Breast Cancer Prediction : A Big Data Approach

Nikita Joshi, Pooja Gunagi, Anurag Devagiri, Prof Leena Sakri

Paper Abstract - BC (Breast Cancer) has become one of the most prevalent cancers diagnosed in women. If BC is identified at an earlier stage, it can help save lives of those affected by Breast Cancer. The tumors are classified into two types - Malignant and Benign. The cancerous tumors are Malignant and the non-cancerous tumors are Benign. In the present scenario, it has become a challenging task to detect the occurrence of Breast Cancer tumors and the process takes several days to predict the same. By the use of proper algorithms pertaining to Big Data and Machine Learning this task can be achieved with much ease. The dataset is taken from UCI repository – Wisconsin data.

1. INTRODUCTION

Breast Cancer can occur both in Men and Women. But, it commonly occurs among Women. When growth of cells become uncontrollable, they form mass tissue called tumor. According to the reports, it has become the major cause for death among most of the affected women. There are two types of tumors. They are M (Malignant) and B (Benign). The Benign tumors do not spread to all body organs and can be completely unfasten without the fear of growing again. They are non-cancerous. But the malignant tumors are cancerous and there are high chances of their occurrence even after complete removal. These cells spread and attach to other body tissues to form new tumors that may cause an effect to entire functionality on the whole. In the process of detection of Breast Cancer, a patient has to undergo several tests such as surgical biopsy – operation. In the present scenario, many Big Data, Data Mining and Machine Learning techniques have made great contributions in the field of Healthcare. Early detection of a particular disease, Early medication to prevent further growth of those harmful cells will help save the lives of lots of affected people. The various supervised and unsupervised learning techniques have proved to be very useful and serve as one of the core parts for healthcare measures in this field.

2. LITERATURE SURVEY

2.1 Necessity for Research

Amongst all non-communicable diseases present, Cancer is one of the dangerous disease and there is a urgent and strict need for research in this domain. The symptoms of Cancer may not be clear at the initial stages and hence pose a great challenge for the medical practitioners and pathologists to predict the occurrence of tumor cells. Recently, in the year 2018 the reports from World Health Organisation(WHO) have shown that India has about 1 million cancer cases and more than 50% of these will be diagnosed in Women or precisely around 17000 more women will fall prey to this disease as compared to men. Therefore, because of all these challenges faced at every step of treatment, including the variations in the reports generated by one expert to another, by the manual identification of images, applying appropriate algorithms to detect the cancer can serve helpful in this domain due to its urgent need.

2.2 Stages of Breast Cancer

This disease basically has four stages: In earliest level, there are cancer cells only inside the inner lining of breast duct. This is called as ductal carcinoma in situ. In Level IA, the tumor is about 20mm. In Level IB, a smaller number of infected cells are found in lymph nodes. Level IIA consists of these cells found in about I to III lymph nodes under the arm. Level IIB, the tumor is larger than 20 mm but not more than 50 mm. Level IIIA consists of cancer cells found in about four to nine axillary nodes. Level IIIB and IIIC consists of tumor that has grown into the muscles of the wall of chest or inner layer of skin. In Level IV, the cancer has spread to almost all other organs of body called as metastatic cancer.

2.3 Signs and Symptoms

Cancer has become one of those precarious diseases across the globe and have become more prominent in women rather than men. Due to the absence of appropriate strategies and fast efficient standard solution for detection of cancer cells in the body, symptoms of breast cancer should be detected as early as possible.

Major signs may include the following:

- New lump starts growing in the breast or underarm.
- For some, it may lead to thickening or swelling or part of breast.
- Slight difference in nipple or breast skin, such as dimpling, puckering or redness.
- A change in nipple appearance or sensitivity.

2. Algorithms for Breast Cancer Detection

2.1 K-Nearest Neighbor

K-Nearest Neighbor strategy is a non-parametric and indolent learning algorithm. Its classification method is slightly different. It adopts the methodology of classifying the huge data based on the resemblance of its neighbors. It is a simple method that supplies all accessible instances and classifies new instances based on the resemblance measure. Here, a test point is selected and distances from the test point to every point in the training set is calculated. The KNN algorithm involves initialization of K followed by computation of distance between the input sample and the training samples. This step is followed by sorting the distances, taking the K nearest neighbors and finally applying simple majority.

2.2 Support Vector Machine

Support Vector Machine is a selective classifier formally defined by the separating hyperploid. When labeled training data is already given, this method results into a optimal hyperploid which further segregates into new instances. By varying the regularization parameter which is the tuning parameter of SVM, we will achieve substantial non-linear classified partition with higher accuracy in quite a reasonable amount of time.

2.3 Naive Bayes

The basis for Naïve Bayes technique is Bayes Theorem. This classifier assumes that the existence of a particular characteristic in a class is not related to the existence of any other attribute. It is a method that predicts the probability of different class depending upon numerous attributes.

