International Journal of Advanced Technology and Innovative Research

www.ijatir.org

ISSN 2348–2370
Vol.09,Issue.01,
January-2017,
Pages:0073-0078

Computer-Aided Diagnosis of Mammographic Masses using Scalable Image Retrieval

SYEDA MERAJ FATIMA[1], AFREEN LAYEEQUA[2], SYEDA MUSHARRAF TASKEEN[3]

Abstract: Mammogram analysis is known to provide early-stage diagnosis of breast cancer in reducing its morbidity and mortality. In this paper, we propose a scalable content-based image retrieval (CBIR) framework for digital mammograms. CBIR is of great significance for breast cancer diagnosis as it can provide doctors image-guided avenues to access relevant cases. Clinical decisions based on such cases offer a reliable and consistent supplement for doctors. In our framework, we employ an unsupervised algorithm, Anchor Graph Hashing (AGH), to compress the mammogram features into compact binary codes, and then perform searching in the Hamming space. In addition, we also propose to fuse different features in AGH to improve its search accuracy. Experiments on the Digital Database for Screening Mammography (DDSM) demonstrate that our system is capable of providing content-based accesses to proven diagnosis, and aiding doctors to make reliable clinical decisions. What's more, our system is applicable to large-scale mammogram database, such that high number analogical cases would be retrieved as clinical references.

Keywords: Breast Masses, Computer-Aided Diagnosis (CAD), Mammography, Content-Based Image Retrieval (CBIR).

I. INTRODUCTION

Breast cancer is the second-most common and deadly cancer among women. Since the cause of breast cancer is undiscovered, for the time being, there are no effective ways to prevent it. Fortunately, due to the adoption of mammography screening, early-stage diagnosis of breast cancer significantly reduces its morbidity and mortality. However, breast cancer diagnosis in mammogram screening involves in error prone decision-making. In a pioneering work, it is reported that up to 30% of lesions are possible to be misinterpreted during routine screening. Computer-aided diagnosis (CAD) can play as a clinical auxiliary in detecting the abnormalities in mammograms. A recent study shows the use of CAD in the interpretation of screening mammogram can increase the detection rate of early-stage malignancies. In the past decades, many CAD techniques related to mammography have been proposed and attracted the attention of both computer scientists and radiologists. Most of these work focused on mass detection/classification, and micro-calcifications (MCs)detection/pattern classification. Regardless of improved detection rate, CAD systems commonly result in excessive false positives of malignancy, which would have adverse effect on clinical decision-making. In recent years, researchers become incrementally interested in content-based image retrieval (CBIR) for medical images. Specifically for mammogram analysis, CBIR can provide doctors with content-based manner to get accesses to clinically analogical cases.

These cases of visual similarities can further facilitate decision-making on breast cancer. Different from CAD which computes the likelihood of malignancy, in practice, CBIR aims at providing radiologists with proven diagnosis and other suitable information, by recalling mammograms of past cases visually relevant to a query. With the popularity of mammography, mammograms are available in ever increasing quantities. Consequentially, leveraging clinical information from large rather than small mammogram database becomes more pivotal. Retrieval on a large number of mammographic cases could provide comprehensive reference to radiologists. However, to the best of our knowledge, few efforts have been devoted to scalable mammogram retrieval. In this paper, we investigate scalable mammogram retrieval system on more than 5222 mammographic ROIs obtained from the Digital Database of Screening Mammography(DDSM). Encouraged by the recent success of hashing methods on scalable web-image retrieval, we employ the Anchor Graph Hashing (AGH) approach. AGH derives compact binary codes from mammograms that preserve neighborhood structure inherent in image feature space with high probability, thus resulting in less memory space and computation complexity. In addition, we propose to seamlessly fuse both holistic and local features in AGH on the distance level. We conduct experiments on the aforementioned mammogram repository, to evaluate both retrieval precision and classification accuracy.

