Breast Cancer Detection Using Convolutional Neural Networks

Aditya Agarwal (agarwal.adi@northeastern.edu)
Pushkar Dhabe (dhabe.p@northeastern.edu)
Himani Thakker (thakker.hi@northeastern.edu)

Northeastern University, College of Engineering,
Data Analytics and Engineering Program, Boston, MA, USA

Abstract

The second leading cause of death for women is breast cancer. Early detection and treatment reduce breast cancer mortality. Mammography is crucial for breast cancer screening because it can find early breast lumps or calcification areas. One drawback of breast mammography is that breast cancer masses are harder to find in unusually dense breast tissue.

The objective is to examine various deep-learning techniques that can be applied to create a system that learns how to recognize breast cancer in mammograms and forecast classification results.

With this project, we can help oncologists detect breast cancer at its early stages, maximizing the patient's survival rate and preventing unnecessary or late treatment of patients.

Introduction

In the past, the number of cases of females with breast cancer has increased up to 30%, mainly due to constant exposure to the growth-promoting effects of the female hormones estrogen and progesterone. It is essential to identify the correct type of malignant. Additionally, the conventional method, which involves a visual examination (such as looking for lumps) and cutting-edge medical tools, is insufficient for producing accurate results. The accuracy that the oncologists achieved with this additional technical support for taking high-resolution images was in the range of 74% - 87%.

A mass may be benign or malignant. The shape of benign tumors, which are round or oval, distinguishes them from malignant ones, which have a partially rounded shape and an irregular outline. The malignant mass will also be whiter than any tissue around it.

To address this problem, machine learning techniques have been very successful. By automating the classification process and enabling quick and inexpensive access to the results with basic devices, the objective is to assist doctors. The results of earlier investigations using conventional machine learning models have been exceptional. Still, state-of-the-art models can produce ground-breaking results due to advancements in deep-learning model architectures.

In this project, we will be using Mendeley Data to put together the dataset for this project. The dataset contains the Mammographic Imaging Analysis Society (MIAS) database and the breast dataset. The mammography dataset includes both benign and malignant masses.

Background

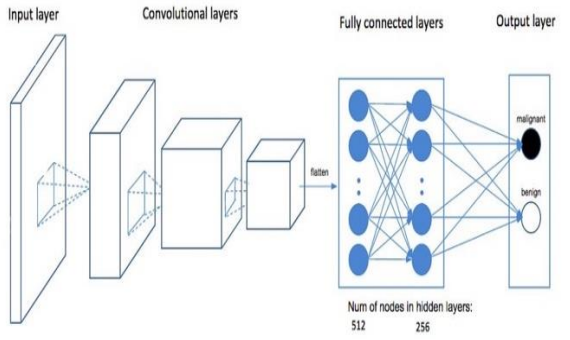
The Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT has developed a predictive analytics model that can predict a patient's risk of developing breast cancer over

several periods. But this model is presently being created. Doctors from Apollo in India, Novant Health in North Carolina, and Barretos in Brazil invested in this technology for detection. This model evaluates the need for additional testing and how frequently women should be screened by considering several factors, including age, hormones, genetics, and breast density.

According to pertinent published data, the precision of risk models used in clinical practice remains constrained despite decades of research and work. Although other Al methods and predictive analytics have shown promise in predicting cancer risk, they frequently perform poorly in new patient populations. It can be challenging to find masses in dense breast tissue.

Approach

One of the most potent neural network architectures, convolutional neural networks, is used extensively in image recognition and classification. They are capable of recognizing objects and images. CNNs have occasionally been shown to be more accurate than humans and capable of producing precise results. In our project, we directly use CNN to extract features and determine the connection between the unlabeled raw pixels and labels.



In this project, we have considered the MIAS and the INbreast dataset. Firstly, we have normalized both datasets in which we convert the image 255 scales to 0-1 scale. We have also partitioned the dataset into training, testing, and validation sets accordingly. For image classification, we have created a baseline CNN model with 2 layers and 32 kernels each and compared it with the advanced CNN model, which has 3 layers with kernel sizes of 32, 64, and 128, respectively. The image below shows the number of layers and number of trainable parameters for the baseline CNN model.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 92, 140, 32)       896

max_pooling2d (MaxPooling2D  (None, 46, 70, 32)        0
)

conv2d_1 (Conv2D)            (None, 46, 70, 32)        9248

max_pooling2d_1 (MaxPooling  (None, 23, 35, 32)        0
2D)