2.4 Decision Tree

They are non-pareil strategy utilized for the purpose of categorization as well as regression. As and how the tree grows deeper and deeper, the complexities regarding the rules of decision tree increases. A decision node consists of 2 or more subdivisions. Non-parent node will represent a categorization. The decision node at the topmost level represents the best predictor and is called as a root node. These trees are capable of handling categorical as well as numerical data.

4. Proposed System

In order to predict breast cancer risks, We have found that, in accordance to the dataset that we have referred from UCI Repository, KNN Algorithm suits the most and has shown the highest accuracy rate of about 97.714% which is greater than other Machine learning models and strategies. Due to highest accuracy rate as shown by this algorithm, majority of the predictions will be yielded correctly.

Dataset description:

Serial No	Name of Attribute
I	Code Number
II	Thickness – Clump
III	Size – Cell
IV	Shape – Cell
V	MA(Marginal Adhesion)
VI	SECS(Single Epithelial Cell Size)
VII	BN(Bare Nuclei)
VIII	BC(Bland Chromatin)
IX	NN(Normal Nucleoli)
X	Mitoses
XI	Class – B(2) or M(4)

Table 1: Data attributes of Breast Cancer

4.1 Evaluation methods:

According to this experiment, we utilized K nearest neighbor strategy and is implemented on Spyder(Python-3.5). Out of a total of 175 instances, We observe that the number of wrong predictions is minimal in case of KNN prediction algorithm yielding highest accuracy rate among all other algorithms. Python is one of those languages which is proved to be both uncomplicated and intensive. It is an easy to learn, powerful programming language. It has efficient data structure and simple as well as effectual methodology to OOPS concept. Its elegance structure and non-static typing together with its interpreted nature make it an perfect language for script and faster app modification in many fields on most of the underlying platforms.

KNN algorithm when implemented in Spyder yielded higher accuracy of about 97.714% compared to several different methodologies.

CM (Confusion Matrix) obtained after applying KNN Algorithm

105	**2**
2	**66**

CM (Confusion Matrix) obtained after applying SVM Algorithm

108	**5**
2	**60**

CM (Confusion Matrix) obtained after applying Naïve Bayes Algorithm

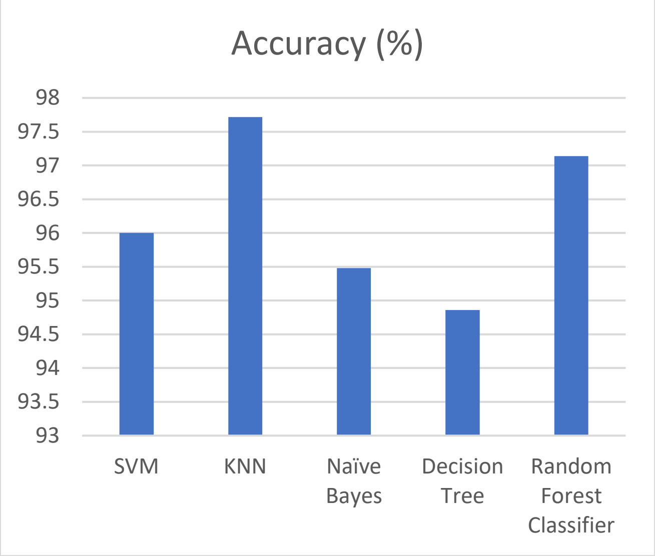
102	**4**
4	**65**

CM (Confusion Matrix) obtained after applying Decision Tree Algorithm

105	**3**
6	**61**

CM (Confusion Matrix) obtained after applying Random Forest Classifier Algorithm

106	**2**
3	**64**



5. CONCLUSION

This paper represents that accuracy rates of various algorithms and the reason for choosing KNN Algorithm for the purpose of predicting Breast Cancer cells to classify into either Malignant or Benign. The experimental results obtained are clearly depicted through the above graphical representation. These results are evaluated with respect to the Wisconsin Breast Cancer Dataset from UCI Repository.

6. REFERENCES

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[2]. Ms. Shweta Srivastava et al., "A Review Paper on Feature Selec- tion Methodologies and Their Applications", International Journal of Engineering Research and Development, Volume 7, PP. 57-61, 2013.

[3]. Animesh et al, "Study and analysis of Breast cancer Cell Detec- tion using Naïve Bayes, SVM and Ensemble Algorithms",International Journal of Computer Applications,vol2, 2016.

[4]. G. Sumalath et al, "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques",International Journal of Innovative Research in Computer and Communication Engineering,vol 5,2017.

[5]. Hiba Asri, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069.

[6]. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. http://www.cancer.org/.

[7]. https://www.r-project.org/