Breast Cancer Detection project report

Submitted by: Jaanhavi Gautam

To: Mr. Shailendra Singh

***INDEX***

1. **ABOUT……….**
2. **KEYWORDS…….**
3. **INTRODUCTION……**
4. **RESEARCH……..**

* **BREAST CANCER**
* **MACHINE LEARNING ALGORITHMS USED FOR DETECTION**

1. **DATASET DESCRIPTION**
2. **CODE SCREENSHOTS WITH OUTPUT SCREENSHOTS**
3. **REFERENCES**

***ABOUT***

India has witnessed 30% of the cases of breast cancer during the last few years and it is likely to increase. Breast cancer in India accounts that one woman is diagnosed every two minutes and every nine minutes, one woman dies. Early detection and diagnosis can save the lives of cancer patients. This project presents a novel method to detect breast cancer by employing techniques of Machine Learning. The authors carried out an experimental analysis on a dataset to evaluate the performance. The proposed method has produced highly accurate and efficient results when compared to the existing methods.

***Keywords***

Machine Learning

Classification

Breast cancer

Deep learning

***Introduction***

Breast cancer (BC) is the malignant tumor that activates in the cells of the breast. A tumor has the potential to spread to other parts of the body [[1]](https://www.sciencedirect.com/science/article/pii/S2405959520300801" \l "b1), [[2]](https://www.sciencedirect.com/science/article/pii/S2405959520300801" \l "b2). BC is a universal disease that hammers the lives of women typically in the age group of 25–50. With the potential rise in the number of BC cases in India, the distress reaching is alarming. During the past five years, the survival rates of BC patients are about 90% in the USA and whereas in India the figure reports approximately 60% [[3]](https://www.sciencedirect.com/science/article/pii/S2405959520300801" \l "b3). BC projection for India during 2020 suggests the number to go as high as two millions [[4]](https://www.sciencedirect.com/science/article/pii/S2405959520300801" \l "b4).

Specialist Doctors have identified hormonal, way of life and environmental factors that may increase an individual’s odds of developing BC. Over 5%–6% of BC patients have linked to gene mutations that went through the ages of the family. Obesity, increasing age, postmenopausal hormonal imbalances are the other factors that cause BC.

As such, there is no prevention mechanism for BC, but early detection can significantly improve the outcome. Further, this can also considerably reduce the costs of the treatment. However, sometimes it is unusual to show cancer symptoms, so early detection is difficult. It is indispensable to employ mammograms and self-breast tests to detect any early irregularities before the tumor gets advanced [[5]](https://www.sciencedirect.com/science/article/pii/S2405959520300801" \l "b5).

The key objective of this paper is to propose a novice method to detect BC. This paper presents a detailed study of existing cancer detection models and presents the highly accurate and efficient results.

***BREAST CANCER***

Breast cancer is the most found disease in the women, worldwide, where abnormal growth of a mass of tissue, cause the expansion of malignant cells leads to acute breast cancer.

These malignant cells are originally created from milk glands of the breast. These malignant cells which are the main reason for breast cancer can be classified into different groups

according to their unusual progress and capability affecting other normal cells [7]. The capability of affecting means whether these malignant cells affect only the local cells or can

spread throughout the full body. The effect of spreading these Journal of Telecommunication, Electronic and Computer Engineering malignant cells throughout the whole body of the patient is called as metastasis . It is very important to prevent this spreading effect by a diagnosis of cancer in the early stages using advanced techniques and equipment. In recent decades,

there are many efforts to employ artificial intelligence and other related methods to assist in the detection of cancer in earlier stages. Early detection of cancer boosts the increase of survival chance to 98% [8]. Figure 1. shows different types of cancers whereby breast cancer is leading with 24% as follows.