II. METHODOLOGY

Given a mammographic ROI, the CBIR seeks out relevant cases in targeted database, based on visual similarities. The framework of our retrieval system is illustrated in Fig.1. It consists of two main phases: offline learning and online query. During the offline phase, we extract image features from mammogram database and compress them into binary codes, by using Anchor Graph and spectral embedding. Such binary codes preserve the similarities in original image feature space with high probability. In the online phase, the image features of a

query ROI are also converted into binary codes, with generalized hashing functions. Then we perform efficient searching in Hamming space to retrieve the most similar ROIs with smallest distance. The proven diagnosis of these retrieved ROIs can facilitate clinical decision-making on the query mammogram.

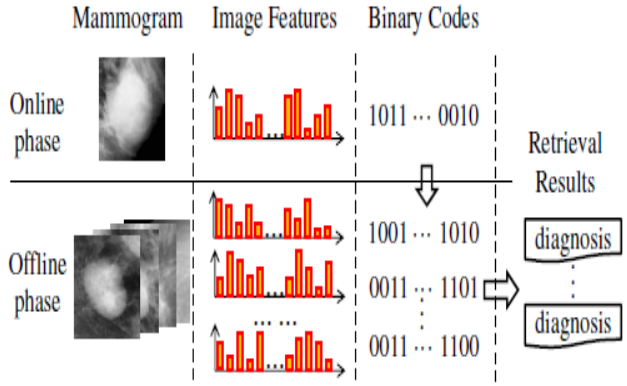

Fig.1. Framework of mammogram retrieval using hashing.

III. PROPOSED APPROACH

In this section, we first introduce the mammographic ROI retrieval framework based on vocabulary tree, then present our refinement on the weights of tree nodes, and finally describe how to make a diagnostic decision using the retrieval set. The overview of our approach is shown in Fig.2.

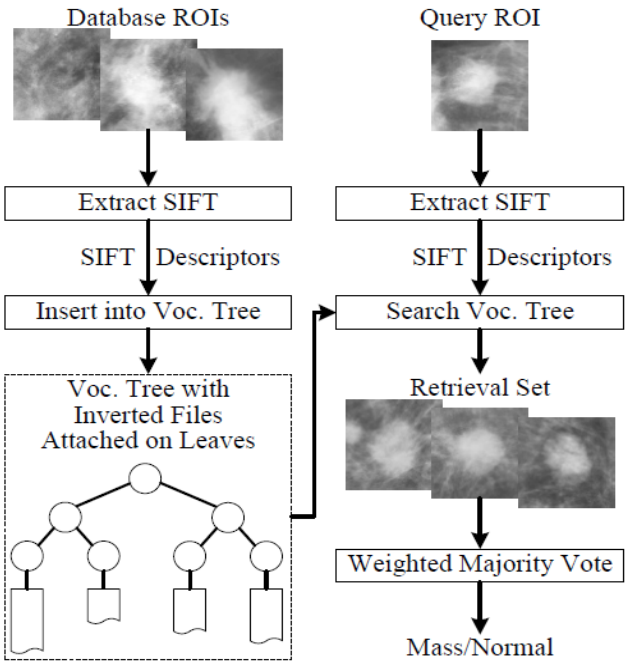

Fig.2. Overview of the proposed approach.

Mammogram Retrieval with a Vocabulary Tree: Our approach builds upon a popular CBIR framework that indexes local image features using vocabulary tree and inverted files. The local feature we choose here is scale-invariant feature transform (SIFT). It has been successfully applied to medical image retrieval and analysis, owing to its excellent robustness and discrimin ability. In this framework, a large set of SIFT descriptors extracted from a separate database are used to train a vocabulary tree through hierarchical k-means clustering. Specifically, k-means algorithm is first run on the entire training data, defining k clusters and their centers. This process is then recursively applied to all the clusters, splitting each cluster into ksub-clusters. After L recursions, a vocabulary tree of depth L and branch factor k is built. Then, all SIFT descriptors extracted from database mammographic ROIs are quantized and indexed using this vocabulary tree and inverted files. Each SIFT descriptor is propagated down the tree by choosing the closest cluster center at each level. The ID of associated database ROI is then added to the inverted file attached to the leaf node. Given a query mammographic ROI q, SIFT features are extracted and quantized. The similarity score between q and a databases ROI d is calculated based on how similar their paths are. The tree nodes are weighted using term frequency-inverse document frequency (TF-IDF) scheme or its variations, where TF means the weight of a node is proportional to its frequency in a query ROI, and IDF indicates that the weight is offset by its frequency in all database ROIs.

