Breast Cancer Classification

Arefin Shamsil, Hagop Pon Dikian, Kalyn Aguiar and Nestor Baptista

Presentation Overview

1. Introduction

Motivation and Objectives

2. Data Curation

Collection, Exploration, Analytics, and Hosting

3. Deep Learning Models

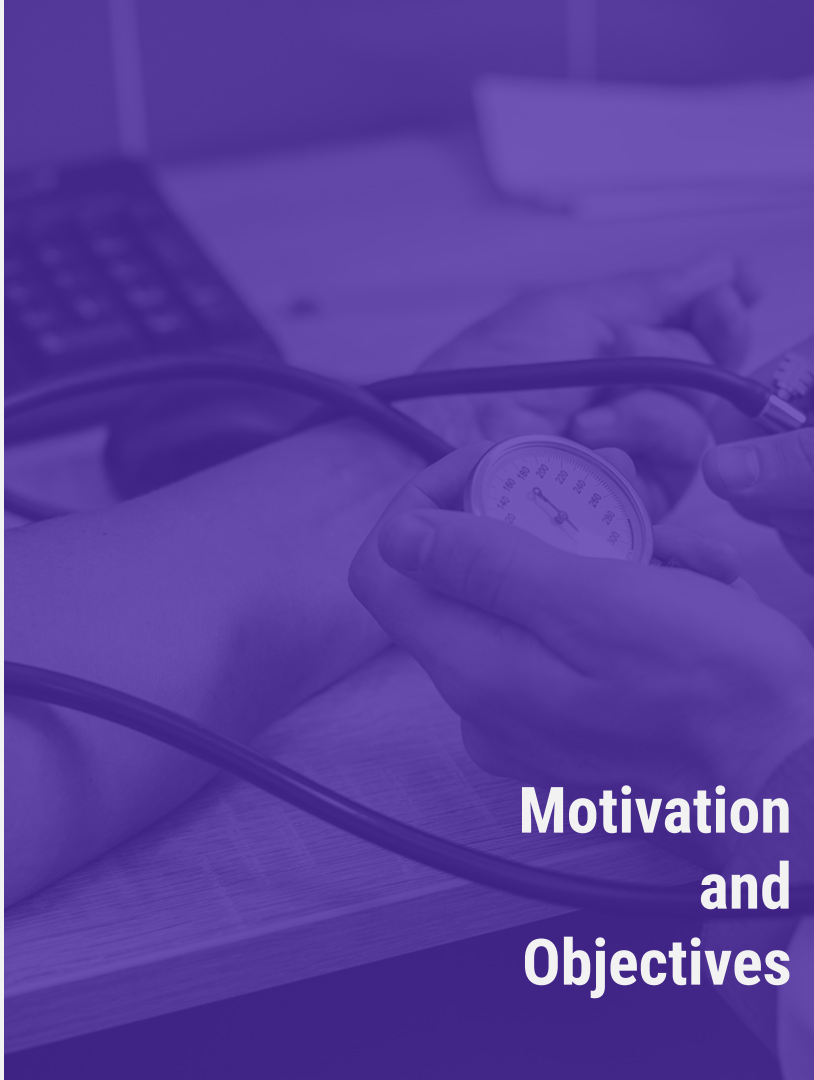
Cancer classification, training, and prediction

4. Concluding Remarks

Summary, Future Work

01

Introduction:



Motivation and Objectives

Breast cancer is the most common cancer diagnosed in women, accounting for more than 1 in 10 new cancer diagnoses each year. It is the second most common cause of death from cancer among women in the world

(Alkabban and Ferguson, Breast Cancer, PubMed - 2022)

Motivation and Objectives

Using Deep learning models to predict classifications of breast cancer through images

How do the images themselves help with classification? What other elements are needed to confirm?

Healthcare issues. Breast cancer one of the most deadliest forms of cancer in recent years

Being able to classify characteristics through images.

Breast Cancer statistical information in the United States from 2020

Analytics

Breast Cancer Statistics for 2020 in the United States

State with highest amount of new cases: California (25, 809)

State with lowest amount of new cases: Wyoming (411)

Average Amount of New Cases in 2020: 4,836

Total Amount of New Cases in 2020: 241,778

Average Number of Recorded Deaths in 2020: 821

Total Amount of Recorded Deaths in 2020: 42,715

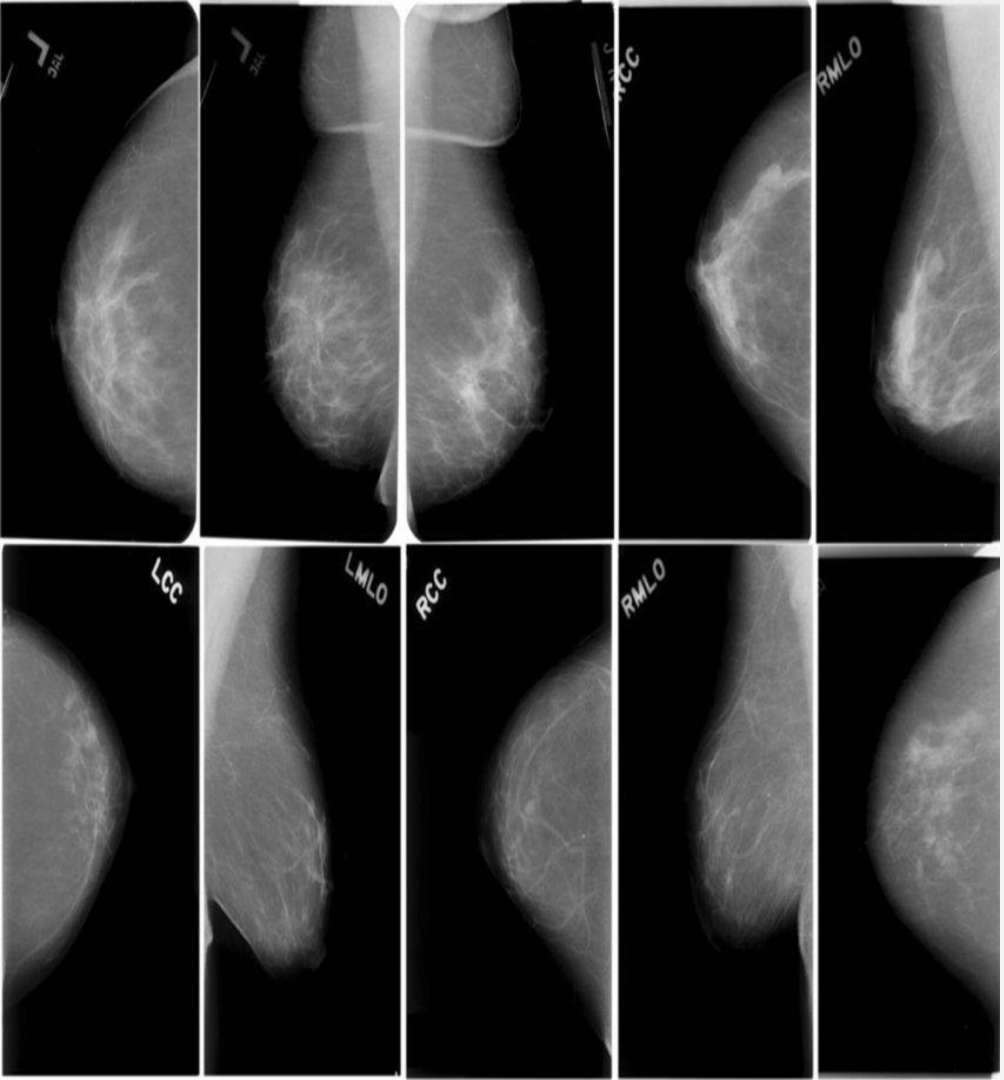
State with the highest recorded Deaths: California (4,520)

State with the lowest recorded Deaths: Alaska (59)

02

Data Curation:

