Supervised Deep Learning applied to Breast Tumor Classification of Mammography images

Nidhi Joshi

SUPERVISORS:Nishant Ravikumar, Sulaiman Vesal, Prathmesh Madhu, Andreas Maier

***Abstract*--In this paper, we implement two deep learning algorithms to classify the mammography images of breast into cancerous and normal. The dataset comprises of 1415 mammograms provided by the Frauen Klinikum, Women's Hospital of the University Hospital Erlangen. We implemented two convolutional neural network (CNN) architectures and experimental results have been summarized. We obtain an accuracy of 68% and 57.14% respectively on the two architectures on the test data.**

***Index Terms*--Breast Cancer, CNN, Mammography, Supervised Learning.**

I. INTRODUCTION

Breast cancer is the most common cancer in women and the main cause of death from cancer among women in the US after skin cancer. Women in US have a "1 in 8" (or about **12 percent**) lifetime risk of getting breast cancer [1].

Any region that does not look like a normal tissue is a possible cause for concern. Particularly, the region of white, high-density tissue is of interest to the radiologists since it is a sign of potential tumor. Mammograms are X-ray images of a breast that can reveal early signs of breast cancer. Mammograms uses low-dose x-rays to detect cancer. A digital breast tomosynthesis computer aided diagnosis (CAD) system can allow detection of a large percentage (89%, 99 of 111) of breast cancers manifesting as masses and microcalcification clusters, with an acceptable false-positive rate (2.7 per breast view) [2]. Hence, CAD systems established methods for robust assessment of medical image-based examination. In this regard, image processing introduced a promising strategy to facilitate tumor grading and staging, while diminishing unnecessary expenses [3].

Although the usage of hand crafted features in computer aided diagnosis was historically very successful, however they still suffered from several drawbacks [6][5][4]: (1) hand crafted features were not generalizable, they worked well for specific cases and failed terribly on even similar problems [7]; (2) they were influenced by the radiologists in their techniques of drawing conclusions and were thus not unbiased (3) multi-modal datasets are only integrated in ad-hoc fashion.

Hence more robust and generalizable techniques were required. Deep learning models could derive predictive transformations via a data-based mathematical optimization process that penalizes inconsistencies between model output and ground truth, without the need of hand crafted features. These models try to mimic human brain which learn and figure out the features on their own. While deep learning models have exhibited encouraging performance in breast cancer imaging analysis tasks [9], these models are often treated as black boxes, and relating high-level features to clinically relevant phenomena has proven difficult [9]. In addition, they help in the detection of the development of secondary malignant growths at a distance from a primary site of cancer [8].

Mehdi Habibzadeh Motlagh in his work on histopathological breast cancer image classification used transfer learning approach wih Resnet. [3]. This framework accurately detected on average 99.8% of the four cancer types including breast, bladder, lung and lymphoma using the ResNet V1 50 pre-trained model. In the next step, ResNet V1 152 classified benign and malignant breast cancers with an accuracy of 98.7%. This indicated that the transfer learning approaches are also effective for breast tumor image classification.

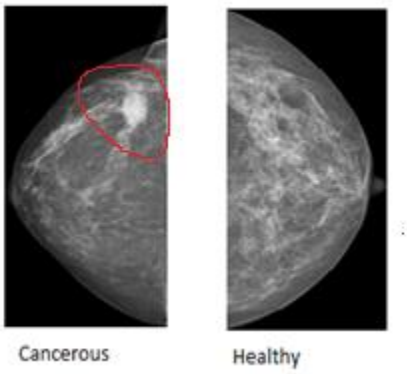
Using convolutional neural networks (CNNs), Darvin Yi was able to achieve an accuracy of 85% and an ROC AUC of 0.91, while leading hand-crafted feature based methods were only able to achieve an accuracy of 71% [10]. They investigate an amalgamation of architectures to show that the best result is reached with an ensemble of the lightweight GoogLe-Nets tasked with interpreting both the coronal caudal view and the mediolateral oblique view, simply averaging the probability scores of both views to make the final prediction [10].

The method proposed by Ribli, 2017 sets the state of the art classification performance on the public INbreast database, AUC = 0.95. The approach described here has achieved the 2nd place in the Digital Mammography DREAM Challenge with AUC = 0.85. When used as a detector, the system reaches high sensitivity with very few false positive marks per image on the INbreast dataset [11].

