Breast Cancer Prediction from Finite Needle Aspiration Data using Machine Learning Algorithms

Sachin SK
Student- PGDM Business Analytics
Reva University
Bengaluru, India
sachink.ba02@reva.edu.in

Neelima Anasuri
Student PGDM Business Analytics
Reva University
Bengaluru, India
neelima.ba03@reva.edu.in

Nidhil CH
Student- PGDMM Business Analytics
Reva University
Bengaluru, India
nidhil.ba02@reva.edu.in

Abstract: Breast Cancer is cancer that develops in cells of the breasts. It is the second most common cancer that diagnosed in women after skin cancer. About 1 in 8 women in the US will develop invasive breast cancer in their lifetime. (https://www.cdc.gov/cancer/breast/statistics/index.htm). About 85% of breast cancers occur in women who have no family history of breast cancer. These occur due to genetic mutations that happen because of the aging process and life in general, rather than inherited mutations.

Breast Cancer diagnosis involves biopsy of the sample from the suspected area. Fine needle biopsy is the standard type of biopsy procedure that is followed. It is carried out by removing the tissue from the suspected area using a fine needle. But, many of the cases, the result of the fine needle biopsy alone will not be sufficient to confirm whether the tumor is benign or malignant. The next level of diagnosis involves core needle biopsy or surgical biopsy. There are many drawbacks of core needle and surgical biopsies. These are more invasive techniques and often involves with the chances of infection and bruising. Surgical biopsy is having longer and more uncomfortable recovery time and often the amount of tissue removed can also change the look and feel of the breast.

If we can predict whether a tumor is malignant or benign finite needle aspiration (FNA) data and identify the key attributes, then further complex diagnostic methods like surgical biopsy can be avoided. We have used the data available in Wisconsin breast cancer data set available in Kaggle for the analysis. We are applying standard Machine learning methods such as Logistic regression, Support Vector Machines, Decision trees on the data to get the insights and predict the type of cancer.

Keywords: Breast Cancer Diagnosis, Machine Learning, Finite Needle Aspiration Data Analysis.

I. Introduction

Breast cancer is the second most common cancer that diagnosed in Women. Breast cancer is caused by uncontrolled growth of cells in breast. The most common symptom of breast cancer is the development of a new lump or mass. But, all tumors may not be cancerous. They can be non-cancerous as well. They are called as a benign tumor. These kind of non - cancerous breast conditions are very common, and most women have them. The other types of tumors, which are cancerous are known as Malignant tumors. They invade and damage the surrounding tissues.

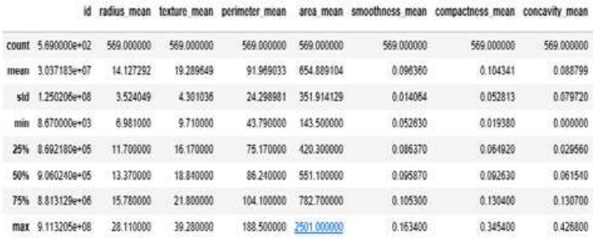
Breast Cancer diagnosis involves biopsy of the sample from the suspected area. Fine needle biopsy is the standard type of biopsy procedure that is followed. It is carried out by removing the tissue from the suspected area using a fine needle. But, many of the cases, the result of the fine needle biopsy alone will not be sufficient to confirm whether the tumor is benign or malignant. The next level of diagnosis involves core needle biopsy or surgical biopsy. There are many drawbacks of core needle and surgical. biopsies. These are more invasive techniques and often involves the chances of infection and bruising. Surgical biopsy is having longer and more uncomfortable recovery time and often the amount of tissue removed can also change the look and feel of the breast.

If we can predict whether the tumor is malignant or benign using finite needle aspiration(FNA) data and identify the key attributes, then further complex diagnostic methods like Surgical biopsy can be avoided.

This paper talks about the application of Machine learning algorithms in identifying the key attributes for diagnosis from the finite needle aspiration data and predicting the nature of the tumor, whether it is benign or malignant. The other focus area is the identification of the probability of cancerous tumors based on each feature using Bayes Theorem.

II. DATASET & VISUALISATION

We have used the WDBC dataset available at UCI Machine Learning Library(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)). The dataset contains 32 tumor features obtained from a digital image of Breast FNA of 569 subjects. The 32 features represent (a) 30 tumor features (b) ID number of subject (c) a class label, which denotes each subject is benign or malignant. Ref Table 1. The dataset is normalized before modelling as the measurements are in different scales.



	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800

Table 1: Data set for Variables

Most of the variables are co-related so we can eliminate highly co-related variables, which helps in stable predictions. Red indicates high co-relation and blue indicates less co-related. Red ones can be eliminated for the accurate prediction. Ref Fig 1.

