

K8. Hand-Held Computer Aided Diagnosis System with Application in Mammography

Mohamed S. ElBially, Mohamed S. Ahmed, Mohamed H. AbdelGawwad, Fatma A. Ali, Fady S. Botros,
Yasser M. Kadah

Biomedical Engineering Department, Cairo University, Giza, Egypt

ABSTRACT

In the recent decades, Egypt has witnessed an increase in the number of mammogram scans, raising the need for cheap and portable systems for computer aided diagnosis (CAD) to reduce examination cost and time which can be achieved through embedded systems. In this work, we introduce an early prototype of a low-cost high-performance hand-held CAD system with an application in diagnosing abnormalities in mammography which can be extended in the future for other modalities such as ultrasound imaging. For this application, we used the BeagleBoard-xM which is known for its high performance and equipped with many peripherals. Specifications of the system are described and the results are presented including the diagnostic performance of the CAD system and the comparison between the computational performance of a reference PC and the new embedded system. The new system has the potential for allowing such high technology to be available at low cost thus providing cost-effective healthcare.

Keywords: *Computer Aided Diagnosis, Embedded Systems, Beagleboard.*

I. INTRODUCTION

Breast cancer is a leading cause of women death worldwide. In Egypt, the number of women suffering from breast cancer is increasing rapidly. According to the official statistics of the National Cancer Institute (Cairo University) in 2011, breast cancer accounts for 35.1% of all the cases of cancer in Egypt and it is the most prevalent cancer among the Egyptian women. The median age of women having a diagnosis for breast cancer in Egypt is ten years less than the corresponding age in the North America and Europe [1]. Presently, mammography is the main imaging technique used for diagnosing breast cancer. Its importance and wide approval comes from its ability to distinguish between soft tissues and all other abnormal structures like microcalcifications and masses.

Computer-aided diagnosis (CADx) is a recent new technology that aims to assist doctors and radiologists in the interpretation of medical images. We are introducing a CADx system which mainly aims to diagnose rather than to detect (CADe) a given region in a mammogram image. A CADx system can reduce the heavy load on doctors to diagnose an enormous number of screens. Therefore, the value and importance of CADx systems as a second opinion to assist radiologists have become more appreciable in recent decade.

To the best of our knowledge, previous work was concerned with developing algorithms and techniques for CADx as only PC-based implementations. This can be very costly and also may not be available for rural regions with low budget clinics that are unable to provide. In this work, we introduce an early prototype of a low-cost high-performance hand-held CAD system with an application in diagnosing cancer in mammography.

In the following sections we cover an overview about our system, starting with features extraction, feature selection method and classifiers used. After that we introduce our system hardware then finally our results which include timing results on the BeagleBoard-xM and statistical results for our system showing our system performance such as accuracy, sensitivity etc.

II. MATERIALS AND METHODS

A. System Overview

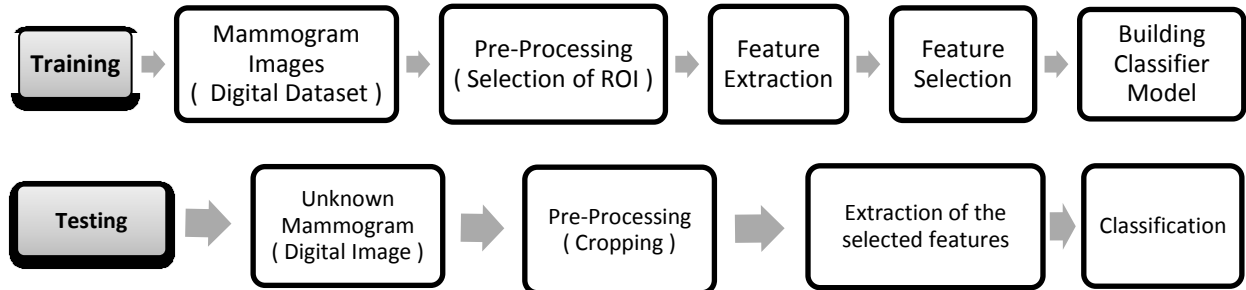


Fig.1. Block diagram of the CAD system

The main stages in the system design are data collection, region of interest (ROI) selection, feature extraction, feature selection and finally classification. Different training and testing sets are used from the well-known mini-MIAS mammography database which offers mammograms digitized at a resolution of 50 micron pixel edge that has been reduced to 200 micron pixel edge and clipped/padded so that every image is 1024 pixels \times 1024 pixels, which is sufficient to view and detect any abnormalities for our CAD system. A block diagram of the system is presented in Fig. 1.

