# Breast Cancer Prediction using Deep learning and Machine Learning Techniques

Monika Tiwari
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
monika.rt@somaiya.edu

Rashi Bharuka
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
rashi.b@somaiya.edu

Praditi Shah
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
praditi.shah@somaiya.edu

Reena Lokare
K.J. Somaiya Institute of Engineering & Information Technology, Sion, Mumbai
Department of Information Technology
University of Mumbai, India
reena.l@somaiya.edu

***Abstract*- Breast Cancer is mostly identified among women and is a major reason for increasing the rate of mortality among women. Diagnosis of breast cancer is time consuming and due to the lesser availability of systems it is necessary to develop a system that can automatically diagnose breast cancer in its early stages. Various Machine Learning and Deep Learning Algorithms have been used for the classification of benign and malignant tumours. The Wisconsin Breast Cancer Dataset has been used which contains 569 samples and 30 features. The paper emphasises on various models that is implemented such as Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN), Multi-Layer perceptron classifier, Artificial Neural Network(ANN)) etc. on the dataset taken from the repository of Kaggle. Each of these algorithms has been measured and compared with respect to accuracy and precision obtained. All the techniques are coded in python and executed in Google Colab, which is a Scientific Python Development Environment. The experiments have shown that SVM and Random Forest Classifier are the best for predictive analysis with an accuracy of 96.5%. To increase the accuracy of prediction, deep learning algorithms such as CNN and ANN have been implemented. The maximum accuracy obtained in the case of ANN and CNN are 99.3% and 97.3% respectively. Activation functions such as Relu and sigmoid have been used to predict the outcomes in terms of probabilities.***

**Keywords- Classification, Machine Learning, Deep learning, Support vector Machine, Random Forest, ANN and CNN.**

## I. INTRODUCTION

Breast Cancer is the second most dangerous cancer, the first being lung cancer. Breast cancer constitutes 11% of new cancer cases approximately out of which close to 24 % are women [1]. People visit an oncologist, in case of any sign or symptom of cancer. The oncologist can diagnose and detect breast cancer through Mammograms, Magnetic resonance imaging (MRI) of breast, ultrasound of X-ray of the breast, tissue biopsy etc. Once breast cancer is confirmed, sentinel node biopsy of the patient is done regularly which helps to detect cancerous cells in lymph nodes. Machine Learning techniques are also used for the classification of benign and malignant tumours. The early detection of Breast Cancer can enhance the prediction and survival rate of the patients. This will help the patients to take necessary treatments at the right time [2]. For benign tumours the patients can avoid unnecessary treatments. Data mining techniques when applied in the medical field can help in prediction of various outcomes, cost minimisation and upgrade the healthcare value to rescue lives of people. The process of classifying tumours can be done by machine learning and deep learning techniques. Deep learning techniques can give better accuracy when data is complex and large compared to machine learning algorithms. Research is being done in the area for various datasets on Breast Cancer.

## II. RELATED WORK

There are many deep learning and machine learning techniques available for cancer detection and prediction. Some of the most used deep learning techniques are Convolutional Neural Network, Recurrent Neural Network and some pre-trained models such as Alex Net, Google Net, VGG16, VGG19, ResNet. Some of the most used dataset available for training and testing are Mammogram image, SEER, UCI, WBCD.

Dongdong Sun et al. have proposed a deep learning (DL) method named D-SVM for the prediction of human breast cancer prognosis. The algorithm effectively learned hierarchical and abstract representation from raw input data and successfully integrated traditional classification method [3].

D. Selvathi et al. have proposed an automated system for achieving error-free detection of breast cancer using a Sparse Autoencoder (SAE) which learns feature representations from the mammogram and a classifier which is cascaded with the SAE performs the classification based on these learned features [4].

Tiancheng He et al. combines natural language processing and deep learning methods to develop an analytic model that targets well-characterized and defined specific breast suspicious patient subgroups rather than a broad heterogeneous group for diagnostic support of breast cancer management [5].

Naresh Khuriwal et al. demonstrates how deep learning technology can be used for the diagnosis of breast cancer using UCI Dataset. The dataset has been collected and pre-processing has been applied to it for scaling and filtering, after which the dataset has been split into training and testing sets and finally graphs have been generated in order to visualise the data. Deep learning algorithms have also been compared with other machine learning algorithms [6].

