BREAST CANCER DETECTION USING PRE-PROCESSED IMAGES

Smt.S J R K Padminivalli V[1], Valluru Komali[2], Nelluri Naga Sai Krishna[3] and Kunchala Sumanth[4]

[1]Assistant Professor, Department of Computer Science and Engineering, R.V.R. & J.C. COLLEGE OF ENGINEERING, GUNTUR, Andhra Pradesh-522019, srivallivasantham @ rvrjc.ac.in,
[2]B.Tech Final Year, Department of Computer Science and Engineering, R.V.R. & J.C. COLLEGE OF ENGINEERING, GUNTUR,Andhra Pradesh-522019, k4669175@gmail.com
[3]B.Tech Final Year Department of Computer Science and Engineering R.V.R. & J.C. COLLEGE OF ENGINEERING GUNTUR,Andhra Pradesh-522019 nelluri421@gmail.com
[4]B.Tech Final Year Department of Computer Science and Engineering R.V.R. & J.C. COLLEGE OF ENGINEERING GUNTUR,Andhra Pradesh-522019 sumanth.kunchala@gmail.com
[†]These authors contributed equally to this work.

Abstract

One of the leading causes of cancer death in women is breast cancer. It occurs when cells in the breast begin to grow uncontrollably, often forming a lump or mass. Although it can happen to both men and women, it affects women more frequently. There are many Machine Learning methods available to detect this cancer but they provide less accuracy and this can potentially be resolved by implementing effective image pre-processing techniques like background removal, noise reduction and image enhancements. The pre-processed images are then sent to models like Convolutional Neural Network(CNN),Decision Tree and K-Nearest Neighbor and the results are combined using various techniques to provide better accuracy.

Keywords: Pre-Processing;CNN;KNN;Decision Tree

Introduction

Breast cancer is the second most common type of cancer among women worldwide, with an estimated 2.3 million new cases diagnosed in 2020 alone. Breast cancer can occur at any age, but it is most common in women over the age of 50.There are several types of breast cancer, including ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC).Risk factors for breast cancer include being female, older age, a personal or family history of breast cancer, certain gene mutations (such as BRCA1 and BRCA2), dense breast tissue, and exposure to estrogen. Hence it is important to increase awareness about breast cancer and encourage women to have a regular checkup and save their lives.

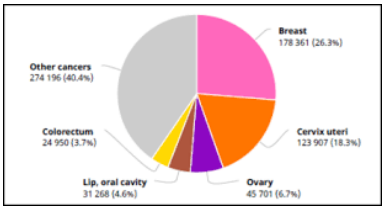

Figure 1 Cancer Statistics

Mammography is considered the gold standard for regular screening. This screened data are to be analyzed which requires radiologists but, the shortage of radiologists leads to delay in treatment. Hence ,it is important to develop an intelligent system that can detect and diagnose abnormalities quickly and accurately.Before developing an intelligent system, it is important to effectively pre-process mammographic images. This involves removing the background, pectoral muscle, and the addition of noise along with the application of image enhancements.

Related Work

Detecting cancer from mammographic images is a difficult, hence important features from these images should be identified. Several approaches exist to detect and segment masses in mammographic images[1] automatically, and the key points and main differences between those strategies are highlighted. The main objective is to highlight the advantages and disadvantages of the different approaches.T his review provides a quantitative comparison in addition to the qualitative description and comparison of different approaches. Two mammographic databases are compared to compare the performance of seven mass detection methods. One is a public digitised database, while the other is a local full-field digital database for local detection. A Receiver Operating Characteristic (ROC) and a Free-response Receiver Operating Characteristic (FROC) analysis is presented.

Since the American Cancer Society (ACS) issued recommendations for early detection of breast cancer in 2003, new information regarding breast magnetic resonance imaging (MRI)[5] screening has become available. A guideline panel evaluated the study and made new recommendations for women at various risk levels. A screening MRI is recommended for women with a lifetime risk of breast cancer greater than or equal to 20-25% of hers. This includes women with a significant family history of breast cancer and women who have been treated for Hodgkin's disease. There is insufficient evidence to support screening in women with a history of breast cancer, women with cancer in situ, women with atypical hyperplasia, and women with very dense breasts on mammography. risk subgroup. The review was assumed to be out of scope including the diagnostic use of MRI.

Radiologists prefer to employ breast ultrasound and mam-

mography imaging modalities to visualize breast cancer. A area of interest (ROI)[11] indicating the tumor is extracted from the image in order to find malignancy. When noise, low contrast, and blurriness are present, segmentation becomes laborious. Before to segmentation, pre-processing is done to improve contrast and remove extraneous information from the image. The categorization of the picture into benign and malignant classifications is also influenced by segmentation. The literature has suggested a number of segmentation approaches to separate the pectoral muscles from the microcalcification region of interest, masses, and breast lesions. This paper offers a thorough analysis of various methods, especially as they pertain to mammography pictures.

Machine Learning algorithms are also applied on datasets of UCI repository for breast cancer detection[3],the work introduced in this paper objectives to analyse the overall performance of a couple of computing device gaining knowledge of classifiers in detecting breast cancer. The acquired consequences genuinely point out the efficacy of linear guide vector classifier and gradient boosting over different classifiers. The findings of this lookup can be similarly utilized for creating extra environment friendly ensemble fashions as properly as optimizing the overall performance of present fashions thereby growing their prediction accuracy.

Later, Computer Aided Diagnosis(CAD) were used to assist radiologists in their work by carrying out a double-reading procedure that offers a second opinion that the doctor might consider during the detection phase.[12] study presents a computer-aided design (CAD) model for the detection of suspicious regions that are later classified as benign or malignant based on a set of features extracted from lesions to describe their visual content by support vector machines, artificial neural networks, and linear discriminant analysis. To identify the subset of features with the highest discriminant power, a genetic algorithm is applied.

Early breast cancer identification and other abnormalities in human breast tissue are now possible thanks to digital mammography. It gives us the chance to create algorithms for computer-aided detection (CAD).3 seperate steps were suggested in [8]. The first stage entails improving the contrast using the contrast limited adaptive histogram equalization (CLAHE) method. After that, create the rectangle to separate the pectoral muscle from the region of interest (ROI), and then use our suggested modified seeded region growing (SRG) technique to suppress the pectoral muscle. The 322 mammography pictures in the MIAS database were all subjected to the proposed algorithms, which resulted in total pectoral muscle suppression in the majority of the scans. As compared to previous segmentation approaches, the suggested algorithm produces better results.

One of the illnesses that causes a significant number of fatalities each year is breast cancer. In the entire world, it is the most prevalent type of cancer and the leading cause of mortality for women. Particularly in the medical industry, where those techniques are frequently utilized to make judgments through diagnosis and analysis. [6] compares the performance of four machine learning algorithms on the datasets: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbors (k-NN). The major goal is to evaluate each algorithm's efficiency and efficacy in terms of accuracy, precision, sensitivity, and specificity in order to determine whether or not the data classification was right. According to experimental findings, SVM provides the highest accuracy and lowest error rate.

Dataset

Mammography is a medical imaging technique that uses low-dose X-rays to examine breast tissue for signs of abnormalities or cancer. It is the most common screening tool used for breast cancer detection. The American Cancer Society recommends that women with an average risk of breast cancer begin annual mammograms at age 45, and then switch to screening every other year at age 55. Women at higher risk of breast cancer may need to start screening earlier and/or have more frequent mammograms. Mammogram images from "cbis-ddsm-breast-cancer-image-dataset" are used to implement the proposed methods. This CBIS-DDSM is a modified and standardized version of DDSM (Digital Database for Screening Mammography). The image is a jpeg of the original dataset (163GB). Original resolution was maintained.The dataset contains 38 attributes like filepath,image-path,BitsAllocated,BitsStored.The dataset contains about 10239 images.It contains normal, benign, and malignant cases with verified pathology information.

Figure 2 Dataset Description

METHODOLOGY

Image Pre-Processing

Image preprocessing is an essential step in image analysis and computer vision tasks. Preprocessing techniques like noise reduction, image enhancement, image segmentation, normalization are applied on the images. Noise is an unwanted and random variation in the pixel values of an image, which can be introduced during image acquisition or transmission. Noise reduction is applied to improve visual quality, enhance image analysis, increase signal-to-noise ratio improving compression. Image enhancement techniques can be used to improve the contrast, image sharpening, making the image clearer and easier to analyze. Normalization is a crucial step in image preprocessing that involves transforming the pixel values of an image to

a common scale. It can be used for improved training, consistent performance ,improve visualization, reduce computational complexity.

Convolutional Neural Networks

CNNs[14] are a type of deep learning model that can automatically learn to recognize complex patterns and features in images.Breast cancer detection typically involves analyzing medical images, such as mammograms or MRI scans. CNNs[15],[10] can be trained on large datasets of labeled medical images to accurately identify potential cancerous regions and distinguish them from normal tissue.One approach is to use a CNN as a binary classifier, where the model is trained to predict whether an image contains cancerous tissue or not. Another approach is to use a CNN as a segmentation model, where the model is trained to identify the exact location and extent of cancerous regions within an image, here we use it as a classifier model.Several studies have shown that CNNs can achieve high accuracy in breast cancer detection and diagnosis.CNN contains convolution layer,max pool layer and dense layers.For breast cancer detection 4 convolutional layers are used with input pattern of (50,50,3) and filters of size 32,64,128 are used.Activation functions[4] like softmax[9],relu[7] are used.

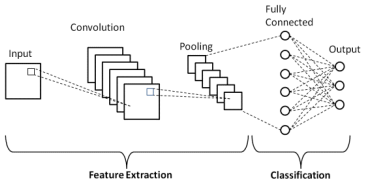

Figure 3 Architecture Of CNN

KNN

K-nearest neighbors (KNN)[13] algorithm is a simple, yet effective method for classification tasks that works well in many situations, including breast cancer classification. KNN is a non-parametric algorithm that does not make any assumptions about the underlying distribution of the data. Instead, it relies on the similarity between the new instance (in this case, a breast cancer case) and the existing labeled data to make a prediction.

In the case of breast cancer classification, the KNN algorithm can be used to classify a new case as malignant or benign based on its similarity to previously diagnosed cases. Specifically, KNN calculates the distance between the features of the new case and the features of the training data. It then selects the k-nearest neighbors based on the smallest distances and assigns the class label of the majority of those neighbors to the new case.

One advantage of using KNN for breast cancer classification is that it does not require a priori knowledge of the underlying probability distribution or parameter estimation. Additionally, KNN can handle large datasets and high-dimensional feature spaces, making it suitable for complex classification problems like breast cancer diagnosis.

Decision Tree

Decision trees[2] are commonly used for classification problems, including breast cancer classification. Decision trees are easy to interpret, as they can be represented graphically, with the decision nodes and branches clearly indicating the decision-making process. This is especially important in medical diagnosis, where it is crucial to understand the reasoning behind the decision. Decision trees can handle mixed data types, such as numerical and categorical data, which is common in medical data. Decision trees can handle missing data by imputing the missing values or creating a separate branch for missing values, which is important in medical data where missing data is common. Decision trees can capture complex relationships between features, which is useful in medical diagnosis, where multiple factors can influence the diagnosis.

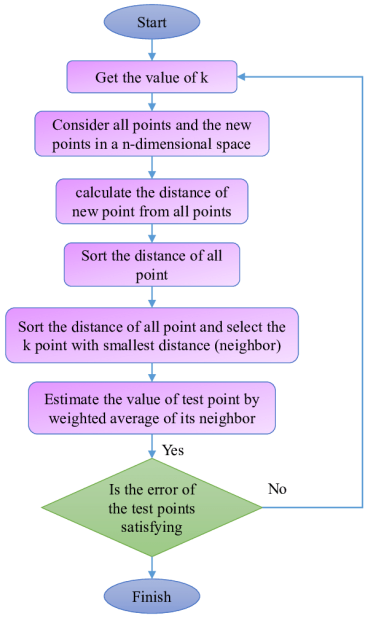

Figure 4 Flowchart Of KNN algorithm

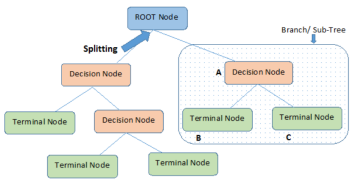

Figure 5 Decision Tree Classification

Fusion Of Results

In general, each classifier method, even the same method with different parameters, focuses on different perspectives and emphasizes certain aspects, and each learning algorithm has its own strengths and weaknesses. Therefore, it is difficult to assess which learning algorithms lead to better performance across all feature sets. The same dataset may be suitable for different algorithms to learn effectively. At the measurement level, there are several methods of fusion. Direct join rules are called votes and simply count the votes for each class and return the class with the most votes. Common algebraic combination rules include mean, maximum, and product.
let, $y \in \{c_1, c_2, \ldots, c_n\}$, m classifier $\{h_1, h_2, \ldots, h_m\}$ are trained in a feature space $D : \{x_1, x_2, \ldots, x_k\}$, the combination rules are defined as follows.

$$P_{\text{mean}}\left(c_i \mid x_1, x_2, \ldots, x_k\right) = \frac{1}{m}\sum_{j=1}^{m} P_{h_j}\left(c_i \mid x_1, x_2, \ldots, x_k\right)$$

$$P_{\max}\left(c_i \mid x_1, x_2, \ldots, x_k\right) = \max_{j=1}^{m} P_{h_j}\left(c_i \mid x_1, x_2, \ldots, x_k\right)$$

where P_{h_j} is the posterior probability of the classifier h_j, P_{mean} is the posterior probability after averaging fusion for m classifiers, $P_{\max}$ is the posterior probability after maximizing for m classifiers, P_{product} is the posterior probability after production for m classifiers.

Firstly, we can obtain the posterior probabilities P_{h_j} of the classifier h_j from the classification results.Once you have two or more sets of classification results, choose an appropriate fusion strategy such as matching, importance, or maximization. Then we can get the predicted posterior probabilities according to the rules of the fusion strategy given in the above equation.

Using simple algebraic rules for classifier fusion can reduce the risk of posterior probability estimation errors and overfitting. Moreover, these algebraic fusion rules have low computational complexity.

Analysis Of Results

Initially, a decision tree classifier is built with entropy as the criterion for splitting, a KNN classifier is built by choosing the 3 nearest values, and a CNN classifier is built with specified parameters. After building these classifiers, the models are trained with the same data to find the accuracy of each model. According to Table 1, each model is accurate to a different degree. It is observed that the accuracy of CNN is higher than the other 2 models.

TABLE 1. The performances of DECISION TREE,KNN and CNN.

Classifier	Accuracy
CNN	0.95
KNN	0.90
DECISION TREE	0.85

Fusing results from different classifiers, also known as ensemble learning or model fusion, is a technique used in machine learning and data mining to combine the predictions of multiple classifiers to improve overall prediction accuracy and robustness. Ensemble learning has been shown to improve prediction accuracy compared to individual classifiers. By combining the predictions of multiple classifiers, the ensemble can benefit from the strengths of different classifiers while mitigating their weaknesses. As a result, predictions that are more accurate and reliable can be made. Table 2 shows the result of the model after combining the results of three models. It is quite evident that the accuracy has increased.

TABLE 2. The final fusion results for DECISION TREE,KNN and CNN.

Fusion Method	Accuracy
mean	**0.97**
max	0.96

Conclusion

This paper presented the principles of three types of classification methods, KNN, Decision Tree, and CNN, and applied them to breast cancer detection, then improved the performance by fusing two perspectives: the fusion of the classification results of KNN, Decision Tree, and CNN using mean and maximum. The results showed that fusion processing is a viable way to improve the performance of breast cancer detection.

In conclusion, fusing results from different classifiers through ensemble learning can lead to improved prediction accuracy, increased robustness, better handling of model uncertainty, increased diversity, better generalization, and increased model stability. Ensemble learning is a powerful technique that can enhance the performance and reliability of prediction models in various machine learning and data mining tasks.

REFERENCES

[1]."A review of automatic mass detection and segmentation in mammographic images" by Arnau Oliver, Jordi Freixenet , Joan Martí , Elsa Pérez , Josep Pont , Erika R.E. Denton , Reyer Zwiggelaar.

[2]. B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in Proc. 1st Workshop Abusive Lang. Online, 2017, pp. 85–90.

[3]."Breast Cancer Detection Using Supervised Machine Learning: A Comparative Analysis" by Akansha Kamboj, Prashmit Tanay, Akash Sinha & Prabhat Kumar.

[4]. C. Gulcehre, M. Moczulski, M. Denil and Y. Bengio, "Noisy activation functions", Proceedings of International Conference on Machine Learning, pp. 3059-3068, 2016.

[5]. D. Saslow, C. Boetes, W. Burke, S. Harms, M. O. Leach, C. D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener, R. A. Smith, E. Warner, M. Yaffe, K. S. Andrews, and C. A. Russell, "American cancer society guidelines for breast screening with MRI as an adjunct to mammography," CA A, Cancer J. Clinicians, vol. 57, no. 2, pp. 75–89, Mar. 2007.

[6].H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Pro-cedia Comput. Sci., vol. 83, pp. 1064–1069, 2016.

[7]. H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2684-2691, doi: 10.1109/IJCNN.2017.7966185.

[8]. I. K. Maitra, S. Nag, and S. K. Bandyopadhyay, "Technique for preprocessing of digital mammogram," Comput. Methods Programs Biomed., vol. 107, no. 2, pp. 175–188, Aug. 2012.

[9]. I. Kouretas and V. Paliouras, "Simplified Hardware Implementation of the Softmax Activation Function," 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 2019, pp. 1-4, doi: 10.1109/MOCAST.2019.8741677.

[10]. Iztok A. Pilih, Dunja Mladenic, Nada Lavrae and Tine S. Prevec3B. Smith, "Using Machine Learning for Outcome Prediction of Patients with Severe Head Injury", Tenth IEEE Symposium on Computer-Based Medical Systems.

[11]. J. Dabass, S. Arora, R. Vig, and M. Hanmandlu, "Segmentation techniques for breast cancer imaging modalities—A review," in Proc. 9th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence), Jan. 2019, pp. 658–66

[12].J. Suckling, J. Parker, and D. Dance, “The mammographic image analysis society digital mammogram database excerpta medica,” in Proc. Int. Congr. Ser., vol. 1069, 1994, pp. 375–378.
[13]. M. Sharma and S. Sharma, "Generalized K-Nearest Neighbour Algorithm- A Predicting Tool", Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 11, Nov. 2013.
[14].S. Pang, A. Du, M. A. Orgun and Z. Yu, "A novel fused convolutional neural network for biomedical image classification", Med. Biol. Eng. Comput., vol. 57, no. 1, pp. 107-121, 2018.
[15]. Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. EMNLP, Oct. 2014, pp. 1746–1751.