JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER DETECTION USING MACHINE LEARNING

Pretty Pramod Kotian
Department of Computer Engineering
Sinhgad Academy of Engineering, Pune
kotianpretty@gmail.com

Snehal Babar
Department of Computer Engineering
Sinhgad Academy of Engineering, Pune
babarsnehal2000@gmail.com

Aishwarya Ningdali
Department of Computer Engineering
Sinhgad Academy of Engineering, Pune
ningdali172000@gmail.com

Muskan Altaf Shaikh
Department of Computer Engineering
Sinhgad Academy of Engineering, Pune
muskanshaikh00060@gmail.com

Abstract : Breast cancer starts in the breast cell. It is a cancerous tumor where cancer cells grow and destroy nearby tissue. It was estimated in 2019 that 270,000 new breast cancer cases were diagnosed which is an alarming rise of cancer in women every year. With the advances of computer technology, we can save a life from cancer at an earlier stage. Hence, we have built the software with the help of machine learning to analyze breast cells before it gets fatal. This project aims to use machine learning algorithms and techniques to detect breast cancer and also do the prediction with Random Forest, KNN (k-Nearest-Neighbor) and Support Vector Machine algorithm. The Breast Cancer Wisconsin original dataset is used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. We will perform data visualization in the form of graphs like histograms, boxplots and also study the correlation between each attribute. In the end, we will develop a classification report and confusion matrix to predict whether the dataset is benign or malignant breast cancer for every machine learning algorithm.

Keywords - Breast Cancer, Machine learning, Prediction, KNN and SVM.

I. INTRODUCTION

Breast cancer is the second leading cause of female death (after lung cancer). Invasive breast cancer will be diagnosed in 246,660 women in the United States this year, with 40,450 women dying. Breast cancer is a cancer that originates in the breast and develops toward other areas of the body. When cells multiply uncontrollably, cancer develops. Breast cancer cells generate a lump that can be felt or seen on an x-ray.

Breast cancer cells can spread to other parts of the body if they enter the blood or lymph system. Changes and mutations in DNA are among the causes of breast cancer. DCIS (ductal carcinoma in situ) and invasive carcinoma are both prominent kinds of breast cancer. Others are less prevalent, such as phyllodes tumors and angiosarcoma. There are a variety of algorithms for evaluating breast cancer outcomes. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer.

There are numerous algorithms for breast cancer classification and prediction. The performance of four classifiers is analyzed in this paper: SVM, Logistic Regression, Random Forest, and kNN, which are among the most popular data mining algorithms. Mammography or a portable cancer diagnostic equipment could be used to detect it early during a screening test. Cancerous breast tissues change as the disease advances, and this can be associated to cancer stage. The stage of breast cancer (I–IV) indicates how far the cancer has spread in a patient. Stages are determined using statistical indications like as tumor size, lymph node metastasis, and distant metastases, among others. Patients must endure chemotherapy to prevent cancer from spreading. Patients must have breast cancer surgery, chemotherapy, radiation, and endocrine therapy to prevent cancer from spreading.

The study's objective is to Identifying and categorizing malignant and benign individuals, as well as considering how to parametrize our classification. As a result, ways to achieve high precision have been developed. We're investigating a variety of datasets to see how Machine Learning may be applied to them. Breast cancer can be classified using machine learning techniques. We wish to lower the mistake rates by using maximum precision JUPYTER employs the 10-fold cross validation test, which is a machine learning technique. To assess and analyze information in terms of efficacy and efficiency

