Manuscript Submitted	November 11, 2020
Accepted	December 2, 2020
Published	December 11, 2020

Early Prediction of Breast Cancer through Machine Learning with Minimal Features

Deepak.R.U[1], N.Hemavathi[2*], R.Sriranjani[2], A.Parvathy[2] and M.Meenalochani[3]

[1]UG Student, School of Electrical and Electronics Engineering, SASTRA Deemed to be University
[2] Assistant Professor, School of Electrical and Electronics Engineering, SASTRA Deemed to be University,
[3] Assistant Professor, Department of Electrical and Electronics Engineering, Kings College of Engineering
**nuhemasen@gmail.com*

Abstract

One of the wide and fast spreading diseases among the younger age groups of women is breast cancer. From recent survey it is revealed that once in four minutes a woman is diagnosed whereas a woman dies once in eight minutes due to late detection. If such cases are detected earlier, their lifetime could have been extended. Hence, the objective of the proposal is to predict the presence of breast cancer earlier through deep learning. Deep learning model is implemented in python programming language by using keras Application Programmable Interface and the accuracies of the popular machine learning models such as Logistic Regression, K Nearest Neighbours, Support Vector Machine (linear), Support Vector Machine (RBF), Gaussian (NB), Decision Tree and Random Forest are computed. Initially, the data set with 30 attributes are considered and then feature selection is carried out through heat map. The model consists of number of hidden layers which performs binary classification on the given dataset to predict whether a person is malignant or benign. The proposal exhibits its supremacy by demonstrating greater accuracy and almost similar confusion matrix and execution time in prediction with reduced attributes obtained through feature selection.

***Keywords*:** *Breast cancer, machine learning, deep learning, feature selection, prediction, performance metrics.*

1. Introduction

Breast cancer is one of the diseases which cause a number of deaths every year over the globe. The number of women affected due to breast cancer is increasing in India especially more in urban areas rather than rural areas (Bansi, Saurabh & Chandra, 2016). By 2030, breast cancer causes maximum death among women in India. In the year 2012, women died of breast cancer is 70, 218 which is the highest in the world. Though early detection can save life, the common modes of detection are feasible only if any symptom occurs. Further, the present detecting methods are Mammography, Ultrasound and Magnetic Resonance Imaging (MRI) etc. The existing methods are costlier and also involve uncomfortable procedures and hence, patients cannot afford it frequently. Literature that deals with the devices employed for breast cancer detection is elaborated below.

Comprehensive review on detection techniques using bio markers, DNA bio markers, bio transducers, micro-RNA is presented and the efficacy of the bio sensors is discussed (Sunil, Hardeepkaur, Nandinigautam, Anilk & Mantha 2017). The biomarkers such as Enzyme Linked Immuno Sorbent Assay, Radio Immuno Assay and Immune Histo Chemistry are used. The development of electro

chemical biosensor and nano biosensor for quantification with high sensitivity and selectivity is elaborated (Hossein, 2016). Monolithic silicon pixel detector for high-precision tracking of x-rays is implemented (Sherwoodrker, Christopher & Vincent 1994). A portable tactile imaging system using sensor array, web camera and a personal computer for breast cancer detection is proposed (Ryu, Heo & Kim, 2010). Ultra wide band based wireless breast cancer detection system by examining breast tissues is proposed using single and dual polarization antennas (Hadi, Emily, Adam, Benoit, Milica & Leslie, 2015). A customized micro wave based breast health monitoring system is proposed with 16 wide band antennas that transmit short pulses into breast tissues and receive back scattered responses (Adam, Emily, Eric, Taylor, Milica & Joshua 2015). However, these methods rely on hardware and the accuracy of detection is based on the quality of hardware used.

