Breast Cancer Prediction Using CNNs

ماشين لرننغ

The ماشين لرننغ Team

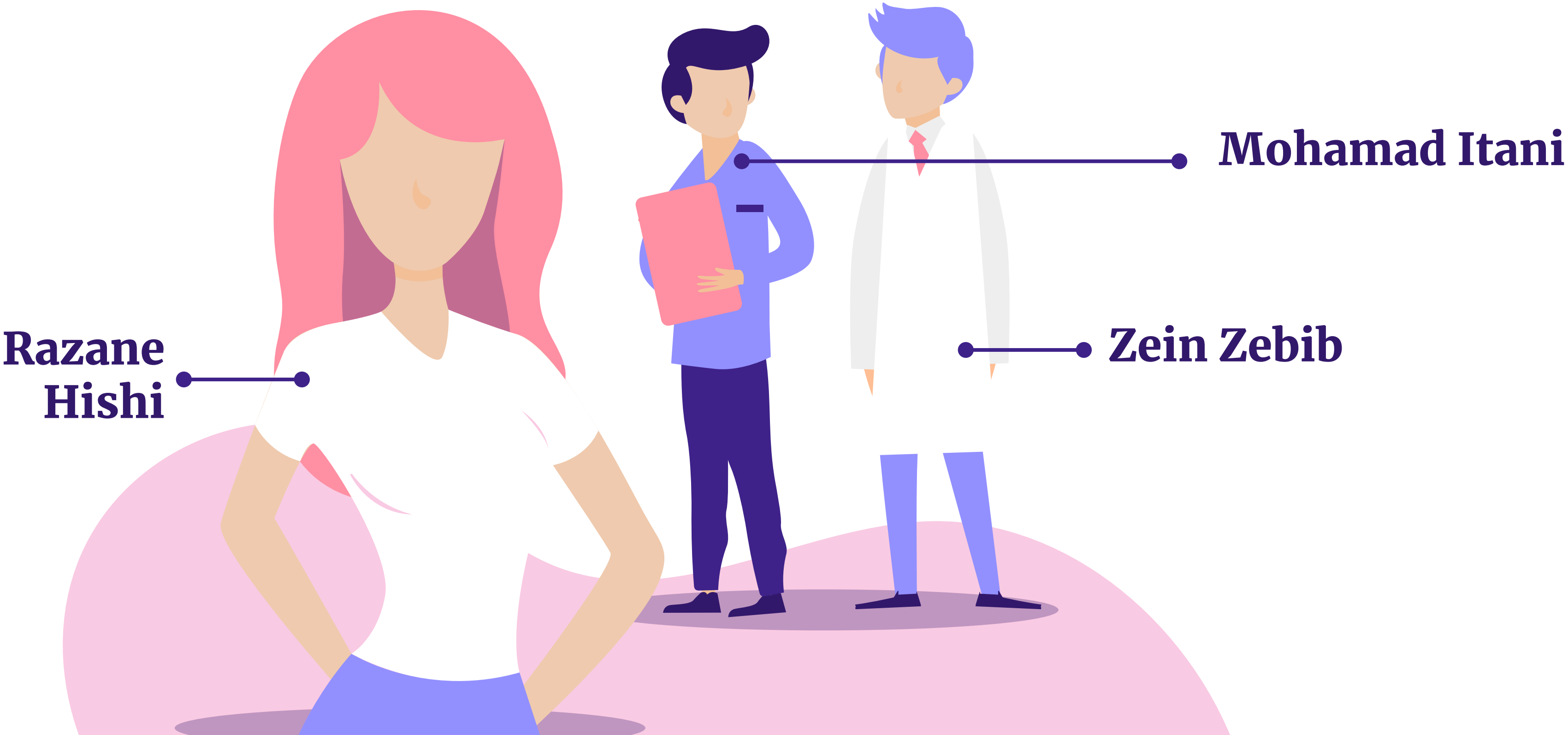
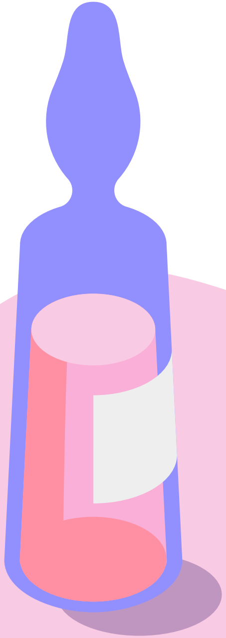