IV. EXPERIMENTS

This section validates the proposed mammogram retrieval and diagnosis approach. First, experimental settings, including dataset, compared methods, and evaluation environment, are described. Then, experimental results are presented and analyzed. Finally, impact of parameters is discussed.

A. Experimental Settings

Our experimental dataset is constructed from the digital database for screening mammography (DDSM).DDSM is currently the largest public mammogram database. It is comprised of 2 604 cases, and every case consists of four views, with two views, CC and MLO, for each breast. The masses have diverse shapes, sizes, margins, breast densities as well as patients' races and ages, and are associated with annotations labeled by experienced radiologists. To simulate practical scenario, a series of ROIs depicting masses and suspicious normal tissues are extracted following the conventions. This process is demonstrated in Fig. 3. First of all, mammograms are mapped from gray level to optical density according to DDSM's instructions to eliminate visual difference caused by different scanners. Second, normalized mammograms are processed for better visual quality using inversion, breast segmentation, and contrast enhancement. Third, 2 340 ROIs centered on masses are extracted. Fourth, 9 213 false positives asserted by a CAD system from healthy cases are used as normal regions. This CAD system is based on a cascade of boosted Haar classifiers and trained on a separate mammogram dataset. Note that compared with experiments, which randomly select normal regions, our experiment setting is more consistent with practice and more challenging. Finally, of the aforementioned ROIs, 500 mass ROIs, and 500 normal ROIs are randomly selected as queries.

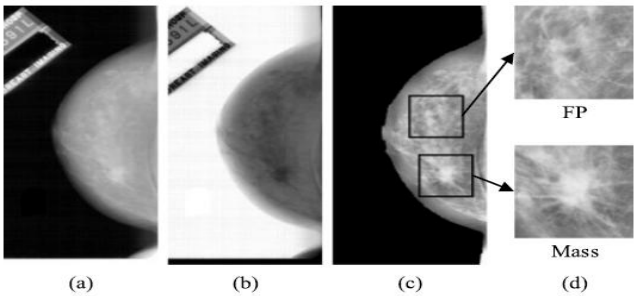

Fig.3. Construction of our dataset. (a) Original mammogram in gray level format. (b) Normalized mammogram in optical density format. (c) Visually enhanced mammogram. (d) Radiologist-annotated mass and CAD-generated false positive.

TABLE I: Retrieval Precision At Different K

K	Method	Mass	Normal	Total
1	NMI	73.5%	75.2%	74.4%
	BoW	76.8%	78.9%	77.9%
	VocTree	82.5%	85.8%	84.2%
	VocTree+AdaptWeight	86.9%	**89.3%**	88.1%
5	NMI	72.6%	74.4%	73.5%
	BoW	76.3%	79.6%	78.0%
	VocTree	82.4%	85.2%	83.8%
	VocTree+AdaptWeight	**87.7%**	89.1%	**88.4%**
20	NMI	68.9%	71.5%	70.2%
	BoW	75.6%	75.3%	75.5%
	VocTree	80.1%	82.2%	81.1%
	VocTree+AdaptWeight	84.5%	86.3%	85.4%