Collection, Exploration, Analytics, and Hosting



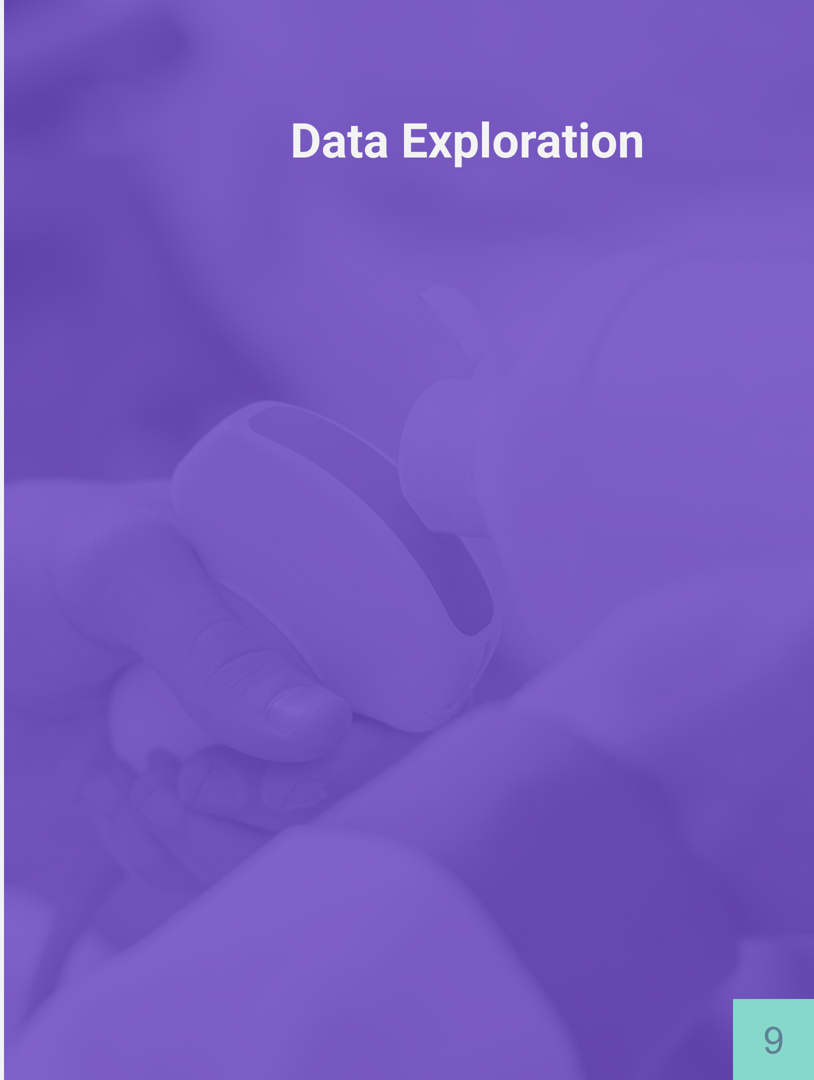
Data Collection

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin (2016)

The Digital Database for Screening Mammography (DDSM) is a database of 2,620 scanned film mammography studies. It contains normal (BWC), benign, and malignant cases with verified pathology information.

Published research results are difficult to replicate due to the lack of a standard evaluation. Most computer-aided diagnosis (CADx) and detection (CADe) algorithms for breast cancer in mammography are evaluated on private data sets

Data Exploration



Three types of image classifications for classifying breast cancer:

Benign

Non-Cancerous

Benign Without Callback

Without callback - the mass or calcification was noteworthy in the eyes of the physician, but patient not at risk of cancer and no follow up needed at this time

Malignant

Cancerous

Data Exploration

Breast Mass Tumor:

A breast tumor refers to an abnormal mass or lump that forms within the breast tissue. Common types of benign breast tumors include fibroadenomas and cysts.

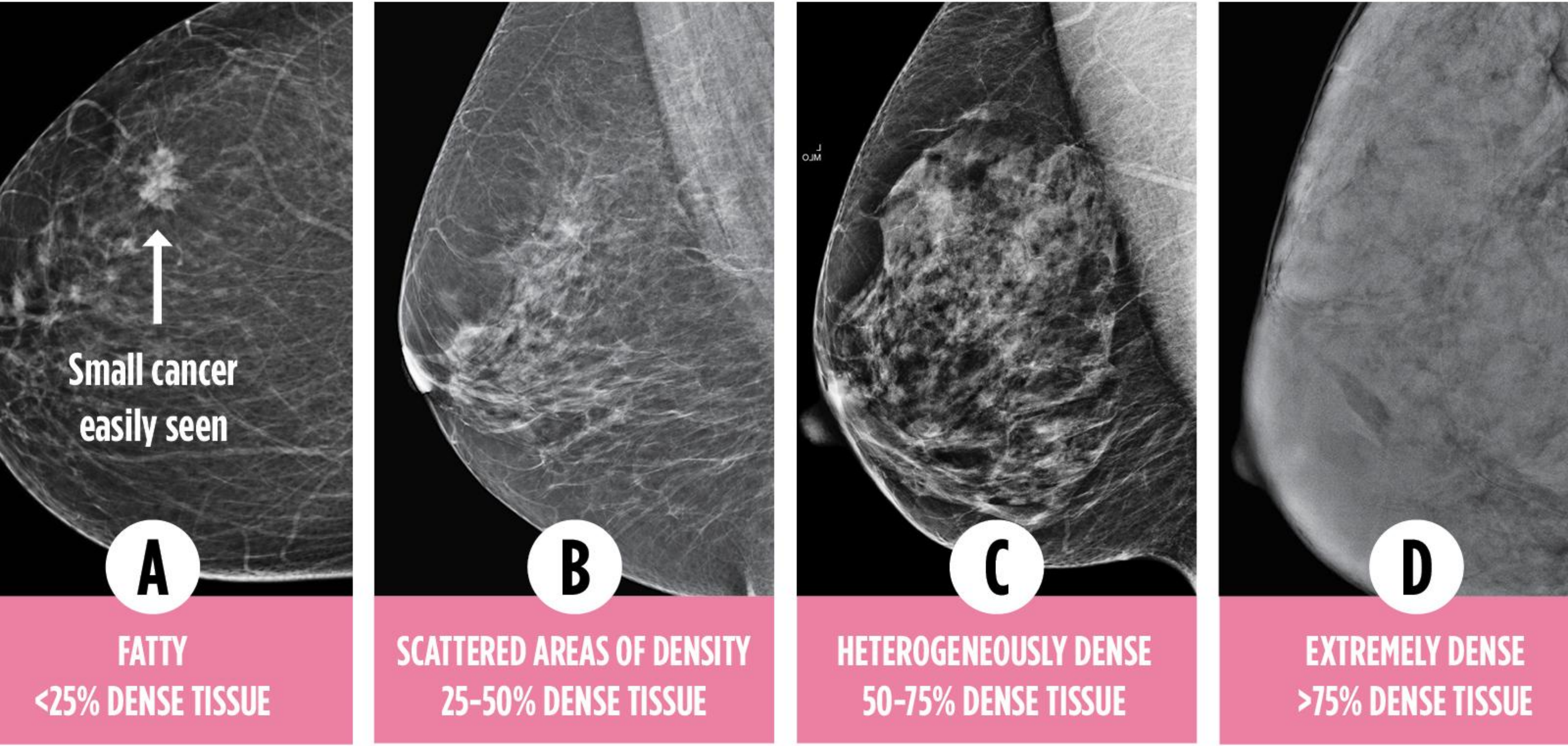
Calcification:

Calcifications in the breast occur when calcium salts build up in the breast tissue, leading to the formation of tiny calcium deposits. It is common for woman to present this after the age of 50

	Calcification DF	Mass DF
patient_id	891	753
pathology	MALIGNANT 673 BENIGN 658 BWC 541	MALIGNANT 784 BENIGN 771 BWC 141
images	1872	1696

Data Exploration

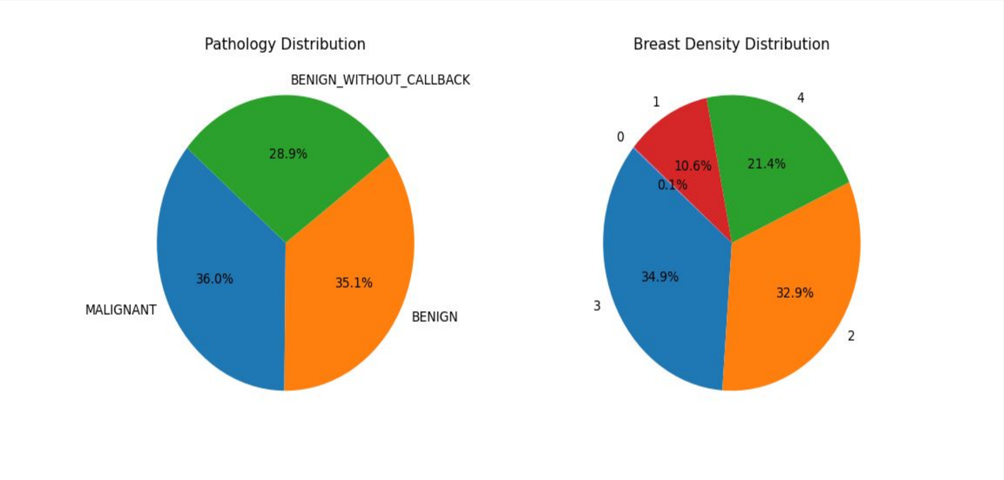
breast_density	Women with dense breast tissue have a higher risk of developing breast cancer than women with little or no dense breast tissue.
left or right breast	The left breast is 5 - 10% more likely to develop cancer than the right breast.
image view	Mediolateral oblique (MLO) view and cranial caudal (CC). Center view and bottom to top view
calc type / calc distribution	Calcifications may appear as bright white spots on mammograms. You can't feel them from the outside
mass shape / mass margins	Circumscribed oval and round masses are usually benign. An irregular shape suggests a greater likelihood of malignancy.



