International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Breast Biopsy Reduction Using Machine Learning Algorithms

Manan Kohli[1]

[1](B.Tech - Student/ ECE, Maharaja Agrasen Institute of Technology)

Abstract - —*One of the most effective ways for prevention of breast cancer is the data derived through mammograms. The next step after a doctor assesses data derived through mammograms is for women to undergo a breast biopsy. An enormous amount of anxiety and surgery arises from false positives among the mammogram results that are not healthy for the mental state of the patient. This paper focuses on reducing the need for unnecessary breast biopsies called upon for women by using the mammographic mass data and applying several machine learning techniques on them to enquire which provides the most accurate results. On comparison, we get 80.7% of accuracy by using supervised machine learning techniques of which logistic regression proves to be the best*

Key Words: **Biopsy, Machine learning, Random forest, K-fold cross-validation, Support vector machine, Logistic regression, K-nearest neighbors, Naive Bayes**

1.INTRODUCTION

Breast cancer is one of the most often researched mammary gland disorders. It has been reported in the literature that a 7:10 ratio of females in developed countries suffer from breast cancer. Thus, It is inferred that 10% of the female population suffers with breast cancer at some point in their lives. Early detection of breast cancer in mammography using machine learning can aid in proper treatment, lowering the risk of death. Chemotherapy is used in the treatment of breast cancer, which has severe side effects. Image enhancement facilitates in the proper detection of mass in mammography images, which improves cancer recognition rate. Erroneous detection occurs due to differences in the size, location, and structure of masses, as well as the poor contrast of mammography. The enhanced mammography images obtained using histogram equalization [1], contrast stretching, and other image enhancement methods [2], [3], [4] but the resulted images shows artifacts in the enhanced image and mass is not detected precisely under different acquisition conditions [5], [6]. Thus, machine learning approach is employed to improve the accuracy of detection of cancer in captured mammography images. The effectiveness of the machine learning approach is to extract the characteristics or features from the captured mammography. The breast cancer treatment is influenced by density [7]. Sometimes occurrence of high density tissue does not means cancer [8]. The large data set requires that assists practitioners in the early diagnosis and treatment of breast cancer in order to minimize mortality [9]

To model the data and anticipate the classification outcomes, several machine learning techniques may be employed and compared that determine the exact location, features and size [10]. To categorize the candidate areas as masses or non-masses, the different learning algorithm is utilized. Any area that corresponds to a tumor, whether malignant (cancerous) or benign(non cancerous), is considered a mass in this context [11]. As a consequence, learning approaches helps to improve accuracy, including decision trees [12], Random Forest [13], K-Fold Cross Validation [14], Support Vector Machine [15], [16], Logistic Regression [17], K-Nearest Neighbours [18], and Naive Bayes [19], [20]. Learning results the difference between cancerous or non cancerous tissue [21], [22]. Benign tumors develop slowly and seldom spread. There are differenttypes of Benign tumors like Adenomas, Fibromas, Hemangiomas, Lipomas, Meningiomas, Myomas, Nevi:, Neuromas, Osteochondromas, Papillomas [23].Benign tumors are caused by a variety of factors, including genetics, diet, stress, a particular region of trauma or damage [24]. These problem in general resolved by using standard treatment plan. Malignant tumors can spread throughout the body, infiltrate and destroy adjacent normal tissues, and develop fast [25].The determination and identification of malignant tumors requires faster detection . To remove and early detection of malignant tumor leads to the reduction of biopsy. There are different forms of malignant tumors include carcinoma, Sarcoma, Leukemia, Lymphoma and multiple myeloma, Central nervous system cancers. Several available therepies including chemotherapy, radiation therapy, and immunotherapy, are all options for treating malignant tumors [26].

The main contribution of this paper is to compare different existing learning approaches and also compare their efficiency. These algorithms differentiate the malignant tumor with benign tumor based on four features. This is a challenging task to differentiate between malignant and benign tumor. The quality of results can be used for further investigation of disease that depends on different approaches and different number of characteristics are sought in order to better tissue characterisation and assist the categorization of these tissues as normal and masses [12]. The organization of paper is as follows is as follows: Section I gives introduction and section II describes how data acquisition uses standard data set and sectionIII and sectionIV presents classification for different machine learning methods and measure its accuracy respectively. In section

V, results are discussed and concluding remarks gives in section VI.

