Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230203

Comparison of machine learning models on breast cancer risk prediction challenge

Mingjie Chen

School of Science, Xi' an Jiaotong - Liverpool University, Suzhou, Jiangsu, 215000, China

Mingjie.Chen2002@student.xjtlu.edu.cn

Abstract. Breast cancer is a malignant tumor that poses a serious risk to women's life and wellbeing. To make matters worse, this cancer is less symptomatic in its early stages. It is not easily diagnosed through traditional means. The topic of this essay is to investigate machine learning for the determination of breast cancer.. The methods based on machine learning are as followed: automated nuclear section segmentation model, BCRecommender System, DNNs, and computer-aided diagnosis model (CADM). The methods studied are all based on the BreCaHad dataset and use a comparative metric.To measure the performance of each model, the accuracy, F1-score, specificity, and precision are used. The result shows that the approaches based on machine learning work well in diagnosing breast carcinoma, with high accuracy. Most of them have a percentage over 90% in accuracy and some of them are even higher than 95%. However, some of the models work poorly, such as layer 1 of BCRecommender with 61.06% accuracy and EfficientNetB0 with 72.96%. With every aspect taken into consideration, computer-aided diagnosis (CADM: Using combined features of HOG, WPD, ResNet as well as PCA + SVM) has the greatest advantage in diagnosing BC.

Keywords: Breast cancer prediction, deep learning, machine learning.

1. Introduction

Breast cancer is a frequent malignancy experienced by young patients, especially among women originating in the breast tissue [1]. This kind of carcinoma is able to cause breast deformity, pain, and trauma, and may eventually cause the death of the patient [2]. Also, previous studies have shown that women with advanced breast cancer have a relatively lower chance of surviving than women with early-stage cancer [3]. A lot of patients have come back to normal life after prompt therapies. Therefore, early diagnosis and treatment are significant and can greatly improve patient survival and quality of life.

There exist some difficulties in diagnosing breast carcinoma. The initial symptoms are not obvious and may be ignored by patients, leading to delayed treatment. Additionally, both histological diagnosis and figuring out the pathological grade and stage are time-consuming and sophisticated tasks. The early method usually used for diagnosing BC is to use the following technique: mammography or tomosynthesis, ultrasound with various optical approaches [4]. However, this approach is not reliable enough under the screen settings, for the sensitivity is not so ideal that the results of BC samples may have a chance to be negative, leading to inaccuracy [5].

Compared with the methods mentioned, machine or artificial intelligence diagnostics take advantage. The machine learning algorithms enhance the accuracy of diagnosing BC significantly, and they can identify females at high risk of triple-negative breast carcinoma (TNBC) for early treatment [6]. Similarly, deep learning is also an effective tool in detecting cancer. The ensemble deep learning method, for example, researchers were able to demonstrate a sensitivity of 97.73% for carcinoma classification and achieved an overall accuracy of 95.29% for histopathological pictures of non-carcinoma and BC carcinoma [7]. In short, AI has higher accuracy and is a more reliable tool for diagnosing BC. The general technical path for machine or artificial intelligence diagnosis involves data collection, pre-processing, feature extraction, model selection, training, and validation, and finally, prediction or decision-making using algorithms such as deep learning, neural networks, or other machine learning techniques [8].

This article aims to compare several methods of machine learning in diagnosing breast cancer, including supervised learning-based cancer detection, automated nuclear section segmentation model, BCRecommender, DL models (including five neural networks), and Computer-aided diagnosis [9-13]. All of the methods used are based on BreCaHAD which is used for detecting breast cancer and histopathological diagnosis [14]. Apart from the accuracy, some other values such as specificity, precision, and F1-score are also taken into consideration. Interestingly, the algorithms and methods that account for machine learning are typical with high accuracy and reliability in diagnosis.

2. Method

In this part, four representative methods are introduced and compared.

