Research, Society and Development, v. 11, n. 5, e3711521941, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i5.21941

# Artificial immune systems applied to clinical diagnosis of breast cancer samples

## Sistemas imunológicos artificiais aplicados ao diagnóstico clínico de amostras de câncer de mama

## Sistemas inmunes artificiales aplicados al diagnóstico clínico de muestras de cáncer de mama

Received: 10/18/2021 | Reviewed: 10/25/2021 | Accept: 03/22/2022 | Published: 03/29/2022

**Simone Silva Frutuoso de Souza**
ORCID: https://orcid.org/0000-0002-4059-0813
Universidade do Estado de Mato Grosso, Brazil
E-mail: simonefrutuoso.mat@gmail.com
**Fábio Roberto Chavarette**
ORCID: https://orcid.org/0000-0002-1203-7586
Universidade Estadual Paulista, Brazil
E-mail: fabio.chavarette@unesp.br
**Fernando Parra dos Anjos Lima**
ORCID: https://orcid.org/0000-0001-8671-1476
Instituto Federal do Mato Grosso, Brazil
E-mail: fernando.lima@ifmt.edu.br

**Abstract**
This work employments a manufactured resistant framework connected for diagnosing breast cancer tests. Taking as premise an immunological prepare, the Negative Selection Algorithm was utilized to segregate the tests, achieving a classification for generous or harmful cases. The most application of the strategy is to help experts within the breast cancer demonstrative prepare, in this manner giving decision-making agility, efficient treatment arranging, unwavering quality and the vital mediation to spare lives. To assess this strategy, the Wisconsin Breast Cancer Determination database was utilized. This is often a real breast cancer database. The comes about gotten utilizing the strategy, when compared with the specialized writing, appear precision, strength and proficiency within the breast cancer demonstrative handle.
**Keywords:** Breast cancer diagnosis; Artificial immune systems; Negative selection algorithm.

**Resumo**
Este trabalho emprega estruturas resistentes fabricadas conectadas para diagnosticar testes de câncer de mama. Tomando como premissa um preparo imunológico, o Algoritmo de Seleção Negativa foi utilizado para segregar os testes, conseguindo uma classificação para casos generosos ou prejudiciais. A maior aplicação da estratégia é ajudar os especialistas da demonstração do câncer de mama a se prepararem, dando assim agilidade na tomada de decisões, tratamento eficiente, qualidade inabalável e a mediação vital para salvar vidas. Para avaliar esta estratégia, foi utilizado o banco de dados Wisconsin Breast Cancer Database. Isso geralmente é um banco de dados real de câncer de mama. Os resultados obtidos utilizando a estratégia, quando comparados com a escrita especializada, aparecem com precisão, força e proficiência no manejo demonstrativo do câncer de mama.
**Palavras-chave:** Diagnóstico de câncer de mama; Sistemas imunológicos artificiais; Algoritmo de seleção negativa.

**Resumen**
Este trabajo emplea estructuras resistentes fabricadas conectadas para el diagnóstico de pruebas de cáncer de mama. Tomando como premisa un preparado inmunológico, se utilizó el Algoritmo de Selección Negativa para segregar las pruebas, logrando una clasificación para casos generosos o dañinos. La mayor aplicación de la estrategia es ayudar a los expertos en la demostración del cáncer de mama a prepararse, dando así agilidad en la toma de decisiones, planificación eficaz del tratamiento, calidad inquebrantable y la mediación necesaria para salvar vidas. Para evaluar esta estrategia, se utilizó la base de datos de determinación de cáncer de mama de Wisconsin. Esta es a menudo una base de datos de cáncer de mama real. Los resultados obtenidos utilizando la técnica, en comparación con la escritura especializada, parecen precisión, fuerza y habilidad en el manejo demostrativo del cáncer de mama.
**Palabras clave:** Diagnóstico de cáncer de mama; Sistemas inmunes artificiales; Algoritmo de selección negativa.

