Breast Cancer Prediction Using Convolutional Neural Networks

Mohamad Itani
Department of Electrical and Computer Engineering
American University of Beirut
Beirut , Lebanon
mdi00@mail.aub.edu

Zein Zebib
Department of Electrical and Computer Engineering
American University of Beirut
Beirut , Lebanon
zhz07@mail.aub.edu

Razane Hishi
Department of Electrical and Computer Engineering
American University of Beirut
Beirut , Lebanon
rjh25@mail.aub.edu

***Abstract*—Pathologist undergo the risky and time consuming job of evaluating hundreds of histopathological images (tissue samples) to diagnose cancer. This paper suggests a simple CNN architecture to classify invasive ductal carcinoma (IDC) in histopathological images. This paper also introduces saliency maps in order to create a more explainable model. Multiple convolutional neural network (CNN) architectures were explored and evaluated. Furthermore, the best architectures were surveyed with classical ML classification algorithms, to validate performance. Our results showed that our architecture had an accuracy of 86% on a smaller and larger training set. The saliency maps were able to capture the highlighting features of the histopathological images.**

I. Introduction

This paper focuses on aiding pathologists' work with detecting breast cancer on histopathological images. The paper explores the use of Convolutional Neural Networks (CNNs) to complete this task. Given the complexity of Neural Networks (NN) they output models that are unexplainable and have a "black box" nature making it hard for humans to understand. This study will also focus on creating models that are explainable using explainable AI (XAI) methods such as post-hoc. The purpose of this paper is to generate a model to assist pathologists ensure that nobody has to suffer from a false diagnosis.

II. Problem Motivation

Pathologists have to undergo a tedious and sensitive process of detecting the precise type and severity of a cancer from tissue samples [2]. This process is called histopathology; it is the study and diagnosis of tissue diseases which involves examining tissues and/or cells under a microscope [6]. A company called PathAI describes this process as "viewing a satellite photo of Boston and trying to find blue cars" [5]. Pathologists have to perform this task hundreds of times a day resulting in many errors. In a study conducted in 2013 approximately 10-20% of all cancer cases were misdiagnosed leading to an estimated 40,000 patients dying [7]. This study will focus on aiding pathologists with diagnosing cancer specifically using breast histopathology images. We chose breast cancer specifically due to it being the most common form of cancer. Moreover, working on new methods to reduce the error and time needed to make a diagnosis will lead to a tremendous impact on many people's lives [3].

III. Problem Definition

The most common type of breast cancer is called Invasive Ductal Carcinoma (IDC) [1]. Pathologists typically diagnose the patient with cancer based on the size, shape, and appearance of the sample. A report of the diagnosis containing the information is then sent to a doctor who can understand why such a diagnosis was made [1]. This paper aims to create a model that can output whether an image of a tissue sample is IDC+ or ICD-. The model is going to use Convolutional Neural Networks (CNNs) to conduct the classification. The model is trained on a dataset of histopathological images.

Due to the complexity of CNNs, the model will operate in a similar manner to a "black box" which means that the user will not understand the reasoning behind the decision. In order to achieve explainability in the model, a post-hoc method called saliency maps is used to highlight the visually captivating unique features. Products in the medical industry must be built upon trust, transparency, and high ethical standards. Thus, it is vital to implement Explainable AI (XAI) in the healthcare industry instead of its "black box" counterpart.

IV. CNN Overview

A CNN is a type of Neural Network that takes an image as input. A typical CNN architecture includes a convolution layer, pooling layer, and a fully connected layer. CNNs take images and assign weights to different objects in the image. [15] The convolution layer has a kernel of a fixed size that highlights a subset of an image and performs matrix multiplication on this portion and saves it. This in turn reduces the size of the image and brings the information into one pixel [16]. The second part is a pooling layer that further reduces the representation and allows the model to extract the most dominant features. The two methods of pooling are max pooling and average pooling. Max pooling returns the max value from the output of the kernel while average pooling averages the values [17]. Before feeding the output of the pooling layer into the fully connected layer, the matrix representation of the output is flattened into an array. The fully connected layer is the final step that classifies the images. The process that the CNN goes through has proven to be very successful as different forms of CNNs are being used in the field of pathology.