CNN is a class of deep neural networks, most commonly applied to analyze visual imagery. They were inspired by biological processes in which the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field[12][13][14]15]. In this way, the network learns the patterns in our image such as edges

This work was supported by the Pattern Recognition Lab of the Friedrich Alexander University, Erlangen, Germany.

and corners in the first few layers and later curves and other complex features.



Fig[1] *Cancerous (tumor marked) and healthy breast mammogram*

II. THEORY

Conv layer is used to compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume [16]. Rectified Linear unit (ReLU) is used to introduce non-linear properties to our network. Nonlinear activation functions are important because the function we are trying to learn is usually nonlinear. Max-pooling is used to down-sample our data so that only the most important features are captured and the rest ignored. Flatten converts the data to a single column form which is further given to the fully connected (Dense) layers. As we perform normalization and scaling for our input data, in the same manner we adjust and scale the activations (outputs of hidden layers) using Batch Normalization technique.

III. METHODS

The deep neural network is trained from scratch using the Keras (backend tensorflow) framework. The work largely involved data preparation, model creation, fitting the prepared data to the model and evaluation of the results. K-fold cross validation is used to evaluate the results on the validation dataset.

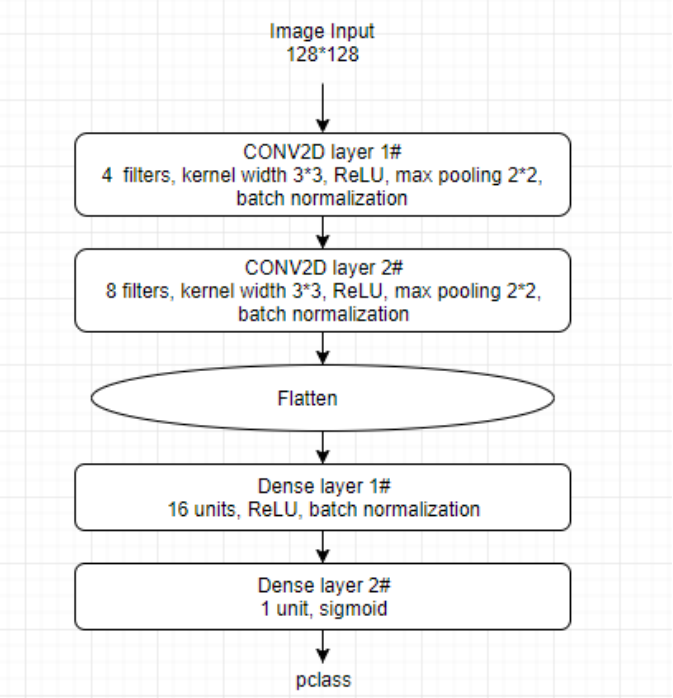
A. Data Preparation

The dataset provided by Frauen Klinikum, Erlangen, consists of 715 cancerous and 700 healthy breast images. The images are divided into training and test set such that randomly 80 percent of both normal and cancerous images go to training and the remaining 20 percent go to the test set. Furthermore, the former training set is again divided in 80-20 ratio for training and validation sets. These training and validation sets are then divided into 5 equal folds to prepare for 5 fold cross validation. All the images have been resized to same size (128*128) for training and testing purposes.

B. Network Architecture

Architecture I.

The first network architecture looks like Fig[3]. A sequential model is used. There are two 2D convolutional layers each followed by two 2D max pooling and two Batch Normalization layers. Following them is a flatten layer. Full connection is achieved with the dense layers as shown in the figure. Final output layer has one neuron with a binary class based classifier- sigmoid function. All parameters for the layers including kernel size and the number of filters are mentioned in the architecture snippet as follows :



