Breast Cancer Image Classification: Impact of Limited Data Example

Michael L Brewer

University of Arkansas at Little Rock  
Little Rock, AR, USA

[mlbrewer@ualr.edu](mailto:mlbrewer@ualr.edu)

https://orcid.org/0000-0002-1455-3162

**ABSTRACT**

**Keywords**

Breast Cancer; CNN; Data Augmentation; Explainable AI

# INTRODUCTION

# Model and Training

The base model, defined in code as base\_model, was constructed from EfficientNetB2 [1]. This model was then enhanced with the following additional layers

model = tf.keras.Sequential([

    resize\_and\_rescale,

    layers.RandomFlip("horizontal\_and\_vertical"),

    layers.RandomRotation(0.2),

    layers.RandomContrast(0.2),

    layers.RandomZoom(0.2),

    base\_model,

    layers.GlobalAveragePooling2D(),

    layers.BatchNormalization(),

    layers.Dropout(.2, name="top\_dropout"),

    layers.Dense(num\_classes, activation='softmax')

])

The first layer simply resizes input images to the expected size (260,260) and converts the anticipated 8-bit images range to [0,1]. This layer is followed by augmentation layers to flip, rotate, adjust contrast, and zoom, respectively.

Then, after the base\_model has been included, additional layers are included to perform

* global average pooling to downsample,
* batch normalization to improve convergence,
* dropput at a rate of .2, and finally
* classification of the three classes using softmax.

Because of these augmentation layers built into the model itself, each of our training attempts will use augmented data. However, the approached described in Section 3 will use additional augmentation prior to the training. Note the aforementioned augmentation, done within the model, does not affect the epoch size nor the affect the balance of the dataset. The additional augmentation discussed below is done using the sample function within an ImageDataGenerator. That sampling is done prior to training and does impact the epoch size and (if desired) the balance of the training dataset.

All training was done using Google Colaboratory, a web-based development environment. See [2]. Training was done with a T4 GPU.

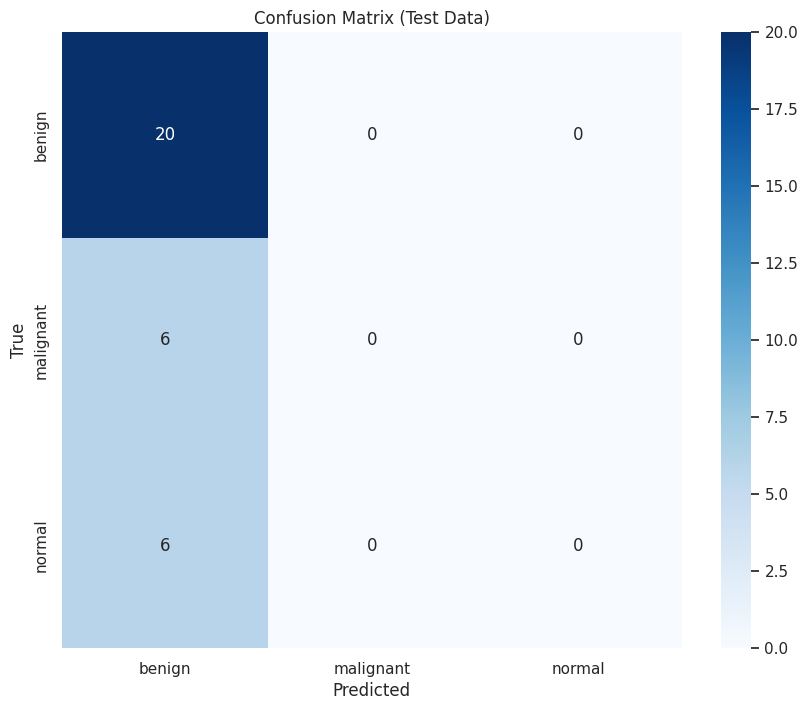
# Training Data

The data used for this is available at Kaggle <https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset> . This dataset contains 780 ultrasound images from 600 patients. These images are classified as “normal”, “benign”, or “malignant” based on a professional classification of the existence and type of breast cancer tumors. See [2]. The image classifications are imbalanced with the following distribution:

* 133 “normal” images
* 437 “benign” images
* 210 “malignant” images

## No Data Augmentation

Using the most naïve approach of no data augmentation, the derived model architecture, when trained, simply learned to classify every image as “benign”, due to the imbalance. On the test set, this resulted in 62.5% accuracy.



## Data Augmentation to Balance

The second approach to data augmentation performed data augmentation to generate a dataset where each class was balanced at the maximum of the original classification counts, from within the training set. In the training set, originally the ‘benign’ classification contained 353 images, so the images for ‘normal’ and ‘malignant’ classifications were augmented to also reach 353.