V. Related Work

Multiple startups and researchers have worked on breast cancer diagnosis over the past few years. PathAI is among the leading startups in the field of medical diagnoses. PathAI is developing assistive technology to allow pathologists to perform accurate breast cancer diagnoses [5]. In addition, they are partnering with Novartis to produce a blackbox model capable of detecting complex hidden patterns that pathologists cannot notice [4].

CNNs have been widely used in the literature to classify breast cancer, with different CNN architectures and techniques being used [20],[21]. In S. Awadh Alanazi [18] a dataset of 270,000 breast histopathology images were classified using three different CNN architectures. The findings concluded that a five-layer CNN resulted in the highest accuracy of 87%. Moreover, the CNN architecture proved to be more accurate than tradition ML approaches such as support vector machines (SVM), logistic regression, and K-means.

In Xia et al[19], a CNN with DenseNet interleaved with Squeeze and Excitation module was used to predict breast histopathology images on 40x, 100x , 200x , and 400x magnification. The total dataset has 7,909 images. The results were shown to have an 89.1±3.6% true positivity rate for the 40x case. In Zhang et al [22], histopathology images were interpreted using captions and visuals.

In this work, saliency maps are proposed to add explainability to a black-box CNN model; something that has not been done yet on histopathology images. Furthermore, a simple CNN architecture trained on equal parts positive and negative cases will be used causing the model not to learn both classes equally well.

VI. Problem Methodology

A. Dataset Description

This dataset originally comes from 162 slide images scanned at 40x magnification. These images are then fragmented into 277,524 patches of size 50x50 pixels. Approximately 30% of the images are IDC+ and the rest is IDC- [8].

TABLE I. HISTOPATHOLOGY IMAGE DATASET

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

B. Libraries Description

1. Pillow

Pillow for loading the images.

2. Numpy

Numpy is used for matrix manipulation as well as changing the images into an array format to be fed into the model.

3. Pandas

Pandas is used for visualizing the data and performing tasks such as one-hot encoding.

4. Matplotlib

Pyplot from Matplotlib is used to print images and visualize loss and accuracy graphs for the model's history.

5. TensorFlow

TensorFlow is used to create and visualize the CNN architecture.

6. Sklearn

Sklearn is used for splitting the dataset into training and testing data. Sklearn is also used for creating various models to be compared with the CNN architecture. Moreover, the classification report function is used to extract the metrics for the various models.

7. Saliency

Saliency library is used for creating and visualizing the vanilla gradient and smooth gradient saliency maps.

C. Model Architecture

The model architecture has two or more convolution layers with ReLU activation and a pooling layer after each convolution. After the last pooling layer, the images are flattened and fed into a dense layer with ReLu activation and a final dense layer with Sigmoid activation.

D. XAI

As mentioned before, generate saliency maps are generated to add explainability to the model. Saliency Maps rank the pixels in an image based on their contribution to the final score of a CNN. [12]When a gradient saliency map is generated; it takes the gradient from the end of the network and propagates it back to the beginning [13].

Two types of saliency maps:

1. Vanilla Gradient

$$s^v = \frac{\delta f}{\delta x}$$

f is a pre-trained NN that takes input feature vector x.

This will quantify the influence of a small change in each input dimension on the output of the network. [14]

2. SmoothGrad

$$s^s = \frac{1}{N}\sum_{i=1}^{N} s^v(x + \Delta_i)$$

Δ_i is the perturbation of input samples by a gaussian distribution.

This improves upon the vanilla gradient by generating clearer maps. It takes the average of the vanilla gradient over perturbed inputs. [14] Both types will used in this paper.

