Mammograms Research

Kevin Lopez
Computer Engineering & Computer Science Department
Kevin.Lopez@student.csulb.edu

April 2024

1 Introduction

Breast cancer is the second most diagnosed cancer among US women, after skin cancers, and is the second leading cause of death among women after lung cancer [7]. A common method of screening for breast cancer is through mammograms. A mammogram is a low-dose x-ray of the breast; it is usually done as a screening method before a patient has any symptoms or feels any lumps. Mammograms can often show abnormal areas in the breasts, like masses and calcifications.

Calcifications(white spots of excess calcium). Macrocalcifications, larger deposits typically linked to non-cancerous conditions such as injuries or inflammations, contrast with microcalcifications, which are smaller calcium specks when arranged irregularly, may indicate cancer risk [6]. Additionally, **Masses** could represent benign conditions like cysts or fibroadenomas, tumors composed of glandular and connective tissue [11].

Mammograms cannot tell for sure if there is cancer, but they can aid providers in deciding if further testing is needed. A lot of the time, after a mammogram has been done, something has been detected, and further testing, such as an ultrasound or biopsy, will proceed. Sensitivity(recall) in traditional mammography is about 87% [15]. It is hard for the human eye to detect if there is cancer in a mass or calcification while looking at a mammogram.

Mammography, along with AI, are tools that can be used to focus on accurately detecting breast cancer. In this paper, I introduce a machine-learning approach to finding breast cancer by only using mammograms with a high recall. High Sensitivity(recall) means that if cancer is present in the person, there would be an 87 percent plus that it gets flagged;

1.1 Related Work

Using AI to help aid the detection of diseases is not a new field; previously, AI has been used to help detect Pneumonia [19], Skin cancer [8], diabetic retinopathy [9], Lymph Node Metastases in Breast Cancer [5], and others [20, 17]

In the case of mammography. The following research has been done. In [20], researchers developed a DNN model that performed about the same as human radiologists. In [13], the

researchers developed a CNN that outperforms a state-of-the-art CAD system (Computer Aid Detection). Among other research [16, 4, 2]

In the domain of convolutional neural networks (CNNs), several competitive architectures have been developed, among which GoogLeNet [21] is particularly notable for its inception layers. Introduced in 2014, this architecture was designed to build deeper networks through inception layers that simultaneously applied multiple filters (1x1, 3x3, 5x5) sizes on the same input data, enhancing the network's ability to capture features at various scales. New iterations of GoogLeNet [22] added enhancements to reduce the number of computations(factorization 5x5 filters into sequences of 3x3 filters) and the incorporation of batch normalization to accelerate training and reduce overfitting. In this work, I'll be using GoogLeNet advancements in CNN to build a model that detects when cancer is present in mammograms.

1.2 Dataset

The mammography dataset used for this project was found in Kaggle [12]. It consists of 299x299x1 images collected from DDSM [10] and CBIS [14].

The data was already split into the following files:

- Training data: Contains 5 .tfrecords files, where each file includes images and their corresponding labels.
- CV data: Includes 2 .npy files, one for images and the other for labels.
- Test data: Consists of 2 .npy files, one for images and the other for labels.

The dataset labels are as follows:

- 0 - Negative (no cancer)
- 1 - Benign calcification (no cancer)
- 2 - Benign mass (no cancer)
- 3 - Malignant calcification (cancer present)
- 4 - Malignant mass (cancer present)

Since I only care for detecting cancer, the multi-class labels were reduced to benign(0) and malignant(1)

1.2.1 CV and test data split:

Discussion on the Kaggle prompted that CV data and test data were not correctly split. CV data only contained *masses*, and test data only contained *calcifications*. To fix this issue, I mixed the data from both sources and produced a new CV dataset and a new test dataset containing both masses and calcifications.

1.2.2 data imbalance

The training data used for this project contained fewer positive test cases(7,289) compared to negative(48,596), which is 13% of the training data. The models trained on this data were often overfitting and not producing good weights for our model. To alleviate this issue, every positive test case was augmented. After augmenting the data and reshuffling, the positive test cases with the negative overfitting was no longer an issue. The cross-validation and test data were untouched.

2 Methodology

For the development of this CNN, I used the idea of GoogLetNet. First, I developed a model with the basic idea of the first Inspection layer V1 [21], and then I developed a bigger custom model on the idea of [22].

Both models were trained with Class Weights[3] and data augmentation to deal with the imbalance data(Positive test cases were 13% of the training data).