B. Image Pre-Processing

In our system we used manual segmentation to extract the ROI, as our proposed work is mainly concerned with diagnosis rather than automatic segmentation and cancer detection. Furthermore we applied histogram equalization on our region of interest to enhance its contrast.

C. Feature Extraction

In our system we mainly depend on features extracted from mammograms using GLCM (gray level co-occurrence matrix), some histogram features and wavelet based features.

1) GLCM Features:

GLCM calculates the probability of a pixel with the gray level i occurring in a specific spatial relationship to a pixel with the value j .

We calculated autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares (variance), sum average, sum variance, sum entropy, difference entropy, information measure of correlation, information measure of correlation², inverse difference moment normalized and inverse difference normalized for GLCM at different distance (3,5,10) and different angle (0,45,90,135) [2] [3].

2) Histogram Features:

We calculated the mean (measure of average intensity), standard deviation (measure of average contrast), smoothness (measure of the relative smoothness of the intensity in a region), third moment (measure the skewness of a histogram), entropy, and kurtosis [4] [5].

3) Wavelet-based Features:

In our system we decompose our signal using 2-D wavelet transform due to its capability for extracting microcalcifications from low-frequency backgrounds and high-frequency noise.

We decompose our mammogram image up to level two using Daubechies orthogonal wavelet of length four because it consumes less computation time compared to the other wavelets in the Daubechies orthogonal family, then for each level we calculate the energy which gives seven different energies, and one value for entropy. [6] [7].

D. Feature selection

Feature selection is an important step in breast cancer diagnosis and classification. Instead of using all available variables (features or attributes) in the data, one selectively chooses a subset of features to be used. There are number of advantages of feature selection: (1) dimension reduction to reduce the computational cost, (2) reduction of noise to improve the classification accuracy, (3) more interpretable features or characteristics that can help identify and monitor the target diseases or function types. In this paper we use Weka software (version 2.7.2) [8] for data mining and feature selection.

In this paper we used SVMAttributeEval[9] for feature selection which evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class separately using a *one-vs-all* method and then dealing from the top of each pile to give a final ranking.

E. Classification

The final stage is classification in which the region of interest (ROI) is classified between normal/mass or normal/microcalcifications, based on a previously classified training set. This process is divided into two phases namely; Training Phase and Testing Phase. In the training phase, features are extracted from images that are already confirmed to belong to a specific class and the classifier is trained. During the testing phase, parameters of the query ROI are calculated then fed to the classifier after training in order to map the unknown ROI to the relevant class.

There are different types of classifiers; in this study we have applied the K-Nearest Neighbour (KNN) classifier and the Support Vector Machine (SVM) classifier.

1) K-Nearest Neighbour (KNN) Classifier

In this algorithm, the query ROI is classified according to the majority vote of its neighbours, with the ROI being assigned to the most common class among its k nearest neighbours; where k is a small positive integer. The neighbours are taken from the training set for which the correct classification is known. We used the Euclidean distance as the distance metric and gave each neighbour a weight of $1/d$ where d is the distance to the neighbour. Many values of k have been applied in order to find its influence on classification and reach the optimal results [10].