2. DATA ACQUISITION AND PROCESSING

The mammographic mass data that is used for this paper was obtained from the UCI Machine Learning Repository [27]. This data was donated by researchers from Germany where data was collected at the Institute of Radiology of the University Erlangen-Nuremberg and contains several attributes that are used to predict the severity (benign or malignant) of mammographic mass. It includes a BI-RADS assessment, the patient'sage, three BI-RADS attributes, and the ground truth (the severity field) for 516 benign and 445 malignant masses found on full field digital mammograms. BI-RADS assessment: 1 to 5 (ordinal, nonpredictive) Age: patient's age in years (integer)

Shape: mass shape: round=1 oval=2 lobular=3

irregular=4 (nominal)

Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5

(nominal)

Density: mass density high=1 iso=2 low=3 fatcontaining=4 (ordinal)

Severity: benign=0 or malignant=1 (binomial, goal field)

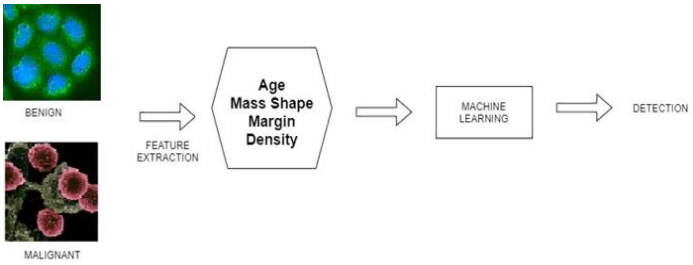
The Breast Imaging-Report and Data System (BI-RADS) is a risk assessment tool that denotes values by the physician after examination of the mammograms. As this is a non-predictive attribute and may contain biases through the physician, it is not used in our comparative study. The data set contains missing values that are randomly distributed and is dropped.

3. CLASSIFICATION APPROACH

In this study, the attributes of the mammography mass data are split into a training and testing set. The four predictive attributes, the patient's age, mammography mass shape, margin and density are used for predicting whether a given mammography mass will be benign or malignant as shown in Fig. 1. In the data set that is being used, missing fields were present but had no correlation among them and were randomly distributed, hence we were able to drop the missing fields. Some of the models used in the study required the input data to be normalised. Preprocessing was applied on the attribute data to get a set of normalized data. Different machine learning techniques are used for the classification of the data of which 75% is used as training data and the rest over which the models are applied.

4. CLASSIFICATION ACCURACY

The comparative study saw different accuracy's for different techniques of machine learning applied on the attribute data. Decision trees are inherently a greedy algorithm and it minimizes entropy. As Decision Trees are very susceptible to over-fitting, several alternate Decision Trees are used to let them vote on the final classification using Random Forests. K-fold Cross Validation was also used to prevent over-fitting after the



entire data was divided into equal k sets. Support Machines work well for higher dimensional data having a lot of features as depict from Table I. Different kernels for SVM's were used to predict the maximum accuracy. However, logistic regression was seen to provide the highest accuracy among all the other techniques used as observed in Table II.

5. RESULT

After all the machine learning techniques were applied to the attribute data to predict whether a given set of mammographic mass data would prove to be benign or malignant, different accuracies were observed. The three algorithms with lowest accuracy were Decision Trees, Random Forests and K-Fold Cross Validation. K-Nearest Neighbours, Naive Bayes and different kernels of Support Vector Machine were in the second category with moderate accuracy. The model with the highest accuracy came out to be Logistic Regression with an accuracy of 80.7%. The limitation of this study is the size of data used.

The number of samples and attributes used for training and testing is low. The analysis of data with respect to clinical settings should be carried out with a larger datas

TABLE I

ACCURACY OF DIFFERENT ALGORITHMS

S. No.	Technique	Accuracy
1	Decision Trees	73.55%
2	K - Fold Cross Validation	73.73%
3	Random Forest	74.21%
	Support Vector Machine Linear Kernel	79.75% 80.12%

4	RBF Kernel Sigmoid Kernel Poly Kernel	 74.57% 79.03%
5	Logistic Regression	80.72%
6	K-Nearest Neighbours	79.15%
7	Naive Bayes	78.55%

TABLE II

LOGISTIC REGRESSION MATRIX

Severity	Precision	Recall	f1-score
Benign	0.87	0.76	0.81
Malignant	0.78	0.88	0.83

6. CONCLUSION

In this paper, we compared different machine learning algorithms on the UCI Mammographic Mass data set. The aim of the paper was to find a machine learning technique that would provide the most accurate results in order to reduce unnecessary breast biopsies. According to the results, Logistic Regression, being the simplest algorithm among all the models, gave the highest accuracy of 80.7%. These models can aid physicians in deciding whether to undertake a breast biopsy or a short-term follow-up examination on a concerning lesion found on a mammogram.

REFERENCES

[1] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," Computer vision, graphics, and image processing, vol. 39, no. 3, pp. 355–368, 1987.

[2] R. Sharma, M. Ravinder, N. Sharma, and K. Sharma, "An optimal remote sensing image enhancement with weak detail preservation in wavelet domain," Journal of Ambient Intelligence and Humanized Computing, pp. 1–12, 2021.

[3] N. Sharma and O. P. Verma, "Gamma correction based satellite image enhancement using singular value decomposition and discrete wavelet transform," in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. IEEE, 2014, pp. 1286–1289.

[4] ——, "A novel fuzzy based satellite image enhancement," in Proceedings of International Conference on Computer Vision and Image Processing. Springer, 2017, pp. 421–428.

[5] R. C. Gonzalez, R. E. Woods et al., "Digital image processing second edition," Beijing: Publishing House of Electronics Industry, vol. 455, 2002.

[6] O. P. Verma and N. Sharma, "Efficient color cast correction based on fuzzy logic." Journal of Engineering Science & Technology Review, vol. 10, no. 3, 2017.

[7] M. Kashif, K. R. Malik, S. Jabbar, and J. Chaudhry, "Application of machine learning and image processing for detection of breast cancer," in Innovation in Health Informatics. Elsevier, 2020, pp. 145–162.

[8] N. Mendhiratta, A. B. Rosenkrantz, X. Meng, J. S. Wysock, M. Fenstermaker, R. Huang, F.-M. Deng, J. Melamed, M. Zhou, W. C. Huang et al., "Magnetic resonance imaging ultrasound fusion targeted prostate biopsy in a consecutive cohort of men with no previous biopsy: reduction of overdetection through improved risk stratification," The Journal of urology, vol. 194, no. 6, pp. 1601–1606, 2015.

[9] J. R. Harris, M. E. Lippman, U. Veronesi, and W. Willett, "Breast cancer," New England Journal of Medicine, vol. 327, no. 5, pp. 319–328, 1992.

[10] M. Hanmandlu et al., "A new entropy function and a classifier for thermal face recognition," Engineering Applications of Artificial Intelligence, vol. 36, pp. 269–286, 2014.

[11] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in 2015 international conference on digital image computing: techniques and applications (DICTA). IEEE, 2015, pp. 1–8.

[12] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E. R. ´Denton, and R. Zwiggelaar, "A novel breast tissue density classification methodology," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp.55–65, 2008.

[13] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in International workshop on machine learning and data mining in pattern recognition. Springer, 2012, pp. 154–168.

[14] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," Journal of machine learning research, vol. 5, no. Sep, pp. 1089–1105, 2004.

[15] T. Subashini, V. Ramalingam, and S. Palanivel, "Automated assessment of breast tissue density in digital mammograms," Computer Vision and Image Understanding, vol. 114, no. 1, pp. 33–43, 2010.

[16] K. Taifi, N. Taifi, M. Fakir, S. Safi, and M. Sarfraz, "Mammogram classification using nonsubsampled

contourlet transform and gray-level co-occurrence matrix," in Critical Approaches to Information Retrieval Research. IGI Global, 2020, pp. 239–255.

[17] R. E. Wright, "Logistic regression." 1995.

[18] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[19] I. Rish et al., "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41–46.

[20] M. Bahl, R. Barzilay, A. B. Yedidia, N. J. Locascio, L. Yu, and C. D. Lehman, "High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision," Radiology, vol. 286, no. 3, pp. 810–818, 2018.

[21] H. Shen, T. Liu, J. Cui, P. Borole, A. Benjamin, K. Kording, and D. Issadore, "A web-based automated machine learning platform to analyze liquid biopsy data," Lab on a Chip, vol. 20, no. 12, pp. 2166–2174, 2020.

[22] P. Stelzer, O. Steding, M. Raudner, G. Euller, P. Clauser, and P. Baltzer, "Combined texture analysis and machine learning in suspicious calcifications detected by mammography: Potential to avoid unnecessary stereotactical biopsies," European Journal of Radiology, vol. 132, p. 109309, 2020.

[23] R. Woodhams, K. Matsunaga, S. Kan, H. Hata, M. Ozaki, K. Iwabuchi, M. Kuranami, M. Watanabe, and K. Hayakawa,"Adc mapping of benign and malignant breast tumors," Magnetic resonance in medical sciences, vol. 4, no. 1, pp.35–42, 2005.

[24] F. Zhao, L. Zhao, L. Wang, and H. Song, "A collaborative lshade algorithm with comprehensive learning mechanism," Applied Soft Computing, vol. 96, p. 106609, 2020.

[25] T. J. Dougherty, J. E. Kaufman, A. Goldfarb, K. R. Weishaupt, D. Boyle, and A. Mittleman, "Photoradiation therapy for the treatment of malignant tumors," Cancer research, vol. 38, no. 8, pp. 2628–2635, 1978.

[26] N. Hjortholm, E. Jaddini, K. Hałaburda, and E. Snarski, "Strategies of pain reduction during the bone marrow biopsy," Annals of hematology, vol. 92, no. 2, pp. 145–149, 2013.

[27] J. YOON and D.-W. KIM, "Uci machine learning repository uci machine learning repository, 2007," IEICE transactions on information and systems, vol. 95, no. 5, pp. 1531–1535, 2012.

[28] S. J. Malebary and A. Hashmi, "Automated breast mass classification system using deep learning and ensemble learning in digital mammogram," IEEE Access, vol. 9, pp. 55 312– 55 328, 2021.

[29] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, "Iaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography," arXiv preprint arXiv:2103.12308, 2021.

[30] S. Batchu, F. Liu, A. Amireh, J. Waller, and M. Umair, "A review of applications of machine learning in mammography and future challenges," Oncology, pp. 1–8, 2021.

[31] M. Heidari, S. Lakshmivarahan, S. Mirniaharikandehei, G. Danala, S. K. R. Maryada, H. Liu, and B. Zheng,"Applying a random projection algorithm to optimize machine learning model for breast lesion classification," IEEE Transactions on Biomedical Engineering, 2021.