Breast Cancer Detection Using Machine Learning

Nikita Chawla[1], Mahek Gangadia[1], Sonal Gehlot[1], Yogita Shelar[2]

[1]Dept. of Information Technology, Atharva College of Engineering, Maharashtra, India
[2]Assistant Professor, Dept. of Information Technology, Atharva College of Engineering, Maharashtra, India

---

Abstract - *The foremost often occurring cancer after lung cancer is Breast Cancer (BC), BC is the second most frequent cause of fatality in each developed and undeveloped world, BC is characterized by the mutation of genes, constant pain, changes within the size, color(redness), skin texture of breasts. Classification of BC leads pathologists to search out a scientific and objective prognostic, usually, the foremost frequent classification is binary (benign cancer/malign cancer). Today, Machine Learning (ML) techniques are being broadly speaking utilized in the breast cancer classification issue. They lay out high classification accuracy and effective diagnostic capabilities. In this paper, we present two different classifiers: The naive Bayes(NB) classifier and the k-nearest neighbor(KNN) for breast cancer classification. The Wisconsin Diagnostic Breast Cancer dataset has been used for training the model. The result was obtained in real-time.*

This system will have a doctor login whenever the doctor can record the case details along with the case history of the patient. The patient will also be given a login using which the patient can also access.

***Key Words*: Breast Cancer, k-Nearest-Neighbor, Naive Bayes, Benign Cancer, Malign Cancer**

1. INTRODUCTION

The most commonly occurring type of cancer is breast cancer. Cancer is a disease that occurs when there are changes or mutations that take place in genes that help in cell growth. It is known to affect over two million women all over the world. There are no prevention techniques for breast cancer but early detection and diagnosis are critical in determining the chances of survival. During the early detection stages of the disease, the symptoms are not presented well and hence diagnosis is delayed. Approximately fifteen percent of the total deaths resulted from all types of cancer among women.

Breast cancer is caused by a combination of factors, including family history, obesity, hormones, radiation therapy, and even reproductive factors. Breast cancer is caused by a mistake or mutation in a single cell, which can be shut down by the system or causes an uncontrolled cell division. If the problem is not resolved after a few months, masses of cells with incorrect instructions form. The only method of improving the results of breast cancer cases is early diagnosis and screening.

In this study, we tailor two machine learning techniques for breast cancer classification. Using the Wisconsin Breast cancer, the goal of this approach is for cancer classification using two classifiers in a data set. Each classification uses two classifiers in the data set. Each classifier's performance will be evaluated in terms of accuracy, training process, and testing process.

2. BACKGROUND

In this section, we discuss breast cancer classification, followed by different machine learning techniques used in our breast cancer classification.

A. Breast Cancer Classification (BCC)

We will evaluate and classify breast cancer in this machine learning project. Malignant type breast cancer and benign type breast cancer.

The main goal of this project is to develop a model using a dataset that can correctly classify whether breast cancer is malignant or benign. Breast cancer classification requires nine characteristics to establish a good prognosis: 1. Determine the layered structures **(Clump Thickness)**; 2. **Uniformity of cell size**; 3. **Uniformity of cell shape;** 4. Normal cells are connected to each other while cancer cells spread all over the organ **(Marginal Adhesion)**; 5. When epithelial cells are enlarged it is a sign of malignancy **(Single Epithelial cell size);** 6. Nuclei in benign tumors are not surrounded by cytoplasm**(Bare Nuclei)**; 7. Nucleus texture in benign cells is homogenous. In malignancies, the chromatin is coarser **(Bland Chromatin)**; 8. The nucleolus is normally inconspicuous and tiny in normal cells. There are more than one nucleoli in cancer cells, and they become much more conspicuous,**(Normal Nucleoli)**; 9. Estimate the number of mitoses that have occurred. The higher the number, the higher the risk of malignancy**(Mitosis)**; Even if one of the nine criteria is very large, the risk of malignancy requires all nine.

B. Machine Learning

Machine learning can be defined as a subset of artificial intelligence that teaches the ability to learn into a system on the basis of a data set used for the purpose of training. There are multiple approaches or techniques that can be taken to make the system learn. Some are neural networks, decision trees, and clustering.

Three machine learning techniques are

a) Supervised Learning: It creates a function that predicts results based on supplied data. The function is created using the training data and serves as a guide for the user. For fresh datasets, a mechanism to generate valuable epiphanies is being developed.

b) Unsupervised Learning: In this the machine is forced to learn from an unlabeled dataset, which is then differentiated on the basis of some characters, allowing the algorithm to respond to that information without external assistance.

c) Reinforcement Learning: The learning process is iterative and continues from the environment. The system finally learns all conceivable system states over a long period of time.

