Building a Machine Learning Model on Breast Cancer Data with Focus on Cross Validation and Accuracy

Sagar Rai[1], Aditya Anand[2] and Kunal Singh[3]

Department of Electronics & Communication Engineering

NIT Jamshedpur, Jamshedpur, 831014, India

E-mail:- *raisagar2102@gmail.com[1], rustyad.anand@gmail.com[2], kunalsingh.ece@nitjsr.ac.in[3]*

DOI:- https://doi.org/10.47531/MANTECH/ECC.2021.37

Abstract

Breast cancer, abbreviated as BC, is one of the most prominent cancers among females globally, consisting of the major percentage of the new cancerous cases and the disease-related fatalities in the world among the gender. This makes the disease a major health-related issue in the current world. Disease's early diagnosis highly upgrades the prognosis and result in a high survival rate among women. This is mainly due to the fact that the early diagnosis may promote timely clinical treatment. Additionally, the correct classification of benign (not risky) tumors saves the patients from going to unnecessary treatments. The unique advantages of Machine Learning (ML) to detect complex relations and critical features have a major advantage over any other traditional method for correct classification of the disease tumor. Research shows that an expert physician can diagnose a case of breast cancer with an accuracy of 79 percent while the accuracy of 91% or above is achieved by using machine learning algorithms. In the conducted project, we have performed various operations (data pre-processing and feature selection) on the raw data collected from the UCI repository to get meaningful data from the raw data. We then trained various Machine Learning models on the meaningful data to achieve great accuracy in the classification of the breast tumor as dangerous or not. The study's main aim was to find an algorithm that has a good cross-validation score along with a high cross-validation score. K-fold cross-validation was used for testing the trained model. This ensured that the model was neither highly biased neither had a high variance. Application programming interface (API) support for the model using Flask is also provided for cross-language usage of the trained model.

Keywords: - *Breast Cancer, Machine Learning, Malignant, Benign, Tumor, Cross Validation, K-Folds, Flask, API*

INTRODUCTION

Breast Cancer as abbreviated usually as BC is a type of common cancer that is the largest cause of cancer-related deaths among women throughout the world. The research shows that the early detection of the disease in its early stages can improve the path of treatment taken and the rate of survival significantly. The disease is mainly classified into two major types, namely malignant and benign. The first type means the disease is risky and may cause severe health-related risks, including the spreading of the tumor to the surrounding tissues. The second type is a type where the tumor does not spread to the various tissues and does not require specific cancer medication, but its treatment is also necessary. The treatment plan can be accurately decided based on the type of tumor.

The major aim of the research work was to use machine learning on the Wisconsin breast cancer dataset for the prediction of the tumor as benign or malignant. The main focus is to do the study using various machine learning models with a major focus on cross-validation score and accuracy.

A. Types of Breast Tumor

As the name of the disease suggests, this disease mainly relates to cancers or malignant tumors appearing in the inner lining of the milk ducts present in the breast tissues. [1] Tumors present within the breast tissue can be cancerous or non-cancerous and hence are divided into the subsequent two types. [2]

Benign Tumors - The tumors which are non-malignant/non-cancerous are referred to as benign tumors. This type of tumor is localised, and it does not spread to the further parts of the body.

These also require treatment as they might grow big in size and hence cause various other issues. Benign tumors can even turn to malignant tumors, which is why their treatment is necessary.

Malignant Tumors - Those tumors which may or have already resulted in cancer or are carcinogenic in any way, i.e., they spread to the underlying and neighbouring tissues are known as the malignant tumors.

LITERATURE REVIEW

1. **Missing data computation using statistical and machine learning methods n a real breast cancer problem: 5** Missing data imputation is an important task in cases where using all the available data is crucial and not discard records with missing values. The imputation methods based on machine learning algorithms outperformed statistical imputation methods in the prediction of patient outcome, which helped us in analysing the data that we were not able to provide to the system.
2. **Applications of Machine Learning in Cancer Prediction and Prognosis: [5]** Among the validated studies and better designed, it was pretty clear that machine learning methods can be used to substantially (about 15-20%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. With the help of this paper, we were able to take various examples that we can use as input for prediction.
3. **Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence: [6]** Most of the medical databases are not analysed for finding valuable and hidden knowledge. Advanced data mining techniques were used to discover hidden patterns and relationships in this paper. Models developed from these techniques were very much useful for medical practitioners to make the right decisions.

TECHNOLOGIES USED

A. Concepts of Machine Learning

Machine Learning is an application of Artificial Intelligence (AI) that provides computers with the ability to make use of learning algorithms that make inferences from data to learn new tasks instead of being explicitly programmed. Machine learning is used to develop computer programs that can access data and use these data to learn for themselves.

The primary aim of Machine learning is to allow computers to learn automatically and take or adjust actions accordingly without any intervention of human beings. The learning process involves observation of data and looking for patterns in them so that it can predict the future outputs based on the inputs that we provide.

- ***Supervised Learning*** - Supervised learning in machine learning is where one has input and output variables, let X and Y, and uses an algorithm to learn the mapping function from the input to output Y= f(X). The goal is that when you have new input data (x), you can approximate the mapping function so well that you can predict the output variables (Y) for that data. [7]

  Supervised Learning algorithms are divided into two types classification algorithms and regression algorithms. We use various classification algorithms in our project for the prediction.

- ***Classification Algorithms*** - These algorithms are used for mapping an x to y using a function f(x) which is based on some parameters depending upon the model used to produce only discrete-valued outputs.

  The assignment of the discrete value is depended on the probability, and it is calculated for different classes, and the final class is selected depending upon the highest probability or a manually selected desired cutoff value in some of the cases. Judgement of the trained model is done on the basis of various parameters like accuracy, precision, recall and f1score.

- We have used various classification models such as Logistic Regression, Decision Trees, Random Forests, K Nearest Neighbors, Support Vector Machines and Perceptron in our study.

B. Various Classification Algorithms

- ***Logistic Regression*** - One of the most important models used for the statistical classification of the Machine Learning problems and study of the data. The sigmoid function represented by ‘g’ is used mainly for the calculation of the probability.
- ***Decision Trees*** - Regression nodes are built-in decision trees. It handles the features in a

bottom-up manner to predict the output. A tree is built with all the features as its nodes, and the leaf nodes are the prediction to classes used to give the classification in a particular class.

- ***Random Forests*** **-** It can be defined as a type of ensemble learning method used in the prediction of output which is used in both regression and classification problems. It can be imagined as a combined cluster of decision trees in which all the decision tree features are used in combination for the prediction of the output.

PROPOSED METHODOLOGY

As indicated by the research, most experienced physicians can detect the breast cancer tumor as benign or malignant with an accuracy of about 79%, while the accuracy of 91% is achieved by using machine learning classification techniques.

Hence, we propose to use various machine learning classification algorithms such as logistic regression, decision trees, random forest, perceptron, k nearest neighbors, support vector machines and perceptron to classify the tumors as benign or malignant by training our machine learning model on the already available data. We have taken our data from the publically available UCI Machine Learning repository, Breast Cancer Wisconsin (Diagnostic) Dataset.

A. Workflow

See block diagram

B. Data Collection

A digitized image of a fine needle aspirate (FNA) of a breast mass is used to compute the features of the dataset. The needle aspirate process is a process which is any process used to simply extract some cells from the tumor. Characters present in the sample is described in the three-dimensional space. [11]

We will train our model on this dataset for the detection of breast cancer in the new cases. The collected data consists of 32 columns and 569 rows. The data is divided into a test-train ratio of 80% to 20%.

C. Data Preprocessing

A technique that is used to convert raw data into an easily understandable format is termed as data pre-processing. [12] The data we get is usually incomplete, inconsistent or may lack in a certain behavior or trends and is more likely to contain many errors. Data pre-processing is used as a proven method for resolving such issues.

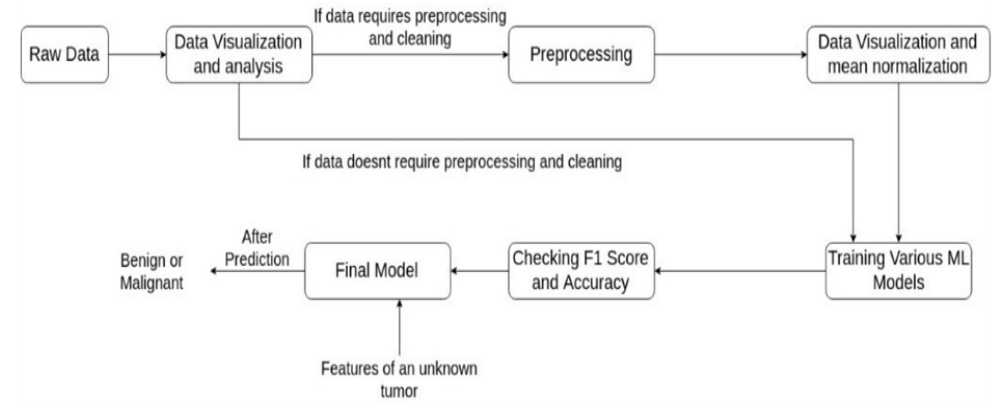
The steps involved in data pre-processing are Data cleaning, Data Integration, Data Transformation, Data Reduction and Data Discretization.

In our Machine Learning model, we saw that the dataset has many features with varying limits which may lead to poor training and performance of our machine learning model. Hence, we apply feature scaling and mean normalization to both our training and testing data to bring all the features within a similar range, which will help the machine learning model to train better.

```
mean      19.153953
std        4.202446
min        9.710000
25%       16.155000
50%       18.770000
75%       21.595000
max       39.280000
Name: texture_mean, dtype: float64
----------------------------------------
count    559.000000
mean      91.731467
std       24.069775
min       43.790000
25%       75.190000
```

Fig. 2: Features before applying feature scaling and mean normalization

BLOCK DIAGRAM



```
mean     -5.866903e-15
std       1.000000e+00
min      -2.247251e+00
25%      -7.136209e-01
50%      -9.136428e-02
75%       5.808632e-01
max       4.789126e+00
Name: texture_mean, dtype: float64
----------------------------------------
count     5.590000e+02
mean     -1.535246e-15
std       1.000000e+00
min      -1.991770e+00
25%      -6.872298e-01
```

Fig. 3: Features after applying feature scaling and mean normalization

By plotting various plots such as joint plots, pair plots, correlation matrix, scatter plots and bar plots, we found that some of the features are highly correlated to each other, and hence we drop those features from our dataset. This is done as some machine learning algorithms such as random forests are good at detecting interactions between various features, but highly correlated features can mask these interactions. See **figure 4**.

This can be viewed as a special case of Occam's razor which states that a simpler model is preferable, and in some cases, a model with fewer features is simpler. This holds true with the concept of minimum description length, which makes this more precise. [13]

We remove the following highly correlated features from the dataset: -

'perimeter_mean', 'radius_mean', 'compactness_mean', 'concave points_mean', 'radius_se','perimeter_se', 'radius_worst', 'perimeter_worst', 'compactness_worst', 'concavepoints_worst', 'compactness_se', 'concavepoints_se', 'texture_worst', 'area_worst'

See **figure 5** and **figure 6**.

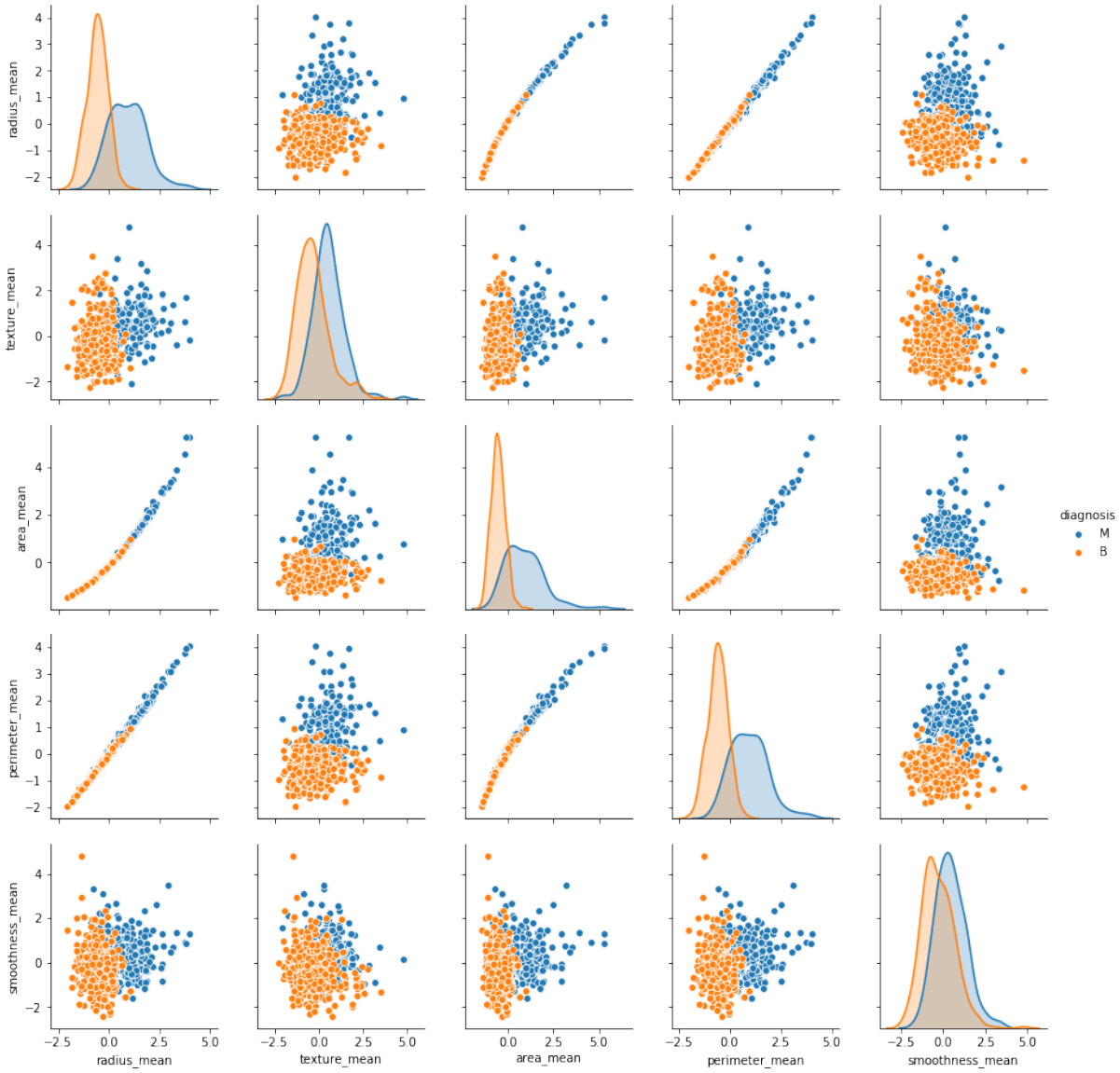


Fig. 4: Pairplot Showing Relation between various Features

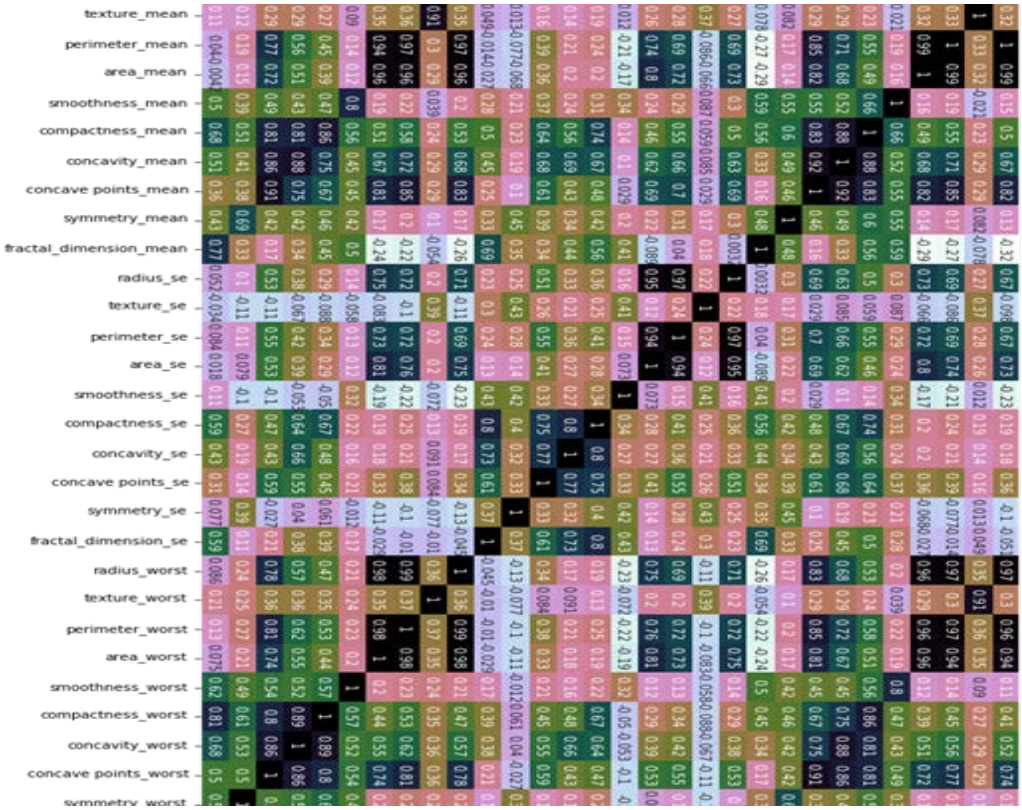


Fig. 5: Correlation Matrix before Removal of highly correlated features

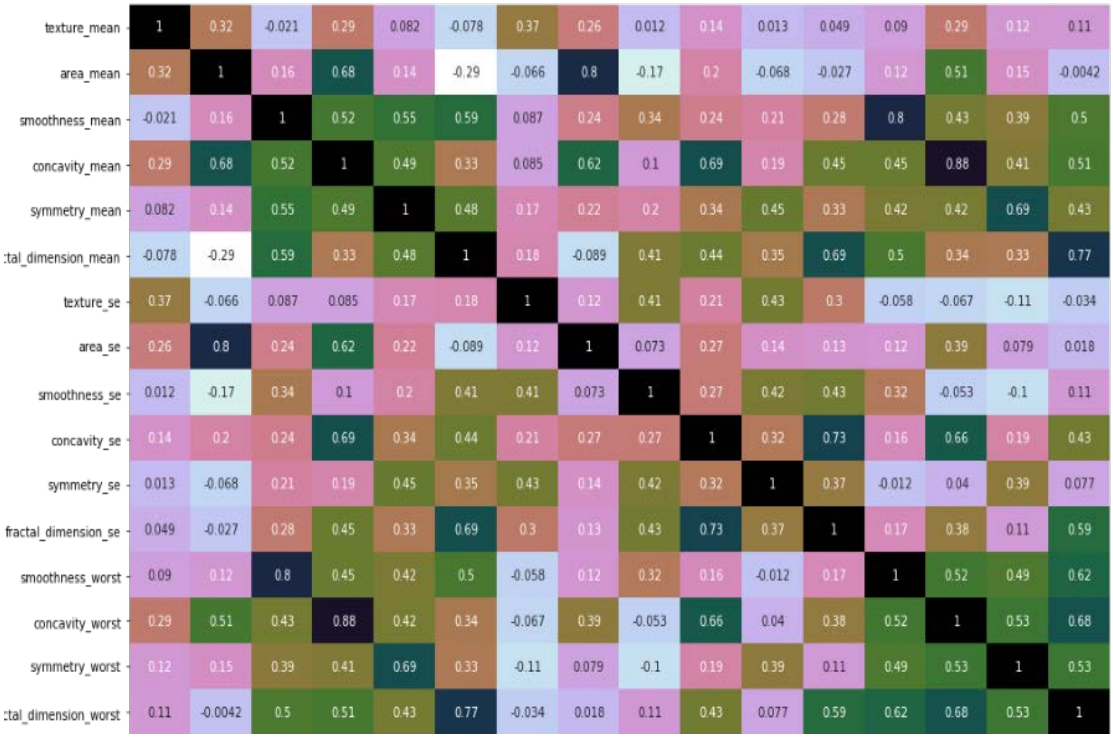


Fig. 6: Correlation Matrix after removal of highly correlated features

D. Model Training

After performing data preprocessing and data cleaning on our testing and training datasets, we now train different machine learning models. We use train_test_split to split our training data into test and training sets in a ratio of 20% to 80% for checking the model accuracy on the test set and finally train our data on the complete training dataset.

We use a k-folds algorithm to check the k-folds cross-validation score. We have not specified a cross-validation set separately as we are using k

folds cross-validation, which splits the training data in n folds and trains the data on n-1 folds and performs cross-validation on the remaining part. If the cross-validation score is high, then only we move forward to test the accuracy of our model on the test set.

After finally training our model, we use our model to predict whether the data in our testing dataset have benign or malignant tumors.

RESULTS AND CONCLUSIONS

A. Results

We trained our ML model on the training dataset using various classifiers to get our final result. We then checked the data on which we had to predict whether the data corresponded to malignant or benign tumors. We used the k folds algorithm for getting the cross-validation score, and the number of splits used in the algorithm was 5. The main basis used in this paper for the comparison of the various models was the cross-validation score and accuracy. The various scores such as precision, recall and f1 score and accuracy were computed using the confusion matrix.

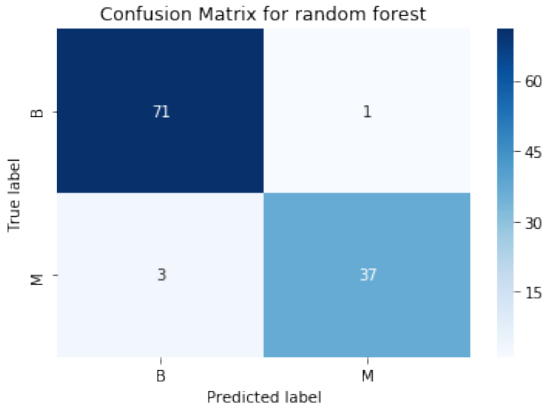


Fig. 7: Confusion Matrix for random forest classifier

The performance of various classifiers is shown in the table below as a comparative study:

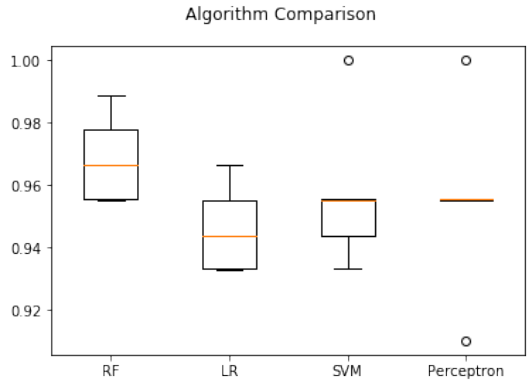


Fig. 8: Algorithm comparison on spread of accuracy scores across each CV fold

Model	Cross Validation Score	Percentage Accuracy
Logistic Regression	[0.96666667 0.97777778 0.98876404 0.94382022 0.98876404]	95.54
Support Vector Machines	[0.96666667 0.96666667 0.955056188 0.92134831 0.88505618]	93.75
Decision Tree	[0.9 0.92222222 0.98876404 0.8988764 0.91011236]	91.07
Random Forest	[0.96666667 0.95555556 0.97752809 0.95505618 0.88764045]	96.43
K Nearest Neighbors	[0.94444444 0.9555556 0.93254827 0.93254827 0.96629213]	90.18
Perception	[0.97777778 0.95555556 0.95505618 0.92134831 0.98876404]	94.64

Fig. 9: Comparison of Various Models

A. Conclusion

We found that by integration of multidimensional data with various feature techniques and classification algorithms helps to achieve great results in this domain. We also were able to achieve our goal of improving the prediction of breast cancer based on tumor features. We were successfully able to train various machine learning models by studying the various features in detail using various plotting and statistical techniques and eliminate highly correlated features for improving the model accuracy as much as we could. There is further scope of research in this field, which includes using various ML algorithms for the classification of benign and malignant tumors based on mammography images. Further research must also be done to make the model predictions with higher accuracy and f1 score on a dataset with more variables. On the ML side, bagging and boosting algorithms may also be applied for improving the tagging results. Furthermore, other medical data used for breast cancer detection, such as Breast ultrasound data, Breast MRI data and Breast ultrasound data, must also be used for improved results of detection.

REFERENCES

1. Ganesh N. Sharma, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and K. K Sharma," VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW",J Adv Pharm Technol Res. 2010 Apr-Jun; 1(2): 109–126.
2. Althuis, M. D., et al., Global trends in breast cancer incidence and mortality1973–1997. Int. J. Epidemiol. 34:405–412, 2005. April 1, 2005.
3. https://www.sciencedirect.com/science/article/pii/S1877050916302575
4. Joseph A. Cruz, David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", Sage Journals, First Published January 1, 2006
5. L. Rokach, O. Maimon, "Top – Down Induction of Decision Trees Classifiers – A Survey", IEEE Transactions on Systems

6. Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.
7. https://medium.com/datadriveninvestor/regression-in-machine-learning-296caae933ec
8. https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7
9. “Breast Cancer Wisconsin Dataset”, Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Original
10. Vikas Chaurasia, Saurabh Pal, “A Novel Approach for Breast Cancer Detection Using Data Mining Techniques”, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014
11. M S Chen, J Han, P S Yu,” Data mining: an overview from a database perspective IEEE Transactions on Knowledge and Data Engineering”, volume 8, issue 6, p. 866 - 883 Posted: 2002
12. Ya-Qin Liu; Cheng Wang ; Lu Zhang, “Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data”, IEEE Xplore, 14 July 2009.
13. Pedro Domingos, “The Role of Occam’s Razor in Knowledge Discovery”,
14. Dipanjan Sarkar, Raghav Bali, Tushar Sharma. "Practical Machine Learning with Python", Springer Science and Business Media LLC, 2018 2019
15. Tamires Brito-Sarracino, Moises Rocha dos Santos, Eric Freire Antunes, Iury Batista de Andrade Santos et al. "Explainable Machine Learning for Breast Cancer Diagnosis", 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 2019
16. Mourad Nasri, Mohamed Hamdi. "LTE QoS Parameters Prediction Using Multivariate Linear Regression Algorithm", 2019. 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2019
17. Hakan Gulmez. "chapter 9 Detection of Chronic Disease in Primary Care Using Artificial Intelligence Techniques", IGI Global, 2020
18. Aastha Sainger, Rishikesh Yadav, Pradnya Tipare, Samidha Waghralkar, Vimla Jethani, Amit Barve. "Chapter 4 Analysis of Light Pollution Prediction Using Mathematical Model and Machine Learning Techniques", Springer Science and Business Media LLC, 2020.