2.1. Automated nuclear section segmentation model (ANSM)

The method is summarized in one sentence: automatic classification, measurement, and detection of Ki-stained nuclear sections for prognostic assessment of breast cancer based on machine learning of nuclear textures. It starts with a pre-processing phase, where the RGB data is converted into HIS values. Then, unsupervised learning and Otsu thresholding segmentation are used to compute the parametric parameters further. The method uses the cell proliferation scoring mechanism to determine whether the cells are benign or malignant. Ki-67 is used as a marker that can be used to calculate the cell proliferation score as well as to describe cell classification. The python programming language is used, and several open-source library packages such as matplotlib are supported [10].

2.2. BCRecommender system (BCRS)

This approach considers feature-based classification and clinical and histopathological sections to diagnose and understand breast cancer. This is a hybrid approach using the BCRecommender recommendation model. The model is structured in layers. Each is used differently and with different results. The system is valid for both early and advanced stages of cancer compared to conventional diagnosis. The methodology is divided into four steps: problem definition, data collection, system design, model selection, and results reporting. Extensive breast cancer data, including clinical trial results, pathology reports, and genetic and lifestyle information, were collected and pre-processed for functional classification using uniform column coding, true feature scaling, and normalization. After data selection, the most efficient classification model was determined [11].

2.3. DNNs

This approach consists of five advanced DL models. Firstly, histopathological data sets using machine learning (ML) and deep learning (DL) visual interpretation methods are compared. Next, the DL and XAI survey is partially reported and covers interpretability metrics. A new histopathology annotation dataset, a new DL method and a new interpretable algorithm for morphological and molecular analysis. The methodology consisted primarily of training on the BreakHis dataset using five DL models, then evaluating the BreCaHAD dataset and using Grad-CAM to assess the interpretability of each DL model. The number of annotations matched to the JSON file was then validated to assess the accuracy of the interpretable regions of the models [12].

2.4. Computer-aided diagnosis model (CADM)

The CADM system uses a novel DCNN (Deep Convolutional Neural Network). There are convolutional layers, a small SE-ResNet module, and fully connected layers for breast cancer classification. The system was compared with well-known pre-trained migration learning models, VGG-16, VGG-19 and ResNet-50. VGG-16 used a practical regression classifier and had the highest accuracy of 92.6%. The BreakHis dataset used in this study included 7909 histopathological images from 82 breast cancer patients. The four steps of the proposed CADM technique are image preprocessing, feature extraction and fusion, feature reduction, and categorization. The basis of CADM is the wavelet packet decomposition (WPD) function, which used a directed gradient histogram to combine the reconstructed features.Then feature data was condensed using principal component analysis [13].

3. Result

3.1. Dataset

It is significant to analyze histopathological tissue before diagnosing the typical carcinoma such as BC (breast cancer). The BreCaHAD histopathological annotation and diagnostic data set for breast cancer allows researchers to test the method and evaluate its effectiveness so that approaches can be optimized timely. The data set has 162 breast cancer histopathology images from surgical pathology, including sufficient severe disease cases [14]. Sample cases were gathered from numerous settings, including corporate structures with clearly specified borders and structures with weakly differentiating characteristics. It is worth mentioning that this data set is publicly accessible among the biomedical imaging community [14]. About the grading system, the data set strictly adhered to the breast cancer grading system. The Nottingham scoring system, also known as the modified Elston-Ellis version of the Scarff-Bloom-Richardson scoring system,is used [15,16]. The system is often used in breast tissue grading. There are three main characteristics: mitotic counting, nuclear pleomorphism, and tubular formation. Each is given a score of 1 to 3, which is added up to a total score ranging from 3 to 9. Then the degree of the BC can eventually be identified [14]. Creating the data set requires the expertise and experience of pathologists with long-time work as well.

3.2. Result comparison

Performances of the aforementioned methods are demonstrated in Table 1, and Table 2 shows the performances measured by f1-value, prevision and specificity. Their corresponding visualization results are illustrated in Figure 1 and Figure 2 respectively for more intuitive understanding of the performances.

Tabel 1. Result comparison measured by accuracy.

Method	Accuracy
ANSM	90.88%
BCRS layer1	61.06%
BCRS layer2	97.52%
BCRS layer3	97.39%
DNNs: MobileNetV2	98.74%
DNNs: NASNetLarge	83.65%
DNNs: EfficientNetB0	72.96%
DNNs: ViT	90.57%
DNNs: DenseNet201	97.48%
CADM	97.10%

Tabel 2. Result comparison measured by F1-score, specificity, precision and accuracy.