1)Naive Bayes

Naive Bayes classifiers are probabilistic classifiers that are based on the application of the Bayes theorem. A Bayesian method is fundamental in probability and statistics that can be defined as a framework for modeling decisions. It is naive because it assumes that all features are independent of each other, which is rarely the case in real-world scenarios, but Naive Bayes still proves to be effective for a wide range of machine learning problems. Variables in Naive Bayes are conditionally independent; Naive Bayes can be used to determine a model from data that directly influence each other.

2)k-Nearest Neighbors(KNN)

K can be thought of as a representation of the data points for training that is close to the test data point that we will use to find the class. A k-nearest-neighbor algorithm is an algorithm used to determine where a data set belongs based on the data sets around it. The method that is used for regression and classification. KNN gathers all the data points nearby to process a new data point. Attributes with a high degree of variation are important in determining the distance.

Given N training vectors in Figure 1, the kNN algorithm identifies the k-nearest neighbors regardless of labels.

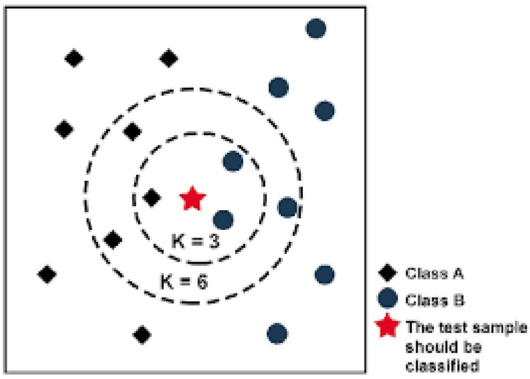


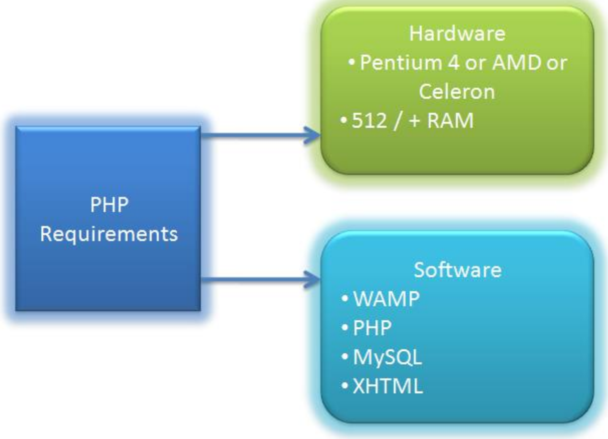
Figure 1: kNN Illustration

C. Web Development

In this section, we discuss the programming languages used for the doctor and patient login website.

1)PHP

PHP could be a server-side scripting language that's used for web development but may additionally be used for general-purpose programming. PHP is now present on over 244 million websites and more than two million web servers. The PHP group now produces the reference implementation of PHP, which was originally created by Rasmus Lerdorf in 1995. While PHP was originally an abbreviation for Private Home Page, it is now a recursive acronym for PHP: Hypertext Preprocessor.



PHP Requirements

Hardware
- Pentium 4 or AMD or Celeron
- 512 / + RAM

Software
- WAMP
- PHP
- MySQL
- XHTML

Figure 2: PHP

2)MySQL

MySQL, officially known as "My Sequel", is the world's most popular open-source relational database management system (RDBMS) that runs as a server and provides multi-user access to a variety of databases, though SQLite like;y has more total embedded developments. SQL is an abbreviation for Structured Query Language

3. RELATED WORK IN BREAST CANCER

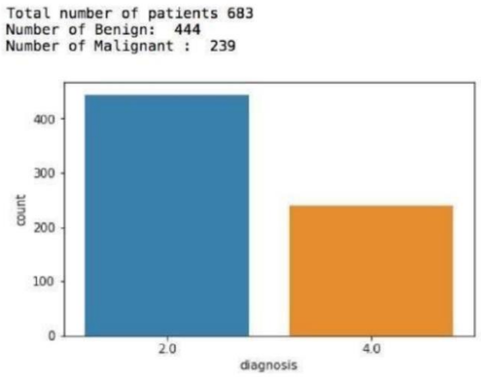
Many studies have been conducted in the field of BCC and MI; some of them used mammography images, which can miss about 15% of breast cancer; other techniques are more specific and use the genome or phenotypes to do classification. Several techniques, including SoftMax Discriminant Classifier(SDC), Linear Discriminant Analysis(LDA), and Fuzzy C Means Clustering, are used to classify breast cancer. The K Nearest Neighbors algorithm is a popular machine learning algorithm. Before classifying a new element, we must use a similarity measure to compare it to other elements. KNN can be used to assess the performance of false-positive rates in cancer classification. In general, naive Bayesian classifiers are used to predict biological, chemical, and physiological outcomes. To determine prognostics or classification models, NBC is sometimes combined with other classifiers such as decision trees in cancer classification.

