Breast Cancer Classification Using Machine Learning

Mayank Pathak
Information Technology
Galgotia's College Of Engineering and Technology
mayankpathak870@gmail.com

Nikhil Kumar
Information Technology
Galgotia's College Of Engineering and Technology
ayaansingh585@gmail.com

Lakshay Vardhan
Information Technology
Galgotia's College of Engineering and Technology
galgo.lakshay@gmail.com

K Rajkumar
Assistance Professor
Information Technology
Galgotia's College Of Engineering and Technology

Abstract--

Breast cancer may be a heterogeneous disease, commonly known to begin as neighborhood lesion within the breast, and so spread gradually. The second leading cause of death among women is cancer. In 2020 within the US there'll be 2,79,100 new cases and 42.690 estimated deaths. Together with 48,530 new cases of noninvasive (in-situ) cancer approximately every one-fifth of the affected women will die because of this disease [1]. This (breast) dual organ inside a woman makes significant changes in size, shape and function in conjunction with pregnancy, lactation, puberty and postpartum so that a protective pregnancy should take place before the age of 30. When the breast grow abnormally it leads to breast cancer. These cells form a lump or mass because they divide quicker than healthy cells do and accumulate [2].
The lymph nodes and other part of the body can also be spreaded by the cells of the breast which is dangerous. Cells can often be seen from X-Ray or felt as a lump. Great information is obtainable so analyzing this huge amount of knowledge to extract the novel and usable information or knowledge is incredibly complicated and time-consuming task the best technique for this is data mining. There is class imbalance within the data. Since the probability of having the disease is much more than not having the disease. The goal of this research paper is to differentiate Malignant and Benign patients. The dataset which we have used here is Wisconsin Dataset (WBC).

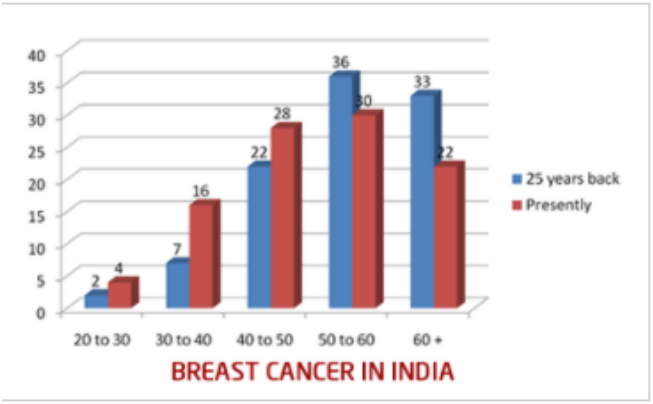
Keywords-- Machine Learning; Breast Cancer; Benign; Malignant;

I. INTRODUCTION

Breast Cancer – A cancer that is formed in the cell of the breast. Breast cancer can occur in women and in men but this occurs rarely in case of men. Symptoms of breast cancer includes
1. There is a lump in breast and underarm (armpit).
2. Breast is thickened or swelled.
3. Irritation or dimpling of breast skin.
4. The nipple area or the breast has redness or flaky skin.
5. the nipple area suffers pain or pulling out of nipple**.**

And many more are the symptoms of breast cancer. Its treatment depends on the stage of the cancer. It's may consist of, radiation, hormone therapy, chemotherapy and surgery. These are the stats showing the number of cases presently and also comparing it with the stats of 25 years ago. The X-axis which is shown by the horizontal line below the bars represents the various age categories: 20 to 30 years, 30 to 40years ,40 to 50 years and so on the percentage of cases in India is shown in the vertical line and it is in the Y-axis. The blue bar represents the cases 25 years earlier, and maroon colored bar represents the today's situation. Earlier out of every 100 patients of breast cancer. The percentage of patients between the age of 20 to 30 years is 2%, between 30 and 40 it is 7%. The patients who are above 50 years age are the most i.e., 69%. Comparing it with current scenario 4% of the patient are between 20 to 30 years age, 30 to 40 ages are 16% and 28% are in 40 to 50% of range. This makes 48 % of patients are below 50.
Earlier before 2008, the cancer of cervix was the most common cancer in women but after 2008 due to technological advancement and awareness of the problem breast cancer is the most common and cervix cancer is still common in rural areas, he. So, the best way of getting treated is by early diagnosis [3].