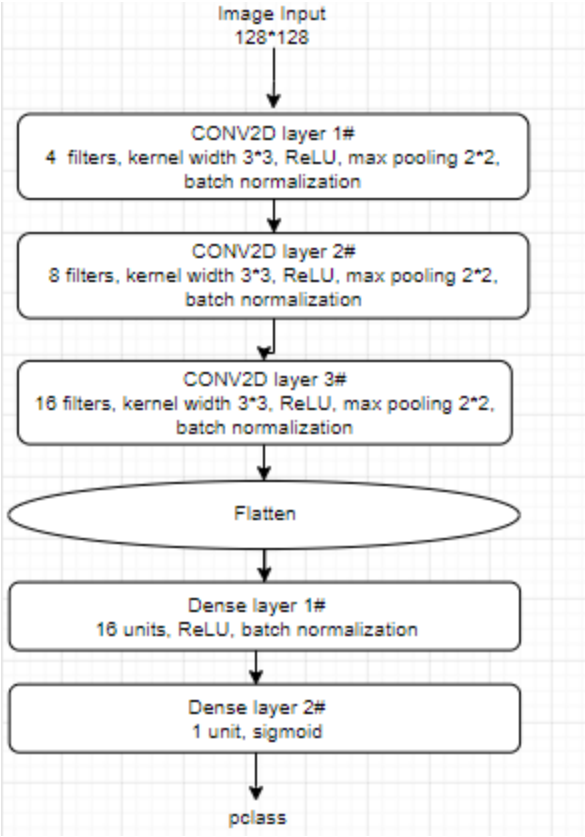
pclass:class confidence

Fig[2] *Architecture I.*

Early stopping with Patience criteria was set to 10 epochs. ‘Adam’ optimizer and binary cross entropy loss function are been used to compile the model.

Architecture II.

The second network architecture looks like [Fig4]. A sequential model is used. There are three 2D convolutional layers each followed by three 2D max pooling and batch normalization layers. Following them is a flatten layer. Full connection is achieved with the dense layers as shown in the figure. Final output layer has one neuron with a binary class based classifier- sigmoid function. All parameters for the layers including kernel size and the number of filters are mentioned in the architecture snippet as follows :



pclass:class confidence

Fig[3] *Architecture II.*

Early stopping with Patience criteria was set to 10 epochs. Again, 'Adam' optimizer and binary cross entropy loss function are been used to compile this model.

IV. RESULTS AND DISCUSSION

C. Performance on validation and unseen test set

The 5 fold validation technique has been used to evaluate and fine tune the model. Precision, Recall and Accuracy, along with F1-score statistics for validation data has been presented in Table 1-4.

Table 1 *: Evaluation for five folds on validation dataset for architecture I.*

	Architecture 1		
Test data	Precision	Recall	F1-score
Model 1	0.8346	0.5904	0.6916
Model 2	0.812	0.8504	0.8308
Model 3	0.8797	0.9435	0.9105
Model 4	0.9699	0.5265	0.6825
Model 5	0.8421	0.7467	0.7915

As observed from the validation results of the 5 sets, Architecture I performs well on the data, and hence in order to validate our hypothesis of generalization, it should also perform well on the totally unseen test data. However, to understand and interpret the results better, we generated test results on all the 5 trained models and they are summarized in Table 2 as below:

Table 2 *: Evaluation for five folds on unseen test dataset for architecture I.*

	Architecture 1		
Test data	Precision	Recall	F1-score
Model 1	0.7662	0.5	0.6051
Model 2	0.7013	0.5625	0.6243
Model 3	0.7143	0.567	0.6322
Model 4	0.7403	0.57	0.6441
Model 5	0.8553	0.6436	0.7345

Table 3 *: Evaluation for five folds on validation dataset for architecture II.*

	Architecture 2		
Test data	Precision	Recall	F1-score
Model 1	0.2556	0.6296	0.3636
Model 2	0.7293	0.533	0.6159
Model 3	0.5263	0.7692	0.625
Model 4	0.8647	0.6928	0.7692
Model 5	0.8947	0.6398	0.7461

Again, we will be testing the generalization capability of the trained model, Architecture II based on its performance on the unseen test data. Evaluation is presented in Table 4 as follows:

Table 4 *: Evaluation for five folds on unseen test dataset for architecture II.*

	Architecture 2		
Test data	Precision	Recall	F1-score
Model 1	0.2208	0.5862	0.3208
Model 2	0.6494	0.495	0.5618
Model 3	0.4545	0.6604	0.5385
Model 4	0.7403	0.57	0.6441
Model 5	0.7092	0.5125	0.6367

In order to understand and discuss the impact of the results more clearly, we also include the confusion matrices for both the architectures on the unseen test set for the best validation fold.

For Architecture I

	True Cancer	True Normal
Predicted Cancer	65	11
Predicted Normal	36	35

Accuracy:68%

Fig[4]

For Architecture II

	True Cancer	True Normal
Predicted Cancer	57	20
Predicted Normal	43	27

Accuracy: 57.14%

Fig[5]

The above neural network architecture I performs fairly well on given dataset. Apart from accuracies, factors like how frequently a model classifies a case of true cancer into normal matters a lot. Those are the cases that are very relevant from medical point of view. The reverse of it, i.e. classifying a true normal to cancerous (False positives) is not something which has to be relatively paid for so heavily. So, to better train the model penalty could be introduced which penalize the False negatives (actually cancerous falsely classified as normal) over the other cases. Again, Architecture I performs better in terms of precision and recall values as seen from Table 1 and 2. Apart from this, other factor to be taken care of is class imbalance but since we did not have that case, we could skip thinking about compensating it.

