RSNA Breast Cancer Detection

AWS Machine Learning Engineer Capstone Project

Alaa Mohamed
1-16-2023

Contents

Definition

Project Overview

- The project is about Breast cancer detection using screening mammography images. Our input to the deep learning network is the screening mammograms of both breasts. We try to leverage the anatomical asymmetry between the patient's breasts to confirm the presence of anomalies. So, the goal is to perform binary classification whether the screening is normal or has cancer.
- The dataset used is Kaggle's RSNA Breast Cancer Detection Dataset. https://www.kaggle.com/competitions/rsna-breast-cancer-detection
- In order to simplify the problem and be able to execute the training on AWS I used a small portion of the data available all the details are present in in RSNA_dataset_prep.ipynb

Problem Statement

Breast cancer is on of the most prevalent cancers worldwide. According to the WHO, it is the second leading cause of death among women. Early detection of breast cancer is crucial for better treatment and survival outcomes. Also, AI would significantly reduce the workload on trained experts and radiologists. However, false positive rate is a major concern. So, a reliable model that accurately detects both classes is the optimal solution.

Metrics

Since this is a binary classification problem, we will evaluate our model by:

- Accuracy: measures the model's ability to predict the correct class. However maybe misleading in imbalanced datasets.
- AUC-ROC: The plot of the true positive rate against the false positive rate. The highest the area under the curve (AUC), the better the model in the classification task.
- F1-score: combines the precision and recall of a classifier into a single metric by taking their harmonic mean.

Analysis

Data Exploration

Tabular data exploration:

	site_id	patient_id	image_id	laterality	view	age	cancer	biopsy	invasive	BIRADS	implant	density	machine_id	difficult_negative_case
0	2	10006	462822612	L	CC	61.0	0	0	0	NaN	0	NaN	29	False
1	2	10006	1459541791	L	MLO	61.0	0	0	0	NaN	0	NaN	29	False
2	2	10006	1864590858	R	MLO	61.0	0	0	0	NaN	0	NaN	29	False
3	2	10006	1874946579	R	CC	61.0	0	0	0	NaN	0	NaN	29	False
4	2	10011	220375232	L	CC	55.0	0	0	0	0.0	0	NaN	21	True

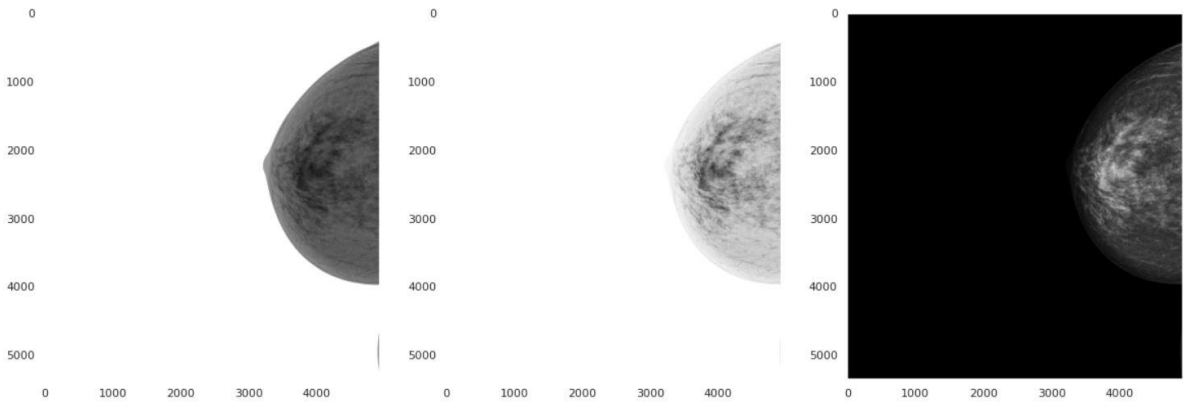
Insights:

1. 54706 images are present, laterality and views are present for all images., density is provided for 29470 only, age for 54669 (good almost all the images), and BIRADS for 26286 which is almost half the images present.
2. Almost all patients have the 4 views. Some have more views but are negligible
3. 54706 images belonging to 11913 unique cases. The data is collected across two sites from different 10 mammography machines
4. Images are in DICOM format but stored as lossless jpeg (needs some extra packages to be read)
5. Images differ by machine, some are inverted, some need VOI LUT application, some already have the VOI but also have some sort of label (LMLO) from the viewer. So, we need to take this into consideration that the dataset is not homogeneous and require careful preprocessing.

