Comparison of machine learning algorithms for early breast cancer detection using histopathological images

Karina Jonina
Supervisor: Dr. Oisín Cawley
Department of Computing, Institute of Technology Carlow
c00278440@itcarlow.ie

INTRODUCTION

According to National Breast Cancer Research Institute, 1 in 9 Irish women will develop breast cancer, however, only 1 in 11 will be diagnosed with breast cancer before 75 years of age (National Breast Cancer Research Institute, no date). Early detection of breast cancer can save many lives.

Routine detection methods for breast cancer include Ultrasound, Mammograms, Computed Tomography (CP), Magnetic Resonance Imaging (MRI).

Histopathology biopsy imaging is considered to be the "golden standard". The biopsy procedure consists of a specialised needle device guided by X-Ray to extract a sample of tissues from the tumorous area. This tissue is placed on a slide, stained by haematoxylin and eosin (H&E) and scanned by electron microscopy to obtain digital pathological images.

Each pathological slide contains millions of cells, which must be viewed by pathologist. These images are complex and diverse, requiring pathologist's prior knowledge and time during work-intensive diagnosis. An improvement in development of process could improve the efficiency of pathologist, as well as overall the accuracy of detection results (Gurcan et al., 2009; Zhong, Piao and Zhang, 2021)

DATASET

BreaKHis dataset consists of 7,909 biopsy images of benign and malignant breast tumours, which was acquired on 82 patients during a clinical study from January 2014 to December 2014 in Brazil and organized into four magnification factors (40x, 100x, 200x, 400x).

Images have been categorized as benign (healthy) breast tumours or malignant (cancerous) breast tumours by a pathologist. Figure 1 below shows slides of breast malignant tumour (stained with HE) seen in different magnification factors and highlighted rectangle (manually added for illustrated purposes only) is the area of interest selected by pathologist to be detailed in the next magnification factor (Spanhol *et al.*, 2016).

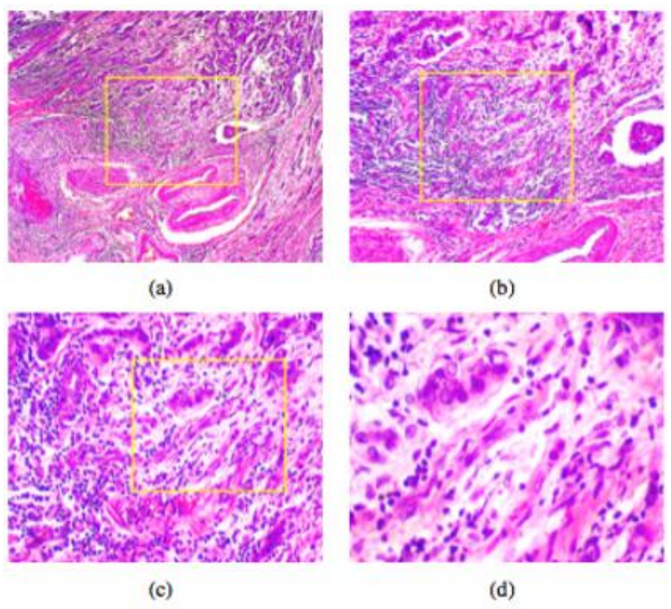

Figure 1. Slides of breast malignant tumour (stained with HE) seen in different magnification factors: (a)40x, (b) 100x, (c) 200x, and (d) 400x (Spanhol *et al.*, 2016).

Magnification	40x	100x	200x	400x	Total	Patients
Malignant	1370	1437	1390	1232	5429	24
Benign	625	644	623	588	24680	58
Total	1995	2081	2013	1820	7909	82

Table 1: Distribution of Magnification Factor and Histological Types (Spanhol *et al.*, 2016)

TOOLS/ TECHNOLOGY

LITERATURE REVIEW

Several researchers have conducted deep learning and neural network analysis on BreaKHis dataset and found that the models built predict benign or malignant best with magnification at 200x.

