Breast Cancer Diagnosis and Prognosis Using Artificial Neural Network

O. S Ogah, and J.O. Ugah,
Dept. of Computer Science EBSU

Abstract

As the mortality rate of breast cancer increases, there is the need to foster efforts to combat this ugly disease and save our mothers and all female folks from untimely death. Various methods being used to diagnosis this disease range from traditional hand sampling of the breast to a more advanced and generally accepted technique known as biopsy, a fine needle aspirations that enable early detection of the disease. Manual analysis of the biopsy image by clinical pathologist have many limitations, hence the need of computer aided approach. Extensive works had been carried out using this approach but more works still have to be done as a better model to rely on have not been achieved. In this thesis, we employed various deep neural network and machine learning frameworks like keras, tensorflow and lightgbm. We also made use of breast cancer histopathology image dataset release by Rui Yan et al, (2019) and Araujo et al, (2017). We deduced a model whose sensitivity score is higher than that of our redecessor. Reversed CRIP-DM methodology by John Rollins and very deep learning feature representation was used in the analysis. Lightgbm was used for the training and validation. At the end a python script was developed to perform real time breast cancer diagnosis.

Keywords: Biopsy, Breast cancer, Histopathology images, Image classification, Deep neural network, Dataset, Feature representations,

---

Introduction

Cancers have become a major health issues in recent time. According to International Agency for Research on Cancer, a body from the World Health Organizations(WHO), and Global Burden of Disease Cancer Collaboration (GBD), cancer cases have increased by 28% between 2006 and 2016, and it is estimated that 2.7 million new cancer cases will emerge before 2030 (National Cancer Registry, 2013). Out of the different types of cancer, breast cancer is the most common and deadly among women. The World Health Organizations (WHO) has it on record that there are about 1.7 million incident cases from which there are 535,000 death recorded(National Cancer Registry, 2013).

In Nigeria, female breast cancer is seen as major cause of morbidity and mortality. Its incidence rate ranges from 36.3 to 50.2 in every 100,000 live births. Reports from the University College Ibadan teaching hospital, shows that there are 600,000 cases of breast cancer yearly in Nigeria and about 80% of the victims lose their lives (DeSantis et al, 2013).

Therefore, early diagnosis of breast cancer has become necessary. The diagnosis of breast cancers had been since 40 years ago. The use of different methods like breast examination, mammogram (low frequency X-ray), Magnetic Resonance Imaging (MRI), and Ultrasound exist. These methods do not produce accurate breast cancer diagnosis (Abdalla et al., 2009). Biopsy techniques are the main methods relied upon for accurate diagnosis and prognosis of breast cancer. The types of biopsy method used are fine needle aspiration, vacuum-assisted and surgical biopsy. The process involves the collection of samples of cells of breast tumour, fix it on the microscope slide, and then stain it

with Hematoxlyin and Eosin (Veta et al., 2014). After that, the pathological images are then analyzed and diagnosis by the clinical pathologists. However, manual analysis of breast cancer images is difficult and time-consuming. It also requires a well trained histopathologist (Spanhol et al., 2016). More so, the result can be affected by the level of experience of the pathologists involved and also, most time the results are subjected to further re-examination because of discrepancy among pathologist.

Therefore, computer assisted approach for the analysis of pathological images plays a significant role in the diagnosis of breast cancer and its prognosis (Aswathy et al., 2017). Traditional feature extraction methods, like the Scale-Invariant Feature Transform (SIFT) and Gray Level Co-occurrence Matrix (GLCM), rely on supervised information. Furthermore, prior knowledge of data is needed to select useful features. But this approach will produce low quality of extracted features and also, the computational load will be very heavy. At the end of the process, the extracted features will have low level and unrepresentative features of pathological images. And as a result it leads to poor classification and generalization.

Artificial Neural Network methods like the convolution neural network (CNN) can extract features and retrieve information from data automatically. It can also learn advanced abstract representations of data. CNN can solve the problems of traditional feature extraction and have been successfully applied in Computer Vision, He et al.(2015); Xie et al.(2018), Biomedical Science Gulshan et al.(2016); Esteva et al.(2017) and many other fields.

