Review of Breast Cancer Detection Method

Jiahao Huang [1, †, *], Qianfeng Qiao [2, †] and Yu Zhou [2, †]

[1] Admiral Farragut Academy-Tianjin, Tianjin, China

[2] Maple Leaf International School-Chongqing, Chongqing, China

* Corresponding author email: jh4611@columbia.edu

†These authors contributed equally to this work and should be considered co-first authors

Abstract. The fierceness of breast cancer is obvious for the patients and the field of medical studies for its malignant destruction in the human body. The failure in diagnosis may put patients in suffering from high financial pressure and may cause the patients' physical health to deteriorate. The key in this research of contrasting different kinds of algorithms and varied types of classifications is to support physicians to make a better judgment to prevent misdiagnosis. The research introduced many aspects of data on breast cancer detection to make the data inputs of algorithms able to satisfy doctors' needs. Moreover, the varied kinds of features will specify and confirm the presence of breast cancer tumours instead of treating them as a part of the benign tissues. With the information above, the research found a difference in accuracy between the classification and tried to make a comparison to conclude that most satisfy and support the actual use of these data to physicians, who may see the symptoms with misunderstanding. Thus, Random Forest was qualified to be evaluated as the most applicable algorithm of its excellence in processing higher dimensional, which is dynamic, data.

Keywords: Breast Cancer Detection; Machine Learning; CNN.

1. Introduction

Breast Cancer, common cancer, among all types of them, showed its high probability to the public. Certain kind of cancer is similar to the therapy of other cancers. However, cancer affects a lot of people and most proportion of them suffered from pain and other symptoms from this type of illness. Moreover, people have to take over large amounts of the bill if wanted to receive a chance of treatment that cannot ensure the recovery of the patient. Thus, a series of research focusing on breast cancer data will be necessary.

Currently, cancer can be weakened or destroyed in several ways: including Surgeries, Chemicals, Radiation and other ways of treatments. According to Rosemary [1], surgeons make surgeries for patients and cut off pathological tissues that were polluted by the cancer cells as the tissue has the presence of a lump, an appearance of a tumour. This will prevent the cancer cells to expand their domination and to spread over the body. Radiation, similarly, kills cancer cells through rays. Chemical therapy, however, is different from the other two ways. It gives the patient a specific chemical and these chemicals can eligible in exterminating the cancer cells. With these methods, people can recover from these diseases even if they were not eliminated, and this leaves a potential danger to the patients.

Although the improvement of modern biochemical science was not enough to take over this particular kind of disease, the statistics that were collected by hospitals and physicians are still useful for researchers today. People can use different kinds of methods to make a judgment on the risk or the probability that cancer is present in people's bodies. These ways varied from different topic, some of them are based on patients' reactions and physical appearances, while others may be analyzed through data. The final result, which gave the physician a credible source of information, will eventually indicate the patient's probable illness.

Triple Assessment is a method of detecting breast cancer that consists of three diagnostic components: "medical history and clinical examination, mammography and/or ultrasound, fine needle aspiration cytology and/or core biopsy" [2]. The advantage of Triple Assessment is its ability

to detect breast cancer quickly and easily at an early stage, which is non-invasive and has a high accuracy rate, it has a sensitivity of 99% [2].

Research statements describe the various data features that can be used to predict breast cancer and their relationship with breast cancer. We mainly focus on the comparison and evaluation of different types of data processes, including pictures and algorithms. The research will set up a contrast in many types of aspects such as the accuracy, and the level of skills that people are required to reach to understand the pieces of information. The advantages and disadvantages will be labelled and analyzed in the research to bring out facts about the algorithms, and this may help the public to decide which way to judge patients' potential risk of getting breast cancer is better for the patients themselves.

