**Mammogram Image Preparation Techniques for Optimal AI/DL-based Breast Cancer Diagnosis using Transfer Learning approach**

**Abstract**:

Applying Convolutional Neural Networks for mammogram classification toward breast cancer diagnosis can be very challenging due to a multitude of reasons. Various health care facilities employ a variety of mammogram apparatus and screening technologies. Therefore the image characteristics do vary across different available datasets and over time as well as new screening technologies evolve. This challenge is particularly relevant to the Transfer Learning stream of AI/DL technology, where in the images’ spatial and pattern characteristics from the source dataset the transfer learning is inherited from most likely differs from that of the target data set being trained upon. This experimental survey research work explores a few research works on transfer learning, identifies some challenges involved, and articulates some methods to have the data set images prepared for better data modeling. The methods include scaling the images, normalizing the pixel values, and curating the images with required augmentations such as vertical and horizontal flips, cropping, etc. In addition, this research work also identifies the importance of the image data augmentation technique that guarantees the improvement of model performance and also helps facilitate to reduce the generalization error, which is otherwise possible only with large dataset sizes. This research work provides the test results of test-time augmentation to showcase how it improves the predictive performance of a fit model as opposed to that without augmentation. Train-time augmentation is not performed or benchmarked.

**Introduction**

Breast carcinoma is one of the most invasive tumors that account for cancer deaths in females1. In 2020, an estimated 684,996 women across the world died from breast cancer16 .Early diagnosis through mammogram screening is key and extremely critical to reducing mortality2, and start early treatment.It is proven that early diagnosis increases survivability and decreases the overall mortality rate.

Recently, Artificial Intelligence (AI) technology has been explored to assist in breast cancer diagnosis. Since most medical records are being uploaded digitally these days, it provides the opportunity to utilize Deep Learning to make more informed decisions. There are various imaging modalities that radiologists, oncologists, and pathologists typically use for breast cancer diagnosis - Digital Mammography (DM), Ultrasound, Magnetic Resource Imaging (MRI), and Histopathology (HP) being the most prominent of them. However, most of this work has been done with curated datasets that are not directly transferable to real-world data encountered in practice. Naturally, they don’t offer any means where a radiologist or patient can use the raw mammogram image to directly get an output in real-time.

This experimental survey research work aims at using the raw mammogram images as a data set in training the Deep Learning-based data model. This survey work attempts to identify the some challenges faced in AI/DL-based technology that serves as an assistive tool in having patients or radiologists take a second opinion on breast cancer diagnosis. The patient or radiologist simply supplies the raw (without any manual clinical annotations) mammogram image to a website that runs the AI-program in the backend. Such a tool becomes very useful, especially in underserved rural towns in developing countries where there are many misdiagnoses and under-diagnosis of breast cancer resulting in breast cancer being unnoticed until it’s too late.

**Challenges to Research Community**

***Challenge 1: Dataset Availability Issue***

Healthcare is a sensitive area when it comes to making mammogram images available to the research community. There exists a handful of facilities and organizations to publish mammogram datasets like CBIS-DDSM14, Mammographic Imaging Analysis Society (MIAS) database4, and INbreast15 towards fostering AI/DL-based research.

***Challenge 2: Dataset Size Issue***

Success of DL lies in being able to automatically learn relevant features from data, which requires a large enough dataset in most cases. Techniques on natural images are not directly transferable to medical AI due to low dataset size and the tedious process of collecting annotations. However, a majority of the available mammogram data sets such as INbreast and MIAS are too small in size (in the low hundreds). This makes the classical DL-based model training, which involves training from scratch, very inefficient in terms of time, computing power and storage needs.

***Challenge 3: Varied Imaging Modality***

Each mammogram dataset surveyed in this work is different from the other data set, in terms of the mammogram technology used and in terms of various aspects of the image pattern. For example, the INbreast dataset uses Full-Field Digital Mammogram (FFDM) while the CBIS-CBIS-DDSM dataset uses Digitized screen-film Mammograms (DFM) technology.

