International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252

Breast Cancer Prediction Using Machine Learning Algorithms

Mrs. P Revathi, Jesintha Praveena A, Porkodi C D

Assistant Professor,Dept. of Computer Science and Engineering Meenakshi Sundararajan Engineering College Chennai, India

Dept. of Computer Science and Engineering Meenakshi Sundararajan Engineering College Chennai, India

Dept. of Computer Science and Engineering Meenakshi Sundararajan Engineering College Chennai, India

Submitted: 25-07-2021 Revised: 04-08-2021 Accepted: 06-08-2021

ABSTRACT— Breast cancer is type of tumor that occurs in the tissues of the breast.It is most common type of cancer found in women around the world and it is among the leading causes of death in women.This becomes very handy, especially in the medical field where diagnosis and analysis are done through these techniques.Wisconsin Breast cancer data set is used to perform a comparison between Support vector machine,Logistic regression,Random forest classifier,K-Nearest neighbor.Based on the result of performed experiments,the Random Forest algorithm shows the highest accuracy (99.76%) with the least error rate.

Keywords—Random forest classifier, SVM, Machine Learning Algorithm.

I. INTRODUCTION

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new casesof invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of of women's death isestimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow outof control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump.Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to otherandparts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are manydifferent types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasivecarcinoma. Others, like phyllodes tumours and angiosarcoma are less common. There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – Fatigue, Headaches, Pain andnumbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classificationand prediction of breast cancer outcomes.

The present paper gives a comparison between the performance offour classifiers: SVM , Logistic Regression , Random Forest and kNN which are among the most influentialdata mining algorithms. It can be medically detected early during a screening examination throughmammography or by portable cancer diagnostic tool. Cancerous breast tissues change with the progression ofthe disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes howfar a patient's cancer has proliferated. Statistical indicators such as tumour size, lymph node metastasis,distant metastasis and so on are used to determine stages. To prevent cancer from spreading, patients have toundergo breast cancer surgery, chemotherapy, radiotherapy and endocrine. The goal of the research is toidentify and classify Malignant and Benign patients and intending how to parametrize our classificationtechniques hence to achieve high accuracy. We are looking into many datasets and how further MachineLearning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates withmaximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in JUPYTERto evaluate the data and analyse data in terms of effectiveness and efficiency.

II. RELATEDWORKS

A. SUPPORT VECTOR MACHINE:

Random forest is an algorithm that integrates multiple trees to form a forest through the idea of ensemble learning.

Its basic unit is decision tree. Each decision tree is a classifier, so for an input sample, N trees will have N classificationresults. Random forest integrates all the classification votingresults and specifies the category with the most voting

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252

timesas the final output. The construction process of random forestis as follows:

1) Let N be the number of samples in the original training set, and M be the number of characteristic attributes. Bootstrap sampling technique is used to extract N samples from the original training set to form a training subset.
2) m features are randomly selected as candidate features (m≤M) from M feature attributes. Each node of the decision tree selects the optimal attributes according to some rules (Gini impurity, information divergence, etc.) to split until all the training samples of the node belong to the same class, and they are completely split without pruning in the process.
3) Repeat the above two steps k times to build k decisiontrees and generate the random forest.
4) Using random forest to make decision, let x be the test sample, hi be the single decision tree, Y be the output variable, which is the classification label, I be the indicative function, H be the random forest model, and the decision formula.

B. Advantages and disadvantages of random forest algorithm

Random forest algorithm has many advantages. As a combination algorithm of classifiers, it can optimize the overall performance of the classification system by synthesizing the capabilities of several weak classifiers, which is better than a single classifier. When generating random forest, each decision tree is independent of each other and generated at the same time. The training speed is fast, and it is easy to make parallel method.

At the same time, the random forest algorithm also has some disadvantages. Because the randomness of the decision tree added by the random forest almost only occurs in the feature selection when the decision tree is generated, the fixity of the decision tree generation rules will lead to a certain degree of over fitting. At the same time, in the face of datawith high and unbalanced feature dimensions, performance ofalgorithm is seriously weakened because high-dimensionaldata usually contains a large number of irrelevant andredundant features.

C. **K-Nearest Neighbour:**

K-Nearest Neighbour (KNN) issaid to be the simplest and the most straightforwardclassification algorithm. Like most machine learningalgorithms, K-NN does not learn anything from theprovided dataset and its attributes, but simply use the pointsfrom the training data and finds the K number of nearestneighbors to that data point using Euclidean Distanceand classify it to the class which has the first K neighborsclosest to it.