The remaining 1 840 mass ROIs and 8 713 normal ROIs, 10 553 ROIs in total, form a large database. The query and database ROIs are selected from different cases in order to avoid positive bias. For a more reliable performance evaluation, the random selection of query ROIs is repeated for five times. After each selection, all the methods are tested, and the average performance from five runs is reported. We also implement two other medical image retrieval systems for comparison. The first one, presented and, it performs a template matching between query ROI and each database ROI based on normalized mutual information (NMI).Experiments are show that NMI obtains good retrieval precision and best diagnosis accuracy among eight information theoretic similarity measures. The second one, similar to, represents each ROI with a SIFT BoW and measures the $\chi 2$ distance between query ROI and each database ROI. While SIFT feature is derived from dense grids or super pixels, our implementation employs the traditional SIFT derived from DoG extreme. This is to better test the vocabulary tree framework under the condition that the same feature is utilized. For this method, a vocabulary containing k_{BoW} =1 000 visual words is constructed using k-means clustering. Our method is tested twice, with the adaptive weighting scheme deactivated for the first time, and activated for the second time. Both of them employ a vocabulary tree of branch factor k =10 and depth L = 6. These four approaches are denoted asNMI, BoW, VocTree, and VocTree+Adapt Weight in the following analysis. All the methods are implemented in C++ and evaluated on a high-performance laptop with Intel Core i7 processor (6Mcache, 2.40 GHz), 16GB memory, and Windows 7 operating system.

B. Results and Analysis

First of all, retrieval precision is evaluated, which is defined as the percentage of retrieved database ROIs that are relevant to query ROI. Overall the precision changes slightly as the size of retrieval set K increases from 1 to 20. The precisions at top K = 1, 5, and 20 retrievals are summarized in Table I. Two retrieval sets returned by VocTree+Adapt Weight are provided

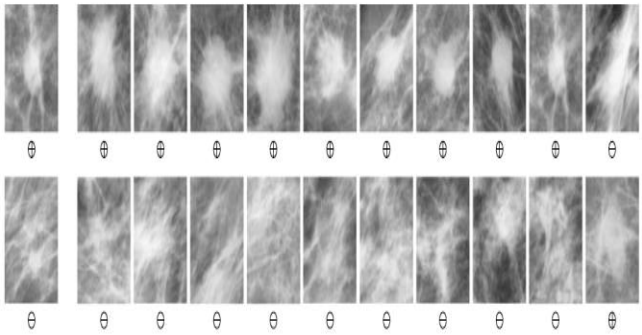

Fig.4. Two query ROIs (left) and their top K=10 retrieved database ROIs calculated by VocTree+Adapt Weight (right). For each ROI, its class is shown below. Both query ROIs are correctly classified according to a weighted majority vote of their retrieval sets.

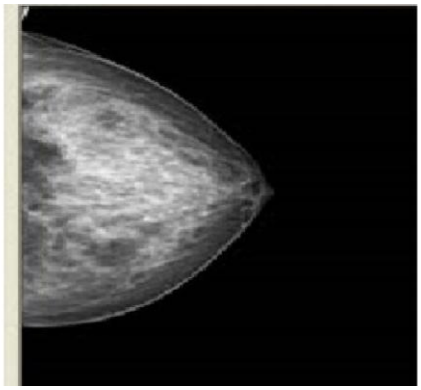

Fig.5. Input Image as Query.

Fig.6. ROI mask Selection of Input Image.

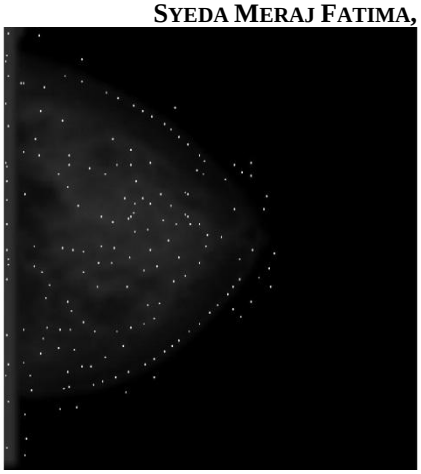

Fig.7. Comparison of Interest Points from ROI of Input Image.

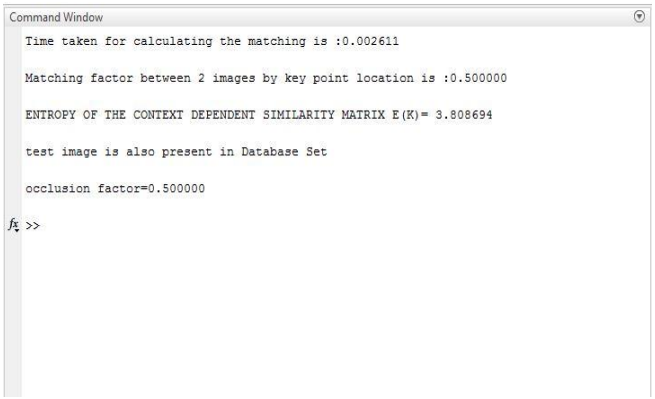

Fig.8. Results of image after comparison.

