Spendylove Apaloo

CSC 485 – 001

Dr. Chou

May 10, 2024

Breast Cancer Classification Using Convolutional Neural Networks

## **Introduction**

Breast cancer stands as the most prevalent form of cancer affecting individuals in the United States, with a staggering number of cases reported annually. According to the breastcancer.org “In 2023, an estimated 297,790 new cases of invasive breast cancer are expected to be diagnosed in U.S. women, along with 55,720 new cases of DCIS. In 2023, an estimated 2,800 new cases of invasive breast cancer are expected to be diagnosed in men” (American Cancer Society). Detection primarily relies on the use of breast examinations or clinical devices such as mammograms, ultrasounds, MRI, and biopsies. However, the devices may not always provide the most accurate results to the patients. Studies show that mammograms have a sensitivity rate of about 87% (Susan G Komen). Additionally, screening mammograms “miss about 20% of breast cancers that are present at the time of screening” (NCI). False results can have negative effects for the patients. For those who receives a false-negative results, there may be delays in early preventative treatments that have the potential to lessen the spread of the cancer cells as well as the negative effects that may occur when the cancers are not identified early on. . However, advancements in technology, notably through the utilization of deep learning models, offer promise in improving prediction accuracy.

By utilizing extensive datasets, these models enhance their predictive powers and may fill the void left by conventional screening techniques. Medical personnel can more accurately detect small indications of cancer by integrating state-of-the-art technology into their clinical practices. This allows for interventions to be made sooner in the disease's course. The objective of this paper is to build an algorithm capable of identifying the presence of breast cancer using convolutional neural networks.

## **Data**

The data for this paper was obtained by the Duke University Health System, however the data set was curated by the National Cancer Institute. Researchers analyzed and annotated analyzed Digital Breast Tomosynthesis (DBT) volumes obtained from Duke Health System. The study included 16,802 DBT studies from 13,954 patients performed between August 26, 2014, and January 29, 2018. These studies were divided into four groups: normal, actionable, benign, and cancer. The normal group comprised screening studies without abnormal findings, the actionable group included studies necessitating further imaging examination, the benign group involved studies with biopsied benign masses or architectural distortions, and the cancer group encompassed studies with biopsied cancerous masses or architectural distortions. The dataset consisted of 22,032 reconstructed volumes from 5,610 studies involving 5,060 patients. The researchers then divided data into distinct training, validation, and test sets. The test set comprised 460 studies from 418 patients, while the validation set included 312 studies from 280 patients. The remaining 4,838 studies from 4,362 patients constituted the training set. Importantly, the selection process for including cases from the benign and cancer groups in the test and validation sets was conducted to maintain a balanced representation of masses and architectural distortions across subsets.

# **Method**

The medical images curated for this data set are DICOM (Digital Imaging and Communications in Medicine) images. DICOM is a standard format used for storing and transmitting medical imaging data, including X-rays, MRIs, and CT scans. However, DICOM images can be challenging to work with in the context of CNN models. DICOM images may have varying resolutions, orientations, and pixel intensities, therefore, there it is necessary to normalize and standardize the images to ensure consistency across the dataset. Finally, the size of DICOM datasets can pose computational challenges, requiring efficient data handling and processing.

To address the challenges posed by DICOM images, labeling data, and the scarcity of patients with cancer cells, several solutions were implemented. Firstly, only the training data provided was utilized as the master data set for this project. The new data was then split into a training, validation, and test set. Secondly, the DICOM images were resized to ensure uniform dimensions across all inputs, which helped streamline processing and analysis. Additionally, the data was preprocessed to generate NumPy arrays, enabling efficient manipulation and utilization within the model architecture. Finally, the data were placed into a categorized as cancer vs. non-cancer because of the limited number of cancer cases – 1.6% of the data set (Buda, Saha et.al).