Wearable fluorescent imaging system is implemented and evaluated for sensitivity and experimented in mouse (Shengkui, Suman, Nan, Rongguang, Samuel & Viktor, 2015). Microwave based breast health monitoring and breast cancer detection is implemented with wearable patient interface (Emily, Hadi, Adam, Benoit, Leslie & Milica, 2016, Ulta, Laila, Zainal & Nur, 2018). Wearable bio impedance and infra red spectroscopy based on optical and electrical properties employed for breast cancer detection is demonstrated (Omar, Mariam, Mohamed & Klaus, 2018). In millimeter-wave imaging, significant dielectric contrast between healthy and tumour tissues is observed. Thermography identifies cancer based on temperature changes in breasts by fixing multiple thermal sensors in various positions. Nonetheless, the above methods are less accurate, complex and time consuming. In addition, as early detection or prediction of such ailment lessens the quantity of deaths, machine learning based early prediction of breast cancer is proposed.
Computation of breast density from breast MRI is attempted through expectation-maximization segmentation of body breast and air breast surfaces (Albert, Michiel, Ritse, Robert & Nico, 2015). Machine learning based cancer detection is implemented (Erwin, Pauline & Marylise, 2018). For the thermal images so obtained from thermography, machine learning techniques such as Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Random Forest (RF) and Decision Tree (DT) are employed to detect the presence of cancer (Vartika, Yamini & Santanu, 2019). Detection of breast cancer through hybrid approach of MAD normalization and AdaboostM1 classification is presented (Kemal & Ümit, 2018). Digital Database for Screening Mammography (DDSM) is employed to detect cancer by employing machine vision techniques (Muhammad, Tufail & Shahzad, 2019. Deep learning based detection using screening mammograms is presented (Li, Margolies, Joseph, Eugene, Russell & Weiva, 2017). Deep learning based detection using histopathological Images by employing transfer learning is implemented (Juanying, Ran, Joseph & Chaoyang, 2019). The mammography images are segmentation through region of interest and threshold and region based methods (Dina, Maha, Stephen & Jinchang, 2019). The segmented image is fed to deep learning model for feature extraction and binary classification is performed using support vector machine. However, the choice of machine learning techniques from the pool of algorithms relies on nature of data, execution time, system specification etc.
To address the pitfalls of the existing systems, deep learning based breast cancer detection system is proposed. The proposal is elaborated in the succeeding section.

2. Proposed system

The system architecture of the proposed system is presented in Figure 1. The input data set of 570 patients with 30 attributes is fed as input to the deep learning model.

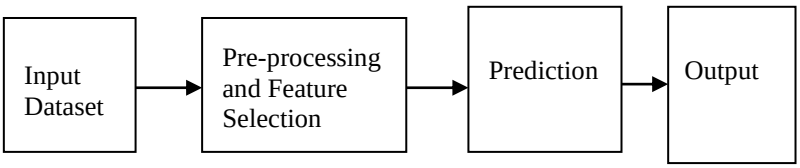


Figure 1. Architecture of Breast Cancer Detection Scheme

The data ought to be pre-processed to remove erroneous and missing data. Further, the pre-processed data when it is utilized with 30 attributes, then the execution time and the system resources involved in computation is also increased which is not desirable. In such scenario, feature selection with significant attributes play a significant role in reducing the execution time and resource utilization. Therefore, to attain, feature selection, heat map is implemented and significant features are chosen. The dataset with significant features are fed as input to prediction algorithm. The deep learning model is constructed, trained and tested using python 3.7 programming language in anaconda platform. Deep learning model is constructed using keras API. Though the constructed model, presence of breast cancer is predicted through a pool of machine learning algorithms such as logistic regression, K Nearest Neighbours, Support Vector Machine (linear), Support Vector Machine (RBF), Gaussian (NB), Decision Tree and Random Forest. The implementation of the proposed breast cancer detection scheme is elucidated in the succeeding section.

3. Implementation and results

There are several types of models available in keras API. This kind of application Sequential model of keras enables us to easily create a deep learning neural network. In sequential model we can add any number of layers of our wish and feed datasets to the network. The dataset used to train the model is taken from the kaggle community named Breast cancer Wisconsin (Diagnostic) dataset. The dataset contains 30 features of the breast cancer tumour and records of 570 patients. Those 30 inputs are listed in Table 1.

The programming part of the deep learning model is done in the anaconda environment and python 3.7. Considering the complexity behind the maths of deep learning keras API is used to simplify the operations. Initially, 30 features of the tumour listed in Table 1 are used to predict breast cancer. For evaluating the model, k-fold cross validation form sk learn is employed. This evaluates the model by a simple resampling technique or in other words this splits the given data into k-parts and trains the model, this will be used to evaluate the performance of the model.

Further, wrapper is an object of a class which is used to create and return the neural network which is termed as fit function that takes arguments such as number of epochs and batch size. The deep learning model so constructed has four layers of fully connected hidden layers with 7 and 3 nodes respectively. The initial weights of the nodes are assigned using random numbers. ReLu (Rectified Linear unit) activation functions are used in the neural network and in the last output layer alone the Samoyed activation function is used in order to get the prediction probability ranging from 0 to 1. The accuracies of the popular machine learning models such as Logistic regression, K Nearest Neighbours, Support Vector Machine (linear), Support Vector Machine (RBF), Gaussian (NB), Decision Tree and Random Forest from the sklearn API are presented in Table 2. From Table 2, it is inferred that random forest algorithm outperforms when compared with other machine learning algorithms. Attaining the accuracy with 30 features is the conventional that leads to resource and time consuming. Hence, the proposal aims to attain such accuracy with a lesser number of features. To reduce the number of features, feature selection algorithm ought to be implemented.

Table 1. Features in the Dataset

Features	Features	Features
Radius mean	Radius largest worst	Texture se
Perimeter mean	Perimeter largest worst	Area se
Smoothness mean	Smoothness largest worst	Compactness se
Concavity mean	Concavity largest worst	Concave points se
Symmetry mean	Symmetry largest worst	Fractal dimension se
Radius se	Texture mean	Texture largest worst
Perimeter se	Area mean	Area largest worst
Smoothness se	Compactness mean	Compactness largest worst
Concavity se	Concave points mean	Concave points largest worst
Symmetry se	Fractal dimension mean	Fractal dimension largest worst

Table 2. Machine Learning Models with Corresponding Accuracy for 30 Attributes

Machine Learning Model	Accuracy
Logistic regression	95.8041958041958
K nearest neighbours	95.1048951048951
Support vector machine linear	97.2027972027972
Support vector machine RBF	96.5034965034965
gaussianNB	91.6083916083916
Decision tree	95.8041958041958
Random forest	98.6013986013986

In addition, the feature selection should be carried out without deteriorating the performance of the machine learning algorithm in terms of accuracy. Hence, the performance of the model with respect to number of features is studied using sk learn and is plotted in Figure 2.

From Figure 2, it is evident that the performance of the machine learning algorithm is found to be correlating when the number of features is almost 11. Further, the choice of features should be identified through some means. This is accomplished through heat map depicted in Figure 3 which correlates between all the 30 features with each other. Based on the highest level of correlation, the heat map yields 14 distinct features which are not related to each other and is tabulated in Table 3.

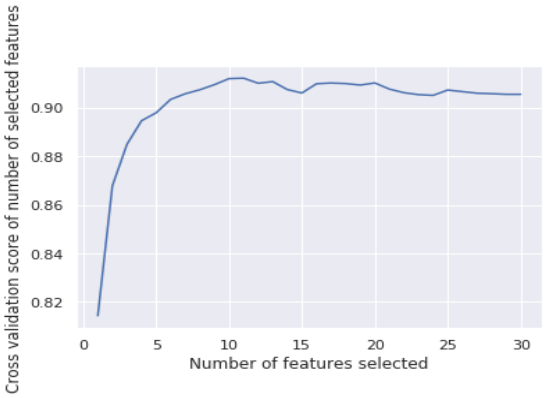
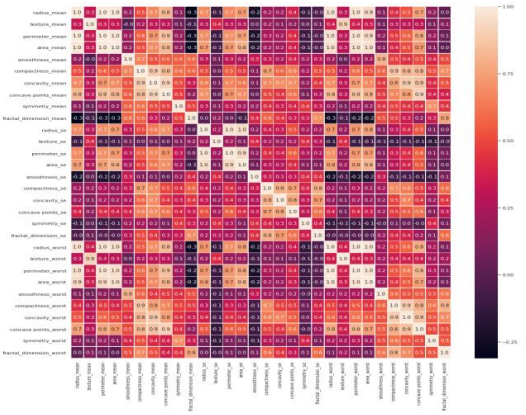


Figure 2. Number of Selected Features Vs Cross Validation Score

Table 3. Selected Features through Heat Map

Features	Features
perimeter_mean	radius_mean'
radius_se	perimeter_se
compactness_worst	concave points_worst
texture_worst	area_worst
compactness_mean	concave points_mean
radius_worst	perimeter_worst
compactness_se	concave points_se

The accuracy of machine learning algorithms correspond to 14 features is computed and is compared with that of 30 features and is depicted in Figure 4.



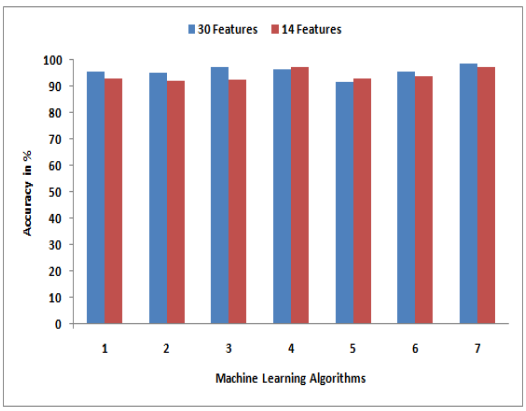


Figure 3. Heat map for feature selection

figure 4. Accuracy of Machine Learning Algorithms

Note:1- Logistic regression 2-K Nearest Neighbours 3-Support Vector Machine (linear) 4-Support Vector Machine (RBF) 5-Gaussian (NB) 6-Decision Tree 7-Random Forest

From Figure 4, it is evident that the accuracy of all the algorithms with 14 features almost coincides with that of 30 features. To emphasize the significance of the proposal, the confusion matrix of all the algorithms correspond to 30 and 14 features are demonstrated in Table 4.

From Table 4, it is apparent that the number of true positive cases for both the cases almost coincides which clearly exhibits the efficacy of the proposal. As the performance of random forest algorithm outperforms when compared with other machine learning algorithms, the performance of the same algorithm having 30 features is compared with 14 features. Their corresponding confusion matrix is presented in Figures 5 (a) and (b) respectively.

Table 4. Confusion Matrix of Algorithms

Confusion Matrix / Algorithms	30 features				14 features			
	TP	**FP**	**FN**	**TN**	**TP**	**FP**	**FN**	**TN**
Logistic regression	87	3	3	50	88	2	4	49
KNN	89	1	6	47	83	7	4	49
SVM (linear)	88	2	2	51	88	2	3	50
SVM (RBF)	88	2	3	50	88	2	2	51
Gaussian (NB)	84	6	6	47	85	5	5	48
Decision Tree	86	4	2	51	83	7	2	51
Random Forest	89	1	1	52	88	2	2	51

Note: True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN)

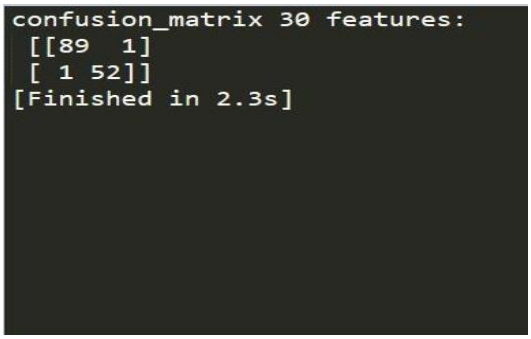


Figure 5 (a). 30 features

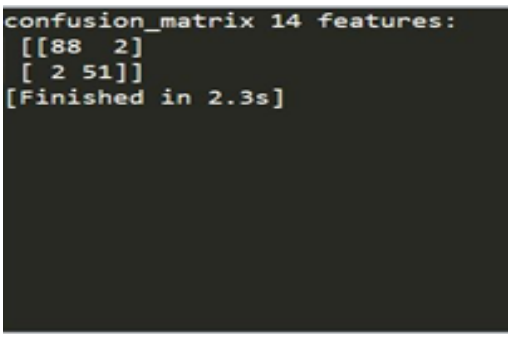


Figure 5 (b). 14 features

Figures 5(a) and (b) Snapshot of confusion matrix of Random Forest Algorithm

From Figures 5(a) and (b), it is evident that true positive with 30 features is 89 whereas for 14 attributes, it is observed as 88

8. Conclusion

Breast cancer is one of the diseases that cause death worldwide. The death rate can be reduced if the disease is predicted at early stage. Hence, deep learning based breast cancer early prediction system is proposed. The proposal is implemented in Python. A pool of machine learning algorithms such as Logistic regression, K Nearest Neighbours, Support Vector Machine (linear), Support Vector Machine (RBF), Gaussian (NB), Decision Tree and Random Forest are implemented and their accuracy are computed with 30 features. Further, for efficient utilization of resources, feature selection is carried out through heat map. The efficiency of the proposed model with 14 features is

evaluated by comparing the accuracy and confusion matrix with that of system with 30 features. The results demonstrate that performance of system with 14 features coincide with that of 30 features. This proves the effectiveness of the proposal. Further, the work can be extended to feature extraction from the scan image to enhance the accuracy.

References

Bansi D Malhotra, Saurabh Kumar & Chandra Mouli Pandey. (2016). Nanomaterials based biosensors for cancer biomarker detection, Journal of Physics: Conference Series. 704, 1-11.

