Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 13, Issue 1, January 2022: 2207-2213

Breast Cancer Detection from Histopathological Images using Deep Learning Algorithms

Prof O.V.P.R Sivakumar[a], Kurapati Kavya Reddy[b], Ponaganti Ashwini[c]

[a]Professor, Department of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology, Hyderabad

[b]Student, Department of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology

[c]Student, Department of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology

Abstract

In India and over the world, Cancer has become a deadly disease that is adversely affecting women from different age groups. According to a survey, one in every 30 women suffers from this deadly disease in their lifetime. This project was first thought of because of the increase in cases of breast cancer and if we can detect it at an early stage then there is an increased chance of it getting cured. This paper lays a foundation in making the detection of cancer automated from histopathological images using deep learning algorithms so that more and more people can get it diagnosed at an early stage and subsequently get cured. To enhance the capability of histopathology image classification, a powerful deep learning algorithm and a sizable, varied dataset are required. In this paper, we have proposed an automatic and precise histopathological image analytical method, to detect breast cancer at an early stage. Deep learning techniques have recently made significant strides and produced outstanding results in the fields of computer vision and image processing, which has encouraged many researchers to use this method for the categorization of histopathology images.

Keywords: breast cancer, lesion classification, machine learning, convolutional neural networks, deep learning, benign, in situ carcinoma, invasive carcinoma, malignant.

1. Introduction

Cancer is a critical public health problem worldwide. When lung cancer is excluded, breast cancer incidence rates are the second highest for women. In addition, compared to other cancers, breast cancer has a particularly high fatality rate. The analysis of histopathological images continues to be the most popular technique for diagnosing breast cancer despite the rapid advancements in medicine. Of all the histopathological image analysis tasks, the most important is the classification task. Because the foundation and bottleneck of other in-depth research, such nuclei localization, mitosis detection, and gland segmentation, is the automatic and precise classification of high- resolution histopathology pictures. Currently, pathologists' manual qualitative analysis is the basic foundation for histopathological imaging in clinical practice. However, this analysis method has at least three drawbacks. The first issue is that there is a lack of pathologists worldwide, particularly in less

developed nations and smaller hospitals. It is vital to find a solution to the imbalanced distribution of resources and resource scarcity. Second, the pathologist's extensive professional expertise and years of acquired diagnostic experience determine whether the histological diagnosis is entirely accurate or not. Due to the pathologist's subjectivity, there are ever more inconsistent diagnoses. Third, pathologists are more likely to become weary and unfocused due to the complexity of the histopathological pictures.

2. Methodology

The most widely used kind of deep learning network, convolutional neural networks (CNNs), excel at both picture categorization and image feature extraction. These results have laid the foundation for the application of CNN in histopathological image classification.

In this paper, we present a strategy that preserves both the short-term and long-term spatial correlations between patches while extracting richer multilevel features and fusing CNN and RNN benefits. The high-resolution pathology images were first divided into small areas. Then, CNN is used to extract the richer multilevel image features of each patch. The patch features are finally fused by the RNN to create the final image categorization. Our average accuracy for the 4-class classification task was 91.3%, outperforming the most recent technique.

It is especially important to note that the classification sensitivity of benign photos dramatically increased from 66.7% to 84.3% as a result of our dataset including as many diverse subsets spanning different age groups as feasible to guarantee appropriate data diversity. This rise suggests that in order to enhance the capability of histopathology image classification, both a high-performance deep learning system and a sufficiently vast and diverse dataset are required.