2. Data Type

2.1 Image-based Data

In recent years the use of chest computed tomography has received increasing attention in the diagnosis of breast cancer. CT has been used for the detection and diagnosis of primary cancers and the detection of localised areas of the body. For cancer detection patients prefer MIRI to CT because of the health risks associated with ionising radiation, but real-time CT has fewer contraindications, less testing time and less cost. In 2008, breast CT was introduced into the initial clinical trial of 10 healthy volunteers and 69 female patients presenting with category 4 and 5 lesions. In the preliminary trial, there was no significant difference in breast CT compared to radiography in terms of the effect between benign and malignant lesions and breast density on the visualisation of lesions [3], but breast CT images had a better comfort level.Fig.1 CT images showing multiple malignant tumour sites in the breast (a) and axilla (b) and clearly showing the various degrees of metastatic spread of cancer [4]. CT images provide a complete picture of the spread of breast cancer, which can be useful in the treatment and diagnosis of cancer.

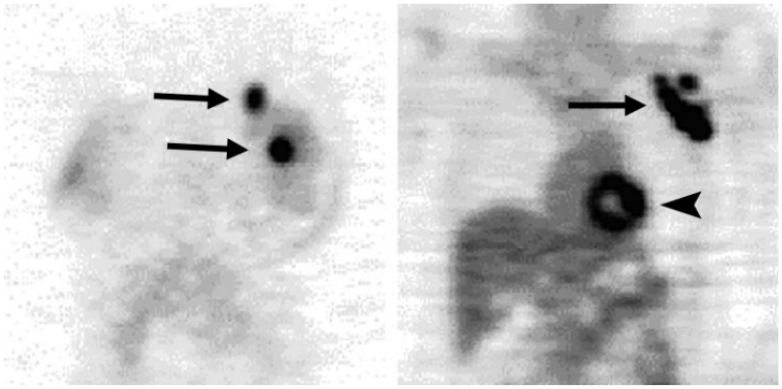

Fig 1. Configuration of breast cancer tumours -- lumps

2.2 Numeric Data

2.2.1 Properties on Tumors

A tumour is a new growth of various tumour-causing substances in the body, and breast cancer as a malignant neoplastic disease is much more prevalent in the female population than in men. Benign tumours are slow growing and do not metastasise easily and usually do not cause death. Malignant tumours, on the other hand, grow rapidly, metastasise easily and grow aggressively. Among the metastatic sites of breast cancer, the axillary lymph node is the most common site of metastasis and metastasis is the main cause of cancer-related death in breast cancer patients [5]. The size of the tumour and the extent of lump involvement are the two most important indicators of breast cancer. However, mortality is significantly increased with increasing lymph node involvement, regardless of tumour

size [6]. Therefore, the status of the axillary lymph node can be used as an indicator of the extent of cancer spread.

2.2.2 Evaluation Toward Stages of Cancer by Recognizing Physical Status of Patients

Cancer is a disease that was created by a certain number of dysfunctional cells gathering in certain body tissues. In Breast Cancer, cancer cells would duplicate and might spread over the organs around the body, eventually killing the host. This duration consisted of four stages, I, II, III and IV. In the case of its mechanism, we decided to research the stages of living in patients to see the influence of breast cancer. According to the form of A Ahmad1, S Barrington2, M Maisey2 and RD Rubens1 [7], usually, patients who suffered from breast cancer are in stage II, III, and IV and experiences significant pain from the disease with sensory loss. They are not able to make self-treatment and are urged to the demand of accompany. These breast cancer patients developed plexopathy—a status of losing neural reflexes such as losing feelings of touch.

A common factor shows that patients who were in the later stage of cancer began to have a stronger intensity of pain and mental depression. Moreover, it may cause a common phenomenon, to be skeletal, making the immune system of the physical body reluctant to react and make a defence. Interestingly, axillary cysts are especially common among patients who have current stages II and III, which might implicate the shifts of cancer cells.

3. Machine Learning Algorithm

3.1 Image Classification Algorithms

Through the graphs of breast cancer, the computers required a way to capture features from the pictures quickly since analyzing pictures with their colours and features individually is a heavy task. Thus, an algorithm is invented and can play the role of catching features in the pictures. Convolutional Neural Network (CNN), as described as a way of conducting information from those images, supported the computer on this particular function. According to O'Shea, Keiron, and Ryan Nash [8], CNN can express colours in a systematic way that was consisted of vectors, which are the output of complex calculations. To start the process of CNN, inputs enter the convolutional layer where the images began to be expressed by an array of numbers by pixels. After the convolution layer, the image began to begin its second process. It will enter the pooling layer that rules of pooling vary. Some of the pooling may maximize the values of datasets, while others may take a mean from the series of data.

