Detection of Breast Cancer from Histopathology image and Classifying Benign and Malignant State Using Machine learning

Dr. Amol Potgantwar[1],Kajal Patil[2], Sunidhi Jain[3], Poonam Jadhav[4], Komal Bhadane[5]

[1] *Professor, Department of Computer Engineering, SITRC, Nashik*
[2] *Student, Department of Computer Engineering, SITRC, Nashik*
[3] *Student, Department of Computer Engineering, SITRC, Nashik*
[4] *Student, Department of Computer Engineering, SITRC, Nashik*
[5] *Student, Department of Computer Engineering, SITRC, Nashik*

ABSTRACT

In today's world, cancer is a major public health concern. Breast cancer is a type of cancer that begins in the breast and spreads to the rest of the body. Breast cancer is one of the leading causes of death in women. Cancer occurs when cells become uncontrollably large. There are several types of breast cancer. The model proposed addressed both benign and malignant breast cancer. Breast cancer identification and classification using histopathology and ultrasound images are critical steps in computer-aided diagnosis systems. Researchers have demonstrated the ability to automate the initial level identification and classification of tumors over the last few decades. Breast cancer can be detected early, allowing patients to receive the appropriate treatment and improve their chances of survival. Deep learning (DL) and machine learning (ML) techniques are used to solve many medical problems. Several previous scientific studies on the categorization and identification of cancer tumors using various types of models have been published in the literature, but they have some limitations. The lack of a dataset, on the other hand, makes research difficult. Using the deep learning technique, the proposed methodology was created to aid in the automatic detection and diagnosis of breast cancer.

Keywords— Mammography, deep learning, convolutional neural network, augmentation.

I. INTRODUCTION

Bosom malignant growth is one of the inescapable types of disease on the planet in ladies with more than one and a half million anticipated findings in 2010 and the reason for death for the greater part more than half a million each year [1]. In Qatar, breast cancer is by far the most widespread cancer type, accounting for 31% of all cancer cases in women [2]. It is shown that the danger of developing breast cancer in women is 56 in every 100,000 [2]. Early identification of bosom malignancy is the most proficient methodology in saving lives as it raises the opportunity of endurance through a powerful treatment prompting a decrease in death rates. Mammography is considered the most common imaging technique used for breast cancer screening and the detection of abnormality in breast tissue.

Currently, radiologists need to examine the entire mammogram of a case, and doctors require a test for biopsy to determine whether a tumor is benign or malignant. Radiologists can determine if the depicted mammogram has cancer or not, but the error rate is between 8% to 16% [3]. Although the current clinical methods to detect breast cancer have dramatically improved in the last few decades [4] [5], there are still many limitations, such as variability among the opinion among radiologists and, additionally, the procedures are time-consuming and invasive.

To overcome such limitations of cancer diagnosis and treatment plan, deep learning (DL), a branch of machine learning, based techniques are gaining attention in the scientific community as well as in clinical setup.

Recently, several deep learning-based techniques applying CNN have started to achieve remarkable performance in the medical area such as chest pathology classification [6], thoracoabdominal lymph node detection, and lung disease identification [7]. For mammography, [8] and [9] studied breast masses tumor detection using a recurrent neural network and random forests. Several works tackled the difficulty of breast mass tumors classification, for example, by selecting a multi-stage approach using textual features and extracted hand-engineered semantic [10]. Wan et al. classified a scanned mammogram by extracting the features from each view of the mammogram and consolidating them to output a prediction [11]. Arevalo et al. performed massive preprocessing steps based on domain knowledge before applying into the CNN training phase [12].

DL based method is also advantageous over regular Computer-Aided Diagnosis (CAD) systems in terms of detection capability. Traditional CAD approaches used a combination of fine-tuned hierarchy with parameters that had to be empirically tested. The research of such systems usually required decades of studies and implementation [13]. DL based systems have already proven the ability to outperform several traditional methods. For particular problem specifications, DL methods led to an exceptional improvement compared to modern CAD systems [14] evaluated on large datasets, e.g. in[6]

II. LITERATURE SURVEY

Authors of [16] presents the method to detect cancer region and classify normal and cancerous patient. Pre-processing operation perform on the input Mammogram image and undesirable part removed from the image, tumor region segmented from the image using morphological operation and highlighted the region on original mammogram image or if mammogram image is normal case then it shows that patient is normal.