## 1. Introduction

Cancer is a chronic disease that affects millions of people worldwide. In view of an aging population and longer life expectancy, higher rates of cancer incidences have been observed, especially breast cancer, which is one of the cancers with the highest incidence rates. This type of cancer occurs mainly in medium and low developing countries where the prevalence

of cancer has become a public health problem, which requires greater vigilance, prevention behaviors, awareness policies, and effective treatments. According to the National Cancer Institute (INCA), a survey conducted in early 2013 showed that breast cancer represents 27.9% of the cases affecting women in 2012, in other words, of all existing types of cancer, breast cancer is the one that affects women most (Inca, 2021).

Worldwide, the breast cancer death rate is very high, probably because the disease is diagnosed at advanced stages. The statistics conducted by the World Health Organization (WHO) indicate that approximately 39% of the cases of women fighting breast cancer will result in death (Oms, 2021).

To reduce these rates, it is vital to conduct awareness and prevention campaigns, and especially conduct breast cancer diagnosis at an early stage. The correct diagnosis at an early stage of the disease enables quick planning decisions and evidently treatment efficiency. However, to properly diagnose cancer is a complex and difficult process because it involves many variables. The correct diagnosis requires a very experienced professional, and especially an accurate classification of the clinical stage of the tumor (cancer stage).

The traditional classification systems used are detailed and complex, and often difficult to use, posing limitations to the pathologists, and a slower decision-making process (Manikantan et al., 2009). Thus, it is vital to develop integrated systems that are capacitated to work with processing techniques and data analysis, and which, when combined with the professionals' experience, provide the necessary assistance to carry out the diagnosis and plan the treatment.

In this regard, the application of Artificial Intelligence (AI) could be a possible solution to the problems in the diagnosis process. The intelligent methods can extract information and knowledge from complex problems, and moreover they are easy to apply. There are several methods based on this concept, which are currently being used to assist professionals to perform the diagnosis of diseases, especially professionals with little experience. These methods provide security, reliability and agility in the diagnosis.

In the specialized literature some articles aimed at the diagnosis of breast cancer stand out. In reference (Pena-Reyes & Sipper, 1999), the authors present a method for diagnosing cancer samples using fuzzy logic and genetic algorithm. In (Wang & Lee, 2002) the ANFIS network (Jung, 1993) was used for the diagnosis of breast cancer, the results were considered satisfactory (96.30%). Article (Meesad & Yen, 2003) proposes a hybrid system which uses a neural network coupled to a specialized fuzzy system. The system showed good performance when applied to the Wisconsin database (Wbcd, 2021). In (Song et al., 2005), a method using a Neuro-Fuzzy Network (ANFIS) is presented for diagnosing breast cancer samples. In reference (Polat et al., 2007) an Artificial Immune Recognition System (AIRS) is presented to classify breast cancer samples. This method is based on the immune system metaphors, and seeks to reproduce mechanisms inspired by immunology, such as competition for resources, clonal selection, affinity maturation and the formation of memory cells. The resource allocation mechanism is based on the fuzzy logic. A specialized system for diagnosing breast cancer was proposed in (Karabatak et al., 2008). In paper (Naghibi et al., 2010), the authors propose a new approach using a hierarchical fuzzy neural network for the diagnosis. In (Hamdi et al., 2010) a TSK-type fuzzy model is presented for diagnosing breast cancer. In (Zhao& Davis, 2011) the authors submit a method based on the clonal selection algorithm, which adds a modification by using radial-based model of partial least squares to assist in the antibody's generation process. This modification allows transferring the sample from an original stage to a final stage by a linear regression process. The key to a good classification refers to the width of the center of the radial base. The authors conducted a series of tests and obtained good results (99.58%). However, depending on the value given to the radial-base width, the classification process behaves in a disorderly manner.

AIRS is a promising technique in the field of AI. Its design is inspired by the Biological Immune Systems. It aims to computationally reproduce its main characteristics, properties and abilities (Dasgupta, 1998). AIRS is a suitable tool for the diagnosis of diseases, a consequence of its ability to detect changes in behavior patterns (Castro & Timmis, 2002). AIRS

allows including new disease patterns in the diagnostic process without having to reboot the memory system, that is, it enables continuous learning, which means that the system can be more efficient as new patterns are available.

This article submits a method for diagnosing breast cancer based on artificial immune systems. From the samples of cancerous lesions, the negative selection algorithm (NSA) is implemented to differentiate between benign and non-benign samples (where there is evidence of malignancy). The samples classified as benign do not pose risks, i.e., they are not harmful to the organism. Those classified as non-benign are samples that need greater attention because they exhibit evidence of malignancy. The data analysis is performed by comparing the previously created detectors and samples, evaluating their affinity. If the affinity among the samples exceeds a predetermined threshold set by the professional, a match is found, and the sample is classified.

To evaluate the performance of the method, tests using a real database were performed, substantially explored in the literature. The database used was the Wisconsin Breast Cancer Diagnosis (WBCD) (Wbcd, 2021). This method afforded satisfactory results, with a high generalization and reliability capacity, and low computational effort. When compared against the other articles in the specialized literature, the results showed that the method is better than the other techniques. The method presented proved to be accurate, with an average success rate of 99.77% in the diagnoses performed.

## 2. Methodology

### 2.1 Wisconsin Breast Cancer Diagnosis (WBCD) Database

The WBCD database was created by William H. Wolberg, a doctor at the hospital of the University of Wisconsin-Madison, Wisconsin, United States (Wbcd, 2021). Between 1989 and 1991 Dr. Wolberg received several cases of breast tumors to be analyzed. In the analyses, the tumors were diagnosed as benign and malignant. Based on the referenced information, a database was assembled with 9 instances representing characteristics of the tumor and of course the classifications for these instances, which totaled 10 variables (Mangasarian et al., 1990; Wolberg & Mangasarian, 1990).

The characteristics stored in the database are as follows:

1. cell mass thickness (CT);
2. cell size uniformity (CS);
3. cell shape uniformity (CH);
4. marginal accession (AD);
5. epithelial cell size (EP);
6. empty nucleus (BN);
7. bland chromatin (CO);
8. normal nucleolus (NN);
9. mitosis (MM);
10. classification (“benign” or “malignant”);

This database has 699 samples, of which 65% represent benign tumors and 35% represent malignant tumors (Bennett & Mangasarian, 1992). Table 1 shows a small sample of the data contained in this base. In this problem, class 2 corresponds to a normal pattern (“benign”) and class 4 corresponds to an abnormal pattern (“malign”).

**Table 1:** Samples of Wisconsin Breast Cancer Diagnosis data.

| ID | CT | CS | CH | AD | EP | BN | CO | NN | MM | CLASS |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |
| 47 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | 2 |
| 69 | 5 | 1 | 3 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 93 | 2 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |
| 135 | 4 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 157 | 4 | 1 | 1 | 1 | 2 | 1 | 3 | 2 | 1 | 2 |
| 186 | 7 | 5 | 10 | 10 | 10 | 10 | 4 | 10 | 3 | 4 |
| 247 | 9 | 10 | 10 | 1 | 10 | 8 | 3 | 3 | 1 | 4 |
| 277 | 8 | 10 | 10 | 10 | 8 | 10 | 10 | 7 | 3 | 4 |
| 343 | 5 | 3 | 3 | 1 | 3 | 3 | 3 | 3 | 3 | 4 |
| 400 | 8 | 5 | 6 | 2 | 3 | 10 | 6 | 6 | 1 | 4 |
| 532 | 6 | 10 | 10 | 10 | 4 | 10 | 7 | 10 | 1 | 4 |

Source: Authors.

### 2.2 Biological immune system (BIS)

BIS is the body's main defense against various infection agents that invade the biological body. In this case, the immune system must act instantly, effectively responding against the invading pathogens, identifying them in order to protect the human body from the eminent threat. There are two types of responses: the response by the innate immune system and the response by the adaptive system.

The innate immune system is the first line of defense with a very fast response, characterized by the phagocytic cells (granulocytes, macrophages, etc.), responsible for ingesting foreign particles in the organism, and other types of defenses such as physical (skin) and chemical barriers. The adaptive immune system is in the second level, which is capable of recognizing microorganisms such as viruses, bacteria, fungi, protozoa, helminths and certain types of worms. The adaptive immune system is called immunological memory. It is capable of recording information from the infectious pathogens, at the initial detection, in order to expedite a response to this same type of agent that may reoccur in the future (Castro & Timmis, 2002).

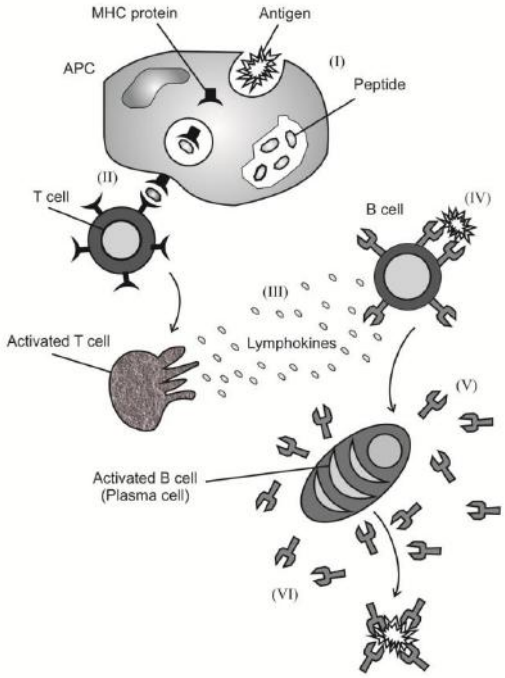
#### 2.2.1 Basic Defense Mechanisms

The biological body, especially human, is protected by various cells and molecules that work in harmony, directing responses to foreign substances to the body, the so-called antigens. Figure 1 shows this complex process in a simplified manner.

In step (I), the process is initiated when a pathogen (infectious agent) is ingested by a specialized antigen-presenting cell (APCs). At this stage, the pathogens are digested, fragmented in antigenic peptides. In step (II), fragments of peptides bind to MHC (major histocompatibility complex) and appear on the surface of the APC molecule. Next, in step (III), T cells, which have receptor molecules on their surface, are able to recognize different MCH/peptides antigens processed by the APC, that is, the recognition causes a status of activation. The third step is the self/nonself-discrimination carried out by the body, differentiating the cells of the infectious agents (Dasgupta, 1998; Castro & Timmis, 2002).

In step (IV), where the system is now activated by having recognized an MHC/peptide antigen, the T cells divide and secrete chemical signals (lymphokine), which signal to other components of the immune system that an antigen was found. After this signaling, in step (V) the B cells which have unique receptor molecules specificity on their surface, are able to recognize the free antigens in the body, without the need for ingestion and digestion of the presenting cells (APC), and are thereby activated. When activated, the B cells, in step (VI), divide and turn into plasmacytes, which secrete high levels of antibodies.

**Figure 1**: Simplified diagram of the recognition and activation mechanisms of BIS.




Source: Authors.

In step (VII), these generated antibodies bond to the antigens found. Thus, the pathogen is neutralized, hence destroying the threat. Some of the T and B cells transform into memory cells, which remain circulating in the system and ensuring a quick and efficient response to a future exposure of the same type of infectious pathogen strain. It should be noted that the entire process is carried out with the cooperation of the set of forming cells of the immune system, each one responsible for a relatively simple function, and the set overall performs an extremely complex process (Dasgupta, 1998; Castro & Timmis, 2002).

### 2.3 Negative Selection Algorithm (NSA)

NSA is a technique based on the pattern recognition process performed by the biological immune system, which is developed as a computational model. The ANS proposed by (Castro, 2001), for detecting changes in the systems, is based on the negative selection of T cells in the thymus. This process is the discrimination of the cells between self and non-self-performed by the organism. The algorithm is executed in two steps as shown below (Castro, 2001):

1. Censor

a) Define the set of chains (S) to be protected;

b) Generate random chains and evaluate the affinity (Match) between each of them and the chains themselves. If the affinity exceeds a threshold set, reject the chain. Otherwise, store them in a set of detectors (R).

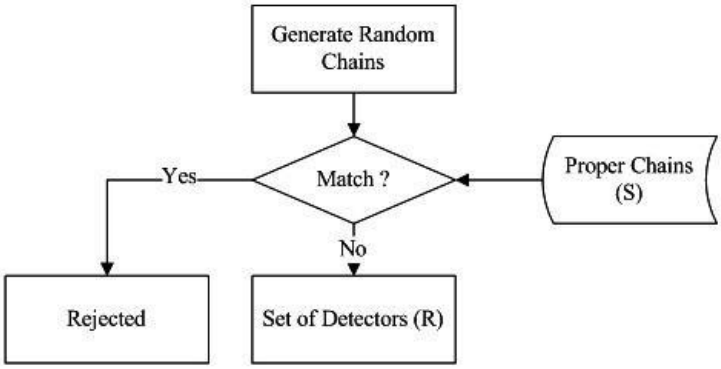
2. Monitoring

a) Given the set of chains to be protected (protected chains), evaluate the affinity between each of them and the set of detectors. If the affinity exceeds a predetermined threshold, then a non-self-element is identified.

Figures 2 and 3 illustrate the flowcharts of the censor and monitoring phase of the negative selection algorithm.

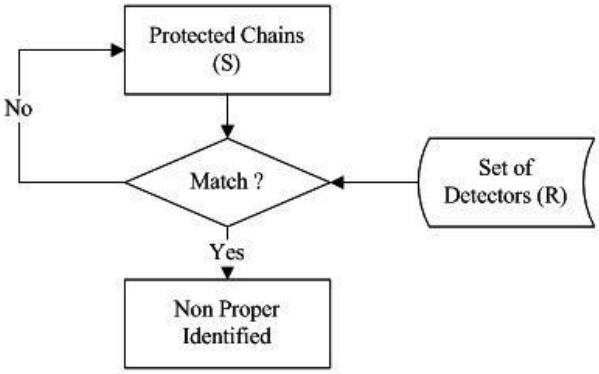
In the censor step of NSA, the detectors are initially defined, which represent a normal condition of the body, known as self-chains (S). The objective of this phase is to generate a set of detector patterns (R), with the ability to recognize any non-self-pattern, in the data monitoring stage. Then, with the reading of the data, the chains are randomly chosen and the affinity is verified by comparing these chains with the set of self-chains (S). Assuming that the affinity exceeds a predetermined threshold, the chain is rejected. Otherwise, this chain is accepted in the set of detectors (R), and will be used to conduct the classifications during the data monitoring. The detectors are similar to the matured T cells type capable of recognizing pathogens, which is to detect practically any not-self element, that is, a change or error in the data to be monitored (Lima et al., 2013).

**Figure 2:** Flowchart of NSA censor phase.




Source: Lima et al. (2013).

**Figure 3:** Flowchart of NSA Monitoring phase.




Source: Lima et al. (2013).

In the monitoring phase, the data is monitored in order to identify changes in the behavior of the samples and then classify these changes using the set of detectors created during the censor phase. Thus, by analyzing the protected chains (S) and comparing them with the set of detectors (R) the affinity between each of the chains is evaluated. If the affinity exceeds a certain threshold, then the non-self-element is detected and classified. It should be noted that the censor and monitoring phases are performed offline and in real-time, respectively (Lima et al., 2013).

Some observations can be made about the negative selection algorithm, highlighted as follows (Dasgupta, 2006; Forrest et al., 1997):

- the size of the set of detectors does not necessarily increase with the number of chains to be protected;

- the probability of detecting anomalies increases exponentially with the number of independent detection algorithms;
- detection is symmetric;
- There is an exponential cost to generate the detectors with regards to the size of the set of chains to be protected (self), because the random generation of chains of the R set results in the repeated generation of several chains;

### 2.4 Matching criterion

To evaluate the affinity between the chains and state that they are similar, a criterion known as a match is used. The match can be perfect or partial (Lima et al., 2013).