TABLE II: Classification Accuracy At Different K

K	Method	Mass	Normal	Total
1	NMI	73.5%	75.2%	74.4%
	BoW	76.8%	78.9%	77.9%
	VocTree	82.5%	85.8%	84.2%
	VocTree+AdaptWeight	86.9%	89.3%	88.1%
5	NMI	73.3%	76.1%	74.7%
	BoW	78.7%	80.3%	79.5%
	VocTree	84.9%	86.7%	85.8%
	VocTree+AdaptWeight	**90.1%**	**91.5%**	**90.8%**
20	NMI	71.2%	74.6%	72.9%
	BoW	77.0%	76.2%	76.6%
	VocTree	81.9%	84.1%	83.0%
	VocTree+AdaptWeight	86.1%	87.7%	86.9%

The results show that our methods, especially VocTree+AdaptWeight, surpass the compared approaches. Detailed results show that much incorrect retrieval is due to the visual similarity between malignant masses and normal ROIs with bright cores and speculated edges as shown in Fig.5. It is also notable that retrieval precisions for normal regions are generally higher than those for masses. A possible reason is that the database has more normal ROIs than masses, therefore it is easier for a normal query ROI to find similar database ROIs. Second, classification accuracy is measured, which refers tothe percentage of query ROIs that are correctly classified. The classification accuracies at top K = 1, 5, and 20 retrievals are reported in Table II. Once again, our methods consistently outperform the other two approaches. In addition, the classification accuracy is even better than the retrieval precision, since irrelevant retrievals would not cause a misclassification as long as they remain a minority of the retrieval set as shown in Fig.6. Especially, Voc-Tree+Adapt Weight achieves a classification accuracy as high as90.8% at K = 5, which is pretty satisfactory.

Finally, efficiency and scalability are investigated. Efficiency is assessed using the average processing time needed to retrieve and classify a query ROI. Since the classification step is merely a vote on the retrieval set, processing time is actually equal to retrieval time. Besides, as K increases from 1to 20, a retrieval procedure only needs to change the size of the max/min heap, which records the similarity/distance scores of the retrieved database ROIs as shown in Fig.7. Therefore, the processing time barely changes as K varies, and we only report the time at in Fig. 4 for visual evaluation.

TABLE III: Performance At Different Database Sizes

Size	Method	Retri. Prec.	Class. Accu.	Time (sec)
2,600	NMI	66.4%	68.8%	3.19
	BoW	70.6%	72.7%	0.71
	VocTree	75.7%	80.1%	**0.29**
	VocTree+AdaptWeight	**81.6%**	**84.4%**	0.34
5,200	NMI	67.9%	69.6%	6.23
	BoW	73.3%	74.5%	1.14
	VocTree	79.6%	81.2%	**0.32**
	VocTree+AdaptWeight	**83.8%**	**86.1%**	0.39
10,553	NMI	70.2%	72.9%	12.31
	BoW	75.5%	76.6%	1.95
	VocTree	81.1%	83.0%	**0.39**
	VocTree+AdaptWeight	**85.4%**	**86.9%**	0.48

K = 20. Scalability of a method is measured by testing how its performance changes as the database expands. To this end, two smaller databases are constructed by randomly sampling a half and a quarter of database ROIs, and all the methods are evaluated again on these two databases. Their retrieval precisions, classification accuracies, and average processing time at top K = 20 retrievals are summarized in Table III. According to this table, we can reach several conclusions. First of all, our methods are consistently superior to the compared approaches with respect to all three evaluation metrics, especially efficiency. VocTree+ AdaptWeight obtains even better retrieval precision and classification accuracy than those of VocTree at the cost of a little more processing time. Second, the vocabulary tree framework demonstrates excellent scalability. In this framework, as we explained in Section III-A, the similarity calculation only needs to consider those database features that fallin to neighbor leaf nodes as the query features do, which account for a small portion of all the database features. What is more, as the database expands, we can use a larger vocabulary tree (with bigger branch factor k and/or depth L) to reduce the portion of database features that need to be considered as shown in Fig.8. Therefore, the time cost

of similarity computation is not only small but also sub linear in database size.