The paper utilizes Convolutional Neural Networks (CNN) to build a deep learning model, due to its superior ability for image classification. Convolutional Neural networks consist of their main layers, namely, the convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for extracting local patterns or features from the input image through a series of convolutional operations. The layer uses a small sliding filter (also known as a kernel) to slide over the input image to get a full convolution using element wise multiplication. over the input image and computing the dot product between the filter and the local patch of pixels at each position. Activation layers are then applied to the output of each convolutional operation to introduce nonlinearity. This paper uses the ReLu (Rectified Linear Unit) activation, which prevents the emergence of the “vanishing gradient” problem.

Pooling layers serve to down sample the feature maps, reducing the spatial dimensions while retaining the most important information. The pooling operations included in this algorithm is the max pooling, which selects the maximum value within each local region. Pooling lessens the number of parameters in succeeding layers, which eases the computational load, and it also decreases in the number of parameters in subsequent layers.

Finally, fully connected layers assign the features that were extracted from the earlier layers to the output classes or labels. Based on the learnt representations, these layers carry out sophisticated reasoning and decision-making, which finally results in the final classification or regression output. The model uses sigmoid activation as a binary classification method In order to reduce the difference between predicted and ground truth labels, CNNs learn to optimize their parameters such as filter weights and biases during training by using gradient descent and backpropagation.

# **Results**

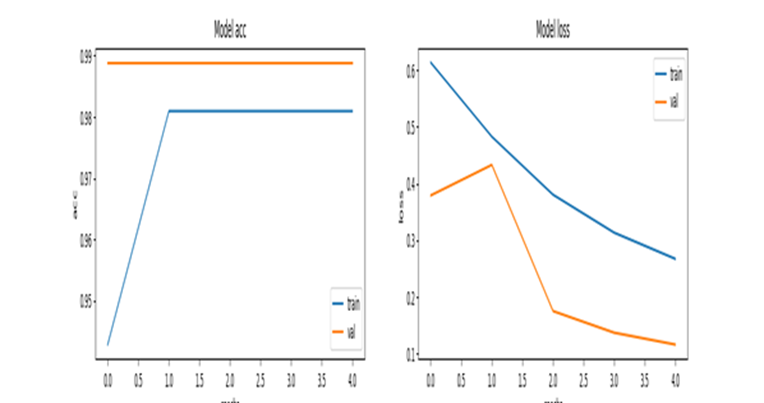
Figure one shows the results of the CNN model. The image indicates that the model performs exceptionally well on the training data with a training accuracy is constant at about 0.98, and a training loss decreases steadily to near zero. However, the model’s performance on the validation data is not as good. The model’s validation accuracy starts at about 0.98 but drops sharply to around 0.55. Similarly, the validation loss starts at a similar point as the training loss but increases sharply to about 0.5 before decreasing slightly. This discrepancy between training and validation performance is a classic sign of overfitting, because of the ratio of cancer to non-cancer cells. To address the issue of overfitting, the algorithm is modified to take half of the cancerous and non-cancerous data when training the model. However, the subsequent output graphs did not yield better results for model classification.

Figure 1 Model Results.

# **Conclusion**

Convolution Neural Networks are superior deep- learning model with the ability for precise and accurate image classification. The models built using can be used as additional tools for breast cancer classification and they can aid in the detection of early breast cancer cells in the patients. Due to the limitation of diversity in the data as well as the lack of access sophisticated computers, the CNN model used in this paper did not yield the desired results. Therefore, there is room for further research and fine tuning in order to produce a model that will yield the desired results. Future research can explore techniques to augment the dataset with additional cancer samples. By addressing these avenues for further research, we can strive towards achieving more accurate and reliable detection of breast cancer, which can ultimately lead to improved patient outcomes and healthcare practices.

Works Cited

American Cancer Society: “Breast Cancer Facts & Figures 2022-2024,” “Cancer Facts & Figures 2023.”

Buda M, Saha A, Walsh R, et al. A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images. JAMA Netw Open. 2021;4(8):e2119100. doi:10.1001/jamanetworkopen.2021.19100

“Mammogram Accuracy - Accuracy of Mammograms.” Susan G. Komen, 9 Dec. 2022.

“Mammograms.” NCI, www.cancer.gov. Accessed 10 May 2024.