Method	Accuracy	Specificity	Precision	F1-score
ANSM	90.88%	68.03%	95.7%	94.74%
BCRS layer1	61.06%	NA	73.0%	74.0%
BCRS layer2	97.52%	NA	94.0%	94.0%
CADM	97.10%	97.33%	97.75%	97.33%

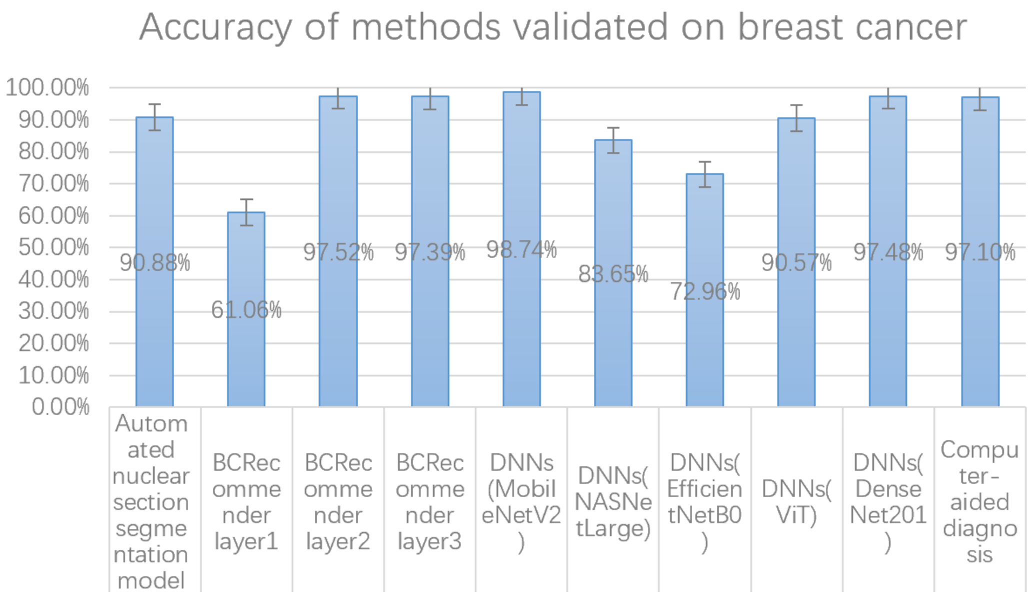

Figure 1. Accuracy performances demonstrated by bar chart.

