Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets

Ratula Ray¹, Azian Azamimi Abdullah², Debasish Kumar Mallick³, Satya Ranjan Dash⁴*
¹School of Biotechnology, KIIT Deemed to be University, Bubaneswar, India
²School of Mechatronic Engineering, Universiti Malaysia Perlis, Malaysia
³,⁴School of Computer Applications, KIIT Deemed to be University, India
⁴sdashfca@kiit.ac.in

Abstract: Breast cancer has been identified as the second leading cause of death among women worldwide after lung cancer and hence, it becomes extremely crucial to identify it at an early stage, which can considerably increase the chances of survival. The most important part in cancer detection is to be able to differentiate between benign and malignant tumors and this is where the work of Machine Learning comes in. Taking all the dependent features upon consideration, Supervised Machine Learning methods allow for classification with higher degree of accuracy and improve upon the misdiagnosis of the physicians, which might occur almost 20% of the time. In our paper, we are focusing towards understanding the shortcomings of digital mammograms in detection of breast cancer and utilize Machine Learning classifiers for the classification of benign and malignant tumors using image analysis. Apart from this, we are also looking into implementing Supervised Machine Learning classifiers such as Decision Tree, K Nearest Neighbour (KNN), Random Forest and Gaussian Naive Bayes classifiers for assessing the risks involved with breast cancer by analyzing the biomarkers that are involved with it. Our aim is to provide a comprehensive view on prediction of breast cancer through Machine Learning through both image and data analyses, which can play a pivotal role in prevention of misdiagnosis in future. Fig. 1. gives a layout for the breast cancer prediction using Supervised Machine learning classifiers.

1. Introduction

There are majorly two types of breast cancer: invasive and non-invasive breast cancer. Invasive form of breast cancer is more dangerous and it metastasizes from one location and spread to the other organs. Ducts and Lobes are the two most common sites where the breast cancer invades and it can rapidly progress from that point on. Based on these two sites, there are about four different types of breast cancer:

- **Ductal carcinoma in-situ (DCIS):** It is a non-invasive form of breast cancer, where the cancer is restricted to the milk ducts. There is always a chance of recurrence of the cancer even if it has subsided.
Lobular carcinoma in-situ (LCIS): Like DCIS, this form of cancer is also non-invasive but is restricted to the milk producing glands of the breast or lobules. LCIS is associated with abnormal growth of cells on the lobules of the breast.

Invasive ductal carcinoma (IDC): This is a metastatic form of breast cancer and also the most common type. This affects not only the milk ducts but also the surrounding layers of fatty tissues and travels further into the lymphatic system. In our paper, we will deal with this form of breast cancer classification in greater detail.

Invasive lobular carcinoma (ILC): Similar to IDC, this form of cancer is also not restricted to the site of occurrence. Though it originates at the lobules of the breasts, it also rapidly metastasizes to the nearby tissues, affecting them considerably.

Histology can be defined as the study of the anatomy of the tissue and the preparation steps of the histology slide involves the steps such as: fixing, processing, embedding, sectioning and staining [1]. Histopathological analysis of breast cancer refers to the analysis of the biopsies of the breast tissues to identify the region where the cancer is concentrated at and this work is mostly performed by a pathologist. Classification tasks using Machine Learning tools gives the computer the ability to differentiate the entire dataset into multiple classes based upon the features they are being associated with. The dataset that we use can either be in the form of numeric data or images (which need to be transformed into vector forms before it can be used for the classification task). In our paper, we are concerned about classification based on two separate aspects:

a. Classification of the presence of breast cancer based on important biomarkers (numeric data classification)

b. Classification of the histopathological images of breast tissue biopsies to identify particular regions of IDC affecting the tissues to confirm presence of breast cancer (image analysis)

2. Literature review

In our literature survey, the first paper we studied deals with the detection of micro calcifications on breast tissues for aiding in the diagnosis of breast cancer by studying the mammogram database collected from the University of Wisconsin. The image dataset used for this purpose adjusts the contrasts by analyzing the histograms of the respective images, which is followed by application of morphology filtering. Finally through K-means clustering, in association with fuzzy rule based classifiers, the classification process is carried out. The suggested algorithms led to the goal of detection of the cancer type [2].

