A Survey on Accurate Breast Cancer Detection and Classification using Machine Learning Approach

D. Sandeep[1,*] and *G. N.* Beena Bethel[2]

[1]MTech Student, Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.
[2]Professor, Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.

Abstract- This survey paper is used to discuss about the detection of breast cancer tissues using different machine learning algorithms. Identification of cancers using scanned images are very important for correct diagnosis. Many algorithms are present for detection of cancer using image processing techniques, all these algorithms have the main goal of detecting those cell tissues. Each algorithm has their own assumptions and advantages, here is a review of some of those algorithms for breast cancer detection. This paper highlights the algorithms and their assumptions of the prior published papers.

1 Introduction

Breast cancers are one of the major health issues for women. Early detection of the cancer tissues can be useful for their diagnosis. Cancers are formed by excessive growth of cells in an uncontrollable manner. They can be two types of tumors benign and malignant. Benign tumors are harmless they do not spread, but malignant tumors are dangerous and these cells form together as a lump they can spread throughout the body if not treated in time. Breast cancer is due to those lumps which are formed in the breast of women. These can be detected by considering personal or family medical history, physical examination, mammograms or ultrasound scan or by biopsy etc. Various methods are used for the detection of breast cancers. Proper diagnosis can reduce the risk of death in the patient.

2 Materials and Methods

In related paper [1] Proposed convolutional neural networks for detection of benign or malignant tumors in breast. Mini-MIAS (Mini-mammographic image analysis society) dataset is used, these images are pre-processed, as a model for machine learning framework Tensor Flow library has been selected. In this CNN is used, the grayscale mammogram image is used as input layer, hidden layer consists of convolutional layer, ReLU (rectified linear unit) layer, pooling layer and fully connected layer are used. It also uses back propagation for update weight for latest their closet value, in this logit layer gives 3 possible types of outputs 0 for normal, 1 for benign, 2 for malignant. It had the accuracy of 82.7%.

In paper [2] Proposed mammographic images by using 2 main angels Craniocaudal (CC) view, Mediolateral-oblique (MLO) view. The pre-trained model VGG-16 network model is used which is proposed by oxford visual geometry group for the ILSVRC competition. Here MIAS (Mammographic image analysis society) and DDSM (Digital database for screening mammography) datasets are used, separate execution is done for different datasets. The VGG-16 model consists of 16 hidden layers which composed of 13 convolutional layers and 3 FC layers. The images from the dataset MIAS, DDSM are pre-processed then by using CNN a new model is trained using transfer learning VGG-16 network, it is A model that extracts features from the input mammograms, then uses these features to train the neural network classifier, and uses the pre-propagated VGG-16 model to detect abnormal areas through backpropagation, thereby updating the several final layers weights. Then after completion the results are obtained, in these the new model results are compared with feature model for MIAS dataset images the performance is increased to 0.88% and for DDSM new model had acquired high accuracy. It concludes that CC view is way better than MLO view where CC view has 0.931 accuracy and MLO view has 0.887 accuracy.

In paper [3] Proposed breast cancer diagnosis in which the entire algorithm has 2 parts one for identification of the cancer tissues and the other part is classification of the cancer tissue. The dataset is collected from the images of http://web.inf.ufpr.br/vr/breast-cancer-database then these images are pre-processed and analysed by using

* Corresponding author: sandeepdulam19@gmail.com

wavelet transform and the benefit features are extracted by using the result of wavelet transform is to obtain the maximum number of functions through standard division, and extract the functions from the pre-processed image. and diagnosis step is used to distinguish between malignant and benign tumors and the benign and malignant tumors are separated. There are two different types of benign (phyllodes and adenosis tumor) and two different types of Malignant (papillary carcinoma and ductal carcinoma). After feature extraction images are analysed using Grey level co-occurrence matrix (GLCM). The outputs are considered as input for fuzzy logic for identifying benign (phyllodes and adenosis tumor) and two different types of Malignant (papillary carcinoma and ductal carcinoma) tumors and the accuracy of 98%.

