International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2021
https://doi.org/10.21276/ijircst.2021.9.3.4
Article ID IRP1148, Pages 27-30
www.ijircst.org

Breast Tumor Detection Using Classification Algorithm

Mitali[1], Dr. Aman Jatain[2], Swati Gupta[3]

[1]Department of CSE, Amity University Haryana, India
[2]Assistant Professor, Amity University Haryana, India
[3]Assistant Professor, Amity University Haryana, India

Correspondence should be addressed to Mitali; mital.milansinha@gmail.com

ABSTRACT- Breast cancer is the most common cancer in women. It is the leading cause of cancer death in developing countries and the second leading cause of cancer death in women in the United States, behind only lung cancer. Females are more likely to develop breast cancer. However, in a few instances, it is clear that males have also been affected. Breast cancer has been discovered. [1]

Breast tumors may be either cancerous or non- cancerous. Benign tumors are easily treated with doctor-prescribed medications. Malignant tumors are a sign of a high risk of breast cancer and should be removed as soon as possible. Early identification and treatment of tumor lowers the risk of cancer-related death. The survival rate of breast cancer patients in America has been reported to be 90% in recent years, but it is as low as 60% in India. [2]

The key cause of this shortcoming is late tumor diagnosis and classification, which causes treatment delays. Using machine learning to perform a task will reduce the workload of physicians and radiologists. According to research, doctors can only diagnose breast cancer with a 79 percent accuracy rate, while machines can diagnose it with a 91 percent accuracy rate. Early diagnosis is the best way to increase the chance of treatment and chance of survival.

Keywords: Malignant, Benign, Diagnosis, Radiologist, Survival, Tumors.

I. INTRODUCTION

The rapid development of machine learning continues to pique the interest of the medical imaging community in using these techniques to increase cancer screening accuracy. [3]

Breast cancer is the second leading cause of cancer death among women in the United States, and screening mammography has been shown to minimize the number of cases. Breast cancer is the product of malignant tumors not being treated at the appropriate time, while benign tumors are expected to stay benign forever, except in the long term. Benign tumors develop slowly and do not spread, but they can become malignant in a short period of time. Malignant tumors can spread quickly and invade other organs, destroying surrounding normal tissues. [4] Machine learning is an artificial intelligence subfield that allows computers to create models from sample data in order to automate decision- making processes using data inputs. My study is focused on developing a model that classifies tumors as benign or malignant using a sample dataset to automate decision-making based on new tumor patient input results. Machine learning is well known for requiring large training datasets to be accurate. To improve the accuracy of breast cancer classification algorithms, it is important to use both the few completely annotated datasets and larger datasets marked with only the cancer status of each data point. [5]

An accurate model assists physicians and radiologists in the classification of tumors and the identification of breast cancer patients with greater precision than radiologists can manually provide. Data from a new patient's tumor can be fed into a computer to determine if it is benign or malignant. It will not only minimize workload, but will also improve tumor recognition and breast cancer detection accuracy. [6]

II. METHODOLOGY

A. Defining the problem statement

Classification of tumor as benign or malignant.

B. Importing dataset

Dataset consist of various attributes that may or may not affect our problem statement. We have to select them which according to us will be an aid in solving the problem. Selection will be based on further steps done in processing of data. Breast tumor benign and malignant cases were chosen from a dataset with a significant amount of data entry. This technique can assess breast masses quickly and identify them in an automated manner. [7]

I used a dataset from UCI which had 569 instances of tumors: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

I sorted the target names (i.e. malignant vs. benign) and the 30 main features of a tumor using the Breast Cancer Dataset from the UCI archive. This information was entered into a data frame that I will use for the analysis.