TABLE IV: Performance of Voctree+Adaptweight At Different L

L	Retri. Prec.	Class. Accu.	Time (sec)
3	73.8%	75.3%	1.69
4	78.7%	80.8%	0.91
5	83.1%	85.2%	0.62
6	85.4%	86.9%	0.48
7	**86.2%**	**87.1%**	**0.47**

On the contrary, NMI and BoW calculate a similarity/ distance score between queries ROI and each database ROI, which takes a linear time regarding database size. (The time for query feature extraction and quantization in BoW remains unchanged for different database sizes.)Last but not least, as the database grows, all the methods obtain better retrieval precisions and classification accuracies. This result agrees with the experiments and confirms our assumption that it is more likely to find relevant cases and make a correct diagnosis using a large database. All the experiments lead to several conclusions. 1) NMI obtains the worst results among all the tested methods. The reason is that masses have diverse shapes, sizes, and cluttered background; therefore, it is not suitable to match two entire ROIs without extracting certain features from invariant keypoints.2) Our method is superior to BoW. Although both employing SIFT feature, they implement different quantization, indexing, and similarity calculation schemes. Specifically, first, BoW uses a single-level k-means clustering for feature quantization, whose computational cost is linear in vocabulary size. As a result, the vocabulary typically has a small size (100, 1 000 and our implementation). Instead, our method utilizes hierarchical k-means, whose computational cost is logarithmic in vocabulary size. Thus, it can afford a much larger and more discriminative vocabulary (106 in aforementioned experiments).Second, BoW performs exhaustive search without the aid of any index. On the contrary, in our model, quantized database features are indexed using inverted files so that only a small portion of them is considered during similarity computation, and the portion of involved features can be further decreased by increasing vocabulary size (L or k). Actually, experiments on general CBIR datasets demonstrate that our method could retrieve in real time from millions of images. Finally, all the visual words in BoW are treated equally, whereas their weights are elaborately adjusted according to the whole database (IDF) and each query (TF and adaptive weighting). 3) Adaptive weighting, which down weights the excessive features extracted from normal regions, could improve retrieval precision and classification accuracy without considerably reducing efficiency.

C. Discussion of Parameters

To test the impact of parameters on our method's performance, we have trained several vocabulary trees of branch factor k = 10 and depth L = 3, . . . ,7. For each vocabulary tree, the performance of VocTree+AdaptWeight at top K = 20 retrievals is measured using five randomly selected query sets, and the average performance from five runs is summarized in Table IV. From this table, we can see that the retrieval precision, classification accuracy, and computational efficiency of VocTree+AdaptWeight improve substantially as L goes from 3to 5, then improve slightly as L increases to 6 and 7. Two conclusions can be drawn from this observation. On the one hand, larger vocabulary trees tend to achieve better performance. As explained, a larger vocabulary tree has smaller and more discriminative leaf nodes, which result in better retrieval precision as well as classification accuracy. Besides, as the total number of leaf nodes increases, the portion of database features that need to be considered during similarity calculation is reduced, therefore, the efficiency is also improved. On the other hand, the vocabulary tree framework could benefit from more training features. The performance gain from L = 6 to7 is very small. It is probably due to the limited number of training features, since nearly a third of leaf nodes are empty during the training process when L becomes 7. These two conclusions are consistent with the observations in general image retrieval.