Figure 1 Breast cancer in different ages



DATASET

We have taken the dataset from Wisconsin Breast Cancer dataset (ORIGINAL) which is taken from UCL machine learning repository. There are approximately 699 instances in the data. There are 10 real valued attributes. The dataset is divided into 2 classes-

1. Benign
2. Malignant

All the features are taken from digitalized image of Fine Needle Aspirate (FNA) of Breast mass. All the characters tics are of the nuclei [4]

Attribute Information:

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

.

II. MACHINE LEARNING

Machine learning is the study of computer algorithms and the use of information, and it is considered a branch of computer science. Machine learning is critical because it allows businesses to see trends in customer behavior and operational patterns. It also assists in the construction of products on the edge. Many of today's leading companies like IBM, YouTube, Uber, Google make machine learning an integral part of their operations. Machine learning has become a major competitive divider for several companies.

There are basically three types of Machine learning methods and they are as follows:

1. Supervised Learning: Supervised learning is a form of learning in which machines are trained using labeled data for training. On the basis of data, the output is predicted by the machine. Some real-world examples of supervised learning algorithm are Credit card fraud detection, Image classification. In supervised learning the dataset is splitted between training dataset, validation and test dataset.

2. Unsupervised Learning: Unsupervised learning is a form of learning in which models can be monitored using a training database. Unsupervised learning is like learning of human brain it groups the data according to the similarities we have only the input data set which is not labelled It is helpful in finding the insights from the data

3.Semi-supervised learning: Unsupervised learning (without any labelled training data and desired output) and supervised learning (with labelled training data and desired output) are the two types of learning (with completely labelled training data and both input and output provided in it). Although certain training examples in training data lack training labels, many machine-learning researchers have discovered that unlabeled data, when combined with a little a priori knowledge, can be useful. Amount of labeled data, can produce a more improved or desired output and more accuracy in learning technique.

4. Reinforcement learning: Reinforcement learning is a part of machine learning which checks how software agents will react or take actions in an environment so on maximize some accuracy of the output close to the user can get the required output. Learning Machines have seen the use of cases ranging from predicting customer behavior to building a self-driving car package.
When it comes to benefits, learning about technology can help businesses understand their customers at a deeper level. By collecting customer data and associating it with behaviors over time, machine learning algorithms can learn organizations and help teams plan development and marketing programs on customer needs.

SUPPORT VECTOR MACHINES(SVM)

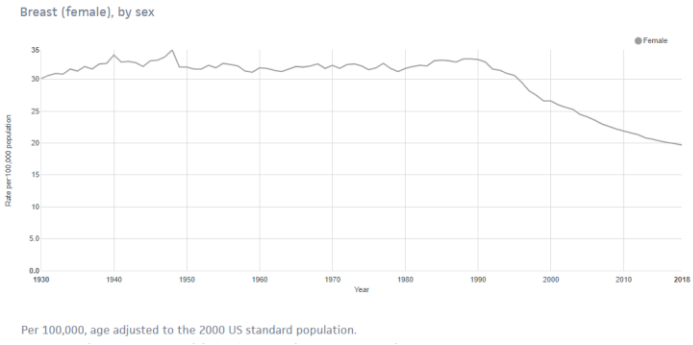
The main objective of using Support Vector Machines is that it creates the best decision boundary that separates the working space into classes so we can easily differentiate in future if we put some new data in it. Hyperplane is known as the best decision boundary. SVM are supervised learning models that uses both the classification and regression technique. It was developed by Vladimir Vapnik at AT&T Bell Laboratories in 1963. Support Vector Machines are one of the most efficient and robust technique which after training the data again can reduce the line of separation between the classes.SVM are classified in two categories that is Linear classification and Non-Linear Classification. SVM became popular in the late 90's because it was able to handle a lot of continuous and phase flexibility. In SVM we must widen the gap between the two categories. In the future the new model will be marked in the same space and predict which side they fall on.

k-NEAREST NEIGHBOURS ALGORITHM(k-NN)

K Nearest Neighbor is a Supervised Learning Algorithm and is one of the simplest machine learning algorithms. KNN

algorithm classifies the data based on the categories and then the algorithm assumes from the available data and put the new cases into the category which is similar to it. The algorithm can be learned by some techniques that is Instance based learning, Lazy Learning Model and Non-Parametric method. KNN also requires data preparation by data scaling, reducing dimensionality and treatment of missing value. Classification and Regression techniques both can be used in this method. KNN algorithm is very useful when the data is large16. There is no such method to determine the value of K but the best preferred value is 5. KNN algorithm is also known as lazy learner algorithm just because it does not train or learn itself in the training phase instead it performs action on the dataset at the time of classification. After saving the dataset KNN algorithm separates the new data into a similar data section [5].