Ch. Shravya et al. focuses a relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN) on a particular dataset. [7].

J. Ferlay et al. briefly explains the sources and methods of estimation, validity and completeness of available data, and possible explanations for the observed patterns of incidence and mortality [8].

M. J. Van De Vijver et al. proposed microarray analysis to evaluate previously established 70-gene prognosis profile and classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis [9].

## III. METHODOLOGY

In this paper various machine learning and deep learning algorithms have been used for the diagnosis of breast cancer. The paper consists of two main parts, pre-processing of the data and creating models for prediction. In this paper, the Wisconsin Breast Cancer Dataset has been used that is publicly available for researchers [10]. This database is generated from biopsy images and contains 569 samples and 30 features.

The Fig 1 highlights the steps to be followed from start to end in order to implement a model that can be used for prediction of breast cancer.

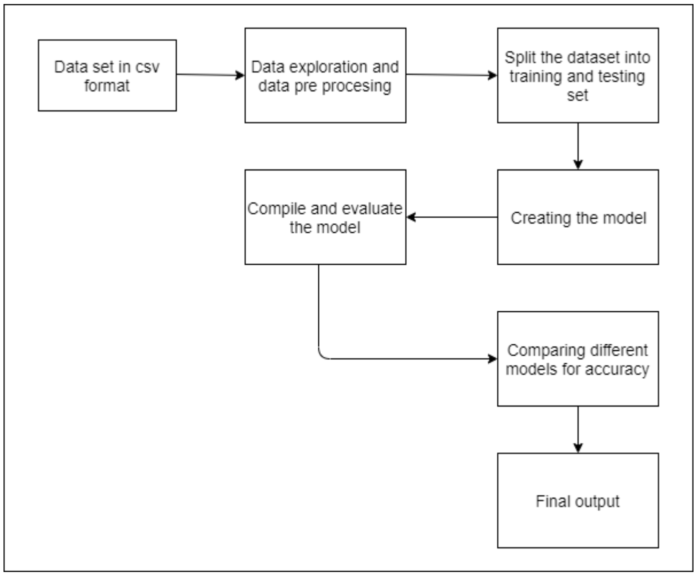
Data set in csv format
Data exploration and data pre procesing
Split the dataset into training and testing set
Creating the model
Compile and evaluate the model
Comparing different models for accuracy
Final output

Fig 1: Block Diagram of Proposed methodology

The initial step is data exploration and pre-processing which includes methods such as Label Encoder and normalisation. Label Encoder is an efficient tool for encoding the levels of the categorical features into numeric values. All the categorical features are encoded. In this paper, malignant and benign values have been classified as 0 and 1. In the Normalizer Method, the values of all the attributes are rescaled in the range of 0 to 1. The formula in equation (1) is used for this purpose.

$$\frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}} \tag{1}$$

Pre-processing is followed by splitting of data into train and test sets for the creation of models. 75% of the data has been used for training and the remaining 25% for testing. Various Machine Learning algorithms such as Logistic Regression, KNN, and SVM etc. have been applied to create models for predicting cancer. [11]. In the dataset used in this project, the outcome can be classified into two values, namely, M (malignant) or B (benign).

K-Nearest Neighbour is a supervised machine learning algorithm because the data given to it is labelled. The test data points classifications depends upon the nearest training data points instead of considering the parameters of the dataset [12]. SVM is also a supervised machine learning algorithm which is used as a training algorithm to study classification and regression rules from data [13]. Random forest algorithm has been applied next on the dataset. This algorithm creates decision trees on data samples, gets the prediction from each of them and finally selects the best solution by the means of voting. The Decision tree technique has also been applied on the data. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. The naïve Bayes classifiers were applied next, which are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

The accuracy achieved after applying these methods is not high enough and hence deep learning techniques such as CNN and ANN algorithms have been used. A Convolution Neural Network can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other [14]. The final algorithm used is ANN. Artificial Neural Networks are widely used in science and information technology due to their notable properties including parallelism, distributed storage, and adaptive self-learning capability. They have also been utilized to solve biomedical problems, especially in the areas of classification and prediction [15].

## IV. IMPLEMENTATION

To implement the project, recent papers on breast cancer prediction were studied. Many deep learning and machine learning techniques are available for cancer detection and prediction. The pre-processing techniques such as Label Encoder and Normalizer Method were implemented to handle the data efficiently. The bar graph shown in Fig 2 depicts the total number of benign and malignant cells in the dataset obtained after applying the Label Encoder method.

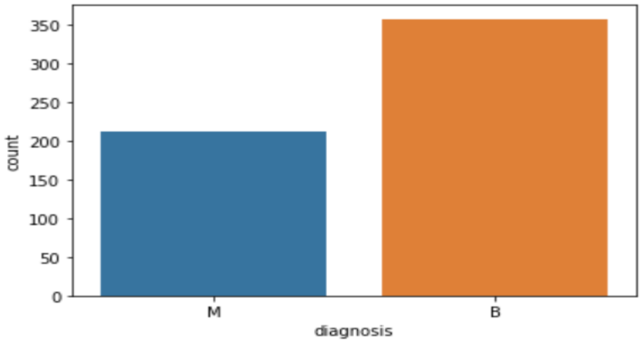

Fig 2: Number of malignant and benign

The Machine learning algorithms used in the project are Logistic Regression, Random Forest, K-Nearest Neighbour, Naive Bayes, Decision Tree, and Support Vector Machine. The Deep learning algorithms that can be used for predicting breast cancer are Artificial Neural Network, Convolutional Neural Network and Recurrent Neural Network. In this project, CNN and ANN models of Deep Learning have been implemented. The table 1.1 and 1.2 explain the parameters used for both CNN as well as ANN models. The various parameters used are number of neurons, number of input, number of epochs for which the model was trained and activation function.

Table 1.1 Parameters used in CNN model

| | |
|---|---|
| Number of Neurons | Con Layer1- 36<br>Con Layer2- 64 |
| Number of Input | 30 |
| Number of epochs | 50 |
| Activation Function | ReLU, Sigmoid |

Table 1.2 Parameters used in ANN model

| | |
|---|---|
| Number of Neurons | 15 |
| Number of Input | 30 |
| Number of epochs | 50 |
| Activation Function | ReLU, Sigmoid |

## V. RESULTS AND DISCUSSION

Various machine learning such as K Nearest Neighbour(KNN), Support Vector Machine(SVM), Decision tree, Naïve Bayes Logistic Regression, Random Forest were used for predicting breast cancer on the Wisconsin dataset. The maximum accuracy achieved was 96.5%, which was given by SVM and Random Forest algorithms. In order to increase the prediction accuracy, deep learning algorithms such as Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) were implemented.

The Fig 3 and Fig 4 below show the model accuracy and loss in graphical format with respect to the number of epochs for ANN model. As the number of epochs increases, the accuracy of the model increases while the loss decreases.

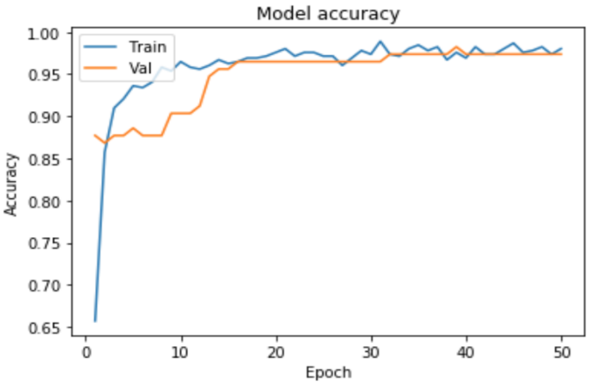

Fig 3: Graph plot for Model Accuracy

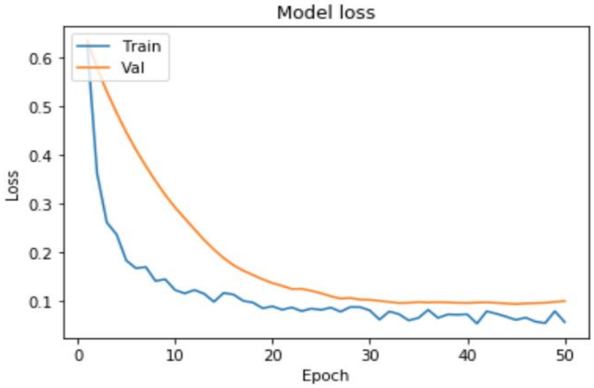

Fig 4: Graph plot for Model Loss

The accuracy obtained in case of CNN model and ANN model were 97.3% and 99.3% respectively, which was more efficient Compared to the above machine learning algorithms. Deep learning proved to be more efficient due to the use of activation functions such as ReLu and sigmoid. Using Activation function, it was possible to obtain the result in terms of probability rather than in Machine Learning algorithms, where the results were simply given in two labels, namely, 0 (for benign) and 1(for malignant). The table 1.3 shows the values obtained using the different algorithms.

Table 1.3: Comparison of ML and DL algorithm

| Algorithm | Accuracy | Precision | Sensitivity |
|---|---|---|---|
| KNN | 0.95 | 0.95 | 0.99 |
| SVM | 0.96 | 0.98 | 0.97 |
| Decision tree | 0.95 | 0.99 | 0.93 |
| Naïve Bayes | 0.92 | 0.93 | 0.94 |
| Logistic Regression | 0.94 | 0.96 | 0.96 |
| Random Forest | 0.96 | 0.98 | 0.97 |
| CNN | 0.97 | 0.97 | 0.98 |
| ANN | 0.99 | 0.99 | 0.99 |

## VI. CONCLUSION AND FUTURE SCOPE

In this paper, various machine learning as well as deep learning algorithms are implemented and their accuracy is compared. Machine Learning algorithms are giving maximum 96.5% accuracy, which was given by SVM and Random Forest Algorithms. In the case of Deep Learning, CNN and ANN were implemented. The accuracy given by CNN is 97.3% and that by ANN is 99.3%. The paper concludes that deep learning models give better accuracy compared to machine Learning Algorithms. Also the output is predictable in terms of probability in deep learning using Activation functions, which was not possible with machine learning algorithms.

In future, these techniques may be implemented on datasets that consist of images. The system may also be integrated with an application or website. The accuracy of the model created may be increased in order to give better predictions

# REFERENCES

[1] World Health Organization. Accessed on: Feb 13, 2020. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu,"Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.

[3] Dongdong Sun, M.Wang, H. Feng and Ao Li , "Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction", 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)

[4] D. Selvathi and A Aarthypoornila, "Performance analysis of various classifiers on deep learning network for breast cancer detection", International Conference on Signal Processing and Communication (ICSPC)

[5] Tiancheng He, M. Puppala, R. Ogunti, J.J. Mancuso, Xiaohui Yu, J. C. Chang, T. A.Patel and S.T. C. Wong, "Deep learning analytics for diagnostic support of breast cancer disease management", 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)

[6] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm", 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).

[7] Ch. Shravya, K. Pravalika, Sk. Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", 2019 International Journal of Innovative Technology and Exploring Engineering.

[8] J. Ferlay, C. Héry, P. Autier, and R. Sankaranarayanan, "Global burden of breast cancer," in Breast cancer epidemiology, ed: Springer, 2010, pp. 119.

[9] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, et al., "A gene-expression signature as a predictor of survival in breast cancer," New England Journal of Medicine, vol. 347, pp. 19992009, 2002.

[10] Breast Cancer Wisconsin Dataset, Kaggle. Accessed on: Feb 13, 2020. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[11] Chao-Ying, Joanne, Peng Kuk Lida Lee, Gary M. Ingersoll – "An Introduction to Logistic Regression Analysis and Reporting ", September/October 2002

[12] Mohammad Bol and raftar and Sadegh Bafandeh Imandoust - "Application of K-Nearest Neighbour (KNN) Approach for Predicting Economic Events: Theoretical Background"- International Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013

[13] Ebrahim Edriss Ebrahim Ali1 , Wu Zhi Feng2- "Breast Cancer Classification using Support Vector Machine and Neural Network"– International Journal of Science and Research(IJSR) , 3March 2016

[14] Mr. Madhan S, Priyadarshini P, Brindha C, Bairavi B, "Predicting Breast Cancer using Convolutional Neural Network", SSRG International Journal of Computer Science and Engineering (SSRG – IJCSE ) – Special Issue ICMR Mar 2019

[15] Ismail Saritas, "Prediction of Breast Cancer Using Artificial Neural Networks", Article in Journal of Medical Systems 36(5):2901-7, August 2011