Because CNN is a gradual process, (it)may repeat several times for its use of recognizing subtle vectors. The computer will eventually generate images with better resolutions.

However, the disadvantages for CNN are also obvious in some parts of the study. The algorithm will fail in having memories and in combining the data that does not match. Videos can also be a disadvantage for CNN, on the other hand, the audio files of the videos may be made CNN pressured on processing its sounds, which may weaken the quality of features.

Compared to the logistic regression method applied by Jabeen Sultana1*, Abdul Khader Jilani [9], CNN requires longer runtime. However, the runtimes' costs declared its accuracy in data processing, eventually making it strong in its robustness.

Deep learning models of CNN may contribute to predicting breast cancer risks. From the example of Hafiz Mughees Ahmad1, Sajid Ghuffar2, and Khurram Khurshid1 [10], the models of application in mammography of breast cancer evaluation based on hundreds of the samples were ResNet50, GoogleNet and AlexNet, which are originated from the CNN algorithm. The statistics showed that the final accuracy of the models, whatever kind of classification, varied slightly, which is an increment of 2-4%. The Alexnet had the lowest accuracy while Resnet had the highest, with support from Dina A. Ragab, Omneya Attallah, Maha Sharkas, Jinchang Ren, Stephen Marshall [11].

Table 1. A reference of experiment in accuracy from the form of Hafiz Mughees Ahmad, Sajid Ghuffar, Khurram Khurshid [10]

Methods		Accuracy	
		Patch-wise Classification	Image-wise Classification
Proposed Methods	AlexNet	79.84	82.3
	GoogleNet	81.07	83.6
	ResNet50	83.60	85.0

3.2 Numeric Classification Algorithms

When working with numerical data we often encounter situations where the dimensionality of the data is too high, which makes storage and analysis more difficult and slows down operations. Therefore, dimensionality reduction can be used to improve efficiency. There are two main types of dimensionality reduction: projection and manifold learning. PCA (principal component analysis) is the most commonly used method and is unsupervised learning. PCA uses projection for dimensionality reduction and can reduce the number of features in a dataset while maintaining the accuracy of the information. Jhajharia et al. proposed a neural network-based model for breast cancer prediction. They pre-processed the data using principal component analysis, extracted relevant features and then trained them using an artificial neural network to achieve 96% accuracy [12].

A decision tree is a classification model. The purpose of a decision tree is to allow the information entropy to decrease when the depth (the number of layers) of the decision tree increases. The classification efficiency of a decision tree depends on how fast the entropy decreases, which increases with the rate of entropy reduction The advantage of a decision tree is its high Interpretability, but the disadvantage is that it tends to overfit, so pruning is required. The common classification algorithm for decision trees is C4.5. Ruggieri improved it and proposed a new algorithm called EC4.5, which uses the same decision tree, but is five times more efficient than C4.5 [13].

Ensemble learning is the combination of multiple learners, which usually has better generalization and accuracy than a single learner. There are two modes of ensemble learning: i. The first is where individual learners are unrelated to each other and can learn independently at the same time, an example of this is bagging and random forests ii. The second is where individual learners are dependent on other learners and need to learn serially, an example of this is boosting [14].

Random forests are based on decision trees, each tree randomly selects a portion of samples and a portion of the random features in the samples therefore the training results are different. We don't know which samples are outliers or which features have a greater impact on the classification, so randomization can reduce the impact of both on the results. Compared to decision trees random forests are less prone to overfitting. Ma Mengwei et al. constructed four machine learning models, namely Random Forest, XGBoost, Logistic Regression and Support Vector Machine, from 11 clinical information such as age, age of menarche, fertility history… and 24 imaging features from X-ray and ultrasound of 200 breast cancers. It was concluded that Random Forest had the best overall performance, but the accuracy still needed to be improved [15].