In paper [4] Proposed different data mining tools for breast cancer prediction and classification, in this WBCO (Wisconsin breast cancer original) dataset from UCI repository had been used the images are then pre-processed, then the classifiers are used which are Bayes classifier (Bayesian Logistic Regression, Naive Bayes), Decision Tree (simple CART, J48) are used for the pre-processed images which classify and analyse the benign and malignant tissues. Naive Bayes is fast, clear and simple classifier which considers attributes that are mutually independent, Bayesian logistic is for the problems which has two class values. Simple CART is a methodology which is widely used for prediction, J48 creates a decision node in the tree for guessing expected value of the class. WEKA tool is used for this process, and classification accuracy for the algorithms are acquired. Naïve Bayes had 95.2654%, Bayesian logistic Regression had 65.4232%, Simple CART had 98.1349%, and J48 had 97.274% accuracy.

In related paper [5] Proposed that identified of masses in the breasts using mammograms with adaption of breast density. In this the DDSM (Digital Database for Screening Mammography) dataset is used. The images are pre-processed and there are different stages proposed for tumor detection. At first the breast density is detected by using adaptive algorithm which is capable of analysing the image and telling if it is dense or non-dense, then a micro-genetic algorithm is used to create a texture proximity mask to select the regions which suspect of containing lesions which is done using segmentation. But in some cases there are excessive segmentation of suspect regions are formed even the healthy regions are marked as tumors this is called as false positive these regions are removed by using DBSCAN and a proximity ranking of textures extracted from ROI, and Local binary patterns (LBP) and SVM (support vector machines) classifiers are used and the tumor tissues are classified thus by the results obtained are Segmentation allowed 96.73% of the lumps to be separated in loose breasts, of which 2031 were not formed, while in dense breasts, 94.07% of lumps were separated, of which 1337 were not formed. In parameter evaluation (training), segmentation was able to separate 97.41% of the masses, but 9613 was not formed in the loose breasts, and 9413 was not formed in the masses. In dense breasts, there are 48% lumps, but there are no 9933 formations.

In related paper [6] Proposed different ML algorithms for cancer detection, here Original Wisconsin Breast Cancer Dataset that is obtained from the UCI Repository dataset is considered. The images in the data set are pre-processed, and then a machine learning algorithm such as a support vector machine (SVM) is used, which selects key patterns from all classes called support vectors and separates them, thereby generating linear functions that make them to a large extent Bayesian Networks (BN) are based on a recursive method to random forest (RF). In this method, each iteration involves selecting a random sample from a data set with replacement and another sample without replacement Select a random sample in the, and then split the resulting data are used for prediction and are used for various breast cancer attributes are considered by these machine learning algorithms and accuracy of the algorithms are calculated which shows as 97% and recall values, precision values and area under ROC (receiver operating characteristic) values are acquired.

In paper [7] proposed various machine learning algorithms which are trained to detect the breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The dataset is pre-processed and machine learning algorithms like GRU-SVM (gated recurrent unit- support vector machine) which is used for binary classification, linear regression is used as classifier which was done by applying threshold, multilayer perceptron consists of hidden layers that enable the approximation of functions, nearest neighbour is used for the optimization, SoftMax regression produces a probability distribution for the classes, support vector machine used as binary classification to determine optimal hyperplane for separating two classes in the dataset and the cancer tumors are identified. The GRU-SVM has the training accuracy of 90.68%, linear regressing has the training accuracy of 92.89%, multilayer perceptron has the training accuracy of 96.92%, SoftMax regression has the training accuracy of 97.36%, support vector machine has the training accuracy of 97.7% and nearest neighbour does not have recorded training accuracy because it does not require training.

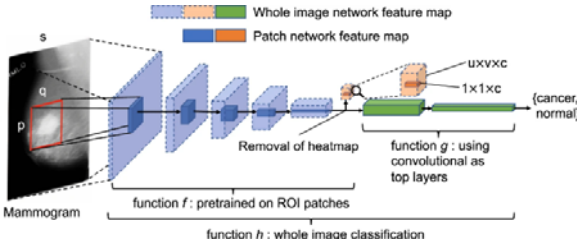

Fig.1. Breast Cancer detection using Convolutional neural network

In [8] authors have proposed detection of breast cancer an end-to-end training approach of deep learning process is proposed here. In this case, only training data

sets with complete clinical annotations or complete cancer status representations are used. A part of the breast mammography screening digital database (CBIS-DDSM) and the INbreast data set are used for cured breast images, and all annotations are only required at the initial stage of training, and only image-level titles used in subsequent steps can delete you When you use rarely available annotated lesions and the folding neural network detects cancerous tissue, you will gain reliance, thus for CBIS-DDSM the sensitivity is 86.1%, specificity is 80.1% and for full field digital mammography (FFDM) images for INbreast database the sensitivity is 86.7% and specificity is 96.1%.

In related paper [9] proposed several Data mining algorithms for early-stage breast cancer prediction. Here Wisconsin breast cancer (WBC) dataset and Breast cancer dataset are used. Use a sampling filter to pre-process the image and sample the data, and then remove missing values from the data set. Three Naive Bayes (NB) classification methods are used to estimate the probability of each class value to which a particular instance belongs. For this type, the J48 algorithm uses information entropy, and uses this information entropy to decompose each data attribute into smaller data sets to check the entropy difference. Minimal order optimization (SMO) replaces all missing values globally and converts the nominal attribute to binary, then detects breast cancer and reports the J48 result: 75.52% of breast cancer Data set and SMO: 96.99% of the WBC data set, then after applying pre-processing techniques accuracy is increased and gets a conclusion that SMO is better than J48 algorithm.

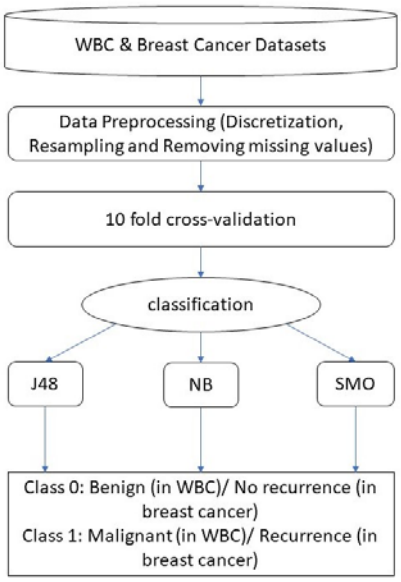

Fig. 2. Flow chart for paper [9]

In paper [10] Proposed machine learning techniques for breast cancer detection. Here Wisconsin Breast Cancer Diagnostic (WBCD) dataset is used. The data is pre-processed by standardizing it which is rescaling method that transforms features with Gaussian distribution, then five non-linear machine learning algorithms are used which are Multilayer Perceptron (MLP) which has three layers using a non-linear activation function, K-Nearest Neighbour (KNN) a new element is compared to other elements using similarity measurement and the distance is used as the weight of the neighbour, Classification and Regression Trees (CART) it is used to develop statistical model which deals with the data that is not fully finished, Gaussian Naive Bayes (NB) which is used when features have continuous values which is a used for classification, Support Vector Machine (SVM) which is used for separating two classes by determining the linear classifier then by showing the results such as Multilayer Perceptron (MLP) got the accuracy of 96.70%, K-Nearest Neighbour (KNN) got the accuracy of 96.27%, Classification and Regression Trees (CART) got the accuracy of 91.0%, Gaussian Naive Bayes (NB) got the accuracy of 93.62%, Support Vector Machine (SVM) got the accuracy of 96.42%.

In paper [11] proposed that breast cancer can be identified by using genetic algorithm. Here breast cancer dataset is considered from UCI which contains a Multi surface Method-tree (MSM-T) which uses linear programming. Then the data is pre-processed in which missing entries in dataset are filled by using the average values. Composed hybrid feature selection (CHFS) architecture is proposed which consists of information gain (IG) which gain measure gives the effect of the features and selects that are larger than the threshold, Gain ration (GR), different classifiers like J48, Naive Bayes and JRIP are used for comparison the accuracy is compared before CHFS is applied and the results with after application of CHFS the results are J48 has 95.32%, Naïve Bayes has 92.98%, and JRIP has 97.07%.

In [12] Proposed a hybrid genetic algorithm for detecting breast cancer. Wisconsin Breast Cancer dataset is used from the UCI machine learning repository. The data is then pre-processed, a hybrid feature selection approach is utilised which is combination of Genetic Algorithm (GA) and Mutual Information (MI) are good indicators of the correlation between features and class names. It is less sensitive to noise or outliers, classifiers such as Support Vector Machine (SVM) which is used to separate two different classed using a hyperplane and K-Nearest Neighbour which considers the distance between different nodes are used and the breast cancer tumors are identified whether it is benign or malignant tumor. For SVM classifier the AUC is 0.9669 and correct rate is 0.9844, for K-NN classifier the AUC is 0.9678 and correct rate is 0.9865.

In the paper [13] used six different machine learning algorithms for prediction of breast cancer, the Wisconsin breast cancer data (original) dataset is used. Then data is pre-processed so that no missing values are present and machine learning algorithms like Support Vector Machine, Naïve Bayes, Random Forest, Decision tree, KNN, Logistics Regression algorithms are used which has the accuracy of SVM has 97.07%, NB and RF has 97%, KNN, DT, LR has 96%.

In [14] proposed a new method for detection of breast cancer, the data set comes from M. Cancer Hospital and Research Institute, Visakhapatnam, India. The data set consists of 8009 histopathological image samples from 683 patients. Then pre-process the data set and use the new DNNS (Value Assisted Deep Neural Network) technology. Using the proposed method, an accuracy of 97.21% is obtained.

By the paper [15] used adaptive ensemble voting scheme for breast cancer detection, the Wisconsin Breast Cancer dataset is used. Different machine learning algorithms like Logistic Regression, Support Vector Machine, K-Nearest Neighbours algorithms are considered. At first the algorithm is applied and projected Ensemble voting techniques for breast cancers detection, then these three algorithms are compared and acquired the precision of 98.50%.

According to [16] used three machine learning algorithms for breast cancer detection, the Wisconsin Breast Cancer Diagnosis (WBCD) dataset is used. Then machine learning algorithms like Support Vector Machine, Decision Tree, K-Nearest Neighbours are applied and SVM acquired the accuracy of 97.9%, K-NN acquired the accuracy of 96.7%, Decision Tree acquired the accuracy of 93.7%.

In research paper [17] proposed a Fuzzy c-means algorithm for early detection of breast cancer, the Wisconsin Breast Cancer Diagnosis (WBCD) dataset is used. The Fuzzy c-means algorithm is applied alone with the pattern recognition model so that tumors can be found accurately and FCM classifier has acquired accuracy of 100% true positive, 87% true negative, 0% false positive, 13% false negative.

In [18] Used Data Mining techniques for identification of breast cancer, the Wisconsin Breast Cancer Dataset is used. Three classification techniques like Sequential Minimal Optimization (SMO), IBK (K Nearest Neighbours classifier), Best First (BF) trees are applied in WEKA and results are acquired. BF tree has the accuracy of 95.46%, IBK has the accuracy of 95.90%, SMO has the accuracy of 96.19%.

Related to paper [19] various machine learning algorithms are applied for prediction of breast cancer, the BCCD and WBCD datasets are used. Then the data is pre-processed and different classification models such as Decision Tree (DT), RF, SVM, Neural Network (NN), Logistics Regression (LR) are applied and for different datasets and for BCCD dataset it has the accuracy of DT has 0.686, SVM has 0.714, RF has 0.743, LR has 0.657, NN has 0.600 and for WBCD it has the accuracy of DT has 0.961, SVM has 0.951, RF has 0.961, LR has 0.937, NN has 0.956.

In reference to [20] three different Machine learning algorithms are used to predict breast cancer, the Iranian center for Breast Cancer (ICBC) from 1997 to 2008 dataset is used. The machine learning algorithms like Decision tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM) are applied and the results are DT has the accuracy of 0.936, ANN has the accuracy of 0.947, and SVM has the accuracy of 0.957.