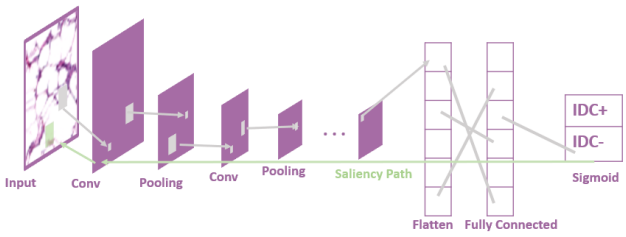

Fig. 1. Model Architecture

E. Experiments

Each experiment includes its own set of challenges that added some complexity to this study.

1. Architecture Modification

This step will involve modifying the CNN architecture to achieve the most balanced and robust CNN. Some modifications include modifying the number of convolutional and max pooling layers and changing the activation from sigmoid to SoftMax.

Challenges: How to simplify the complexity of the model while retaining accurate predictions?

2. Model Surveying

In order to validate the CNN model, different models will be used to compare the best CNN architecture outputted from the architecture modification experiments.

Challenges: Identifying which other models would be a good comparison with the CNN?

3. Saliency Map Evaluation

The best CNN architecture from the architecture modification experiment will have various saliency maps evaluated to determine if the saliency maps are giving intuitive and explainable maps.

Challenges: How are the saliency maps meant to be evaluated when the development team has no experience in pathology?

VII. Experimental Results

A. Experimental Setup

The experiments are run on Google's Colab platform using a GPU hardware accelerator. Due to the limitation of RAM on Colab, the dataset had to be limited to a size below 80k images. Moreover, all experiments shown were run with 50 epochs and SIZE equal to 20k. (SIZE represents the number of images of each class) This approach was chosen due to hardware limitations and to ensure a balanced dataset.

B. Experiment Metrics

In order to evaluate our model, the following metrics were used:

1. *Accuracy*
2. *F1 score for IDC- and IDC+*
3. *Recall*
4. *Recall for IDC+*

These metrics are examined in conjunction to choose a robust and accurate model. A higher recall is valued because the cost of false negatives in this case is very high as it leads to a patient potentially losing their lives.

C. Architecture Modification Results

Two different models were experimented on by varying the number of convolutional and max pooling layers. The first model has the architecture described above in *Fig. 1* with a one-dimensional vector output, sigmoid activation function, and an Adam optimizer (Archi 1). The second model has a similar architecture to *Fig. 1*. However, the output is a two-dimensional vector output, with SoftMax activation function, and a sgd optimizer (Archi 2). Both architectures also have l2 regularization applied on each convolutional layer to avoid overfitting.

TABLE II. Archi 1 and 2 Evaluations

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 2 Metrics					
# Of Conv Layers	***F1 IDC-***	***F1 IDC+***	***Accuracy***	***Recall IDC+***	***Recall***
2	86	87	86	89	86
3	81	85	83	94	83
4	85	87	86	92	86
5	84	87	86	93	86

The underlined entries in TABLE II are the models that are evaluated as best in the respective architectures. They will be referred to as Model 1 (Archi 1 with 4 convolution layers) and Model 2 (Archi 2 with 4 convolution layers)

An experimental architecture that is similar to a DCGAN discriminator was also experimented with. The outcome of this model had an accuracy of 93%. However, the validation loss kept increasing which indicated overfitting. Further research into this architecture could be done.

The challenges mentioned before were addressed by testing different architectures of a simple CNN to arrive at the one with the best performance.

D. Model Surveying Results

TABLE III. Model Surveying Results

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

Logistic regression and random forests were implemented using Sklearn library. The dataset was also reshaped to fit the data into the respective models. An implementation of support vector machines was tested at SIZE equal to 4k and resulted in an accuracy of 90%. However, this model proved to be very bad when SIZE was scaled to 20k. These models were chosen due to their common applications in classification.