2. Related Works

Breast cancer diagnosis based on image analysis has been studied for more than 40 years, and there have been several notable research achievements in this area. The authors adopted approaches ranged from traditional machine learning algorithm to a more advances deep learning algorithm. The primary aim is to train an existing dataset so as to achieve a reliable accuracy and sensitivity score. Most notable authors like Kowal et al,(2013), Filipczuk et al,(2013),Jackowsli et al,(2012), George et al,(2014), Wang et al,(2016), Asri et al,(2016) made use of clustering algorithms like K-means (KM), fuzzy C-means (FCM), Competitive Neural Network (CNN) and Gaussian Mixture Model (GMM) to perform nuclei segmentation from microscopic image dataset at their disposal. After which that they used different classification approaches namely: K-Nearest Neigbor, Naive Bayes, and Decision Tree, Support vector Machine (SVM) to classify these images into benign and malignant tumours. Though, their research findings showed good accuracy. But there are some challenges:

1. The numbers of datasets used is small.
2. Data collections are from about 50 to 100 patients
3. The dataset are patched microscopic image instead of whole slide image.
4. They rely on binary classification. Either Begnin or Malignant.

With the availability of BreakHis dataset produced by Spanhol et al,(2016), the likes of Spanhol et al,(2016) and Bayramoglu et al, (2016) carried out a histopathology image classification using the BreakHis dataset though still based on binary classification but with improved accuracy as compared to works done using machine learning. Also they made use of deep learning architectures like the AlexNet produced by (Krizhevsky et al., 2012). But the need to carry out multi class classification is paramount. A given predictive model should be able to classify a breast cancer ample into benign or malignant types (i.e InSitu carcinoma or Invasive carcinoma). On this note Araujo et al.(2017), Han et al,(2017), Nawaz et al,(2018), Motlagh et al,(2018), Jiang et al,(2019) used different deep learning algorithms to perform breast cancer multi class classification making use of the Bioimaging2015 and the BreakHis dataset. There were an improve accuracy and sensitivity score. But the number of patients through which the dataset was collected is still considerably small. i.e BreakHis is collected from 82 patients and Bioimaging 2015 is gotten from 249 patients. Rui et al,(2019) published an article with an improve whole side image of 3472 image from more than 1000 patients of diverse age group. Deep trained model deduce from such dataset is believed to robust for future inference of breast cancer patients. Rui et al, (2019), achieve a validation accuracy score of 91.3% and sensitivity score of 89.2%. The aim

of our research work is to improve on their results using their dataset in combination with that of Araujo et al. (2017), deep learning feature representation and machine learning lightGBM algorithm.

TABLE 1: Summary of the Reviews

Authors	No of Data sets	Accuracy
Kowal et al,2013	500 images	94% on binary class
Fillipczuk et al,2013	737 images	98.51% on binary class
George et al,2014	94 image of 1502 of cell nuclei	94% on binary class
Wang et al,2016	68 image of 3600 cell nuclei	96.19% on binary class
Asri et al,2017	699 images	97.13 on binary class
Spanhol et al,2016a	BreakHis dataset	85% on binary class
Spanhol et al, 2016b	BreakHis dataset	87% on multi class
Bayramoglu et al 2016	BreakHis dataset	83.25% multiclass
Araujo et al,2017	Bioimaging2015	77.883% on multi class
Han et al,2017	BreakHis dataset	95% on multi class
Nawaz et al,2018	BreakHis dataset	95.4% on multi class
Motlagh et al,2018	BreakHis dataset	96.4% on multi class
Jiang et al,2019	BreakHis dataset	98.13% on multi class
Rui et al	3771 whole slide images from above 1000 patients	91.3 on multi class

3. Datasets

Deep learning algorithms learn from big amount of data. The emergence of Imagenet, a deep learning model that was produced using about 100k domestic images of 1k classes acts as a break-through in this field of discipline.