2) Support Vector Machine (SVM) Classifier

SVM classifier considered a supervised machine learning algorithm first described by VapnikChervonenkis [11]. It does the same functionality as neural networks but with high quality of generalization as it tries to minimize the error of generalization rather than the training error besides ease of training [12]. It can easily model high complex real world problems like medical imaging analysis and it is characterized by its high performance with data sets that have many attributes even with few numbers of training cases unlike other classification algorithms. Another property of SVM is its ability to do kernel transformation trick which maps input data space to a higher space for more accurate and easy classification when the data sets are not easily separable [13]. This kernel can be linear like a linear equation that can be used when there is a large number of features and small training set and as a non-linear kernel like Radial Basis Function kernel (RBF) when number of features is small but there is a reasonable number of training sets. In this paper we used the RBF kernel as we have a small number of features relative to the training set, in this kernel there are two main parameters that can describe the classifier model firstly regularization parameter (C) which is a positive number that describes the trade-off between model complexity and classification allowed error, secondly standard deviation (γ) of the Gaussian RBF kernel.

F. Implementation

1) PC based implementation

Our PC implementation is based on the following programs:

- Image Processing: Cropping region of interest (ROI) manually using ImageJ(open-source) version1.47d.
- Feature Extraction: MATLAB (Mathworks, version R2009b).
- Feature Selection: WEKA (Waikato Environment for Knowledge Analysis) Machine LearningProgram.
- Training and Testing: Orange data mining and machine learning software suite (version 2.6a2).

2) *Embedded implementation*

In our embedded based implementation we used BeagleBoard-xM (Texas Instrument, revision B)(BB-xm) 8cm x 8cm which is equipped with DM3730 platform and other interface peripherals like 4 USB ports, Ethernet,512-MB LPDDR RAM and microSD slot [14] [see Fig. 2]. DM3730 is a heterogeneous platform as it consists of General Purpose Processor (GPP) and Digital Signal Processor (DSP) as co-processor.

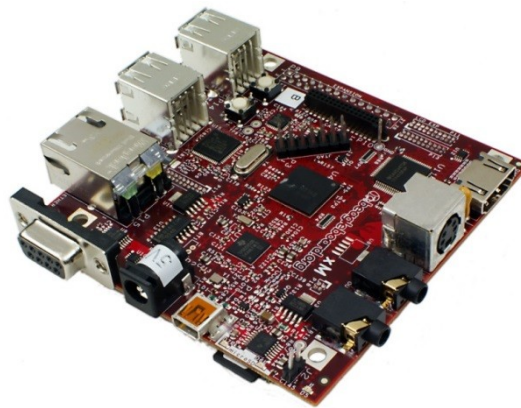


Fig.2. BeagleBoard-xM revision B with its peripherals [15].

2.1) *GPP side*

It's a 1GHZ ARM cortex-A8 RISC processor architecture operated with minimal image Angstrom distribution operating system which we build using Yocto Project [16] and a baluster kernel , This Linux based distribution characterized by best utilization of the embedded resources with pre-packaged installed software , all of these are loaded on a mini SDcard which is divided into two partitions , boot partition that contains booting related files like UImage, u-boot.img, uRnv.txt, MLO, boot.src and boot.cmd and rootfs partition that contains root directory files for user [17].

2.2) *DSP side*

A 800 MHZ Very Long Instruction Word (VLIW) TMS320C64+ DSP core that can execute up to eight 32 bit instructions per cycle, C64+ CPU has 64 general purpose 32 bit registers and eight functional units; these units are: six Arithmetic and Logic Units (ALU) and two multipliers [18] operated with a Real Time Operating System (RTOS) called SysBIOS 6.x which is open source and its kernel can be modified and rebuild.

2.3) *Interprocessor communication protocol (IPC)*

In order to speed up some functions performance we use DSP core due to its specialized architecture dedicated for extensive complicated mathematical operations but to do so we need some sort of communication between GPP side (wherethe main program flows) and DSP side, this communication process is performed using Syslink2.20 as it uses grid topology so that any processor can talk to any other processor directly unlike DSPlink which use star topology that allow all processors to talk to a central unique processor, the communication process between GPP and DSP is shown in Fig.3.Syslink contains more than one protocol for data transfer between processors like MessageQ, RingIO and FrameQ, in our project we used MessageQ protocol which is illustrated in Fig. 4. As it is more robust and less time consuming [19].