Both implemented architectures in this paper outperformed the other implemented models based on all metrics.

E. Saliency Map Evaluation Results

The saliency maps were generated using Archi 2 with 4 convolutional layers. This was done because the Saliency library required the output of the model to have SoftMax.

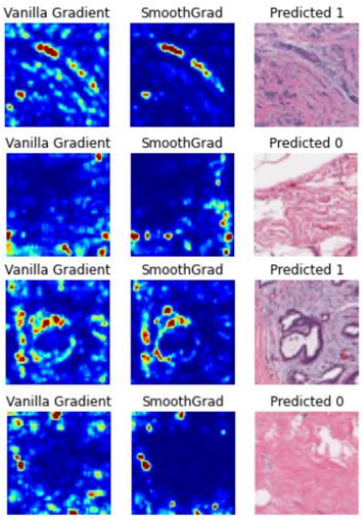

Fig. 2. Saliency Maps

Given the lack of expertise in Pathology in the team, the saliency maps were not evaluated accurately. However, the team noticed by careful inspection that the IDC+ cases had a more purple color. The saliency maps are able to capture this initial observation. As seen in the first and third saliency maps in *Fig. 2*, the dark purple outlines were highlighted by the maps indicating that the model was able to abstract this simple rule. In the second and third saliency maps in *Fig.2*, the tiny purple spots were highlighted indicating that the model is looking for these spots when evaluating breast cancer. Comparing the Vannila Gradient maps to the SmoothGrad maps, it is pretty evident that the SmoothGrad maps captured less noise and showed the more defining features of the histopathological image indicating it is a better map. This observation indicates that the saliency maps are a success and hence a supplement to the model's prediction. This will allow pathologists to get an inside view into how a model is "thinking".

F. Further Discussion and Additional Experiments

The main incentive behind why Archi 2 was developed is to implement saliency maps using the Saliency library. This output however provided the team with two good working models. They are evaluated as "good" due to the balance in metrics and higher recall score compared to the other proposed models of the respective architectures. As mentioned previously, both models were evaluated using SIZE of 20k. Further experiments were conducted by varying the variable SIZE, in order to see how the respective models will perform on bigger datasets, further disecting each model's performance. This experiment further addresses the challenges mentioned with ensuring we have a simple yet stable model.

TABLE IV. MODEL 1 AND MODEL 2 WITH VARYING SIZE

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 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

This experiment allowed both models to be compared to more exposure on data. The data in TABLE IV shows that as SIZE increases Model 2 is able to perform better. The model was able to maintian a constant accuracy as SIZE increased, while Model 1 had decreasing accuracy as SIZE increased. This shows that Model 2 is able to scale better with a larger dataset. Furthermore, Model 2 outclassed Model 1 in every metric except the Recall for IDC+. Making the decision to sacrifice a bit of the recall score was prefered due to Model 2's better scalability. As mentioned before further increase in SIZE would make Colab crash.

To further insepct the difference between Model 1 and Model 2, the loss vs epoch and accuracy vs epoch graphs were recorded.

