Comparitive Analysis Of Different Classification Models To Detect Breast Cancer

Bhargav Muppalla
Department of Computer Science
College of Arts And Sciences
bmuppalla1@student.gsu.edu

***Abstract*—According to global statistics, breast cancer (BC) is one of the most frequent malignancies among women globally. Early detection of BC improves the prognosis and chances of survival by allowing patients to get timely therapeutic therapy. Patients may avoid unneeded therapies if benign tumors are classified more precisely. Motivated by this, in our work, using BC dataset we classify whether a person has malignant tumor or benign tumor using logistic regression, KNN classifier, support vector machine, Kernel SVM, Naive Bayes, Decision Tree and Random Forest Classification models and compare their performances to determine which is the best model for this problem.**

I. Introduction

The use of classification and data mining technologies to categorize data is quite successful. Particularly in the medical industry, where such procedures are commonly employed in diagnosis and decision-making. Early detection of BC improves the prognosis and chances of survival by allowing patients to get timely therapeutic therapy. Patients may avoid unneeded therapies if benign tumors are classified more precisely. As a result, accurate BC diagnosis and categorization of individuals into malignant or benign groups is a hot topic of research. Machine learning (ML) is widely regarded as the approach of choice in BC pattern classification and forecast modeling due to its unique benefits in detecting essential characteristics from complicated BC datasets.

A. Recommended screening guidelines

Mammography. The mammography is the most essential breast cancer screening test. A mammogram is a type of X-ray that is used to examine the breast. It can identify breast cancer up to two years before you or your doctor can feel the growth.

A mammography should be done once a year for women aged 40–45 who are at an average risk of breast cancer.

Starting at the age of 30, high-risk women should receive annual mammograms and an MRI.

B. Some Risk Factors For Breast Cancer

Age. As women become older, their chances of developing breast cancer increase. Breast cancer is seen in about 80% of women over the age of 50.

Personal experience with breast cancer. A woman who has had breast cancer in one breast is more likely to get cancer in the other breast.

Breast cancer runs in the family. If a woman's mother, sister, or daughter had breast cancer when she was young, she has an increased chance of breast cancer (before 40). Having additional relatives who have been diagnosed with breast cancer might further increase your risk.

Genetic factors. Women with particular genetic abnormalities, such as alterations in the BRCA1 and BRCA2 genes, have a greater lifetime chance of getting breast cancer. Other gene variations may also increase the risk of breast cancer.

II. Data Preparation

For breast cancer datasets, we used the UCI Machine Learning Repository.
http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

Dr. William H. Wolberg, a physician at the University Of Wisconsin Hospital in Madison, Wisconsin, produced the dataset that was utilized in this work. Dr. Wolberg used fluid samples collected from patients with solid breast masses and an easy-to-use graphical computer tool called Xcyt to analyze cytological characteristics based on a digital scan to construct the dataset. The software computes 10 features from each of the cells in the sample using a curve-fitting technique, then calculates the mean value, extreme value, and standard error of each feature for the picture, returning a **30 real-valued vector**.

A. Attribute Information

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter).
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation” — 1)

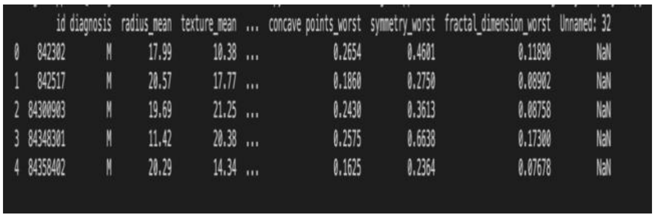
For each picture, the mean, standard error, and "worst" or largest (mean of the three largest values) features were computed, yielding **30 features**. For example, field 3 represents Mean Radius, field 13 represents Radius SE, and field 23 represents Worst Radius.

B. Objectives

The goal of this study is to identify which traits are most useful in predicting whether a cancer is malignant or benign, as well as to look for general trends that can help with model selection and hyper parameter selection. The objective is to determine if the cancer is benign or malignant. I did this by fitting a function that can predict the discrete class of fresh input using machine learning classification algorithms.

III. Data Exploration

Class distribution:
Benign: 458 (65.5%)
Malignant: 241 (34.5%)



```
        id diagnosis  radius_mean  texture_mean  ...  concave points_worst  symmetry_worst  fractal_dimension_worst  Unnamed: 32
0   842302         M        17.99         10.38  ...                0.2654          0.4601                  0.11890          NaN
1   842517         M        20.57         17.77  ...                0.1860          0.2750                  0.08902          NaN
2 84300903         M        19.69         21.25  ...                0.2430          0.3613                  0.08758          NaN
3 84348301         M        11.42         20.38  ...                0.2575          0.6638                  0.17300          NaN
4 84358402         M        20.29         14.34  ...                0.1625          0.2364                  0.07678          NaN
```

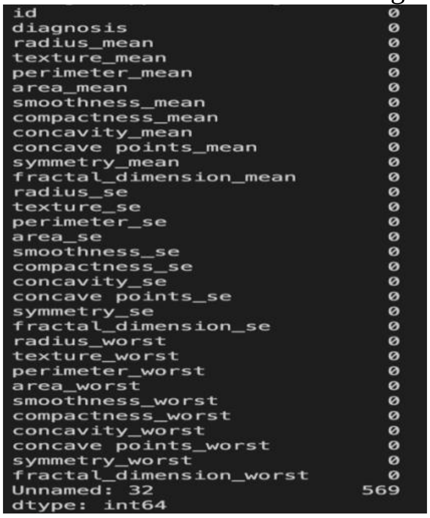
Fig: Top 5 Data of our dataset.

Cancer data set dimensions: (569, 32)

The data set has 569 rows and 32 columns, as can be seen. The column we'll anticipate is 'Diagnosis,' which indicates whether the malignancy is M = malignant or B = benign. The number 1 indicates that the cancer is malignant, whereas the number 0 indicates that it is benign. We can see that 357 of the 569 people are designated as B (benign) and 212 are labeled as M (malignant).

A. Missing Data

The dataset has no null or missing values.



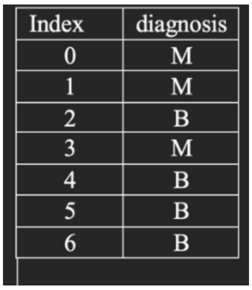
```
id                           0
diagnosis                    0
radius_mean                  0
texture_mean                 0
perimeter_mean               0
area_mean                    0
smoothness_mean              0
compactness_mean             0
concavity_mean               0
concave points_mean          0
symmetry_mean                0
fractal_dimension_mean       0
radius_se                    0
texture_se                   0
perimeter_se                 0
area_se                      0
smoothness_se                0
compactness_se               0
concavity_se                 0
concave points_se            0
symmetry_se                  0
fractal_dimension_se         0
radius_worst                 0
texture_worst                0
perimeter_worst              0
area_worst                   0
smoothness_worst             0
compactness_worst            0
concavity_worst              0
concave points_worst         0
symmetry_worst               0
fractal_dimension_worst      0
Unnamed: 32                569
dtype: int64
```

Fig: Observing missing data

B. Categorical Data

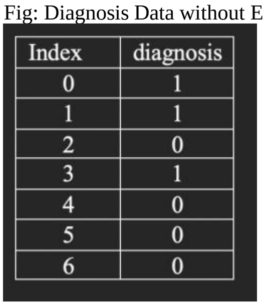
Categorical data consists of variables with label values rather than numeric values. The number of available values is frequently restricted to a small number.
Users are often classified by nation, gender, age group, and other factors.

Label Encoder was used to label the categorical data. Label Encoder is a Python module that is used to transform categorical input, such as text data, into numbers that our predictive models can understand better.



Index	diagnosis
0	M
1	M
2	B
3	M
4	B
5	B
6	B

Fig: Diagnosis Data without Encoding



Index	diagnosis
0	1
1	1
2	0
3	1
4	0
5	0
6	0

Fig: Diagnosis Data after Encoding

IV. Splitting the Dataset

The data we utilize is divided into two categories: training data and test data. The training set comprises a known output, and the model learns from it to generalize to new data in the future. To test our model's prediction on this subset, we have the test dataset (or subset). Data is split in the ratio of 75:25 (75% for training and 25% for testing)

We did this using SciKit-Learn library in Python using the train_test_split method.

V. Feature Scaling

Dataset will almost always contain features with a wide range of magnitudes, units, and ranges. However, most machine learning algorithms compute the Euclidian distance between two data points. All characteristics must be brought to the same magnitude level. This may be accomplished by scaling.
This means data transformed so that it fits within a specific scale, like 0–100 or 0–1.

VI. Model Selection

Supervised learning is a form of system in which both the desired input and output data are given. To offer a learning framework for future data processing, input and output data are labeled for categorization. Regression and classification issues are two types of supervised learning tasks.

When the output variable is a real or continuous value, such as "salary" or "weight," you have a regression problem.

When the output variable is a category, such as filtering emails as "spam" or "not spam," the problem is called a classification problem.

Unsupervised Learning: Unsupervised learning is when an algorithm uses data that hasn't been classed or labeled and allows it to operate on it without being guided.

In our dataset, the outcome variable, or dependent variable, Y, has just two sets of values: M (Malign) or B (Balance) (Benign). As a result, we used the supervised learning method Classification.

A.Classification Algorithms

We used 7 different classification algorithms to solve the problem.
1.Logistic Regression
2.KNN (K nearest neighbors)
3.SVM (support vector machines)
4.Kernel SVM
5.Naive Bayes
6.Decision Tree Algorithm
7.Random Forest Classification

We used sklearn library to import all methods of classification algorithms.

VII. Results

We Applied different classification models on the same dataset and compared their accuracies. We Validated the robustness of models by removing some of the features and observed how particular features affecting the performance of the model.

A. Accuracies with all features

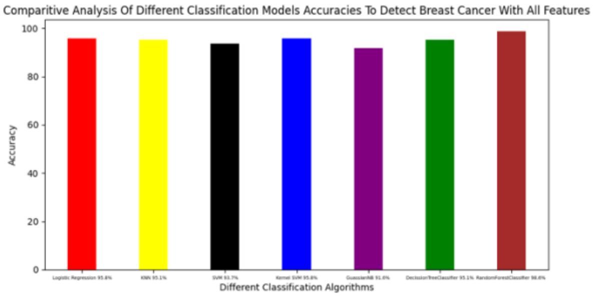


Fig: Comparative Analysis of Different Classification Models in Detecting Breast Cancer including all 32 features available.

With all 32 features Random Forest Algorithm performs better with an accuracy of 98.6%.

B. Without mean feartures

Dataset contains features like radius_mean, texture_mean, perimeter_mean etc., these are the mean values of raw values of radius ,texture, perimeter. We excluded these features and applied models to check their accuracies.

With mean features excluded, accuracy of Random Forest Algorithm drops, and Kernel SVM performs best in this case. From which we can observe that different algorithms perform differently with different kind of features.

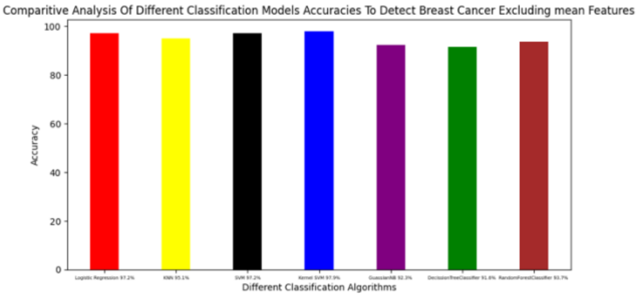


Fig: Comparative Analysis of Different Classification Models in Detecting Breast Cancer Excluding all mean features.

C. *Without standard error features*

Dataset contains features like radius_se, texture_se, perimeter_se etc., these are the standard error values of raw values of radius ,texture, perimeter. We excluded these features and applied models to check their accuracies.

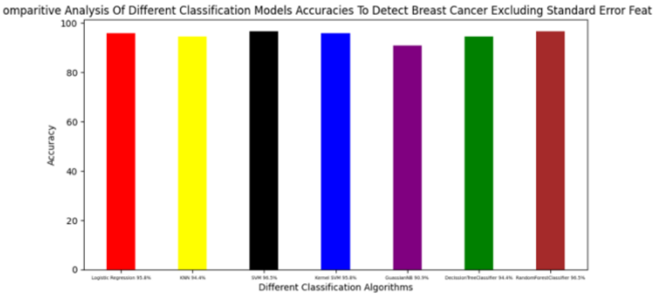
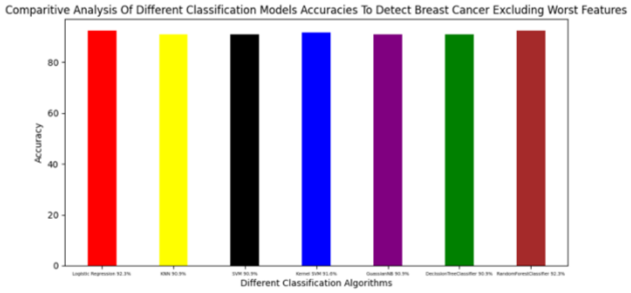


Fig: Comparative Analysis of Different Classification Models in Detecting Breast Cancer Excluding all standard error features.

With mean features excluded, SVM and Random Forrest models performs.

D. *Without Worst features*

Dataset contains features like radius_worst, texture_worst, perimeter_worst etc., these are the worst values of raw values of radius ,texture, perimeter. We excluded these features and applied models to check their accuracies.



From observations of the results, we find the features tagged as worst are most influential features. Performance of all the models decreased without these features. Random forest performs best among all models with 92.3% accuracy.

VIII. Conclusion

Finally, we've created our classification model, and we've discovered that the Random Forest Classification method produces the best results for our dataset. It isn't necessarily relevant to all datasets, though. To select a model, we must first assess our dataset before using our machine learning model. We also tested our models Robustness with subsets of features.