Figure 2 Stats per 10000 in different years



DATA PREPROCESSING

In preprocessing phase, we are loading the data and check for the missing values if the values are missing in the dataset, then we will be working it by removing the values with negative values so that there is less error in accuracy part. The data available is so huge and missing data can lead to errors in the efficiency.

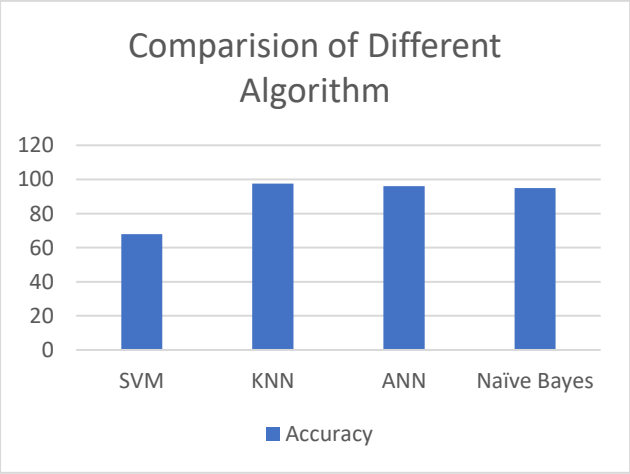
III. COMPARATIVE STUDY OF EXISTING CLASSIFICATION AND TECHNIQUES

TABLE 1 COMPARATIVE STUDY OF EXISTING CLASSIFICATION AND TECHNIQUES

S.NO	PAPER TITLE	ALGORITHM	DATASET	RESULT
1.	COMPARATIVE STUDY ON DIFFERENT CLASSIFICATION TECHNIQUES FOR BREAST CANCER DATASET [6]	J48, MLP Rough Set	BREAST CANCER DATASET	J48: - 79.97% MLP: -73.35% ROUGH SET: - 71.36%
2.	INTEGRATION OF DATA MINING CLASSIFICATION AND ESEMBLE LEARNING FOR PREDICTION THE TYPE OF BREAST CANCER RECURRENCE [7].	NB, SVM, GRNN and J48	UCI BREAST CANCER DATASET	GRNN & J48 accuracy: 91% NB & SVM: 89%
3.	BREAST CANCER PREDICTION VIA MACHINE LEARNING [8]	KNN, SVM, RANDOM FOREST, GRADIENT BOOSTING	BREAST CANCER WISCONSIN(DIAGNOSTIC) DATA	KNN, SVM, RANDOM FOREST, GRADIENT BOOSTING accuracy: 70%

4.	BREAST CANCER DETECTION IN DIGITAL MAMMOGRAMS [9]	MOMENT IN VARIATION+FRACTAL DIMENSION, REGION BASED, TAMURA FEATURES	MINI-MIAS DATA SET	MOMENT IN VARIATION+FRACTAL DIMENSION accuracy:96.92 REGION BASED accuracy:91% TAMURA FEATURES accuracy:78.6
5.	BREAST CANCER CLASSIFICATION AND DETECTION [10]	RANDOM FOREST	CT SCAN DATASET	RANDOM FOREST accuracy: 95%
6.	BREAST CANCER CLASSIFICATION USING DEEP LEARNING [11]	MULTILAYER PERCEPTRON ALGORITHM	BREAST CANCER DATASET	MULTILAYER PERCEPTRON ALGORITHM accuracy: 96.5%
7.	COMPARISION OF MACHINE LEARNING METHODS FOR BREAST CANCER DIAGNOSIS [12]	SUPPORT VECTOR MACHINE, ANN	WISCONSIN BREAST CANCER DATASET	SUPPORT VECTOR MACHINE accuracy: 96.9% ANN accuracy: 95.4%
8.	CLASSIFICATION OF BREAST CANCER DATA USING MACHINE LEARNING ALGORITHM [13]	KNN, SVM	WISCONSIN BREAST CANCER DATASET	KNN, SVM accuracy: 96%

