A Survey on Analysis and Classification of Breast Cancer

Nandish Sonali¹, Prathibha Ramapura Javaregowda²

¹Department of Computer Science and Engineering, JSS Science and Technology University, Mysuru 570006, India; ²Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570006, India.

ABSTRACT

Introduction: The reports obtained after pathological examinations on breast cancer have been digitized by sophisticated machines and stored as Electronic Health Records (EHRs). These EHRs contribute to Computer-Aided Diagnosis and for better clinical decision support.

Objectives: Understanding of various machine learning techniques used for the classification of standard and real-time breast cancer datasets. Reviewing of classification of various types of dataset like images, text, numerical values etc., on breast cancer.

Methods: In this paper, a rigorous literature survey has been made on the classification of the dataset on breast cancer using various machine learning methods on standard datasets like Breast Cancer Dataset, Wisconsin Breast Cancer Dataset, Wisconsin Breast Cancer Diagnostic Dataset, Wisconsin Breast Cancer Prognostic Dataset and Surveillance Epidemiology and End Results Dataset etc. In the literature, it has been observed that some of the authors have worked on the classification of datasets that are collected from different hospitals. Images of the breast have been analyzed by looking at the property of luminance, colour and shape variation, texture, reaction to biomarkers and many other factors. For understanding proliferation in breast cancer, various scoring systems are used. They include Bloom-Richardson Score, Masood Score, Modified Masood Score, Robinson’s Score and many others. The EHRs containing the records in text form on breast cancer have been interpreted using Natural Language Processing approaches like text segmentation, named entity recognition and part of speech tagging etc., and classified using machine learning approaches.

Results: Classification of breast cancer has been made on different types of datasets using machine learning methods and the range of accuracy obtained is between 75.60% and 99.86%.

Conclusion: Most of the existing classifiers are binary classifiers to classify breast cancer datasets into benign and malignant classes. However, it is necessary to design multiclass classifiers for building a precise clinical decision support system and to provide targeted therapy for cancerous patients using cost-effective diagnostic methods.

Key Words: Breast Cancer, Digital Image Processing, Scoring System, Natural Language Processing, Pathology, Cytopathology, Histopathology

INTRODUCTION

Cell division is a normal process in the human body where cells grow old or become damaged and die and new cells take their place. Breast cancer develops from breast tissue when cells in the region grow out of control.¹

Diagnosis and prognosis of breast cancer are very important for its effective clinical management and treatment. It is noted that the lack of proper detection of breast cancer has increased the number of cases in India and the World.

Breast cancer affects nearly 34% of women between the age of 20 and 100 years worldwide² and it is expected to cross 1,00,000 patients annually in India.³

Computer-Aided Diagnosis (CAD) is used to enhance the diagnosis and prognosis of breast cancer. Under pathology, breast cancer detection is mainly based on the cell morphology and architecture distribution. Breast cancer can be classified into benign and malignant by considering nuclei detection, nuclei segmentation and the number of nucleoli using Digital Image Processing.⁴ The standard approach to
analysing the image dataset includes analysis of Haematoxylin (H) & Eosin (E) images. 7 Natural Language Processing approaches are used for text report segmentation, clinical information retrieval, part-of-speech tagging, named entity recognition and context extraction. 8 We can observe that authors Fushman et al., 9 have discussed that Clinical Decision Systems (CDS) have improved practitioner’s performance by 60% in the reviewed cases of breast cancer. Most commonly, the binary classification is made which divides breast cancer into benign and malignant classes. Another approach used for breast cancer diagnosis is the grading system. 8, 9, 10 A review of the classification of breast cancer on different types of datasets using machine learning approaches is given in the next section.

**Methodology**
An extensive literature review has been made on the analysis of breast cancer considering Histopathology and Cytology data in numeric, image and text form. The data is obtained from standard datasets, cancer registries and real-time data from various hospitals.