Table of Contents



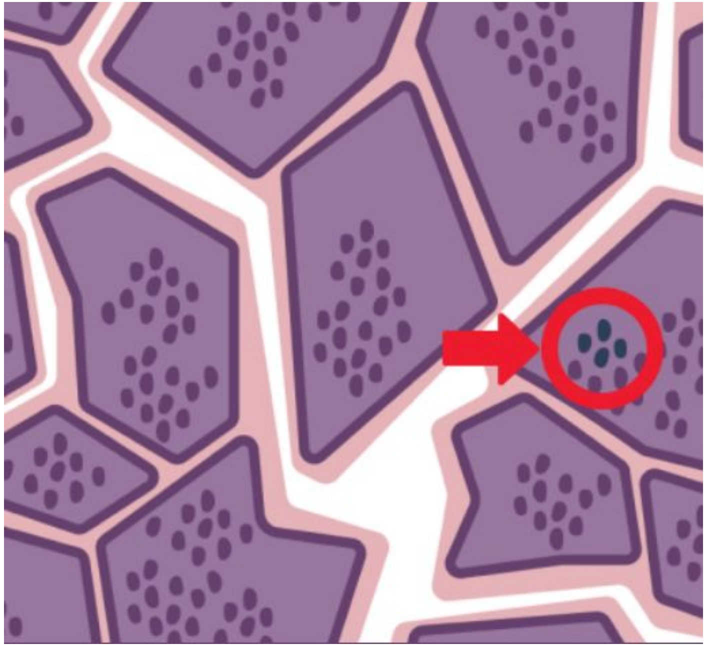
01

Problem Motivation

40,000

Patients died due to cancer misdiagnosis in 2013

01 Problem Motivation & Challenges



- Cancer Detection is described as spotting 3 blue cars from a satellite photo of a city.
- Pathologists do this for hundreds of tissue samples
- Breast Cancer accounts for 30% of all cancer in women

How can we detect cancer efficiently and allow pathologists to understand the output?

02 Problem Definition

Classification of breast cancer as IDC+ or IDC- using images of tissue samples.

Adding explainability to the AI model

03 Dataset

DESCRIPTION

Diagnosis	Number of Images
Positive	78,786
Negative	198,738

- Dataset consists of 162 slide images of breast cancer specimen scanned at 40x
- 277,524 patches of size 50x50 extracted
- Dataset taken from Kaggle

04 Methodology

CNN to classify tissue samples as IDC+ or IDC-.
Implemented using tensorflow library

Saliency maps to view into the models "thinking" process. SmoothGrad and Vanilla Gradient maps.
Implemented using saliency library

We assumed we could tell if a saliency map is highlighting important features. *We aren't pathologists*

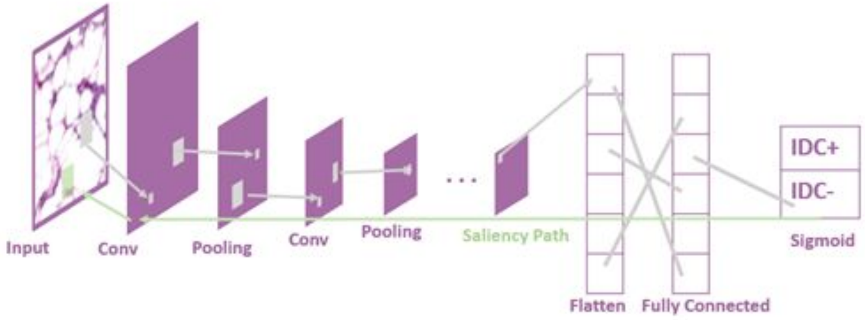
Model

Input

Labeled Images of IDC+ and IDC- samples
Images are of size 50x50 and colored.

