

Software System for Automated Assessment of the Breast Cancer Risk from Mammograms

Ivan Radosavljević*, Aleksandra Mitrović*, Zora Konjović**, Đorđe Obradović**, Vuk Malbaša*

* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

** Singidunum University, Novi Sad, Serbia

ivanradosavljevic@uns.ac.rs, aleksandramitrovic@uns.ac.rs, zkonjovic@singidunum.ac.rs, djobradovic@singidunum.ac.rs, vmalbasa@gmail.com

Abstract—This paper will present a new software system that enables automated assessment of the breast cancer risk from the mammographic images. The aim of the system is to provide for a quick and efficient detection of all potentially malignant abnormalities from the mammograms and determination of preliminary malignancy risk for each. Such preliminary risk assessment could, further, be used either to terminate the procedure (small risk), or as a priority by which the patient will be scheduled for following procedural steps. The system extracts the data necessary for risk assessment from the mammogram. For that purpose methods for digital image processing and analysis are used. Interpretation of the extracted data and the final risk assessment require a method capable to deal with imprecise and incomplete knowledge because no exact mathematical model for estimation of potentially malignant abnormalities risk exists. Therefore, risk assessment is implemented by means of a fuzzy controller capable to imitate the work of a human expert / radiologist. The paper presents architecture of the system for automated assessment of the breast cancer risk from mammograms, some specific details of its implementation, and discussion of the system's accuracy.

I. INTRODUCTION

Mammography is a noninvasive procedure used to detect early signs of breast cancer. This procedure is conducted by first taking an X-ray image of a breast. The image is then visually inspected. The inspection of these images is done by expert radiologists who have the goal of discovering any harmful irregularities. Efficient discovery of these irregularities relies on factors such as the quality of the acquired image and the experience of the radiologist conducting the inspection, both of which can vary. The quality of the image can be reduced by human error e.g. not positioning the breast correctly during the imaging or not setting up the imaging device correctly. It can also be reduced by structural properties of a breast such as dominant presence of dense glandular tissue. Even when the quality of the image is acceptable the inspection of the image can take up to an hour. This amount of time is acceptable during the usual inspection of mammograms of single patients. However, mammography is also used during screenings, when a large number of patients in a short period of time are subjected to X-ray imaging. The amount of images acquired in this manner can be vast, but it is required that these images be accurately inspected as soon as possible. This combination of requirements can put a lot of stress on the individuals inspecting the images which can lead to subpar inspections of the aforementioned X-ray images [1].

In this paper, a possible solution in the form of a system for automatic risk assessment, will be presented. The goal of this system is to detect possible signs of irregularities visible on the mammograms and to assess the risk for each of the detected irregularities. This system could be used to prioritize mammograms based on the assessed risk, enabling radiologists to focus their attention on inspecting high risk mammograms and then later on the low risk mammograms.

Since the data required to assess the risk is in the form of an image, it has to be extracted in a format appropriate for computer analysis. For this purpose the techniques of digital image processing and analysis will be utilized.

The extracted data then has to be interpreted. This interpretation has to be carried out in such a manner that it allows the use of incomplete and imprecise knowledge. The reason this requirement exists is the lack of an exact mathematical model that could be used to assess the risk. Instead it is necessary to model the process of inspection carried out by the expert radiologists who use qualitative measures such as far, close, large, small etc. and not precise quantitative measures to describe the features of the image they are inspecting. Thus a fuzzy logic controller [2] will be used for the purpose of interpreting the extracted data and risk assessment.

In the following chapter an outline of the possible problems encountered in creating a system for automated assessment of breast cancer risk, as well as some previous research in this area will be presented. The third chapter describes the methodology used to solve some of the aforementioned problems. The following chapter shows the details of the implementation. The fifth chapter shows the achieved results of the solution presented in this paper. The final chapter present the conclusion as well as possible further research directions and improvements.

II. RESEARCH QUESTIONS

The problem addressed in this research is automation of the process of medical examination intended for early breast cancer detection, when massive screenings using mammography are conducted. The critical point of the process is image interpretation which should result in reliable risk estimate. Automation of this activity could satisfy two contradictory requirements, reliability of risk estimation and limited time available for image processing.

The image processing itself presents a problem since the quality of the images can greatly vary between patients. Also the high resolution of the images means that the hardware requirements for fast processing are very high. These can be lowered by downscaling the image, but at a cost of losing many of the fine details in the image required to extract useful features.

Extraction of these useful features such as chest muscle, glandular tissues, calcifications and potential tumors is greatly impaired by the difficulty of discerning between these features. For example a large calcification can appear to be a tumor or glandular tissue, which can lower the precision of risk assessment.

Even if the image quality is acceptable and the features are extracted successfully, the risk assessment is not a trivial task. Since there is no exact mathematical model to describe this risk assessment, the model has to be created in such a manner that it mimics the expert radiologist. This means that the model has to be capable to operate with incomplete and imprecise knowledge as well as qualitative variables used by expert radiologists to describe features of the X-ray images.

Similar research in this field was conducted by a number of authors and is discussed in [3]. This paper shows the comparison of different approaches with respect to their accuracy and the possibility of clinical use. The best performing methods presented in this paper are shown in Table I.

Although the accuracy of these algorithms is high, the paper emphasizes that these algorithms are not suitable for clinical use. In the paper it is also noted that the high accuracy is usually achieved on specific datasets, and that these methods cannot be generalized. The authors also point out that conventional methods of image processing and analysis have lower performance than the methods employing domain specific knowledge.

III. METHODOLOGY

The solution presented in this paper is separated into two main components:

- image processing software, and
- risk assessment software.

Both of these components are a separate software solution dedicated to a subset of problems.

The image processing software is used to extract important features from the raw images. These features include:

- the breast,

- the chest muscle,
- forbidden areas, and
- high density areas.

The extraction of these features is carried out by separate modules of which the image processing software is composed. Figure 1 shows each of the modules and the data flow through those modules.

The features extracted using the image processing software are then fed into the risk assessment software. The risk assessment software utilizes a fuzzy logic controller to assess the risk. This piece of software is also used to show the results of assessment and to visualize those results.

IV. SOLUTION

In this chapter the implementations of image processing and risk assessment software will be presented.

A. Image processing software

Before the images can be used for any processing they first have to be converted into an appropriate format. The format of a raw image is shown in Table II. The height and width are 32-bit values representing the height and width of an image in pixels. The following 16 bits are used to store the intensity threshold used to calculate the final intensities of pixels. The values of interest lookup table (VOI LUT) length, a 32-bit value, stores the number of elements in the lookup table. The lookup table itself is an array of 32-bit values used to calculate the intensities of pixels. The final array are the values of raw intensities of pixels each represented as a 16-bit integer.

After this raw image is loaded, the minimal threshold and the VOI LUT are applied to all of the pixels. The result of this operation is an image without noise created by an oversensitive sensor used in the imaging device.

The image acquired in this process is suitable for further processing, however it is not suitable to be used as a display image. This problem was solved by creating a heatmap image from the previously acquired image. An example of a generated heatmap is shown in Figure 2.

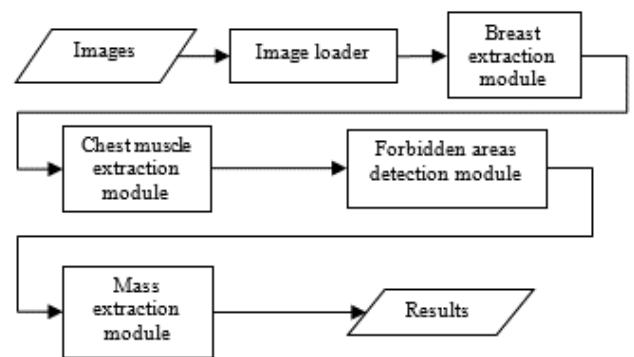


Figure 1. Data flow through the image processing software

TABLE I. THREE BEST PERFORMING BREAST CANCER DETECTION METHODS

Method	Accuracy
Region based edge-profile acutance measure [4]	92%
Model based vision algorithm. Difference of Gaussians and texture features [5]	92%
Wavelet features in combination [6]	90%

TABLE II. INPUT IMAGE FILE FORMAT

Width	Height	Intensity threshold	VOI LUT length	VOI LUT	Intensities
4B	4B	2B	4B	(VOI LUT length*4)B	(Width*Height*2)B

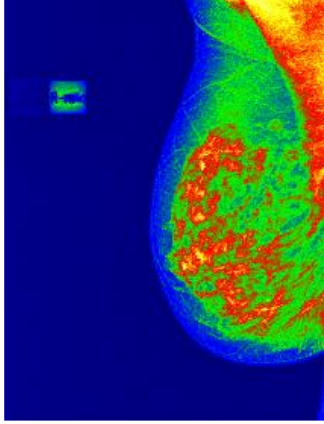


Figure 2. Breast image heatmap

1) Breast extraction module

After the image has been prepared for processing the first step of the actual processing is to extract the breast from the image and create a binary image of this region of interest (ROI). This is accomplished by first applying the Gaussian filter to the original image to reduce the high frequency noise [7]. The following step is to determine the threshold to segment the image into the breast area and the background area. For this purpose the Triangle method [8] is used. After the threshold has been calculated, each of the pixels of the image are compared to it. Pixels below the threshold are classified as background, and those above the threshold are classified as foreground. Figure 3 shows the results of segmentation. As can be seen from this image there are multiple areas that were above the threshold. Some of these areas may be labels often found on X-ray images while some are just noise. For the purpose of this solution it is assumed that the largest area, the one with the highest white pixel count, is the breast area and it is labeled as such. The extraction of individual white areas is done by using

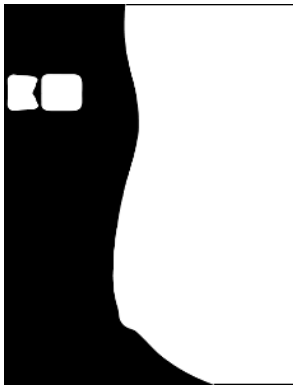


Figure 3. The results of the image segmentation

the depth first search (DFS) algorithm to extract the connected components. During the extraction, the total count of pixels is calculated for each component. After the extraction the largest component is used to create a new binary image representing a mask for the ROI, i.e. the breast.

Besides extracting the breast, this step is also used to create a polygon from each of the detected regions. This is done to simplify the visualization of the extracted regions and to provide a more compact representation of the extracted regions than the high resolution binary images. The process starts by first detecting the edges of each binary region. This is achieved by applying a morphological edge detection [9]. The following step is the extraction of vertices of the polygon. This is done by removing all lines on an angle of 0° , 45° , 90° or 135° and leaving only their endpoints. These endpoints represent the vertices of a polygon. The extracted vertices are then sorted in clockwise order. This was necessary to draw the extracted polygons correctly.

2) Chest muscle extraction module

Like in the breast extraction module, the first step in this module is to apply the Gaussian filter in order to reduce the high frequency noise. The next step is the application of a mask acquired in the previous module to the filtered image. The resulting image contains only the ROI. From this image, a histogram of pixel intensities is created. This histogram is then split into eight equal bins and is used to create a new image. This image is created by replacing the values of pixels with the number of bin to which the given pixel value belongs. The result is an image with reduced range of pixel intensities, in this case only eight. Figure 4 shows the result of this operation. In the following step the edges that separate the muscle from the rest of the breast are extracted from the previously created image. For this purpose the Sobel filter is used [7]. As can be seen from Figure 5 the Sobel filter detects all the edges, not only the ones of the muscle. In order to detect which of these edges belong to the muscle it is necessary to first define constraints which these edges have to satisfy to be considered muscle edges. In the case of this paper it is assumed that these constraints are:

- the edge must be continuous,
- the angle of the edge must be between 20° and 70° for the left muscle,

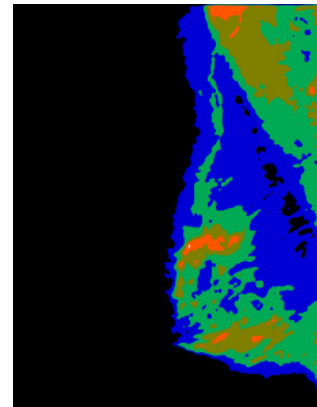


Figure 4. The image with reduced range of pixel intensities



Figure 5. Edges extracted using the Sobel filter

- the angle of the edge must be between -70° and -20° for the right muscle, and
- the muscle edge must be the longest edge satisfying the previous three constraints.

These constraints can further be simplified by assuming that the edges are straight lines. Thus the edge that satisfies these constraints can be found by using the Hough transform feature extraction technique to detect all lines [7] and then finding the one line that fits those constraints the most. The next step in finding the muscle area is subtracting the edges from the binary image of the breast, thus acquiring an image of parts of the breast with different intensities. The position of these individual parts is then compared to the chest muscle line. The parts whose area is at least 70% on the left side of the line are considered to be on the left side, otherwise the parts are considered to be on the right side i.e. to belong to the chest muscle or to belong to the breast, respectively. All of the parts that were classified as parts of the chest muscle are then joined into a binary image that represents a chest muscle area. Figure 6 shows the end result of chest muscle extraction.

3) Forbidden areas detection module

Forbidden areas are areas where no high density masses should appear. If any high density masses appear in these areas they are considered high risk masses. These areas include the areas around the outside of the chest muscle edge and around the inner side of the breast edge. The process of extracting these areas starts with detecting the edges of the chest muscle and the breast. This is done by applying the morphological edge detection to the binary images of the breast and the chest muscle. In the following step the distance transform is calculated for the image containing the edges of the chest muscle and the breast. The resulting image contains the pixels whose value is equal to the distance of that pixel to the nearest edge. Since the pixels with a distance of less than 300 pixels from the edges are the only ones considered relevant the rest of the pixels are labeled as having a infinite distance. The result of this process is shown in Figure 7.

4) Mass extraction module

High frequency noise is removed by using the Gaussian filter. The result of subtracting the binary image of the chest muscle from the binary image of the breast is then applied as a



Figure 6. Extracted chest muscle

mask on the filtered image. The resulting image is then transformed to reduce the range of pixel intensities to the interval $[0, 8]$ by using the method described in I.A.2).

The segmentation of the resulting image is carried out by applying a predetermined threshold. The threshold value used in this solution is 4 units of pixel intensity. Any pixel whose intensity surpasses this threshold is considered high density tissue and as such is extracted from the image. Regions of interest are extracted from the resulting binary image by utilizing the DFS algorithm. The extracted regions are then filtered according to their area. Any region with an area less than 25 mm^2 is discarded.

For each of the remaining ROI, their parameters are calculated. Those parameters are later utilized in the risk assessment for that region. These parameters are:

1. the area of the region,
2. the center of mass,
3. the angle along the major axis,
4. eccentricity of the mass region,
5. the distance from the edge of the breast, and
6. the distance from the edge of the chest muscle.

This procedure is applied to the images of the left and the right breast, resulting in a separate set of masses for each image. This is followed by the pairing of related masses from both sets.

The pairing of masses is done by comparing the distances between the masses from the left image and those from the right image. The next step is to find a mass detected on the right image that is closest to the transformed center of the mass

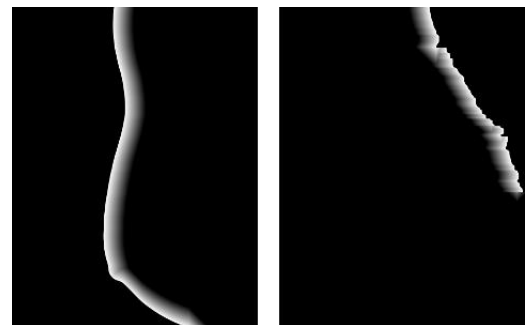


Figure 7. Result of the applied Distance transform

detected on the left image while not being further away from the center than 3cm. This is done for each mass detected on the left image. The transformed center is calculated by mirroring the coordinate of a center of a mass on the left image to the right image. If, for a given mass on the left image, no mass on the right image satisfies these conditions, that mass is considered to not have a pair. This procedure is repeated in the same manner for each mass on the right image.

After the pairing is completed, three additional mass parameters are calculated:

1. the distance from the closest mass,
2. the difference between the angles along the major axis, and
3. the difference between the areas.

For each mass, a polygon is formed for later use in visualization. An example of masses extracted in the aforementioned process is shown in Figure 8.

All of the data acquired from the image processing software component is forwarded to the risk assessment software component.

B. Risk Assessment Software

The risk assessment software component uses a fuzzy logic controller to estimate the risk value of a mass. The fuzzy controller has the following input parameters:

1. the distance from the edge of the breast,
2. the distance from the edge of the chest muscle,
3. the area of the mass,
4. eccentricity of the mass region,
5. the distance between a mass and the mass closest to it,
6. the difference between angles along the major axis, and
7. the difference between the area of a mass and the area of the mass closest to it.

These values are then fuzzified and used to determine the degree to which a previously constructed set of rules is satisfied. These rules are:

1. IF (*close to breast edge* OR *close to chest muscle*) AND *eccentricity low* OR *difference between angles along the major axis large* OR (*large area* OR *large*

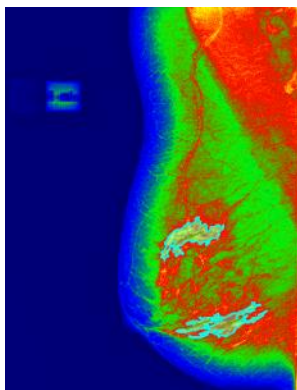


Figure 8. Result of mass extraction

area difference) AND *large distance between a mass and the mass closest to it*) THEN *risk high*,

2. IF (*far from breast edge* OR *far from chest muscle*) AND *eccentricity low* AND *difference between angles along the major axis small* OR (*medium area* OR *medium area difference*) AND *large distance between a mass and the mass closest to it*) THEN *risk medium*, and
3. IF (*far from breast edge* OR *far from chest muscle*) AND *eccentricity high* AND *difference between angles along the major axis small* OR (*small area* OR *small area difference*) AND *small distance between a mass and the mass closest to it*) THEN *risk low*.

The rule that is satisfied the most is considered the only satisfied rule. The degree to which this rule is satisfied represents the risk estimate value for a given mass.

V. RESULTS

The evaluation of the segmentation was conducted by visual inspection of each of the 25 pairs of images. The segmentation of a pair of images is considered correct if every ROI is successfully extracted. An extraction is successful if the ROI is extracted in its entirety and without visible errors. Errors include: missing parts of the ROI, merging two or more ROI, failing to detect ROI, falsely declaring parts of the image to be ROI.

As is evident from Table III, in 6 out of 25 images, the breast was not successfully extracted. In all of the six images the problem manifested as a merger of the visible part of the patient's arm with the breast and the labels on the image. In 5 cases the chest muscle was only partially extracted.

The extraction of masses had the worst performance. The masses were successfully extracted in only 6 out of 11 cases. Failure to correctly extract masses adversely affected the subsequent pairing of masses. These failures happened due to the similar densities of the masses and the surrounding tissue.

The accuracy of risk assessment depends on the accuracy of the segmentation process. Thus, in every case where the segmentation was not successful, the risk estimation was incorrect. In the remaining cases, it was evident that the risk estimation was leaning towards overestimating the risk values, especially in the case of paired masses. This was due to the fact small differences in paired masses caused large differences in risk estimation. The problem was additionally aggravated by the inflexibility of the ruleset for risk estimation.

VI. CONCLUSION

The solution presented in this paper is a software architecture and an implementation of a system for the segmentation and

TABLE III. PERFORMANCE ACHIEVED ON TEST DATA

Total images	Brest extraction errors	Chest muscle extraction errors	Mass extraction errors	Overall errors
25	6	5	11	12

evaluation of mammograms. The implemented system is divided into two components: an image processing software and a risk assessment software. The image processing software is used to extract important features from raw images, including the breast, chest muscle, forbidden areas, and high density areas. It is also responsible for pairing up the high density areas, and the calculation of their features. The results from the image processing software are then additionally processed and displayed in the risk assessment software component. Risk assessment for every extracted mass is also done during this processing. The assessment is implemented with a fuzzy controller whose inputs are features of the extracted masses. Every mass is assigned a risk value using a predetermined set of rules.

The performance of both software components has also been evaluated. These evaluations show that the segmentation software had the lowest error rates during the segmentation of the breast and chest muscles, 24% and 20% respectively, and the highest error rate of 44% during the segmentation of masses. These results had an adverse effect on subsequent processing steps, leading to inflated risk values.

Plans for the future improvement of the solution presented in this paper include:

- improvement of high density area segmentation by utilizing blob detection techniques,
- improvement of the method for chest muscle detection by using a curve to model the muscle edge,
- the introduction of extra parameters for risk estimation, for example: the difference in density between extracted masses, the difference between a breast's average density and the density of extracted masses etc.,
- the introduction of additional rules for risk estimation,
- the acquisition of additional data such as the age of the patient, their medical history, smoking habits etc., and
- the detection of calcifications and the inclusion of their densities as a parameter for risk assessment.

ACKNOWLEDGMENTS

The authors would like to thank the “Dr Milenko Marin” hospital in Loznica, for providing the data necessary to conduct this research. They are also grateful to Vesna Lazić, M.D. for sharing her knowledge and experience with the authors.

REFERENCES

- [1] P. Hogg, J. Kelly, and C. Mercer, Eds., *Digital Mammography: A Holistic Approach*. Springer International Publishing, 2015.
- [2] H.-J. Zimmermann, *Fuzzy Set Theory and Its Applications*. Dordrecht: Springer Netherlands, 2013.
- [3] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K.-H. Ng, “Computer-Aided Breast Cancer Detection Using Mammograms: A Review,” *IEEE Reviews in Biomedical Engineering*, vol. 6, pp. 77–98, 2013.
- [4] R. M. Rangayyan and S. G. Elkadiki, “Algorithm for the computation of region-based image edge profile acutance,” *J. Electronic Imaging*, vol. 4, pp. 62–70, Jan. 1995.
- [5] W. E. Polakowski *et al.*, “Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency,” *IEEE Trans Med Imaging*, vol. 16, no. 6, pp. 811–819, Dec. 1997.
- [6] H. YOSHIDA, “Computer-aided diagnosis in mammography: Detection of clustered microcalcifications based on multiscale edge representation,” *Computer Assisted Radiology, 10th International symposium and Exhibition*, pp. 390–395, 1996.
- [7] S. E. Umbaugh, *Digital Image Processing and Analysis: With MATLAB and CVIptools*. Taylor & Francis, CRC Press, 2017.
- [8] G. W. Zack, W. E. Rogers, and S. A. Latt, “Automatic measurement of sister chromatid exchange frequency,” *J Histochem Cytochem.*, vol. 25, no. 7, pp. 741–753, Jul. 1977.
- [9] B. Jähne, *Digital Image Processing*, 6th ed. Berlin Heidelberg: Springer-Verlag, 2005.