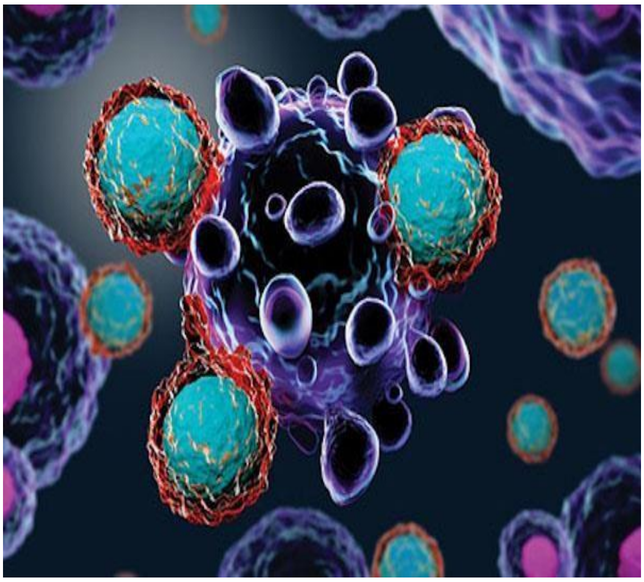


Fig. 1: Breast cancer

C. *Visualizing the data*

Visualizing the data helps to select those attributes that affect our problem statement. The texture function provides the structural and dimensional detail, as well as the image's strength. Typically, a feature detector can tell you whether or not a specific feature is present in the picture. The field, convex hull, and centroid are all structural features that provide details about the feature's structure and orientation. [8]

The dataset's mean, median, and standard deviation always provide useful information about the dataset and its distribution. It is classified as a statistical characteristic. After that, I used multiple methods from the seaborn library to visualize the data, including pair plots and heatmaps. These factors aided me in deciding which main features I would use in my experiment, and I chose mean smoothness and mean field. [9]

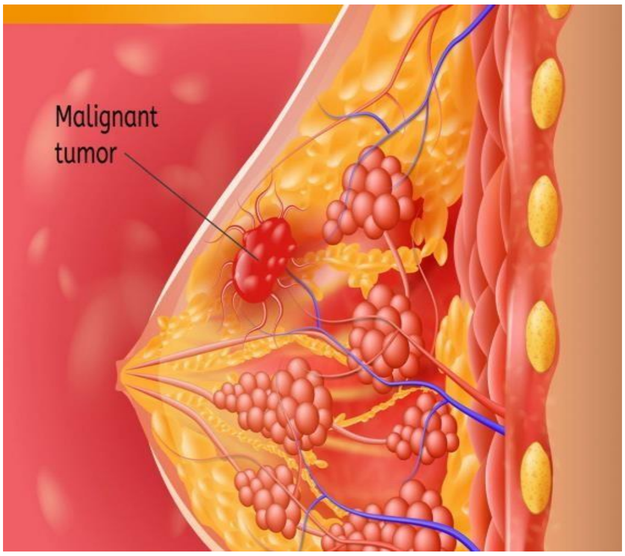


Fig. 2: Malignant tumor

D. *Model Training*

Training the model is very crucial. There are various algorithms to do so. We must find the one that gives the best result on validation and test data. The model will then be trained by generating a train data set and leaving the rest for testing. I imported the train test split from the scikit learn library. I trained the model with 80% of the data and tested it with 20% of the data. I used Standard scaler mixture for feature scaling. Then, using the SVM model without any optimization, I made a heat map using an uncertainty matrix. As can be seen in the picture, the model did not correctly classify the malignant tumors and will need to be improved. [10]

Overfitting and underfitting was also checked using validation method. Support vector Machines proved to provide the highest accuracy of about 96% which was greater than other machine learning classification algorithms likes logistic regression, K- nearest neighbor, Kernel SVM, Decision tree algorithm etc.

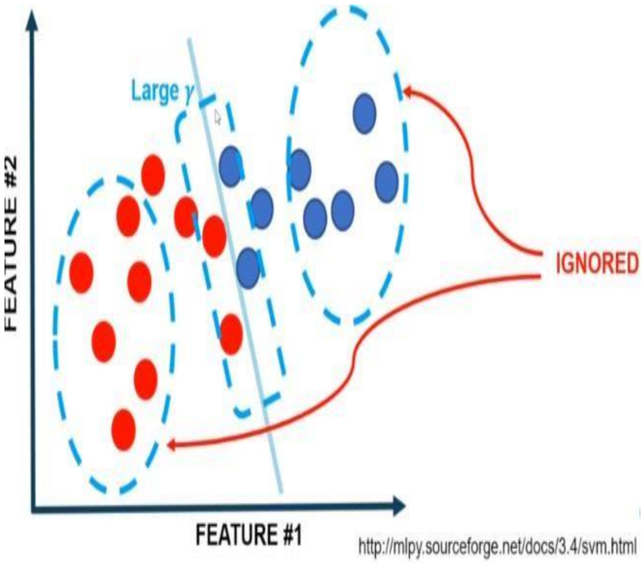


Fig. 3: Large Gamma

E. *Evaluating the model*

This step is very important as it informs us whether a particular algorithm is fruitful and is serving its purpose or not. [11]

The feature set is then prioritized using three strategies in order to reduce the classifier model complexity and computational time- Filter is a feature selection process that does not evaluate any classifier algorithms. Chi-square, Fisher ranking, and other examples.

F. *Wrapper*

It is focused on the success of a specific classifier in terms of evaluation. Embedded- It uses both a filter and a wrapper to create a classifier. After that, I optimized the result using the SVM method and two parameters.[12]

The C parameter, which governs trade-off between training points by "penalizing" the dataset for misclassification, was the first one I used. The boundary lines are smooth with a small value of C because the cost of misclassification is low, but as the value of C increases, the boundary lines become established because the cost of misclassification or the "penalty" is high. I used the Gamma parameter for more optimization.[13]

The Gamma parameter determines how far a single training set's effect extends. In other words, it determines a dataset's distribution. We will concentrate on the points closer to the hyper plane, or the line dividing the malignant and benign datasets, if we use a broad gamma.

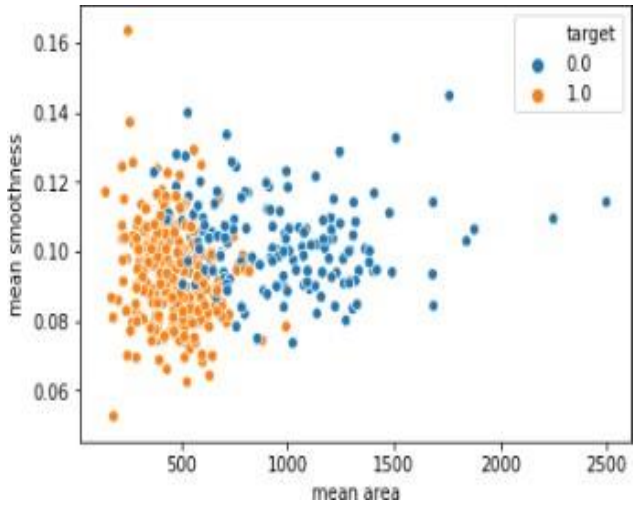


Fig. 4: Non- normalized data

Improving the model - Improving the model to achieve high train, validation and test accuracy and slight difference between them is a crucial and important part of solving the problem statement. In this case study, the misclassification was a Type 1 Error. In other words, three benign tumors were mistakenly identified as malignant. Type 2 errors, which are much more dangerous, were not present. This would mean that benign tumors would be categorized as benign and malignant tumors would be classified as malignant. Doctors' workloads will be reduced, and lives will be saved, particularly in developing countries. Combining computer vision and machine learning methods to specifically classify cancer using a tissue image will enhance the technique even further. [14]

F

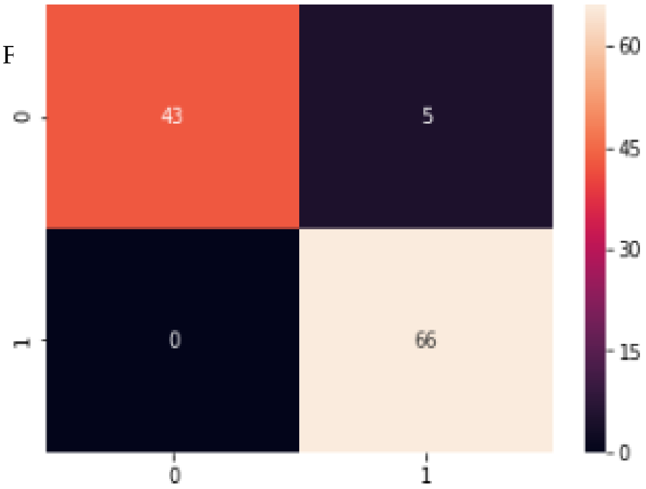
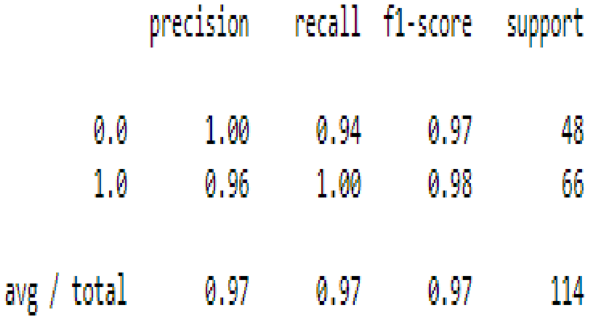


Fig. 5: Improved confusion matrix

I agreed on the adjustments I need to make in order to improve my data. I started by using the Unity-Based normalization method to normalize my data. The discrepancies between the datasets after normalization are mentioned below. [15]

This normalized dataset was used to retrain my model using the SVM function. My accuracy had improved significantly to 96 percent as a result of this. An uncertainty matrix helped me imagine this in another heatmap. Following that, I use the SVM model for both the C and Gamma parameters. Now I had to select the best C and Gamma values, which would be difficult in such a huge dataset. I found a feature in scikit-learn called Grid Search CV that would help me find the best parameters in a grid. I chose a kernel of radial basis function and set some ranges for C and Gamma in this grid. Now I'd fit the normalized training data to the grid. Then I used the grid best params_ function to get the best values, which gave me a C value of 10 and a gamma value of 0.1. Finally, I employed the use of a feature grid. To get the best predictions, predict on my normalized results. I used a Confusion Matrix to plot this in a heat map.



```
             precision    recall  f1-score   support

        0.0       1.00      0.94      0.97        48
        1.0       0.96      1.00      0.98        66

avg / total       0.97      0.97      0.97       114
```

Fig. 6: Final precision score

III. CONCLUSION

I was able to see my performance by making a classification report that demonstrated my accuracy.

I achieved a precision of 34% for the initial non-optimized results. I got a precision of 96 percent for the second normalized results.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] J. E. T. Akinsola, S. O. Kuyoro, O. Awodele& F. A. Kasali, Performance Evaluation of Supervised Machine Learning Algorithms Using Multi-CriteriaDecision Making Techniques. International Conference on Information Technology in Education and Development (ITED) Proceedings, 17 – 34, 2019.

[2] Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong, "Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation", Biomedical Research, vol. 29, no. 10, 2018.

[3] Mohamad Mahmoud Al Rahhal, "Breast Cancer Classification in Histopathological Images using Convolutional Neural Network", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no.3, 2018.

[4] Vikas Chaurasia, Saurabh Pal and BB Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational

Technology 2018, vol. 12, no. 2, pp. 119-126,2018.

[5] et al., "A parallel random forest algorithm for big data in a Spark cloud computing environment", IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4, pp. 919-933, 2017.

[6] Chetak Kandaswamy et al., "High-content analysis of breast cancer using single-cell deep transfer learning", Journal of biomolecular screening, vol. 21, no. 3, pp. 252-259, 2016.

[7] Pedro Henriques Abreu et al., "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review", ACM Computing Surveys (CSUR), vol. 49, no.3, 2016.

[8] T Choudhury, V Kumar, D Nigam and B Mandal, "Intelligent classification of lung & oral cancer through diverse data mining algorithms", International Conference on Micro-Electronics and Telecommunication Engineering, 2016.

[9] M. H. Yap et al., "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks", IEEE Journalof Biomedical and Health Informatics, vol. 22, no. 4, pp. 1218-1226, July 2018.

[10] Silva, J., Lezama, O.B.P., Varela, N., Borrero, L.A.: Integration of data mining classification techniques and ensemble learningfor predicting the type of breast cancer recurrence. In: Miani, R., Camargos, L., Zarpelão, B., Rosas, E., Pasquini, R. (eds.) GPC 2019. LNCS, vol. 11484, pp. 18–30. Springer, Cham (2019).

[11] Ojha U., Goel, S.: A study on prediction ofbreast cancer recurrence using data mining techniques. In: 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, pp. 527–530, 2017.

[12] Pritom, A.I., Munshi, M.A.R., Sabab, S.A., Shihab, S.: Predicting breast cancer recurrence using effective classification and feature selection technique. In: 19th International Conference on Computer and Information Technology (ICCIT), pp. 310–314. IEEE (2016).

[13] Asri, H., Mousannif, H., Al, M.H., Noel, T.: Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Sci. 83, 1064–1069 (2016)

[14] Darrab, S., Ergenc, B., Vertical pattern mining algorithm for multiple support thresholds. In: International Conference on Knowledge Based and Intelligent Information and Engineering (KES), Procedia Computer Science, vol. 112, pp. 417–426 (2017).

[15] Mohammed S.A., Darrab S., Noaman S.A.,Saake G. (2020) Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan Y., Shi Y., Tuba M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234. Springer, Singapore

ABOUT THE AUTHORS

Mitali, Dedicated, efficient and hardworking final year student of MCA.

Dr. Aman Jatain, Dedicated, efficient and goal-oriented educationist with rich academic experience of 12 years in renowned Industrial Organizations/ Colleges/Institutions. She has dome B.Tech, M.tech and PhD. in computer science.

Swati Gupta, Dedicated, efficient and goal-oriented educationist with rich academic experience of 12 years in renowned Industrial Organizations/ Colleges/Institutions. She is pursuing her PhD. in database.