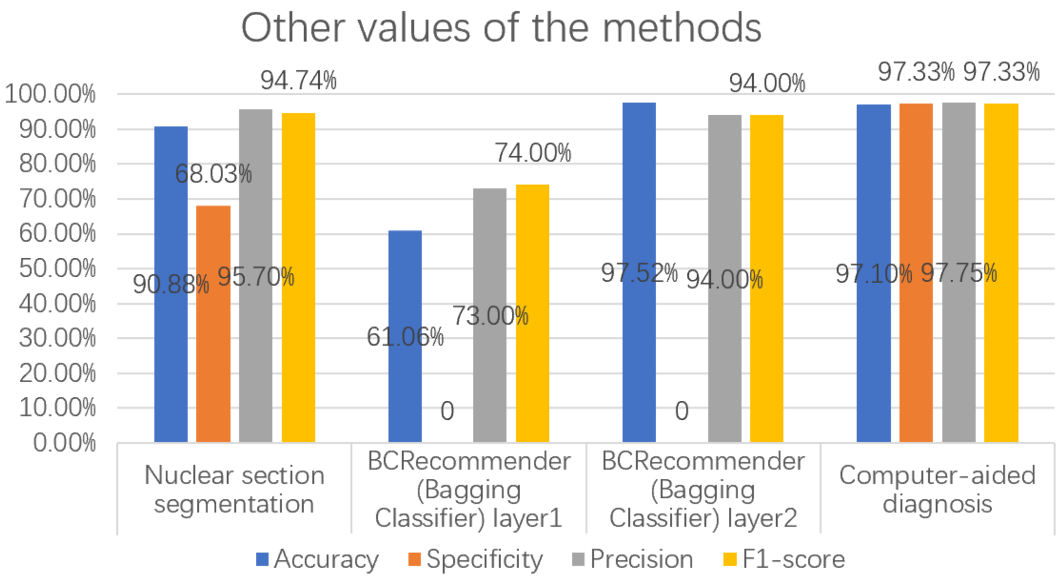

Figure 2. Performances measured by other values demonstrated by bar chart.

As all the methods used are based on the BreCaHAD Data Set, the results are relatively objective. As the figures (table and bar chart) show above, the methods found have pretty high accuracy in diagnosing breast carcinoma and some of them even reached an accuracy exceeding 95%. One of the DL models called MobileNetV2 has the highest accuracy (98.74%) of all types of models.

Additionally, the accuracy of DenseNet201 (DL model) is the third highest, reaching 97.48%, merely 0.04% less than the next best (97.39%). As for the rest of the DL models: NASNetLarge (DL models), EfficientNetB0 (DL models), and ViT (DL models), the accuracy rates are 83.65%, 72.96%, and 90.57%, respectively. Although layer 1 of the BCRecommender (Bagging Classifier) has an accuracy of only 61.06%, layer 2 and layer 3 has an accuracy of higher than 97%. It is worth mentioning that the nuclear section segmentation model (90.88%) thresholding and the computer-aided diagnosis (97.10%) are both with great accuracy in evaluating Breast Cancer. According to the specificity, precision, F1-score, and accuracy, the technique called computer-aided diagnosis takes the dominant position, for four values are all higher than 97%. Compared with the CADM, the nuclear section segmentation model has a lower specificity (68.03%) but is still high in precision (95.7%) and F1-score (94.74%). Therefore, CADM should be the most reliable method for predicting breast cancer statistically.

Traditionally, the trained CNN classifier is used for extracting features and transferring the images to the previously created classifier in order to classify the test image [9]. Whereas, the CADM system employs a new DCNN with convolutional layers, a compact SE-ResNet module, and fully connected layers for breast cancer classification. Moreover, the system can be further improved by incorporating additional features or optimizing the hyperparameters [13]. By comparing, the nuclear section segmentation model is low in specificity. However, this method also has some obvious advantages, because it relies on pixel clustering, mathematical parameters, and unsupervised machine learning. The algorithm can reduce the time duration of a pathologist to take an effective decision [10]. It also enhances the efficiency and accuracy of automatic segmentation by understanding the issue of covering core segments in test pictures [10]. With respect to the BCRecommender, it gives layer-by-layer proposals for breast cancer conclusion, making it a productive instrument for the convenient recognizable proof of cancer and guaranteeing the next chance of effective treatment [11]. Due to the different methods used in each layer, the accuracy of other layers is relatively considerable except for the first layer. Actually, MobileNetV2, NASNetLarge, EfficientNetB0 (DL models), ViT, and DenseNet201 are all advanced DL models. The advantage of the method is that it provides a way to not only evaluate the accuracy of the models but their capacity to recognize the proper districts where the tumor cores are found as well, which can aid pathologists with tumour diagnoses [12].

4. Conclusion

Through comparison, the MobileNetV2 model takes the dominant position in accuracy, with a proportion of 98.74%. Also, many models have an accuracy higher than 90% (ViT: 90.57% and Automated nuclear section segmentation model: 90.88%). Some of the percentages are even over 95%, such as layer 1 and layer 2 of the BCRecommender (Bagging Classifier). In short, the models based on machine learning demonstrate impressive accuracy and effectiveness. Considering all the values, including specificity, precision and F1-score, the CAD is the most reliable model for detecting breast cancer. The proposed CAD system employs a novel DCNN architecture and compares favorably with other powerful models. In addition, systems can be further refined by incorporating additional features or optimizing the hyperparameters. It is proved that the DCNN has the capability for the categorization of breast cancer, especially the time trained on large datasets like BreCaHAD.