Mammograms miss about 40% of breast cancers in the densest breasts. (Breast Cancer Institute, 2018)

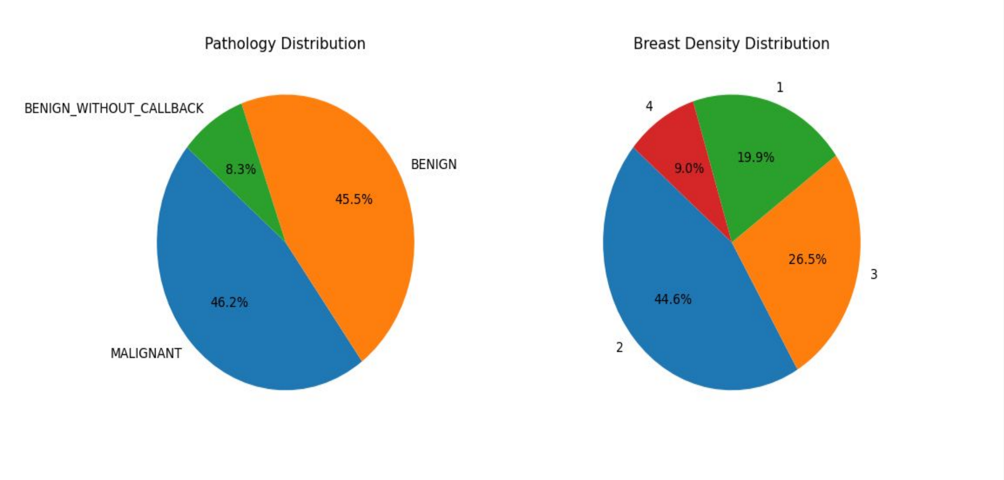
Data Exploration

CALC



As shown in multiple researches, it is easier to identify whether a calcification is Benign or Malignant. Our dataset also shows this. Calcification occurs in all breast densities with more likeness in the densest breasts.

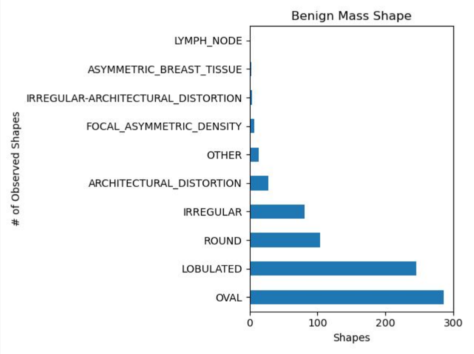
MASS



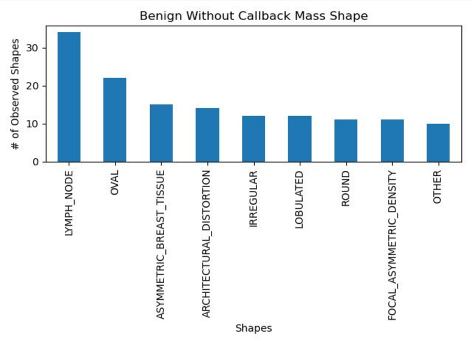
Almost always having a visible mass requires further investigation. That is why there is a small percentage of BWC cases. The density distribution makes the case that is easier to identify the mass with less dense breasts

Exploratory Data Analysis

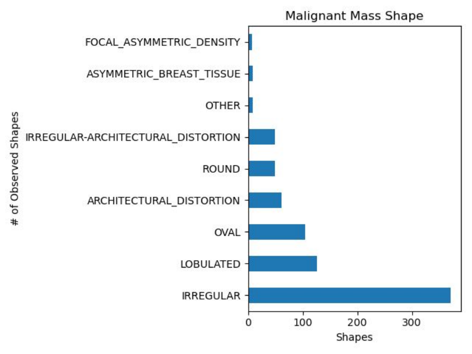
01



02



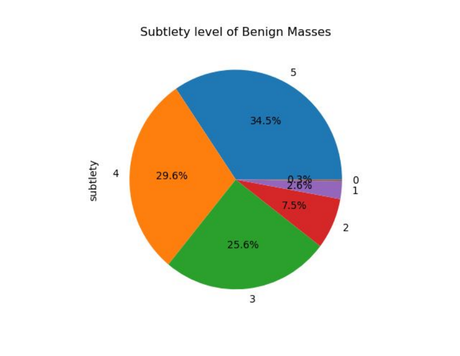
03

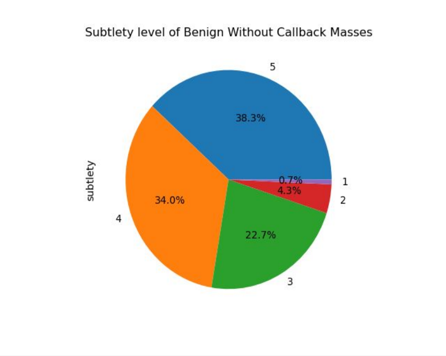


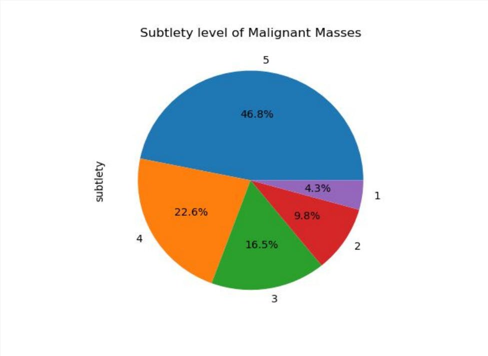
Most seen mass shape
Benign: Oval
Benign without callback: Lymph Node
Malignant: Irregular

Clear distinction between shapes for all three classifications

Easier to reliably classify cancer classification of masses







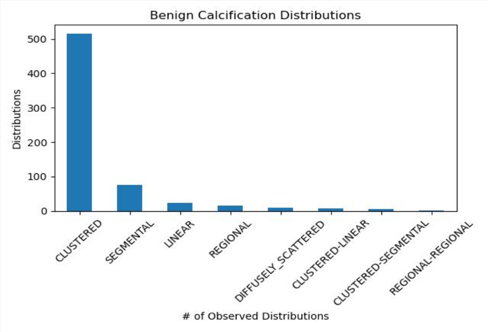
Exploratory Data Analysis

Most Frequent Subtlety Level of Masses
Benign: 5 (34.5%)
Benign without callback: 5 (38.3%)
Malignant: 5 (46.8%)

Subtlety Level on a scale from 1 to 5. 5 being the most unsubtle. Masses are noticeable on the images.

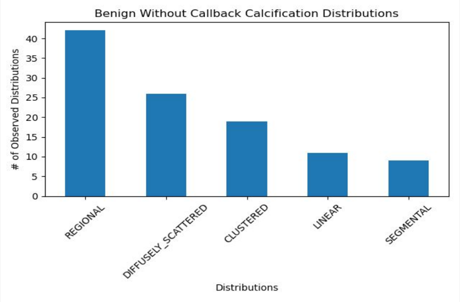
Easier for classifications to be interpreted from image.

01



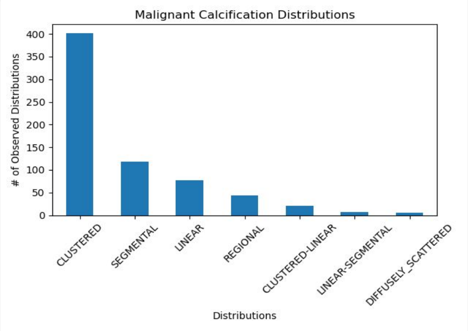
Exploratory Data Analysis

02



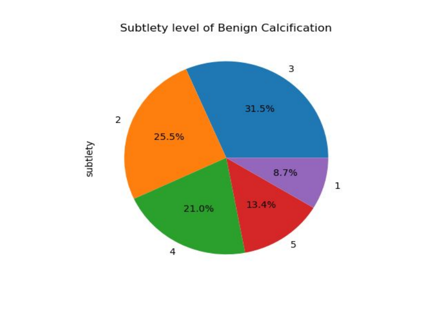
Most seen Calcification Distribution
Benign: Clustered
Benign without callback: Regional
Malignant: Clustered

