Voting Classification Approach for Breast Cancer Detection

Ravi Kumar Barwal, Neeraj Raheja, and Pankaj Kumar

Abstract The cancer of breast is the most frequent type of cancer among ladies worldwide following by the lung cancer. A cycle of regeneration procedures in body is maintained with the cells. In general, the natural functioning of the body is maintained by balancing the development and death rate of cells. The pipeline of breast cancer detection includes many tasks such as preprocessing, segmentation, feature extraction and classification. In this research work, voting classification approach is applied for the breast cancer detection. The proposed approach improve performance in comparison to existing models in compliance with precision, recall and accuracy.

Keywords Breast cancer · GLCM · Threshold segmentation · Voting classification

1 Introduction

The technique of image processing is becoming progressively advanced and the trend is to develop more and more automation. The technology of image processing aims to improve crude imagery got from various sources or images clicked in typical day-to-day life to serve different purposes. Digital Image Processing (DIP) is an area of special importance in CSE and has its roots to the multiple areas as well. The significance of image processing technology in the healthcare sector is not hidden from anyone. This technology process images in digital manner using artificial intelligence algorithms. The effect of digitized pictures on present day culture is so incredible, and is a crucial part of science and technology. As of now, the sample of bloods are taken to laboratory and prepared with different substrates to generate outcomes. Image indexing and extraction based on content- has been a significant exploration zone in software engineering throughout the previous few years. Numerous digitized images are usually clicked and stored in different databases in different formats. Thus,

R. K. Barwal (✉) · N. Raheja
CSE Department, Maharishi Markandeshwar (Deemed to be University), Mullana 133207, India

P. Kumar
CSE Department, Government College for Women, Shahzadpur, Ambala, India

N. Marriwala et al. (eds.), *Emergent Converging Technologies and Biomedical Systems*,
Lecture Notes in Electrical Engineering 841,
https://doi.org/10.1007/978-981-16-8774-7_55

enormous datasets containing images are being made and utilized in numerous applications [1]. The cancer of breast is the most frequent type of cancer among ladies worldwide following by the lung cancer. A cycle of regeneration procedures in body is maintained with the cells. In general, the natural functioning of the body is maintained by balancing the development and death rate of cells. In any case, this is not generally the situation. Once in a while, an unusual circumstance takes place which leads to abnormal growth in some cells. Consequently, cancer occurs in a specific area of the body and spreads to other areas [2]. Distinct kinds of cancer may occur in individual's body and the cancer of breast is one that is considered as a serious health concern. Women have more susceptibility against breast cancer as compared to men because of the anatomy of the human body. Several reasons may cause breast cancer such as age, medical history, fatness and drinking, etc. The breast of a woman is made up of nipples, ducts and fatty tissues and lobules. In general, epithelial tumors are developed inside the lobules and ducts which cause tumor in the breast later on. The cancer after its initiation disseminates to the other areas of the body.

Breast cancers generally are of two types: Benign and Malignant. The first case has non-cancerous cells that do not cause death. However, in some scenarios, these cells could turn into cancerous cells. The defense system of body referred to as "sac" normally is useful in segregating the benign tumors from harmful cells and expels them from the body. The second type of cancer initiates when cells are growing abnormally and quickly. These cells can spread to the neighboring tissues [3]. The malignant tissue's nuclei is often larger than the regular tissue and can be life threatening [4]. Breast cancer is considered as the heterogeneous disease whose formation is done with different entities such as distinctive biological, histological and clinical attributes. The radiology images are deployed to carry out the clinical screening such as mammography, ultrasound imaging and MRI.

The preprocessing aids the localization of region for irregularity detection. The major concern in mammogram preprocessing faces is to outline the Pectoral Muscle (PM) boundary from the remaining breast area. The PMs generally occur in MLO views of the mammograms. The occurrence of PMs in the MLO view may disturb automated recognition of lesions and can intensify the false positive (FP) alarms. Most of the studies backed the elimination of PMs for making the diagnostic accuracy of the CAD system better. Therefore, it is highly important to successfully remove PMs for preventing false detection [5]. Additionally, it not only decreases the time complexity but also improves the accuracy besides preventing the intra-observation inconsistencies.

Mammograms do not deliver good contrast between healthy glandular breast tissues and cancerous tissues and between the malignant lesions and the background particularly in dense breasts. Poor contrast is inherent to mammography images. The Beer-Lambert equation states that the thicker the tissue is, the rarer the photons go through it. This implies that the X-ray beam goes through malignant breast tissues and normal glandular tissues [6] in dense breast tissue will not create attenuation with huge difference between the two tissues, and therefore less contrast between the healthy glandular and cancerous tissues will occur. Noise is the other common issue in mammograms. The non-uniform image brightness in the parts representing the same tissues causes noise in mammograms [7]. The non-uniform distribution of photons may be the reason of noise. This is known as quantum noise. This noise degrades the quality of images particularly in small objects with low contrast, for example, a mini tumor in a dense breast. Increase in exposure time may reduce quantum noise. The occurrence of noise in a mammogram makes it grainy. The grainy appearance decreases the discernibility of some features in the image specifically for small objects with low contrast representing the scenario for a mini tumor in a dense breast.

2 Literature Review

Wang et al. [8] projected a densely deep supervision technique for increasing the detection sensitivity in efficient manner. To achieve this, multi-layer attributes were employed. In addition, a threshold loss was put forward for presenting voxel-level adaptive threshold to determine the image as discerning cancerous or non-cancerous. Consequently, great sensitivity with was attained with less false positives (FPs). A dataset gathered from 219 patients that comprised 745 cancer regions were executed to calculate the accuracy of projected technique. The outcomes of experiment proved that the sensitivity provided through the projected technique was counted 95% with 0.84 false positives. An effectual cancer detection system was obtained from the projected technique to analyze the breast with the help of ABUS.

Khasana et al. [9] intended to implement watershed transform (WT) approach during segmentation procedure for creating the position of the tumor and distinguishing the matters on the basis of background. Afterward, the thresholding binaries were employed for splitting the tumor image in the form of an object. In the last stage, the area of cancer was computed. The outcomes indicated that the intended algorithm had provided the accuracy and error rate of around 88.65 and 11.35% of the overall tested data. The testing results showed that the intended algorithm was useful to detect the breast cancer with the help of ultrasound image.

Kavya et al. [10] put forward a strategy to detect the breast cancer in which imaging schemes namely, mammography and thermography were implemented. The computer aided diagnosis (CAD) tool was deployed as an effective technique for segmenting and classifying the digital images. The data taken from hospital was analyzed using this technique. The Cyber-physical system (CPS) was exploited to collect the data and share the details to particular systems. Furthermore, the network was integrated, human was interacted with system and system was made flexible, scalable and optimized with the help of CPS. The suggested approach was capable of detecting the breast cancer and providing high safety for the patients for which errors rates were mitigated and the data was monitored.

Varma et al. [11] introduced an approach for analyzing the digital mammograms. For this purpose, texture segmentation was employed for visualizing and detecting the images. Afterward, the attributes were extracted effectively due to which enhanced detection was obtained and proper action was taken for diminishing the risks related to breast cancer. Moreover, the introduced approach was adaptable to alleviate the processing time and enhance the processing speed. The final output images demonstrated that the introduced approach was efficient to outline the anomalies in the breast tissue and successfully detected the breast cancer.

Soliman et al. [12] designed an effective system for detecting the breast cancer with image processing methods. The significant attributes of the breast were extracted from the region of interest (ROI) whose segmentation was done from the thermal input image. Afterward, a neural network (NN) classification algorithm was utilized to categorize the image as cancerous or normal on the basis of these features. A benchmark dataset was applied to quantify the designed system and a success rate was obtained around 96.51%. The outcomes revealed that the designed system was efficient.

Liu et al. [13] presented two diverse microwave imaging techniques for detecting the breast cancer. In the initial technique, SAR an high-quality imagery technique was utilized and the second technique was planned on the basis of inverse scattering quantitative imaging technique for reconstructing the relative permittivity for the breast cancer. The presented techniques were tested on 2D simplified breast phantom. The outcomes of testing depicted the efficiency of presented techniques for detecting the breast cancer cell with good resolution.

Razavi et al. [14] recommended a computer aided diagnosis (CAD) method in which preprocessing median and Gaussian filters were deployed first of all. The cells of interest areas were segmented using an adaptive thresholding technique and watershed algorithm. Thereafter, the recommended technique was utilized to compute the ratio amid green and red FISH signals of all decomposable cells. Enormous Fluorescence in situ hybridization (FISH) images were deployed for presenting the automatic gene expression of epidemic growing feature receptors-2 (HER2) status. The outcomes exhibited the effectiveness of the recommended method to specify the HER2 state of probable patients.

Sangeetha et al. [15] established a new mechanism for detecting the breast asymmetry and micro calcification cancer cells in which various methods of digital image processing (DIP) were integrated that had not utilized in this research area. The established mechanism assisted in detecting the breast cancer in initial phase. An end to end (E2E) solution was acquired from this mechanism. The outcomes generated through the established mechanism were highly accurate in terms of true positive (TP) and true negative (TN).

Yin et al. [16] developed a new RAR algorithmic approach that planned a neighborhood pair-wise correlation-based weighting for dealing with the negative impacts of both artifact and glandular tissues. The maximum combination of these coefficients was applied to weight, sum and time-shift the backscattered signals. The three-dimensional finite-difference-time-domain models were exploited to evaluate the developed algorithm in anatomic and dielectric manner. The developed RAR algorithm had potential to recognize and detect the cancer under various scenarios. The outcomes validated that the developed algorithm was applicable for breast cancer screening.

3 Research Methodology

This noise degrades the quality of images particularly in small objects with low contrast, for example, a mini tumor in a dense breast. Increase in exposure time may reduce quantum noise. The occurrence of noise in a mammogram makes it grainy. The grainy appearance decreases the discernibility of some features in the image specifically for small objects with low contrast representing the scenario for a mini tumor in a dense breast.

Various phases are executed to predict the breast cancer and these phases are defined as.

3.1 Data Acquisition

This phase includes the collection of data from distinct clinical organizations to conduct tests.

3.2 Data Preprocessing

The entirety is accomplished and the data is analyzed to deploy the machine learning methods and the preprocessing is performed on the data. The redundant attributes are removed from the dataset in order to transmit the clean and de-noised data for enhancing the efficacy of training framework.

3.3 Feature Selection

The GLCM algorithm is applied in this phase for the feature extraction. This algorithm can be used for extracting multiple texture features. For this purpose, the considered metrics are as follows:

G denotes applied no. of Gray levels.

μ indicates the mean of P.

μ_x, μ_y represent the means while σ_x, σ_y denote standard deviation P_x, and P_y, respectively. $P_x(i)$ Signifies the ith entry in the marginal-probability matrix achieved after adding the rows of $P(i, j)$.

Contrast: This metric measure the local variations of an image.

$$CONTRAST = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G} \sum_{j=1}^{G} P(i, j) \right\}, \quad |i - j| = n \tag{1}$$

This metric that computes variance in contrast or local intensity favors aids from $P(i, j)$ outside the diagonal: $i \neq j$.

Homogeneity: This metric measures the nearness of the spreading of components in the GLCM to the diagonal of GLCM.

$$\sum_{i}\sum_{j}\frac{P_d[i, j]}{1 + |i - j|} \tag{2}$$

Local Homogeneity, Inverse Difference Moment (IDM):

$$IDM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{1}{1 + (i - j)^2}P(i, j) \tag{3}$$

The image homogeneity also affects IDM. IDM gets mini support from inhomogeneous regions $(i \neq j)$ as a result of the weighting factor $\left(1 + (i - j)^2\right)^{-1}$.

The low value of IDM for inhomogeneous images, and a comparatively greater value for homogeneous image is obtained as the result.

Entropy: This metric measures the content of the information. It is a measure of the unpredictability of distributed intensity. Inhomogeneous images have low 1st order entropy. A homogeneous image. On the other hand, has a great entropy.

$$-\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}P(i, j) \times \log(P(i, j)) \tag{4}$$

Correlation: This metric measures the gray level linear dependency amongst the pixels at the definite positions regarding each other.

$$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{\{i \times j\} \times P(i, j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \tag{5}$$

Sum of Squares, Variance:

$$VARIANCE = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(1 - \mu)^2 P(i, j) \tag{6}$$

This attribute lays comparatively high weights on the components different from the average rate of $P(i, j)$.

4 Voting Classifier

The voting classifier is a machine learning model that trains the integration of multiple models and predicts the output (class) according to their higher chances for the selected class as a result.

It simply summarizes the findings of each classifier transmitted to the voting classifier and predicts the output phase according to the number of votes. The idea is that instead of building separate dedicated models and obtaining each of them, we create a single model that trains these types and predicts the output according to their combined votes for each output phase (Fig. 1).

Voting Classifier supports two types of votings.

4.1 Hard Voting

In hard voting, the predicted output class is a class with the highest majority of votes, i.e., the class which had the highest probability of being predicted by each of the classifiers.

In this case, the class that received the highest number of votes $N_c(y^t)$ will be chosen. Here, we predict the class label $\hat{y}$ via majority voting of each classifier.

$$\hat{y} = \arg\max(N_c(y_t^1), N_c(y_t^2), \ldots, N_c(y_t^n)) \tag{7}$$

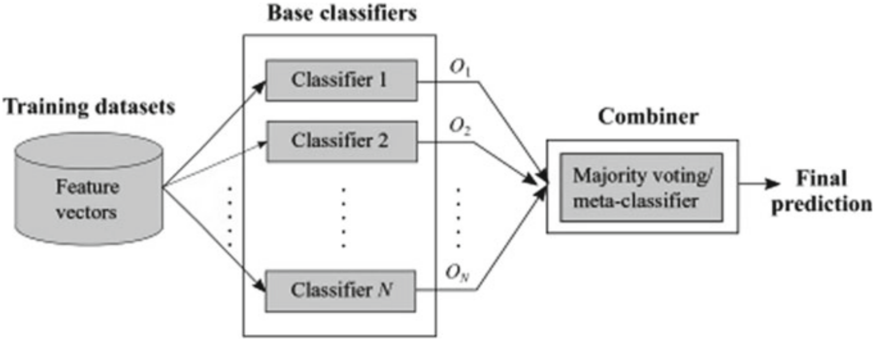

Fig. 1 Voting classifier

4.2 Soft Voting

For soft voting, the withdrawal phase is a prediction based on the proportion of opportunities given to that class.

$$\hat{y} = \arg\max \frac{1}{NClassifiers} \sum\nolimits_{Classifiers} (p1, p2, p3, \ldots pn) \tag{8}$$

This algorithm employs the input from the extracted features. Two classes are defined in this research work. The microarray cancer denotes that the person has probability of occurrence of microarray cancer. The normal is utilized for the person without any possibility of microarray cancer (Fig. 2).

5 Result and Discussion

The dataset is collected from the kaggle. The dataset is generally categorized into two labels. The anaconda is the tool of python which is used for the implementation of the

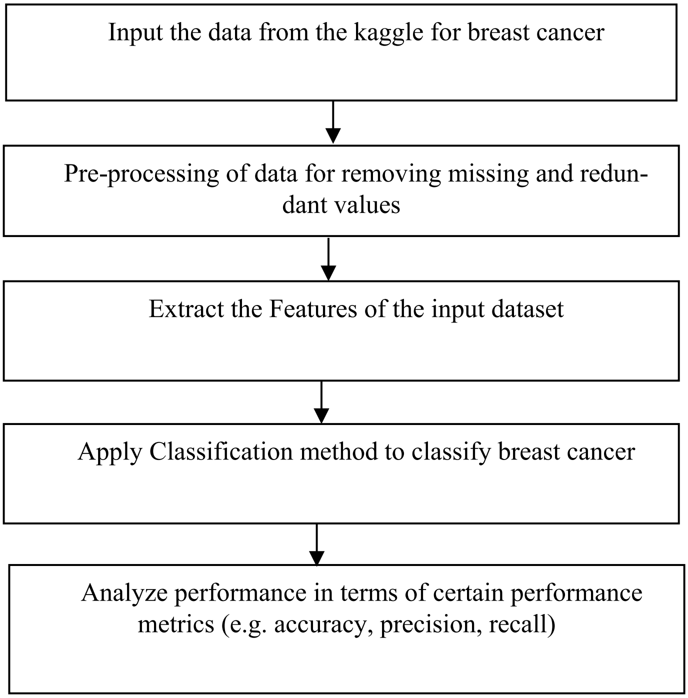
Input the data from the kaggle for breast cancer
Pre-processing of data for removing missing and redundant values
Extract the Features of the input dataset
Apply Classification method to classify breast cancer
Analyze performance in terms of certain performance metrics (e.g. accuracy, precision, recall)

Fig. 2 Proposed methodology

proposed methodology. The various performance analysis parameters like accuracy, precision and recall.

Various presentation metrics are used to calculate the system performance of the classification system (existing and proposed system). These are described below:

$$Accuracy = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \tag{9}$$

The precision main motive is to calculate the true +ive (TP) units relative to false +ive (FP) units.

$$Precision = \frac{Tp}{Tp + Fp} \tag{10}$$

The main motive of recall is to calculate true +ive (TP) units in relation to false –ive (FN) units that are not at all classes. The arithmetic performance or expression form of recall parameter is declared in:

$$Recall = \frac{Tp}{Tp + Fn} \tag{11}$$

As shown in Fig. 3, the performance of random forest is compared with voting classification for the breast cancer prediction. It is analyzed that accuracy of voting classification is 96%, precision value is 76% and recall value is 78% which improve performance up to 5% (Table 1).

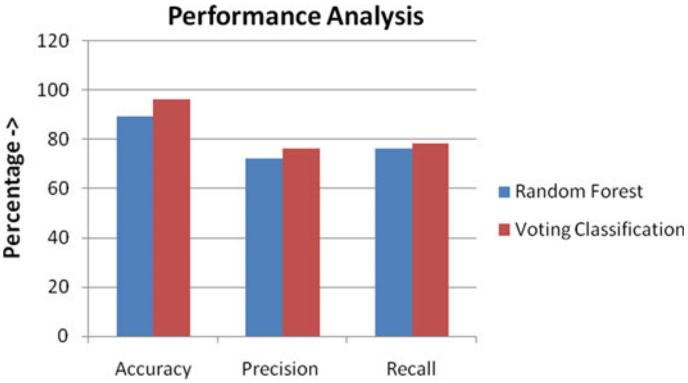

Fig. 3 Performance analysis

Table 1 Performance analysis

Performance parameter	Random forest (%)	Voting classification (%)
Accuracy	89	96
Precision	72	76
Recall	76	78

6 Conclusion

The survival rate of breast cancer has expanded, and a significant fall in the number of deaths due to these diseases has been noticed in the last few years. There are several factors behind this. One of the most considerable factors is the detection of this disorder in the early stage. The timely detection of this disease contributes significantly in the disease healing. It also offers deep insight of the disorder which further leads to the conclusion. The breast cancer detection has various phases which include preprocessing, segmentation, feature extraction and classification. The threshold-based strategy is applied for the segmentation. The textural feature extraction algorithm called GLCM is applied for the feature extraction and in the last voting classification method is applied for classifying the cancer of breast. The performance of the developed model is compared with standard models in the context of accuracy, precision and recall. It is been analyzed that results are optimized up to 7% for the breast cancer detection.

References

1. Sahni P, Mittal N (2019) Breast cancer detection using image processing techniques. In: Advances in interdisciplinary engineering, pp 813–823
2. Chtihrakkannan R, Kavitha P, Mangayar karasi T, Karthikeyan R (2019) Breast cancer detection using machine learning. Int J Innov Technol Explor Eng (IJITEE) 8(11). ISSN: 2278-3075
3. Patel VK, Uvaid S, Suthar AC (2012) Mammogram of breast cancer detection based using image enhancement algorithm. Int J Eng Technol Adv Eng 2(8):143–147
4. Yadav BK, Panse MS (2018) Virtual instrumentation based breast cancer detection and classification using image-processing. Int J Res Sci Innov (IJRSI) 5(4):25–31
5. Kumar R, Srivastava R, Srivastava S (2015) Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features. J Med Eng 1–14
6. Ramani R, Suthanthiravanithy S, Valarmathy S (2012) A survey of current image segmentation techniques for detection of breast cancer. Int J Eng Res Appl (IJERA) 2(5):1124–1129
7. Radha M, Adaekalavan S (2016) Mammogram of breast cancer detection based using image enhancement algorithm. Int J Adv Res Comput Commun Eng 5(7):218–221
8. Wang Y, Wang N, Xu M, Yu J, Qin C, Luo X, Yang X, Wang T, Li A, Ni D (2020) Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound. IEEE Trans Med Imaging 866–876
9. Khasana U, Sigit R, Yuniarti H (2020) Segmentation of breast using ultrasound image for detection breast cancer. In: International electronics symposium (IES)

10. Kavya N, Usha N, Sriraam N, Sharath D, Ravi P (2018) Breast cancer detection using non-invasive imaging and cyber physical system. In: 3rd international conference on circuits, control, communication and computing (I4C), pp 1–4
11. Varma C, Sawant O (2018) An alternative approach to detect breast cancer using digital image processing techniques. In: International conference on communication and signal processing (ICCSP), pp 0134–0137
12. Soliman OO, Sweilam NH, Shawky DM (2018) Automatic breast cancer detection using digital thermal images. In: 9th Cairo international biomedical engineering conference (CIBEC), pp 110–113
13. Liu H, Shang X, Ye X (2018) Breast cancer detection using synthetic aperture radar imaging and distorted born iterative method. In: International applied computational electromagnetics society symposium - China (ACES), pp 1–2
14. Razavi S, Hatipoğlu G, Yalçın H (2017) Automatically diagnosing HER2 amplification status for breast cancer patients using large FISH images. In: 25th signal processing and communications applications conference (SIU), pp 1–4
15. Sangeetha R, Murthy K (2017) A novel approach for detection of breast cancer at an early stage by identification of breast asymmetry and microcalcification cancer cells using digital image processing techniques. In: 2nd international conference for convergence in technology (I2CT), pp 593–596
16. Yin T, Ali FH, Reyes-Aldasoro CC (2015) A robust and artifact resistant algorithm of ultrawideband imaging system for breast cancer detection. IEEE Trans Biomed Eng 62(6):1514–1525