D. **Support Vector Machine:**

Support vector machine (SVM) is a quite simple classification algorithm. This classifier is named so because it takes the help of vectors in the feature space to classify the class of a new vector [9,10]. The Maximum Margin Hyper-plane (MMH) decides whether the new vector belongs to class one or class two. If the data point lies beyond the negative hyper-plane or to the left of MMH then it belongs to the class one, else it belongs to the class two, where class one and two are two different classes in a given situation. SVMs can also be used if there are more than two classes.It is a supervised learning algorithm which is used forboth classifification and regression problems. It consists of theoretical and numeric functions to solve theregression problem. It provides the highest accuracy ratewhile doing prediction of large dataset. It is a strongmachine learning technique that is based on 3D and 2Dmodelling.

E. **Logistic Regression:**

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation like Linear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables

The general workflow is:
(1) get a dataset
(2) Train a classifier
(3) make a prediction using such classifier

F. **Data Collection**:

To predict breast cancer, authors used Wisconsin Diagnostic dataset collected from the UCI Machine Learning Repository. There were 699 instances with total of eleven features. Out of eleven features, ten features as input features and remaining one feature is treated as output feature. The whole dataset is divided into training and testing instances in the ratio of 80:20. It means out of 699 instances, 560 instances were used as training dataset and the remaining 140 instances were used as a testing dataset.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252

G. Data Pre-processing

Wisconsin Diagnostic dataset for breast cancer prediction has some missing values. To handle these missing values data pre processing was also used on the mentioned dataset. The attribute like Bare Nuclei column has missing features in the form '?' string which need to be inputted. 16 such instances of missing values were found in this feature. These missing values were replaced by the average/mean values of the features. On the other hand, the attributes like sample code number have no relevance in predicting breast cancer so such types of attributes have been dropped from the dataset.

H. **Flask:**

Flask is a python web framework, basically a python library for developing web applications.It was developed by Armin Ronacher. Flask is based on Web Server Gateway Interface(WSGI) and Jinga2 template engine. A WSGI object can be created as follows:

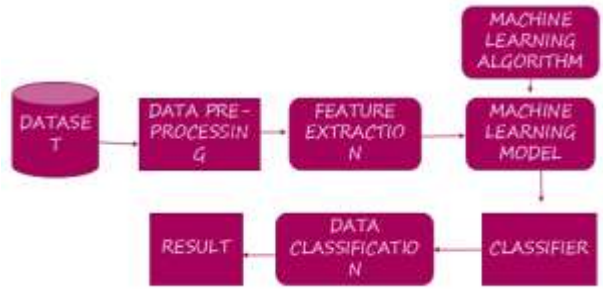
```
from flask import Flask
app = Flask(__name__) # Flask constructor
@app.route('/')
def hello():
    return 'HELLO'
if __name__=='__main__':
    app.run()
```

The run() method is used for starting the flask application.

F . PYTESSERACT:

Python-teseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images. Optical Character Recognition involves the detection of text content on images and translation of the images to encoded text that the computer can easily understand.

III. SYSTEM ARCHITECTURE



The system explains how the user acquires the required information of the patients by means of collecting the data and after gathering the information the details has to be processed using data pre processing finally the pre processed data has been extracted according to their feature given and it has been applied with the machine learning algorithms finally it has been classified and analysed according to the classification algorithms such as Support Vector Machine,K-Nearest Neighbor ,Logistic Regression, and Random Forest Classifier.After analyses using these algorithms it has been predicted that the Breast Cancer patient is in benign or in malignant stage.

IV SYSTEM IMPLEMENTATION

For the implementation of the ML algorithms,the data set was partitioned into the training set and testing set.A comparison between all the four algorithms will be made.The algorithm that gives the best results will be supplied as a model to the website.The website will be made from a python framework,called Flask.And it will host the database on inbuilt Python and Flask libraries.This data set is available on the UCI Machine Learning Repository.It consists of 30 real world attributes which are multivariate.The total number of instances is 569 and there are no missing values in this data set.

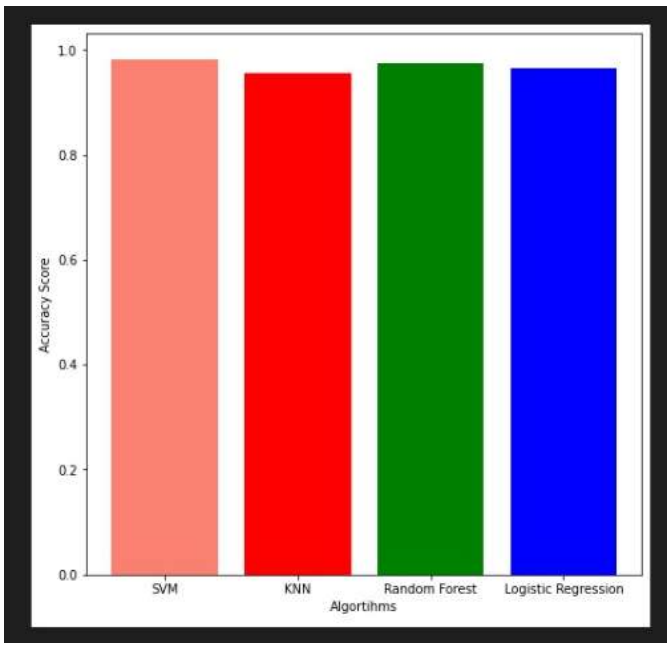
The patient books an appoinment through our website.The patient will then meet the doctor offline for the respective appoinment. The doctor will first check the patient manually,then perform a breast mammogram or an ultrasound.That ultra sound will show an image of the breast consisting the lumps or not.If the lumps are detected, a biopsy will be performed. The digitised image of the Fine Needle Aspirate(FNA) is what forms the features of the data set.Those numbers will be provided to the system by the doctor and the model will detect if its a benign or a malignant cancer.Breast cancer if found at an early stage will help save lives of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252

thousands of women or even men.The report will be then forwarded to the patient in their respective account.

A. GRAPH

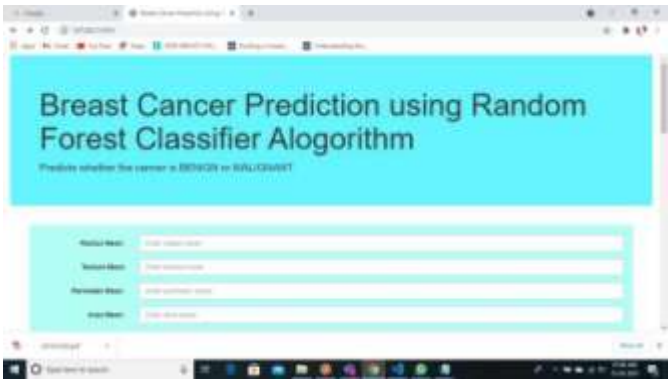
This graph shows the comparivity of all the four algorithms that has been analysed for the prediction of breast cancer type.In this graph,The Machine Learning algorithm such as Support Vector Machine,Logistic Regression,Random Forest Classifier and K-Nearest Neighbor were implemented.The accuracy given by SVM is 97.35% , KNN is 95.23%, LR is 95% and that by RFC is 98.6%.This paper concludes that Random Forest Classifier (RFC) have better accuracy of 98.6%



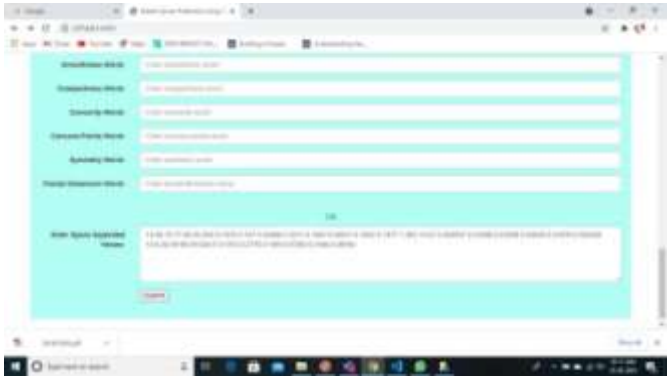
B. GIVE INPUT IN THE WEBPAGE

The user enters the featured details of him in the home page which includes features such as radius mean,perimeter mean,texture mean,area mean and 30 more features to identify whether the patient is in benign or malignant.This is the homepage of the website that is used to predict the breast cancer type using the algorithm that provided more accuracy on comparison with few other machine learning algorithms.

Random forest classifier provided better solution on comparison.