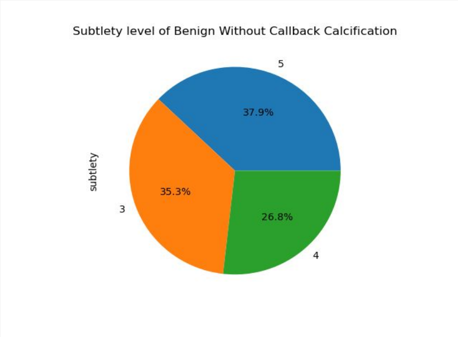
03

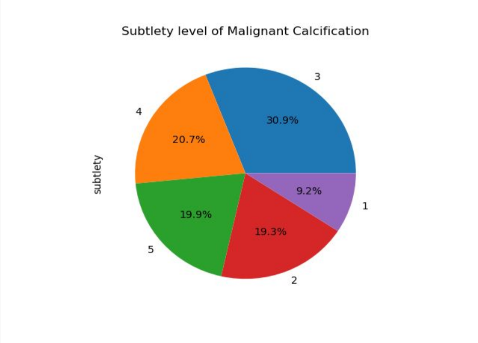


Benign and Malignant share the same most seen calcification distribution. Clustered

Could lead to images misinterpreting due to characteristics being very similar







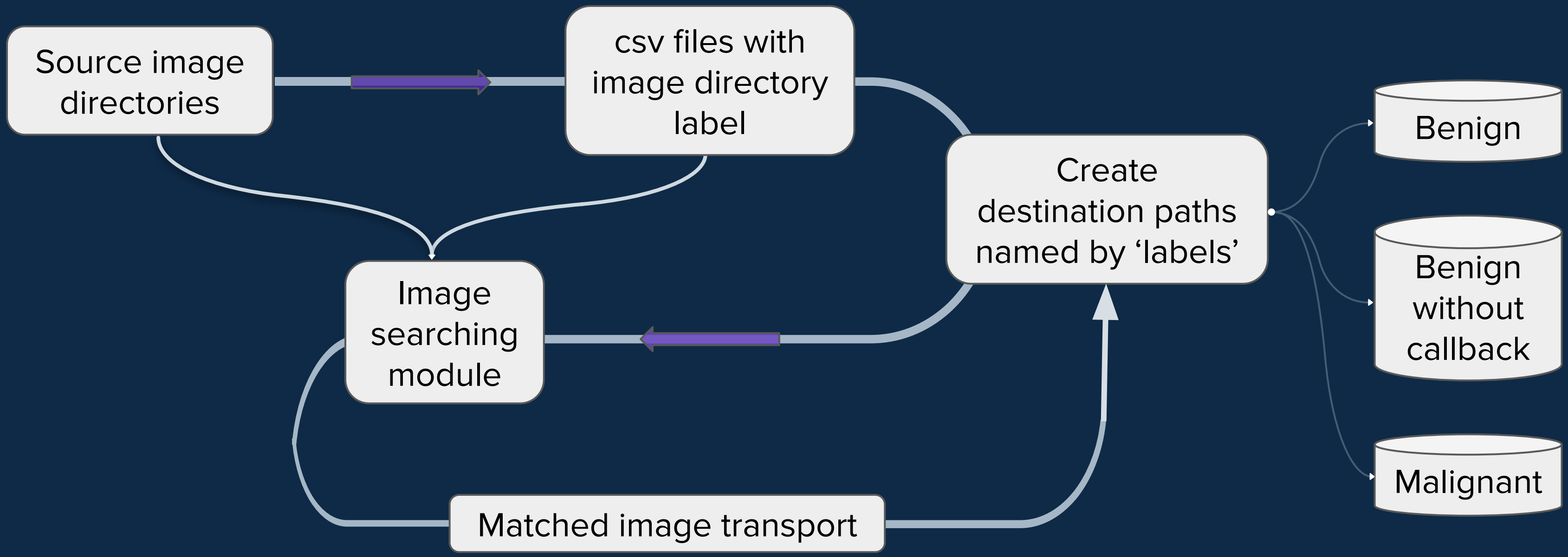
Exploratory Data Analysis

Most Frequent Subtlety Level of Calcifications
Benign: 3 (31.5%)
Benign without callback: 5 (37.9%)
Malignant: 3 (30.9%)

Both Benign and Malignant share the same most frequent subtlety level of 3. Further supports images potentially being misinterpreted.

Subtlety levels seems fairly well distributed across calcifications compared to masses.

Image Data Cleaning Pipeline



Data Hosting

Initially Hosted on Google Colab

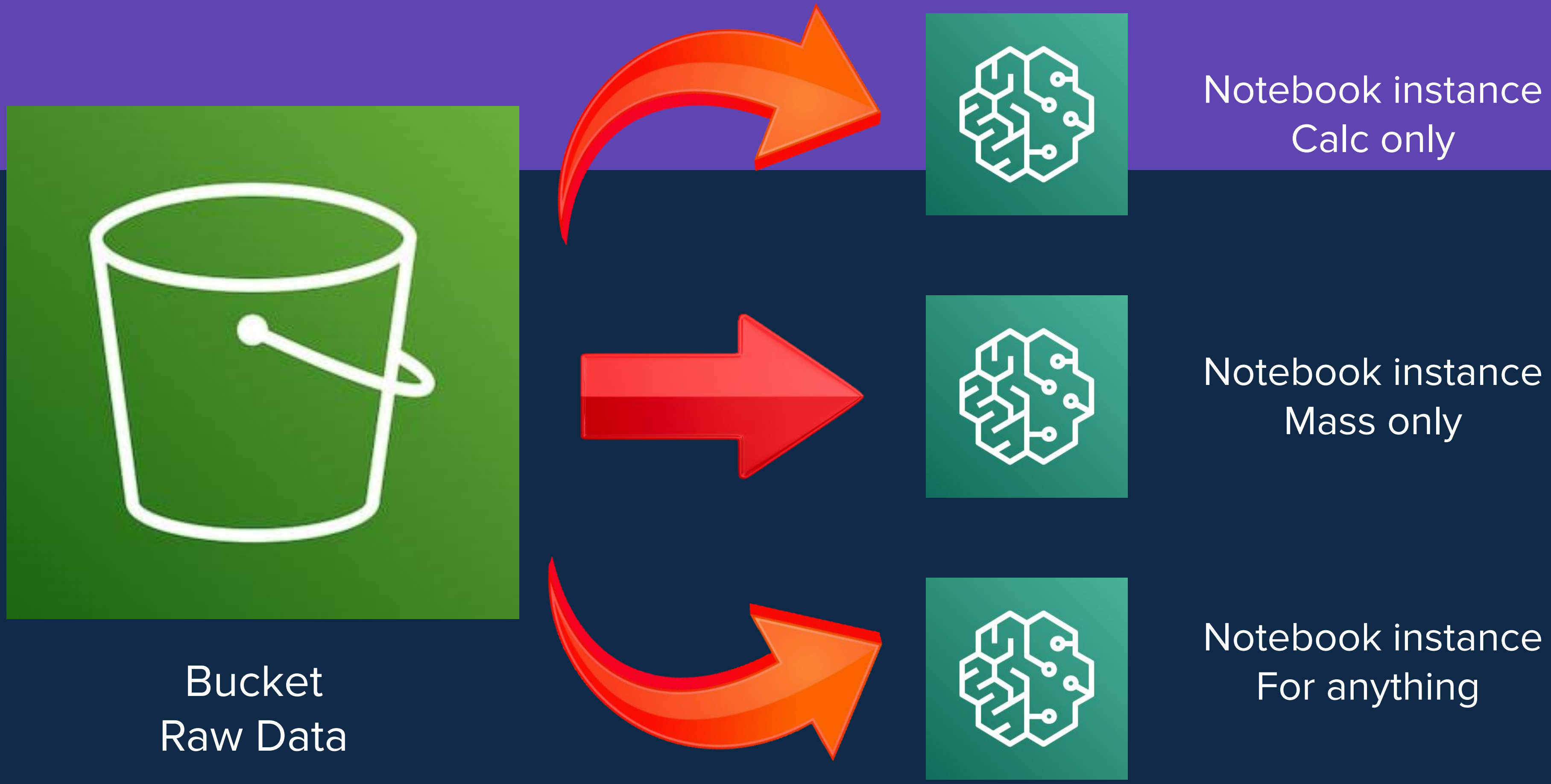
Data, images, and models initially hosted on google colab to test functionality