Before now there are very few dataset of medical image dataset. Most times, the available ones are not properly labeled because the cost of sourcing clinical experts to label the image is very high. According to Veta et al, (2013), there is the need to have a large, labeled and open pathological dataset to make image analysis in this domain effective. After the publication of Veta et al (2013), there has been some open challenge of medical image analysis in the domain. Amongst them are:

1. The International Conference on Image Analysis and Recognition (ICIAR).
2. The International Symposium for Biomedical Imaging (ISBI),
3. The International Conference on Pattern Recognition (ICPR)
4. The Medical Image Computing and Computer-Assisted Intervention (MICCAI),

The pathological image dataset used in these conferences have demonstrated a precise definition of the tasks assessment metric scores to help in performance comparison. But the challenge is that the dataset is relatively small in size and also the number of patients through which the dataset was obtained is few.

On this noted we decided to rely on:

1.Rui et al, (2019), who in-cooperated with Peking University International Hospital to release a new pathological image dataset of breast cancer. The dataset is in the same format with ICIAR2018 dataset.

2. International Conference on Image Analysis and Recognition (Araujo et al, 2017), from the 15th International Conference on Image Analysis and Recognition.The histopathology images are all digitized with the same acquisition conditions and also with a magnification of 100x or 200x and pixel size of 0.42 × 0.42. Then each image is labeled with one of the classes. We are more concerned about the Benign, InSitu and Invasive classes in this research work.

The overall dataset consists of 3772 high-resolution images of 2048×1536 pixels. The datasets are annotated with Hematoxylin and

Eosin (H&E) stained. The Hematoxylin highlights nuclei by staining DNA and Eosin highlights other structures by staining proteins. used in the pathological routine.

The preparatory procedure for pathological image used in this work was the standard paraffin process, which is widely

TABLE 2: Summary of Dataset

Dataset	Benign	InSitu	Invasive	Total
ICIAR2018	100	100	100	300
Rui et al,(2019)	1106	1066	1300	3472
Overall	1206	1166	1400	3772

In particular, the structure of breasts varies with age, such as puberty, sexual maturity, pregnancy, lactation and old age. To ensure the diversity of data to improve the learning ability in the machine learning algorithm, our dataset covers as many different subsets spanning different age groups as possible, which can fully reflect the morphology of breast tissue

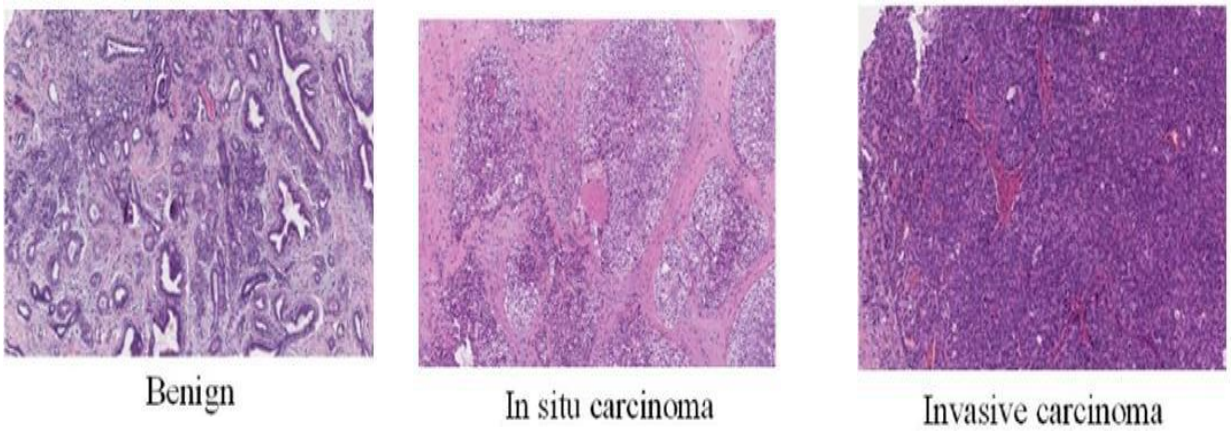


Fig. 1: Examples of microscopic biopsy images in the dataset 4

METHODS

Given a breast cancer pathology image, our goal is to classify it into its classes. That is whether it is Benign, InSitu or Invasive. Many deep convolution Neural Network architectures like the ResNet50, Vgg16 etc have achieved the state-of- the- art result when training on ImageNet of 10million image of 1k classes. Training these neural networks from scratch requires a large number of images as compare to the dataset we have at our disposal. As a result of this, we will experience overfitting if we go ahead to training our dataset with these very deep CNN. In order to avoid this, we employ transfer learning strategy, (a process by which only a part of the pre-trained neural network is being fitted to a new dataset).