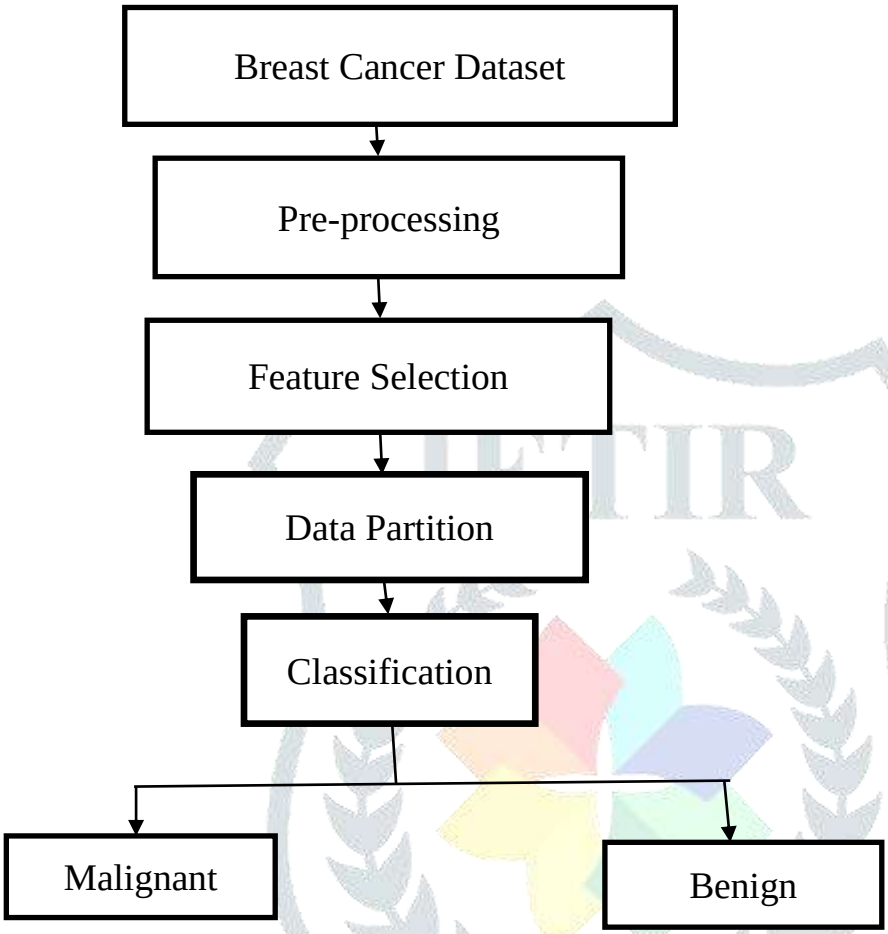
Breast cancer is second most type of cancer after lung cancer to cause of death. Invasive breast cancer will be diagnosed in 246,660 women in the United States this year, with 40,450 women dying. Breast cancer is a cancer that originates in the breast and develops toward other areas of the body. When cells multiply uncontrollably, cancer develops. Breast cancer cells generate a lump that can be felt or seen on an x-ray. Breast cancer cells can spread to other parts of the body if they enter the blood or lymph system. Changes and mutations in DNA are among the causes of breast cancer. DCIS (ductal carcinoma in situ) and invasive carcinoma are both prominent kinds of breast cancer. Others are less prevalent, such as phyllodes tumors and angiosarcoma. There are a variety of algorithms for evaluating breast cancer outcomes. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer. There are numerous algorithms for breast cancer classification and prediction.

The performance of four classifiers is analyzed in this paper: SVM, Logistic Regression, Random Forest, and kNN, which are among the most popular data mining algorithms. Mammography or a portable cancer diagnostic equipment could be used to detect it early during a screening test. Cancerous breast tissues change as the disease advances, and this can be associated to cancer stage. The stage of breast cancer (I–IV) indicates how far the cancer has spread in a patient. Stages are determined using statistical indications like as tumor size, lymph node metastasis, and distant metastases, among others. Patients must endure chemotherapy to prevent cancer from spreading. Patients must have breast cancer surgery, chemotherapy, radiation, and endocrine therapy to prevent cancer from spreading. The study's objective is to Identifying and categorizing malignant and benign individuals, as well as considering how to parametrize our classification

As a result, ways to achieve high precision have been developed. We're investigating a variety of datasets to see how Machine Learning may be applied to them. Breast cancer can be classified using machine learning techniques. We wish to lower the mistake rates by using maximum precision JUPYTER employs the 10-old cross validation test, which is a machine learning technique. To assess and analyze information in terms of efficacy and efficiency.

II. METHODOLOGY

i. Proposed methodology:



Firstly we take the WBCD dataset i.e. Wisconsin Breast Cancer Dataset and then Pre-processing the data by using discretize filter and resampling it and make a Cleaned data for further process, After that there is feature selection process basically this process removes the non-informative data from the model. Next there is data partition process typically this process involves partitioning of the data into a training set and testing set. Then we use different classification techniques with machine learning and finally using this classification techniques, we predicts which type of cancer patient have are Malignant or Benign