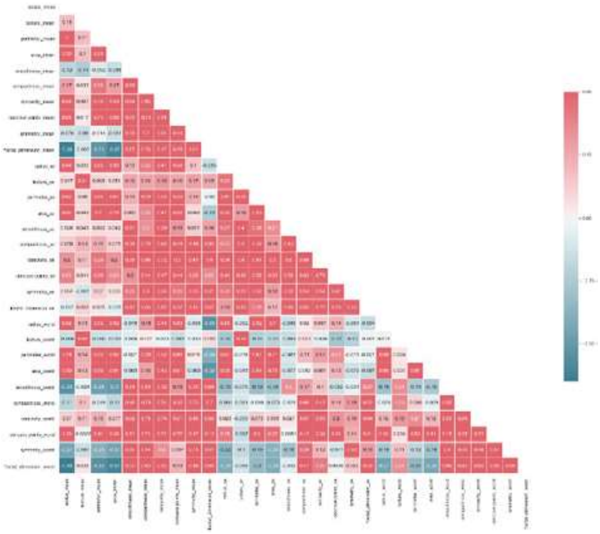


Fig1: Explains Kernel Density Estimation (KDE) plots to check the distribution of malignant and begin cases for various features.

Histography distribution is considered for all the variables, the fig.2 is shown for radius_mean, which shows that it is right skewed. So, we have taken the log to eliminate skewness.

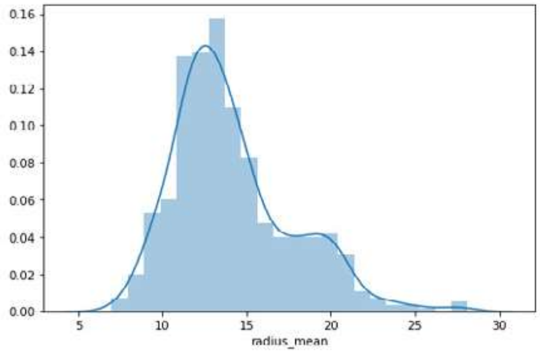


Fig 2: Distribution of radius mean

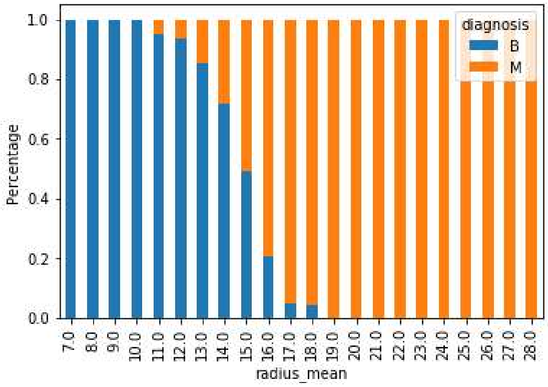


Fig 3: Bivariant Analysis on radius mean vs Diagnosis

From the bivariant graph, it clearly shows if the tissue radius is more than 19 then it is confirmed for cancer. Ref fig 3.

III. Related Work

There exist many methods for breast cancer classification using machine learning.

Quinlan (1996) developed a method using C4.5 decision tree. They used 10-fold cross-validation. Bhardwaj and Tiwari (2015) proposed a method using neural networks. In 2010, Asha Gowda Karegowda et al. Asha Gowda Karegowda et al.,2010 proposed a wrapper approach with genetic algorithm for generation of a subset of attributes with different classifiers such as C4.5, Naïve Bayes, Bayes Networks and Radial basis functions. The above classifiers experiment on the datasets Diabetes, Breast cancer, Heart Stat log and Wisconsin Breast cancer. Vikas Chaurasia and Saurabh Pal11 compare the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores the accuracy of 96.84% in Wisconsin Breast Cancer (original) datasets.

IV. Method of approach

To reduce the features, the technique of principal component analysis was applied. We reduced the dimension to 10 features that could explain 95% of the variance. Ref fig.4

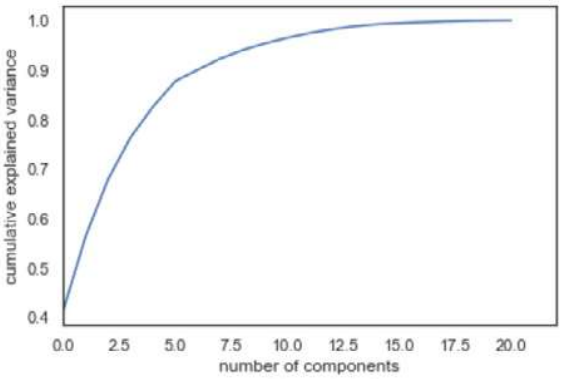


Fig4: AUC Graph – Area Under Curve

By plotting the correlation matrix, we eliminated some of the highly correlated features and statistical analysis was performed to establish the values of each feature that could potentially indicate signs of malignancy. Potential indicators were derived from the Bayes table

Machine learning algorithms such as Decision Tree, SVC, KNN, Logistic Regression were applied and then their corresponding accuracy and recall were recorded. Hyperparameter tuning was also performed for each model to improve accuracy. Finally, the predictions from all the classifiers were combined to create a Vote classifier which has two variants a) hard vote where the majority vote decides the class b) soft vote where the class probabilities are aggregated, and the class is predicted based on who gets the highest-class probability.