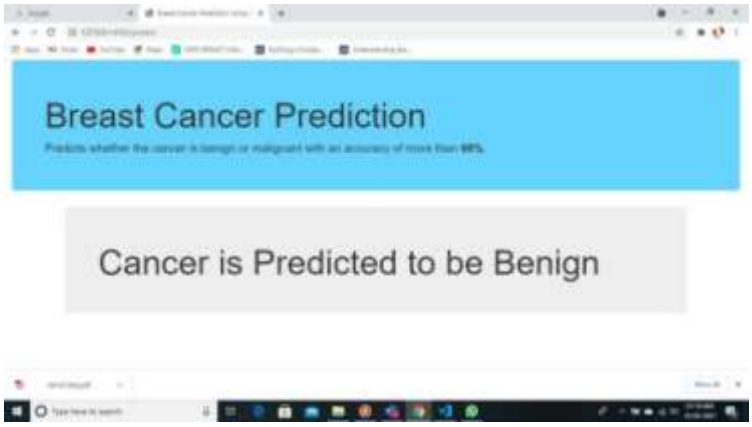


International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252



C. BENIGN TYPE CANCER

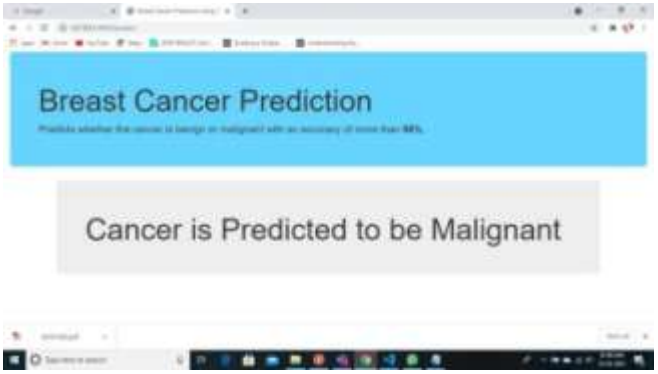
Once when the features has been entered it checks whether the breast cancer type is benign or malignant.It checks with the parameter of test data sets along with the trained data sets.The web page shows that the cancer is predicted to be benignBenign (non-cancerous) breast conditions are very common, and most women have them. In fact, most breast changes are benign. Unlike breast cancers, benign breast conditions are not life-threatening. But some are linked with a higher risk of getting breast cancer later on.



C.MALIGNANT TYPE CANCER

The patient can check for the other type of cancer using the featured data sets available.The data for malignant type cancer has been entered and checked and it shows cancer is predicted to be malignant.If a tumor is found to be malignant, you have breastcancer or another form of cancer. Malignant tumors can be aggressive and may spread to other surrounding tissues.A biopsy may be done on a suspicious lump, which can identify whether it is a tumor, as well as whether it is benign or malignant.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4088-4093 www.ijaem.net ISSN: 2395-5252



V.CONCLUSION AND FUTURE WORK

In statics analysis of Breast Cancer Prediction, Machine Learning algorithms have been used to train classifiers with features.The Machine Learning algorithm such as Support Vector Machine,Logistic Regression,Random Forest Classifier and K-Nearest Neighbor were implemented.The accuracy given by SVM is 97.35% , KNN is 95.23%, LR is 95% and that by RFC is 98.6%.This paper concludes that Random Forest Classifier (RFC) have better accuracy of 98.6%.The Scope of our project is to detect the patient is in benign or in malignant stage with the help of Machine Learning Algorithms.In future these techniques may be implemented on data sets that consists of images.The system may also be integrated with an Android application.The accuracy of the model created may be increased in order to give better predictions.

REFERENCES

[1]. Zerina Masetic, [2016] "Congestive detection using random forest classifier", vol.130,ISSN 0167-2607.

[2]. S.Gokhale , [2018] "Ultrasound characterisation of breast masses",India.The indian journal of radiology imaging,Vol.19,pp.242-249.

[3]. Abien Fred M.Agarap [2019]"An Application of Machine Learning Algorithms" on the Wisconsin Diagnostic Dataset for Breast cancer Detection,7th February.

[4]. L.J .Grimm, J.R.Marks [2020] "prediction of upstaged ductal carcinoma in situ",IEEE trans, June,vol.67,no.6,pp.1565-1572.

[5]. Haifeng Wang and Sang Won Yoon [2017] "Breast Cancer Prediction Using Data Mining Method", Department of System Science and Industrial State University of New York at Binghamton, May.

[6]. Priyanka Gupta and Prof.Shalini L [2018] "Analysis of Machine Learning Techniques for Breast Cancer Prediction" VIT University, Vellore, 5 May.

[7]. Sep [2019] "Breast Cancer Wisconsin (Diagnostic) Data Set".https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Di agnostic)

[8]. "Analysis of Machine Learning Techniques for Breast Cancer Prediction". International Journal of Engineering and Computer Science, http://www.ijecs.in/index.php/ijecs/article/view/4071.,May,Vol.7,no.05, [2018]

[9]. F.C.C.Garcia,I.Muga [2016] "Random forest for malware classification",aiXiv preprint:1609.07770