# BREAST CANCER PREDICTION WITH HYBRID ML MODELS

Dr. S. Sridevi

**ABSTRACT**
In recent times various types of cancer propagation in humans are alarmingly increasing and especially women are prone to and threatened by breast cancer with high morbidity and mortality. The absence of robust prognosis models results in difficulty for physicians to prepare a treatment plan that may extend patient survival chances and time. Hence, the need of the time is to develop the technique which offers minimum error with increased accuracy. Different legacy algorithms like SVM, Regression, are compared with the proposed hybrid prediction model outcome. All experiments are executed within a parallel environment and conducted in anaconda python platform with relevant libraries. This is helpful in domains like. prediction of cancer before diagnosis, prediction of diagnosis and outcome during treatment. The proposed work combining detailed pre-prepressing stages over a deep neural network model with tuned hyper parameters, validated to yield needed accuracy. This can be used to derive and compare the outcome of different techniques and suitable one having max accuracy and stability, can be used depending upon requirement. Different data sets are tried and analysed for prediction with different parameters and results are compared.

**Keywords** — Breast Cancer detection, machine learning, feature selection, classification, hybrid deep learning, image classification, KNN , Random Forest, ROC.

## 1. INTRODUCTION

According to cancer.org[1], breast cancer is the most common cancer in women. In the America, there is a chance that one in 8 women will develop breast cancer. The 2nd major cause of women's death is breast cancer (second to lung cancer). Approximately 0.25 million women's new cases of invasive breast cancer are registered in the USA during 2016 and forty thousand of women’s death has happened due to deferred diagnosis and treatment. Breast cancer is a type of cancer that originates in the cells of the breast. While normal controlled cell growth is needed for human metabolism and evolution, cancerous growth starts when cells begin to grow out of proposition and control. Breast cancer cells typically starts form a tumour that can often be seen on an x-ray or felt as an unusual lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The main cause of breast Cancer includes changes & mutations in DNA.

There are many different types of breast cancer like ductal carcinoma in situ (DCIS) & invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less prevalent. Some of the stages of classification are shown in below figure 1.

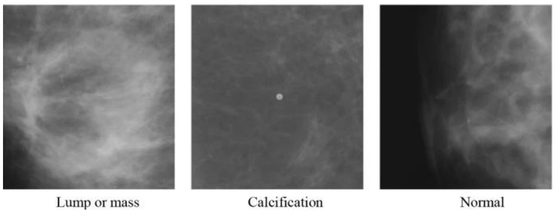

Fig.1 Mammogram images for cancer detection

There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – localized pain, fatigue, headaches & numbness (peripheral neuropathy), bone loss and osteoporosis (bones become fragile). It can be medically detected early during a screening examination through breast mammography or by portable cancer diagnostic tool.

Cancerous breast tissues change with the progression of the disease over time, which can be directly linked to different stages of cancer. The 4 stage of breast cancer describes the extent to which the patient’s cancer has proliferated. Statistical parameters & indicators such as lymph node metastasis, tumour size and distant metastasis and so on are used to determine stages.

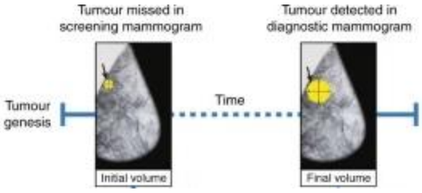
Tumour missed in screening mammogram
Tumour detected in diagnostic mammogram
Tumour genesis
Time
Initial volume
Final volume

Fig 2. Tumour growth & need for early detection

To arrest cancer from spreading, patients have to undergo breast cancer surgery, radiotherapy, chemotherapy and endocrine. The aim of the paper is to identify and classify Malignant and Benign patients and intending how to parametrize the classification to achieve high accuracy. With comparison of legacy algorithms and proposed model we aim to reduce the error rates with maximum accuracy with analysis on data in terms of effectiveness and efficiency.

