability to download them at a variety of qualities and scan factors. The images used were downloaded with a scan factor of 40 which was the maximum available option. A quality of 80 was used due to only small differences between this image quality and images of quality 100, which allowed the images to be slightly smaller. The images were uploaded to an Amazon Web Services S3 bucket for analysis and processing. The images, which were downloaded as TIFF files, range from approximately 500 MB to 3.9 GB. Due to the extremely large sizes of these images, they would not be suitable for analysis using machine learning models as is. Therefore, the main goal of the data processing step is to tile these images into smaller chunks which can be better analyzed by various models. A few tile sizes were tested including (in pixels) 2000 x 2000, 1000 x 1000, 512 x 512, and 256 x 256. This tiling was done by using a custom Python script.

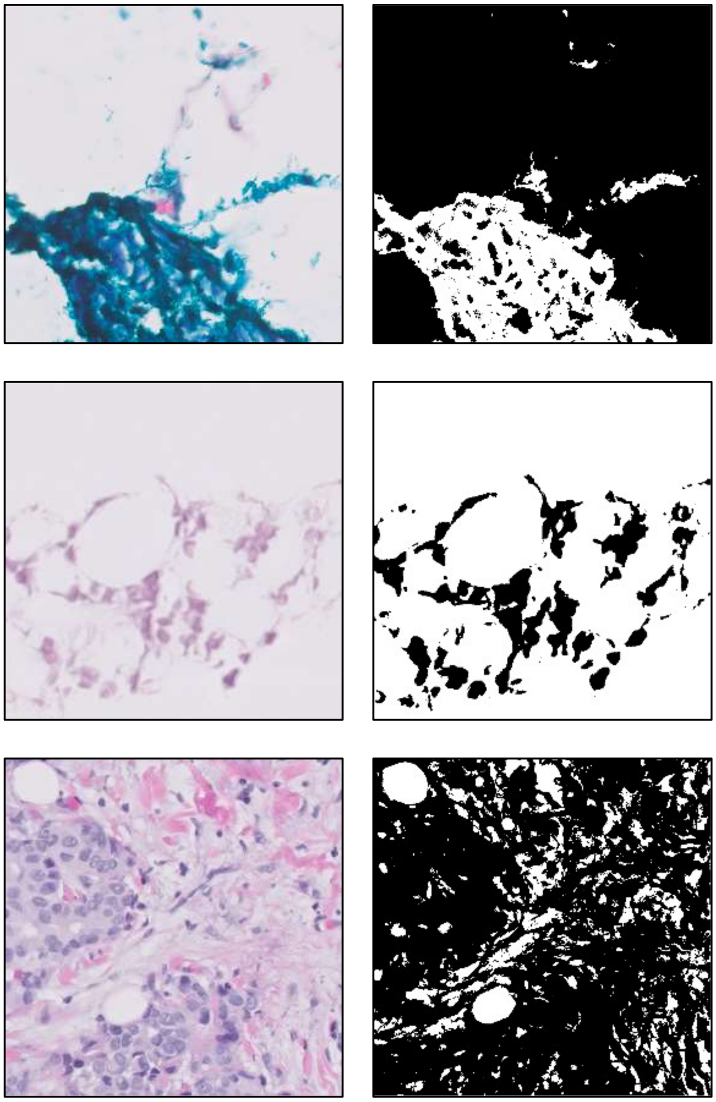


(a) An example of a tile of size 512 x 512 pixels

(b) The same tile as (a) represented by tiles of size 256 x 256

Figure 1 : Larger Tiles vs Smaller Tiles. The 512 x 512 tile size is considered a larger tile and the 256 x 256 tile size is considered smaller. Using smaller tiles makes for faster processing per tile during the classification step but exponentially increases the number of tiles per image. As seen above, 4 non-overlapping 256 x 256 tiles are needed to represent the same data as a single 512 x 512 tile.

