Breast Cancer Detection: A CNN-Based Classification of Breast Histopathology Images

Margaret Njenga
24846577
margaret.n.njenga@stu.mmu.ac.uk

I. Introduction

Breast cancer is one of the most common cancers among women worldwide. According to an article by the WHO, in 2022, 2.3 million women were diagnosed with breast cancer globally, and among them, 670,000 died from it.[1] There is no known cause of breast cancer, but certain factors increase its risk. Some of these factors include age, obesity, alcohol, and history of radiation exposure. A breast lump is one of the most commonly recognized signs of breast cancer, and women are often encouraged to perform self-examinations to detect lumps early. Not all breast lumps are cancerous, hence the need to analyse images to distinguish between benign (non-cancerous) and malignant (cancerous). Other signs may include a change in nipple appearance, change in size, shape, or appearance of the breast. These signs and symptoms start developing in later stages, and a person may not experience any signs or symptoms until the tumours have metastasized.[1]

Early and accurate detection is very crucial for effective treatment and improved survival rates. Doing this manually for millions of patients can be expensive and time-consuming. It is also prone to human error. Adopting models such as CNNs proves to be efficient and accurate.

This project aims to leverage deep learning techniques, particularly custom CNNs and modern CNNs, to develop an automated image classification system that can accurately distinguish between benign and malignant breast tissues.

Histopathological image analysis, which involves examining tissue samples under a microscope, is a key method for diagnosing breast cancer. The input to the model is a colored RGB histopathological image of breast tissue with 3 channels. The input is trained on a convolutional neural network to output a predicted cancer type. The output is binary, where 0 is benign(non-cancerous) and 1 is malignant (cancerous).

II. Dataset

A. Data Description

Some ways to detect breast cancer are to use mammograms, ultrasound, biopsy, and histopathological image analysis. All these involve examining images of breast tissues. The Breast Cancer Histopathological Database (Break His) is extracted from Kaggle online datasets.[1] It comprises 7,909 images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). The dataset contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). [2]

It has been organized to train, validate, and test data, with each containing benign and malignant images with different magnifying factors. The training data has 5,536 samples, which is 70% of the entire dataset. The validation data has 1,186 samples, which is almost equal to the test data, which has 1,187 samples, both contributing to 30% of the dataset.

B. Data Preprocessing and Augmentation

The original data has varying image sizes. CNNs require fixed input sizes for batch training to work efficiently. The images are resized to 128 x 128 when building the CNN from scratch, to speed up training, then later on changed to 224 x 224, which is the standard size when using pretrained models on ImageNet such as ResNet18 and VGG16. The images are cropped 80% to 100% of the original image so that meaningful features, such as cancer cells, which are quite small, are not cut out.

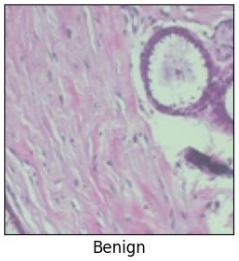
A random horizontal flip is also applied to flip images left and right with a 50% chance to teach the model to pay attention to small variations.

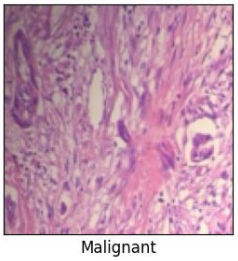
The images are converted to PyTorch's image format then to Tensor and then normalisation is applied. For normalisation, the images are normalized using the mean and standard deviation of the ImageNet dataset. This helps the model generalize better, especially when using pretrained models like ResNet18 and VGG16.

Data Augmentation is not performed on the test and validation data to avoid introducing artificial variations that could affect result evaluation. The only transformations done on the test and validation data are resizing and normalization to match the training data.

Mini-Batch Data Loading

After defining the augmentations for the images, the data is read, and transformations are applied. The data contains 7,909 images, and training on one image at a time or all images at once is computationally expensive. So the model is trained on a small group(batch) of images at a time. This reduces memory load and improves efficiency.