Different classification techniques for breast cancer diagnosis were developed, and the accuracy of many of them was evaluated using data from the Wisconsin breast cancer database. For example, the optimized learning vector methods performance was 96.7%, the highest of any LVO method and SVM's accuracy for cancer diagnosis was 97.13%, the highest in the literature.

4. THE PROPOSED ALGORITHMS

A. *Datasets*

We used a Breast Cancer Dataset(BCD) that was donated to the University of California, Irvine(UIC). There are 11 attributes, the first of which is ID, which we will remove(it is not a feature we actually want to feed in our classification). The nine criteria are intended to determine whether a tumor is benign or malignant; the final feature contains a binary value(2 for benign tumor and 4 for malign tumor). There are 699 clinical cases in the set. Because the initial BCD contains missing data for 16 observations, our dataset is limited to 683 samples.



```
Total number of patients 683
Number of Benign:  444
Number of Malignant :  239
```

Figure 3: Wisconsin Breast Cancer Dataset

Figure 3 shows that 444(65%) of the tumor are benign, while 239(35%) are malignant.

B. *Breast Cancer Detection using the K-Nearest neighbors*

1)Algorithm:

1. Enter the dataset and divide it into two parts: training and testing.
2. Choose one of the instances from the testing sets and compute its distance from the training set.
3. Arrange the distances in ascending order.
4. The instance's class is the most common class of the three first pieces of training instances(k=3).

2)Description:

A sample of N examples and their classes is provided. We divided the data into stages for cross-validation and testing. In KNN, there is no training stage because we compare each new instance. To forecast the outcome of a new instance, we compute the Euclidean distance between the instance and every point in the training set.

C. *Breast Cancer Detection using the Naive Bayes Classifier*

1)Algorithm:

1. Divide the information into two blocks of two classes each, with two sets of features T and classes D.
2. Find the mean and standard deviation for every feature and class.

3. Create a summary for every feature and class.
4. Using the density of distribution, compute the probability of every feature.
5. Multiply the chances of all features to calculate the probability of every class.
6. Calculate the probability of every class to predict the category of an instance from the testing set.

2)Description:

We used the same Naive Bayes primitive in our algorithm, and we divided the dataset into testing and training sets first. The training phase begins by dividing the set into two distinct sets: D is the presence of the tumor and T is a set of features to be tested, and then the D set is divided into two classes: Malignant and Benign(4 or 2).

D. *Methodology*

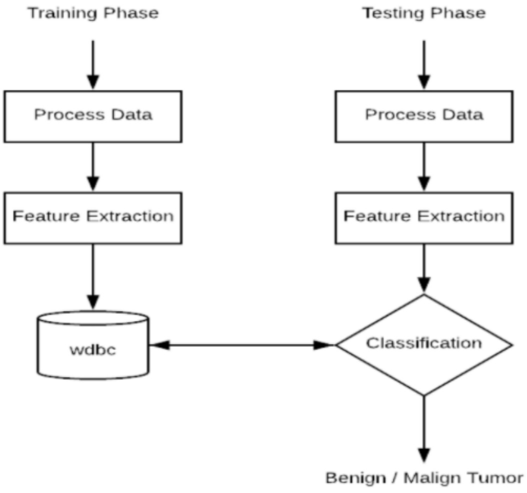


Figure 4: Proposed Model.

As shown in Figure 4, we separate the dataset into two sections for training and testing. Then the data is processed from which we extract 32 features such as radius, texture, parameter, area, and many more. Then those extracted features are classified by the Wisconsin dataset for breast cancer whether the tumor is benign or malignant.

As we classify more and more the dataset is also constantly improved.

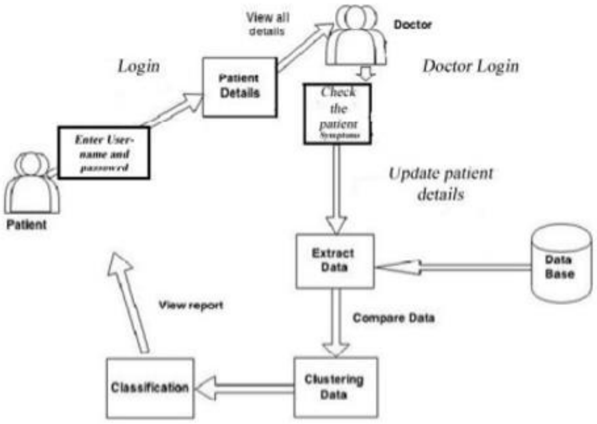


Figure 5: Proposed Web App

We also created a web app to make this model easily usable for doctors as well as patients.