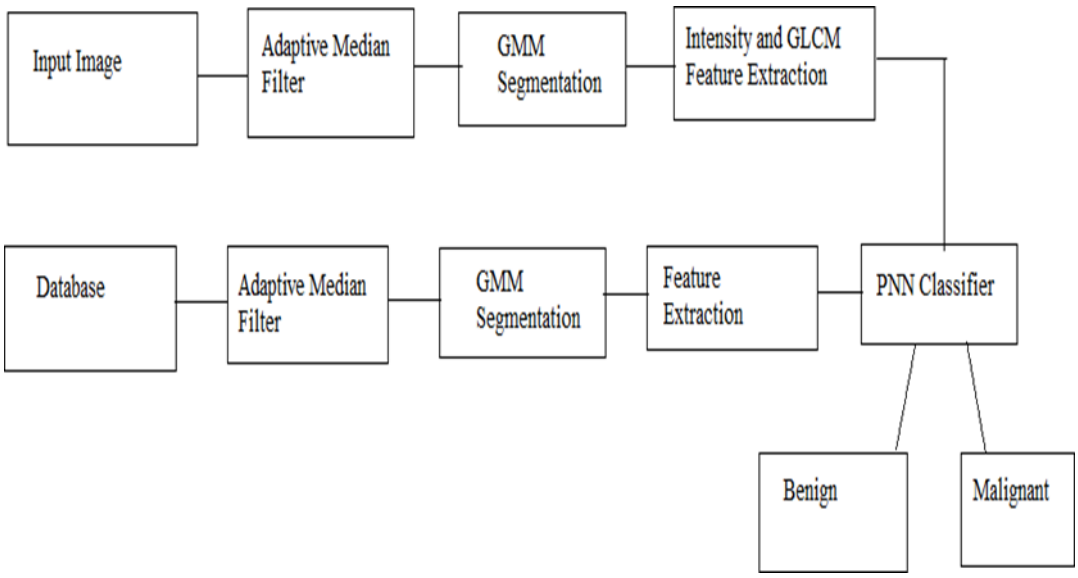

Figure.1 Block diagram for detection of breast cancer

2.1 Adaptive Median Filter

We have used an adaptive mean filter to remove noise from the image since it is better among all the spatial filters and distinguishes fine details from noise. To identify which pixels in a picture have been impacted by impulsive noise, the filter performs spatial processing. In order to determine whether a

pixel is noise or not, the Adaptive Median Filter compares it to each of its neighbours in the surrounding area.

Both the neighbourhood's size and the bar for comparison are movable. Impulse noise is defined as a pixel that differs from the majority of its neighbours and is not physically aligned with the pixels to which it is similar.

The median pixel value of the nearby pixels that passed the noise labelling test is then used to replace these noise pixels. We are initially converting the image into a grayscale image using the rgb2gray() function then applying adaptive mean filtering to the resulting image and then converting the image into unsigned integer 8 using the unit8() function.

2.2 GMM Segmentation

Gaussian Mixture Model (GMM) is a widely used approach for background subtraction and moving object detection. In this way, we pre-process the image. Then we perform GMM segmentation (Gaussian Mixture Model) on the pre-processed image with the number of regions 2 and number of GMM components 2 and a maximum number of iterations 10. We performed k-means segmentation with k=2. The Expectation-Maximization Algorithm of the HMRF-EM (Hidden Markov Random Field Model) was then put into practice.

2.3 GLCM Extraction

The statistical distribution of the observed combinations of intensities at specific points relative to one another in the image is used to compute the texture characteristics in statistical texture analysis. Statistics are divided into first-order, second-order, and higher-order statistics based on the number of intensity points (pixels) in each combination. Second-order statistical texture features can be extracted using the Gray Level Co-occurrence Matrix (GLCM) technique. The approach has been used in several applications, Third and higher order textures consider the relationships among three or more pixels. Although theoretically feasible, these are rarely used because of the lengthy calculating process and the complexity of the interpretation. A GLCM is a matrix with the same number of columns and rows as there are grayscale levels.

3. Problems Faced in Existing Models

The method that was previously utilized for histopathological image classification breaks an entire image into smaller patches, then uses a CNN to classify each patch. The classification results are then combined using a method like a majority vote. The feature representation vector of each patch is likewise extracted using a CNN, and the classification outcome of the entire histopathology picture is then produced using a conventional machine learning classification technique, such as a support vector machine (SVM).

This traditional method faces three challenges. First, rather than being fully utilized to increase the classification accuracy, the high-resolution features of histopathological pictures have had significant detrimental consequences. The primary cause is that the best patch-based approach in use right now does not sufficiently integrate these patches to get a classification result for the entire histopathological image.

Specifically, these methods integrate only the short-distance dependency between patches but ignore the long- distance spatial dependency, which is very helpful for the contextual understanding of the whole image. Second, the pathological picture patch's feature representation is not sufficiently richer. As a result, prior to image-wise fusion, a significant quantity of information is lost, rendering the fusion insufficient. Finally, there are a lot of dataset-related issues.

4. Dataset and Pre-Processing

One main characteristic of the deep learning method is that it can learn from large amounts of training data. Breakthrough results in the computer vision field were obtained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) based on the ImageNet dataset. In contrast, the domain of medical images has few publicly accessible large-scale image databases. Furthermore, the majority of these datasets lack labels. Data labelling by a limited number of medical specialists is quite expensive. Moreover, traditional methods of annotating natural images, such as crowdsourcing, cannot be transplanted to the medical image domain because these tasks are very complex and often require long-term professional training and extensive domain knowledge. Because other large datasets are typically not made available to the public, the majority of the early research on breast cancer pathological image analysis is conducted on a small dataset.

Related to breast cancer pathological image classification, one of the largest open datasets Providing an automatic and accurate categorization for each input breast cancer pathology image was the aim of this task. Our image dataset consists of 3771 high-resolution (2048 × 1536 pixels) and annotated hematoxylin and fluorescein (H&E) stained breast pathological images. Hematoxylin highlights nuclei by staining DNA and fluorescein highlights different structures by staining proteins. All images have equivalent acquisition conditions:100x or 200x magnification. The preparation procedure for pathological sections employed in this work was the standard paraffin method, which is widely employed in the pathological routine. According to the cancer type in every image, each image is labeled as normal, benign, in situ carcinoma, or invasive carcinoma.

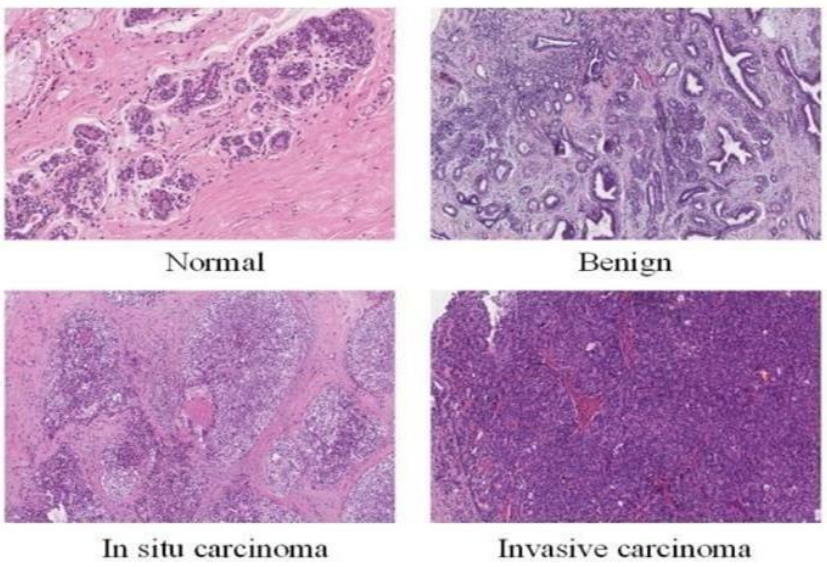

Figure.2 Examples of breast cancer pathological images in our dataset.

In the training stage, the pathological images are pre-processed and enhanced to improve quantitative

analysis. After pre-processing, we first fine-tune the pre-trained Inception-V3 model. For each image, the trained patch- wise model is used to extract the feature representation vectors of 12 patches. These 12 vectors are then utilized as input to train the long short-term memory for images (LSTM).

One diseased image is divided into an average of 12 tiny patches during testing. The patch-wise picture characteristics are then extracted using an improved version of Inception-V3. Every patch is extracted to a feature vector with dimensions of 1 x 5376. That is, from a single problematic image, 12 feature vectors can be retrieved. Finally, the 12 feature vectors (12 × 1 × 5376) are input into a bidirectional LSTM to fuse the features of the 12 small patches to make the final complete image-wise classification. The short-term and long-term spatial correlations between patches can be kept since our method combines the benefits of CNN and RNN.

5. Results

We demonstrate the effectiveness of our suggested algorithm on our publicly available dataset in this part. All experiments in this paper are finished using the TensorFlow framework. We mostly utilise accuracy and sensitivity to assess how well our method works. Positive images are those that fit the category; negative images are those that don't.

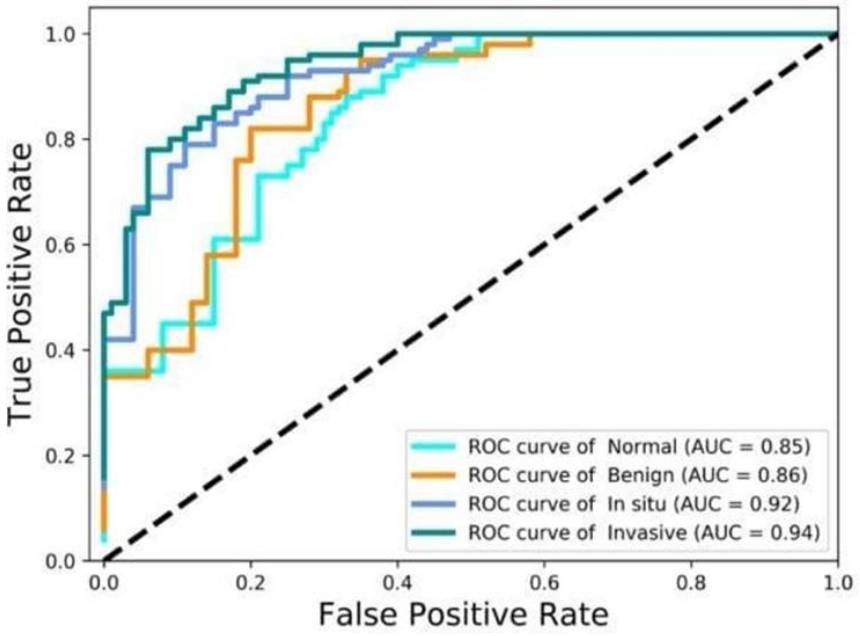

Fig. 3 Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values for normal tissue, benign carcinoma, in situ carcinoma, and invasive carcinoma.

The classification accuracy of normal and benign categories is only 86% and 87%. What is more significant is that 10% of normal categories are misclassified as benign categories and 10% of benign categories are misclassified as normal categories. The Receiver Operating Characteristic (ROC) and Area Under the Curve figures show the same behaviour (AUC). It shows the mean AUC value of 89.25%, corresponding to 85%, 86%, 92%, and 94% for the four classes based on receiver operating characteristic analysis.

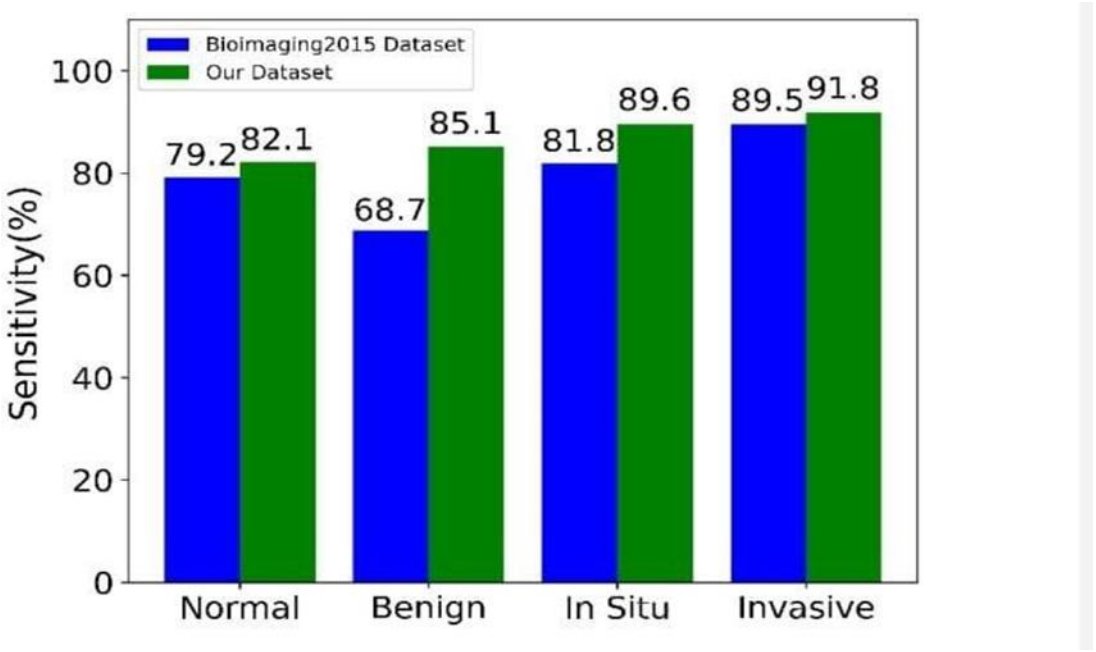

Figure.4 Comparison of sensitivity uses the same method across different datasets

The chart shows how each class of sensitivity improves when a larger dataset is used, but the classification sensitivity of benign photos improves the most, going from 68.7% to 85.1%. Many previous works have described the problem that the classification sensitivity of benign images was relatively low because the characteristics of benign images are not salient, they can be subdivided into many subcategories. Additionally, as they become older, their qualities become more varied.