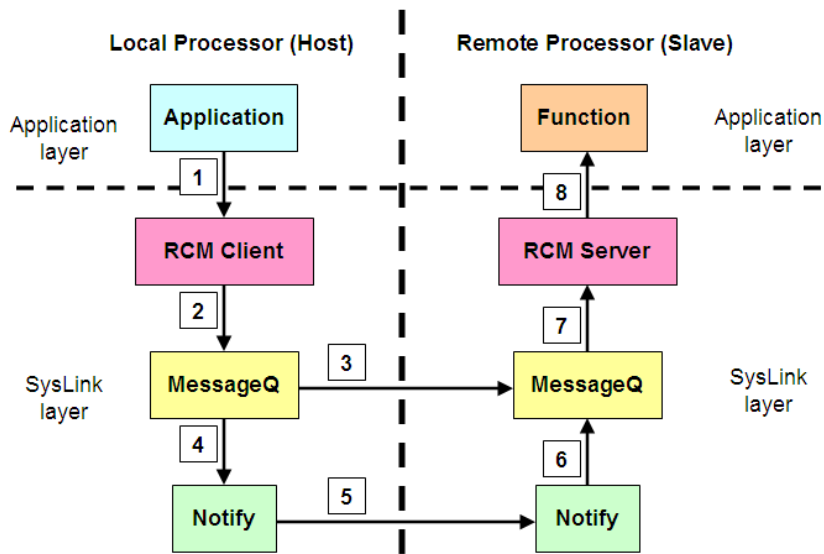


Fig. 3. Communication process between DSP (Local Processor) and DSP (Remote Processor) [20].

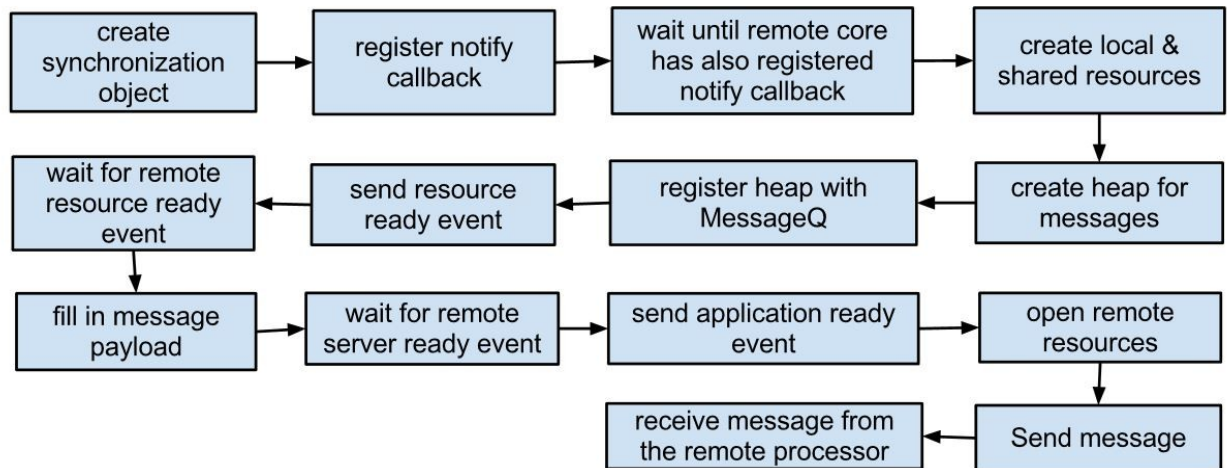


Fig.4. MessageQ protocol procedure.

III. RESULTS AND DISCUSSION

Table 1: shows Time elapsed by 3 different type processor for CAD Algorithm Stages in μ Second.