Different neural network algorithms and deep learning algorithms have been used to predict malignant tumours with various results. Majority of previous researchers focused on predicting target class (benign or malignant) with all 4 magnification factors. There is evidence to suggest that several proposed models achieved the highest the accuracy with magnification at 200x than other magnification factor (Chattopadhyay *et al.,* 2022). However, these studies always discuss the small sample in each magnification dataset as a limitation of a study.

ETHICS

Ethical implementation of incorrect prediction are very high. If a prediction results in False Positive, where the model incorrectly predicts a malignant (cancer) instead of benign (healthy), the patient will feel distraught until further tests are run, and model's prediction is disproved.

However, a False Negative prediction, where the model incorrectly predicts benign (healthy) instead of malignant (cancer), may discourage further medical examination and intervention, and could result in patient patient's death. Hence, there will be a strong emphasis on reducing False Negatives.

RESEARCH OBJECTIVES

- If the number of images in each magnification sub-dataset is increased through image augmentation, does this improve the accuracy of detection results?
- By focusing machine learning algorithms specifically targeting the images of magnification, does this lead to improved accuracy of detection results (than other machine learning algorithms that create general machine learning algorithms?
- Does one machine learning algorithms out of chosen machine learning algorithms lead to better accuracy of detection results?

METHODOLOGY

Data Segmentation
Although BreaKHis dataset provided sub-types of histological types, focus will be strictly 2 categories: malignant (cancer) and benign (healthy), multiple classes will be dropped.

Data Augmentation
With small dataset for each magnification, there is a fear that dataset will cause models to overfit due to its small size, so decided to process and augment each magnification dataset to increase in size by 9 folds (Mamun *et al.,* 2021):
- Resize to 224 x 224 pixels
- Rotate by 10 and 20 degrees
- Flip both vertically and horizontally

- **Data Modelling and Evaluation**
- Split the dataset into training set and test set
- Develop neural network machine learning models to compare
- Evaluate models with accuracy of detection results and focus on reducing False Negative

NEXT STEP

The following steps need to be implemented:
- Images to be processed and augmented as described.
- Specifics of machine learning algorithms need to be decided upon.
- Machine learning algorithms are to be built and used to analyse images
- Fine-tuned machine learning algorithm
- Re-evaluate the accuracy of detection results as appropriate.

REFERENCES

Chattopadhyay, S., Dey, A., Singh, P.K. and Sarkar, R. (2022) 'DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images', *Computers in Biology and Medicine*, 145. doi:10.1016/j.compbiomed.2022.105437.

Gurcan, M. n., Boucheron, L. e., Can, A., Madabhushi, A., Rajpoot, N. m. and Yener, B. (2009) 'Histopathological Image Analysis: A Review', IEEE Reviews in Biomedical Engineering, Biomedical Engineering, *IEEE Reviews in, IEEE Rev. Biomed. Eng.*, 2, pp. 147–171. doi:10.1109/RBME.2009.2034865.

Mamun, R.A., Rafin, G.A., Alam, A. and Sefat, Md.A.I. (2021) 'Application of Deep Convolution Neural Network in Breast Cancer Prediction using Digital Mammograms', *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, pp. 1–7. doi:10.1109/IISEC54230.2021.9672368.

Spanhol, F. A., Oliveira, L. S., Petitjean, C. and Heutte, L. (2016) 'A Dataset for Breast Cancer Histopathological Image Classification', *IEEE Transactions on Biomedical Engineering, Biomedical Engineering,*, 63(7), pp. 1455–1462. doi:10.1109/TBME.2015.2496264.

Zhong, Y., Piao, Y. and Zhang, G. (2021) 'Hybrid Attention Mechanism Guided Convolutional Neural Network for Breast Cancer Histology Images Classification', 2021 *IEEE 9th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 1–7. doi:10.1109/ICCSNT53786.2021.9615395.