## 2. LITERATURE SURVEY & RELATED WORK

Breast Cancer mainly caused by changes and mutations in DNA of human cells in uncontrolled manner. Breast cancer cells usually form a tumour that can be detected on an x-ray or felt as a lump. Earlier detection and rapid inference through machine learning methods help physicians to get more meaningful insights to act upon faster. There are many algorithms for classification & prediction and earlier detection of breast cancer outcomes. random forest, deep learning, SVM and different image processing methods can be applied to detect malignant breast masses. The typical process will have 3 parts like detecting breast masses segmentation and finally whether a mass is benign or malignant. This paper gives a comparison between the performance of different classifiers, which are among the most influential data mining algorithms.

Akbugday et al.,[2] analysed classification on Breast Cancer Dataset by using KNN, SVM and achieved accuracy of 96.%. Wang et al.[3] used Logistic Regression and achieved an Accuracy of 96.4 %. KELES et al., [4] used data mining & random forest & achieved accuracy of 92%. Vikas et al., [5] compare the performance criterion of supervised learning classifiers; like RBF neural networks, Naïve Bayes, SVM-RBF kernel, Decision trees (J48) and simple CART, to find the best classifier in breast cancer datasets.

Kavithal et al.,[6] used ensemble methods with Neural Networks and achieved accuracy of 96.3% lesser than previous studies. Dalen et al. [7] used ADABOOST and achieved accuracy of 97% better than Random Forest. As per Sinthia et al.,[8], backpropagation provides 94 % accuracy. The experimental result shows that SVM-RBF kernel is more accurate than others; it scores accuracy of 96.8% in Wisconsin datasets.

Prediction and prognosis of cancer development are focused on 3 major domains: prediction of cancer vulnerability, prediction of cancer deterioration, and prediction of cancer survival rate. The first domain comprises prediction of the probability of developing certain cancer prior to the patient diagnostics.

The second issue is related to prediction of cancer recurrence in terms of diagnostics and treatment, and the third case is aimed at prediction of several possible parameters characterizing cancer development and treatment after the diagnosis of the disease: survival time, life expectancy, progression, drug sensitivity, etc. The survivability rate and the cancer relapse are dependent very much on the medical treatment and the quality of the diagnosis. Thus to address above issues a robust perdition model for breast cancer tumour progression at earlier stages is necessary which can learn dynamically and tune itself. Thus the proposed model is envisaged to have the above attributes to obtain better efficacy.

## 3. PREDICTION MODEL ARCHITECTURE

Figure 2 below depicts the stages of hybrid prediction model envisaged. The first step before modelling is the data pre-processing stage where data mining techniques are used to filter data converting real world data in a usable format.

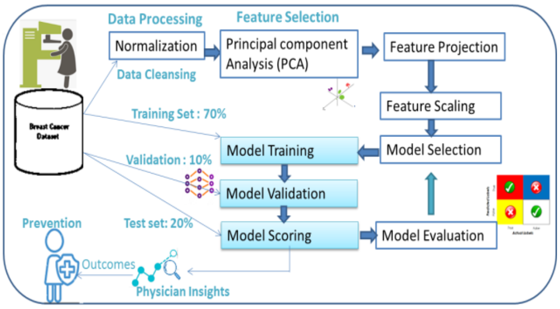

Fig. 3 Prediction model stages

Data pre-processing is a proven method of resolving such issues and different techniques like normalization, standardization & log transformation methods are used. This is followed by feature selection methods like PCA[9] where the identification of a set of axes (vectors) that conveys the maximum quantity of information (most influential input vector(s) over output) in the characteristic space of a multidimensional dataset. Suitable model like MCP, SVM, KNN or ensemble methods are chosen and results are compared. The same mammogram images data set is fed to above different models. The selected features (columns / attributes of cancer parameters) are distributed in 70:10:20 for training, validation and scoring data set (refer below python code)

```
datasets = [
("training", trainPaths, config. TRAIN_PATH),
("validation",  valPaths, config.VAL_PATH),
("testing", testPaths, config.TEST_PATH) ]
```

The model scoring and evaluation compares the predicted vs actual outcomes and repeated until the accuracy, precision and recall factors in desired range. With the trained model, physicians can feed the new patient's mammogram images for getting needed insights and start earlier treatment models to prevent further prognosis of the tumours in the patient.

## 4. DATA & FEATURES PREPARATION

In the data collection phase we do is to collate the data of mammogram images of different patient distribution. We used Kaggle mammogram images repository [10]. Detailed pre-processing, normalization techniques are applied to substitute missing / skewed values and convert them to numerical and categorical format. Real-world data is typically incomplete, inconsistent, and lacking certain to contain many errors. Typically Keras BatchNormalization and other tools are imported for these functions.

Some of the image transformation methods involves, flipping the images along a horizontal axis, shifting vertically/horizontally within a width range of 0.1 to 0.2 and rotating randomly within a $20^o$ range.

Data pre-processing is a proven method of resolving these issues in raw data obtained from different sources. This step is very crucial for the model accuracy, because the quality and quantity of data that you gather will directly determine outcome of the predictive model. For Feature selection and identifying most influential input parameters impacting output PCA techniques are used.

Feature projection or dimensionality reduction is a technique to transform high-dimensional space data to a lower dimensional space (from large input attribute set to few enriched/focused attributes). This can be either linear or nonlinear reduction techniques can be used suitable for the type of relationships among the features in the dataset

Data Set from chosen repository are fed to PCA and out of 30+, parameters about 8 parameters are selected. The output parameter is breast cancer diagnosis – malignant or benign. The Key features found by the study are: Concave points worst, Area worst, Area se, Texture mean, Smoothness worst, Smoothness mean, Symmetry mean, Radius mean, Texture worst etc.

Typically any dataset will contain attributes which are highly varying in magnitudes, units and range. As most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of normalized magnitudes. This can be achieved by feature scaling methods.

## 5. HYBRID PREDICTION MODEL

Typical model construction involves below steps for any ML solution to real-world problems:

1. Construction of the baseline model & iterative performance evaluation.

2. Training of different popular models with various architectures, & selection of best models.

3. Deployment of regularization (Penalize over-fitting) and data augmentation methods to develop the performance.

4. Tuning of hyper-parameters on the final model to get the desired level.

To perform segmentation or classification on large complex images, a classifier is used in sliding window manner to recognize local patches on an image to generate a grid of probabilistic outputs. Subsequently another process to summarize the patch classifier's outputs to give the final classification or segmentation outcome.

For a given input image patch $X \in IR^{p\times q}$ & a patch classifier which is a function *f* so that $f(X) \in IR^c$, where the function's output satisfies $f(X)_i \in [0,1]$ and $\Sigma^c_{i=1} f(X)_i = 1$ and *c* is the number of classes of the patches. Here, *c* is 5 and the classes are: benign &, malignant calcification, benign & malignant mass and background for each patch from a mammogram.

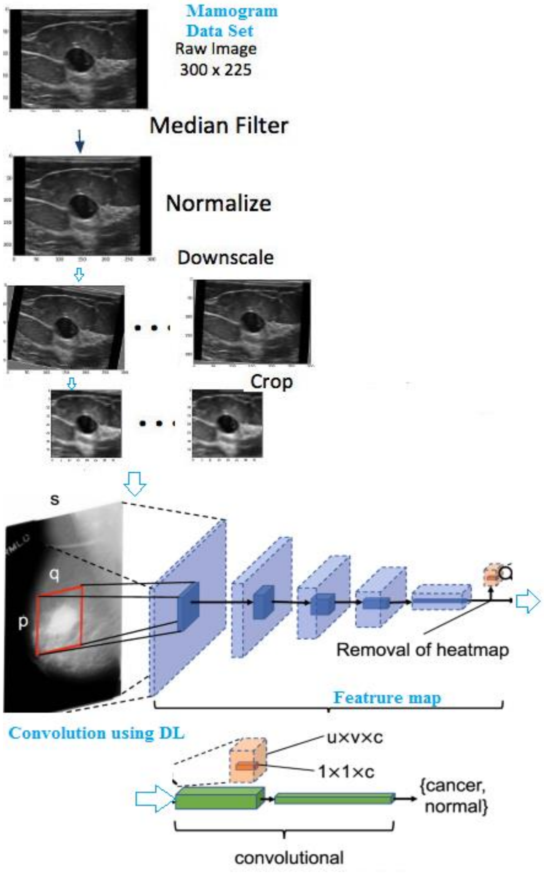

Figure 4. Feature extraction & Convolution

Assuming the input patch is extracted from an image M $\in$ IR$^{r\times s}$ where $p \ll r$ and $q \ll s$. If the function *f* represents a convolutional neural network, then *f* can be applied to *M* without changing the network parameters so that f(M) $\in$ IR$^{u\times v\times c}$ , where $u > 1$ and $v > 1$ depend on the image size and the stride of the patch classifier as in below figure.

An advanced CNN is typically constructed by stacking convolutional layers above the input, followed by 1 or more fully connected layers to join with the classification output. Max pooling layers are often used amid convolutional layers to improve translational invariance & to reduce feature map size. Prevalent CNN structures like: the VGG network[11] & the residual network (Resnet)[12] typically comes handy to formulate this detection problem. Subsequent network layers can be typically grouped into blocks, so that the feature map size is decreased, either at the beginning or at the end of a block but stays the same elsewhere in the block. For eg, a VGG model, is a stack of multiple 3 × 3 convolutional layers with the same depth followed by a 2 × 2 max pooling layer which reduces the feature map size with a factor of two.

Sample Python code for Use 3×3 CONV filters & max pooling:

```
model.add(SeparableConv2D(64, (3,3),
padding="same"))
model.add(Activation("relu"))
model.add(BatchNormalization(axis=
channelDim))
model.add(SeparableConv2D(64, (3,3),
padding="same"))

model.add(Activation("relu"))
model.add(BatchNormalization(axis=
channelDim))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.25))
```

The model is trained using below python code & parameters:

```
M = model.fit_generator( trainGen,
steps_per_epoch = lenTrain//BS,
validation_data = valGen,
validation_steps = lenVal//BS,
class_weight = classWeight,
epochs = NUM_EPOCHS)
```

The trained model is used classify the new images during scoring phase and its performance is evaluated

## 6. RESULT ANALYSIS

The below charts depicts the trend of accuracy over the training cycles(epochs) and the error function output (loss) reduction trend over epochs, which are in desired ranges. The blue and orange lines represent training and validation phase performance.

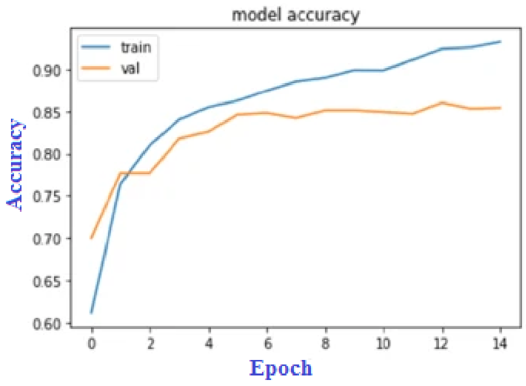

Figure 5. Epoch vs Accuracy Trend

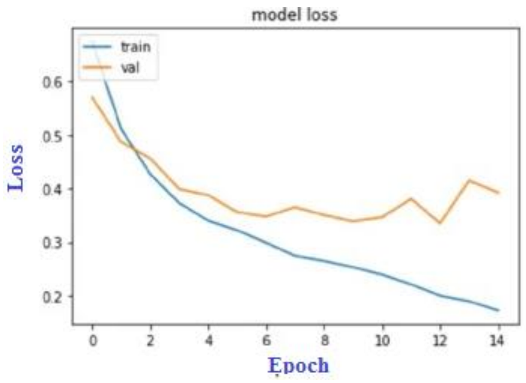

Figure 6. Epoch vs error loss Trend

The ROC figure (7) of the model evaluation provides an AUC of 93.3%, with repeated tuning, under 2 different studies

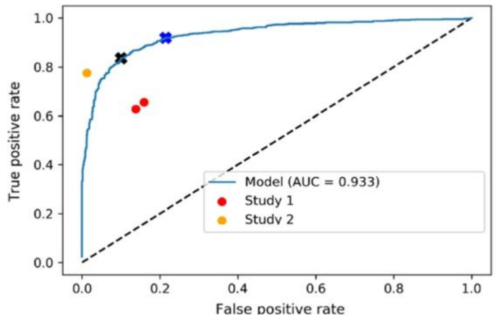

Figure 7. ROC Curve trend

Below figure 8 represents the True positive Vs other confusion matrix attributes overlay above the malignant tissue size in the output image for better visual understanding.

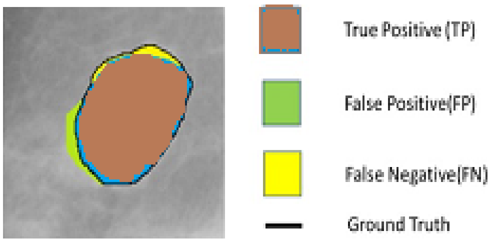

Figure 8. Result accuracy

The performance comparison with other legacy models with the proposed hybrid mode is shown in below figure 9. As conclusion, this model with its well-structured validations, sequence of ensemble algorithms augmenting the real power of deep neural nets, yields the needed accuracy. The estimated clinically relevant threshold and relevant data would be well suited to sufficiently reduce errors, especially false negatives, in the clinical setting.

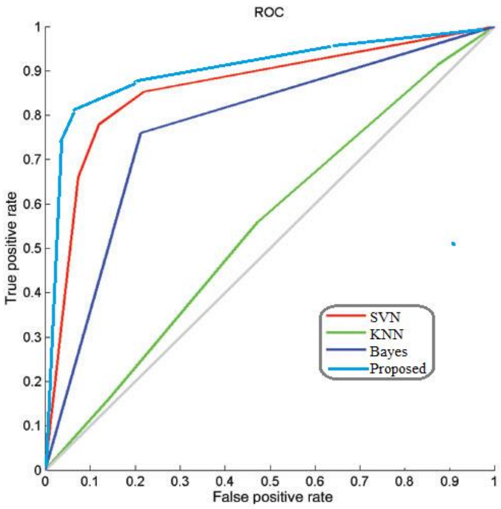

Figure 9. ROC comparison with legacy models

## 6. CONCLUSION & FUTURE RESEARCH

There are different approaches to enhance current model in future. More categorical classification based on the BI-RADS scores, tissue density data and others can be adopted, which has a constraint to have the masses and calcification to have merged.

The proposed approach was found to obtain correct results that might decrease human errors in the diagnosis process & reduce the cost of cancer diagnosis. The limitation of this study is to use the secondary database like Kaggle, and future study could be based on primary real-time hospital data for more accuracy of the results linked to breast cancer identification. The availability of additional, very detailed data set containing not only the BI-RADS scores but also other medical explanations / insights over the images help to tune the model for better precision & accuracy in future, which will help the high demand for research in this popular domain of non-invasive breast tumour detection models.

## REFERENCES

[1] cancer.org url : https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[2] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[3] Wang, D. Zhang and Y. H. Huang “Breast Cancer Prediction Using Machine Learning” (2018), Vol. 66, NO. 7

[4] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[5] Vikas Chaurasia and S.Pal, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis” (FAMS 2016) 83(2016) 1064 – 1069

[6] R. K. Kavitha1, D. D. Rangasamy, “Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm Adaboost and CART Algorithm” Volume 3, Special Issue 1, February 2014

[7] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

[8] P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, “Breast Cancer detection using PCPCET and ADEWNN”, CIEEE’ 17, p.63-65

[9] H. -J. Chiu, T. -H. S. Li and P. -H. Kuo, "Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and SVM," in IEEE Access, vol. 8, pp. 204309-204324, 2020

[10]https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images

[11] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556* [*cs*], 1409.1556 (2014).

[12] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv:1512.03385* [*cs*], 1512.03385 (2015).