Authors of [13] proposed in this paper it is possible to detect the breast cancer at a very early microcalcification stage itself and the result of this proposed methodology will be of very high accuracy leading to true positive and true negative results. The methodology proposed in this paper provides end to end solution. However, authors of [14] feel that digital image recognition of plant diseases is one of the thrust areas and hence came out with a model which comprises of back propagation networks and probabilistic neural networks. It is further depending on color features, shape features and text features extracted from disease image.

Also, the work of authors [15] discusses an approach for automatic detection of abnormalities in the mammograms. Image processing techniques have been applied to accurately segment the suspicious region-of-interest (ROI) prior to abnormality detection. Unsharp masking has been applied for enhancement of the mammogram. Noise removal has been done by using median filtering. Discrete wavelet transform has been applied on filtered image to get the accurate result prior to segmentation. Suspicious ROI has been segmented using the fuzzy-C-means with thresholding technique.

III. CLASSIFICATION METHODS

In this section, we begin our discussion by describing the basic building blocks of CNN, and then we go on and discuss two approaches to improve their performances in case of a limited dataset, namely, data augmentation and transfer learning.

a. Convolutional Neural Networks

Exceptional progress has been performed in image recognition, essentially due to the availability of annotated datasets and deep convolutional neural networks (DCNNs). CNNs empower learning data-driven and hierarchical image features from sufficient training data. Due to the availability of large, annotated datasets, CNNs have become a dominant approach to solve different image classification tasks over the past decade, leading to a state of art performances in this area. CNNs are typically made up of concatenation of several layers. Each one of them is composed of a few subunits, including a bank of learned filters, an element-wise nonlinearity, and a pooling operator to reduce the dimensionality, Fig. 1 demonstrates a convolutional filter used for edge detection, and Fig. 2 shows a sample CNN architecture. CNN models are trained to learn a mapping from a set of training inputs to their corresponding set of outputs via an optimization process to minimize a loss function such as Cross-entropy or mean-squared error.

b. Data Augmentation

Typically, CNNs have millions of parameters that require a proportional number of training data. However, in many situations, it is not possible to get real dataset samples and thus, data augmentation is applied to generate more samples from the existing dataset [16]. Data augmentation is achieved by applying different transformations to the input images without altering the perceived object classes. These transformations can be, for example, rotation, flipping, or subsampling with different crops and scales. Additionally, noise can be added to the input images as a form of data augmentation [16].

c. Transfer Learning

Transfer learning aims to transfer the knowledge of a CNN model trained on a large dataset to another dataset. One approach that has been shown to improve the performance of a CNN is to use a pre-trained CNN model (e.g., GoogLeNet, ResNet, AlexNet, etc.) on a different dataset and use it for the weight initialization for the classification problem in hand. Using transfer learning can be difficult, especially with medical datasets that tend to be of small size and are unbalanced. Many learning strategies to fine- freeze, freeze weights play an essential part in the CNN accuracy and performance.