Support vector machines are a class of supervised learning, a type of binary classification model that is now being used in cancer detection. The gradual maturity of modern techniques provided the application for machine learning tools in breast cancer diagnosis is increasing and the SVM is constantly being optimized and improved. Research has shown that experienced doctors can diagnose breast cancer with an accuracy of 79% while using machine learning techniques can achieve an accuracy of 91%. [16] To extract more efficient extraction of information and diagnosis, An algorithm called K-SVM has been developed, which is mixed, or can be called hybrid, an algorithm using K-means to incorporate SVM, [17] in which K-means is mainly used to identify benign and malignant

patterns of cancer and compute them, SVM is then responsible for detection in the WDBC breast cancer dataset. which contains 10 categories of 32 distinct features of different classes of cell nuclei. The K-SVM method was successful in improving the accuracy of the WDBC data set and achieved a deduction of 97.38%.

Table 2. Performance of four algorithms constructed from imaging and clinical information of 200 patients

Model	Sensitivity	Specificity	Accuracy
RF	60.00	67.40	68.60
XGBoost	55.20	59.80	61.50
LR	64.80	60.30	64.90
SVM	68.10	62.50	65.10

However, there are limitations between individual models, and to address this problem using twelve different support vector machines, A hybrid approach is used, which is an improved blend based on WAUCE.[18]This method has higher accuracy and lower variance compared to other SVMs on the same dataset (SEER).

4. Discussion

Different kinds of CNN algorithms also provided people to have a wider range of selection between them when they have specialized demands. The accuracies of different kinds of classification from the CNN models have slight differences. Although the accuracy of some CNN's algorithms showed its disadvantage according to the form from Hafiz Mughees Ahmad1, Sajid Ghuffar2, Khurram Khurshid1 [10], the deficiency from the accuracy will be replaced by other aspects. From the experimental conclusion of Adyasha Sahu *, Pradeep Kumar Das, Sukadev Meher [19] using the algorithm of Hybrid 5, although the lower accuracy affected its quality of data, it is more applicable than other algorithms.

Random forest complements decision trees, which are prone to over-fitting. The random process of the random forest makes it superior in handling high-dimensional data, generalization ability and resistance to interference [15]. Coupled with the independence between every tree, it can be trained simultaneously and has a low computational cost. It can therefore be preferred for most data.

The SVM algorithm is primarily useful in the detection of breast cancer classified as malignant or benign. With different improvements to the SVM algorithm, the SVM algorithm is highly accurate in dealing with small-scale data detection. In contrast to other algorithms, SVM can use different kernel functions for different decision functions. This gives it the flexibility to handle both low and high-latitude data and is insensitive to outliers in the dataset. It performs poorly for large samples and high latitude parts of the data; the high flexibility along with sensitivity to the choice of parameters and kernel functions is also a drawback of its algorithm. However, it outperforms other algorithms in terms of the generalizability of the algorithm.

5. Conclusion

The resnet algorithm, which belongs to the category of CNN, has a stronger capability in providing professional and accurate information while physicians use the data as a reference for breast cancer diagnosis. The memory part of the algorithm was a problem; however, this was a commonly distributed problem that was widely in the CNN algorithm--it was part of its feature.

The WAUCE-based hybrid algorithm within the SVM algorithm has a high level of accuracy and stability of the calculated results. The fact that different decisions are often used when diagnosing different data is a characteristic of the algorithm's handling of data and the advantage of its flexible data handling capabilities.

Ensemble learning tends to achieve higher accuracy, which is represented by random forests. For data with a large number of features, the Random Forest algorithm is recommended because it is good at handling high-dimensional data and does not require feature selection.

Finally, the researchers contend that the range of algorithm mentioned in this paper does not meet its expectation in amount for its deficiency in gathering information, which is a problem that cannot be solved in the current stage, and future studies may help in filling the blank.

To conclude, the models of classification in different algorithms showed their support toward the patients and the physicians by providing accurate data, especially after the specific kind of technique has been applied to the use of modern medical fields of study. The materials of this particular study remained potentially abundant, which were highly possible to be created by the theories from further research by scholars. However, the progress may not appear within a short period. Efforts and passion will eventually lead the researchers to reach a wider range of knowledge when they began their studies in this particular field.