V. CONCLUSION

In this paper, we propose to use scalable CBIR for the automatic diagnosis of mammographic masses. To retrieve efficiently from a large database, which leads to better retrieval precision and diagnostic accuracy, we employ the vocabulary tree framework to hierarchically quantize and index SIFT descriptors. Furthermore, contextual information in the vocabulary tree is incorporated into TF-IDF weighting scheme to improve the discriminative power of tree nodes. Query mammographic ROIs are classified using a weighted majority vote of its best matched database ROIs. Experiments are conducted on a database including 2174 mass ROIs and 2831 CAD generated false positive ROIs, which is the largest dataset to the best of our knowledge. Excellent results demonstrate the retrieval precision and diagnostic accuracy of our method. Future endeavors will be devoted to refine retrieval set using spatial contextual information of SIFT features. Diagnostic information can also be taken into consideration using feature selection and fusion methods. This work is partially supported by grant NSF-MRI-1229628.

VI. REFERENCES

[1] Menglin Jiang, Shaoting Zhang, Member, IEEE, Hongsheng Li, Member, IEEE,and Dimitris N. Metaxas, Senior Member, IEEE, "Computer-Aided Diagnosis of MammographicMasses Using Scalable Image Retrieval", IEEE Transactions on Biomedical Engineering, Vol. 62, No. 2, February 2015.

[2] American Cancer Society, Breast Cancer Facts & Figures 2013-2014.Atlanta, GA, USA: American Cancer Society, 2013.

[3] N. Howlader, A. M. Noone, M. Krapcho, J. Garshell, N. Neyman,S. F. Altekruse, C. L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, H. Cho,A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin, SEERCancer Statistics Review,

1975-2010. National Cancer Institute, Bethesda,MD, USA, 2013.
[4] H.-D. Cheng, X.-J. Shi, R. Min, L.-M. Hu, X.-P.Cai, and H.-N. Du, "Approachesfor automated detection and classification of masses in mammograms,"Pattern Recog., vol. 39, no. 4, pp. 646–668, 2006.
[5] A. Oliver, J. Freixenet, J. Mart´ı, E. P´erez, J. Pont, E. R. E. Denton, andR. Zwiggelaar, "A review of automatic mass detection and segmentationin mammographic images," Med. Image Anal., vol. 14, no. 2, pp. 87–110,2010.
[6] P. Skaane, K. Engedal, and A. Skjennald, "Interobserver variation in theinterpretation of breast imaging," ActaRadiol., vol. 38, no. 4, pp. 497–502,1997.
[7] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review ofcomputer-aided diagnosis of breast cancer: Toward the detection of subtlesigns," J. Franklin Inst., vol. 344, pp. 312–348, 2007.
[8] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles,"Mammographic characteristics of 115 missed cancers later detected withscreening mammography and the potential utility of computer-aided detection,"Radiology, vol. 219, no. 1, pp. 192–202, 2001.
[9] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detectionof radiographic abnormalities in mammograms by means of opticalscanning and computer analysis," Radiology, vol. 89, no. 2, pp. 211–215,1967.
[10] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computeraideddetection and diagnosis of breast cancer with mammography:Recentadvances," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 2, pp. 236–251,Mar. 2009.
[11] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, andK.-H. Ng, "Computer-aided breast cancer detection using mammograms:A review," IEEE Rev. Biomed. Eng., vol. 6, pp. 77–98, Mar. 2013.
[12] S.-C. Tai, Z.-S.Chen, and W.-T. Tsai, "An automatic mass detectionsystem in mammograms based on complex texture features," IEEE J.Biomed. Health Informat., vol. 18, no. 2, pp. 618–627, Mar. 2014.

Author's Profile:

Syeda Meraj Fatima, Received M.Tech in Digital Systems and Computer electronics from Shadan women's college of Engineering and technology, Presently she is working as an Assistant Professor in Mahaveer Institute of Science and Technology.

Afreen Layeequa, Received M Tech from Shadan women's college of engineering and Technology, Hyderabad. Currently working as an Assistant Professor in Mahaveer Institute of Science and Technology.

Syeda Musharraf Taskeen, Received M.Tech in VLSI and Embedded System from Medak College of engineering and technology, Medak. Presently working asAssistant Professor in Mahaveer Institute of Science and Technology.