Benign


Malignant

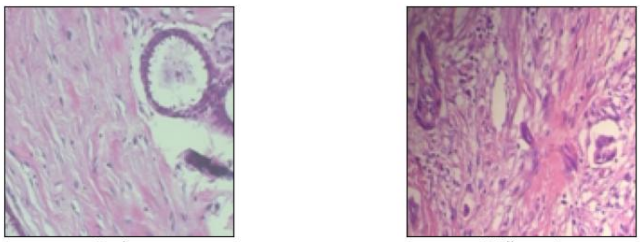


Fig. 1. Sample images without augmentation(above), sample images with augmentation(below)

III. Methodology

A. Model

To classify whether a breast tissue image is benign or malignant, the features of the images are extracted from the training data that contains ground truth data of whether an image shows benign or malignant cancer types. This is a convolution. Convolutional Neural Networks have been widely used in image classification and computer vision to make predictions. This project leverages custom CNNs and modern pretrained CNNs to develop a model that accurately classifies benign and malignant breast tissues.

During augmentation, images are converted to tensors. In a two-dimensional convolutional layer, the convolutional layer takes an input tensor and a kernel tensor and combines them to produce an output tensor through cross-correlation. [3]The challenge that arises once this is done is that the output becomes smaller. To resolve this issue, the input image is padded with 0 along the borders so there is enough space to shift the kernels. During convolution, the kernel moves one step at a time, which is the stride. For an input with multiple channels, such as a breast tissue image, that has 3 RGB channels, the number of kernels needs to be the same as the number of channels. ReLU introduces non-linearity, allowing the network to learn complex patterns. Pooling subsamples the pixels and keeps the most important pixels. It reduces the size and computation and improves spatial invariants.

The layers are stacked to include different batches of features. A fully connected layer expects a 1D matrix, so the final layer is converted to a 1D matrix. This is where the features are combined to make predictions and the outputs a prediction of the probability of an input image being benign(0) or malignant(1).

ResNet18 is a pretrained model on ImageNet1K and was used to compare performance to the custom CNN.

B. Loss

The CNN model was trained on the Binary Cross Entropy loss with logits, which combines sigmoid and Binary Cross Entropy Loss. Since this is a binary classification problem, the BCEWithLogitsLoss was chosen over CrossEntropyLoss.

The loss function compares the output from the model to the true labels of the images. It tells the model how wrong it is. A higher loss shows that the model performs poorly and vice versa. The goal is to minimize loss. When trained on 10 epochs, the CNN model achieved a minimum train loss of 0.22 and a corresponding minimum validation loss of 0.30. The ResNet18 pretrained model was trained on 20 epochs and achieved a minimum loss of 0.009 at 17 epochs.

C. Optimization

The Optimizer is the algorithm that updates the model's weights to reduce the loss. The concept behind optimization is that the model makes a prediction, the loss is calculated, and the optimizer adjusts the model weights to make the next prediction better. The optimizer used in this project is the Adam optimizer, as it performed better than the Stochastic Gradient Descent.

IV. Experiment

A. Evaluation Metrics

The results of both the custom CNN and the ResNet18 pretrained model are evaluated on the same evaluation metrics to ensure consistency and to better compare the performance of both models.

Confusion Matrix

A confusion matrix is commonly used to evaluate the performance of classifiers. The idea behind it is to count the number of times instances of one class are classified as the other class. [4] Each row of the confusion matrix represents an actual class, while each column represents a predicted class.

The first row of the confusion matrix shows that the model predicted 308 benign images correctly as benign and classified 64 benign images as malignant. The second row shows that the model predicted 86 malignant images as benign and 729 malignant images as malignant.

Precision

Having a model that classifies malignant images as benign can give false hope to say, a patient who has been told they have a benign tumour but actually has a malignant tumour. Precision calculates the accuracy of the positive predictions. It is calculated as

$$Precision = TP / (TP + FP))$$

The precision of the custom CNN is 0.92. About 92% of the samples predicted as malignant are malignant. Precision is typically used with Recall.

Recall

It is the ratio of positive instances that are correctly classified by the model. It is calculated as

$$Recall = TP / (TP + FN)$$

The recall score is 0.89, which means that 89% of all the malignant samples were predicted correctly.