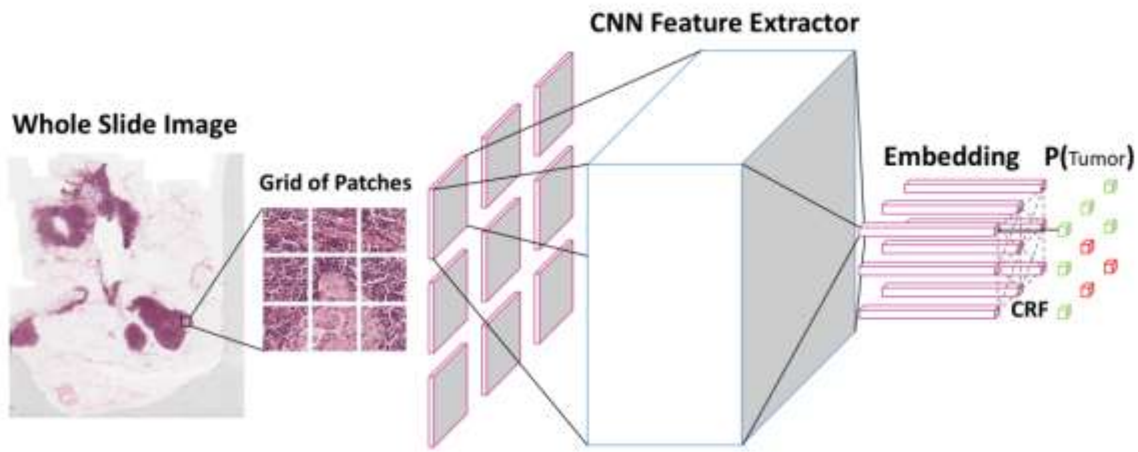

Figure 1 System Architecture

The proposed ResNet50 model achieved an accuracy of 85.71% with a precision rate of 85.7% and a recall rate of 87.3%. A specificity rate of 84% and an F1-score rate of 86.5% were also achieved. Fig. 10 shows the confusion matrix for the proposed model, while Fig. 11 presents the AUROC curve for the altered model. Finally, Fig. 6 and Fig. 7 illustrates the training and loss curves, respectively. The proposed Resnet50 model performed slightly better than InceptionV3 with an increase in the accuracy rate by 6.6% and in the precision rate by 10.3%. However, the recall rate for InceptionV3 was slightly better than the Resnet50 model, with a difference of 1.8%. Based on the outcome obtained from the experiments, we can conclude that using a few numbers of epochs, ten in our case, can produce better results than training the network for larger epochs as the model tends to overfit the data because ten epochs are not a big number. By consequence, the training time for each model was sufficient.

IV. CONCLUSION

This work presented two different end-to-end deep convolutional learning models to classify pre-segmented breast tumor masses. The results of an evaluation of the modified InceptionV3 and ResNet50 models on the classification task were discussed. Additionally, the study illustrated how specific pre-processing, data augmentation, and transfer learning techniques can overcome the dataset size bottleneck, which is popular in the medical computer vision tasks. The two proposed models are very promising to use in real-life clinical practice to support the medical expert's decision. Our results showed that high accuracy levels could be achieved with simple modifications applied to different pre-trained CNN models.

In this study, we did not apply significant changes to InceptionV2 and ResNet50 architectures, only adding two fully connected layers and replacing the final output layer to accommodate the introduced classification problem. In the future, we would like to use another approach reported in literature survey. In this approach, the authors proposed to train many CNN and obtain the average of their prediction results, which would improve the classification performance in the testing phase. Additionally, we could use two models, such as InceptionV3 and ResNet50 or different snapshots from the same model, such as ResNet50 in the ensemble. Another approach is the majority voting, although it requires training an odd number of CNN models to finally decide on the tumor class based on the classification outcome from each model. Also, since cross validation was proven to produce promising results, we can consider cross-validation with various k-folds and observe the resulting models' performance. In the future, we will use our model against available datasets based on the Qatari population. We believe our model will improve the quality of breast cancer treatment plans for the Qatari population.

V. References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," CA. Cancer J. Clin., vol. 69, no. 1, pp. 7–34, 2016.

[2] Hamad Medical Corporation, "Breast Cancer Awareness Campaign," 2018. [Online]. Available: https://www.hamad.qa/EN/your health/bcac/Pages/default.aspx.

[3] E. G. Joann et al., "Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy," Breast Dis., vol. 21, no. 4, pp. 330–332, 2009.

[4] C. Ryan, K. Krawiec, U.-M. O'Reilly, J. Fitzgerald, and D. Medernach, "Building a Stage 1 Computer Aided Detector for Breast Cancer using Genetic Programming Conor," Search Methodol. Introd. Tutorials Optim. Decis. Support Tech. Second Ed., no. May, pp. 143–186, 2014.

[5] J. Dheeba, N. Albert Singh, and S. Tamil Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," J. Biomed. Inform., vol. 49, pp. 45–52, 2014.

[6] T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," Med. Image Anal., vol. 35, no. August, pp. 303–312, 2017.

[7] H. C. Shin et al., "Deep Convolutional Neural Networks for ComputerAided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[8] N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9349, pp. 605–612, 2015.

[9] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests," 2015 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA 2015, pp. 1–8, 2015.

[10] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning," Sci. Rep., vol. 6, no. February, pp. 1–9, 2016.

[11] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9351, pp. 652–660, 2015.

[12] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," Comput. Methods Programs Biomed., vol. 127, no. October 2018, pp. 248–257, 2016.

[13] J. E. Ball and L. M. Bruce, "Digital mammogram spiculated mass detection and spicule segmentation using level sets," Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc., no. April, pp. 4979–4984, 2007.

[14] J. Tang, R. M. Rangayyan, J. Xu, I. E. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 2, pp. 236–251, 2009.

[15] Y. LeCun and others, "LeNet-5, convolutional neural networks," URL http//yann. lecun. com/exdb/lenet, vol. 20, p. 5, 2015.

[16] Poonam Kathale; Snehal Thorat , "Breast Cancer Detection and Classification", 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)