IV.RESULT AND DISCUSSION



In this paper, we've applied SVM and KNN techniques for prediction of the classification of carcinoma to seek out which machine learning methods performance is best. Vector Machines (SVMs) Support Vectors were first described by Vladimir Vapnik and therefore the effectiveness of SVMs has been observed in many patterns of pattern detection. SVMs can show better separation performance compared to many other separation strategies[14]. SVM is one among the most popular machine learning classification technique that is used for the prognosis and diagnosis of cancer. According to SVM, the classes divided by hyperplane with support vectors are critical samples for all classes. The experimental results are demonstrated in Table 1 and Table 2

Table 1: The experimental results of KNN technique

KNN (Nearest Neighbor)	**The result of performance**		
	Precision	Recall	F1 Score
	0.98	0.97	0.98
	0.98	0.96	0.97
Accuracy	97.8571%		

Table 2: The experimental results of Support Vector Machine classification techniques

Support Vector Machine	**The result of performance**		
	Precision	Recall	F1 Score
2	0.68	1	0.81
4	0.00	0.00	0.00
Accuracy	67.8571%		

V. CONCLUSION

Women worldwide are suffering from breast cancer on large scale and it is the second largest in numbers of death among women. This paper introduces us using two algorithms Support vector Machines and KNN Algorithm in breast cancer classification. We are getting best accuracy in KNN method but both the systems are giving excellent accuracy. The accuracy in KNN method is 97.5%. If we increase the size of data set than the running time will also increase and we can achieve higher accuracy. Shuffling the training data again can fluctuate the accuracy very sharply. For future research we can use CADe and CADx technologies which take high resolution images and neural network can be used for detecting it faster[15]. Later we will be working on detection side to give more help in detection by just giving the data and directly getting result about cancer. In future we can use this classification technique in helping the patients and doctors to detect it more accurately. It can be done by automating the process and creating mobile application which can be more user friendly. By this we will reduce the manual error and can be very useful in medical sciences.

VI. ACKNOWLEDGEMENT

I would like to extend my special thanks to my research guide and panel members who encouraged me and supported me to complete the research paper. I would also thank to our College (Galgotia's College of Engineering and Technology) for providing me with all the facility that was required.

REFERENCES

1. https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html
2. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097-0029%2820010115%2952%3A2%3C204%3A%3AAID-JEMT1006%3E3.0.CO%3B2-F
3. https://www.breastcancerindia.net/statistics/trends.html
4. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
5. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML
6. Sayeth Saabith, Elankovan Sundararajan, Azuraliza Abu Bakar," Comparitive Study on Different Classification techniques for Breast Cancer Dataset, IJCSMC, Vol. 3, Issue. 10, October 2014, pg.185 – 191
7. Jesús Silva, Omar Bonerge Pineda Lezama, Noel Varela,Luz Adriana Borrero," Integration of Data Mining Classification and Esemble Learning For Prediction The Type Of Breast Cancer Recurrence", International Conference on Green, Pervasive, and Cloud Computing GPC 2019: Green, Pervasive, and Cloud Computing pp 18-30
8. Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, Dr. A. Sivasangari," Breast Cancer Prediction via Machine Learning", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART
9. Kanchan Lata Kashyap, Manish Kumar Bajpai, Pritee Khanna," Breast Cancer Detection in Digital Mammograms", 978-1-4799-8633-0/15/ ©2015 IEEE
10. Poonam Kathale, Snehal Thorat," Breast Cancer Detection and Classification", 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 78-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/
11. Jasmir, Siti Nurmaini, Reza Firsandaya Malik, Dodo Zaenal Abidin, Ahmad Zarkasi, Yesi Novaria Kunang, Firdaus," Breast Cancer Classification Using Deep Learning", International Conference on Electrical Engineering and Computer Science (ICECOS) 2018, 978-1-5386-5721-8/18
12. Ebru Aydındag Bayrak, Pınar Kırcı, Tolga Ensari," Comparison of Machine Learning Methods for Breast Cancer Diagnosis", 978-1-7281-1013-4/19/$31.00 ©2019 IEEE
13. Burak Akbugday," Classification of Breast Cancer Data Using Machine Learning Algorithms", 978-1-7281-2420-9/19/$31.00 ©2019 IEEE
14. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
15. https://link.springer.com/article/10.3758/s13414-016-1250-0
16. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning