Using Artificial Immune Recognition Systems in Order to Detect Early Breast Cancer

C.D. Katsis
Technological Educational Institute of the Ionian Islands, Dept. of Applications of Information Technology in Administration & Economy, Lefkada Greece
Email: ckatisis@teiion.gr

I. Gkogkou
University Hospital of Ioannina, Dept. of Radiology, Ioannina, Greece
Email: iokastigo@yahoo.com

C.A. Papadopoulos
University Hospital of Ioannina, Dept. of Radiology, Ioannina, Greece
Email: shine921@gmail.com

Y. Goletsis
University of Ioannina, Dept. of Economics, Ioannina, Greece
Email: cgoletsis@cc.uoi.gr

P.V. Boufounou
Technological Educational Institute of the Ionian Islands, Dept of Business Administration, Lixouri, Kefalonia Greece
Email: vivian.boufounou@gmail.com

G. Stylios
Technological Educational Institute of the Ionian Islands, Department of Business Administration, Lefkada Greece
Email: gstylis@teiion.gr

Abstract—In this work, a decision support system for early breast cancer detection is presented. In hard to diagnose cases, different examinations (i.e. mammography, ultrasonography and magnetic resonance imaging) provide contradictory findings and patient is guided to biopsy for definite results. The proposed method employs a Correlation Feature Selection procedure and an Artificial Immune Recognition System (AIRS) and is evaluated using real data collected from 53 subjects with contradictory diagnoses. Comparative results with commonly used artificial intelligence classifiers verify the suitability of the AIRS classifier. The application of such an approach can reduce the number of unnecessary biopsies.

Index Terms—Artificial Immune Recognition System, Breast Cancer, Correlation Feature Selection, Decision Trees, Multilayer Perceptron Artificial Neural Networks, Support Vector Machines

I. Introduction

Cancer begins with the uncontrolled division of one cell and results in a visible mass named tumor. Tumor can be benign or malignant. Malignant tumor grows rapidly and invades its surrounding tissues through causing their damage. Breast cancer is a malignant tissue beginning to grow in the breast. The abnormalities like existence of a breast mass, change in shape and dimension of breast, differences in the color of breast skin, breast aches, etc., are the symptoms of breast cancer. The aforementioned disease is the second leading cause of cancer deaths in women today (after lung cancer) [1] and is the most common cancer among women, excluding non-melanoma skin cancers.

During the last decade, breast cancer outcomes have improved with development of more effective diagnostic techniques and improvements in treatment methodologies. The long-term survival rate for women in whom breast cancer has not metastasized has
increased, with the majority of women surviving many years after diagnosis and treatment. A key factor in this trend is the early detection and accurate diagnosis of the disease [2]. For that reason, women are subjected to screening, by means of mammography (MG). In many cases, lesions discovered need further evaluation, accomplished by means of Ultrasonography (US) and Contrast-Enhanced Magnetic Resonance Imaging Tomography (CE-MRI). From all the above mentioned modalities, underlying lesions are evaluated, determining the possibility of malignancy. During the imaging routine, lesions are characterized using specific features related with breast cancer risk. In some cases, occult or controversial findings between the various modalities can be met, resulting in equivocal lesions’ assessment, leading to unnecessary core or open breast biopsy. Especially in these cases of diagnostic dilemmas between the MG, US and CE-MRI modalities, there is a lack of evidence regarding the correlation of these features with breast cancer.

The last decade, the use of classification systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. Expert systems and different artificial intelligence techniques for classification also help experts in a great deal. Classification systems on the one hand help to minimize possible errors that can be done because of fatigued or inexperienced physician and on the other hand, provide medical data to be examined in shorter time and in more detail. Automated diagnostic systems have been applied to and of are interest for a variety of medical data, such as electrocardiograms (ECGs), electromyograms (EMGs), electroencephalograms (EEGs), ultrasound signals/images, X-rays, and computed tomographic images [3-13]. Moreover, the economic and social values of breast cancer diagnosis are very high. Therefore, the problem has attracted many researchers in the area of computational intelligence recently [14-19]

A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task (i.e. breast cancer). In this framework, a feature selection method has been applied aiming at reducing the set of features that efficiently describe the dataset and in this way at providing a simpler classification model. The CFS algorithm, proposed by Hall [20], is based in the central hypothesis that good feature sets contain features that are highly correlated with the class (malignancy or benignity), yet uncorrelated with each other. CFS is a filter approach [21] independent of the classification algorithm by considering the individual predictive ability of each feature along with the degree of

<table>
<thead>
<tr>
<th>Modality</th>
<th>MG</th>
<th>US</th>
<th>CE-MRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracted features</td>
<td>density</td>
<td>US-margins</td>
<td>CE-MRI margins</td>
</tr>
<tr>
<td></td>
<td>margins</td>
<td>Acoustic shadow</td>
<td>Time signal intensity curve</td>
</tr>
<tr>
<td></td>
<td>Architectural distortion</td>
<td>vascularity</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>microcalcifications</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In this work, we propose a methodology that ranks the multimodal extracted features of the lesions and acts as a decision support system which provides a prognosis of malignancy. In the following paragraphs, we first outline the steps of our methodology. We then present our experimental results; finally a comparison of our proposed, Artificial Immune Recognition System (AIRS) based, method with other commonly used artificial intelligence classifiers is provided.

II. Proposed Methodology

The proposed methodology uses a Correlation Feature Selection (CFS) procedure to rank the extracted multimodal features and an Artificial Immune Recognition System (AIRS) classifier in order to support breast cancer diagnosis. Table 1 provides the lesions’ features extracted from the MG, US and CE-MRI modalities. It must be noticed that no special attributes are necessary to be extracted for our methodology since the same features are used in the daily clinical routine by the physicians to diagnose breast cancer. The overall methodology schema is illustrated in Figure 1.
redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

Classification occurs using an unweighted k-Nearest Neighbor approach [24, 35]. According to this approach the class of a new instance is determined as the class of the majority of the k nearest training examples. Proximity of examples is calculated using a measure of distance, commonly Euclidean distance in the case of continuous variables and Manhattan distance in the case of nominal variables. Algorithm 1 summarizes the training procedure of the AIRS algorithm [25]. The evaluation results achieved by the proposed methodology are provided next.

![Diagram](image)

**Fig. 1:** The proposed methodology schema: For each patient Lesion’s features are extracted through the data acquisition module which consists of the Mammography, Ultrasound and the Magnetic resonance modalities. To provide a simpler classification model a CFS Feature extraction module is used. The classification module is based on the AIRS classifier in order to support breast cancer diagnosis (malignancy or benignity).

The natural immune system is a complex, robust, biological system within an organism that protects against disease by identifying and killing pathogens. It is able to distinguish organism’s own healthy cells and tissues from a wide variety of viruses and parasitic worms. It is adaptive, complex, capable of maintaining memory of previous encounters, to name just a few of its more attractive computational properties. AIRS, is a supervised, immune inspired learning algorithm [22-23]. AIRS algorithm aim is to prepare a pool of memory (recognition) cells representative of the training data the model is exposed to, and suitable for classifying unseen data. The recognition cells in the memory pool are stimulated by an antigen and each cell is allocated a stimulation value. The memory cell with the greatest stimulation selected as the best match memory cell for use in the affinity maturation process.

**III. Results**

To evaluate our methodology, we have gathered data arising from 53 subjects out of 4726 cases. The specific subjects presented lesions that were not highly suggestive of benignity or malignancy when evaluated on every modality used. In all cases biopsy was conducted and the biopsy results were used as golden standard to validate our methodology. The constructed dataset consists of the features presented in Table 1 as well as the biopsy results (malignancy or benignity) for all 53 subjects. All data were collected in the University Hospital of Ioannina, Greece.

The performance of the AIRS classifier on the above dataset has been evaluated. The parameters of the classifier have been selected according to the literature and experimentally. Specifically, the initial ARB cell pool size was set to 1, the number of mutated clones to create of an ARB was set to 80 and the maximum number of resources that can be allocated to ARBs in the ARB pool was set to 300. According to [35] the usual numbers of k are in the range from 1 to 7. For this reason 7 variations of AIRS have been tested, according to the value of k. Having applied the Correlation based Feature extraction the selected features are: (i) Mammography architectural distortion, (ii) Ultrasound margins, and (iii) Ultrasound acoustic shadow.
The importance of the selected features was evaluated by two experienced breast radiologists. According to them the mammography architectural distortion feature is of great importance since invasive carcinoma distorts the interfaces between fat and normal breast parenchyma due to the response of host tissues to the malignancy. Especially, in the very dense breast, the tumor mass can be so obscured by adjacent benign tissues as to be invisible, leaving as the only indication of underlying malignancy an area of focal architectural distortion [36]. Moreover, the ultrasound margins have been the most commonly reported findings in the literature during the past 20 years. The presence of angular margins is a hard finding, indicative of invasive malignancy in most instances. Finally, the acoustic shadow is a finding that reflects the surrounding tissue’s reaction induced by malignant masses. It can occur with either invasive malignancy or ductal carcinoma in situ [37]. We have compared the AIRS performance (for k=1 to 7) with the results obtained by widely used classification methodologies such as Multilayer Perceptron (MLP) Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Decision Trees (C4.5 algorithm). AIRS and the comparative classification scheme is evaluated in two modes: (i) using the full set of features, (ii) using a subset of features obtained by applying the Correlation based Feature Selection (CFS) method. In order to minimize the bias associated with the random sampling of the training and testing data samples, 10fold cross validation is applied. The obtained results are provided in Tables 2 and 3.

### Table 2: AIRS classification results obtained for different k values

<table>
<thead>
<tr>
<th>k #</th>
<th>Using the full set of features</th>
<th>Using the subset of CFS selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIRS (k=1)</td>
<td>73.67 ±1.99</td>
<td>68.00±5.92</td>
</tr>
<tr>
<td>AIRS (k=2)</td>
<td>67.67 ±3.52</td>
<td>59.67±5.75</td>
</tr>
<tr>
<td>AIRS (k=3)</td>
<td>83.33 ±6.63</td>
<td>81.32±5.34</td>
</tr>
<tr>
<td>AIRS (k=4)</td>
<td>75.33 ±5.41</td>
<td>60.00±7.20</td>
</tr>
<tr>
<td>AIRS (k=5)</td>
<td>81.00 ±8.19</td>
<td>67.67±10.66</td>
</tr>
<tr>
<td>AIRS (k=6)</td>
<td>73.67 ±6.21</td>
<td>69.67±6.66</td>
</tr>
<tr>
<td>AIRS (k=7)</td>
<td>77.33 ±5.77</td>
<td>77.00±2.36</td>
</tr>
</tbody>
</table>

### Table 3: The AIRS, MLP, SVM and C4.5 classification results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%) ± STD</th>
<th>Using the full set of features</th>
<th>Accuracy (%) ± STD</th>
<th>Using the subset of CFS selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIRS</td>
<td>83.33±6.63</td>
<td></td>
<td>81.32±5.34</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>73.67±4.32</td>
<td></td>
<td>70.90±4.81</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>70.00±6.33</td>
<td></td>
<td>68.92±3.97</td>
<td></td>
</tr>
<tr>
<td>Decision trees (C4.5)</td>
<td>66.57±2.21</td>
<td></td>
<td>66.15±3.18</td>
<td></td>
</tr>
</tbody>
</table>

As it can be seen from Table 3, the AIRS classifier (using 3-NN) achieves high classification rate compared to the ANN, MLP, SVM and C4.5 approach both using the full set of features or using CFS.
IV. Conclusion

In this work, we have presented a methodology that evaluates the multimodal extracted features of the lesions and provides information to the radiologist regarding breast cancer diagnosis. Moreover we have constructed a dataset containing exclusively equivocal findings between the MG, US, and CE-MRI modalities. The applicability and performance of the Artificial Immune Recognition Systems to our dataset was examined. The classification accuracy of the AIRS algorithm was superior compared to conventional classification schemas. A direct comparison with other methodologies is not feasible since according to our knowledge there is no published work using a combination of MG, US and CE-MRI modalities in obscured findings. The achieved initial results are promising keeping in mind that our constructed dataset consists exclusively of equivocal cases. Our Future work will concern to enrich the constructed database with more equivocal findings and to provide a decision support system useful to the clinical practice aiming to decrease the number of unnecessary biopsies, and by this way to reduce the cost and the rate of complications.

References

[19] S. Belcici, E. El-Darzi, A partially connected neural network-based approach with application to...


Authors’ Profiles

Dr. Christos D. Katsis was born in Preveza, Greece. He received the Diploma Degree in Physics in 1998 and the PhD degree in Medical Physics from the University of Ioannina, Ioannina Greece in 2008. He joined the Dept. of Applications of Information Technology in Administration & Economy of the Technological Educational Institute of Ionian Islands in 2009 and he is currently an adjunct associate professor. He has many years of research experience in biomedical engineering, computer modeling, computational medicine, automated diagnosis and telemedicine. He has published many papers in refereed journals, international conferences and books.
I. Gkogkou, MD, was born in Ioannina, Greece. She received her diploma degree in Medicine in 2004 from the University of Ioannina. Since 2007 she is with the Dept. of Radiology, University Hospital of Ioannina, Greece. Her research interests include Ultrasonography, Contrast-Enhanced Magnetic Resonance Imaging Tomography and breast cancer diagnosis.

C.A. Papadopoulos, MD was born in Athens, Greece. He received his diploma degree from the University of Ioannina in 2000. He is currently with the University Hospital of Ioannina, Dept. of Radiology. His research interests include Ultrasonography, Magnetic Resonance Imaging Tomography and breast cancer diagnosis. He has published many papers in refereed journals, international conferences and books.

Dr. Yorgos Goletsis holds a Diploma Degree in electrical engineering and the Ph.D. degree in Operations Research, both from the National Technical University of Athens, Athens, Greece. He is a Lecturer in the Department of Economics, University of Ioannina. His research interests include operations research, decision support systems, multicriteria analysis, quantitative analysis, data mining, artificial intelligence, project evaluation.

Mr. George Stylios received his diploma at Physics from the Aristotle University of Thessaloniki in 1994 and his MSc in Information Systems & Technology from the City University, London, in 2002. Currently he works as a Lecturer at the Technical Educational Institute of Ionian Islands, Department of Information Technology Applications in Management and Economics, Lefkada, Greece and he is head of the department. He is a PhD student at the Databases Laboratory, Department of Computer Engineering & Informatics of the University of Patras. His research interests include: E-government, data mining, digital citizen, e-citizen; processing data from web forums, blogs, public consultation to parse texts and opinion mining; data mining algorithms within the procedure to support business decisions; web mining. He has published numerous articles in international scientific journals and conference.

Dr. Paraskevi Boufounou was born in Athens Greece. She is Assistant. Prof. of Finance in the Dept. of Business Administration in the Technological Educational Institute of Ionian Islands. She has obtained valuable extensive high level managerial experience (including Chairman, CEO positions) in various private and public Financial Institutions. She has a broad academic experience teaching Finance and Business Development in many Universities at Undergraduate and Postgraduate level. She has published many papers in refereed journals, international conferences and books.