American Journal of Electronics & Communication

Swapnamoy Bhattacharjee et. al., American Journal of Electronics & Communication, Vol. IV (2), 19-22

An Improved Approach on Breast Cancer Detection Using Machine Learning

Swapnamoy Bhattacharjee, Shreetanu Banerjee
Swapnil Panigrahi
Soujanya Bose and Souvick Das
Department of Computer Science & Application
Institute of Engineering & Management
Salt Lake, Sector - V
swapnomoy.2002@gmail.com

Soham Goswami and Saikat Mondal

Assistant Professor at Department of Computer Science & Application
Institute of Engineering & Management
Salt Lake, Sector - V
{soham.goswami & saikat.mondal}@iem.edu.in

***Abstract* - Cancer begins when changes called mutations take place in genes that regulate cell growth. The cells can expand and divide uncontrollably thanks to the mutations. The type of cancer that arises in breast cells is called breast cancer. Generally, breast ducts or lobules are where breast cancer first appears. The ducts that convey the milk from the glands to the nipple are where the milk is created by lobules. Moreover, cancer can develop in the breast's fatty tissue or fibrous connective tissue. Unchecked cancer cells can travel to the lymph nodes under the arms and frequently invade nearby healthy breast tissue. After the cancer has reached the lymph nodes, it has a pathway to spread to other organs, parts of the body.**

As per a 2013 WHO study, "it is projected that more than 508,000 ladies passed away all around the world in 2011 because of bosom disease". Early breast cancer development may be treated and prevented. Nonetheless, a lot of women receive a malignant tumor diagnosis after it has advanced past the point of no return.

The objective of this paper is to present several approaches to investigate the application of multiple algorithms based on Machine Learning for early breast cancer detection.

Index Terms* - *Breast Cancer, Dataset, Random Forest, Logistic Regression, Machine Learning.

I. Introduction

Cancer is the most prominent cause of fatalities around the world, according over one crore deaths in the past one year out of which 22.6% deaths were due to breast cancer. It is the most common type of cancer among women, accounting to 14.7% of cancer cases in India. Early detection happens to be a fruitful way to control breast cancer. There are ample cases that are handled by the early detection and decrease the mortality rate. The most common as well as efficient technique that is used in the field is Machine Learning in this report we specifically used Logistic Regression and Random Forest Classifier.

II. Literature Preview

The related research on machine learning-based breast cancer diagnosis that has been done in the past is covered in this section.

Arpita Joshi and Dr. Ashish Mehta [4] compared the classification outcomes obtained using KNN, SVM, Random Forest, and Decision Tree approaches (Recursive Partitioning and Conditional Inference Tree). Wisconsin Breast Cancer dataset from UCI repository was the one used. The best classifier, according to the simulation results, was KNN, followed by SVM, Random Forest, and Decision Tree.

Using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset, David A. Omondiagbe, Shanmugam Veeramani, and Amandeep S. Sidhu [5] studied the effectiveness of Support Vector Machine, Artificial Neural Network, and Nave Bayes by integrating these machine learning approaches with feature selection/feature extraction methods to find the best suited one. As a result of its higher computational time, SVM-LDA was preferred above all the other approaches, according to the simulation outcomes.

For better dataset processing, Kalyani Wadkar, Prashant Pathak, and Nikhil Wagh [6] conducted a comparative research on ANN and SVM which included multiple classifiers like KNN, CNN, and Inception V3. According to the experimental outcomes and performance analyses, ANN performed more efficiently than SVM, making it a better classifier.

Using machine learning techniques such as the Naive Bayes classifier, SVM classifier, bi-clustering Ada Boost algorithms (HA-BiRNN), RCNN classifier, and bidirectional recurrent neural networks, Anji Reddy Vaka, Badal Soni, and Sudheer Reddy K. [7] proposed a novel method to identify breast cancer. The proposed methodology (Deep Neural Network with Support Value) and machine learning techniques were compared, and the simulated results showed that the DNN algorithm was better in terms of performance, efficiency, and image quality, factors that are critical in today's medical systems, while the other techniques didn't work as expected.

By combining Deep Learning, Artificial Neural Network, Convolutional Neural Network, and Recurrent Neural Network

approaches with Machine Learning techniques including Logistic Regression, Random Forest, K-Nearest Neighbor, Decision Tree, Support Vector Machine, and Naive Bayes Classifier, Monica Tiwari, Rashi Bharuka, Praditi Shah, and Reena Lokare [8] have developed a novel method to diagnose breast cancer. According to a comparison of machine learning and deep learning techniques, the accuracy achieved by ANN and CNN models (99.3% and 97.3%, respectively) was higher than that of the machine learning models.

On the Wisconsin Breast Cancer (original) datasets, K.Anastraj, Dr. T. Chakravarthy, and K. Sriram [9] conducted a comparative analysis between different machine learning algorithms: back propagation network, artificial neural network (ANN), convolutional neural network (CNN), and support vector machine (SVM). For feature extraction and analysis of benign and malignant tumours, ALEXNET was utilised in conjunction with deep and convolutional neural networks. According to the simulation results, support vector machine is the best strategy and has produced superior outcomes (94%).

According to S. Vasundhara, B.V. Kiranmayee, and Chalumuru Suresh's [10] proposal, mammography pictures can be automatically classified as benign, malignant, or normal utilising a variety of machine learning techniques. Support Vector Machines, Convolutional Neural Networks, and Random Forest are compared and contrasted. The simulation results showed that CNN produces intuitive classification of digital mammograms utilising filtering and morphological procedures, making it the best classifier.

The dataset from Dr. William H. Walberg of the University of Wisconsin Hospital was used by Muhammet Fatih Ak [11]. This dataset was subjected to data visualisation and machine learning methods such as logistic regression, k-nearest neighbours, support vector machine, naive bayes, decision tree, random forest, and rotation forest. These machine learning methods and visualisation were implemented using R, Minitab, and Python. All the techniques were compared in a comparative analysis. The best classification accuracy (98.1%) was obtained using the logistic regression model with all features included, and the suggested method demonstrated improved accuracy performances.

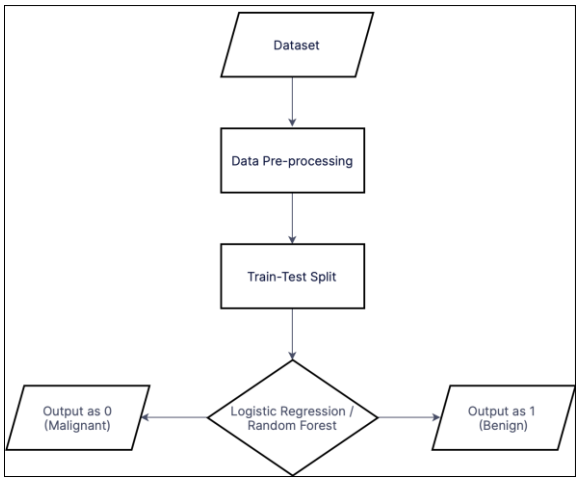
III. Objective

The objective of this model is to find the best features for the process of detecting breast cancer using machine learning, and also to find the effectiveness of the model where the values of the features could be same for both malignant case and benign Case.

IV. Libraries & Dataset

We Imported modules that are needed (Sklearn {train_test split, datasets, logistic regression}, pandas, numpy and we also imported Matplotlib and Seaborn for purpose of data representation. We used the Breast Cancer Wisconsin (Diagnostic) Data Set from sklearn, which will be reffered to as the “dataset.” This database is also available in the UW CS ftp server and kaggle. The dataset contains 31 features 'mean radius,' 'mean texture,' 'mean perimeter,' 'mean area,' 'mean smoothness,' 'mean compactness,' 'mean concavity,' 'mean concave points,' 'mean symmetry,' 'mean fractional dimension,' 'radius error,' 'texture error,' 'perimeter error,' 'area error,' ‘smoothness error,' 'compactness error,' 'concavity error,' 'concave points error,' 'symmetry error,' 'fractal dimension error,' 'worst radius,' 'worst texture,' ‘worst perimeter,' 'worst area,' 'worst smoothness,' 'worst compactness,' 'worst concavity,' 'worst concave points,' 'worst symmetry,' 'worst fractal dimension,' & 'target' The dataset from sklearn will be in numpy array format. In this paper we propose a new method to check the accuracy of the machine learning models.

V. Methodology



Then we are converting the numpy data to data frame (df) using pandas. The “target” feature contains the data if whether the case is malignant of benign represented as 0 & 1 (0 for malignant, 1 for benign) we changed the feature name from ‘target’ to ‘label.’ As every value except for ‘label’ is in float64 and ‘label’ is int64.

1) *Checking for missing values*: We checked for any null values or missing values in the columns.
2) *Find instances of the target feature*: Then we are counting how many instances of 0’s are there and how many instances of 1’s are there in total.
3) *Splitting:* Then we will split the input features and the label.
4) *Input and Test feature:* All the 30 columns except label will be the input features and label test feature.
5) *Creating X & Y:* The input features are taken as X and label which is the target feature is taken as Y.
6) *Train test splitting:* Then we are creating four arrays as x_train, x_test, y_train, y_test. Then we used the train_test_split () function. By that we are splitting the x array into two parts and the y array into two parts.test_size=0.2 means 80% of the dataset will be used for training and 20% will be used for testing.

Random_state will be 2 which means we are randomizing the dataset. We can write any number instead of 2(like 42).

7) *Fitting into Logistic Regression:* Then we are fitting the x_train and y_train data into logistic regression. Then we first give the training data in the model and find the accuracy of the training data, then we give the test data to find the accuracy in the test data.
8) *Fitting into RandomForestClassifier:* We again fit the X_train and Y_train in the RandomForestClassifier prediction model. Then we checked its accuracy on the test data.
9) *Finding feature priority and feature dependence:* Next we found the priority of different features in the dataset out hundred percent. Next we need to know the co-relation of every feature with other feature. For that we used two methods, 'spearman' & 'pearson' method.
10) *Feature selection:* For feature selecting we used the feature_importances_ method from which we manually selected most needed features. This process will reduce the number features from 31 to 7.
11) *Splitting and fitting with selected features*: Then again did train test splitting with the new selected and reduced number of features. 80% of the data was used as training and the rest 20% was used as test data.
12) *Fitting the data into prediction models:* We fitted the new training and test data into LogisticRegression & RandomForestClassifier Model. We again find out the accuracy of the model with reduced number of features.
13) *Finding out overlapping values affecting benign and malignant cases:* Here we plotted a graph to find out which values of the selected features were possible for both malignant case and a benign case and finding out how many such instances were there in the dataset.
14) *Finding accuracy with the overlapped values:* We found out every overlapping values of every selected feature in the dataset and the used that in the prediction models and then checked their accuracy.

IV. Observations

The dataset didn't contain any null or missing values. There were 357 benign cases or 0's & 212 malignant cases or 1's. After the first train test splitting the training data contained 455 rows and the test data contained 114 rows.

These are the following accuracies for the prediction models:

LogisticRegression: 92.10%

RandomForestClassifier: 94.73%

After finding the feature priority we can see that 'worst area' has the highest priority of 15.07% among all the features followed by 'worst perimeter' with priority of 13.67% among all the features.

After feature selection there were 7 features selected for the next process those were 'mean concavity,' 'mean concave points,' 'area error,' 'worst radius,' 'worst perime ter,' 'worst area,' 'worst concave points' and the target feature of 'label.'

After splitting and fitting these features to the prediction model we get the following accuracies:

LogisticRegression*:* 91.22%

RandomForestClassifier; 93.85%

As we can see after feature selection the accuracy of the models are reduced by approx. 1%.

After we plotted the count of malignant and benign cases with respect to the values of the features.

We can also see some overlapping of values. Where for the same value a case can be benign or malignant. Such as for mean concavity the values between 0.020000 & 0.15000 have cases of both benign & malignant, and for 'worst area' the overlapping range is between 490 & 1250. Like this we found the overlapping values for all 7 features and put them in a dataset and used them as test data for our prediction models.

These are accuracies after using overlapping values as test data:

LogisticRegression*:* 80.76%

RandomForestClassifier: 95.38%

Here we found something interesting, the accuracy of LogisticRegression reduced significantly due to a hard to predict data whereas the accuracy of RandomForestClassifier has increased. We then gave manual input to both the models to predict if a case is malignant or benign and both the models were able to predict it successfully.

The reason for the increased accuracy could be that the RandomForestClassifier is overfitting. For that we propose to give more emphasis on developing algorithms on models like Logistic Regression so that new and hard to predict data can be predicted with more accuracy, for that use image processing along with these prediction could also help to make the system more comprehensive.

Serial No.	Comparisons between the models with different values & features		
	Model name	Accuracy	Feature type
1	Logistic Regression	92.10%	Data before feature selection.
2	Random Forest Classifier	94.73%	Data before feature selection.
3	Logistic Regression	91.22%	Data after feature selection.
4	Random Forest Classifier	93.85%	Data after feature selection
5	Logistic Regression	80.76%	Selected features with overlapping values that appear in both malignant & benign cases
6	Random Forest Classifier	95.38%	Selected features with overlapping values that appear in both malignant & benign cases

V. Future scope

With the breast cancer diagnostic tool we are looking forward in the future by reducing the price of cancer detection to some extent. It will be easy to use detection tool which everyone can use and run even in low budget devices. If the tools that uses A.I. & Machine Learning are implemented in the medical sector, it can be used in various hospitals and in homes without undergoing a lot of hassles. Most the time cancer becomes deadly and incurable because it's not diagnosed properly in the first place, but with AI tool one can expect reliability and accuracy in the future which will help in detecting cancer cells in the first place and the patient can start treatment as soon as possible.

VI. Conclusion

The most common cause of death for women is breast cancer, a condition that can be fatal to female patients. Breast cancer, which accounts for 23% of all cancer deaths in postmenopausal women, is one of the most common malignant diseases overall. Though A.I. & machine learning has come a long way in predicting these diseases

More data & new features will be required to make this process faster and less expensive and also to make the model more accurate. Breast cancer detection and screening have improved as a result of increased public attention, breast cancer awareness, and advancements in breast imaging has also made a positive impact on recognition and screening of breast cancer.

Through these methods of prediction and implementation of A.I. in medical space, we can help many to get the treatment for cancer at the right time.

On this basis, have presented this model, by which breast cancer can be recognized using machine learning algorithms.

Acknowledgment

We would like to express our special thanks of gratitude to our Guide Saikat Mondal who helped us a lot in this project, his valuable suggestions helped us to solve tough challenges and without his help this report could not have been completed in time. A special thanks to our Head of Department Prof. Abhishek Bhattacharya who gave us the golden opportunity to do this wonderful report on the topic **"**A Report on Breast Cancer Detection Model Using Machine Learning**"**, which helped us to gain a significant knowledge in the aforesaid subjects. Secondly, we would like to thank our friends who helped us a lot in finalizing this report.

Refernces

[1] Kumar Sanjeev Priyanka, "A Review Paper on Breast Cancer Detection Using Deep Learning," 2021 IOP Conf. Ser.: Mater. Sci. Eng.

[2] Sarthak Vyas,Abhinav Chauhan, Deepak Rana,Mohd Noman, "Breast Cancer Detection Using Machine Learning Techniques," Researchgate, 2022

[3] Yash Amethiya, Prince Pipariya, Shlok Patel, Manan Shah, "Comparative analysis of breast cancer detection using machine learning ad biosensors," Intelligent Machine, voL 2, pp. 69-81, May 2022 .

[4] Arpita Joshi and Dr. Ashish Mehta "Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer," 2017.

[5] David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu "Machine Learning Classification Techniques for Breast CancerDiagnosis," 2019.

[6] Kalyani Wadkar, Prashant Pathak and Nikhil Wagh "Breast Cancer Detection Using ANN Network and Performance Analysis with SVM," 2019.

[7] Anji Reddy Vaka, Badal Soni and Sudheer Reddy "Breast Cancer Detection by Leveraging Machine Learning," 2020.

[8] Monika Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare "Breast Cancer Prediction using Deep learning and Machine Learning Techniques".

[9] K.Anastraj, Dr.T.Chakravarthy and K.Sriram," Breast Cancer detection either Benign Or Malignant Tumor using Deep Convolutional Neural Network With Machine Learning Techniques," 2019.

[10] S.Vasundhara, B.V. Kiranmayee and Chalumuru Suresh "Machine Learning Approach for Breast Cancer Prediction," 2019.

[11] Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," 2020.

[12] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning," 2019.

[13] Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime, Thomas Noeld "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," 2016.

[14] Dana Bazazeh and Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis," 2016.

[15] Vishabh Goel, "Building a Simple Machine Learning Model on Breast Cancer Data," Towards Data Science, Sep 29, 2018.