It has a doctor login where the doctor can put the details regarding patient reports during every visit and check the result of it using the present dataset. The patient also has a login where they can view the reports.

5. IMPLEMENTATION AND RESULT ANALYSIS

KNN(k-Nearest-Neighbor) and Naive Bayes algorithms were used in a comparison study. It is implemented in a computer with an Intel Core i7 processor and 16GB of RAM. We used Python's open-source machine learning packages NumPy, pandas, and Scikit-learn. The program is executed using Jupyter Notebook, an open-source web application. The k fold cross-validation method was used to test the classifier. We used the ten-fold method, which is the most effective.

The data was divided into ten separate parts. Nine out of the ten are used for training and the last set is used for testing and analysis reasons. For the training set, we've got 398 observations while 171 observations were used for the testing set out of the 569 observations.

Table 1 - Comparison of Existing and Proposed Model

SCENARIO	EXISTING MODEL	PROPOSED MODEL
Operation	In-depth Study	A practical application that can be used in health care and personal use.

Accuracy	Static	Real-Time
Dataset	The dataset is restricted	The quality of the dataset improves due to its dynamic nature.

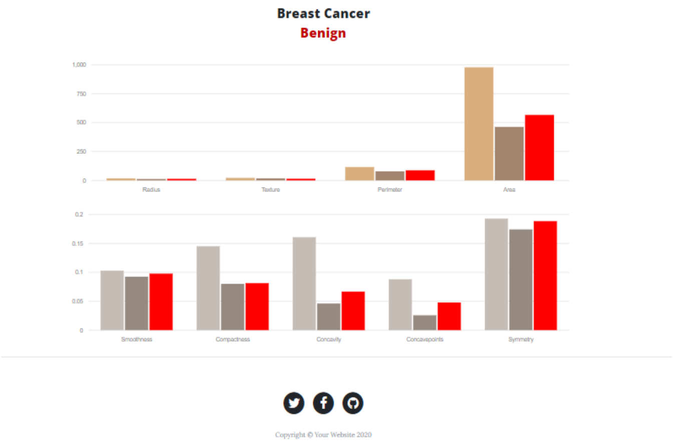


Figure 6: Result for Benign Tumor

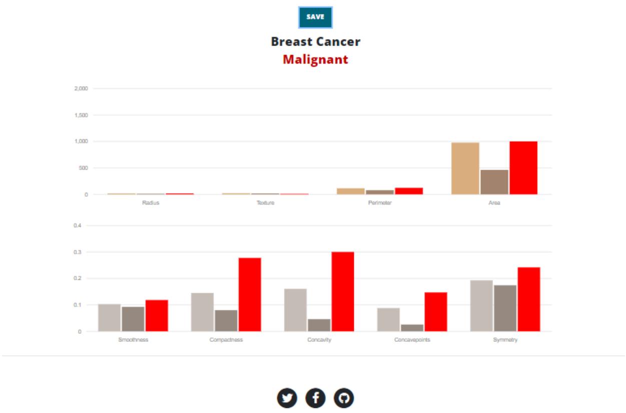


Figure 7: Result for Malignant Tumor

6. CONCLUSION

We learned to build a breast cancer tumor predictor on the Wisconsin dataset. The selection of appropriate algorithms with a good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from suitable databases like the Wisconsin dataset.

7. REFERENCES

1. S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, DOI: 10.1109/CTEMS.2018.8769187

2. Citation: Wu, J.; Hicks, C. Breast Cancer Type Classification Using Machine Learning. J. Pers. Med. 2021, 11, 61. https://doi.org/10.3390/jpm11020061.

3. P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICCSIT), 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.

4. Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier To cite this article: P R Anisha et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1116 012187.

5. Breast Cancer Classification Using Machine Learning Meriem AMRANE1 Ikram GAGAOUA3, Saliha OUKID2.

6. National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/

7. Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast. cancer" 2007.

8. T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm - International Journal of Advanced Research in Computer Science and Software Engineering, 2015.

9. A. Alarabeyyat, A.M., "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm", in 9th International Conference on. IEEE, v.i.e.E. (DeSE), pp. 35-39, 2016.

10. S.K. Prabhakar, H. Rajaguru, "Performance Analysis of Breast Cancer Classification with Softmax Discriminant Classifier and Linear Discriminant Analysis", In: Maglaveras N., Chouvarda I., de Carvalho P. (eds)Precision Medicine Powered by pHealth and Connected Health. IFMBE Proceedings, vol 66. Springer,

Singapore, 2018.

11. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data