Class weights were used to increase the recall of the negative test cases, the class weights were calculated as follows, where x represents the number of positive or negative test cases:

$$weight_x = \frac{1}{x} \cdot \frac{\text{total}}{2}$$

Data Augmentation was used to deal with overfitting; every malignant test case was duplicated 4 times, while the benign test cases were only duplicated once.

The details of each model are below.

2.0.1 Model based on V1:

As shown by Figure 2 (model 1). The model starts with a gray-scale image and applies convolutional, max pooling, and normalization layers. The core of the model consists of two sets of V1 Inception modules. After each set of Inception modules, an average pooling layer is applied to reduce the dimensions. The model concludes with a global average pooling layer to further reduce the spatial dimensions, a dropout layer to prevent overfitting, and, in the end, a dense layer for binary classification to detect if cancer is detected.

While this model was good, it was not perfect. Therefore, a newer model was trained to improve on what had already been created.

2.0.2 Model based on V3:

As shown in Figure 2 (model 1). This module starts with two 5x5 Conv layers, followed by the inception V3 modules. These modules are not exactly the same as [22]. These Inception layers have two extra layers: a 5x5 layer and a 7x7 layer made up of three 3x3 size filters, as shown in Figure 1. While not mentioned in the paper[22], there is BatchNormalization before the activation function; this was noticed in the implementation of Inception_v3 in Keras. Modifications were made based on the V2/V3 Inception layers, with an additional drop-out layer in each inception layer to prevent overfitting.

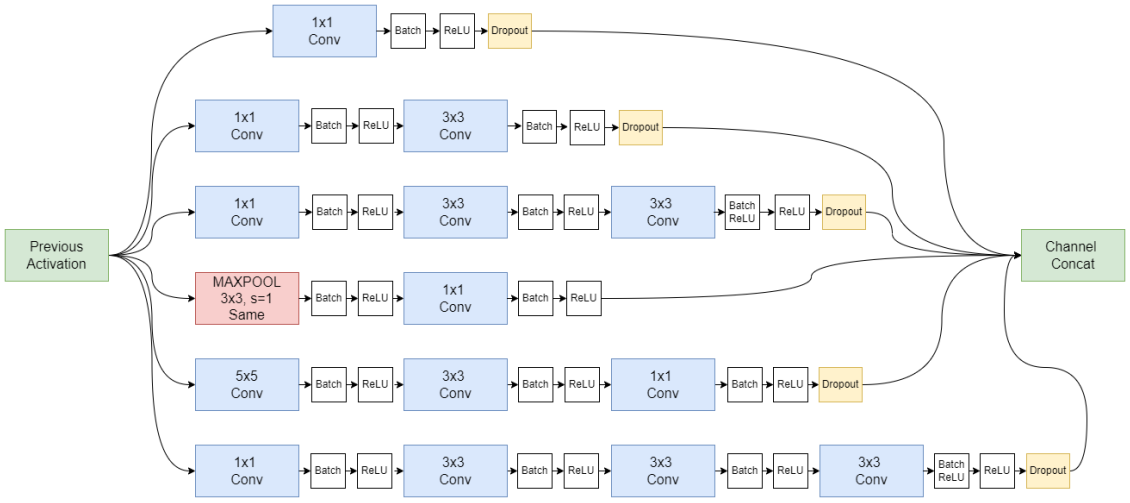

Figure 1: Inception Layer V2/3 with Dropout

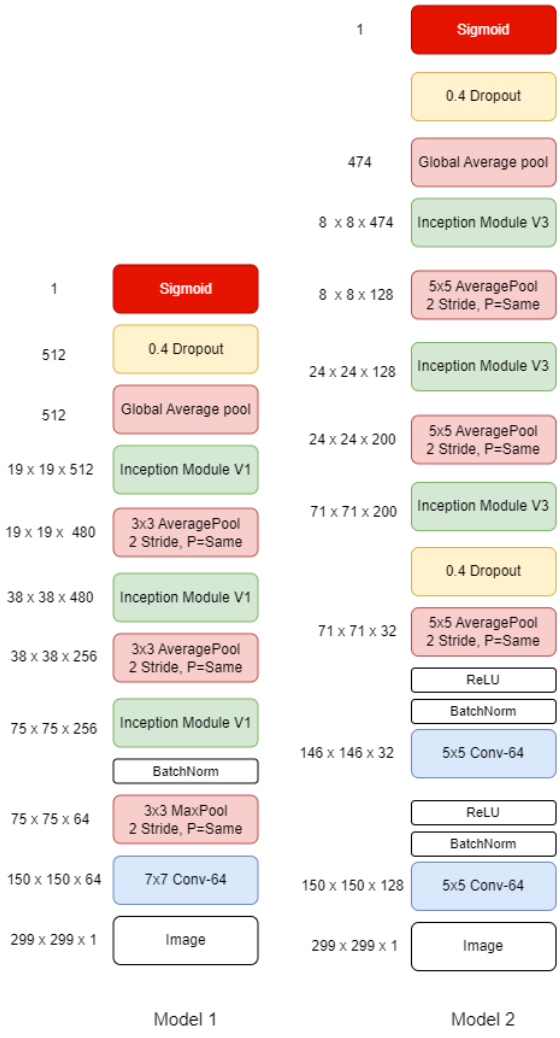

Figure 2: CNN Model Architecture

3 Experimental Setup

Model 1 was trained on a Windows 11 machine with a Razen 7940HS, 8 cores, 16 threads, 64GB of RAM, and a 4060 GPU. Model 1 was trained inside a Docker container distributed by TensorFlow to fully utilize the GPU. Model 2 was trained in Google Colab because of its size; it required a large amount of GPU in order to train the model in a timely manner. The models were created using Keras and Tensorflow [1] with the dataset mentioned in Section 1.2. The data was split when it was posted in Kaggle [12]. The split was split into 11% cross-validation, 11% test, and 78% training.

3.1 Measurement

When dealing with healthcare cancer detection, it is important to always catch cancer when it is present(malignant cases); in other words, it is important to have a high recall, and it is the single most important metric for this project. This means that precision took a hit. The metrics recorded were Loss, recall, precision, and accuracy.