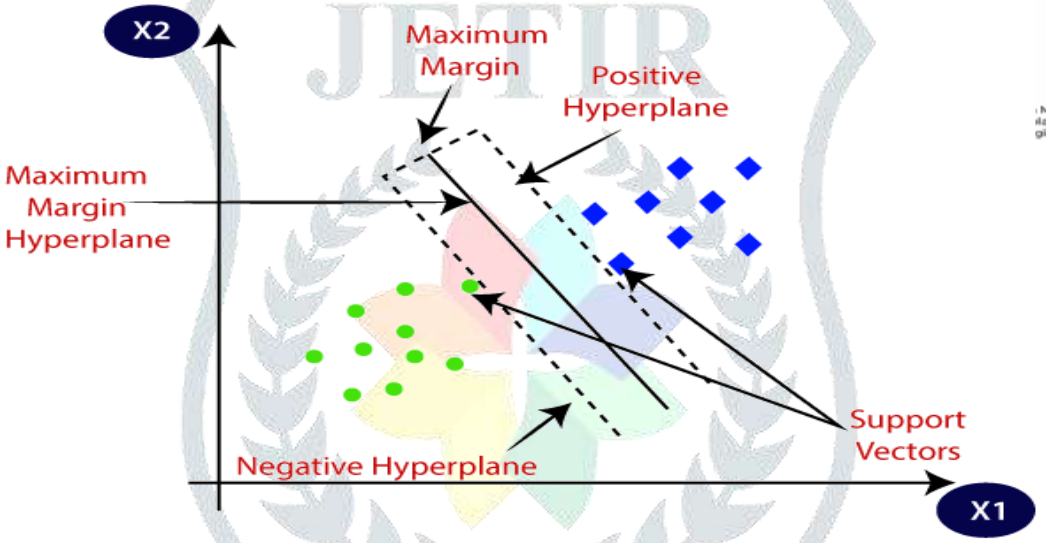
ii. Machine learning Algorithms Used:

k-Nearest Neighbour (kNN):-

k-Nearest Neighbour is a supervised learning method used to tackle problems mostly in the classification and regression fields. Its ease of use comes with a few disadvantages of its own; Depending on precision on data quality, sensitivity to large-scale data, and slow prediction, as well as the fact that its calculation is maintained permanently, makes it difficult. As a result, it necessitates a large amount of memory, making it quite demanding. For most average datasets, a generic k-NN is frequently used to categorized the means of a particular cluster set.

We will discuss the following topics in this paper: Try to train it for pattern recognition and use it for the main goal of making predictions. The k-NN method is a non-parametric technique.
Because k-NN is instanced based learning, both scenarios are trained in a feature space. In output is given after classification. In classification, the output is determined by picking the most votes cast by neighboring clusters of a specified k, where k is an arbitrary number value or a predetermined value based on the goal Feature extraction is a way of extracting output from input data after transformation.
The following is a common approach for how the k-NN Algorithm works:
1)Assigning a value to k or assigning an arbitrary value to it.
2)On both the testing and training datasets, calculations are performed.
3)Classification based on the information provided.
4)Finishing the results received and converting them to output.

SVM (Support Vector Machine) :-



It is one of the most prominent and commonly used Supervised Learning Algorithms. The end goal and objective of this method is to produce a decision boundary that can set apart and isolate n-dimensional space into classes, putting them into current data points in the correct category, and this process is referred to as 'hyperplane'. Memory efficiency and high dimensional space are only a few of the advantages it has over others. It has the capacity to work with both linearly separable and non-separable data.
It is said that Kernel tricks, often known as generalized dot product, are a method of calculating the dot product of two vectors to see how much they affect each other. The possibilities of linearly non-separable data sets have higher probability in higher dimensions, according to Cover's Theorem.

iii. PROPOSED METHODOLOGY RESULT AND DISCUSSION:

All tests and experiments on the classifiers and algorithms described and laid out so far in this study were conducted using JUPYTER notebook libraries, Python 3, version 6.1.5, and scikit learning machine. We have divided and segregated our dataset in this study in a 70:30 ratio. Training accounts for 70% of the

budget, while testing accounts for 30%. We chose JUPYTER because it includes a reputable collection of machine learning algorithms for pre-processing, clustering, classification, and regression that we could utilize on our dataset.

After training the model on our own dataset, we used the k-fold cross validation test to quantify the model's skill on new data or unexplored data sets.

III. RELATED WORK

The main cause of breast cancer is that when cells of breast begin to grow abnormally. These cells divide rapidly than healthy cells then continue to accumulate, forming a lump or mass. Cells may spread through patients breast to their lymph nodes or to other parts of patients body.

In Prateek P. Sengar, Mihir j. Gaikwad we uses Logistic Regression and Decision Tree Classifier for accuracy Measures. In the paper titled " Breast Cancer Detection Using Machine Learning Algorithms " used Random Forest, KNN(K-Nearest-Neighbor) and Naive Bayes algorithm and achieved accuracy 94% from each algorithm. Anusha Bharat and Pooja N & R Anisha Reddy(IEEE 2018) [4] compare the performance criteria of supervised classifier such as KNN, Naïve Bayes, Logistic regression and SVM and achieved accuracy of 99.1%.

In Kampreet S.Bhangu ,& Luxmi Sapra, International Conference on parallel, Distributed and Grid Computing (PDGC,2020) [4] here KNN ,Naïve Bayes , Logistic Regression, SVM and their accuracy came to be 98%. And last Nirdosh kumar, Gaurav Sharma [5] in that KNN, Logistic Regression, SVM, Naïve Bayes algorithms are used here we got logistic Regression=96.49%, SVM=98.24% KNN=97.20% Naïve Bayes=94.74% here we got better accuracy of SVM. We have used classification methods like KNN, SVM, Naive Bayes, Random Forest .

Literature survey:

Sr no	Literature Review	Author	Attributes	ML Algorithms	Accuracy Measures
1	Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction	Prateek P. Sengar, Mihir J. Gaikwad & Prof. Ashlesha S. N agdive (ICSSIT 2020)	Benign and Malignant	Logistic regression , Decision Tree Classifier	a. Training Data: 75% b. Testing Data: 25% Here it can almost pinpoint accuracy using Decision Tree Classifier algorithm.
2	Breast Cancer Detection Using Machine Learning Algorithms	Shubham Sharma, Archit Aggarwal & Tanupriya Choudhury (IEEE 2018)	Diagnosis, Radius_mean, Texture_mean etc	Random Forest, KNN (kNearest-Neighbor) and Naïve Bayes algorithm	94% is accuracy from each algorithm. KNN is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms.
3	Using Machine Learning algorithms for breast cancer risk prediction and diagnosis	Anusha Bharat, Pooja N & R Anishka Reddy (IEEE 2018)	Perimeter, Smoothness, Compactness etc	KNN, Naive Bayes, logistic regression and SVM (Support Vector Machine)	Accuracy of 99.1% for all algorithm. SVM using Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer.
4	Improving diagnostic accuracy for breast cancer using prediction-based approaches	Kamalpreet S. Bhangu, & Luxmi Sapra, International Conference on Parallel, Distributed and Grid Computing (PDGC,2020)	Concavity_mean, Concave points_mean, Symmetry_mean etc	KNN, Naive Bayes, logistic regression, SVM, XG Boost etc	Accuracy came to be 98%. XG Boost classifier with value of 0.99 performed better among all of them
5	The Machine Learning based Optimized Prediction Method for Breast Cancer Detection	Nirdosh Kumar, Gaurav Sharma & Lava Bhargava, (ICECA-2020)	Fractal dimension, Compactness etc	KNN, Naive Bayes, logistic regression, and SVM	Logistic Regression=96.49%, SVM=98.24 %, KNN=97.20 % and Naive Bayes=94.74 %. Here we have better accuracy of SVM

IV. CONCLUSION AND FUTURE SCOPE

We proposed the implementation of breast cancer diagnosis model using two different machine learning algorithms, namely: SVM and KNN in Google Collab using Python Language.

These above-given algorithms gave satisfactory results.

In our predictions the accuracy came to be 98 percent a reasonable success of the experiment executed with the help of Machine Learning Algorithms.

Performance comparison of the machine learning algorithms techniques has been carried out using the Wisconsin Diagnosis Breast Cancer data set.

Hence, given the features according to that of this dataset, breast cancer can be predicted with almost pinpoint accuracy using our SVM and KNN algorithm.

Furthermore, this can also be implemented on a cloud platform for ease of usage.

V. REFERENCE

1. Prateek P. Sengar, Mihir J. Gaikwad & Prof. Ashlesha S. Nagdive. On Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. In the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020)
2. Shubham Sharma, Archit Aggarwal & Tanupriya Choudhury. On Breast Cancer Detection Using Machine Learning Algorithms. In International Conference paper IEEE 2018
3. Anusha Bharat, Pooja N & R Anishka Reddy. On Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. In 2018, IEEE Third International Conference on Circuits, Control, Communication and Computing
4. Kamalpreet S. Bhangu, Jasminder K. Sandhu & Luxmi Sapra. On Improving diagnostic accuracy for breast cancer using prediction-based approaches. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)
5. Nirdosh Kumar, Gaurav Sharma & Lava Bhargava. On The Machine Learning based Optimized Prediction Method for Breast Cancer Detection. In Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)