flatten (Flatten)            (None, 25760)             0

dropout (Dropout)            (None, 25760)             0

dense (Dense)                (None, 50)                1288050

dropout_1 (Dropout)          (None, 50)                0

dense_1 (Dense)              (None, 2)                 102

=================================================================
Total params: 1,298,296
Trainable params: 1,298,296
Non-trainable params: 0
```

Since all the images are non-linear, we have used the Relu activation function to bring non-linearity. We have used Maxpooling after each convolution layer to downsample the image.

Also, we have used Categorical cross-entropy as our loss function to distinguish the probability distribution of benign and malignant from each other. We will use Adam optimizer to update the weights and learning rate accordingly. To evaluate the model performance, we have used "Accuracy" as our performance metric. The image below shows the number of layers and number of trainable parameters for the advanced CNN model (with more number of layers).

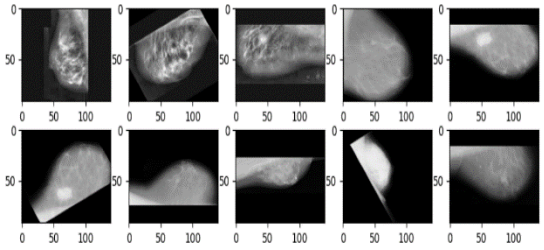
Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 92, 140, 32)	896
max_pooling2d (MaxPooling2D)	(None, 46, 70, 32)	0
conv2d_1 (Conv2D)	(None, 46, 70, 32)	9248
max_pooling2d_1 (MaxPooling2D)	(None, 23, 35, 32)	0
conv2d_2 (Conv2D)	(None, 23, 35, 64)	18496
max_pooling2d_2 (MaxPooling2D)	(None, 11, 17, 64)	0
conv2d_3 (Conv2D)	(None, 11, 17, 128)	73856
max_pooling2d_3 (MaxPooling2D)	(None, 5, 8, 128)	0
dropout (Dropout)	(None, 5, 8, 128)	0
flatten (Flatten)	(None, 5120)	0
dropout_1 (Dropout)	(None, 5120)	0
dense (Dense)	(None, 50)	256050
dropout_2 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 2)	102

Total params: 358,648
Trainable params: 358,648
Non-trainable params: 0

Results

The dataset to be used in this project is put together by Mendeley Data. The dataset contains Mammographic Imaging Analysis Society (MIAS) database and INbreast dataset. The mammography dataset includes both benign and malignant masses.

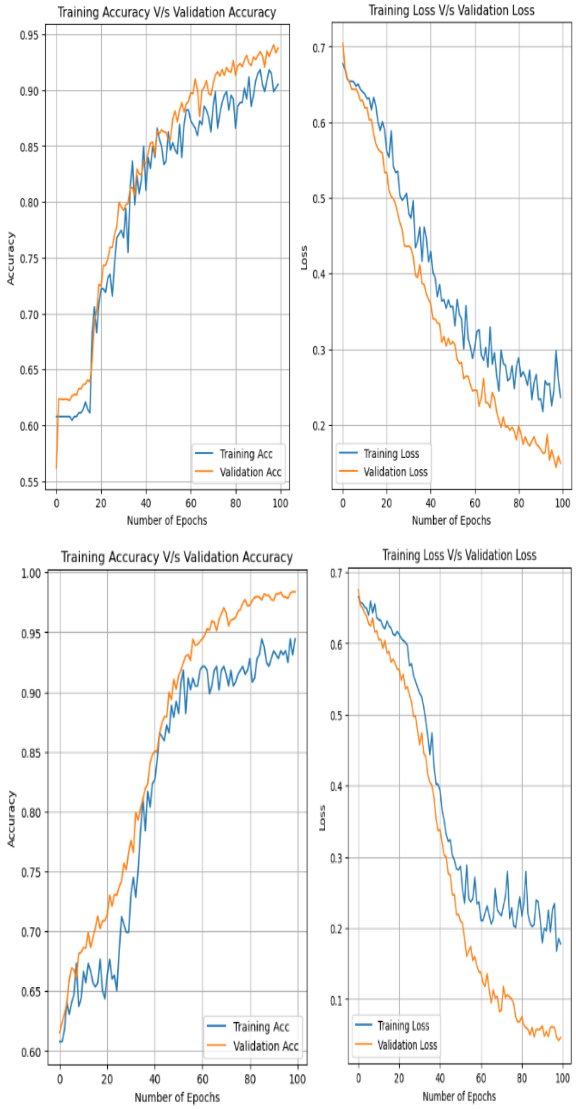
To create the images for this dataset, 106 masses from the INbreast dataset, 53 masses from the MIAS dataset were first extracted. INbreast dataset has 7632 images and MIAS dataset has 3816 images.



