Review on deep learning and machine learning techniques to support early detection of breast cancer

[1]Melwin D Souza, [2]Ananth Prabhu G
[1]HOD and Senior Assistant Professor, [2]Professor
[1]Department of Computer Science and Engineering, [2]Department of Computer Science and Engineering
[1]Moodlakatte Institute of Technology, Kundapura, India, [2]Sahyadri College of Engineering and Management, Mangaluru, India

Abstract: Breast cancer is a global disease and nowadays the number of breast cancer cases is gradually increasing ageing as well as younger age women. Though the majority of underlying causes and other features are usually uniform around the world, every region has its own uniqueness for that cancer. In this paper, review on different machine learning and deep learning based techniques applied on different datasets for segmentation and classification to identify benign or malignant feature of image. The objective of this paper is to discuss and analyze different dataset using various algorithms and summarize the performance by computing accuracy level on detection of breast cancer. Study on analysis of algorithms leads us propose a new research work in the early stage of breast cancer detection

***Keywords*- CNN, k-NN, MIAS, segmentation, Micro calcifications, deep learning, VGG-16, ResNet**

I. INTRODUCTION

The breast cancer is the most universally occurring type of cancer and it is known to distress over two million women every year. This cancer is one of the leading causes of death for women in the worldwide. The rate of death and occurrences of breast cancer increases with age of patient. As per the report of World Health Organization (WHO), breast cancer cases are going to raise more than 19 million in 2025. Survey conducted by cancerindia.org published that every 8 minutes results one death of woman because of cervical cancer in India. According to [2] in 2018, over 0.6 million fatalities were caused by breast cancer. The number is approximately 15% of the total deaths resulting from all types of cancer among women. Cancer is a disease instigated by the deviations arisen in cells spread uncontrollably. These cells develops tumour and named after the part of body in which it originates. Breast cancer usually shows no pain in its starting stage and it can be easily treated, that's why screening is important for early detection

During the primary phases of the cancer, the decease signs are not presented well and hence diagnosis is delayed. It is recommended by the NBCF (National Breast Cancer Foundation) that women with the age of more than forty years should get a mammogram once a year and mammogram is an X-ray of the breast. It is a medical technique used for the detection breast cancer with no side effects believing the method as safe. The magnetic resonance imaging (MRI) is the smartest alternative to mammogram and test is done when the radiologists want to confirm about the existence of the tumour. There are many deep learning and machine learning techniques are available for cancer detection and prediction. Some of most used techniques are for breast cancer diagnosis are support vector machines (SVM), artificial neural networks (ANN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN).

In recent years, several researchers studied and proposed methods for breast mass classification in mammography images. Sharkas et. al used the discrete wavelet transform (DWT), the contour let transform, and the principal component analysis (PCA) methods for feature extraction [3]. The system was able to detect and classify normal and abnormal tissues. Additionally, it classified benign and malignant MC tumors. The achieved rate was almost 98%. Ragab et.al proposed the DWT as a feature extraction technique to detect mass abnormalities in the breast [4]. In addition, an evaluation between support vector machines (SVM) and artificial neural networks (ANN) for classifying normal, abnormal tissues, benign and malignant MCs tumors was introduced. The achieved detection rate was 96% for ANN and 98% for SVM [4]. This paper proposed to work on DITI technique which is widely adopted by the medical specialists to record and analyze the breast malignancy due to its risk-free as well as contactless nature [6].

II. BREAST ABNORMALITIES

Three broad categories of abnormalities that affecting the breast tissues: *Masses/Lumps, Micro calcifications, Architectural distortions.* Knowledge of these abnormalities motivates to get better results in the proposed work.

Masses/Lumps. Confined swelling, protuberance, bulge or bump grows in the breast and which are distinguished from the breast tissue around the area. They are characterized by their breast masses into benign or malignant mass based on the characteristics such as shape, outline, and density [11, 12].

Micro calcifications. Small calcium accumulation in the breast tissues results breast micro calcifications and they are separated into benign, suspicious or high possibility of malignancy [13].

Architectural distortions. The normal breast architecture is slanted without any allied mass and which appears like arrangement of unusual tissues radiating from a point, focal retraction or somewhat random pattern. These distortions may be the central opacity or solid Centre and also results other abnormalities intra-mammary lymph node and asymmetric tubular structure [14].

III. DATA SET DESCRIPTIONS AND PERFORMANCE METRICS

WDBC Processing. [1] Sharma *et. al* proposed a system which is used a Wisconsin Diagnosis Breast Cancer data set which has been acquired from 'UCI ML' repo which has 569 instances and 32 attributes. The resulted variable is either benign (357 observations) or malignant (212 observations). The k-fold cross-validation is utilized in which the presented data is divided into k equally sized bits. System covered performance metrics which includes the parameters such as accuracy, recall, precision, FI score (It is the weighted average of Precision and Recall). Table 3 analyzed by Q. Zhang *et. al* illustrates performance evaluation between different models considering above parameters and IRMA dataset. Model descriptions discussed in later sections[7].

Table 3: Comparison of performance between different methods of classification

	VGG16	ResNet 50	Q. Zhang et. al
Precision	89%	88%	82%
Recall	99%	94%	86 %
Accuracy	94%	91.7%	83.2%

MIAS Processing. Naresh Khuriwal and Nidhi Mishra proposed a system to detect breast cancer from Histopathological images using deep learning [2]. They made use of data set of Mammographic Image Analysis Society (**MIAS**) database which is publicly available to download for research purpose. This data set includes 200 histopathology images. System used convolutional neural network for diagnosis breast cancer and includes two main parts: predict models and Pre- processing data. The final stage that is classification covers deep learning neural network algorithm with four-layer convolutional model on dataset of MIAS to work on a human biological method. First input layers accept 12 neurons and 8 neurons transferred to hidden layers and finally output layer results 1 neuron. Also additional functional units are added in the processing before getting final output.

Table 4. CNN Model Summary

Layer Name	Type	Output Shape	Parameters
Convolutional Layer	Dense	12	496
Hidden Layer	Dense	8	136
Fully Connected Layer	Dense	1	9
Total	params:	641	
Trainable	params:	641	
Non-trainable	params:	0	

Mammographic Image Database for Automated Analysis (MIDAS). Images with DICOM format are accumulated from Janice Lamas Radiology Clinic. The database comprises of nearly 600 digital mammograms and each contains two images per breast, patent information, abnormality descriptions, breast composition, BIRAD categories, and overall impression [5].

3.1 Segmentation

Marker Controlled Watershed Segmentation. The transformation process of this algorithm work on images to find "Catchment basins" and "watershed ridge lines" and considers the surface where light pixels are high and dark pixels are low [17]. Figure 1 shows how watershed segmentation works efficiently when foreground object and background locations are marked or identified.

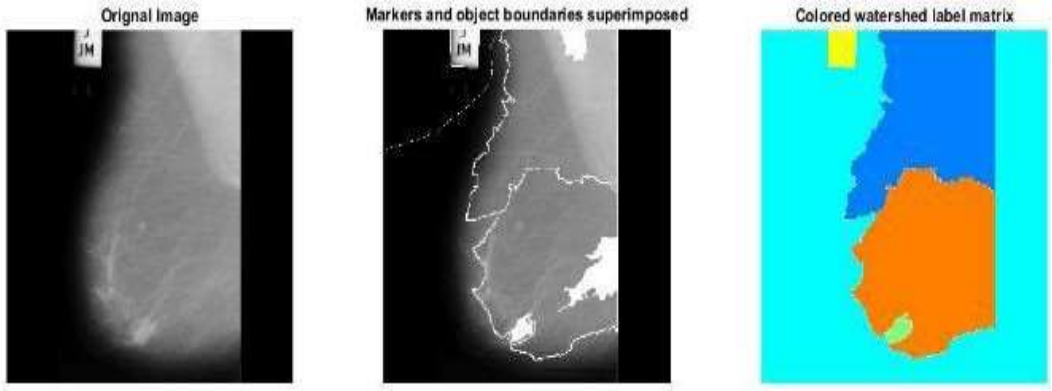

Figure 1: Marker Controlled Watershed Segmentation.

Texture Segmentation Using Texture Filters. [17] Use standard function, entropy, to create texture images. Function return an array value in which each output pixel comprises the entropy cost of the 9x9 pixel around the corresponding pixel in the input image. Median filters and histogram equalization method are used in second stage to improve image quality and 12 features are extracted to train the model.

IV. METHODS AND MODELS

4.1 Machine Learning and Deep Learning Algorithms

Random Forest. It is a *supervised learning* algorithm which forms a collective of decision trees and used to train the system. Each instance of iteration includes data set and a random sample with size N is chosen from the data set [1, 9].

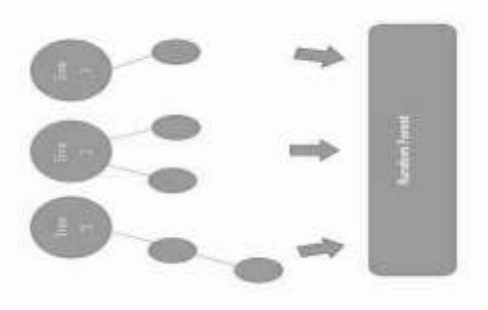

Figure 2: How Random Forest works

The Division of dataset into training and testing sets results 398 observations for training set and 171 observations for testing. The estimators count is fixed to 72 therefore it is confirmed that every single observation in process is predicted at least a few times. The confusion matrix of random forest shows the five misclassified observation which is five for benign and four for Malignant. Finally random forest results the accuracy equals 94.74%.

Table 1: The confusion matrix of random forest

		Predicted	
		Benign	**Malignant**
Actual	**Benign**	103	5
	Malignant	4	59

K-Nearest-Neighbor (k-NN). K may be viewed as the representation of the data points for training in closeness to the testing data point for determining the class. This algorithm belongs to supervised learning approach used for regression and classification. K-Nearest-Neighbor technique gathers all the data points for processing a new data point. Attributes which have a higher degree of deviation are key factors in finding the distance. In the Figure 3, in the view of N training vectors, k-NN algorithm calculates the k closest neighbors of regardless of labels [1, 8].

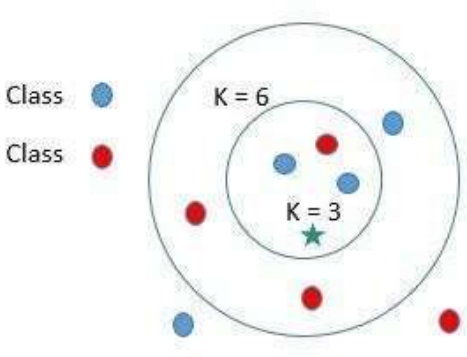

Figure 3: k-NN Illustration

The accuracy of k-NN is found to be 95.90%, there is only one observation that is misclassified as Benign and four observations are misclassified as Malignant as represented in Table 2. The results are comparatively better than Random Forest algorithm. Table 2 represents one observation that is misclassified as Benign and four observations are misclassified as Malignant. K-Nearest-Neighbor results accuracy 95.90% which shows better performance than Random Forest algorithm.

Table 2: k-NN Confusion Matrix

		Predicted	
		Benign	**Malignant**
Actual	**Benign**	107	1
	Malignant	6	57

4.2 Neural Network and Analysis

[2] Figure 4 shows the fundamental activity of neural network. If perceptron bias b = threshold then perceptron rule rewritten. In ongoing process the fired neurons afterward fixed position is called threshold condition. Convolutional neural network is preferable to use for dataset classification to get better accuracy.

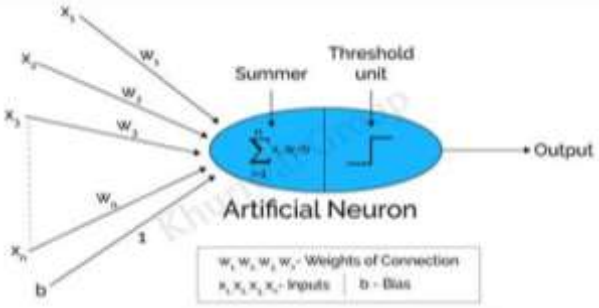

Figure 4: Neural Network Functionality

4.3 Convolutional neuron Networks

Convolution neural network (CNN) functions as Multilayer perceptron (MLP) and both are made up of neurons which have biases and learnable weights. Three layers of CNN are arranged at different levels as shown in Figure 5 [10, 15].

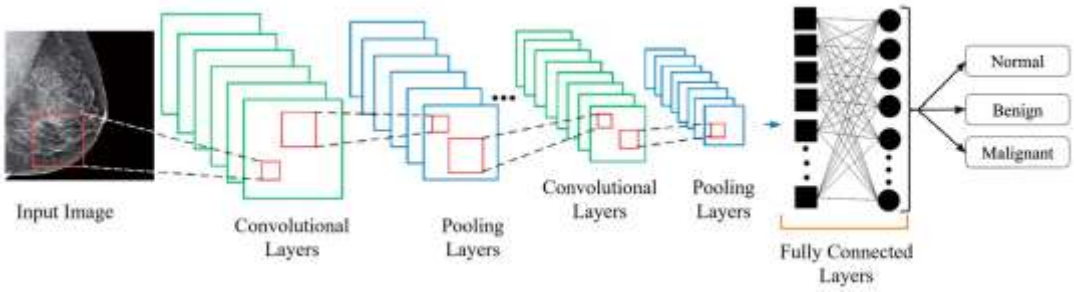

Figure 5: Typical CNN architecture

Convolution Layer: Consist of independent set filters that are convolved individually inside the image and resultant in feature maps for each of the filters.

Subsampling/Pooling Layer: Purpose of this layer is to progressively reduce the spatial size of the representation so that the number of computation is reduced Pooling layer operates independently on each feature map.

Fully Connected Layer: Connected at the end of CNN after several convolutional and max pooling layer. Here, all the neurons are connected to all activations in the previous layer.

Region based Convolution Neural Network (RCNN). R-CNN works for image segmentation where image is accepted as input and results bounding boxes and labels for each instance. In region proposal stage, bounding boxes are created by RCNN with the help of selective search and also adjacent pixels are grouped as texture, color, or intensity. The regions are wrapped to a standard square size and finally an SVM is added to classify at the end of the network.

Fast RCNN. It is faster objection detection technique. In Fast RCNN, the input image is used as input to a CNN to generate a convolutional feature map and the proposals region is identified. and is wrapped into squares by using an ROI pooling layer supports to wrap the regions into squares and a softmax layer predicts the class of the proposed region along with offset values for the bounding box. In Fast R-CNN one time convolution process is done to generate a feature map.

Faster RCNN. Faster RCNN eliminates selective search technique and make the network to learn the region proposals. In faster RCNN, CNN accepts the image as input and results a convolutional feature map. The predicted region proposals can be reformed by ROI pooling layer and this will be considered to classify the images inside the proposed region and predicts the offset value. This algorithm was able to provide a much faster execution than its predecessors.

You Only Look Once (YOLO). A single convolution neural network is used for the full image. This network will partition the image into several regions and attempt to predict bounding boxes. The predictions of each region are then computed and will be used to increment the bounding boxes.

4.4 CNN Models

Visual geometry group (VGG): VGG-16 network model is used to train on the ImageNet database that has large number of images. Deep 16 layers of model categorize 1000 instances by classifying images in the database. Learning characteristics of VGG16 supports to represent the feature over a wide range of images and accepts the standard image input size of 224 by 224 as shown in figure 10 [16].

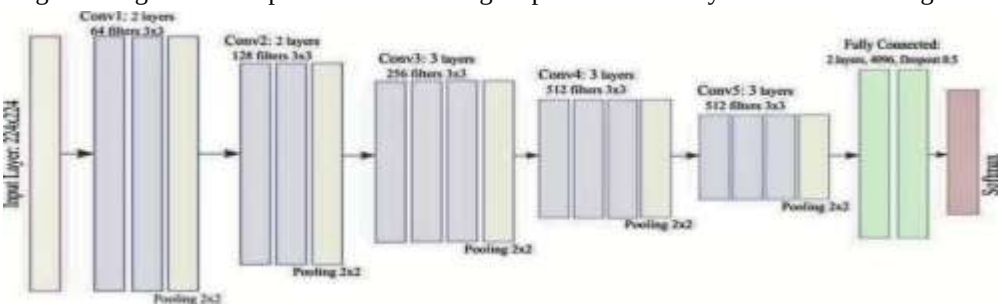

Figure 6: Architecture of VGG16

ResNet: ResNet is well known for its depth layering i.e. 152 layers, and residual blocks model. The residual model discourse the training problem because of deep architecture and introduction of identity skip connection. Such conception provides the layers to copy their input to the next layer [16]. Figure 7 shows the architecture of ResNet50.

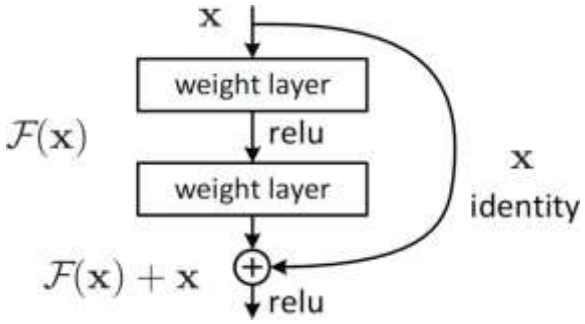

Figure 7: Architecture of ResNet50

V. RESULTS AND DISCUSSIONS

Analysis on the performance of the techniques covered in this paper provides the ideas to select the dataset and flow of processing to detecting the breast cancer by classifying benign or malignant tumor. This paper analyzed the results of research work published in different papers to work on proposed research work. Objective of Proposed work is to use deep learning based technique on thermal images to add more feature train the system. Flow diagram as shown in figure 8 covers all the steps in the breast cancer detection system.

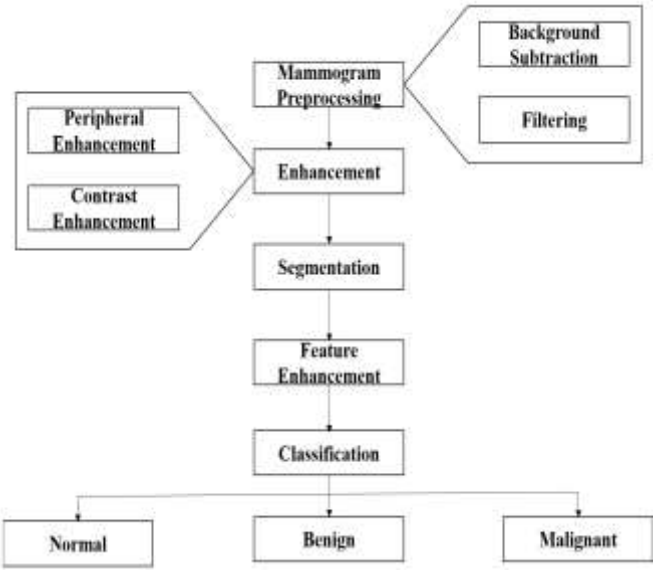

Figure 8: Basic steps CAD System for the breast abnormality

The CAD system implements combination of many image processing techniques as shown in figure 8. Artifacts and labels removed along with noise in the first phase by adopting filtering technique [18]. The second phase covers enhancement of contrast and edges in mammogram image and leads to perform segmentation. Next phase is feature extraction to calculate the region of interest (ROI). The final step is to classify the data images are classified into a normal, benign or malignant in the final phase.

Khuriwal et. al [17] achieved 98% accuracy in their research work on deep learning based breast cancer detection. Confusion matrix of generated using random forest and k-NN algorithms classifies the benign and malignant images with different accuracy level based on actual and predicted values. [8] Comparison of accuracy shows better results in k-nearest-neighbor than random forest (95.90%). According to Zhang Q et. al performance between VGG-16 and ResNet50 results the value of three parameters, accuracy, recall, and precision using IRMA dataset [7]. Comparison on resulted values of these parameters shows better result in VGG-16 than ResNet model.

VI. CONCLUSIONS

In this study various algorithms are discussed on different datasets to analyze the performance and to get more accurate results. Firstly we concentrated on collecting the details of breast cancer detection by processing various machine learning and deep learning models. Based Performance results deep learning based neural networks leads important stages to classify the normal and abnormal feature of breast cancer. The survey conducted by many researchers based on harmfulness and cost, concluded that Digital Infrared Thermal Image (DITI) is widely considered to record the breast malignancy during the screening process.

This paper proposes to work on breast cancer detection in the initial stage so lot of valuable life can be saved by reducing the rate. Also one system to be developed to which perform complete process by adopting features to train and to learn the things provide better results.

References

[1] S. Sharma, A. Aggarwal and T. Choudhury. 2018. "Breast Cancer Detection Using Machine Learning Algorithms," International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, pp. 114-118.

[2] WHO breast cancer statistics [Online]. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breast- cancer/en/

[3] M. Sharkas, M. Al-Sharkawy and D. A. Ragab. 2011. "Detection of Microcalcifications in Mammograms Using Support Vector Machine". UKSim 5th European Symposium on Computer Modeling and Simulation, Madrid, pp. 179-184.

[4] Ragab D, Sharkas M, Al-sharkawy M. 2013. "A comparison between support vector machine and artificial neural network for breast cancer detection 2 the cad system. 171–176.

[5] Fabiano Fernandes, Rodrigo Bonifácio, Lourdes Brasil, Renato Guadagnin and Janice Lamas (2012). MIDAS – Mammographic Image Database for Automated Analysis, Mammography-Recent Advances, Dr. Nachiko Uchiyama (Ed.), ISBN: 978-953-51-0285-4, InTech.

[6] V. Rajinikanth, N.S.M. Raja, S.C. Satapathy, N. Dey, and G.G. Devadhas, 2017 "Thermogram assisted detection and analysis of ductal carcinoma in situ (DCIS)," International conference on intelligent computing, instrumentation and control technologies (ICICICT), IEEE.

[7] Q. Zhang, I. U. Haq, A. Jadoon, A. Basit, and S. Butt, "Classification of mammograms for breast cancer detection based on curvelet transform and multi-layer perceptron.," vol. 28, no. 10, pp. 4311–4315, 2017

[**8]** Time complexity and optimality of kNN [Online] Available:https://nlp.stanford.edu/IR-book/html/htmledition/ time-complexityand book/html/htmledition/time-complexityand-Optimality-of-knn-1.html

[9] Gilles Louppe. 2015. "Understanding Random Forests from theory to practice". University of Liege. arXiv:1407.7502v3 [stat.ML]

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. arXiv:1409.4842

[11] http://www.birads.at.

[12] http://santemedecine.commentcamarche.net/contents/cancer/13_lecancer-du-sein.php3#les-statistiques-alarmantes-du-cancer-du-sein.

[13] Patricia Lorena Arancibia Hernández, Teresa Taub Estrada, Alejandra López Pizarro, María Lorena Díaz Cisternasy Carla Sáez Tapia. Breast calcifications: description and classification according to BI-RADS 5th Edition. BI-RADS Rev Chil Radiol 2016; 22(2): 80-91

[14] https://www.ajronline.org/doi/full/10.2214/AJR.12.10153

[15] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[16] A. Soudani and W. Barhoumi, 2019. "An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction" , *Expert Syst. Appl.*, vol. 118, pp. 400–410

[17] N. Khuriwal and N. Mishra, 2018, "Breast Cancer Detection from Histopathological Images Using Deep Learning", 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-4

[18] O. V. Singh and P. Choudhary, 2018, "A Study on Convolution Neural Network for Breast Cancer Detection," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 2019, pp. 1-7