Platform / process	Core I3- Processor	ARM-Processor	DSP-Processor
Image to Matrix	213 \pm 29	656 \pm 30	NA
GLCM	313 \pm 4	3464 \pm 70	5000
Feature Extraction	389 \pm 6	2960 \pm 114	3175
Classification	75 \pm 4	1175 \pm 44	3000

Table 2, 3: shows results of binary classification between normal-mass cases forSVM,KNN classifiers

Number of Features selected	SVM	Mass for [74 case]						
		CR	Sens	spec	TP	FP	TN	FN
9	ROI = 16	82%	88%	79%	28	9	33	4
	C = 5, γ = 1							
4	ROI = 32	86%	86%	86%	32	5	32	5
	C = 5, γ = 1							
10	ROI = 64	94.6%	97.3%	91.9%	36	1	34	3
	C = 0.5, γ = 2							

Number of Features selected	KNN	Mass for [74 cases]						
		CR	Sens	spec	TP	FP	TN	FN
10	ROI = 16	66%	68.8%	64%	22	15	27	10
	K = 3							
4	ROI = 32	80%	80.6%	79%	29	8	30	7
	K = 3							
10	ROI = 64	93%	90.0%	97%	36	1	33	4
	K = 3							

Table 4, 5: shows results of binary classification between normal-calcification cases for SVM,KNN classifiers

Number of Features selected	SVM	Micro-Calcification for [73 cases]						
		CR	Sens	spec	TP	FP	TN	FN
10	ROI = 16	95%	94%	95%	34	2	35	2
	C = 1, γ = 0.9							
4	ROI = 32	81%	82%	79%	28	8	31	6
	C = 2, γ = 0.5							
10	ROI = 64	77%	78%	76%	28	9	29	8
	C = 5, γ = 4							

Number of Features selected	KNN	Micro-Calcification for [73 cases]						
		CR	Sens	spec	TP	FP	TN	FN
9	ROI = 16	90%	91%	89%	32	4	34	3
	K = 3							
4	ROI = 32	73%	74%	72%	25	11	28	9
	K = 3							
8	ROI = 64	74%	70%	80%	31	6	24	13
	K = 3							

A. CAD Performance

The Size of the region of interest is a very important factor in our results due to the weak resolution of the used dataset (MIAS), so it is obvious from results shown in Tables 2, 3, 4 and 5 that smaller ROI are much better for micro calcifications cases (see Tables 4,5), but for mass diagnosis the results go bad with the decrease of the ROI size due to the lack of shape features when we zoom into the image, so mass diagnosis results much better for large ROI (see Tables 2,3).

SVM model parameters C , γ have been calculated from 10-fold cross-validation using grid search over an exponential growing set of C , γ to find the best combination that maximize the classification accuracy and minimize the generalization error [21]. Tables 2, 4 show the best combinations, and we can see that C values were small because as this value increases we go closer to over fitting and increases system complexity. Also the KNN parameter K was chosen based on the same statistical method 10-fold cross-validation and Tables 3, 5 shows that the best results came when K is not large this indicates that the classes overlap after this value which reduce the classification accuracy.

Results in Tables 2, 3, 4 and 5 show that the SVM classifier is more accurate and give better results for both microcalcifications and masses this is because as we said before SVM always tries to minimize the generalization error not just the training error like KNN, and also because of SVM kernel trick that causes more separation between the classes.

As mentioned before, the MIAS dataset resolution is not very high compared to other digital mammography databases such as the DDSM database. The algorithm will not be very different and the results are expected to carry on to any other database.

B. Timing Performance of System

We can see that timing results of ARM processor is longer than PC, which was expected because the PC processor has a faster clock speed (2.27GHz) than the ARM processor (1GHz) in addition to the availability of significantly more memory on the PC platform than the embedded platform that reduces the cache memory misses. Also DSP processor shows a long timing performance compared to ARM and CPU processors, this is because DSP needs a special optimization incoding so that we can make use of its power in accelerating the addition multiplication calculations. Here we show up some optimizations paradigms (on different levels) that we can follow for further optimization [22]:

- *C Code level*: Using intrinsic, pragmas, word-wide optimization, loop unrolling and compiler feedback.
- *Linear ASM*: For functions or sections that is time consuming like GLCM calculation section.
- *ASM*: For critical sections, where we can use parallel instructions to be executed in the same cycle like the feature extraction section.