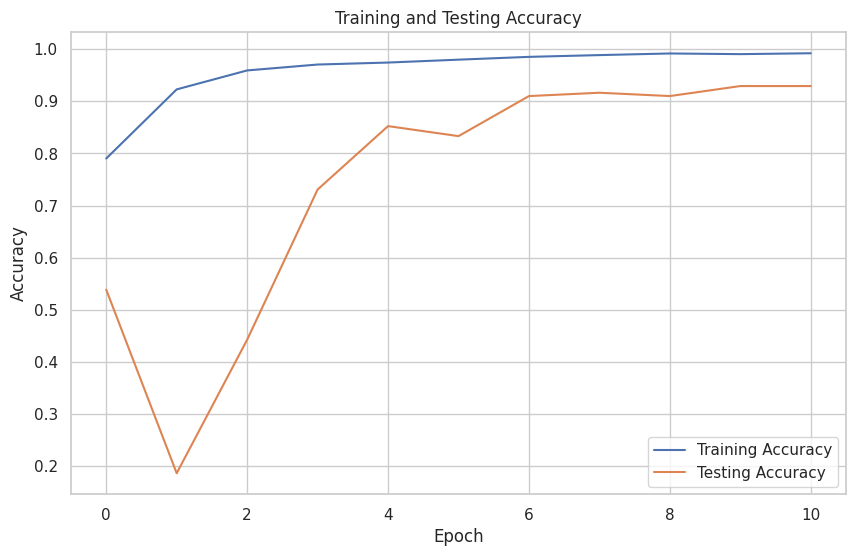
This approach similarly failed to produce a usable model. In fact, the trained model was worse, in terms of accuracy, classifying all images as ‘normal’. Note that because the augmentation was done after the train/test split of the data, the test data is not balanced. This approach resulted in 31.25% accuracy.

A diagram of a graph

Description automatically generated with medium confidence

## Data Augmentation to Enhance

For the 3rd approach, we augment the training data with an additional 2,000 images, per classification type, beyond the balanced dataset used for the 2nd model. The resulting trained model produced a potentially viable. The accuracy as a function o the training accuracy is shown below with the final model, the best model at epoch 8, providing an accuracy of 90.62% on the test set.



The associated confusion matrix is shown below.

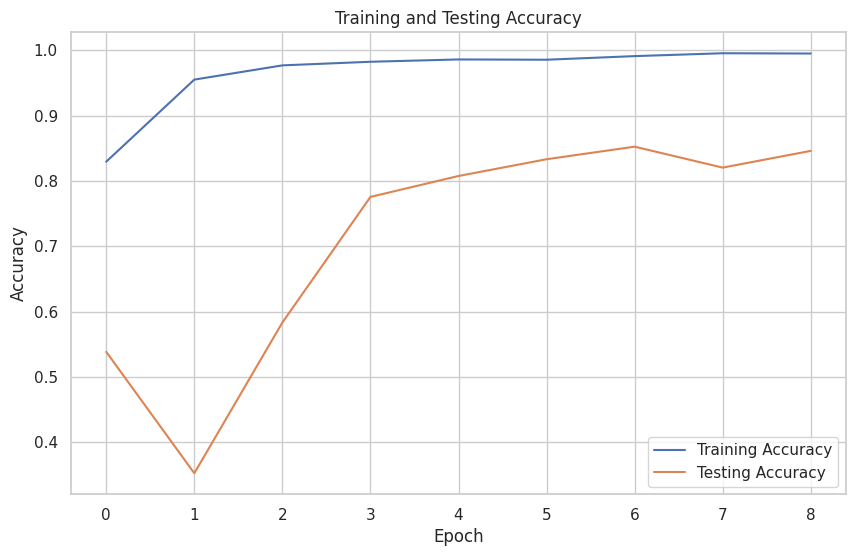
A blue squares with white text

Description automatically generated

## Artificially Reduced Malignant Class

For the fourth approach, to mimic a very limited training set for one class we artificially reduce our ‘malignant’ training data from 167 images (after train/test split) to 33 images. Aftwards though, we use our previously described augmentation method to again provide 2,000 images over our re-balanced set of images. This results again I each class having 2,353 (ie, 2,000 images more than the maximum classification, the ‘benign’ class, which had 353 images).

For this training, the accuracy as a function of the training epochs is given below, with the best model (epoch 6) having an accuracy of 78.12% on the test set.



The confusion matrix is given below:

A blue squares with white text

Description automatically generated

Table . Table captions should be placed above the table

|  |  |  |  |
| --- | --- | --- | --- |
| **Graphics** | **Top** | **In-between** | **Bottom** |
| Tables | End | Last | First |
| Figures | Good | Similar | Very well |

## References and Citations

Footnotes should be Times New Roman 9-point, and justified to the full width of the column.

Use the “ACM Reference format” for references – that is, a numbered list at the end of the article, ordered alphabetically and formatted accordingly. See examples of some typical reference types, in the new “ACM Reference format”, at the end of this document. Within this template, use the style named *references* for the text. Acceptable abbreviations, for journal names, can be found here: <http://library.caltech.edu/reference/abbreviations/>. Word may try to automatically ‘underline’ hotlinks in your references, the correct style is NO underlining.

The references are also in 9 pt., but that section (see Section 7) is ragged right. References should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e. you can give the address to obtain the report within your citation) and may be obtained by any reader. Proprietary information may not be cited. Private communications should be acknowledged, not referenced (e.g., “[Robertson, personal communication]”).

## Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

# FIGURES/CAPTIONS

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7”).

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., “Table 1” or “Figure 2”), please note that the word for Table and Figure are spelled out. Figure’s captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

# SECTIONS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.



Figure 1. Insert caption to place caption below figure.

.

## Subsections

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

# ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

# REFERENCES

|  |  |
| --- | --- |
| [1] | M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *CoRR,* 2019. |
| [2] | G. Al-Dhabyani W, "Dataset of breast ultrasound images. Data in Brief," 2020. |