***III. MACHINE LEARNING METHODS***

Machine Learning is a process that machines (computers) are trained with data to make the decision for similar cases [9]. ML is employed in various applications, such as object

recognition, network, security, and healthcare. There are two ML types i.e. single and hybrid methods like ANN, SVM, Gaussian Mixture Model (GMM), K-Nearest Neighbor

(KNN), Linear Regressive Classification (LRC), Weighted Hierarchical Adaptive Voting Ensemble (WHAVE), etc. Following are the used ML algorithms:

***A. Artificial Neural Network (ANN)***

ANN is a model like human brains nerve system that has a

large number of nodes connected to each other. Each node

has two states: 0 means active and 1 means active. Also, each

node has a positive or negative weight that adjusts the

strength of the node and can activate or deactivate it. ANN

provides samples of data to train the machine. The trained

machine is used to detect the pattern of hidden date. It can

search for patterns among patients’ healthcare and personal

records to identify high-risk lesions [10].

***B. Support Vector Machine (SVM)***

SVM is a supervised pattern classification model which is

used as a training algorithm for learning classification and

regression rule from gathered data [11]. The purpose of this

method is to separate data until a hyperplane with high

minimum distance is found. SVM is used to classify two or

more data types. SVM include single or hybrid models such

as Standard SVM (St-SVM), Proximal Support Vector

Machine (PSVM), Newton Support Vector Machine

(NSVM), Lagrangian Support Vector Machines (LSVM),

Linear Programming Support Vector Machines (LPSVM),

and Smooth Support Vector Machine (SSVM).

***C. K-Nearest Neighbors (KNN)***

KNN is a supervised learning method which is used for

diagnosing and classifying cancer [12]. In this method, the

computer is trained in a specific field and new data is given

to it. Additionally, similar data is used by the machine for

detecting (K) hence, the machine starts finding KNN for the

unknown data. It is recommended to choose a large dataset

for training also K value must be an odd number.

***D. Decision Tree (DT)***

DT is a data mining technique used for early detection of

breast cancer. It is a model that presents classifications or

regressions as a tree. In this model, the data set is broken to

small sub-data, then to smaller ones. As a result, the tree is

developed and at the last level, the result is revealed. In a tree

structure, the leaves characterize the class labels whereby the

branches characterize conjunctions of feature leading to the

class labels Hence, DT is not sensitive to noise [13].

***E. Random Forest (RF) Algorithm***

RF algorithm is used at the regularization point where the

model quality is highest, variance and bias problems are

compromised [14]. RF builds numerous numbers of DTs

using random samples with a replacement to overcome the

problem of DTs. Each tree classifies its observations, and

majority votes decision is chosen. RF is used in the

unsupervised mode for assessing proximities among data

points.

***F. AdaBoost Classifier***

This algorithm is used for classification and regression to

predict breast cancer existence. It converts weak learners to

strong ones by combining all weak learners to form a single

strong rule. It gets the weight of the node and changes it

continuously until an accurate result is found. However, it is

sensitive to noise and quality of features [15].

***G. Naïve Bayes (NB) Classifier***

Naïve Bayes refers to a probabilistic classifier that applies

Bayes’ theorem with robust independence assumptions [16].

In this model, all properties are considered separately to

detect any existing relationship between them. It assumes that

predictive attributes are conditionally independent given a

class. Moreover, the values of the numeric attributes are

distributed within each class. NB is fast and performs well

even with a small dataset. However, it is difficult to find

independent properties in real life. [16]. have deployed NB

classifier for breast cancer detection and it gave the maximum

accuracy with only five dominan.

***PORTION OF DATASET USED***

**COLUMN NAMES**

(ID, CLUMP THICKNESS, UNIFORM CELL SIZE, UNIFORM CELL SHAPE, MARGINAL ADHESION, SINGLE EPITHELIAL SIZE, BARE NUCLEI, BLAND CHROMATIN, NORMAL NUCLEOLI, MITOSIS, CLASS)

1000025,5,1,1,1,2,1,3,1,1,2

1002945,5,4,4,5,7,10,3,2,1,2

1015425,3,1,1,1,2,2,3,1,1,2

1016277,6,8,8,1,3,4,3,7,1,2

1017023,4,1,1,3,2,1,3,1,1,2

1017122,8,10,10,8,7,10,9,7,1,4

***SCREENSHOTS OF CODE AND OUTPUT***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | |  |
|  |  |  |  |  |

***REFERENCES***

<https://en.wikipedia.org/wiki/Breast_cancer>

<https://ieeexplore.ieee.org/abstract/document/7818560>

<https://ieeexplore.ieee.org/abstract/document/7818560>