Sunil Mittal, Hardeepkaur, Nandinigautam, Anilk & Mantha. (2017). Biosensors for breast cancer diagnosis: a review of bioreceptors, biotransducers and signal amplification strategies, Biosensors and Bioelectronics, Elsevier. 88, 217–231.

Hossein Naderi-Manesh. (2016). Early detection of breast cancer using electrochemical and nano biosensor, Cancer Diagnostics Conference & Expo, Italy. 8(5), 88.

Sherwood I. Parker, Christopher J. Kenney & Vincent Z. Peterson. (1994). Breast cancer calcification measurements using direct x-ray detection in a monolithic silicon pixel detector, IEEE Transactions on Nuclear Science. 41(6), 2862-2873.

Ryu, J, M.S. Heo & H.C. Kim. (2010). Development of portable breast self-examination device using enhanced tactile feedback, Electronics Letters. 46(25), 1651 -1653.

Hadi Bahramiabarghouei, Emily Porter, Adam Santorelli, Benoit Gosselin, Milica Popovi´C & Leslie A. Rusch. (2015). 'Flexible 16 Antenna array for microwave breast cancer detection, IEEE Transactions on Biomedical Engineering.62(10), 2516-2525.

Adam Santorelli, Emily Porter, Eric Kang, Taylor Piske, Milica Popovi´C & Joshua D. Schwartz. (2015). A time-domain microwave system for breast cancer detection using a flexible circuit board, IEEE Transactions on Instrumentation and Measurement, 64(11), 2986-2994.

Shengkui Gao, Suman Mondal, Nan Zhu, Rongguang Liang, Samuel Achilefu & Viktor Gruev. (2015). A compact nir fluorescence imaging system with goggle display for intraoperative guidance, IEEE International Symposium on Circuits and Systems (ISCAS).1622-1625.

Shengkui Gao, Suman Modal, Nan Zhu, Rongguang Liang, Samuel Achilefu & Viktor Gruev. (2015). Live demonstration: a compact nir fluorescence imaging system design with goggle display for intraoperative guidance, IEEE International Symposium on Circuits and Systems (ISCAS). 1910-1910.

Emily Porter, Hadi Bahrami, Adam Santorelli, Benoit Gosselin, , Leslie A. Rusch & Milica Popović. (2016). A wearable microwave antenna array for time-domain breast tumor screening, IEEE Transactions on Medical Imaging. 35(6), 1501-1509.

Ulta Delestri, Laila Fadhillah, Zainal Abidin, Nur Afikah et al. (2018). Development of a low-cost wearable breast cancer detection device, 2nd International Conference on Biosignal Analysis, Processing and Systems (ICBAPS).41-46.

Omar Farag, Mariam Mohamed, Mohamed A. Abd El Ghany & Klaus Hofmann. (2018). Integrated sensors for early breast cancer diagnostics, 21st International Symposium on Design and Diagnostics of Electronic Circuits and Systems. 153-157.

Albert Gubern-M´Erida, Michiel Kallenberg, Ritse M. Mann, Robert Mart´I & Nico Karssemeijer. (2015). Breast segmentation and density estimation in breast mri: a fully automatic framework, IEEE Journal of Biomedical and Health Informatics, 19(1), 349-357.

Erwin Halim , Pauline Phoebe Halim & Marylise Hebrard. (2018). Artificial intelligent models for breast cancer early detection, International Conference on Information Management and Technology (ICIMTech).517-521.

Vartika Mishra, Yamini Singh & Santanu Kumar Rath. (2019). Breast cancer detection from thermograms using feature extraction and machine learning techniques, 5th IEEE International Conference for Convergence in Technology (I2CT).1-5.

Kemal Polat & Ümit Şentürk. (2018). A novel ML approach to prediction of breast cancer: combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier, 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).1-4.

Muhammad Aaqib, Muhammad Tufail & Shahzad Anwar. (2019). A novel deep learning based approach for breast cancer detection, 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). 1-6.

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell B. Mcbride & Weiva Sieh. (2017). Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography, Scientific Report. 1-14.

Juanying Xie, Ran Liu, Joseph Luttrell IV & Chaoyang Zhang. (2019). Deep learning based analysis of histopathological images of breast cancer, Frontiers In Genetics.10, 80, 1-19.

Dina A. Ragab, Maha Sharkas, Stephen Marshall & Jinchang Ren. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines, Bio Informatics and Genomics.1-23.