Another goal of the data processing in this step was to discard irrelevant tiles to speed up processing time and avoid negative impacts on the classification model's decision making. One way this can be done is to remove tiles containing large amounts of whitespace as this is irrelevant information that will not be helpful to the classification model. Blue dye can also be removed in some cases as there are instances of blue dye being left along the outside of the sample, likely left by a pathologist. This could influence the model so it can be removed for increased accuracy and faster processing. The Python script performing this tiling also computed the amount of whitespace and blue dye for each image and automatically discarded tiles that contained a percentage of these above a given threshold. This was done using the OpenCV library and its masking functions. Whitespace values in the images were determined to be in the threshold [210, 210, 210] to [255, 255, 255]. Blue dye pixels were determined to be in the threshold [0, 0, 90] to [80, 255, 255]. These ranges were selected by first determining the general range for blue and white and then finetuning based on the specific shades that appear in the images. A mask of the image is then created by mapping any pixels in the given ranges to white (grayscale pixel value 255). Pixels not in the defined color range are mapped to black (grayscale pixel value 0). The number of white pixels in each mask was calculated and divided by the total pixels in the original image to determine the percentage of pixels in the original image in the defined range. A similar process is repeated to determine the percentage of blue dye in the image.



(a) Tiles with blue dye covering more than 15% of the image were discarded. This tile contains 20.6% blue dye and was discarded.

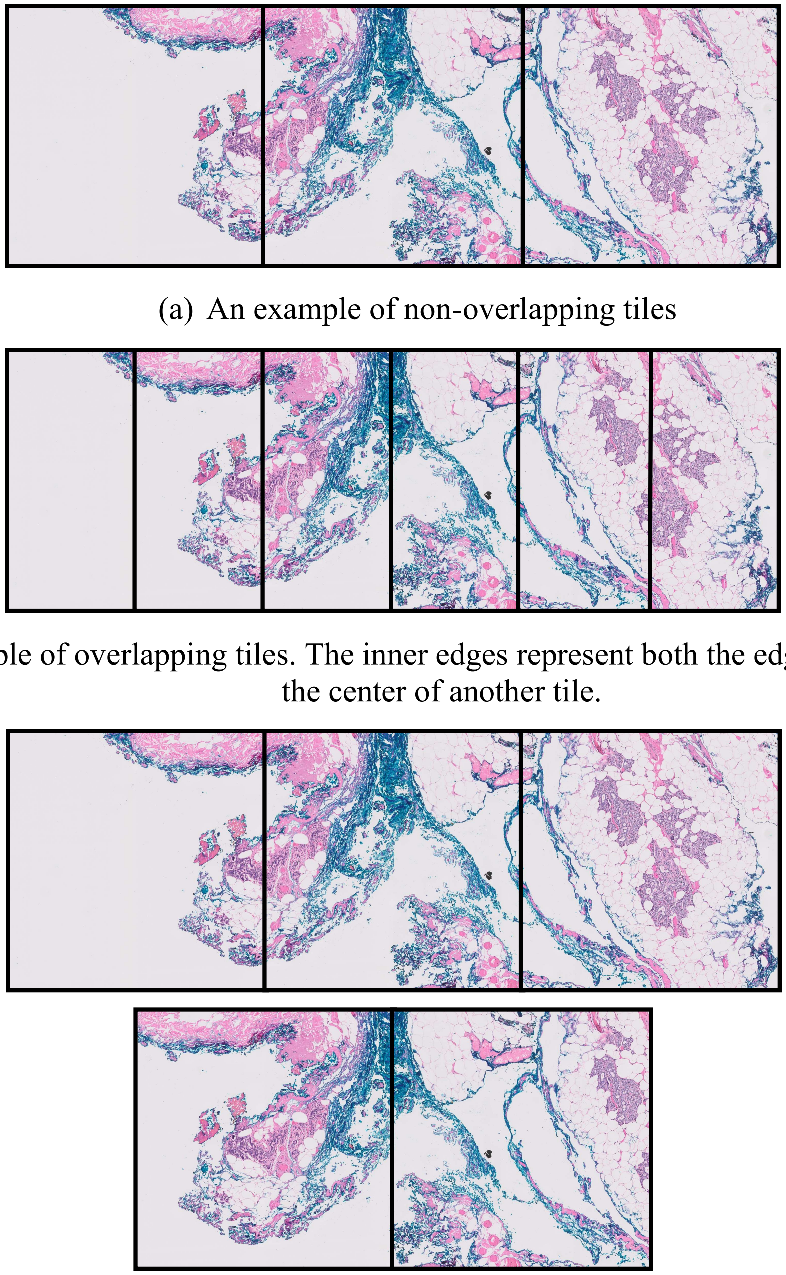
(b) Tiles with greater than 75% whitespace were discarded. This tile contains 86.8% whitespace and was discarded.

(c) Tiles under the threshold for whitespace and blue dye were kept. This tile contains 18.6% whitespace and 0% blue dye and was kept in the dataset.