In the paper [21] certain machine learning algorithms are used for prediction of breast cancers in Chinese women, the Breast Cancer Information Management System (BCIMS) present at West China Hospital of Sichuna University is used. Then different novel machine learning algorithms are used such as XGBoost, Random Forest, and Deep Neural Network are used and the results are acquired, XGBoost has AUC of 0.742, Random Forest has AUC of 0.728, and Deep Neural Network has AUC of 0.728.

In paper [22] various machine learning algorithms are used for the diagnosis of breast cancer, two publically available benchmark datasets are used the Fine Needle Aspirate of Breast Lesions and Fine Needle Aspirates of Breast Lumps (FNAB) are used, machine learning algorithms like Support Vector machine of poly and Radial Basis Function, K-Nearest Neighbours, Probabilistic Neural Network are applied for both datasets and overall accuracy of SVM- Poly is 97.09%, SVM-RBF is 98.80%, KNN is 96.37%, and PNN is 97.23% for dataset I and for FNAB overall accuracy are SVM-Poly is 95.0%, SVM-RBF is 96.33%, KNN is 88.47%, and PNN is 93.39%.

In paper [23] authors used adaptive PSO algorithm, artificial neural network is used for classification of software defects. From paper [24] authors used k-means clustering and ANN for detecting leaf disease and acquired average classification accuracy of 92.5%. According to [25] authors used lifting wavelet transform technique for image transformation for inserting watermark. In paper [26] authors used bacterial foraging particle swarm optimization algorithm for detection of heart failure patients which uses different classification techniques such as KNN, SVM, and neural network classifiers which acquires high accuracy.

3 Comparison

Research made for the breast cancer detection which uses different types of algorithms for detection are differentiated and compared given in the following table:

Table 1. Comparison of the Methods used

Year	Algorithms used	Results	Reference
2017	CNN	82.7%	Reference [1]
2017	CNN using VGG-16	93.1% 88.7%	Reference [2]
2018	Fuzzy Logic	98%	Reference [3]
2018	Naïve Bayes, Bayesian logistic Regression, Simple CART, and J48	95.2654%, 65.4232%, 98.1349%, 97.274%	Reference [4]

2015	Genetic algorithm, Phylogenetic trees, Local binary patterns (LBP) and SVM	92.99%, 83.70%	Reference [5]
2016	SVM, RF and BN	97.0%, 96.6%, 97.1%	Reference [6]
2019	GRU-SVM (gated recurrent unit-support vector machine, Linear Regression, Multilayer, Softmax Regression, and SVM	90.68%, 92.89%, 96.92%, 97.36%, 97.7%	Reference [7]
2019	Region based Convolutional Neural Network (R-CNN) end-to-end training approach	86.7%, 96.1%	Reference [8]
2020	Sequential Minimal Optimization (SMO), J48	96.99%, 75.52%	Reference [9]
2019	MLP, KNN, Classification and Regression Trees (CART), NB, SVM	96.70%, 96.27%, 91.0%, 93.62%, 96.42%	Reference [10]
2020	J48, Naïve Bayes, JRIP	95.32%, 92.98%, 97.07%	Reference [11]
2016	Genetic Algorithm SVM, KNN	96.69%, 96.78%	Reference [12]
2020	SVM, NB, RF, Decision tree, KNN, Logistics Regression	97.07%, 97%, 96%	Reference [13]
2020	Deep Neural Network with Support Value (DNNS)	97.21%	Reference [14]
2019	Logistic Regression, SVM, KNN	98.50%	Reference [15]
2018	SVM, Decision Tree, KNN	97.9%, 96.7%, 93.7%	Reference [16]
2013	Fuzzy c-means algorithm (FCM)	TP-100%, TN-87%, FP-0%, FN-13%	Reference [17]
2014	Sequential Minimal Optimization (SMO), IBK (K Nearest Neighbours classifier), Best First (BF)	96.19%, 95.90%, 95.46%	Reference [18]
2018	Decision Tree (DT), RF, SVM, Neural Network (NN), Logistics Regression (LR)	96.1%, 95.1%, 96.1%, 95.6%, 93.7%	Reference [19]
2013	Decision tree (DT), Artificial Neural Network (ANN), and SVM	93.6%, 94.7%, 95.7%	Reference [20]
2020	XGBoost, Random Forest, and Deep Neural Network	74.2%, 72.8%, 72.8%	Reference [21]
2010	SVM-Poly, SVM-RBF, KNN and PNN	95.0%, 96.33%, 88.47%, 93.39%	Reference [22]

4 Conclusion

In this paper a partial survey for breast cancer detection is done. Various techniques which are proposed earlier are mentioned here along with their usage of algorithms and assumption for execution of the problem. The dataset used for different techniques is also mentioned and the results obtained by using those techniques are mentioned. Since detection of tumor is a difficult task various algorithms produce different results. The computation time is considered along with the accuracy. Based on the survey mentioned in every reference, the possible algorithm combinations are examined. This work can be used for further review of breast cancer detection and can be assessed using all possible methods.

References

1. Y. J. Tan, K. S. Sim, and F. F. Ting, ICORAS, **1-5**, (2017)
2. Shuyue Guan and Murray Loew, IEEE AIPR, **1-8**, (2017)
3. Shaker K. Ali, Wamidh K. Mutlag**,** JATIT & LLS, **Vol.96**. **No 17**, (2018)
4. Dr. S. N. Singh, Shivani Thakral, ICCCA, (2018)
5. Wener Borges de Sampaioa, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Marcelo Gattass, *Elsevier*, (2015)
6. Dana Bazazeh and Raed Shubair, ICEDSA, (2016)
7. Abien Fred M. Agarap, ICMLSC, (2019)
8. Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh, www.nature.com/scientificreports, (2019)
9. Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, and Gunter Saake, *link.springer.com*, (2020)
10. Ali Al Bataineh, *semanticscholar.org*, (2019)
11. Ahmed Abdullah Farid, Gamal Ibrahim Selim1, and Hatem A. Khater, www.preprints.org , (2020)
12. Abeer Alzubaidi, Georgina Cosma, David Brown, A. Graham Pockley, IEEE, (2016)
13. F. M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, Rozina Akter, IJSTR, (2020)
14. Anji Reddy Vaka, Badal Soni, Sudheer Reddy K, KICS, (2020)
15. Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru, IJRTE, (2019)
16. Omar Ibrahim Obaid, Mazin Abed Mohammed, Mohd Khanapi Abd Ghani, Salama A. Mostafa, Fahad Taha AL-Dhief, *International Journal of Engineering and Technology*, (2018)
17. Indira Muhic, *Southeast Europe Journal Of Soft Computing*, (2013)
18. Vikas Chaurasia, Saurabh Pal, *International Journal of Innovative Research in Computer and Communication Engineering*, (2014)
19. Yixuan Li, Zixuan Chen, *Applied and Computational Mathematics*, **VOL.7, NO.4,** (2018)
20. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, JHMI, (2013)
21. Can Hou1, MPH, DPhil; Xiaorong Zhong, DPhil, MD; Ping He, DPhil, MD; Bin Xu, MSc; Sha Diao, MSc, Fang Yi, MSc; Hong Zheng, DPhil, MD; Jiayuan Li, DPhil, *JMIR MEDICAL INFORMATICS*, (2020),
22. Alireza Osareh, Bita Shadgar, IEEE, (2010)
23. B.Dhanalaxmia, G.Apparao Naidu, K.Anuradha, ICICT, (2014)
24. Ch. Usha Kumari, S. Jeevan Prasad, G. Mounika, ICCMC, (2019)
25. Singamaneni Kranthi Kumar, Pallela Dileep Kumar Reddy, Gajula Ramesh3, Venkata Rao addumala, *Traitementn du signal* IIETA, (2019)
26. Padmavathi Kora and Sri Ramakrishna Kalva, *Springer plus*, (2015)