6. Conclusion

Using a hybrid convolutional and recurrent deep neural network, we developed a new technique for classifying breast cancer pathology images in this research. Based on the richer feature representation of the pathological image patches, our method considered the short-term and the long-term spatial correlations between patches through an RNN, which is right behind a richer multilevel CNN feature extractor. As a result, both the short- and long-term spatial correlations between patches were taken into account. It has been demonstrated through comprehensive testing and comparisons that our novel technology outperforms the most recent technique. For future work, to improve the accuracy of classification, outstanding deep learning algorithms and large enough as well as diverse datasets are indispensable. In terms of algorithms, the use of attention mechanisms in deep learning algorithms is a direction that can be tried, because it has achieved outstanding performance in natural image processing. Because it has achieved remarkable performance in natural image processing, the employment of attention mechanisms in deep learning algorithms is a direction that can be tried in terms of algorithms. Hardware advancements are crucial as well, of course. At the same time, we are trying to extend this approach to whole slide images which will be more difficult but will produce greater value in clinical practice.

References (APA)

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer [Internet] 2020.

[2] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, CA: A Cancer J. Clinic. 68 (2018) 7–30.

[3] M. Veta, J.P. Pluim, P.J. Van Diest, M.A. Viergever, Breast cancer histopathology image analysis: a review, IEEE Trans. Biomed. Eng. 61 (2014) 1400–1411.

[4] G. Litjens, T. Kooi, B.E. Bejnordi, S. Aaa, F. Ciompi, M. Ghafoorian, V.D.L. Jawm, G.B. Van, C.I. SÃnchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[5] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221– 248.

[6] A. Holzinger, B. Malle, P. Kieseberg, P.M. Roth, H. Müller, R. Reihs, K. Zatloukal, Towards the augmented pathologist: challenges of explainable-ai in digital pathology, arXiv preprint arXiv (2017).

[7] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng. 63 (2016) 1455–1462.

[8] https://www.bing.com/ck/a?!&&p=33fb1e87793a719fJmltdHM9MTY2NDE1MDQwMCZpZ3VpZD0yODlkMzBlNy0yMzUzLTZmNTQtMmEyMi0yMmViMjI1MzZlYTImaW5zaWQ9NTE3Nw&ptn=3&hsh=3&fclid=289d30e723536f542a2222eb22536ea2&u=a1aHR0cHM6Ly93d3cuaWJtYy51cC5wdC9ldmVudHMvdXBjb21pbmctZXZlbnRzL2Jpb2ltYWdpbmcyMDE1LXN5bXBvc2l1bS1hcHBsaWVkLWJpb2ltYWdpbmc&ntb=1

[9] C. Loukas, S. Kostopoulos, A. Tanoglidi, D. Glotsos, C. Sfikas, D. Cavouras, Breast cancer characterization based on image classification of tissue sections visualized under low magnification, Comput. Math. Methods Med. 2013 (2013) 1–7.

[10] C. Bilgin, C. Demir, C. Nagi, B. Yener, Cell-graph mining for breast tissue modeling and classification, in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007, pp. 5311–5314.

[11] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, J. Tomaszewski, Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features, in IEEE International Symposium on Biomedical Imaging: From Nano To Macro, 2008, pp. 496–499.

[12] P. Filipczuk, T. Stevens, A. Krzyzak, R. Monczak, Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies, IEEE Trans. Med. Imaging 32 (2013) 2169–2178.

[13] Y. Zhang, B. Zhang, F. Coenen, W. Lu, Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles, Mach. Vis. Appl. 24 (2013) 1405–1420.

[14] Y. Zhang, B. Zhang, F. Coenen, J. Xiao, W. Lu, One-class kernel subspace ensemble for medical image classification, EURASIP J Adv Sign Process 07 (2014) 1–13.

[15] A.D. Belsare, M.M. Mushrif, M.A. Pangarkar, N. Meshram, Classification of breast cancer histopathology images using texture feature analysis, in TENCON 2015–2015 IEEE Region 10 Conference, 2016, pp. 1–5.