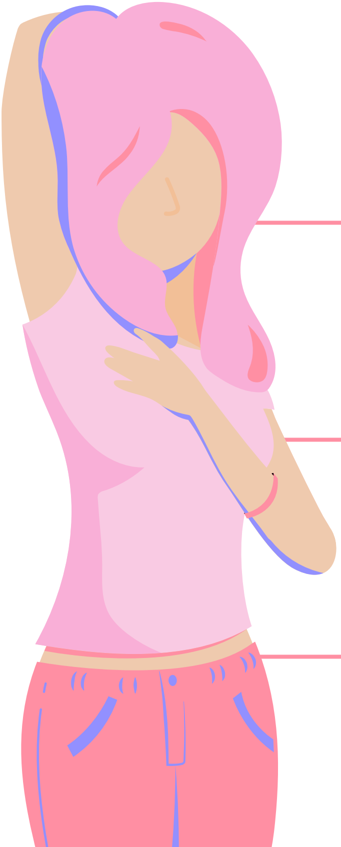
Output

Classification into one of the two categories. 1 dimensional vector output



We assumed the base model architecture will work on saliency maps from the saliency library.
This assumption lead to us experimenting with different architectures.

05 Experiments



A
Architecture Modification

B
Models Survey

C
Saliency Map Evaluation

Architecture Modification Results

Two different variations of a CNN were experimented on.

Archi 1 Metrics					
# of Conv Layers	*F1 IDC-*	*F1 IDC+*	*Accuracy*	*Recall IDC+*	*Recall*
2	84	86	85	88	85
3	87	87	87	87	87
4	86	87	87	89	87
5	84	87	85	94	85

Archi 1
Output : 1D vector
Activation : Sigmoid
Optimizer: Adam

Archi 2 Metrics					
2	86	87	86	89	86
3	81	85	83	94	83
4	85	87	86	92	86
5	84	87	86	93	86

Archi 2
Output : 2D vector
Activation : SoftMax
Optimizer: sgd

In both cases, four convolutions layers performed best. They will be referred to as Model 1 and Model 2.

Model Surveying Results

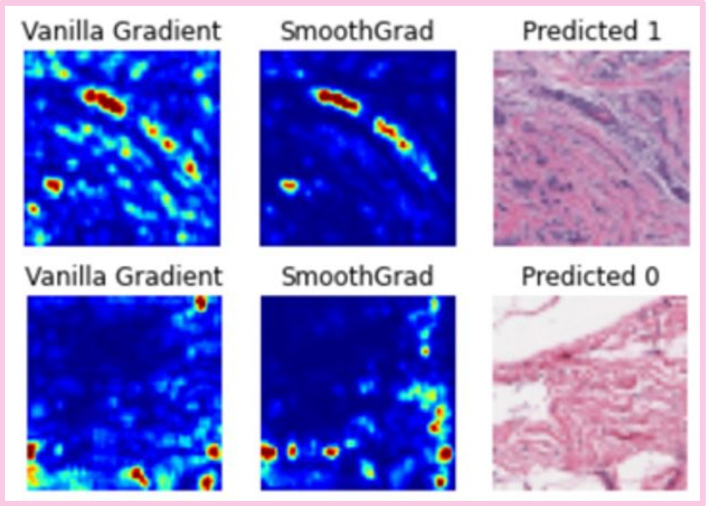
The model was compared to a Logistic Regression model, an SVC model, and a Random Forest Classifier model

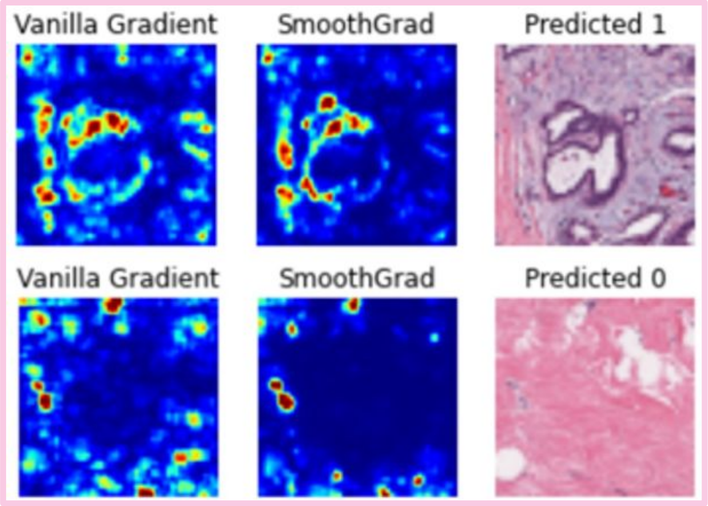
Model	*F1 IDC-*	*F1 IDC+*	*Accuracy*	*Recall IDC+*	*Recall*
Logistic Regression	76	75	76	78	76
Random Forest Classifier	82	82	82	84	82
Model 1	86	87	87	89	87
Model 2	85	87	86	92	86

Both Model 1 and 2 outperformed the other implemented models in all metrics

Saliency Maps Results

The saliency maps were generated using Model 2. The Saliency library required the output of the model to have to have SoftMax.





Saliency maps were able to highlight the features the small and big purple spots in the tissue samples.

Further Experiments Results P1

SIZE represents the number of images of each class

Model 1 Metrics					
SIZE	*F1 IDC-*	*F1 IDC+*	*Accuracy*	*Recall IDC+*	*Recall*
5k	78	82	81	92	81
10k	87	84	86	75	85
20k	86	87	87	89	87
30k	85	84	85	81	85
40k	83	86	84	92	84

Model 1

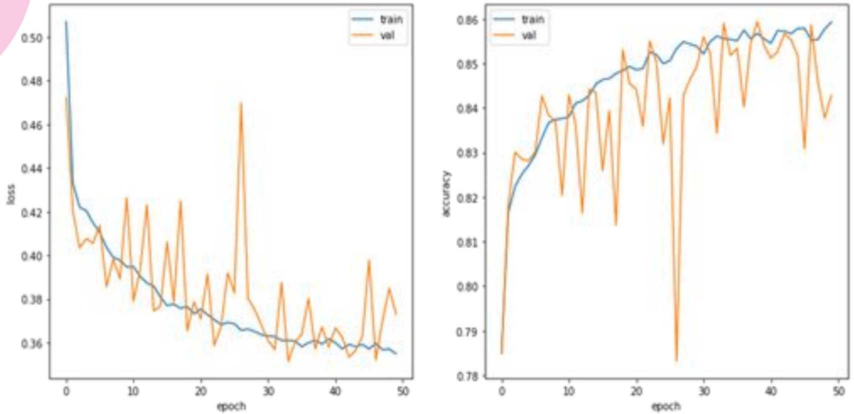
Model 2 Metrics					
5k	77	81	79	89	79
10k	87	86	87	85	87
20k	85	87	86	92	86
30k	85	87	86	91	86
40k	86	87	86	86	86

Model 2

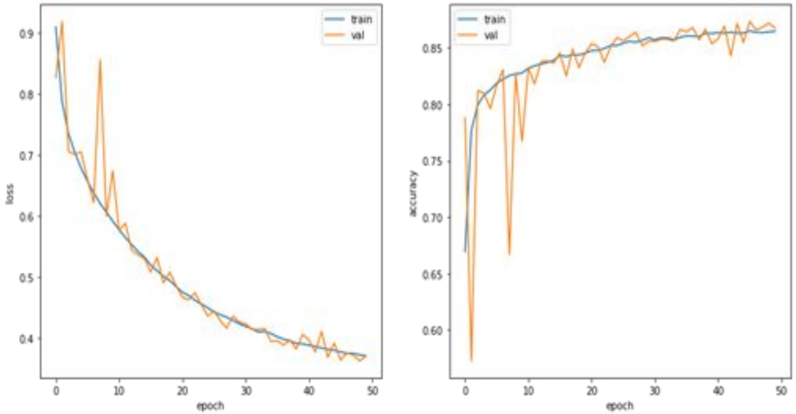
As SIZE increased, Model 2 performed better because it had a more stable accuracy.

Further Experiments Results P2

To further validate the difference between the models, the following graphs were compared.



Model 1



Model 2

Model 2 had less "spikes" and a more stable graph than Model 1's graphs.
Model 2 was able to abstract the rules needed to predict the two classes.

06 Conclusion and Future Work

Model 2 had good accuracy and showed to be stable when exposed to more data

Dataset outliers were included in the training process because they required a lot of manual work

Hardware limitations halted further development of the model.

Future study is required with increasing SIZE to include all the IDC+ images.

06 Conclusion and Further Work +

There should also be further study in checking how magnification of the images affect the accuracy of the model. This was tested in related work.

Pathologist should be included in the process of evaluating saliency maps and to further understand what is used in classifying cancer.

Thank You for Listening !!!

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