The perfect match is when the two chains under analysis have the same values n all positions, that is, the two chains must necessarily be equal. However, in the partial match there is no need for all positions of the chains to have the same value. In the partial match, only a number of positions between the chains must have the same value to affirm the match, with this amount being previously set. This quantity is known as the rate of affinity.

In this article, we chose to use the partial match proposed in (Bradley & Tyrrell, 2002), in which the rate of affinity represents the degree of similarity that must occur between the two chains under analysis for the match to be confirmed.

The rate of affinity is defined by the following equation (Bradley & Tyrrell, 2002):

$$Af = \left( \frac{An}{At} \right) * 100 \quad (1)$$

where:

$Af$ : affinity rate;

$An$ : number of normal samples;

$At$ : total number of samples.

### 2.5 Proposed method

The breast cancer diagnostic system presented in this section is based on the artificial immune systems, especially in the negative selection algorithm, which was presented in section 3 of this article.

The negative selection algorithm is a computational method that seeks to reproduce the negative selection process performed by matured T cells within the human body. In summary, this process performs a self/nonself-discrimination, which represents the main diagnosis mechanism of the human body, which is responsible for identifying unknown pathogens by the immune system (viruses, bacteria, fungi, etc.).

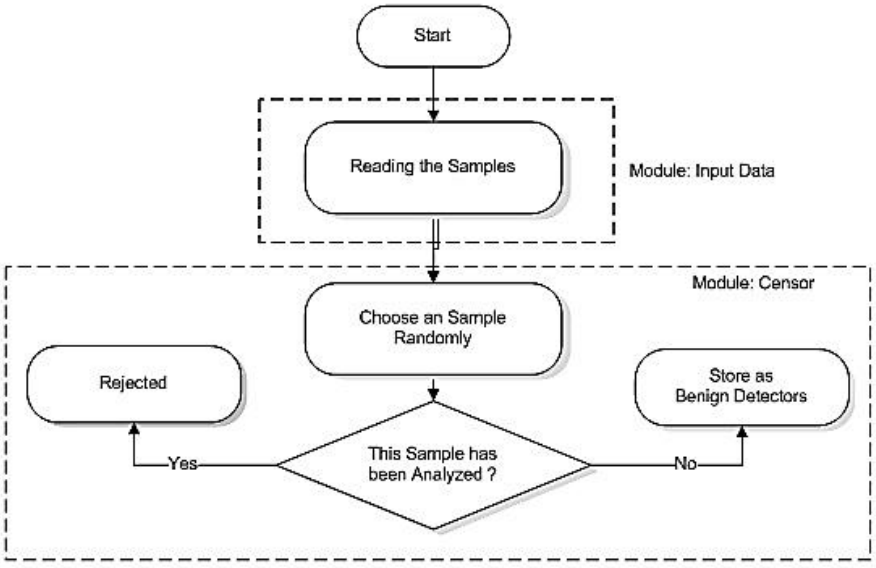
Based on this algorithm, the diagnostic system proposed in this paper performs the self/non-self-discrimination, a benign and non-self-tumor in which no evidence of malignancy was identified.

The proposed method consists of two phases: censor and data monitoring. In the censor phase, a data census is carried out in order to transfer knowledge, creating chains of detectors to identify abnormalities (non-self) in the monitoring process. In the monitoring phase the data is analyzed and compared with the detectors created during the censor phase, in order to present a diagnosis through the self/non-self-discrimination. The following are the censor and monitoring stages of the breast cancer diagnosis system.

### 2.6 Censor

At this stage, the detectors are generated to be used by the AIRS during the monitoring process. Due to the fact that the problem in question has two classes of patterns, the algorithm only needs to have knowledge of the normal samples (self), so that based on this information the discrimination is performed for what is self (benign) and what is non-self (malignant). Thus, the censor phase for this system has a single step. In this step the self-pattern is defined (benign) through the routine shown in Figure 4, which illustrates how the process to generate the self-detectors is performed.

**Figure 4:** