International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 5th Jan 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 01

DOI: 10.48047/IJIEMR/V12/ISSUE 01/18

Title **A Comparative Analysis of Machine Learning Algorithms for Prediction of Breast Cancer**

Volume 12, ISSUE 01, Pages: 185-191

Paper Authors

G. Karuna, Peddanna Sumanjali , Pranav Joshi , Samiya Sadiq ,G.Kalpana

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A Comparative Analysis of Machine LearningAlgorithms for Prediction of Breast Cancer

G. Karuna[1], Peddanna Sumanjali[2] , Pranav Joshi[3] , Samiya Sadiq[4] ,G.Kalpana[5].

[1,]Professor of CSE, GRIET, Hyderabad, Telangana.
[2,3,4] UG Student, Department of CSE, GRIET, Hyderabad, Telangana, India.
5)Assistant Professor, Department of CSE ,VJIT ,Hyderabad ,Telangana, India.
[a)]Corresponding author: karunavenkatg@gmail.com , anjuu1908@gmail.com pranavjoshi1108@gmail.com, samiyasadiq2002@gmail.com, gkalpanavjit.ac.in

Abstract
Among the most typical cancers in women is breast cancer. As a result, a prediction approach is required. The effectiveness of a few machine learning algorithms for detecting breast cancer, including Naïve-Bayes, Random-Forest, and K-nearest-Neighbor, is compared and contrasted in this survey. It explains why using machine learning techniques is preferable to using traditional methods. The Wisconsin data set is the dataset used, and the Wisconsin data set affects how accurate certain machine learning algorithms are. The smoothness, concavity, and radius of the tumor are some of the factors taken into account while determining if it is malignant or benign. Random Forest turned out to perform exceedingly well with 94 %.

Keywords: machine learning algorithms, Random Forest, Naïve Bayes, K -Nearest Neighbors

Introduction

There are different types of breast cancer. It occurs when cells in breast becomes out of control. Breast cancer is the most common malignancy among Indian women with 25%of cases and 14.7 of deaths. **Invasive ductal carcinoma** and **Invasive lobular carcinoma** are the two most prevalent types of breast cancer. Breast cancer can develop from various parts of breast. The breast is composed of three components: connective tissue, ducts, and lobules. Through blood and lymph vessels, breast cancer can develop outside of the breast cancer. There are various screening techniques available to detect breast cancer like

Mammography, Biopsy, MRI, and Thermography etc.

Mammography: A mammographic image is an X-ray of breast [fig.1.1]. Mammograms are used by doctors to discover early symptoms of breast cancer. Breast cancer is most successfully detected early by routine mammography, which can detect it up to 3yrs before symptoms appear. Most women find getting a mammogram painful. Some women experience agony. It's sensitivity to index cancer ranges from **63% to 98%**

Biopsy: Surgeons, interventional radiologists, and interventional cardiologists usually perform this [fig 1.2]. To diagnose a disease, samples of tissue or cells are removed for testing. Not all women who require a breast biopsy have cancer. Its sensitivity is typically cited as being **90–99%**. It may result in bleeding, bruising and infection.

MRI: Breast MRI [fig 1.3], often known as breast MRI, is a test that looks for abnormalities in the breasts such as breast cancer. Multiple pictures of your breast are taken during a breast MRI. A computer combines breast MRI scans to produce detailed images. Usually, a breast MRI is carried out following a cancer-positive biopsy. Doctors can determine the severity of the condition with the help of a breast MRI. MRI presented a diagnostic accuracy of **86.9%**, as well as a sensitivity of 95.5%.

Thermography: As a result of Thermography (R), an infrared image shows patterns of heat near or on the skin's surface. It only gives a little information [fig 1.4]. Although thermography can reveal changes vascular and thermal characteristics, it cannot reveal changes to the breast. It can pick up alterations that aren't malignant, therefore a mammogram would be necessary to determine the results. Thermography's accuracy was obviously inferior to mammography (69.7% vs. 76.9%).

The most advocated method in predicting the early detection of breast cancer is machine learning. The modeling and training can help in the early cure of breast cancer with the availability of datasets that have information about the features collected from the mammograms as well as other imaging modalities. These early diagnostic techniques can concentrate on giving cancer patients prompt access

to treatment, enhancing their quality of life.

Naive Bayes: One of the best classification methods is naive Bayes. Ranking performance is a more fascinating concept than simple classification in many decision-making systems. Thus, the weighted notion is implemented to enhance the performance of classic Naive Bayes.

Random Forest: It is a well-known method of supervised machine learning. Both classification and regression problems can be solved using it. Ensemble learning encompasses the idea of combining multiple classifiers in order to increase model efficiency and deal with complicated problems.

K-nearest Neighbor: The K-nearest neighbors algorithm, a supervised learning also known as KNN. It predicts the grouping of data points. It is often used as a classification technique, though it is also used for regression problems as well.

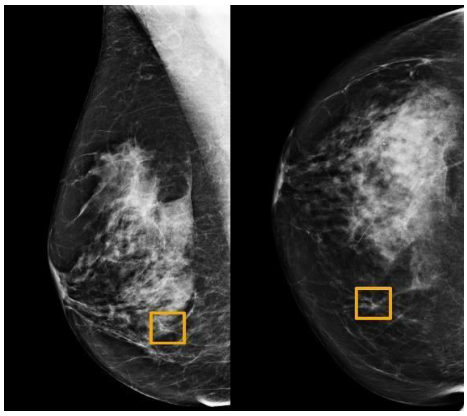


Figure 1.1

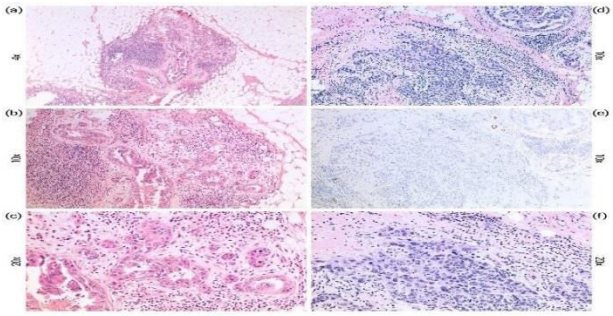


Figure1.2

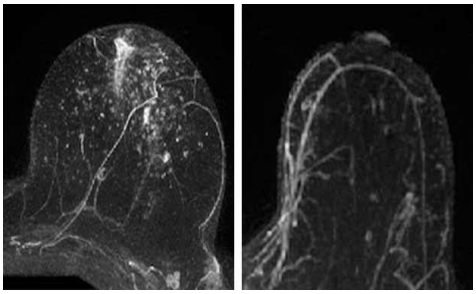


Figure1.3

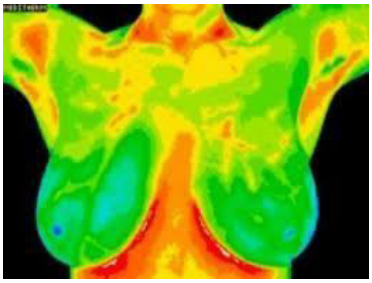


Figure 1.4

Paper objectives:

1 To improve prediction accuracy for fatal breast cancer at an earlier stage using the most appropriate machine learning method.

As the work progresses, 1.1 numerous articles in the same topic are first studied, and a proper description is given in the literature review.

The technique part of 1.2 clearly explains the method.

1.3 The result analysis offers insights into the results that were acquired.

In the section after the conclusion, we compare the findings.

2. Depending on accuracy, the optimal algorithm is chosen.

Literature Review

In [1], "authors compare five supervised machine learning methods for predicting breast cancer using the Wisconsin Breast Cancer dataset, including support vector machines (SVM), K-nearest neighbors, random forests, artificial neural networks (ANNs), and logistic regression. ANNs achieved a maximum accuracy of about 98.57%". For the prediction of breast cancer [3], "the following algorithms were used: Logistic Regression, Naive Bayes, and Random Forest; as well as the feature selection techniques Sequential Forward Feature Selection, Recursive Feature Elimination, f-test, and correlation combination results in better performance and it was found that Random Forest produced predictions with a higher degree of accuracy. Additionally, Sequential Forward Selection performs better on the larger dataset than f-test does on the smaller sample". In [2] "In addition to feature scaling, cross-validation, and ensemble machine learning models, we also use a bagging technique. KNN with decision trees outputs 100% accuracy". The decision tree model produces 100% accuracy when the train-test dataset is divided 90:10 with 300 bags of trees. In [4] the "authors Sri Hari Nallamala, ...et., have worked with logistic regression, KNN, SVM algorithms their contributions involving machine learning algorithms which proved to bring about 98.50% accuracy. Their methodologies involve panda ,numpy ,matplotlib, sci-kt-learn", in [5]a paper by Seid Hussain Yesuf, have shown, " SVM is the most accurate cancer diagnosis method when it comes to breast tumors. It outperforms Bayesian networks, k-nearest neighbors, and artificial neural networks in accuracy by 98.8%". The performance of four machine learning

algorithms—Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbors (KNN) is compared in [6] and is concluded that "SVM gives the highest accuracy of 97.13% with lowest error rate".

Methodology

In this paper we have used Wisconsin (Diagnostic)Dataset . The link to download dataset is provided below: *https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data*

The data from the above link is trained on three algorithms that are: Naive Bayes, Random Forest, and K-nearest Neighbor. On observing the dataset, the following conclusions were made:

Dataset general information:

1. Dataset Characteristics: Multivariate
2. Attribute Characteristics: Real
3. Attribute Characteristics: Classification
4. Number of Instances: 569
5. Number of Attributes: 32
6. Missing Values: No
7. Class distribution: 357 benign, 212 malignant

Attribute Information:
1) ID number

2) Diagnosis (M = malignant, B = benign)

While running dataset using Naive Bayes "id" and "Unnamed: 32" features were removed because they are not required while diagnosing whether the patient has a cancer or not. The data corresponding to malignant (M) and benign (B) were collected and visualized with the help of matplotlib in python. Then the data is tested and trained. Then our trained variables are passed to fit() (Note: we import GaussianNB from sk-learn.Naive_Bayes). It was observed that using Naïve Bayes we got an accuracy of 93.5%. In our Random Forest approach, we drop the unwanted variables. By adjusting the distribution's mean to zero and standard deviation to unit variance, we are normalizing the dataset in this case. (Note: Neglect any feature with low variance). We identify the number of malignant and benign patients and found that around 65% were having benign tumor and also found how features correlates. After further processing, training and testing of data it was concluded that random forest gave an accuracy of 94%. For KNN we consider only necessary attributes and train and test them after similar preprocessing and got an accuracy of 88%.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

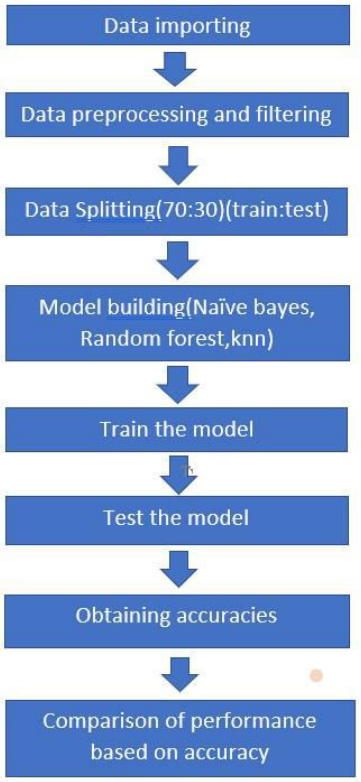


fig 2.1

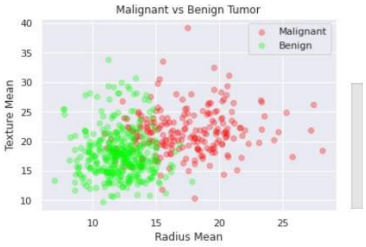
fig2.1 depicts the model deployment and analysis, initially the Wisconsin dataset is preprocessed and necessary attributes are chosen, additionally the data is split in a way that it assures majority of data is used to train the Naïve Bayes, Random Forest and KNN, further respective models are tested and accuracy scores are extracted for each model and compared. higher the accuracy score better the model.

Result and Analysis

upon running the respective models over the same dataset:

Naïve bayes analysis:

```
[ ] plt.title("Malignant vs Benign Tumor")
    plt.xlabel("Radius Mean")
    plt.ylabel("Texture Mean")
    plt.scatter(M.radius_mean, M.texture_mean
    plt.scatter(B.radius_mean, B.texture_mean
    plt.legend()
    plt.show()
```




```
[ ] ve Bayes score: ",nb.score(x_test, y_test)

    Naive Bayes score:  0.935672514619883
```

Accuracy:93.567%

Random forest analysis:

```
[ ] looking at the number of patients with Mal
    tas.diagnosis.value_counts().plot(kind='b
    t.title("Diagnosis (M=1 , B=0)", fontsize
    t.ylabel("Total Number of Patients")
    t.grid(b=True)
```

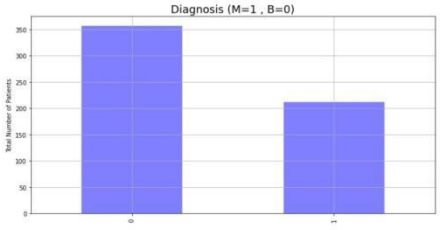


fig.3.2

```
rf = RandomForestClassifier(n_estimators=

rf = rf.fit(X_train, y_train)

predicted = rf.predict(X_test)

acc_test = metrics.accuracy_score(y_test,

print ('The accuracy on test  is %s' % (r

The accuracy on test data is 0.94
```

Accuracy:94 %

KNN analysis:

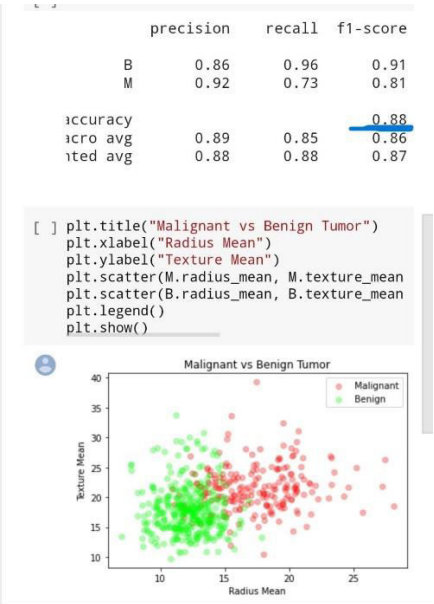


fig.3.3

Accuracy: 88 %.

N NAÏVE BAYES	RANDOM FOREST	KNN
93 93.567%	94 %	88%

Conclusion

Among Random Forest, Naïve Bayes, KNN machine learning algorithms, Random Forest excelled with a accuracy of 94 % over naïve bayes and KNN with 93.5 % and 88% respectively, the key objective here is to determine if the lump formed is malignant or benign. The work can also be further extended by bringing into picture the other machine learning algorithms and doing a performance analysis and expanding the size of dataset.

Reference Section:

1. Md. Milon Iam1, Md. Rezwanul Haque1, Hasib Iqball, Md. Munirul Hasan, Mahmudul Hasan, Muhammad Nomani Kabir 'Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques', Springer on 18th Aug 2020
2. Naveen, Dr. R. K. Sharma, Dr. Anil Ramachandran 'Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models', 978-1-7281-0630-4/19/@2019 IEEE
3. Dhanya R, Irene Rose Paul, Sai Sindhu Akula, Madhumathi Sivakumar and Jyothisha J Nair," A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection", 978-1-5386-8113-8/19 @ 2019 IEEE
4. Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru" Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8" BREAST CANCER DETECTION USING MACHINE LEARNING TECHNIQUES", Volume, Issue-2S3, July 2019.
5. Seid Hassen Yesuf, 10, No. 5, September-October 2019 International Journal of Advanced Research in Computer Science
6. Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime c, Thomas Noeld, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Published by Elsevier B.V (2016)