F1 Score

The final metric used is the F1 score which is the harmonic mean of precision and recall. The score is 0.91

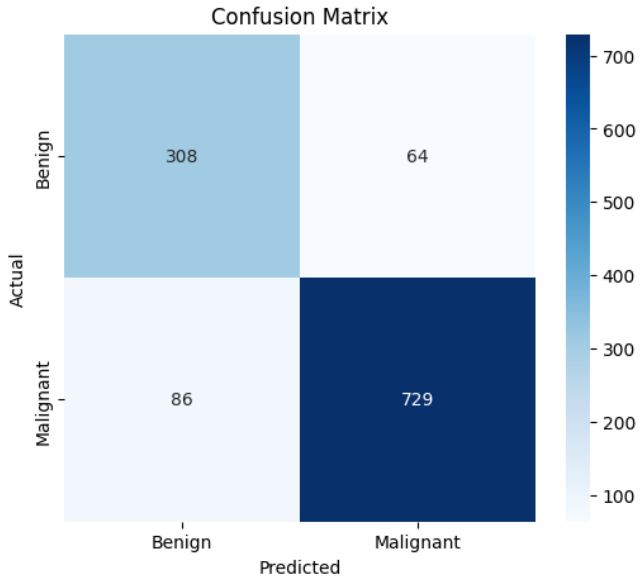


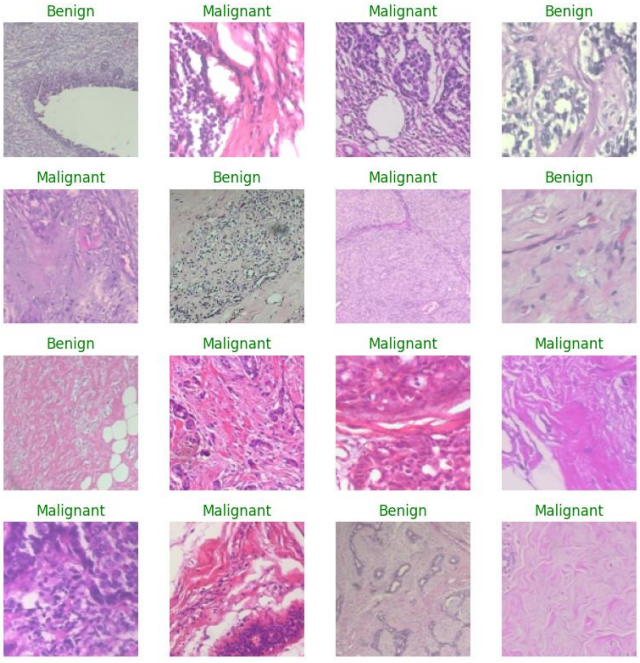
Fig. 2. Visual Confusion Matrix of the custom CNN model

B. *Quantitative Results*

TABLE I.

Metric	Custom CNN
Accuracy	87.36%
Precision	91.93%
Recall	89.45%
F1 score	90.67%

C. *Qualitative Results*



The model made almost perfect predictions of the benign and malignant image samples. The code was run thrice to shuffle the data and the model still showed correct labels. From these results, there is a distinct difference between benign and malignant breast tissue samples. Benign tissue samples are less pigmented and show some structure while the malignant tissue samples are more pigmented.

D. *Comparisons*

METRIC	CUSTOM CNN	RESNET18
ACCURACY	87.36%	96.38%
PRECISION	91.93%	95.51%
RECALL	89.45%	98.28%
F1 SCORE	90.67%	97.39%

The ResNet18 pretrained model outperforms the custom CNN model and achieves the best recall score of 98%.

REFERENCES

[1] World Health Organization, "Breast cancer," World Health Organization, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Bukun, "BreakHis-BreastCancerHistopathological Database," Kaggle, 2020.[Online].Available:https://www.kaggle.com/datasets/ambarish/breakhis/data

[3] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into Deep Learning", 1st ed. Cambridge: Cambridge University Press, 2021. [Online]. Available: https://d2l.ai/chapter_convolutional-neural-networks/conv-layer.html

[4] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.