V. EVALUATION AND RESULTS

Out of all the machine learning algorithms, Logistic regression performed better than other sophisticated models such as SVM, Decision Tree etc. Logistic regression recorded an accuracy of 97.0% and a sensitivity of 94% for the Malignant class. Refer Table 3.

Increase in symmetry_worst,symmetry_sd error are likely to increase in the case of benign cancer whereas radius mean and concavity worst are likely to increase in malignant cases

The mean and median smoothness of malignant cells on an average is greater than the mean & median smoothness of benign category cells.This implies mean smoothness could be an important feature for differentiating between benign and malignant cells .

Radius mean is highly correlated to texture mean, radius_worst, concave points etc

Odds ratio calculated from the logit classifier also reveals that radius means, concave points are the top predictors for a tumor to be considered malignant. Refer to Table 2.

The variable importance plot indicates that fractal dimension mean is the most important feature for splitting the data followed by concavity_worst, ,smoothness_mean etc Refer Fig 5.

FEATURE	ODDS RATIO
radius_mean	[-1.70710281802]
radius_sd_error	[-0.107616790861]
radius_worst	[-0.0874010129088]
teature_mean	[0.00672901063862]
texture_sd_error	[0.124733105678]
perimeter_mean	[0.50381484364]
perimeter_sd_error	[0.263108156184]
perimeter_worst	[0.247641034584]
area_mean	[0,020979458427]
area_sd_error	[-0.044216787569]
area_worst	[-0.962745090388]
smoothness_mean	[0.0887884558258]
smoothness_sd_error	[0.105130076283]
smoothness_worst	[0.00835197816851]
compactness_mean	[-0.112863294301]
compactness_sd_error	[0,0390862520374]
compactness_worst	[0.0293612797992]
concavity_mean	[0.0306929370696]
concavity_sd_error	[-0.0090512484805]
concavity_worst	[-1.42199601442]
concave_points_mean	[0.28983482624]
concave_points_sd_error	[0.2558782071]
concave_points_worst	[0.0189721938966]
symmetry_mean	[0.214596498501]
symmetry_sd_error	[0.99194261563]
symmetry_worst	[1.48331571047]
fractal_dimension_mean	[0.542716089865]
fractal_dimension_sd_error	[0.601972180877]
fractal_dimension_worst	[0.105876809414]

Table 2

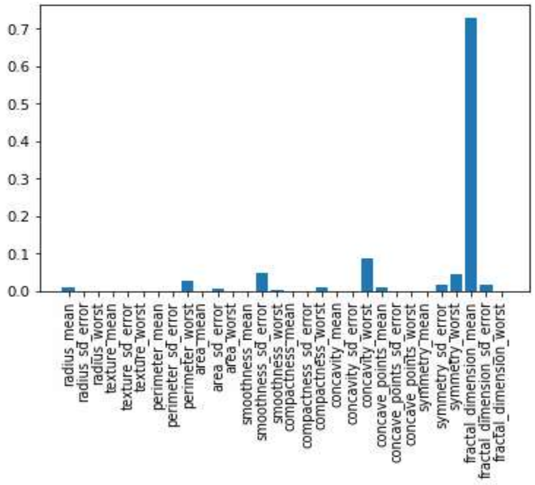


Fig 5-Variable Importance Graph

Classifier	Accuracy	Precision	Recall	ROC_AUC
Logistic Regression	0.97	0.97	.94	0.99
Random Forest	0.92	0.9	0.88	0.98
Support Vector Machine	0.95	0.96	0.91	0.98
Gaussian Naïve Bayes	0.9	0.92	0.82	0.97
K-Nearest Neighbors	0.95	0.96	0.9	0.98
Votes Classifier	0.98	0.97	0.95	0.99

Table 3.