The adopted approach is called deep convolutional feature extraction. So, we made used of ResNet50 architecture to carryout unsupervised feature extraction. This helped us to obtain the image sparse descriptor of low dimensionality. This dimensionality reduction plays great roles in combating overfitting during the supervised training stage.he extracted rich feature is then trained using lightgbm, a distributed gradient boosting framework for machine learning. We used lightgbm because of its high performance and also tolerance to overfitting

4.1 Data Normalization

The dataset was stained by ***Hematoxylin*** (H) and ***Eosin*** (E) during preparation. So in other to make it ready for any improved quantitative analysis, the amount of **H&E** in each image needs to be normalized. The normalization algorithm was deduced from the work of Ruifrok, et al, (2001). The aim of the algorithm was to deconvolve the color formation acquired with red-green-blue (RGB) in a given color image and then calculates the contribution of each of the applied **H&E** stains based on stain-specific RGB absorption.

.

So, the **H&E** contained in the image is adjusted by decomposing the RGB color of the tissue into **H&E** color space, followed by multiplying the magnitude of **H&E** of each pixel by two uniform variables with the range between 0.7 and 13 (Ruifrok et al, 2001). After the normalization, Hematoxylin (**H**) stained image and eosin (**E)** stained image were also obtained. But we are more interested the **normalized** image to carryout out analysis. Fig.24 below diagram below shows the normalized image

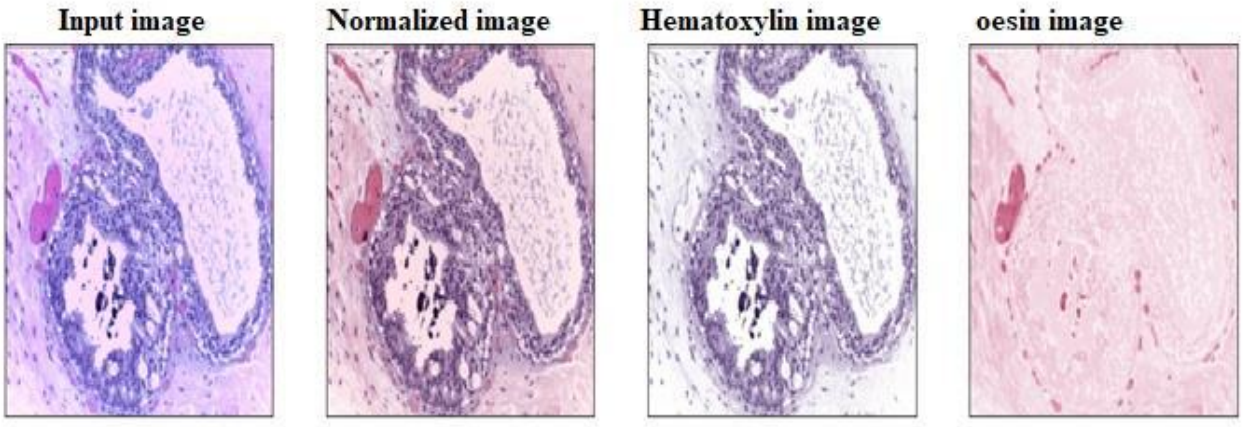


Fig.2: Normalized sample of H&E image

4.2 Data Splitting

After the image normalization process, there is the need for data splitting. As noted earlier, the image has high resolution of 2048 by 1538 pixels. Such amount of pixels cannot be fed into any neural network because of the computational resources and cost. Even if such graphic processing unit (GPU) exists, the cost will be very high. For every data science modeling, effective use of computational resources is a primary goal. On this note, the large size image was split into patches of smaller size. To determine the number of patches each image will be split into without overlapping, we adopted the normal binary

.

notation of: ***2, 4, 16, 32, 64,128, 256 and 512 etc.*** We also consider that the input size of all deep neural networks which varies from ***224 by 224 to 256 by 256.***

Given that the Height of the image: (***H***) = 2048 and the Width (***W***) = 1536.

If we adopt 512 binary notation, then, H will be 2048/512 = 4 and W will be 1536/512 = 3.

Such that, the number of patches that each image will be split into is H*W= 4*3 = 12. The resized height and width is 512 by 512 non-overlapping pixels

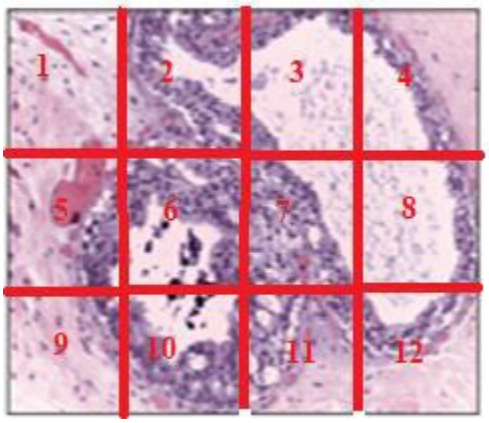
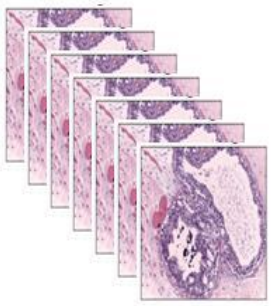


Fig3. Showing normalize whole slide image and generated patches



According to Adrian R, (2017), data augmentation is the process by which a new dataset is generated from existing dataset and that new dataset is then used in the model training instead of the old one. Data augmentation comprises wide range of techniques used to generate new training sample by applying random ***jitters and perturbation***. The process also ensures that the class label of the data was not changed. If a dataset was augmented, it will allow the network to continuously see new data with slight changes from the original which will make it learn more robust features. So, in each patched image, we performed 10 random data augmentation with the downscale of 512 X 512 pixels. As noted in the split patches, each image represents 12 crop patches. The patches are then encoded into 12 descriptors.

In order to preserve the spatial dimension of the image, each split image is saved into a numpy array folder before carrying data augmentation. All these processes are done programmically.

4.3 Data Augmentation and Pre-processing

Based on the formula below as deduced by Boureau et al, (2010), the set of 12 descriptors is combined through a 3 normal pooling to form a single descriptor. They suggested a hyperparameter of p=3.

Eqn 1.

From the above formula, **N** is the number of patch per image, **d*i*** is the descriptor of a patch and **dp** is a pooled descriptor of the image. So, the p-norm of a vector gives the average for p = 1 and the max for p---infinity. This means that for every whole slide image, we obtained 10 (number of augmentation) X 1 (patch size of 5) X 3 (CNN encoder) = 30 descriptors