Figure 2: Cleaning the dataset: Tiles with large amounts of whitespace and blue dye were discarded to clean the dataset. This eliminates irrelevant data and data that can skew the classification model.

The final percentage of whitespace and blue dye to allow is still under analysis. However, through visual analysis, current test runs have the threshold at 75% for whitespace and 15% for blue dye. Whitespace is considered irrelevant information while the blue dye tiles could negatively impact the model's decision making. Further analysis will be done in the future to determine if tiles containing blue dye should only be discarded when they also contain a substantial amount of whitespace and are likely on the 'edge'. Removing tiles with whitespace and blue dye can greatly decrease the total number of tiles in an image as shown later in Figure 4.

Another consideration when tiling the images is whether to use overlapping image tiles. With non-overlapping tiles, the edge of images align and contain no overlapping or repeated data. When using a non-overlapping method, the number of tiles created is much smaller which will allow for faster processing in the classification step. However, this means there could be relevant data that is contained on the edge of a tile and can not be fully seen in a single tile. This is where the use of overlapping tiles can be useful. Overlapping tiles is a technique that involves starting the edge of the next tile in the middle of the previous tile. This allows the data that was on the edge of the previous tile to be in the center of the next tile and fully contained. The benefit to this technique is that data that may be missed on the edge of non-overlapping tiles will be in the center of the next tile. However, this greatly increases the number of tiles created from a single image. Smaller overlap amounts can also be considered such as starting the edge of the next tile at the ¾ mark of the previous tile. Analysis is being performed on total tile counts for overlapping and non-overlapping tiles. Ultimately, the decision of how much overlap to use if any will consider both processing time and relevant data on the edge of each tile.



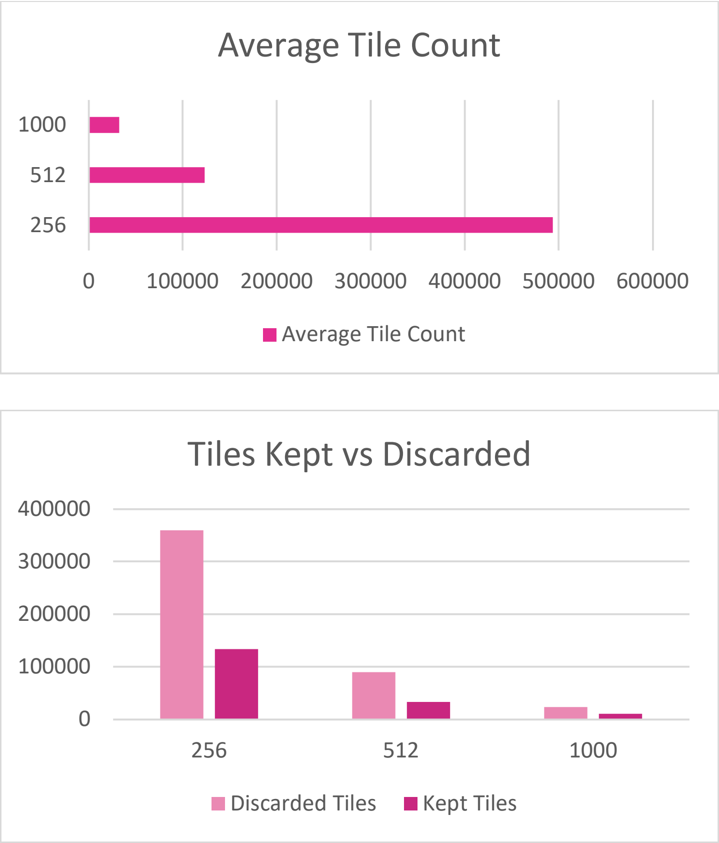
(a) An example of non-overlapping tiles

(b) An example of overlapping tiles. The inner edges represent both the edge of one tile and the center of another tile.

(c) This shows the extra tiles that were generated from the overlapping technique beneath the 3 non-overlapping tiles. The data that is at the edge of the first non-overlapping tile is at the center of the overlapping tile.

Figure 3: Overlapping vs. Non-overlapping Tiles. Overlapping tiles allow for regions of the data that may be on the edge of a tile to be the center of the next tile. This means that relevant data is not cut off or excluded. However, each image will produce significantly more tiles. The example above goes from requiring 3 tiles to requiring 5 tiles to cover the same region of data.