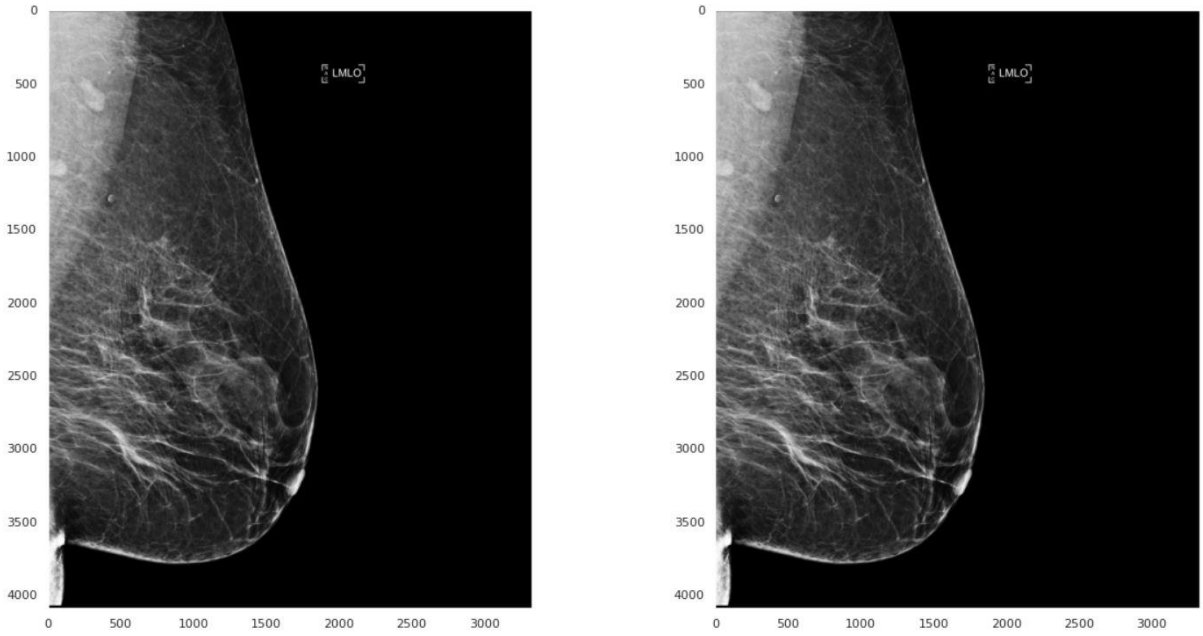
	age
count	54669.000000
mean	58.543928
std	10.050884
min	26.000000
25%	51.000000
50%	59.000000
75%	66.000000
max	89.000000

Dataset Examples:

1. An example of a CC view where VOI LUT was needed and photometric corrections as well.



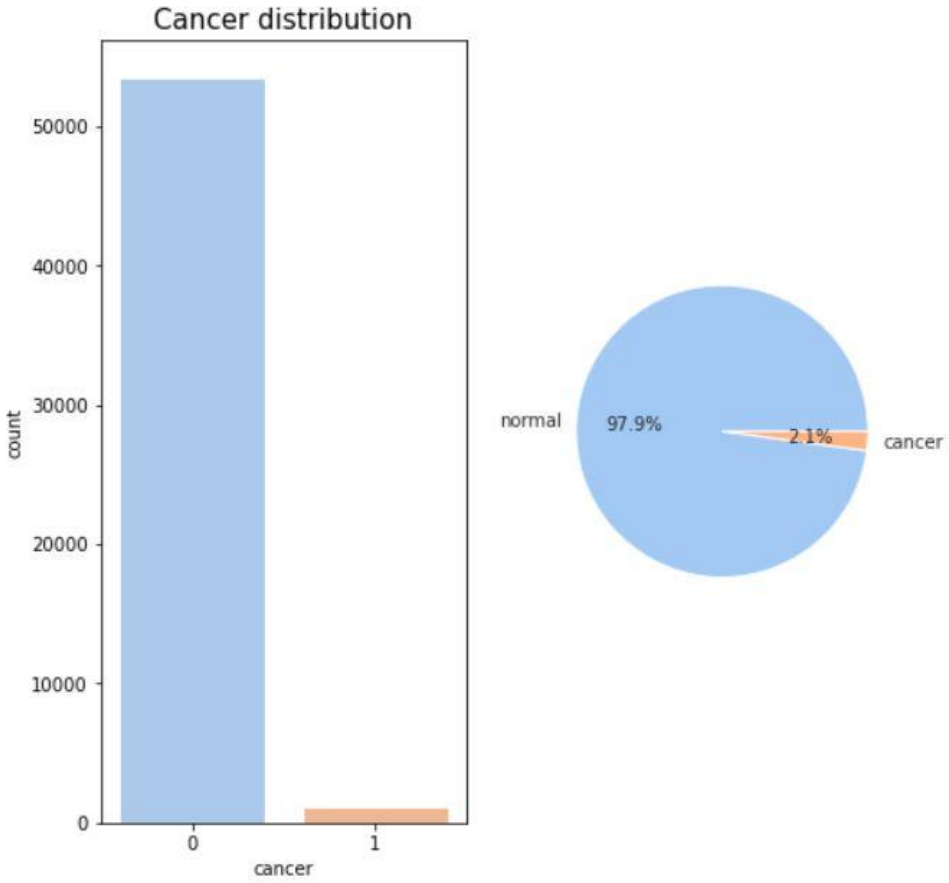
2. Example where neither the VOI nor photometric corrections were needed but artifacts are present



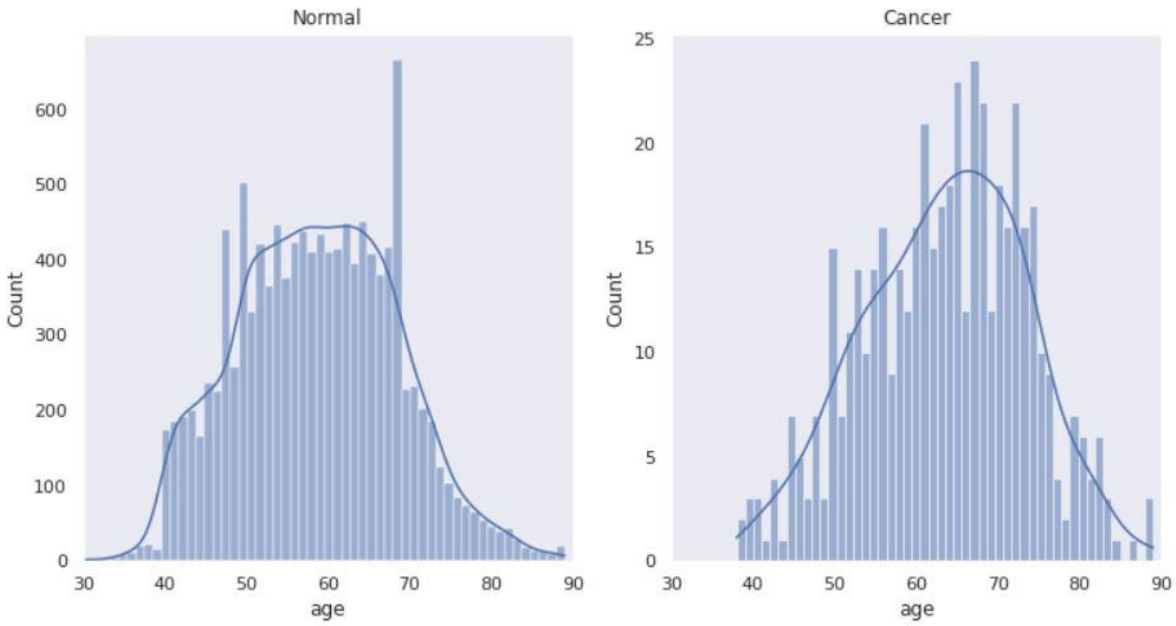
Exploratory Visualization

All the EDA is found in RSNA_dataset_EDA.ipynb. Some major findings are reported below.

1. Class Imbalance: dataset is extremely imbalanced where cancer cases represent about 2% of the images.



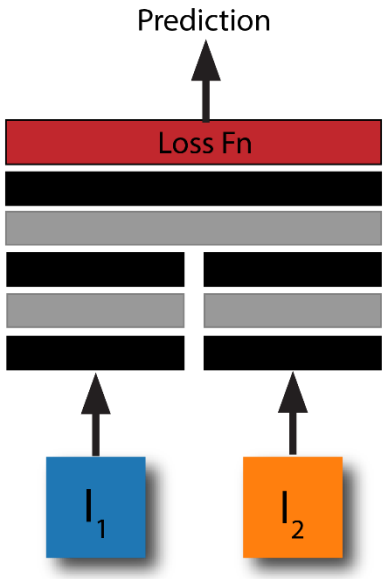
2. Age distribution per class. The age in the cancer class is negatively skewed a little. Which makes sense medically since age is a risk factor of breast cancer.



Algorithms and Techniques

I propose using Siamese neural networks. Siamese neural networks are popular in computer vision problems like signature verification and face recognition. Recently, it was utilized in the medical domain since many organs are anatomically symmetrical.

Siamese network's architecture consists of a twin CNN with shared weights and then a fully connected classifier. Siamese can automate the radiologists' approach in comparing the two breasts for confirming any suspicious lesions. It may as well reduce false positives.



Benchmark

Results can be compared to a conventional CNN.

Methodology

Data Preprocessing

1. DICOM is read. (Pixel array as well as metadata)
2. Perform VOI LUT to adjust the intensity scale.
3. Perform photometric correction if needed.
4. Normalize the images
5. Save as PNG for convenience

Implementation

I implemented the training script in TensorFlow and Keras (found in train_model.py)

The training job was done using AWS's SageMaker SDK (found in train_and_deploy.ipynb). I faced some complications as this was my first time using TensoFlow on AWS so it was a little tricky to know which version works in the training job and can be deployed. Also, I didn't know

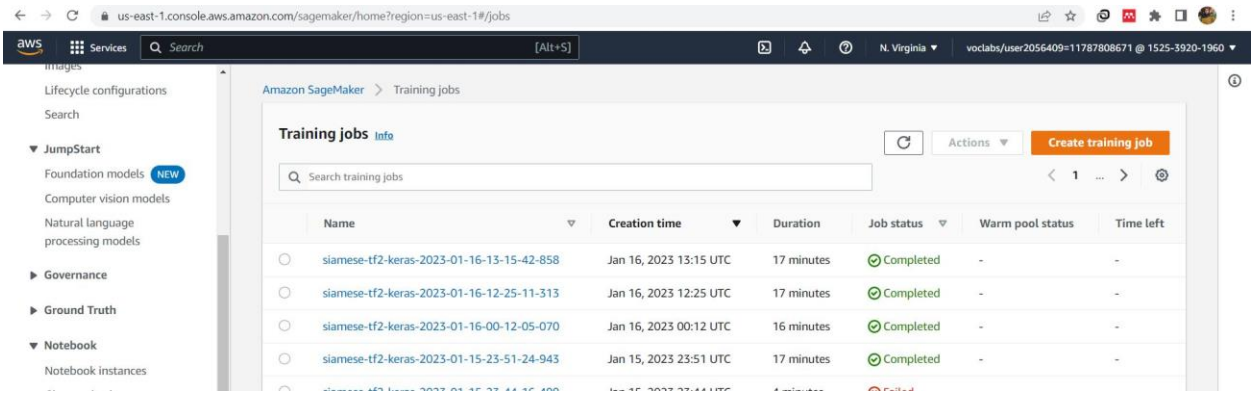
that I have to save the model in a certain way in order for SageMaker to be able to copy it into S3.

```
role = get_execution_role()

estimator = TensorFlow(
    role=role,
    base_job_name="siamese-tf2-keras",
    instance_count=1,
    instance_type="ml.p2.xlarge",
    entry_point="train_model.py",
    framework_version="2.3",
    py_version="py37",
    hyperparameters=hyperparameters
)
```

Refinement

Many training iterations were done trying to find the optimal architecture and hyperparameters.

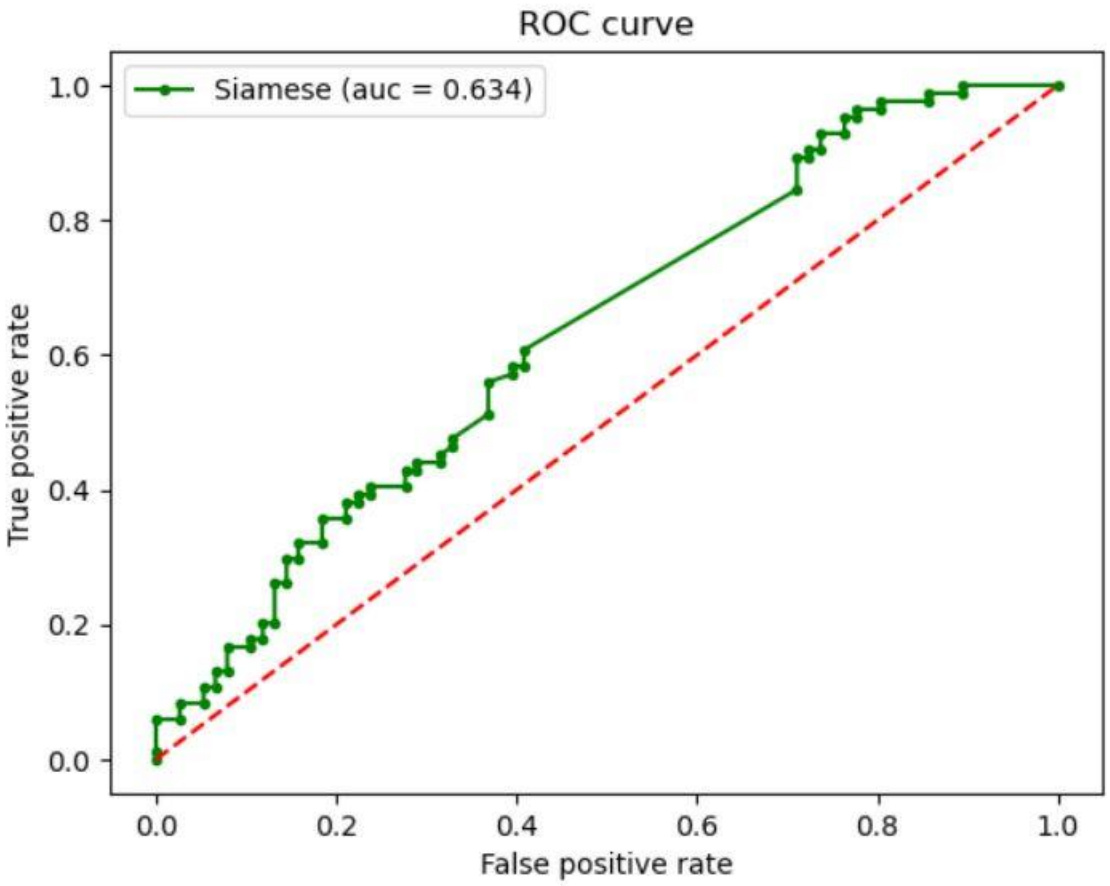


Results

Model Evaluation

Metric	Value
Accuracy	56.25%
AUC	0.63
F1-Score	0.52

The results are not optimal. But in a hard problem as breast cancer detection, the results seem to indicate that siamese neural networks could serve as a robust classifier.



Conclusion and Future Work

Siamese Neural Networks seems to be a reliable classifier.

However, some limitations include:

- The dataset size I chose, but this serves as a proof of concept.
- Keeping the aspect ratio of the input images and removing the artifacts.

Therefore, a more thorough implementation and analysis is warranted on the whole dataset.

- Maybe try some other loss functions like contrastive loss to better test my hypothesis.