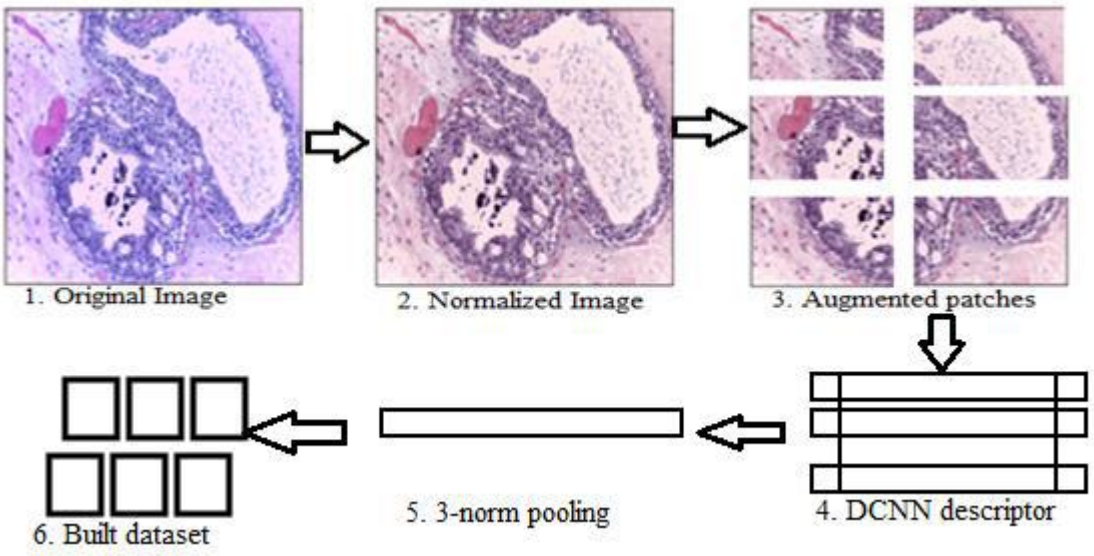


Fig 4 Data preprocessing pipeline

4.4 Feature extraction

This is the process in which the number of features in a dataset is reduced by creating new features from the existing ones and then does away with the old ones. The new features must have the quality of the original set of features (i.e., it is in linear combination of the existing features).

Deep convolution neural network with millions of parameter have achieved the state-of the-art results in image recognition and classification. Amongst these architectures is

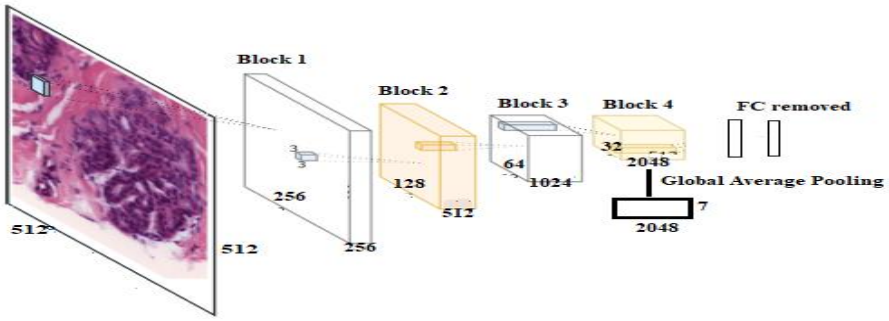


Fig. 5: Schematic overview of the Resnet 50 network architecture for deep feature extraction

Figure **3** below. The output volume of the feature map via the global average pooling will be

So, instead of allowing the network to propagate through the entire network, we stop it at the global average pooling layer, extract the values from the specified layer and then treat the values as a feature vector. The extracted feature vectors are then used for the training using lightgbm.

data was split into ten (10) folds so as to maintain the class distribution. Augmentations increased the size of the dataset by 30(1 patch sizes x 3 encoders x 10 affine augmentations). During this process, we made such that the descriptors of each image remained correlated. To make sure that there is no information leakage, the entire descriptors of each image in the dataset was saved in the same fold. For each combination of the encoder, the crop size and the scale we trained 10 gradient boosting

7x7x2048 making it a one-dimensional feature vector of 100,352-dim per image for a 32 bit float

models with 10-fold cross-validation. , this leads to the increase in the diversity of the models with little data bagging. Furthermore, we recycled each dataset 5 times with different random seed. This means that a total 10 (number of folds) ×5 (seeds) ×4 (scale and crop) ×3 (CNN encoders) = 600 was trained in the gradient boosting models.

During the cross-validation stage, all folds whose models not trained on this fold was predicted. We also extract 30 descriptors for each image and use them with all models trained for particular patch size and encoder. The predictions were then averaged over all augmentations and models. In the end, the predicted class was determined by the maximum probability score.

5. Results

The validation was done using the 10-folds stratified cross-validation. The accuracy was 93% and area under the ROC curve was 0.99. Figure 4 below show the results

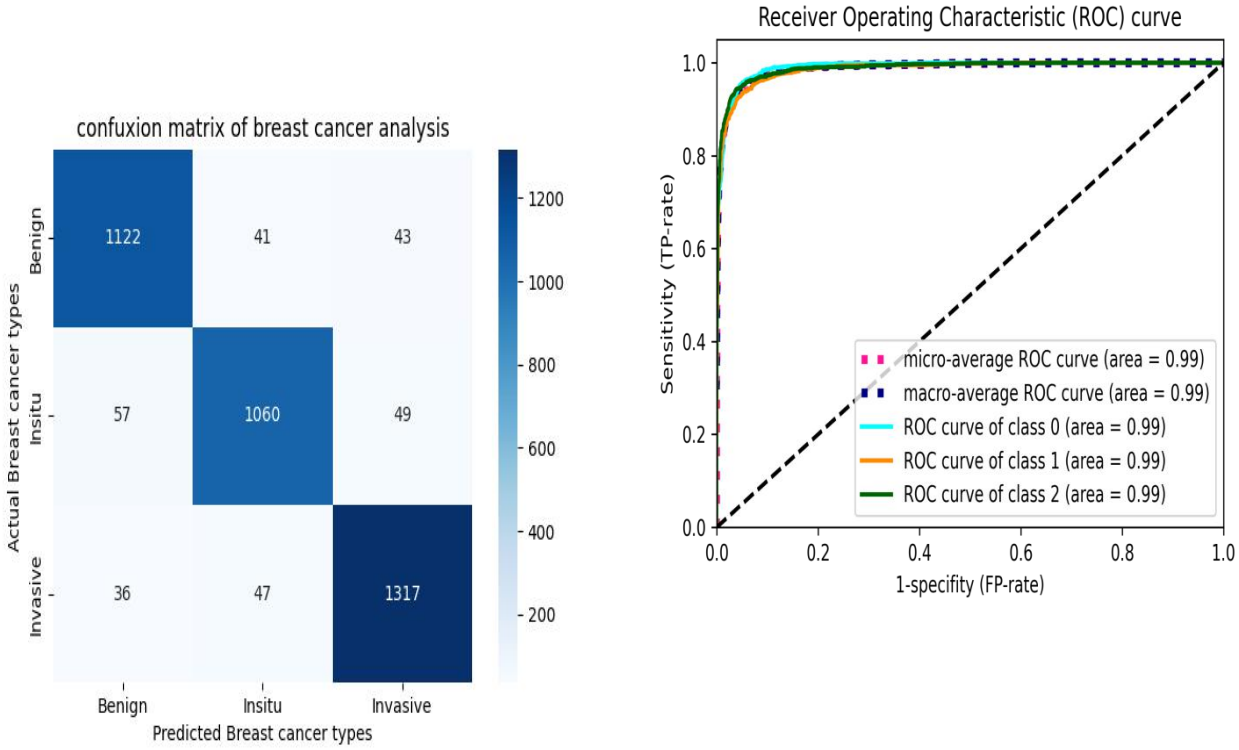


Figure 6: showing Confuxion matrix and RUC-ROC curve

5.1 Result Comparisons

Our primary aim in the research work was to train a breast cancer image classification model that will have improved accuracy and sensitivity score as compare to the work of Rui et al (2019). Table 3 below gives a quick summary of the comparisons

Table 3: Comparison of Sensitivity and Accuracy

Methods	**Image-wise Accuracy (%)**	**Sensitivity of all classes (%)**
Rui et al, 2019	91.3	89.2
Our Model	**93**	**99**

6. Conclusion & Recommendation

In this paper, we proposed an effective method for the classification of H&E stained histological breast cancer images. We made used strong data augmentation and deep convolutional features extracted using ResNet50 pertained model on ImageNet and applied light gradient boosting algorithm for the training. This enabled us to achieved high accuracy and sensitivity. To the best our knowledge, the reported results are superior to the work done by on. our predecessor that used same dataset. Also, as new framework, deep learning architecture and libraries is evolving on daily bases, the best of breast cancer histopathology image classification is yet to be achieved. More high quality image dataset will prove a robust model at that can be globally accepted. So, future works can utilize more dataset and new architecture like the vision transformer to see if the accuracy and sensitivity of model can be improved