Analysis on the tile counts of each image was performed. Using a tile size of 512 x 512 yielded an average of approximately 39,600 tiles while images tiled with a size of 256 x 256 pixels yielded an average of approximately 9,900 tiles per image when using non overlapping tiles. When using overlapping tiles, a tile size of 512 x 512 yielded an average of approximately 123,000 tiles per image. The 256 x 256 tile size for overlapping tiles resulted in an average of approximately 493,000 tiles per whole slide image. Using a tile size of 256 x 256 produces four times the number of tiles as using a 512 x 512 tile size on an image. This is because it takes four 256 x 256 tiles to cover the same area covered by a 512 x 512 tile as demonstrated in Figure 1. The number of tiles exponentially increases when using overlapping tiles in comparison to non-overlapping tiles. Figure 4 provides a visualization of the amounts of tiles yielded by various tile sizes. This analysis is important in choosing the optimal tile size. Using a tile size that is too large may lead to longer processing time for the model. Using a tile size that is too small may result in extremely large amounts of data that cannot all be utilized. Analysis is still being done on the optimal tile size. Image tiles were also analyzed to determine how many were being kept and how many were being discarded. The script is currently being used with the 75% whitespace and 15% blue dye thresholds and discarded an average of 72% of tiles created per image. Whitespace is the main cause for discarding a tile as blue dye made up only 1% of the discards.



(a) This chart shows the average number of tiles created per image by different tile sizes. There is 1 bin per image to denote the average tile count.

(b) Analysis was performed to determine the number of tiles being kept or discarded. Most of the discarded tiles were due to whitespace. Only 1% of discarded tiles were due to blue dye.

Figure 4: Image Tile Counts by Tile Size. Analysis was performed on a random selection of 50 images to find average tile counts per tile by size used. The numbers shown above were calculated by considering a tiling using overlapping tiles.

Clustering

The clustering step in the pipeline serves as an extension of the data processing stage. Each whole slide image contains only a single recurrence score that is associated with the cancer cells in the image. However, since the image will be tiled, there may be some tiles from the image that contain only healthy cells. If these tiles with healthy cells were associated with large recurrence scores, it could negatively impact the classification model. There are two options to handle this issue. The first is to remove the tiles that contain only healthy cells from the dataset and use the cancerous tiles and their recurrence scores in the classification step. Another option is to assign low recurrence scores to the healthy cell tiles. A heatmap could then be generated for each tissue sample image to determine an overall recurrence score given both the cancerous and healthy cells contained in the image. The clustering step will aim to split the tiles remaining from the data processing step into two groups to identify the healthy cells so that one of these approaches can be applied. This clustering will be done with a machine learning algorithm in Amazon SageMaker using Python. Images sorted into the "cancerous cells" group will then move forward, along with their image's recurrence score, to the classification step. Tiles sorted into the "healthy cells" group will either be discarded or assigned a new, low recurrence score, depending on the method chosen. If the second option is selected, heatmaps of the tissue samples will need to be generated prior to the classification step.

Classification

The classification step is the final step in the pipeline and will be where the recurrence prediction occurs. This step will also utilize Amazon SageMaker. If the option of removing healthy cell tiles from the dataset is chosen, the remaining cancerous tiles will be split into training, testing and validation groups. Once the data has been split, the training dataset will be used to train the model. The model will be provided with both tiled images and their recurrence scores as training data. The trained model can then be tested on unseen data to determine accuracy of the model. Fine tuning and adjustments may be made to the model to increase overall accuracy. If the heatmap option is selected, the classification model will be provided with an image's heatmap along with the associated recurrence score for training. Testing will be performed by providing only the heatmap and having the model predict the recurrence score for the whole image.

Future Work

The next step in the process for this work is to determine the optimal tile size and overlap amount. Analysis on these is being done by analyzing other similar studies to see what was most successful. Also, tests are being run on the dataset to determine the number of tiles various tile sizes and overlap produce. The optimal size will be one that includes enough relevant data and is small enough that it is easy for the classifier to process but also does not produce so much data that only a small portion of it can be used. Once these have been determined, the scripts for the data processing can be implemented into the AWS SageMaker pipeline to begin tiling the images and preparing for the clustering stage. Another decision that will need to be made is if the "healthy cell" tile group from clustering will be removed from the dataset or if it will be used to create heatmaps for each tissue sample. These are the decisions that will need to be made in order to continue moving forward to the classification step.

Related Work

A 2017 study proposed a convolutional neural network architecture specifically designed to recognize the structure of breast tissue and nuclei in pathology images. This work investigated classifying images using both a two and four label classification method. The four-label classification differentiated tissue between normal, benign lesions, in situ carcinoma and invasive carcinoma while the two-label classification divided images into carcinoma vs. non-carcinoma classes. In this study, non-overlapping image patches of size 512 x 512 pixels were fed into a convolutional neural network (CNN) model and a CNN+SVM (support vector machine) model using three different patch probability fusion methods. This CNN architecture included five convolutional and three max pooling layers with kernel sizes of 3x3 and two max pooling layers with a 2x2 filter. Three fully connected layers were also included in the architecture. The accuracy was approximately 65% for four-label classification and 77% for two-label classification for the CNN only model. The CNN + SVM showed improved results with an accuracy of 83.3% (Araújo T).

Another similar study attempted to predict Oncotype DX recurrence scores in breast cancer MRIs in 2019. This work chose convolutional kernels of size 3x3 and max pooling kernels of size 2x2 to prevent overfitting. 134 cases were used in the study and were split into 80% training and 20% testing data. Patients were classified into three different recurrence categories based on likeliness of recurrence in 10 years. A score of 18 or below was considered low risk. Scores between 18 and 30 fell into the moderate risk category while scores above 30 were classified as high risk. Of the 134 cases, 77 were low risk, 40 were moderate risk and only 17 were considered high risk for recurrence. Two different experiments were conducted. One with three classifications for recurrence scores (low, moderate, high). The second experiment packaged moderate and high recurrence scores together with low-risk scores remaining their own class. The three-class experiment yielded an accuracy score of 81%, while results for the two-class experiment were slightly better with an accuracy of 84% (Ha, Chang and Mutasa).

Conclusion

Recurrence score testing has helped decrease chemotherapy rates among breast cancer patients. The Oncotype DX test for recurrence is helping oncologists make more informed decisions about treatment for their patients. However, high costs and high demand of the test limits its availability (Siow, De Boer and Lindeman). The three-step pipeline discussed in this paper lays out a plan to utilize machine learning for recurrence prediction using a custom HE slide dataset to automate the task of recurrence prediction to aid pathologists. There are many benefits to utilizing machine learning in this area. Eventually machine learning developments in breast cancer recurrence prediction may reduce human error and help pathologists learn more about markers for recurrence. The overall goal of this work is to further research machine learning methods for medical image diagnostics and work towards making recurrence tests more accessible for breast cancer patients.

Acknowledgements

Thank you to Dr. Joseph Geradts from the Brody School of Medicine for providing the idea for this project and the dataset. Thank you, Amazon Web Services, for providing computational resources and support.

References

American Cancer Society. *Breast Cancer HER2 Status*. 25 August 2022. <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-her2-status.html>.

—. *Breast Cancer Hormone Receptor Status*. 8 November 2021. <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html>.

Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. "Classification of breast cancer histology images using Convolutional Neural Networks." *PLoS ONE* (2017).

CDC. *Basic Information About Breast Cancer*. 2022. 7 August 2022.

Giaquinto, Angela N, et al. "Breast Cancer Statistics, 2022." *CA: a cancer journal for clinicians* (2022): 524-541.

Ha, R, P Chang and S Mutasa. "Convolutional Neural Network Using a Breast MRI Tumor Dataset Can Predict Oncotype DX Recurrence Score." *Journal of Magnetic Resonance Imaging* (2019).

Kumar, Neeraj, et al. "Convolutional Neural Networks for Prostate Cancer Recurrence Prediction." *SPIE Medical Imaging*. 2017.

Pouyanfar, Samira, et al. "A Survey on Deep Learning Algorithms, Techniques, and Applications." *ACM Computing Surveys* (2019): 1-36.

Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations*. San Diego, 2015.

Siow, Rong Zhen, et al. "Spotlight on the utility of the Oncotype DX® breast cancer assay." *International Journal of Women's Health* (2018): 89-100.

Susan G. Komen. *Oncotype DX*. 5 January 2022. <https://www.komen.org/breast-cancer/diagnosis/factors-that-affect-prognosis/oncotype-dx/>.

Szegedy, Christian, et al. "Rethinking the Inception Architecture for Computer Vision." *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

Zhang, Aston, et al. *Dive into Deep Learning*. Corwin, 2021.