A Conceptual Method for Breast Tumor Classification using SHAP Values and Adaboost

Sulthana Rinsy A. P.[1], Anish Kumar B.[2]

[1]Dept. of Computer Science and Engineering, MEA Engineering College, Kerala, India
[2]Assistant Professor, Dept. of Computer Science and Engineering, MEA Engineering College, Kerala, India

---***---

Abstract - *Breast cancer is one of the most tremendous cancers among women worldwide. It is the most common, and the leading cause of cancer related deaths among women between 20 to 59 years of age. This is an attempt aiming to assist clinicians in improving the accuracy of diagnostic decisions by classifying the breast cancer as benign and malignant tumors from ultrasound using computer-aided diagnosis (CAD) system with human-in-loop. In this method, feature acquisition is performed on the basis of Breast Imaging Reporting and Data System (BI-RADS) lexicon and experience of doctors by an user-participated feature scoring scheme. The classification is done by combining SHAP value mining and Adaboost algorithm. SHAP value based mining is done because it can form meaningful pattern clusters on the training data. The patterns frequently appearing within the tumors with an equivalent label are often considered a possible diagnostic rule. Subsequently, the diagnostic rules are utilized to construct component classifiers of the Adaboost algorithm via a completely unique rules combination strategy. Finally, the Adaboost learning is performed to discover effective combinations and integrate them into a strong classifier. The experimental results show that the proposed method yielded the simplest prediction performance, indicating an honest potential in clinical applications.*

Key Words: BI-RADS lexicon, SHAP Value mining, Adaboost

1. INTRODUCTION

Breast cancer is the one of the most common cancer among women worldwide and about 2.1 million women are undergoing treatment on breast cancer worldwide according to global estimates of cancer 2018 [1]. Global burden of cancer worldwide using GLOBOCAN 2018 estimated that the leading cause of cancer death in women is breast cancer whereas in men that is lung cancer. According to their survey the death rate due to breast cancer is getting in the peak rapidly. However the early detection and diagnosis can reduce the rapid growth in mortality and thereby can improve the survival rate. Breast cancer can occur both in men and women, but it's far more common in women. Finding carcinoma early and getting state-of-the-art cancer treatment are the foremost important strategies to stop deaths from carcinoma. Breast cancer that's found early, when it's small and has not spread, is simpler to treat successfully. Getting regular screening tests is the most reliable thanks to find carcinoma early. After detecting the cancer the key challenge faced by clinicians and doctors are in classifying the cancer into benign and malignant where machine learning techniques can play a vital role in the classification of tumor by applying proper classification algorithms.

Different tests are often wont to search for and diagnose carcinoma. Number of imaging technologies have been demonstrated to be of great help to early diagnosis for breast cancer [2] [3]. Mammography is the most commonly used screening method for breast cancer in early stage. However mammography has some limitations, not all breasts look the same on a mammogram, a woman's age or breast density can make cancers more or less difficult to see. In general, screening mammograms are less effective in younger women because they tend to have dense breast tissue and also radiation from mammography does harm to the patient's body and can significantly increase the risk of breast cancer [4].

Ultrasonography has become a popular alternative to mammography in clinical practice. Comparatively speaking, ultrasonography has the advantages of being non-radioactive, non-invasive, low cost and more convenient in practice [5] [6]. In addition, ultrasonography is not only more sensitive to dense breast tissues, it also has higher accuracy in discriminating malignant and benign tumors.

Breast Imaging Reporting and Data System (BI-RADS) is another helpful tool frequently used in clinical practice [7]. The system is developed to standardize the reporting of characteristics descriptions in mammography, ultrasound or MRI, so as to promote communication among clinicians. However, there is still a high misdiagnosis rate in the clinical application due to the subjective dependence and experience variation among clinicians. Hence, computer aided diagnosis (CAD) system has important research value in helping clinicians improve the accuracy of diagnosis of breast tumors.

2. RELATED WORK

In recent years the research in the area of analysis and classification in machine learning plays a vital role. For the tumor detection and analysis there are various methods used in the recent studies [8], for the breast tumor classification a large number of CAD approaches have been proposed in recent years Different CAD system uses different methods for classification such as some of the system uses

SVM [9] [10] and some of the system used robust phase based texture descriptor [11]and some used fuzzy cerebellar model neural network [12] which classify tumor through a learning mechanism to imitate the cerebellum of human being and some uses CNN [13]. A CAD system used SVM with 28 features in the ultrasound image and achieved a high accuracy of 94.3% in the classification of tumor as benign and malignant [14]. For automatically detecting the tumor a CAD system used fuzzy SVM and regression feature selection [15]. For efficient use of the principle component analysis and image retrieval a 3D system for breast nodules diagnosis is proposed in [16]. Some studies used a clustering method and affinity propagation for identifying breast tumors as benign and malignant.

Recently almost all systems have been designed to work in an automated way for both feature extraction and classification, However in some cases that will lead to complicating the situation by providing false positive results. Feature extraction and selection are the two important stages in the classification even in the image processing and pattern classification on [17] [18] where poor quality image leads to unstable performance. Therefore the human-in-loop CAD system performs better than that a fully automated system.

3. PROPOSED METHOD

Before going to the technical details of the proposed conceptual framework it might be useful knowing about various types and characteristics of breast cancer. There are various types of breast carcinoma, according to WHO classification the breast carcinoma falls into fifteen categories and when diagnosing with different breast tumor each shows distinct image characteristics in BUS images. By using BI-RADS lexicon which covers all these characteristics clinician can analyse the breast tumor. However BI-RADS has a limitation that it will only consider a few features of a particular breast carcinoma and it will not show a distinct pattern for other features. That is it considers a subset of few features for diagnosing the tumor which resulting BI-RADS feature will not be all consistent for different types of carcinoma in practice, in contrast this limitation leading to reduce the use of BI-RADS features for classification as an accurate one.

The framework here used for classifying the breast carcinoma is a CAD method which includes BI- RADS features along with human- in- the loop strategy for improving the classification accuracy. It has diagnostic rule mining and ensemble learning algorithm. There is a lot of feature extraction and learning algorithm in machine learning, from that we extracted the proper mining method and learning algorithm by analyzing and testing different methods along with a human-in-the-loop. We propose a CAD framework involving diagnostic rule mining and ensemble learning algorithm along with the involvement of human as an operator. It became different from a typical CAD system in the case of feature extraction, where an operator also involved for rating the feature along with BIRADS lexicon. This framework uses both the prepossessed trained data along with knowledge and experience of doctor which makes this system dominant over a typical CAD system.

Usually clinicians predict some findings on the basis of some features not considers all the features, likewise here for extracting such subset of features as diagnostic rule we propose the mining method using the shapely addictive explanation (SHAP) values. SHAP values can be applied to the rated feature data from that we can predict the high risk group malignant and low risk group benign. The SHAP values can be applied at global and local level, from that we can predict how much the contribution of each feature on the final prediction on average. It helps us to analyze and interpret our model intuitively and creates clusters for benign and malignant diagnostic rules by extracting the pattern from each feature.

Our motivation is to use SHAP method sufficiently for diagnostic rule mining and use an efficient ensemble learning algorithm. Adaboost algorithm is a relevant and an appropriate learning algorithm in ensemble learning. This Adaboost algorithm is implanted to make SHAP based classifier together into a strong robust classifier. The notable feature of this method is data mining using SHAP values along with an operator and also in an automated way and hence good interpretability of the final model. Those SHAP based rules used to build the ensemble classifier would help understand the relevant clinical manifestation related to breast tumors in ultrasound.

3.1 System design

In this section, we describe the novel SHAP based ensemble learning method. The workflow of the proposed method is given in Figure 1, we start by how the SHAP value can be applied for discovering diagnostic rules from the data matrix obtained after feature scoring. After that a conceptual novel strategy is applied to build component classifier and finally the Adaboost algorithm is briefly described.

3.2 Data preparation

Machine learning techniques train and test the data before implementing, for that first we have to collect and rate the data using BI-RADS lexicon with the help of a human-in-loop CAD system.

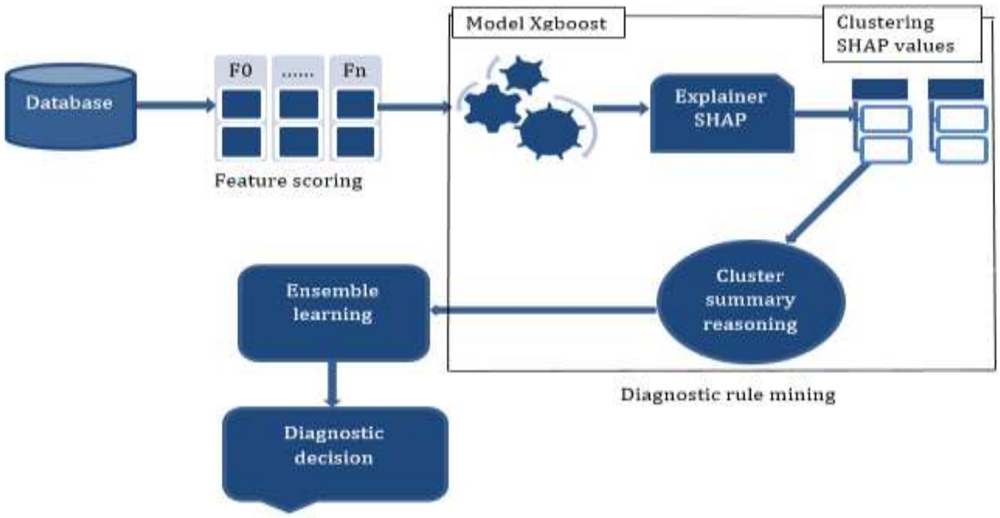

Figure 1: Workflow of proposed system

3.3 BI-RADS feature scoring scheme

BI-RADS feature scoring was widely used in mammograms and also in recent ultrasound images. Generally clinicians guess and classify a breast tumor using a BI-RADS descriptors on the basis of clinician or doctor experience. BI-RADS lexicon is a tool comprising all the necessary features such as shape, orientation, margin, echo pattern, posterior acoustic characteristics, calcification etc. Here in this research we have considered 24 features from that we have concluded that 12 features are more convenient where the other 12 features are more common in both benign and malignant cases. These 12 important features show different behavioural characteristics in patients with different tumors. These 12 features form a new feature subset for showing malignant and benign results. Here the rating features can be done using BI-RADS reference table which rates the value from 0 to 4. Higher the value more it inclined towards malignant otherwise benign. The clinician rate the feature after getting the BUS image on the reference table. For showing malignant benign it shows the values 1 and 0 respectively. Instead of image processing and feature extraction method this proposed method adopts a straightforward method, because some image processing methods show some limitations such as highly sensitive to the quality of breast ultrasound image and the type of noises. BI-RADS feature scoring scheme helps the clinicians for analysing and interpreting the diagnostic rule in an efficient way. From the perspective of a doctor such a system would be more convenient and can be applied in real time medical application than a fully automated system with complex image processing and feature extraction tasks.

3.4 Mining of diagnostic rule using SHAP values

For applying SHAP value first fit the data into a complex model like Xgboost. It is done for getting the SHAP values for each example in the data and then cluster them to find the pattern. This can lead to finding the intervention of diagnostic rules of malignant and benign cancers. SHAP values can create meaningful clusters because SHAP values for all features are on the same scale (log odds for binary Xgboost).

3.5 The classification of different feature spaces

For performing the ensemble learning a similarity classifier containing both malignant and benign rules needs to be constructed, for constructing such a weak classifier a novel combination strategy is used. After separating the benign and malignant rule anyone from benign rule and anyone from malignant rule can be matched together into a weak classifier that works on the basis of similarity principle for classifying the new tumor. In this research for finding the distance measure between different feature spaces, Feature space dependent normalized distance which is calculated using the following Equation 1 for evaluating the similarity of a test instance and diagnostic rule.

Equation 1

$$FSDND(T,R) = \frac{||V_T - V_R||_2}{||V_1 - V_0||_2}$$

Where V_T and V_R are stands for vectors of test instance and diagnostic rule respectively. V_1 and V_0 are the maximum and minimum vector, the denominator in the Equation 1 plays a normalization in the feature space of diagnostic rule. To take a decision on tumor, FSDNDs of the test instance malignant and benign rules are measured and for smaller FSDND more similar they are. With the FSDND a classifier can be created from a pair of benign and malignant rule by SHAP method using the Equation 2. Let R_m and R_b be the malignant and benign rules respectively and x be the new tumor instance then as shown below the predicted class of such a similarity classifier S(x) is the same as the attribution of the rule with smaller FSDND.

Equation 2

$$S(x) = \begin{cases} 1, & FSDND(x, R_b) > FSDND(x, R_m) \\ 0, & else \end{cases}$$

3.6 Ensemble learning

Ensemble learning is a method which can comprise multiple models such as classifier to solve a particular computation problem. Adaboost is one of the important and relevant algorithms in ensemble learning algorithms used in machine learning. It can improve the accuracy of classification by combining multiple weak classifiers. Here, similarity classifiers are treated as weak classifier. In this method, on various training examples the classifier should be trained interactively. In each iteration it tries to minimize the training error by providing an excellent fit for those examples. The weight of those correctly classified instances become lower whereas the other become higher. After the all iteration all the component classifiers are combined together to get the final hypothesis.

4. EXPERIMENTS

In this section the effectiveness of tumor classification based on SHAP values and Adaboost algorithm is investigated. First the parameter setup and experiment procedure are introduced. Then the evaluation metrics of each experiment is performed. Here 150 malignant tumor affected patient feature scored data set and 100 benign tumor affected patient feature scored data set is collected and used to train and test the system. The operation is performed in anaconda Spyder python 3.4 version. After the testing the performance is evaluated in terms of accuracy, specificity and sensitivity. The performance comparison of different classifiers are given below in the Table 1. Evaluation indices are shown in the Table 2.

Table 1: COMPARISON RESULT AMONG DIFFERENT CAD SYSTEMS

Classifier	Accuracy	sensitivity	specificity
SVM	94.4%	94.3%	94.4%
Fuzzy SVM	94.25%	91.67%	96.08%
Fuzzy cerebellar model NN	92.31%	93.55%	91.18%
Judgement by experience	85.62%	93.8%	72.97%
SHAP value + Adaboost	96.41%	96.72%	95.75%

Table 2: EVALUATION INDICES

Accuracy	(TP+TN)/(TP+TN+FP+FN)
Sensitivity	(TP)/(TP+FN)
Specificity	(TN)/(TN+FP)

5. CONCLUSION

In this paper a conceptual human-in-loop framework using SHAP values and Adaboost algorithm is implemented for classifying benign and malignant breast tumors. It is an innovative method which adopts an operator based feature scoring scheme rather than fully automated image processing and feature extraction methods. In contrast this method introduces the experience of clinicians during the feature extraction, which is easily acceptable to doctors in real application and that improves the robustness of our system.

ACKNOWLEDGEMENT

The work was supported by Computer science and engineering department of MEA Engineering College Perithalmanna, Malappuram, India.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 68, no. 6, pp. 394–424, 2018.

[2] Z. Chen, H. Strange, A. Oliver, E. R. Denton, C. Boggis, and R. Zwiggelaar, "Topological modeling and classification of mammographic micro-calcification clusters," IEEE transactions on biomedical engineering, vol. 62, no. 4, pp. 1203–1214, 2014.

[3] B. Al-Shargabi and F. Al-Shami, "An experimental study for breast cancer prediction algorithms," in Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems, 2019, pp. 1–6.

[4] T. Yin, F. H. Ali, and C. C. Reyes-Aldasoro, "A robust and artifact resistant algorithm of ultrawideband imaging system for breast cancer detection,"IEEETransactionsonBiomedicalEngineering,vol.62,no.6, pp. 1514–1525, 2015.

[5] T. Ungi, G. Gauvin, A. Lasso, C. T. Yeo, P. Pezeshki, T. Vaughan, K. Carter, J. Rudan, C. J. Engel, and G. Fichtinger, "Navigated breast tumor excision using electromagnetically tracked ultrasound and surgical instruments," IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 600–606, 2015.

[6] M. Costantini, P. Belli, R. Lombardi, G. Franceschini, A. Mul`e, and L. Bonomo, "Characterization of solid breast masses: use of the sonographic breast imaging reporting and data system lexicon," Journal of ultrasound in medicine, vol. 25, no. 5, pp. 649–659, 2006.

[7] S. Ackerman, "Acr bi-rads for breast ultrasound," Ultrasound in Medicine and Biology, vol. 41, no. 4, p. S95, 2015.

[8] S. Zhou, J. Shi, J. Zhu, Y. Cai, and R. Wang, "Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image," Biomedical Signal Processing and Control, vol. 8, no. 6, pp. 688–696, 2013.

[9] Q. Huang, F. Yang, L. Liu, and X. Li, "Automatic segmentation of breast lesions for interaction in ultrasonic computer-aided diagnosis," Information Sciences, vol. 314, pp. 293–310, 2015.

[10] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and adaboost for breast tumor classification," IEEE Transactions on Knowledge and Data Engineering, 2019

[11] L. Cai, X. Wang, Y. Wang, Y. Guo, J. Yu, and Y. Wang, "Robust phasebased texture descriptor for classification of breast ultrasound images," Biomedical engineering online, vol. 14, no. 1, p. 26, 2015.

[12] C.-M. Lin, Y.-L. Hou, T.-Y. Chen, and K.-H. Chen, "Breast nodules computer-aided diagnostic system design using fuzzy cerebellar model neural networks," IEEE Transactions on Fuzzy Systems, vol. 22, no. 3, pp. 693–699, 2013.

[13] R. Nicole, "of paper with only first word capitalized j. name stand. abbrev."

[14] Y.-L. Huang, K.-L. Wang, and D.-R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," Neural Computing & Applications, vol. 15, no. 2, pp. 164–169, 2006.

[15] X. Shi, H.-D. Cheng, L. Hu, W. Ju, and J. Tian, "Detection and classification of masses in breast ultrasound images," Digital signal processing, vol. 20, no. 3, pp. 824–836, 2010.

[16] Y.-L. Huang, S.-H. Lin, and D.-R. Chen, "Computer-aided diagnosis applied to 3-d us of solid breast nodules by using principal component analysis and image retrieval," in 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. IEEE, 2006, pp. 1802–1805.

[17] Y.Li,W.Liu,X.Li,Q.Huang,andX.Li,"Ga-sift:Anewscaleinvariant feature transform for multispectral image using geometric algebra," Information Sciences, vol. 281, pp. 559–572, 2014.

[18] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," IEEE Transactions on Cybernetics, vol. 44, no. 6, pp. 793–804, 2013.