Timing profiling shown in Table 1 will help in deciding which functions to be executed on ARM side and which are better to be done on DSP side.

IV. Conclusions

In this work, we introduce an early prototype of a low-cost high-performance hand-held CADx system with an application in detecting cancer in mammography. The results of the developed system performance and implementation to classify images from a standard mammography database were presented. The embedded CAD system was shown to offer processing power that is practical for clinical use and can be extended to other modalities like ultrasound imaging. The new system shows encouraging potential to provide cost-effective healthcare.

References

- [1] Breast cancer in Egypt: The challenges include education and detection [Online]. Available: <http://communities.washingtontimes.com/neighborhood/egypt-pyramids-and-revolution/2011/oct/11/breast-cancer-egypt-challenges-education-treatment/>.
- [2] R.Nithya, and B.Santhi, "Mammogram classification using maximum difference feature selection method," *Journal of Theoretical and Applied Information Technology*, vol. 33, pp. 197-204, 2011.
- [3] R.Nithya, and B.Santhi, "Comparative study on feature extraction method for breast cancer classification," *Journal of Theoretical and Applied Information Technology*, vol. 33, pp. 220-226, 2011.
- [4] H.S.Sheshadri and A.Kandaswamy, "Breast tissue classification using statistical feature extraction of mammogram," *Medical Imaging and Information Sciences*, vol.23, pp105-107, 2006.
- [5] N. Pradeep, H. Girisha, B. Sreepathi, and K. Karibasappa, "Feature extraction of mammograms," *International Journal of Bioinformatics Research*, vol. 4, pp. 241-244, 2012.
- [6] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," *IEEE Trans.Med.Imag.*, vol. 19, pp. 115-126, 2000.
- [7] T.C. Wang and N.B. Karayiannis, "Detection of microcalcifications in digital mammograms using wavelets," *IEEE Trans.Med.Imag.*, vol. 17, pp. 498 - 509, 1998.
- [8] Weka 3: Data Mining Software in Java [Online]. Available: <http://www.cs.waikato.ac.nz/ml/weka/index.html>.
- [9] J. Guyon, S. Weston, and V. Barnhill, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.
- [10] W.A Mohamed and Y.M.Kadah, "Computer aided diagnosis of digital mammograms," *Proc. International Conference on Computer Engineering & Systems, 2007 (ICCES '07)*, pp.299 – 303, 2007.
- [11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, "Numerical Recipes 3rd Edition: The Art of Scientific Computing," Cambridge University Press, Cambridge, 2007.
- [13] Introduction to support vector machine by Dustin Boswell [Online]. Available: <http://www.work.caltech.edu/~boswell/IntroToSVM.pdf>.
- [14] BeagleBoard-xM Rev. C System Reference Manual [Online]. Available: http://beagleboard.org/static/BBxMSRM_latest.pdf.
- [15] BeagleBoard-xM Product Details [Online]. Available: <http://beagleboard.org/hardware-xM>.
- [16] Linux-based systems for embedded products regardless of the hardware architecture [Online]. Available: <https://www.yoctoproject.org>.
- [17] Beagle Board Beginners [Online]. Available: <http://elinux.org/BeagleBoardBeginners>.
- [18] TMS320C64x/C64x+DSP CPU and Instruction Set Reference Guide [Online]. Available: <http://www.ti.com/lit/ug/spru732j/spru732j.pdf>.
- [19] C.Rinaldi, *Performance evaluation and optimization of an OMAP platform for embedded SDR systems*, Master of Science Thesis, School of Information and Communication Technology, Royal Institute of Technology (KTH), Stockholm, Sweden, 2011.
- [20] Syslink Project [Online]. Available: http://omappedia.org/wiki/Syslink_Project.
- [21] A Practical Guide to Support Vector Classification [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- [22] TMS320C6000 Optimizing C Compiler User's Guide [Online]. Available: <http://www.ti.com/lit/ug/spru187o/spru187o.pdf>.