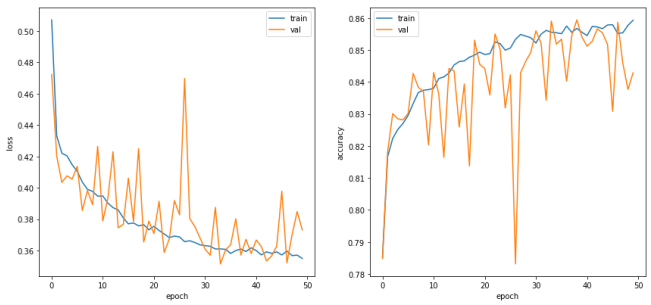

Fig. 3. Model 1 {SIZE:40k , epochs:50}

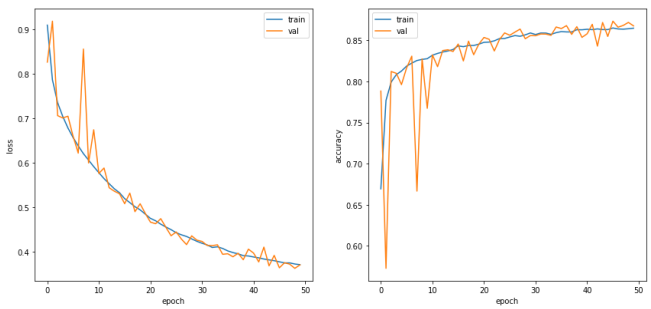

Fig. 4. Model 2 {SIZE:40k , epochs:50}

As evident from the graphs, Model 2 had less "spikes" and a more stable graph than Model 1's graphs. This indicates that Model 2 was able to abstract the rules needed to predict the two classes better.

VIII. Conlusion

As mentioned earlier hardware limitations halted further development of the model. Future study is required with increasing SIZE to include all the IDC+ images. (SIZE = 78k). The dataset should not be increased further to encompass all the IDC- images in order to avoid the model learning how to predict IDC- better than IDC+ resulting in an imbalanced model. There should also be further study on examining the effect of image magnification on the accuracy of the model. In Xia et al [19] the model had true positive rates compared with magnification. The model in this study had a decreasing true positive rate as magnification increases. Further research should be conducted on the effect of magnification on the model in our paper. Furthermore, the study had a fairly complex algorithm that probably lowers the explainability of the model. Although the model in our paper is simple, it's metrics were comparable to the study.

Another issue discovered is that the dataset included some images that have no value, as seen in the outlined samples in *Fig. 5*.

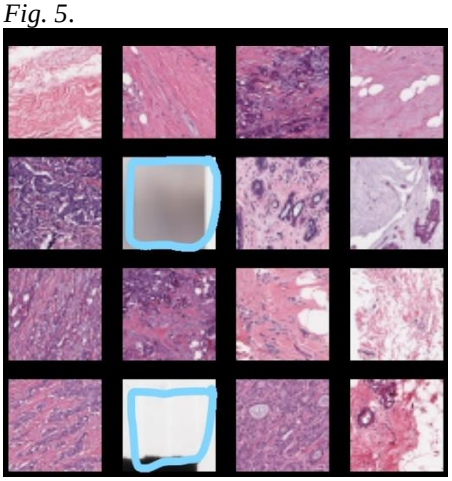

Fig. 5. Dataset Outliers

These discrepancies appeared a few times while running experiments and they acted as outliers to the model. Further preprocessing is required to weed out these outliers.

In comparison with the state-of-the-art model in S. Awadh Alanazi et al [18], which was trained on the whole dataset and the only recorded metric was accuracy, the model in our paper had less layers and a more balanced dataset and resulted in a similar accuracy.

Finally, pathologists need to be included in the development process. If a pathologist is included, the saliency maps can be better evaluated. Moreover, insights on how a pathologist actually diagnoses breast cancer will surface. The insights from the pathologist could allow a the team to create a model that encompasses the proper diagnosing process. The team in this paper is composed of engineers who have no knowledge on the evaluation of any histopathological smaple, other than through mere observation.

References

[1] A. C. S. Medical and Editorial Team, "Types of breast cancer: Different breast cancer types," *American Cancer Society*. https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html.

[2] Memorial Sloan Kettering Cancer Center. 2021. *The Role of Pathology*.https://www.mskcc.org/cancer-care/diagnosis-treatment/diagnosing/role-pathology

[3] National Cancer Institute. 2021. *Common Cancer Types*. https://www.cancer.gov/types/common-cancers

[4] Novartis, "Artificial intelligence decodes cancer pathology images," *Novartis*. https://www.novartis.com/stories/artificial-intelligence-decodes-cancer-pathology-images.