---

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional

neural networks, in Advances in neural information processing systems, pp. 10971105, 2012.
2. Abdalla F, Boder J, Markus R, Hashmi H, Buhmeida A, Collan Y. Correlation of nuclear morphometry of breast cancer in histological sections with clinic pathological features and prognosis. Anticancer Res. 2009 May; 29(5):1771-6. PMID: 19443402.
3. Adrian Rosebrock. Practical Python and openCV + Case Studies. PyImageSearch.com, 2016, Url: https://www.pyimagesearch.com/practical-python-opencv/(cited on pages19, 38).
4. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., et al. (2017). Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 12:e0177544. doi: 10.1371/journal.pone.0177544.
5. Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 83, 10641069. doi: 10.1016/j.procs.2016.04.224.
6. Aswathy, M., and Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: present status and future possibilities. Inform. Med.Unlocked 8, 7479. doi: 10.1016/j.imu.2016.11.001.
7. Bayramoglu, N., Kannala, J., and Heikkilä, J. (eds) (2016). Deep learning for magnification independent breast cancer histopathology image classification, in 23rd International Conference on Pattern Recognition (ICPR), 2016. (Cancun: IEEE).
8. Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 111-118(2010).
9. DeSantis C, Ma J, Bryan L, et al., 2013. Breast cancer statistics, 2013. CA Cancer J Clin, 64(1), page:52-62.
10. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017).Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115118. doi: 10.1038/nature21056.
11. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A.,et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316, 24022410. doi: 10.1001/jama.2016.17216.
12. Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., and Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. 7:4172. doi: 10.1038/s41598-017-04075-z.
13. He, K., Zhang, X., Ren, S., and Sun, J. (eds) (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in proceedings of the IEEE international conference on computer vision(Santiago)
14. He, K., Zhang, X., Ren, S., and Sun, J. (eds) (2016).Deep residual learning for image recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas, NV).
15. Jiang Y, Chen L, Zhang H, Xiao X (2019)Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. PLoS ONE 14(3):e0214587. https://doi.org/10.1371/journal.pone.0214587.
16. K. Jackowski, B. Krawczyk, and M. Woźniak, "Cost-sensitive splitting and selection method for medical decision support system," in Intelligent Data Engineering and Automated Learning—IDEAL 2012, ser. Lecture Notes in Computer Science, H. Yin, J. A. Costa, and G. Barreto, Eds. Berlin: Springer, 2012, vol. 7435, pp. 85085.. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images, Comput. Biol. Med., vol. 43, no. 10, pp. 15631572,2013.
17. Motlagh, N. H., Jannesary, M., Aboulkheyr, H., Khosravi, P., Elemento, O.,Totonchi, M., et al. (2018). Breast Cancer Histopathological Image Classification: a Deep Learning Approach. bioRxiv 242818. doi: 10.1101/242818.

18. National Cancer Registry, ⟨Http://85.128.14.124/krn⟩, 2012 (accessed on 19.10.2019).
19. Nawaz, M., Sewissy, A. A., and Soliman, T. H. A. (2018). Multi-class breast cancer classification using deep learning convolutional neural network. Int. J. Adv. Comput. Sci. Appl. 9, 316322. doi: 10.14569/IJACSA.2018.090645.
20. P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies, IEEE Trans. Med. Imag., vol. 32, no. 12, pp. 21692178, Dec. 2013.
21. P. Wang, X. Hu, Q. Liu, X. Zhu, and Y. Li, Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Process., vol. 122, pp. 113, May 2016.
22. R. Yan, F. Ren, Z. Wang, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, F. Zhang, Breast cancer histopathological image classification using a hybrid deep neural network, Methods (2019), doi: https://doi.org/10.1016/j.ymeth.2019.06.014.
23. Ruifrok, A.C., Johnston, D.A., et al.: Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23(4), 291-299(2001).
24. Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016a). A dataset for breast cancer histopathological image classification. IEEE Transac. Biomed.Eng. 63, 14551462,doi: 10.1109/TBME.2015.2496264.
25. Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (eds) (2016b). Breast cancer histopathological image classification using convolutional neural networks, in 2016 International Joint Conference on Neural Networks (IJCNN). (Vancouver, BC: IEEE).
26. Veta M, Pluim JP, van Diest PJ, Viergever MA. Breast cancer histopathology image analysis: A review. Biomedical Engineering, IEEE Transactions on. 2014 May;61(5):1400-11.
27. Xie J, Liu R, Luttrell IV J and Zhang C(2019) Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Front. Genet. 10:80.doi: 10.3389/fgene.2019.00080.
28. Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury, Remote computer-aided breast cancer detection and diagnosis system based on cytological images, IEEE Syst. J., vol. 8, no. 3, pp. 949964, Sep. 2014.