The dataset is further divided into training and testing sets. The training set consists of 90% of the data, and the test has 10% of the data points. Next, we normalized both the datasets and designed a CNN network and train the model over 100 epochs. The reason for keeping 100 epochs was because the data points available are very few for each class.

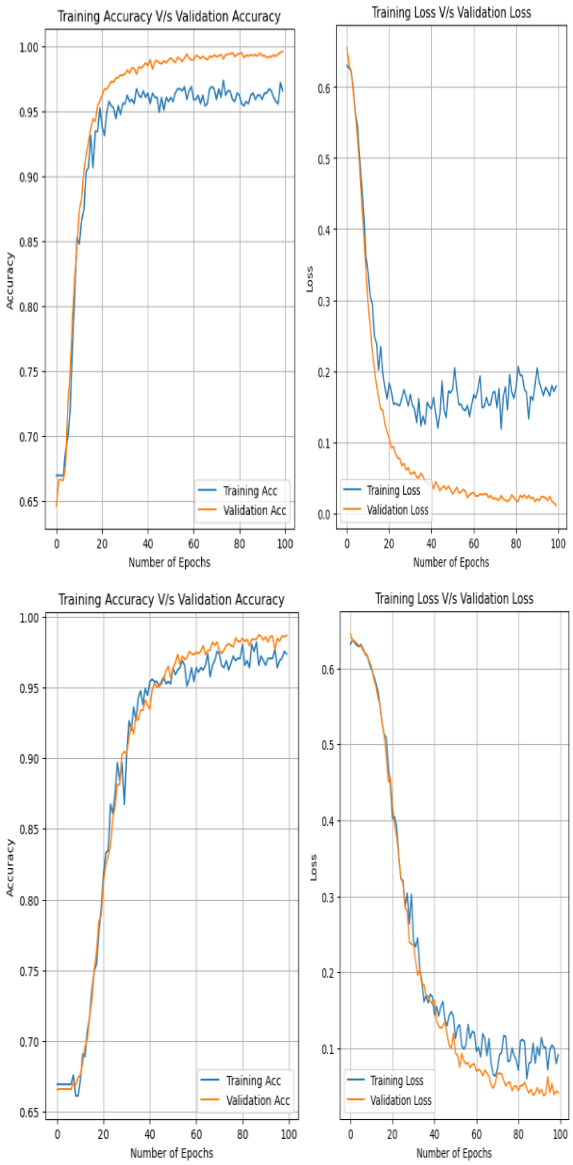
For both MIAS and INbreast dataset, the difference between training accuracy and testing accuracy is minimal, as can be seen in the chart below. However, the disparity between the loss of the training set and the validation set grows as the number of epochs increases.

For MIAS dataset, the accuracy and loss values for baseline, and advanced CNN model respectively –



We can also depict that the accuracy starts to increase with the number of epochs, and ultimately saturates. And there is no sign of underfitting and overfitting, as validation accuracy and training accuracy curves are similar in distribution.

For INbreast dataset, the accuracy and loss values for baseline and advanced CNN model are shown below –



Conclusion

The project's results were encouraging, and it appears that neural networks are performing effectively for the detection of breast cancer. In this project, we majorly focused on using CNN model without pre-training on large datasets. Despite the biased dataset, which had some classes with fewer data points than others, the model appears to perform remarkably well. Given that the model was trained using the CPU, less computation was needed.

Future research may employ more sophisticated systems to train the model more quickly and complete the necessary computations. We can also use Res-net model architecture to predict the data points more accurately.

References

1. Breast cancer detection using deep convolutional neural networks and support vector machines by Dina A. Ragab, Maha Sharkas, Stephen Marshall, Jinchang Ren https://peerj.com/articles/6201/
2. Classification of Breast Cancer from Mammogram images using Deep Convolution Neural Networks by Sobia Shakeel; Gulistan Raja https://ieeexplore.ieee.org/document/9393191
3. Building a convolutional neural network using Tensorflow-keras by Analytics Vidhya https://www.analyticsvidhya.com/blog/2021/06/building-a-convolutional-neural-network-using-tensorflow-keras/