**Challenge 4: Small Region Of Interest Issue**

Achieving a fully automated AI/Dl-based breast cancer diagnosis has been a challenging problem as the typical DL-based generic image classification task can’t be directly applied to mammogram images. This is mainly because, in a given mammogram image, the actual Region Of Interest (ROI) representing a potential malignant region takes only a very small portion ( can be as small as 120x120 pixels) of the whole image of the breast. Surrounding this reason, some of the research studies8-13 scoped their work that applies to classifying based on manually pre-annotated lesions.

**Challenge 5:**

For AI/DL-based breast cancer diagnosis programs that employ transfer learning, the research community faces various challenges in terms of having the mammogram images in a certain format, certain dimension, and require certain pre-processing. This research work identifies these main challenges and proposes the techniques and solutions with some test results.

**Transfer Learning**

The premise behind transfer learning33 is that researchers can take a model already trained on a large dataset and transfer the learnt features to a smaller target dataset. While AI/DL-based breast cancer detection via mammogram image classification usually employs full model training based on Convolutional Neural Networks (CNN) based architecture, the approach of transfer learning involves freezing the early convolutional layers of the network and only training the last few layers which make a prediction. The idea is that the convolutional layers extract general, low-level features that are applicable across images — such as edges, patterns, and gradients — and the later layers identify specific features within a mammogram image such as malignant lesion pattern, benign lesion pattern, mass pattern, etc. There exist some state-of-the-art architectures like VGG16 and ResNet50 which are used as genesis pre-trained models for applying transfer learning techniques. VGG16 or ResNet50 architectures provide a pre-trained model that’s trained with millions of images (ImageNet7), and any image classification problem with a smaller dataset can directly leverage many image semantics and characteristics from such a humongous image data set and produce much better metrics related to model predictions.

 The research work by Li Shen3 develops a deep learning algorithm that can accurately detect breast cancer on screening mammograms using an “end-to-end” training approach that efficiently leverages training datasets with either complete clinical annotation or only the cancer status (label) of the whole image. In this approach, lesion annotations are required only in the initial training stage, and subsequent stages require only image-level labels, eliminating the reliance on rarely available lesion annotations. The largest public database with ROI annotations for Digitized Film Mammograms – CBIS-DDSM – contains several thousand images with pixel-level annotations, which can be exploited to train a patch classifier function ‘f’ that is first trained on patches and then refined on whole images. The research work by Li Shen demonstrates that a whole image classifier trained using end-to-end approach on the CBIS-CBIS-DDSM digitized film mammograms can be transferred to INbreast30 FFDM images using only a subset of the INbreast data for fine-tuning and without further reliance on the availability of lesion annotations. These findings show that automatic deep learning methods can be readily trained to attain high accuracy on heterogeneous mammography platforms

**Transfer Learning Chain:**

This survey work explores and identifies best practices in applying the transfer learning process to mammogram image classification problems. Transfer learning takes the shape of the chain process, where in we start with a state-of-the-art trained data model on generic large image datasets, then transfer its model knowledge to train the next biggest available mammogram dataset, and so on. The following chain applies to this survey and research work that corresponds to AUC ROC test results on INbreast Dataset:



ImageNet

Dataset

**Model Training**

Using VGG16 architecture with top 2 layers :

[512-512-1024]x2

Saved .h5 model file

CBIS-DDSM MammogramsDataset

**Transfer Learning**

Use pre-trained model

+

Fine Tune model parameters specific to CBIS-DDSM dataset

Saved .h5 model file

**Transfer Learning**

Using pre-trained model on to the INbreast Dataset and Fine Tune model parameters specific to INbreast dataset

INbreast Mammograms Dataset

**Mammogram Data Preparation Pipeline**

DICOM–format mammogram images

PNG-format image files ready for Mammogram Image Preparation Phase in model training

Unified Image Converter Utility Program

PGM –format mammogram images

…….