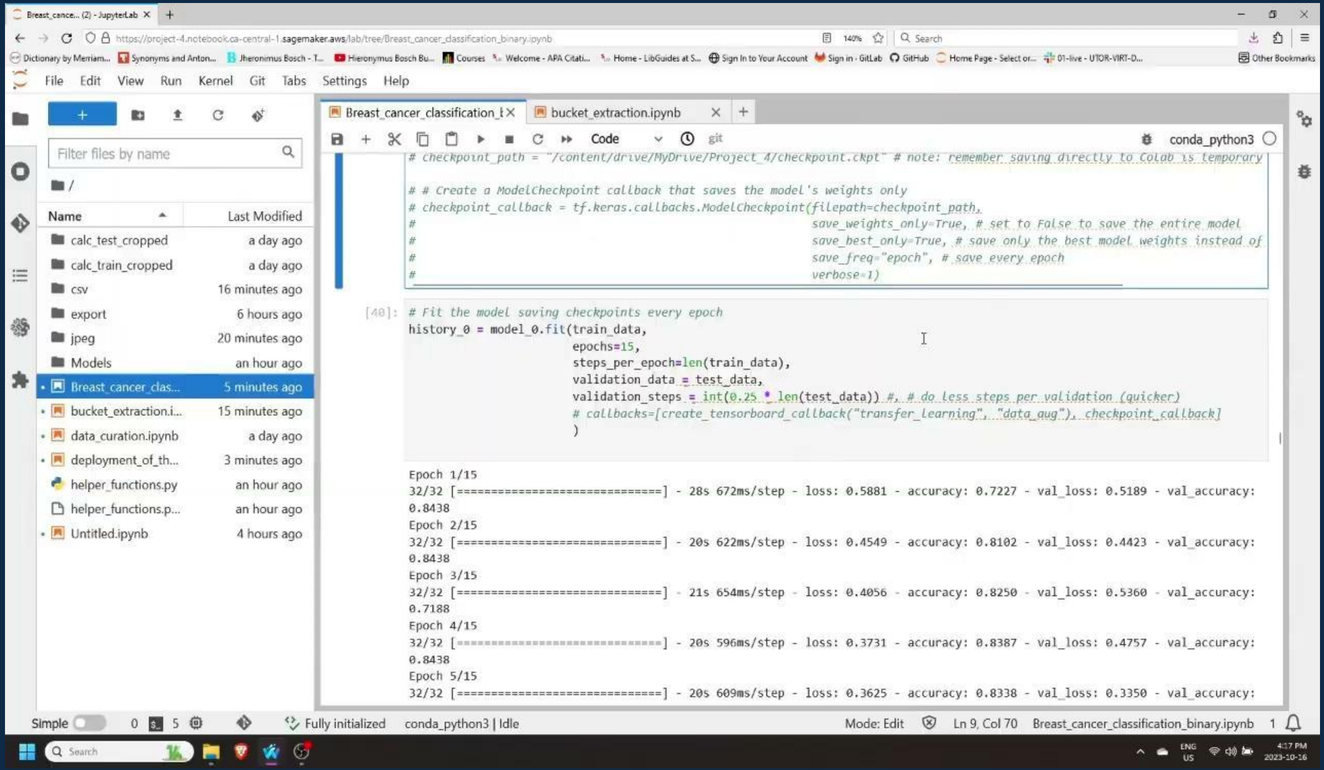
Five gigabytes worth of image data

Moved to AWS

Image data and models hosted correctly on AWS

Image data held in buckets. Models pulled image data from buckets and executed code successfully


Notebook instance
Calc only
Notebook instance
Mass only
Bucket
Raw Data
Notebook instance
For anything

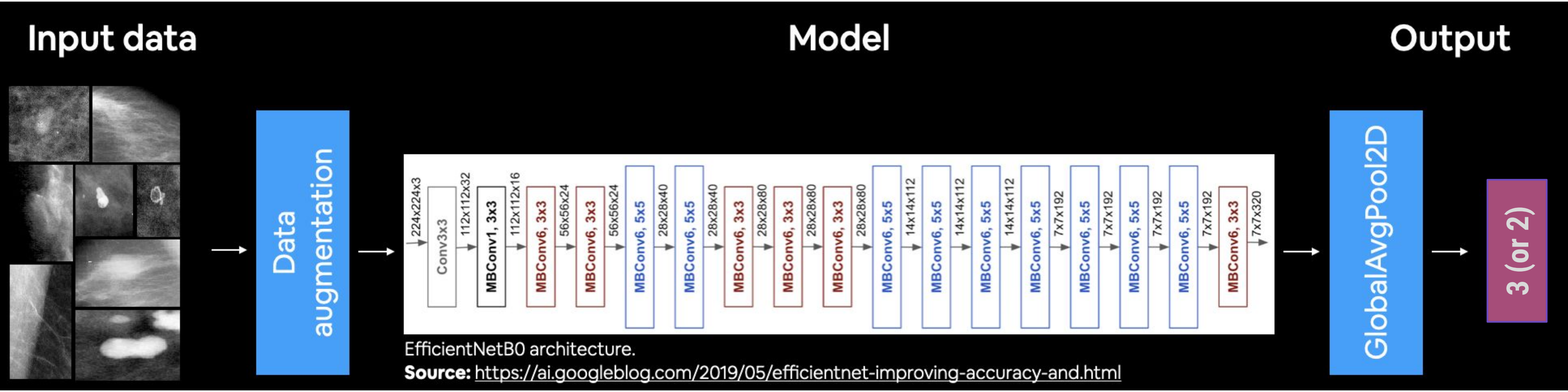

File Edit View Run Kernel Git Tabs Settings Help
Breast_cancer_classification_binary.ipynb
bucket_extraction.ipynb
Filter files by name
Name
Last Modified
calc_test_cropped a day ago
calc_train_cropped a day ago
csv 16 minutes ago
export 6 hours ago
jpeg 20 minutes ago
Models an hour ago
Breast_cancer_clas... 5 minutes ago
bucket_extraction.i... 15 minutes ago
data_curation.ipynb a day ago
deployment_of_th... 3 minutes ago
helper_functions.py an hour ago
helper_functions.p... an hour ago
Untitled.ipynb 4 hours ago
checkpoint_path = "/content/drive/MyDrive/Project_4/checkpoint.ckpt" # note: remember saving directly to Colab is temporary
Create a ModelCheckpoint callback that saves the model's weights only
checkpoint_callback = tf.keras.callbacks.ModelCheckpoint(filepath=checkpoint_path,
save_weights_only=True, # set to False to save the entire model
save_best_only=True, # save only the best model weights instead of
save_freq="epoch", # save every epoch
verbose=1)
[40]: # Fit the model saving checkpoints every epoch
history_0 = model_0.fit(train_data,
epochs=15,
steps_per_epoch=len(train_data),
validation_data = test_data,
validation_steps = int(0.25 * len(test_data)) #, # do less steps per validation (quicker)
# callbacks=[create_tensorboard_callback("transfer_learning", "data_aug"), checkpoint_callback]
)
Epoch 1/15
32/32 [==============================] - 28s 672ms/step - loss: 0.5881 - accuracy: 0.7227 - val_loss: 0.5189 - val_accuracy: 0.8438
Epoch 2/15
32/32 [==============================] - 20s 622ms/step - loss: 0.4549 - accuracy: 0.8102 - val_loss: 0.4423 - val_accuracy: 0.8438
Epoch 3/15
32/32 [==============================] - 21s 654ms/step - loss: 0.4056 - accuracy: 0.8250 - val_loss: 0.5360 - val_accuracy: 0.7188
Epoch 4/15
32/32 [==============================] - 20s 596ms/step - loss: 0.3731 - accuracy: 0.8387 - val_loss: 0.4757 - val_accuracy: 0.8438
Epoch 5/15
32/32 [==============================] - 20s 609ms/step - loss: 0.3625 - accuracy: 0.8338 - val_loss: 0.3350 - val_accuracy:
Simple 0 5 Fully initialized conda_python3 | Idle Mode: Edit Ln 9, Col 70 Breast_cancer_classification_binary.ipynb 1

03

Deep Learning Models:

Deep learning Models

- Adopted transfer learning method using the EfficientNetV2B0 architecture
- Fit the model to our desired output classes
- Tested on multi-class classification
- Tested on binary classification



EfficientNetB0 architecture.
Source: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

Multi-class classification training and validation (B - BWC - M)

Calcification dataset

Base model as is

Validation loss:

0.903

Validation accuracy:

0.509

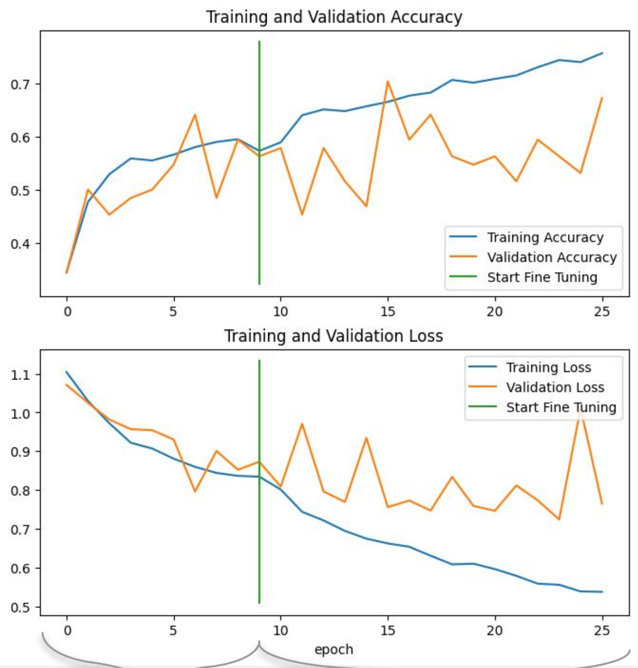
Unfreeze top 10 layers

Validation loss:

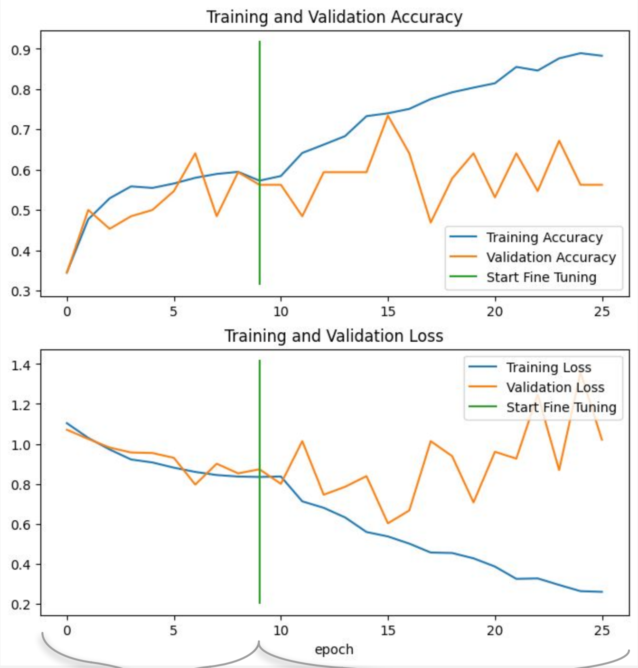
0.863

Validation accuracy:

0.536



Base model as is **Unfreeze top 10 layers**



Base model as is **Unfreeze all layers**

Unfreeze all layers

Validation loss:

1.017

Validation accuracy:

0.571

Multi-class classification training and validation (B - BWC - M)

Mass dataset

Base model as is

Validation loss:

0.898

Validation accuracy:

0.569

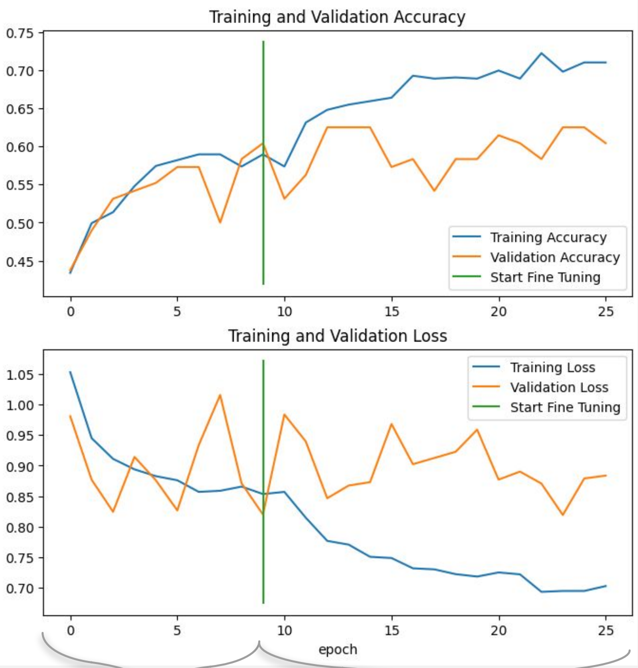
Unfreeze top 10 layers

Validation loss:

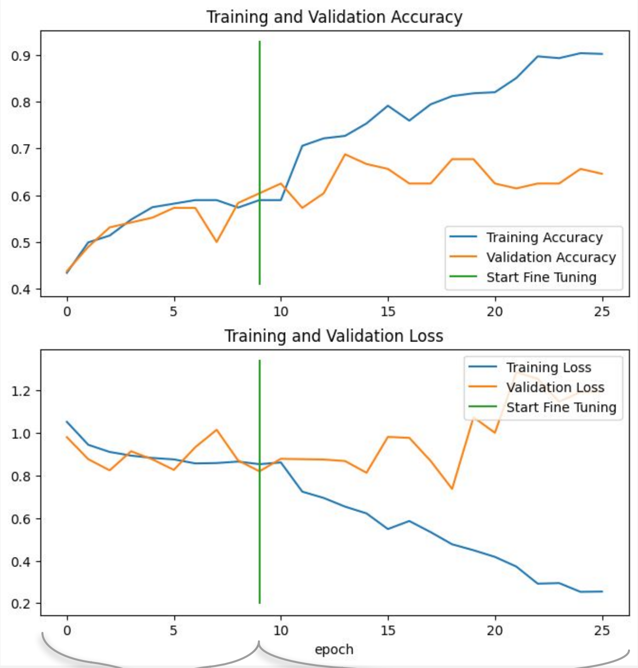
0.875

Validation accuracy:

0.629



Base model as is **Unfreeze top 10 layers**



Base model as is **Unfreeze all layers**

Unfreeze all layers

Validation loss:

1.253

Validation accuracy:

0.635

Binary classification training and validation (BWC vs M)

Calcification dataset

Base model as is

Validation loss:

0.349

Validation accuracy:

0.852

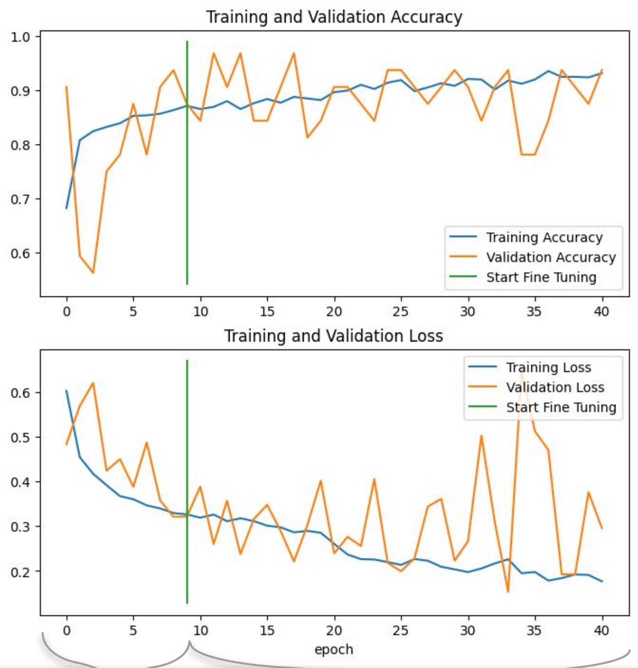
Unfreeze top 10 layers

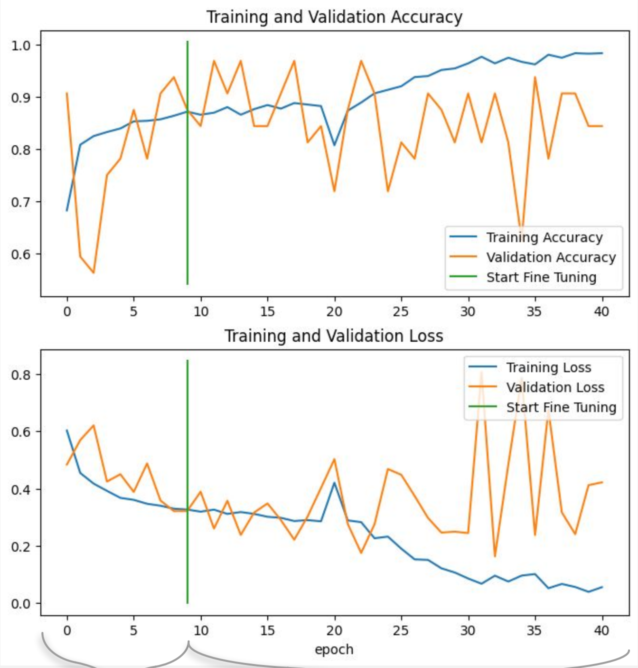
Validation loss:

0.329

Validation accuracy:

0.877





Unfreeze all layers

Validation loss:

0.320

Validation accuracy:

0.872

Base model as is **Unfreeze top 10 layers** **Base model as is** **Unfreeze all layers**

Binary classification training and validation (BWC vs M)

Mass dataset

Base model as is

Validation loss:

0.487

Validation accuracy:

0.825

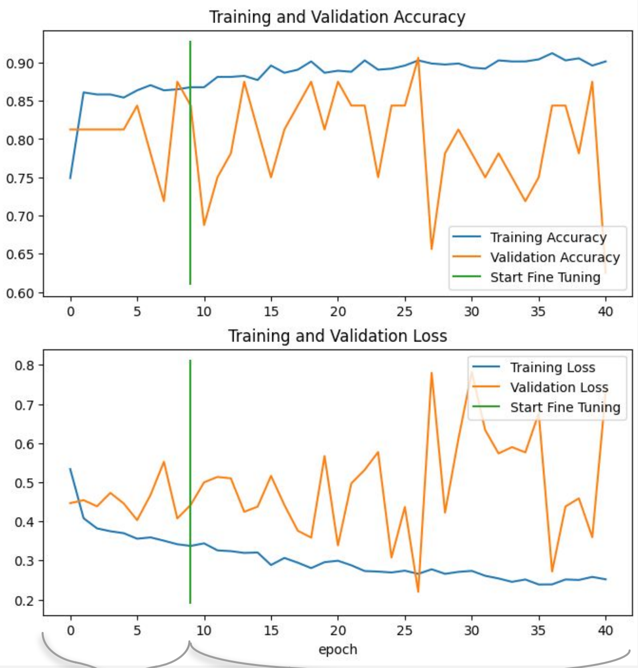
Unfreeze top 10 layers

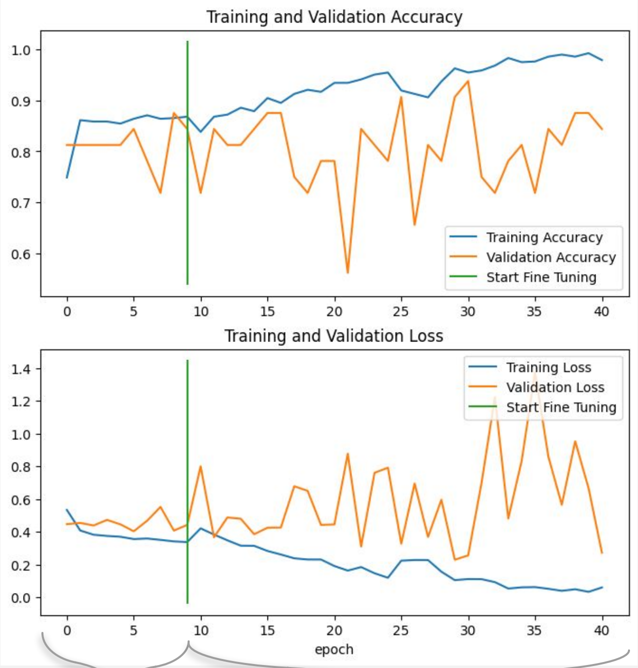
Validation loss:

0.496

Validation accuracy:

0.804





Base model as is **Unfreeze top 10 layers** **Base model as is** **Unfreeze all layers**

Unfreeze all layers

Validation loss:

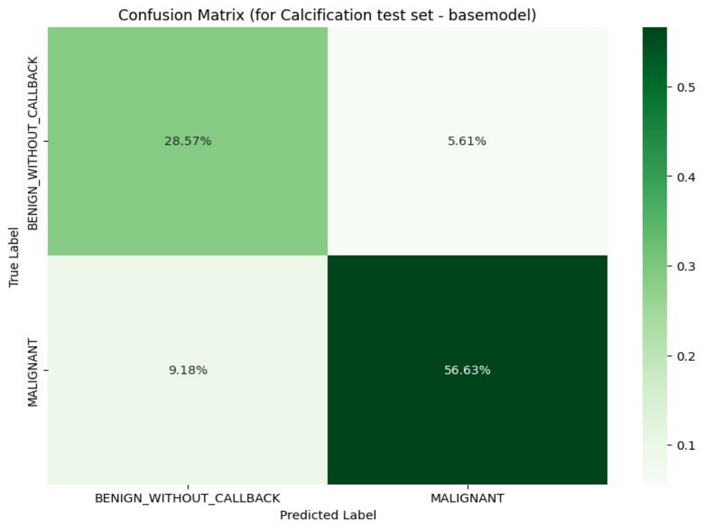
0.586

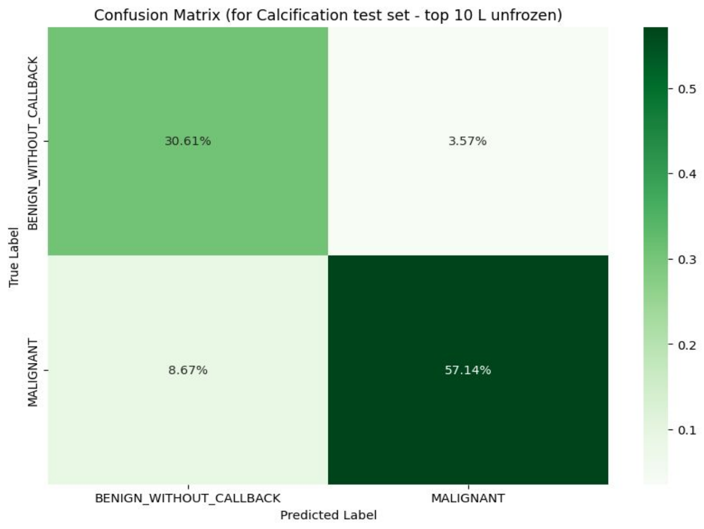
Validation accuracy:

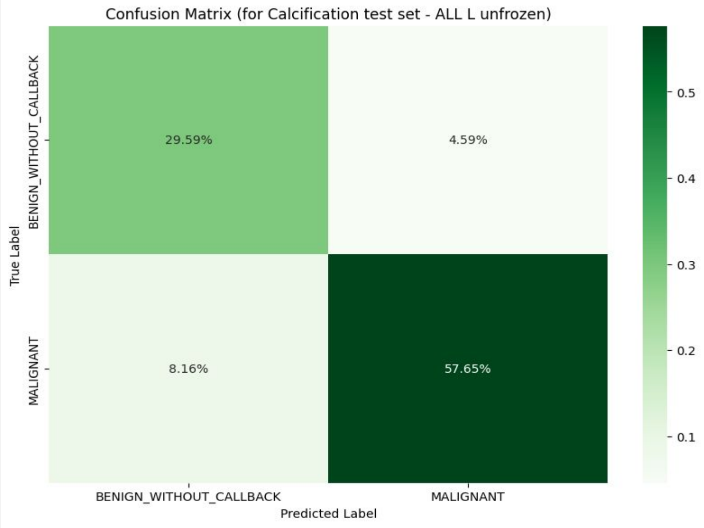
0.799

Confusion matrices

Calcification test set

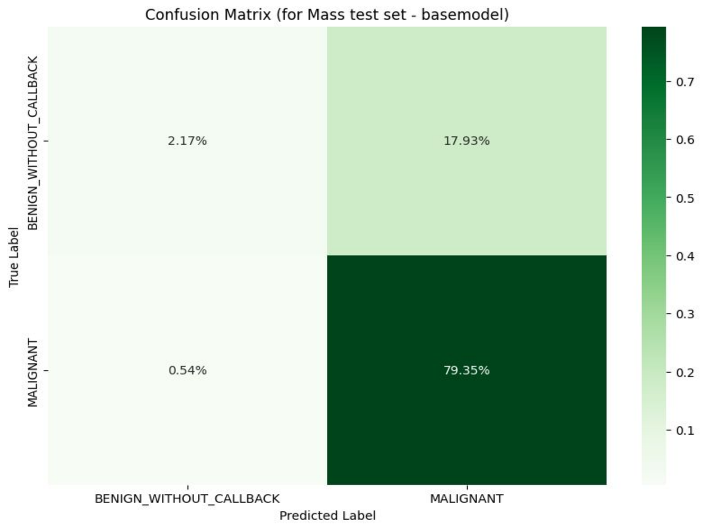


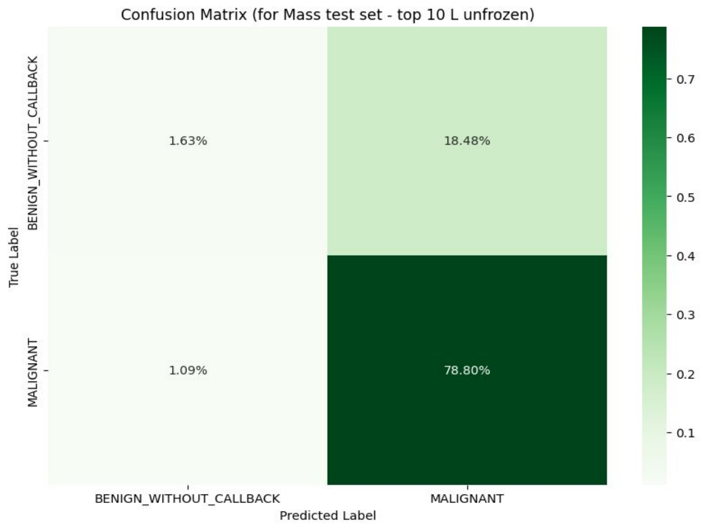


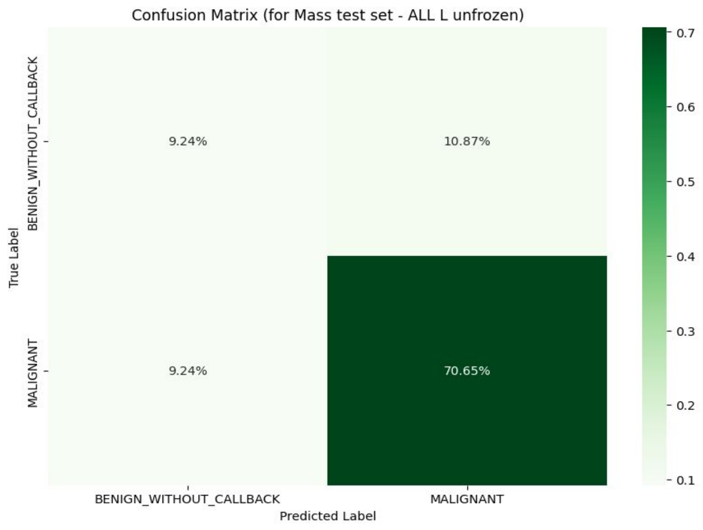


Confusion matrices

Mass test set







06

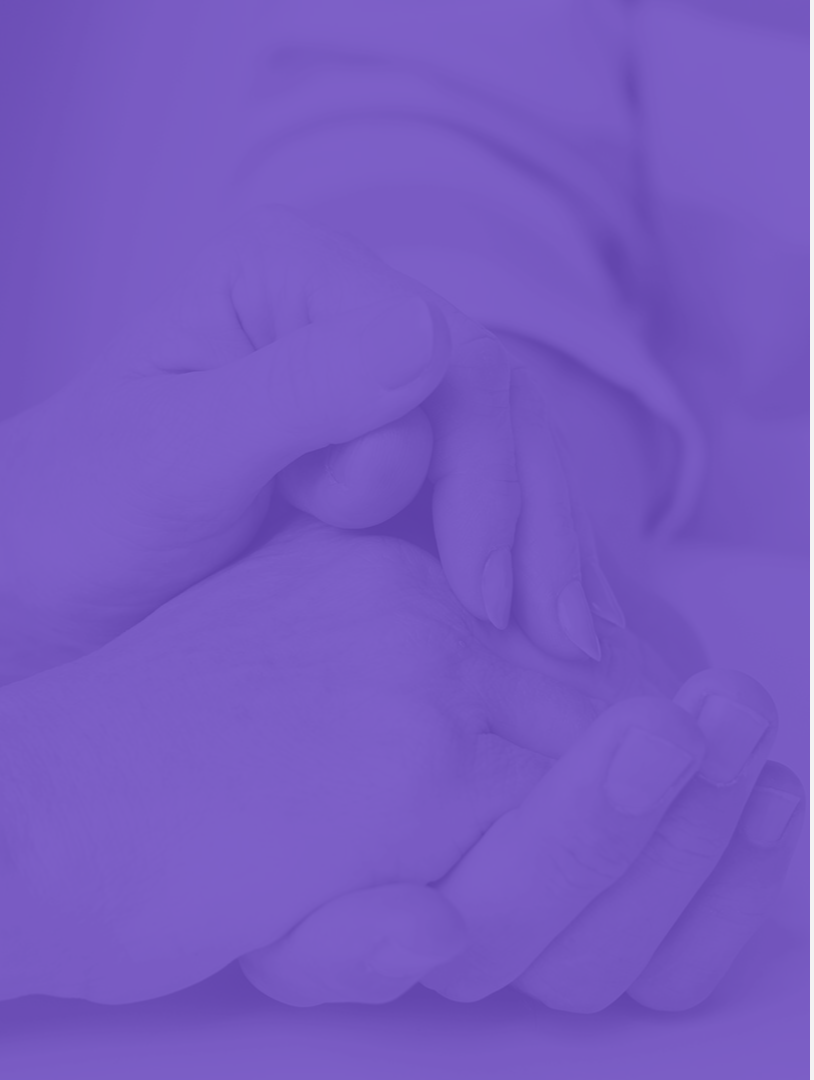
Concluding Remarks

Summary

- Presented breast cancer mammography image dataset
- Performed data exploration, feature analysis, and analytics on the textual and image data
- Hosted image data on AWS S3 buckets, trained and tested using AWS SageMaker
- Implemented a transfer learning method using the EfficientNetV2B0 model
- Attempted optimization using fine tuning by unfreezing top 10 layers and then all layers for both mass and calcification datasets
- The models can predict only the malignant cases with satisfactory accuracies

Future Work

- The textual labels and the image labels could be combined in a more sophisticated architecture to improve accuracies



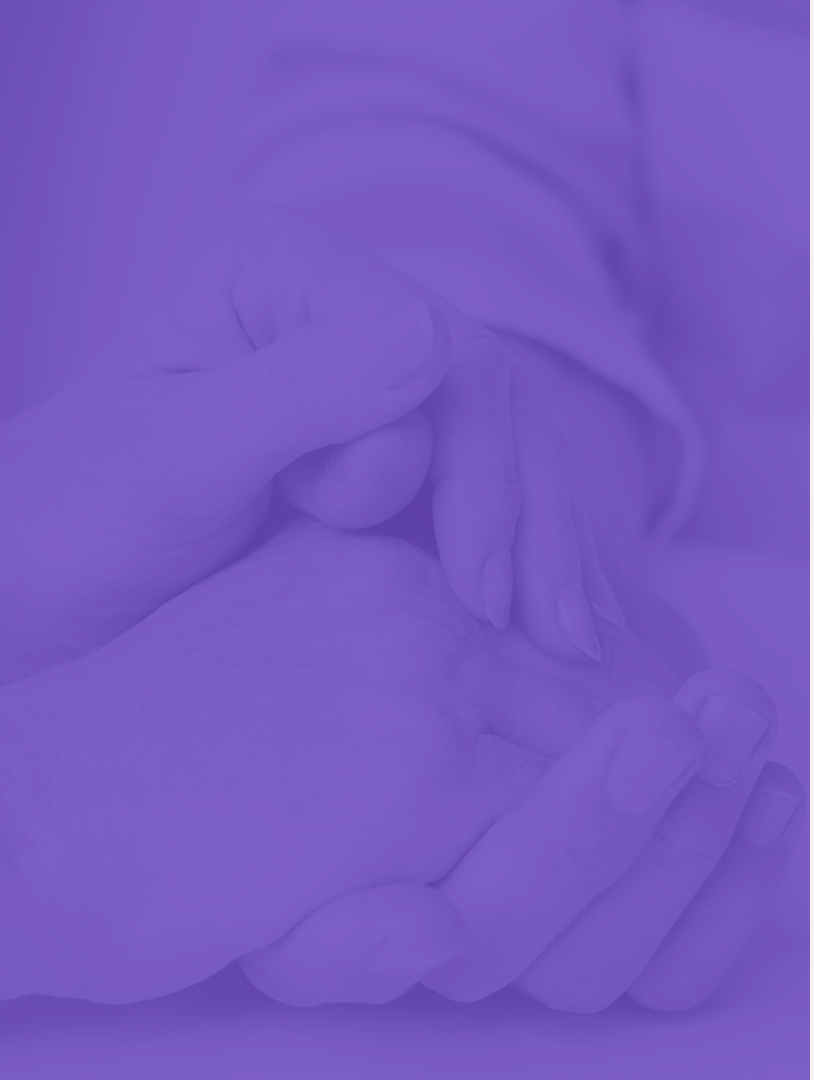
Conclusions

- Analysis shows potential of features influencing how the model interprets images and classifies classification of breast cancer
- Crucial for health professionals to continue using other tools and methods for classification of breast cancer
- Features such as shape, description not only have an effect on predicting an image but, an effect as whole on classifying classifications of breast cancer. Subtlety levels also showed that most of these masses or calcifications are being caught on the images as it was a very small percentage at level 1. 1 being completely subtle.

Thank you for your Attention!

Questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.



References

- https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset/data
- https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/
- https://plotly.com/python/choropleth-maps/
- https://towardsdatascience.com/simplest-way-of-creating-a-choropleth-map-by-u-s-states-in-python-f359ada7735e
- https://stackoverflow.com/questions/66572349/python-sub-state-names-for-abbrev-via-python-dict-with-re-subblisher
- AUTHOR (YEAR). *Title of the publication*. Publisher