Literature also suggests medical image analysis performed for the breast cancer dataset, which is followed by the meta-training model. For this they have set up a SOTA for weakly supervised breast screening problem. Several baseline and multiple tasks trained by DenseNet and some fine-tuning by using MIL framework have been applied. But MIL's performance does not provide improved results. In another proposed meta-training of same dataset model, CL sampling yields more accurate result than random sampling, whose performance is closer to that of DenseNet [3].

Another paper also reveals case–control studies of 1171 breast cancer cases with 1659 controls. The digitized film-screen analysis is done by checking for Mammographic breast density and 46 breast texture features are taken into account which includes first and second order feature, fractal analysis and Fourier transformation. The logistic regression model is used to evaluate the features involved in detection of breast cancer after adjustment of the details. Fractal dimensions for the thresholds around 10% and 15% indicate increasing of the risk, whereas fractal dimensions of 60% to 95% indicates decreasing of risk of cancer. Overall, the result of the risk involved in this case depends upon the density of mammographic image [4].

One of the most significant drawbacks of mammographic images is the high rate of false positive for breast cancer diagnosis. Ultrasound data yields more accurate results compared to mammograms in this case. Paper suggests the use of computer-aided system (CAD), which can be used to find out the area
affected by cancer more accurately. Utilizing FSVM (Fuzzy Support Vector Machine) with ultrasound dataset in association with feature extraction and selection techniques have also been developed. For this purpose, results have shown that the accuracy of ANN classifier is 88.51%. Moreover, it has also been seen that when SVM classifier yields an accuracy of 87.36%, in that case, FSVM yields 94.25% accuracy [5].

Computer aided detection system is also used to identify the position of the area of the breast affected by cancer using the sum-of-square error function, which improves the efficiency and robustness in identification of the micro calcification clusters. By using artificial neural network, convolution neural network and linear discriminant analysis, selection of the features has been done in cluster method. The result was found by analyzing 96 medical cases with 192 images dataset collected from University of Michigan, and it yielded an accuracy of 96% [6]. Method for development for reduction of false positive results have been developed using computerized detection based on analysis of bilateral information. The bilateral computer aided detection can be used for reducing false positives by training [7]. XR mammography detected 31 patients who were classified for breast cancer according to BIRADS (Breast Imaging-Reporting and Data System), along with their biopsy size and fractal dimension. The parenchyma fractal dimension lesions area showed complex breast calcifications that were consistent with their morphological pattern. The Fractal dimension analysis was used for getting result [8].

3. Implementation

3.1. Dataset

We have utilized the Breast Cancer Coimbra dataset for our prediction analysis of the numeric dataset, which takes into consideration the different biomarkers for breast cancer prediction [9]. For image analysis, we have used the Kaggle dataset for Breast Histopathology images [10].

3.2. Methodology

Figure 1 gives a layout for the methodology used for classifying benign and malignant breast cancer using Supervised Machine Learning classifiers.
3.3. Results

We have used the Precision, Recall and F1-score as the metrics for statistical analysis of the datasets. Table 1. gives the result for the numeric data used for analysis of breast cancer based on different biomarkers such as Age, BMI, Glucose, Insulin, HOMA, Lectin, Adiponectin, Resistin, MCP.1.

The metrics used for this purpose are as follows:

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\]

\[
\text{F1-score: } \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}
\]

| Classifiers                      | Precision | Recall | F1-score |
|----------------------------------|-----------|--------|----------|
| Gaussian Naive Bayes Classifier  | 66%       | 60%    | 59%      |
| Decision Tree Classifier         | 70%       | 70%    | 70%      |
| K-nearest neighbors Classifier(n=5) | 48%     | 48%    | 48%      |
| Random Forest Classifier         | 71%       | 71%    | 70%      |

Out of the different classifiers used, we find Decision Tree and Random Forest classifiers showing promising results with an accuracy score of 70%. We hence used Random Forest classifier to check the most important features linked with the prediction of the end result. Figure 2 shows the graphical representation of the ranking of the biomarkers with their relevance to the presence of breast cancer. Out of them, BMI(Body Mass Index) and Resistin show the most dependency on the end result.

![Figure 2. Graphical representation of the feature ranking according to their importance with the data label](image-url)
The performance of the same sets of classifiers taken for histopathological image analysis for breast cancer tissues are depicted in Table 2. The analysis was done by taking into account the variability of color change upon staining.

### Table 2: Statistical results determining the performance of the classifiers for the histopathological image of breast tissues dataset

| Classifiers                          | Precision | Recall | F1-score |
|-------------------------------------|-----------|--------|----------|
| Gaussian Naive Bayes Classifier     | 86%       | 83%    | 84%      |
| Decision Tree Classifier            | 87%       | 87%    | 87%      |
| K-nearest neighbors Classifier (n=5)| 88%       | 88%    | 88%      |
| Random Forest Classifier            | 91%       | 92%    | 92%      |

### 4. Conclusion

Histopathological image analysis is the microscopic analysis of breast tissue derived from biopsy specimens. In our case, we have used segmentation of the cellular components on the tissue samples based on color of the staining. Identification of the cancer nuclei is the basic aim of this computational based analysis and this can be done by sectioning of the region of interest from the background information. In case of H and E staining of breast tissue samples, blue color denotes the nuclei and the cytoplasm, whereas the stroma is stained with pink and the RBCs are stained in red. Our result depicts that Random Forest Classifier shows the best result with F1-score of 92%, followed by K-nearest neighbors classifier. Future prospects in this field is huge and there are many aspects of segmentation of the images that can be carried out, like region-based segmentation, shape-based segmentation and texture-based segmentation, which can give much improved and reliable results. Computational constraint for using large dataset for image classification is the reason we have used train-test split as the validation method instead of k-fold cross validation in this case. For the numeric dataset, we have seen that Random Forest and Decision Trees give comparable results. Our findings also suggest that BMI and Resistin are the most important biomarkers that affect the breast cancer results. BMI acts as a major biomarker for breast cancer and it is linked differently in pre and postmenopausal women. It is basically due to the difference in production of estrogen and progesterone and thus in case of premenopausal women, higher the BMI; lower is the risk of developing breast cancer. However, in case of post-menopausal women, the risk of breast cancer is positively linked with increasing BMI. Resistin is an adipocytokine and is known to be positively related with the increasing risk of breast cancer.

In this paper, we have highlighted the different classifiers powered by Machine learning techniques in order to detect the breast cancer using both numeric and image datasets. We are hopeful that these techniques will assist the pathologists and doctors to make much more precise and informed decisions and improve the detection process significantly.

### References

[1] Aswathy, M. A., & Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities. Informatics in Medicine Unlocked, 8, 74-79.

[2] Guardado-Medina, R. O., Lepe, R. R., & Magaña, B. O. (2012). Intelligent detection of microcalcifications in mammograms based on fuzzy techniques.

[3] Maicas, G., Bradley, A. P., Nascimento, J. C., Reid, I., & Carneiro, G. (2018, September). Training medical image analysis systems like radiologists. In International Conference on
Medical Image Computing and Computer-Assisted Intervention (pp. 546-554). Springer, Cham.

[4] Malkov, S., Shepherd, J. A., Scott, C. G., Tamimi, R. M., Ma, L., Bertrand, K. A., ... & Norman, A. (2016). Mammographic texture and risk of breast cancer by tumor type and estrogen receptor status. Breast Cancer Research, 18(1), 122.

[5] Shi, X., Cheng, H., & Hu, L. (2006, October). Mass detection and classification in breast ultrasound images using fuzzy SVM. In 9th Joint International Conference on Information Sciences (JCIS-06). Atlantis Press.

[6] Ge, J., Sahiner, B., Hadjiiski, L. M., Chan, H. P., Wei, J., Helvie, M. A., & Zhou, C. (2006). Computer aided detection of clusters of microcalcifications on full field digital mammograms. Medical Physics, 33(8), 2975-2988.

[7] Wu, Y. T., Wei, J., Hadjiiski, L. M., Sahiner, B., Zhou, C., Ge, J., ... & Chan, H. P. (2007). Bilateral analysis based false positive reduction for computer aided mass detection. Medical physics, 34(8), 3334-3344.

[8] Verma, G., Luciani, M. L., Palombo, A., Metaxa, L., Panzironi, G., Pediconi, F., ... & Todde, V. (2018). Microcalcification morphological descriptors and parenchyma fractal dimension hierarchically interact in breast cancer: A diagnostic perspective. Computers in biology and medicine, 93, 1-6.

[9] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC cancer, 18(1), 29.

[10] Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics, 7.