Any Other format mammogram images

**Mammogram Images Preparation in the Transfer Learning centric Image Classification - Best Practices**

**Data Preparation Phase:**

Convert any 3-channel images to Grayscale 2D images

Pixel Scaling – center & normalize the pixel values

Resize all images to one uniform shape.

Select the image format (PNG, JPEG etc) and fixed size

Train time - **Data Augmentation Phase:**

Compile all the images ready for training

Modify Images:

-Shift

-Flips

(horizontal& vertical)

Resize and Crop the input images

Test time - Conventional style  **Data Augmentation Phase:**

Each image is ready for prediction on the fit model

Crop each rescaled image

Rescale each image

Test time - **Modified Data Augmentation Phase:**

Crop each rescaled image

Perform 4 flips (Horizontal and vertical) to produce 4 versions of each image

Predict on each of the 4 versions of the same original image.

Take a mean value of the 4 predictions and assign and represent this mean predicted value for the image.

Rescale each image

**Computational Environment**

Hardware: This work used Google Colab Pro+, with GPU hardware accelerator and High-RAM

Software: This work used the Keras Framework v2.0.8, TensorFlow 1.3, and Python 3.5

**Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Test Data set size | Image Dimension Used | Base Architecture of the pre-trained model( Transfer Learning) | Top 2 Layers ( 2 blocks) | AUC conventional-style Test-time Data Augmentation | AUC with Modified Test-time Data Augmentation technique |
| INBREAST | 410 images | 1152 x 896 | VGG16 | [512-512-1024]x2 | [0.93894267 0.93894267] | [0.95820923 0.95820923] |

**Code**

The source code that corresponds to the techniques mentioned in this work and to obtain the above results are available to the public at this GitHub account.

<https://github.com/Ritz313/end2end-all-conv>

**References**

1. American Cancer Society. How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/howcommon-isbreast-cancer.html (2018).

2. Oefnger, K. C. et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA 314, 1599–1614 (2015).

3. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep. 2019 Aug 29;9(1):12495. DOI: 10.1038/s41598-019-48995-4. PMID: 31467326; PMCID: PMC6715802.

4. Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., et al. (2015). *Mammographic Image Analysis Society (MIAS) database v1.21* [Dataset]. https://www.repository.cam.ac.uk/handle/1810/250394

5. Srikanth Tammina (2019); Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images; International Journal of Scientific and Research Publications (IJSRP) 9(10) (ISSN: 2250-3153), DOI: http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9420

7. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

8. Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Analysis 35, 303–312 (2017).

9. Jamieson, A. R., Drukker, K. & Giger, M. L. Breast image feature learning with adaptive deconvolutional networks. Proc. SPIE 8315, (6–13 (2012).

10. Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L. & Lopez, M. A. G. Convolutional neural networks for mammography mass lesion classifcation. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 797–800, https://doi.org/10.1109/EMBC.2015.7318482 (2015).

11. Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L. & Guevara Lopez, M. A. Representation learning for mammography mass lesion classifcation with convolutional neural networks. Comput. Methods Programs Biomed. 127, 248–257 (2016).

12. Lévy, D. & Jain, A. Breast Mass Classifcation from Mammograms using Deep Convolutional Neural Networks. arXiv preprint arXiv:1612.00542 (2016).

13. Dhungel, N., Carneiro, G. & Bradley, A. P. The Automated Learning of Deep Features for Breast Mass Classification from Mammograms. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G. & Wells, W. (eds) Medical Image Computing and ComputerAssisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II, 106–114 (Springer International Publishing, Cham, 2016).

14. Lee, R. S., Gimenez, F., Hoogi, A. & Rubin, D. Curated Breast Imaging Subset of DDSM. The Cancer Imaging Arch (2016).

15. Moreira, I. C. et al. INbreast: Toward a Full-field Digital Mammographic Database. Acad. Radiol. 19, 236–248 (2012).

16. cancer.net : Breast Cancer - Statistics

https://www.cancer.net/cancer-types/breast-cancer/statistics#:~:text=It%20is%20estimated%20that%2043%2C780,world%20died%20from%20breast%20cancer.