V. CONCLUSION

1. Adding extra convolution and pooling layers did not improve the results. In fact it overfits to the model due to unavailability of very large image dataset. In case of architecture II, as the training accuracies reach very high values, validation accuracies start falling which we conclude is a case of overfitting. The number of trainable parameters play a very important role in this case. If kernel size and number of filters are too many the model memorizes the results however overfits to it and cannot generalize the scenario well.

2. It is much easier to handle medical images if the imaging modalities have a standard imaging and resolution system. Else pre-processing techniques like image resizing and bounding box generation have to be developed.

3. The data should be representative of its labels. Only then the deep learning models can learn them well.

VI. REFERENCES

[1] American Cancer Society. Cancer Facts and Figures 2019. Atlanta, Ga: American Cancer Society; 2019.

[2] Lia Morra , Daniela Sacchetto, Manuela Durando, Silvano Agliozzo, Luca Alessandro Carbonaro, Silvia Delsanto, Barbara Pesce, Diego Persano, Giovanna Mariscotti, Vincenzo Marra, Paolo Fonio, Alberto Bert. From the Department of Research and Development, im3D, Via Lessolo 3, 10153 Turin, Italy (L.M., D.S., S.A., S.D., D.P., A.B.); Department of Radiology, University of Turin, Turin, Italy (M.D., G.M., P.F.); Department of Diagnostic Imaging and Radiation Therapy, Radiology University of Torino, Azienda Ospedaliero Universitaria Città della Salute e della Scienza di Torino, Turin, Italy (M.D., G.M., P.F.); Unità di Radiologia, IRCCS Policlinico S. Donato, Milan, Italy (L.C.); C.d.C. Paideia, Rome, Italy (B.P.); and Department of Radiology, Sant'Anna Hospital, Turin, Italy (V.M.).

[3] Breast Cancer Histopathological Image Classification: A Deep Learning Approach by Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha.

[4] Alan Bekker and et.al. Multi-view probabilistic classification of breast microcalcifications. IEEE Transactions On Medical Imaging, 35(2):645–653, 2016.

[5] Jinghui Chu, Hang Min, Li Liu, and Wei Lu. A novel computer aided breast mass detection scheme based on morphological enhancement and slic superpixel segmentation. Medical Physics, 42(7):3859–3869, 2015.

[6] Maryellen L Giger, Heang-Ping Chan, and John Boone. Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. Medical physics, 35(12):5799–5820, 2008.

[7] David A Gutman and et.al. Mr imaging predictors of molecular profile and survival: multiinstitutional study of the tcga glioblastoma data set. Radiology, 267(2):560–569, 2013.

[8] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718, 2016.

[9] Thijs Kooi and et.al. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis, 35:303–312, 2017.

[10] Darvin Yi and et.al Optimizing and Visualizing Deep Learning for Benign/Malignant Classification in BreastTumors, arXiv:1705.06362v1 [cs.CV] 17 May 2017.

[11] Dezs˝o Ribli and et.al. Detecting and classifying lesions in mammograms with Deep Learning, arXiv:1707.08401v3 [cs.CV] 9 Nov 2017.

[12] Fukushima, K. (2007). "Neocognitron". Scholarpedia. 2 (1): 1717. doi:10.4249/scholarpedia.1717.

[13] Hubel, D. H.; Wiesel, T. N. (1968-03-01). "Receptive fields and functional architecture of monkey striate cortex". The Journal of Physiology. 195 (1): 215–243. doi:10.1113/jphysiol.1968.sp008455. ISSN 0022-3751. PMC 1557912. PMID 4966457.

[14] Fukushima, Kunihiko (1980). "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position" (PDF). Biological Cybernetics. 36 (4): 193–202. doi:10.1007/BF00344251. PMID 7370364. Retrieved 16 November 2013.

[15] Matusugu, Masakazu; Katsuhiko Mori; Yusuke Mitari; Yuji Kaneda (2003). "Subject independent facial expression recognition with robust face detection using a convolutional neural network" (PDF). Neural Networks. 16 (5): 555–559. doi:10.1016/S0893-6080(03)00115-1. PMID 12850007. Retrieved 17 November 2013.

[16] CS231n: Convolutional Neural Networks for Visual Recognition, cs231n.stanford.edu.