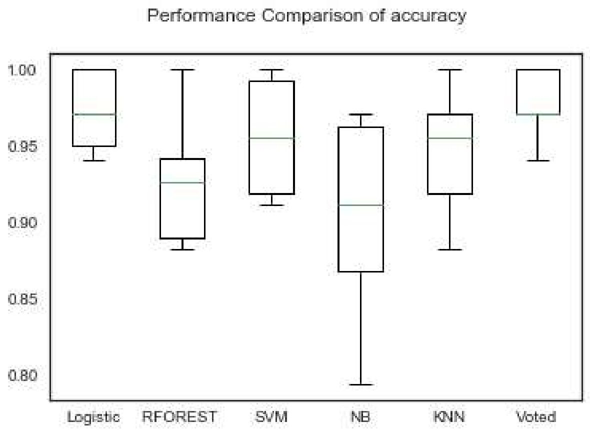


Fig 6-Performance Comparison Accuracy

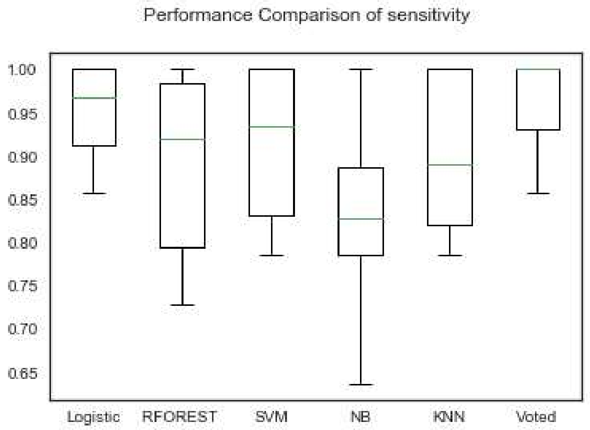


Fig 7- Recall Comparison

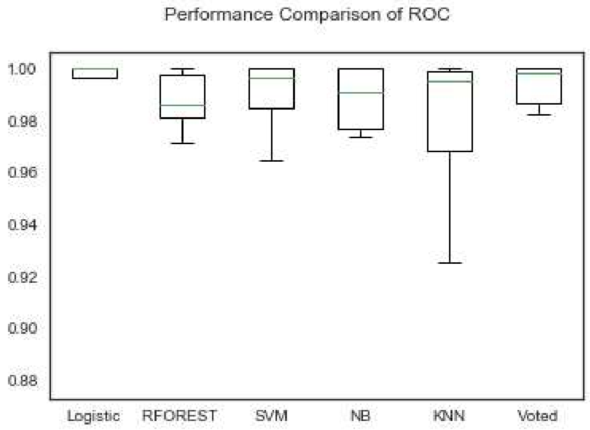


Fig 8- ROC Comparison

VI. Conclusion

As per the analysis done on the dataset, we could derive the potential indicators of malignant cancer. Using feature selection and a vote classifier, the highest accuracy of 98% was achieved. In addition to that, the highest precision and recall of 97% and 95% was achieved. The prediction accuracy of the logit classifier is strong. In the future work, we can add, and cross-validate regularization to the logit classifier. These models are strong & reliable assets when used along with clinical investigation to differentiate malignant and benign tumors.

VII. REFERENCES

[1] Dora, Lingraj; Agrawal, Sanjay; Panda, Rutuparna; Abraham, Ajith. Optimal breast cancer classification using Gauss-Newton representation-based algorithm. *Expert Systems with Applications*, *Volume 85* – Nov 1, 2017

[2] Nilashi, Mehrbakhsh; Ibrahim, Othman; Ahmadi, Hossein; Shahmoradi, Leila, A knowledge-based system for breast cancer classification using fuzzy logic method, *Telematics and Informatics, Volume 34 (4) – Jul 1, 2017*

[3] Bache K.,Lichman M.,2013. *UCI Machine Learning Repository.*

[4] J.R. Quinlan,"c4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc, 1992.

[5] Quinlan, J.R, "Induction of decision trees". *Journal of Machine Learning 1(1986) 81-106.*

[6] Bhardwaj A.Tiwari A.,2015.Breast cancer diagnosis using genetically optimized neural network model. *Expert Syst. Appl.42(10),4611-4620.*

[7] Bichen Zheng Sang WonYoonSarah S.Lam. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms., *Expert Systems with Applications. Volume 41, Issue 4, Part 1, March 2014, Pages 1476-1482*

[8] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," *International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 3, Issue. 1, January 2014 – 22 vol. 3, no. 1, pp. 10–22, 2014.*

[9] Chen, Hui-Ling; Yang, Bo; Liu, Jie; Liu, Da-You, : "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", *Expert Systems with Applications, Volume 38 (7) – Jul 1, 2011*

[10] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16,906-914.*

[11] Karen E. Shafer-Peltier Abigail S. Haka Maryann Fitzmaurice Joseph Crowe Jonathan Myles Ramachandra R. Dasari Michael S. Feld; Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo;, *Journal Of Raman Spectroscopy July2002.*

[12] RudySetiono;, Generating concise and accurate classification rules for breast cancer diagnosis;, *Artificial Intelligence in Medicine Volume 18, Issue 3, March 2000, Pages 205-219*