The research aims to find the effect of machine learning in evaluating BC and which types of models work better. The article describes each method and the dataset used quite visually, as well as the results they produced. The research is informative for those who want to diagnose aspects of breast cancer through machine learning or machine learning-based models. Certainly, there are some limitations of the research. For example, the models involved are not enough. Moreover, this study wanted to highlight the benefits of machine learning-based models in diagnosing breast cancer. Although very high accuracy can be demonstrated in terms of data, no comparison with other diagnostic methods is made to highlight the advantages of machine learning models. Consequently, the improvement direction could be adding other models for comparison. The prerequisite is that these models are based on the BreCaHAD dataset and that they are compared in various ways.

References

[1] Gabriel, C. A., & Domchek, S. M. (2010). Breast cancer in young women. Breast cancer research, 12, 1-10.

[2] Nelson, H. D., Pappas, M., Cantor, A., Griffin, J., Daeges, M., & Humphrey, L. (2016). Harms of breast cancer screening: systematic review to update the 2009 US Preventive Services Task Force Recommendation. Annals of internal medicine, 164(4), 256-267.

[3] Kerlikowske, K., Zhu, W., Tosteson, A. N., Sprague, B. L., Tice, J. A., et, al. (2015). Identifying women with dense breasts at high risk for interval cancer: a cohort study. Annals of internal medicine, 162(10), 673-681.

[4] Tagliafico, A. S., Piana, M., Schenone, D., Lai, R., Massone, A. M., & Houssami, N. (2020). Overview of radiomics in breast cancer diagnosis and prognostication. The Breast, 49, 74-80.

[5] Phi, X. A., Tagliafico, A., Houssami, N., Greuter, M. J., & de Bock, G. H. (2018). Digital breast tomosynthesis for breast cancer screening and diagnosis in women with dense breasts–a systematic review and meta-analysis. BMC cancer, 18, 1-9.

[6] Ling, L., Aldoghachi, A. F., Chong, Z. X., Ho, W. Y., Yeap, S. K., et, al. (2022). Addressing the Clinical Feasibility of Adopting Circulating miRNA for Breast Cancer Detection, Monitoring and Management with Artificial Intelligence and Machine Learning Platforms. International Journal of Molecular Sciences, 23(23), 15382.

[7] Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J., & Maria Vanegas, A. (2020). Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors, 20(16), 4373.

[8] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity, 25, 1315-1360.

[9] Sikder, J., Das, U. K., & Chakma, R. J. (2021). Supervised learning-based cancer detection. International Journal of Advanced Computer Science and Applications, 12(5), 863-869.

[10] Kumar, A., & Prateek, M. (2020). Automated Detection and Classification of Ki-67 Stained Nuclear Section Using Machine Learning Based on Texture of Nucleus to Measure Proliferation Score for Prognostic Evaluation of Breast Carcinoma, 1, 1-5.

[11] Bhargava, H., Makeri, Y. A., Gyamenah, P., Gupta, S., Vyas, G., Sharma, A., & Chatterjee, S. (2022). BCRecommender System for Breast Cancer Diagnosis using Machine Learning Approaches, 1, 1-13.

[12] Macedo, D. C., De Lima John, W. S., Santos, V. D., LO, M. T., et, al. (2022). Evaluating Interpretability in Deep Learning using Breast Cancer Histopathological Images. In 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images, 1, 276-281.

[13] Anwar, F., Attallah, O., Ghanem, N., & Ismail, M. A. (2020). Automatic breast cancer classification from histopathological images. In 2019 International conference on advances in the emerging computing technologies, 1, 1-6.

[14] Aksac, A., Demetrick, D. J., Ozyer, T., & Alhajj, R. (2019). BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis. BMC research notes, 12(1), 1-3.

[15] Elston, C. W., & Ellis, I. (1991). I. The value of histological grade in breast cancer: experience from a large study with long - term follow - up. Pathological prognostic factors in breast cancer. Histopathology, 19, 403-410.

[16] Bloom, H. J. G., & Richardson, W. (1957). Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. British journal of cancer, 11(3), 359.