[5] Pathai.com. 2021. *PathAI* https://www.pathai.com/what-we-do/

[6] Pathologists, T., 2021. *Histopathology*. [online] Rcpath.org https://www.rcpath.org/discover-pathology/news/fact-sheets/histopathology.html

[7] Paul and Perkins. 2021. *Cancer Statistics by Type of Cancer*. https://paulandperkins.com/cancer-statistics/?cfchlmanaged_tk=pmd_h019yW9N8zmusFLY20.YrY8GtyLVSL2BTGh4Orl1QQ4-1633802094-0-gqNtZGzNAyWjcnBszRMl

[8] P. Mooney, "Breast histopathology images," *Kaggle*, 19-Dec-2017 https://www.kaggle.com/paultimothymooney/breast-histopathology-images.

[9] ResearchGate, "Issues in training a convolutional neural network model ..." https://www.researchgate.net/publication/334541650_Issues_in_Training_a_Convolutional_Neural_Network_Model_for_Image_Classification.

[10] S. Alanazi et al., "Boosting Breast Cancer Detection Using Convolutional Neural Network", 2021. .

[11] T. S. Sheikh, Y. Lee, and M. Cho, "Histopathological classification of breast cancer images using a multi-scale input and multi-feature network," *Cancers*, 24-Jul-2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465368/.

[12] *Saliency maps in tensorflow 2.0*. Saliency Maps in Tensorflow 2.0 · UR Machine Learning Blog. (n.d.). Retrieved November 10, 2021, from https://usmanr149.github.io/urmlblog/cnn/2020/05/01/Salincy-Maps.html.

[13] Mundhenk, T. N., Chen, B. Y., & Friedland, G. (2020, March 9). *Efficient saliency maps for explainable AI*. arXiv.org. Retrieved November 10, 2021, from https://arxiv.org/abs/1911.11293

[14] Lu, X., Tolmachev, A., Yamamoto, T., Takeuchi, K., Okajima, S., Takebayashi, T., Maruhashi, K., & Kashima, H. (2021, August 30). *Crowdsourcing evaluation of saliency-based XAI methods*. arXiv.org. Retrieved November 10, 2021, from https://arxiv.org/abs/2107.00456

[15] Saha, S. (2018, December 17). *A comprehensive guide to Convolutional Neural Networks-the eli5 way*. Medium. Retrieved November 10, 2021, from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[16] *What is a convolutional layer?* Databricks. (2020, May 15). Retrieved November 10, 2021, from https://databricks.com/glossary/convolutional-layer.

[17] Pokharna, H. (2016, July 28). *The best explanation of Convolutional Neural Networks on the internet!* Medium. Retrieved November 10, 2021, from https://medium.com/technologymadeeasy/the-best-explanation-of-convolutional-neural-networks-on-the-internet-fbb8b1ad5df8.

[18] Alanazi, S., Kamruzzaman, M., Islam Sarker, M., Alruwaili, M., Alhwaiti, Y., Alshammari, N. and Siddiqi, M., 2021. Boosting Breast Cancer Detection Using Convolutional Neural Network. hindawi.com

[19] X. Li, X. Shen, Y. Zhou, X. Wang, and T.-Q. Li, "Classification of breast cancer histopathological images using interleaved DenseNet with Senet (idsnet)," PLOS ONE. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0232127#pone.0232127.ref009

[20] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact CNNS - BMC Medical Informatics and Decision making," BioMed Central, 22-Oct-2019. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0913-x

[21] Spanhol F A, Oliveira L S, Petitjean C, Heutte L. Breast cancer histopathological image classification using convolutional neural networks[C], IEEE, 2016: 2560–2567.

[22] Zhang, R.; Weber, C.; Grossman, R.; Khan, A.A. Evaluating and interpreting caption prediction for histopathology images. In Proceedings of the 5th Machine Learning for Healthcare Conference, in PMLR, Online Metting, 7–8 August 2020; Volume 126, pp. 418–435