**Review on Classification of Breast Cancer using Standard Datasets**

**Classification of Breast Cancer using Histopathological Numerical Data**
Authors Delen et al., have made survival analysis of breast cancer on Surveillance, Epidemiology and End Results (SEER) dataset and classified them into survivability and non-survivability of the patient using Artificial Neural Network (ANN), Decision Tree (DT) and Logistic Regression (LR) approaches and obtained accuracy of 91.2%, 93.6% and 89.2% respectively. 10 Authors Qin et al., have analyzed the imbalanced SEER dataset and obtained an accuracy of 76.59% and Area Under the Curve (AUC) of 76.78%, respectively. 11 Authors Rajesh et al., have used the C4.5 classification algorithm on SEER dataset to classify patients into either the ‘Carcinoma in situ’ group or ‘Malignant potential’ group and obtained an accuracy of 92.2%. 12

Rathore et al., have analyzed SEER dataset by using an ensemble classification approach using DT, Naïve Bayes (NB), Multiple Association Rule for predicting the survivability of breast cancer patients and the best accuracy obtained is 71.87%. 13 Swetha Karya has analyzed SEER dataset with DT classifier and achieved a classification accuracy of 93.62% with a sensitivity of 96.02% and a specificity of 90.66%. 14 The authors Umesh et al., have analyzed SEER dataset using the association rule mining method and obtained a sensitivity of 56.32%, specificity of 91.78% and accuracy of 87.72%. 15, 16 The authors Yeulkar et al., have used C4.5 and NB classifier on SEER dataset containing breast cancer samples for the period 1975 till 2013 and obtained an accuracy of 98.1% for C4.5 and 95.85% for NB approach. 18

**Classification of Breast Cancer using Cytological Numerical Dataset**

**BC and WBC dataset**
Lavanya et al., have used the Classification and Regression Technique (CART) for analysis on BC and WBC datasets and obtained an accuracy of 69.23% and 94.84% respectively. 20 Authors Paulin et al., have analyzed the WBC dataset using Feedforward Neural Network (FNN) to obtain the highest diagnostic performance of 99.26% using 6 neurons. 21 The authors Salama et al., have analyzed the WBC dataset and obtained the highest accuracy of 97.5% by using an ensemble classifier containing five classifiers viz., i) J48, ii) MultiLayer Perceptron (MLP), iii) NB, iv) SVM and v) k-NN classifier with Principal Component Analysis (PCA). 22 The authors Inan et al., have analyzed the WBC dataset using a hybrid approach including Apriori Algorithm and PCA together with ANN classifier. They have used 10-fold cross-validation and obtained average classification accuracy of 98.29%. 23 Authors Tintu et al., have analyzed the WBC dataset using MLP, SVM, NB, Fuzzy C-Means (FCM) for breast cancer diagnosis. The best accuracy was obtained for FCM with a training accuracy of 97.13% and a testing accuracy of 98.62%. 24 Ravikumar et al., have analyzed the WBC dataset and obtained the best results for the SVM classifier with an accuracy of 97.59%, the sensitivity of 98.10% and specificity of 96.60%. 25

The authors Grewal et al., have analysed the WBC dataset and obtained a sensitivity of 95% and specificity of 98.8%. 26 The authors Kathija et al., have analysed the WBC dataset by using NB and SVM classifier along with 10-fold cross-validation technique. The best accuracy of 95.6% is obtained with the sensitivity of 97% and specificity of 100% using the NB Classifier. 27 Chaursia et al., have proposed prediction of benign and malignant conditions on standard WBC dataset. The authors have used six classifiers viz., i) NB, ii) RBF, iii) J48, iv) SVM, v) K-NN and vi) RBF tree. The highest accuracy obtained is 97.36% for NB. 28

**WBCD dataset**
Lavanya et al., have used CART for analysis on the WBCD dataset and obtained an accuracy of 92.97%. The authors Salama et al., analyzed the WBCD dataset and obtained the
highest accuracy of 97.7% by using an ensemble classifier containing SVM and MLP classifier. Shweta Karya has analyzed the WBCD dataset using a decision tree classifier and obtained the best accuracy of 93.62%. Menaka et al., have analyzed WBCD datasets using SVM with RBF and obtained an accuracy of 97.37% respectively.

The author Leena Vig has applied SVM, NB and Random Forest (RF) classifiers with 100 decision trees on the WBCD dataset and achieved the best accuracy of 95.64% with a sensitivity of 97% and specificity of 94%. Hazra et al., have analyzed data by considering only 5 features on 32 features from the WBCD dataset using an ensemble of NB and SVM classifiers and obtained an accuracy of 97.4%. The author Agarap has proposed a model with MLP that gave the best performance measure with an accuracy of 99.04%.

**WBCP dataset**
Authors Tintu et al., have analyzed the WBCP dataset using MLP, SVM, NB, FCM for breast cancer prognosis and obtained 100% True Positive (TP) and 87% True Negative (TN) rates. The authors Wolberg et al., have built a neural network model on the WBCP dataset for prognosis prediction. They obtained a probability that 50% of patients would be disease-free when the period considered for breast cancer recurrence was less than or equal to 5 years from the time of occurrence of cancer and 90% of patients would be disease-free when the period considered was greater than 5 years. Senturk et al., have made their analysis on these standard datasets using Rapid Miner 5.0 data mining tool with an accuracy of 98.4%.

The best performance obtained by various approaches on each type of standard dataset is given in Table 1.

**Review on Classification of Breast Cancer using Grades or Scoring System**
In recent years, the analysis of breast cancer has been expanded from binary to multiclass classification. Hence, the concept of grading or scoring the lesions has been considered. The known methods of grading or scoring include Bloom-Richardson Score (BRS), Modified Bloom-Richardson Score (MBRS), Masood Score (MS), Modified Masood Score (MMS) and many others. Under these methods, the characteristics of breast lesions are measured and an interval of value is fixed with a particular grade or score.

**Classification of Breast Cancer on Histopathological Numerical Dataset**
The authors Meyer et al., have classified 631 patients from St. Luke’s Hospital, USA using BR Score and have obtained a kappa statistic of 0.38. Rekha et al., have proposed an MBR grading system on 50 breast carcinoma cases from a tertiary centre at Mysore, India and have obtained a histopathological correlation of 86%.

**Classification of Breast Cancer using Cytological Numerical Dataset**
Authors Mridha et al., have used Masood’s score on 62 breast cancer patients from the All India Institute of Medical Sciences and obtained specificity for FNAC technique for carcinoma between 89% to 98% and sensitivity between 93% to 98%. Nandini et al., have proposed an MMS system to classify 100 lesions samples into four categories with an accuracy of 96%. Sheeba et al., have made the comparison of both MS and MMS methods on 100 cases collected at Kilpauk Medical College, Chennai, India. The Cyto-histological correlation is 88% and the accuracy of MMS is 84%. The authors Cherath et al. have collected a dataset of 207 cases in a tertiary health centre in South India, to analyse the samples using the MS and MMS approaches and obtained an overall accuracy of 97.5%, the sensitivity of 94.5% and specificity of 100%. It is also validated that MMS is a better scoring system than MS.

**Review on Classification of Breast Cancer using Image Dataset**
A computer-Aided Diagnosis (CAD) algorithm has been developed for the detection and prediction of diseases and to assist the pathologist for better clinical decision making. Under pathology, histopathological analysis is considered as the golden standard by pathologists.

**Classification of Breast Cancer using Histopathological Image Dataset**
The authors Jelen et al., have used a database that consists of 110 FNA Biopsy (FNAB) images from the University of Wroclaw, Poland. There are 44 images with high malignancy and 66 images with intermediate malignancy. They have used the SVM framework to assign a malignancy grade based on pre-extracted features with an accuracy of up to 94.24%. The authors Cosatto et al., have used 208 histopathological images from St Luke’s Hospital, Chesterfield, USA, to identify the Cancer Nuclei, using the Hough transform and Active Contour Model for segmentation. The authors have used an SVM classifier for morphology and texture-based classification and obtained 92% of True Positive Rate and 72% of Kappa statistical measure. Fatakadawala et al., have considered H&E stained breast biopsy cores at The Cancer Institute of New Jersey. For a total of 62 HER2+ breast biopsy images, the Expectation-Maximisation based segmentation with Geodesic Active Contour with Overlap Resolution (EMaGACOR) was found to have a detection sensitivity of over 90% and a positive predictive value of over 78%.

The authors Basavanahally et al., have used a total of 41 H&E stained breast biopsy samples from 12 patients at The Cancer Institute of New Jersey.
Jiang et al. have developed Breast 62-65 58-59,60 49-55 54-57 56, 57 54 Authors Zeng et al., have used 150 de-identified reports from the Spanhol et al., 56, 57 66 54 Authors Carell et al., have 93% of sensitivity and 95% of specificity. The clinical Text Analysis and Knowledge Extraction System (cTAKES) method is used for analysis and achieved 1472 patients obtained from Group Health Research Institute (NCI) cohort, with 248 samples and the Vancouver General Hospital (VGH) cohort, with 328 samples where both cohorts had the value of Probability P ≤ 0.001.50 53 Wang et al., have done colour recognition by applying a fuzzy inference system and combining RGB and CIE LAB colour space. The approach is used on a standard ICPR12 dataset with the combination of Hand Crafted (HC) features and features derived from Convolutional Neural Networks (CNN). The data has been analyzed using a combination of HC and CNN and obtained an F-Measure of 73.45%.51 Spanhol et al., have classified breast cancer images of Breakhis dataset by using deep features also called DeCAF. DeCAF features are neither HC nor fully automated in nature. They have obtained a classification accuracy of 90%.52 The authors Beevi et al. have proposed a Krill Herd Algorithm to differentiate mitotic and non-mitotic groups. They have obtained a Precision of 62.50% and a Recall of 93.75%.53 Jiang et al., have developed Breast Cancer Histopathology image Classification Network (BHC-Net) for binary classification of images using Breakhis dataset and obtained performance between 98.87% and 99.34%.54 The authors Dabeer et al., have analyzed 7909 images stained by H&E and paraffin on Breakhis dataset and classified using CNN and obtained an accuracy of 99.86%.55

**Review on Classification of Breast Cancer containing Text Data**

Electronic Health Records are most commonly available in the form of text data. Text data contains a lot of valuable clinical information to ascertain the exact cause and status of any disease. For diagnosis and prognosis of breast lesions, clinical data in the form of text can be extracted to obtain conclusions about the exact condition of breast cancer using machine learning approaches.56, 57

**Classification of Breast Cancer using Histopathological Text Data**

The authors Carell et al., have designed an abstraction search for breast cancer recurrence using clinical notes of 1472 patients obtained from Group Health Research Institute, Seattle, USA to identify the recurrence of breast cancer. The clinical Text Analysis and Knowledge Extraction System (cTAKES) method is used for analysis and achieved 93% of sensitivity and 95% of specificity.

**RESULTS**

In literature, it is observed that the breast cancer dataset has been analyzed by researchers using various machine learning approaches on standard datasets and obtained the best accuracy of 99.26%. For image data, the best accuracy of 99.86% is obtained by using CNN on 7909 Histopathology images that are collected from Breakhis dataset. Considering the grade or scoring system, the best accuracy of 97.5% is obtained by MMS on 207 samples collected from a tertiary centre in South India. The best accuracy of 96% is obtained by using TIES on 249 reports that are taken from the cancer registry database.
DISCUSSION

From the literature review, it is observed that breast cancer is manifested by abnormal growth of tumours in various parts of the breast. Some of the common areas of tumour growth observed under histopathology include the nipple, areola, lymphocytes, nodes, etc. Under cytology, it is observed by morphological changes in cell, nucleus and nucleoli.

CONCLUSION

Most of the existing classifiers are binary classifiers to classify breast cancer data into a benign and malignant classes. However, it is necessary to design multiclass classifiers on breast cancer datasets for precise clinical decision support to provide targeted therapy for cancerous patients.

Source of Funding: We hereby declare that there is no funding involved in our work.

Conflict of Interest: Nil

ACKNOWLEDGEMENT

The authors are also grateful to authors/editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

Authors’ Contribution

The author Nandish Sonali has contributed by reviewing the papers related to the topic and submitted the inference. The author Prathibha Ramapura Javaregowda has contributed by editing the content of the paper.

REFERENCES


42. Nandini NM, Rekha TS, Manjunath GV. Evaluation of scoring system in cytological diagnosis and management of breast lesion with the review of the literature. Indian J Cancer. 2011 Apr 1; 48(2):240-5.


Table 1: Best Performance obtained by various approaches on each type of standard dataset

<table>
<thead>
<tr>
<th>Dataset/ No. of Samples</th>
<th>Author and Year</th>
<th>Analysis Method</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEER</td>
<td>Yeulkar, Sheikh,2017</td>
<td>C4.5</td>
<td>Acc = 98.10%</td>
</tr>
<tr>
<td>BC /(286)</td>
<td>Lavanya, Usha Rani,2011</td>
<td>Classification And Regression Technique</td>
<td>Acc = 69.23%</td>
</tr>
<tr>
<td>WBC /(699)</td>
<td>Paulin, Santhakumaran, 2010</td>
<td>Feed-forward Neural Network</td>
<td>Acc = 99.26%</td>
</tr>
<tr>
<td>WBCD/(569)</td>
<td>Agarap,2018</td>
<td>Ensemble approach using GRU-SVM,LR,MLP,NN,SR,SVM</td>
<td>Acc = 99.04%</td>
</tr>
<tr>
<td>WBCP/ (198)</td>
<td>Tintu, Paulin,2013</td>
<td>Fuzzy C Means</td>
<td>Acc = 98.26%</td>
</tr>
</tbody>
</table>