References

[1] Rosemary A. Walker. The complexities of breast cancer desmoplasia[R]. Leicester: University of Leicester, Breast Cancer Research Unit, Clinical Sciences, Glenfield Hospital, 2001: 2.

[2] Pragnya Chigurupati. Triple Assessment for Breast Cancer[N]. Dr. Pragnya Chigurupati, 2021-3-25.

[3] Lindfors K K, Boone J M, Nelson T R, et al. Dedicated breast CT: initial clinical experience[J]. Radiology, 2008, 246(3): 725.

[4] Rosen E L, Eubank W B, Mankoff D A. FDG PET, PET/CT, and breast cancer imaging[J]. Radiographics, 2007, 27(suppl_1): S215-S229.

[5] Liu Z, Ni S, Yang C, et al. Axillary lymph node metastasis prediction by contrast-enhanced computed tomography images for breast cancer patients based on deep learning[J]. Computers in Biology and Medicine, 2021, 136: 104715.

[6] Carter C L, Allen C, Henson D E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases[J]. Cancer, 1989, 63(1): 181-187.

[7] A Ahmad, S Barrington, M Maisey, RD Rubens. Use of positron emission tomography in evaluation of brachial plexopathy in breast cancer patients[J]. British Journal of Cancer, 1999, 79(3/4): 478-479.

[8] O'Shea, Keiron, and Ryan Nash. An introduction to convolutional neural networks[R]. arXiv preprint arXiv:1511.08458,2015: 2-10.

[9] Jabeen Sultana1, Abdul Khader Jilani. Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers[J]. International Journal of Engineering & Technology, 2018, 7 (4.20) (2018), 22.

[10] Hafiz Mughees Ahmad1, Sajid Ghuffar, Khurram Khurshid. Classification of Breast Cancer Histology Images Using Transfer Learning[R]. Islamabad:1WISP Lab, Deptartment of Electrical Engineering, 2GREL Lab, Department of Space Science Institute of Space Technology, 2019:2-4.

[11] Dina A. Ragab, Omneya Attallah, Maha Sharkas, Jinchang Ren, Stephen Marshall. A Framework for Breast Cancer Classification using MultiDCNNs[R]. Glasgow and Alexandria: Electronics & Communications Engineering Department, Arab Academy for Science, Technology, and Maritime Transport (AASTMT), Alexandria 1029, Egypt, and the Electronic & Electrical Engineering Department, University of Strathclyde,2021:10-24.

[12] LIANG Shutong, GUO Maozu, ZHAO Lingling. Survey on medical decision support systems based on machine learning[J]. Computer Engineering and Applications, 2019, 55(19):1-11.

[13] Houssein E H, Emam M M, Ali A A, et al. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review[J]. Expert Systems with Applications, 2021, 167: 114161.

[14] MA Mengwei, QIN Genggeng, XU Weimin, LIU Renyi, WEN Chanjuan, ZENG Hui, CHEN Weiguo. Machine learning models for predicting molecular types of breast cancer based on X-ray and ultrasound breast imaging reporting and data system[J]. Medical Imaging Technology in China, 2020, 36(12): 1814-1819.

[15] Zhou Zhihua. Machine Learning[M]. Bejing: Tsinghua University Press, 2016: 1-425.

[16] Lamidi A. Breast Cancer Classification Using Support Vector Machine (SVM)[J]. 2018.

[17] Zheng B, Yoon S W, Lam S S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms[J]. Expert Systems with Applications, 2014, 41(4): 1476-1482.

[18] Wang H, Zheng B, Yoon S W, et al. A support vector machine-based ensemble algorithm for breast cancer diagnosis[J]. European Journal of Operational Research, 2018, 267(2): 687-699.

[19] Adyasha Sahu ∗, Pradeep Kumar Das, Sukadev Meher. High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets[J]. Elsevier, 2023, Biomedical Signal Processing and Control 80(2023) 104292: 6-13.