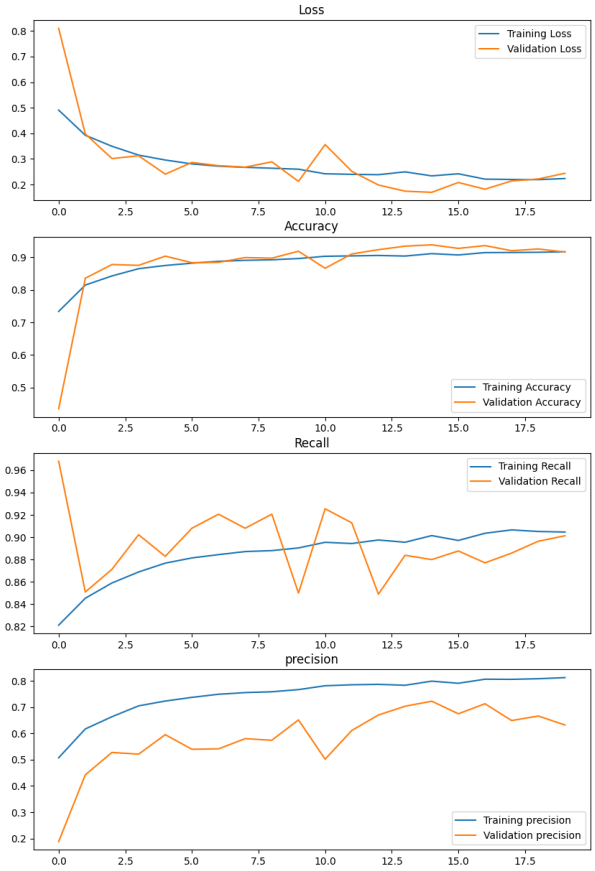

Figure 3: Training Graphs V1

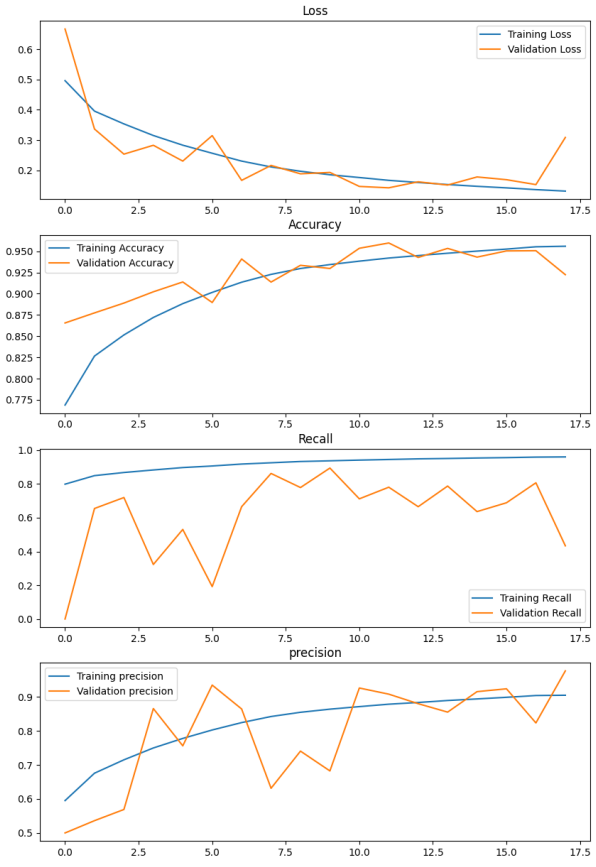

Figure 4: Training Graphs V3

3.2 Results Analysis, Intuitions, and Comparison

From these models, 2 states were saved, as by the following metrics

1. The latest model
2. The model with the lowest loss

Between these 2, the validation results are shown below.

	Model 1 - Best Loss	Model 1 - Latest	Model 2 - Best Loss	Model 2 - Latest
Loss	0.1908	0.2407	0.1422	0.3086
Accuracy	0.9238	0.9143	0.9598	0.9224
Recall	0.8673	0.2739	0.7803	0.4337
Precision	0.6667	0.6232	0.9077	0.9760

Table 1: Model Comparison using validation data

After comparing multiple states of the models against the validation dataset, I select the one with the highest recall (since I care to always detect when cancer is present). Model 1 has the best metrics in terms of recall.

The metrics for this model compared against the test dataset are shown below.

	Model 1 - Best Loss	Model 1 - Latest
Loss	0.1564	0.3626
Accuracy	0.9503	0.9081
Recall	0.8435	0.2739
Precision	0.7807	0.9963

Table 2: Metrics using test Data

4 Conclusion

Early detection of cancer is crucial for patient survival, and mammograms are one of the primary methods used for this purpose. Ensuring accurate detection in these initial tests is vital, especially when cancer is present.

In this project, I have designed and fully trained a convolutional neural network that is capable of detecting cancer with a high sensitivity or recall value. This means that if a patient has cancer, this model is able to detect it with high precision. For those patients that are mislabeled, they can move on to more tests, and later would find out that it is not cancer.

Detecting cancer in an image is good, but there might still be some skepticism, and it will be better to know exactly where cancer is present. Future work can involve semantic segmentation to determine what Convolution Neural network is considered malignant; this can improve the network and provide guidance for the doctors on what to look for.

5 Code contribution

All of the work done in this project was done by Kevin Lopez (me). The code is located on GitHub [18].

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org/`, 2015.

[2] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, 2016.

[3] TensorFlow Authors. Imbalanced data. `https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#class_weights`, 2023. Accessed: 2024-05-13.

[4] Anton S Becker, Magda Marcon, Soleen Ghafoor, Moritz C Wurnig, Thomas Frauenfelder, and Andreas Boss. Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative Radiology*, 52(7):434–440, 2017.

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[6] Amanda Demetri-Lewis, Priscilla J. Slanetz, and Ronald L. Eisenberg. Breast calcifications: The focal group. *American Journal of Roentgenology*, 198(4):W325–W343, 2012. PMID: 22451569.

[7] C. DeSantis, J. Ma, L. Bryan, and A. Jemal. Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 64:52–62, 2014.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[9] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[10] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer. The digital database for screening mammography. In M.J. Yaffe, editor, *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001.

[11] N. Houssami, M. N. Cheung, and J. M. Dixon. Fibroadenoma of the breast. *The Medical Journal of Australia*, 174(4):185–188, Feb 2001.

[12] Scott Kooch. Ddsm mammography. `https://www.kaggle.com/datasets/skooch/ddsm-mammography`, Accessed 2024.

[13] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[14] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of ddsm, 2016.

[15] Constance D. Lehman, Robert F. Arao, Brian L. Sprague, Janie M. Lee, Diana S. M. Buist, Karla Kerlikowske, Louise M. Henderson, Tracy Onega, Anna N. A. Tosteson, Garth H. Rauscher, and Diana L. Miglioretti. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49–58, 2017.

[16] Daniel Levy and Anil K Jain. Breast mass detection in mammography using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*, 2016.

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[18] Kevin Lopez. Mammographyai. `https://github.com/KevinDLopez/MammographyAI`, 2024. Accessed: 2024-05-13.

[19] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[20] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan,

Thomas Mertelmeier, Matthew G Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M Mann, and Ioannis Sechopoulos. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 03 2019.

[21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.