CLASSIFICATION AND BREAST CANCER SUBTYPES DETECTION USING HISTOPATHOLOGY IMAGES

Cancer is one of the leading causes of human death worldwide currently. For women, breast cancer-related deaths are higher compared to the other types of cancer-related deaths and this type of cancer causes thousands of deaths each year worldwide. It has been reported that the incidence rate of breast cancer ranges from 19.3 per 100,000 women in East Africa, to 89.7 per 100,000 women in Western Europe. The number of new cases has continued to grow in recent years, and this number is expected to increase to 27 million in 2030.

Breast cancer develops from breast tissue identified by lump in the breast and there are some changes in normal conditions. Clinical screening includes mammography, breast ultrasound, biopsy and other method. A biopsy is the only diagnostic procedure that can definitely determine if the suspicious area is cancerous. The pathologists diagnose by visual inspection of histological slides under the microscope, which is considered as confirmatory gold standard for diagnosis. However, the traditional manual diagnosis needs intense workload by experts with expertise. Diagnostic errors are prone to happen with the pathologists that have not enough diagnostic experience. It is shown that the use of Computer-aided diagnosis (CAD) to automatically classify histopathological images can not only improve the diagnostic efficiency, but also provide doctors with more objective and accurate diagnosis results.

Deep Learning is a growing technology in the field of machine learning and it has got the attention of many researchers. The Convolutional Neural Network (CNN) has achieved great success in a large-scale image and video recognition.

RELATAED WORKS

Spanhol et al used AlexNet to classify breast cancer pathology images for both benign and malignant categories. Their classification results are 6% higher than traditional machine learning classification algorithms. As another method, author mentions that previously trained CNN reuse is used as a feature vector, and DeCAF features are extracted. Then, the DeCAF feature is used as an input to the classifier trained for the new classification task. It achieved an average of 84% accuracy on breast cancer case images.

Kausik et al. proposed a multiple instance learning (MIL) framework for CNN. They introduced a new pooling layer that helped to aggregate most informative features from patches constituting a whole slide, without necessitating inter-patch overlap or global slide coverage. An accuracy of about 88% was obtained on breast cancer case images. In Another research, the author proposed a structured deep learning model for solving the subordinates of breast cancer, with the best classification result reaching 92.19%. In [16], the authors proposed that hybrid CNN unit could make full use of the local and global features of an image, so as to make a more accurate prediction. The author also introduces the bagging strategies and hierarchy voting tactic to help improve the performance of the classifier. Finally, 87.5% classification accuracy was obtained on the multiple classifications of breast cancer. Akba et al. [17] propose a novel regularization technique for CNNs, and named it as the transition module, which captures filters at multiple scales, and then collapses them via global average pooling to ease network size reduction from convolutional layers to FC layers. The transition module was able to adapt to a small data-set successfully by achieving accuracy rates of 91.9%.

Wei et al. proposed that the class and subclass labels of breast cancer should be used as apriori knowledge to suppress the feature distance of different breast cancer pathological images. At the same time, a data augmentation method was proposed, and the accuracy of the binary classifications was reached 97%. In another research method, the author introduces two methods. The first method is based on the extraction of a set of handcrafted features encoded by two coding models (bag of words and locality constrained linear coding), and then support vector machines were trained for classification. The second method is based on the design of convolutional neural networks. The experiment result shows that the convolutional neural network is superior to the classifier based on manual features. The accuracy of the two classifications is 96.15% and 98.33% respectively, and the accuracy of multi-classification is 83.31% and 88.23% respectively.

RESEARCH GAP

Although successful detection of malignant tumors from histopathological images largely depends on the long-term experience of radiologists, experts sometimes disagree with their decisions. Image diagnosis using Computer-aided diagnosis provides a reliable for experts' decision-making. Hence automatic and precision classification for breast cancer histopathological image is of great importance in clinical application for identifying whether it is benign and malignant tumors and the subtypes of the tumors from histopathological images.

Challenges for using convolutional neural networks for automatic classification of pathological breast cancer images

(1) Over-fitting of the model due to the continuous deepening of the model, the number of parameters of CNN also increases rapidly.
Solution to minimize overfitting - A large number of breast cancer histopathological images are usually required as training data for training CNN. However, the cost of obtaining a large number of labeled breast cancer images is expensive. Therefore, in case of limited breast cancer image data, we need to reduce the model over-fitting risk from the perspective of reducing CNN parameters and using data augmentation methods.

(2) It is well known that various hyperparameters have a great influence on the performance of the CNN model, especially the learning rate. In the process of model training, it is often necessary to adjust the learning rate parameters to obtain better performance manually, which makes it difficult to apply the algorithm in real life applications by non-expert users.

Hence author proposes a better framework to classify the histopathological images accurately without overfitting.

DATASET DESCRIPTION

The Breast Cancer Histopathological Image Classification (BreakHis) is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format

Characteristics

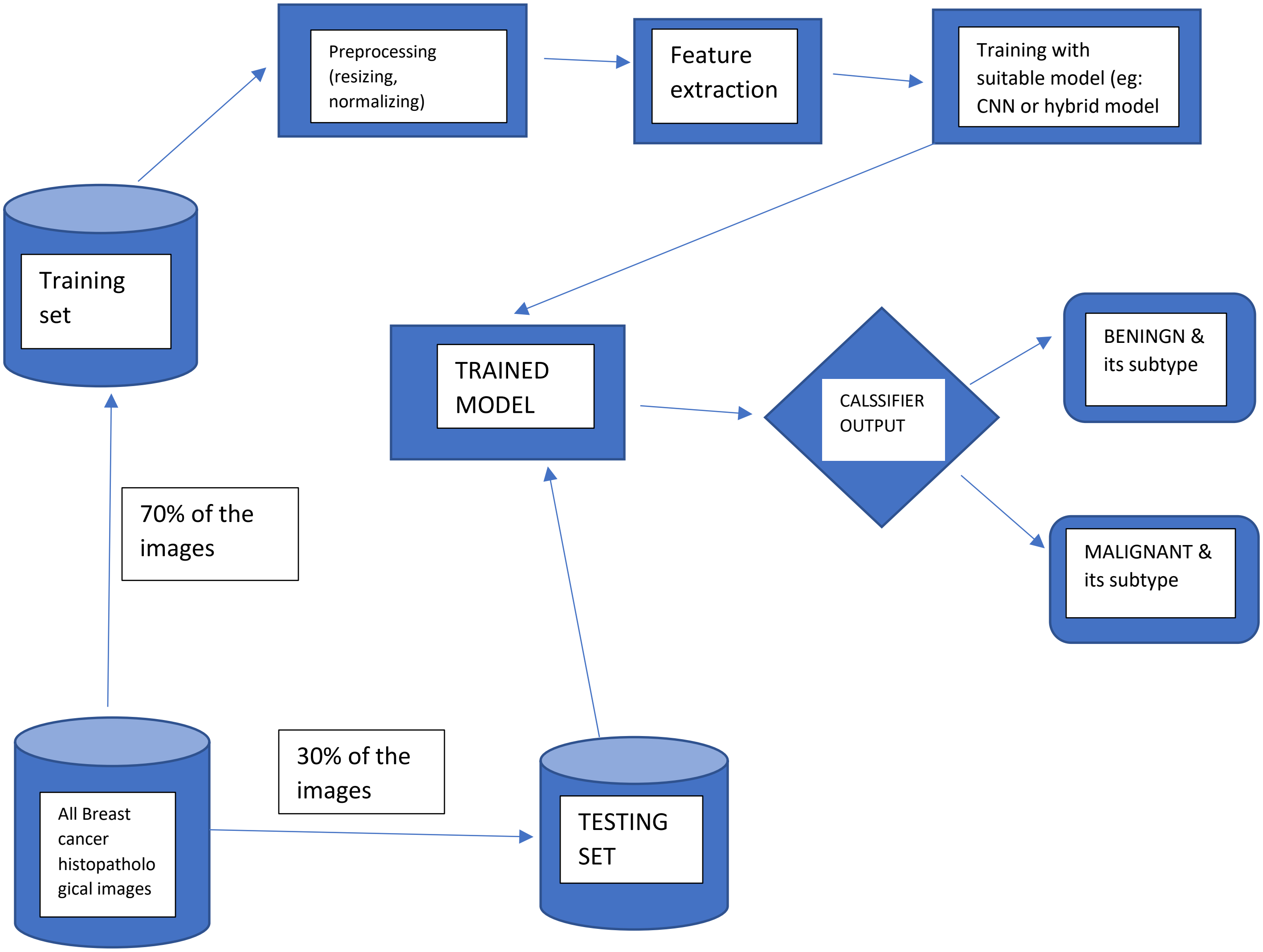
The dataset BreaKHis is divided into two main groups: benign tumors and malignant tumors. Histologically benign is a term referring to a lesion that does not match any criteria of malignancy – e.g., marked cellular atypia, mitosis, disruption of basement membranes, metastasize, etc. Normally, benign tumors are relatively "innocents", presents slow growing and remains localized. Malignant tumor is a synonym for cancer: lesion can invade and destroy adjacent structures (locally invasive) and spread to distant sites (metastasize) to cause death.

In current version, samples present in dataset were collected by SOB method, also named partial mastectomy or excisional biopsy. This type of procedure, compared to any methods of needle biopsy, removes the larger size of tissue sample and is done in a hospital with general anesthetic.

Both breast tumors benign and malignant can be sorted into different types based on the way the tumoral cells look under the microscope. Various types/subtypes of breast tumors can have different prognoses and treatment implications. The dataset currently contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC).

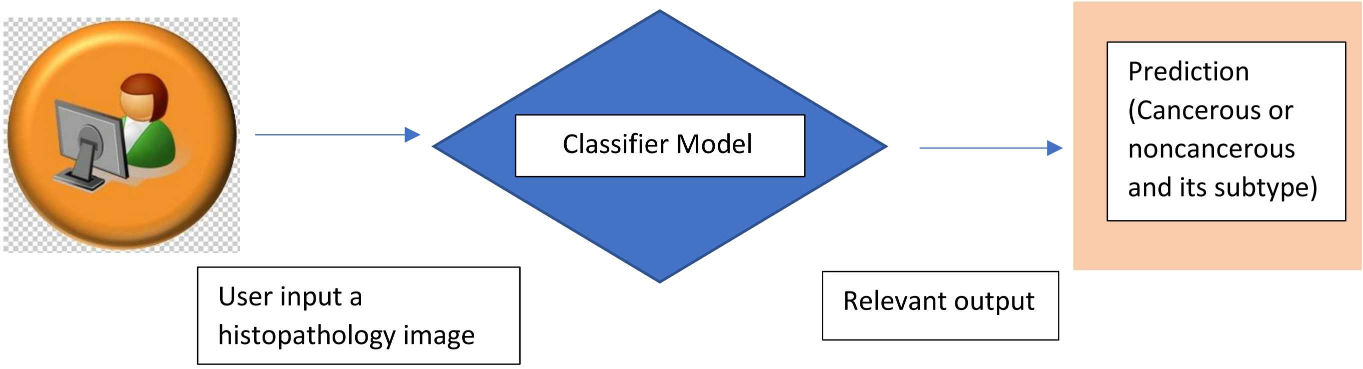
Link to download: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

PROPOSED MODEL



*If feature extractions possible, the model will be trained with the specific features extracted.

END PRODUCT


Classifier Model
Prediction (Cancerous or noncancerous and its subtype)
User input a histopathology image
Relevant output

REQUIREMENTS

1. Utilizing the image preprocessing techniques
2. Feature extractions in order to train the model
3. Choosing the best fit model (dep learning model) in order to predict the cancerous or non-cancerous histopathology image with higher accuracy
4. Utilizing the best fit algorithm in order to predict the tumor subtypes with higher accuracy
5. Test set results
6. Validation (Evaluation techniques to see the prediction results)
7. A software-based application deploying the trained model which enables user to input the histopathology image (breast cancer) and gives the output as benign or malignant and its subtypes accurately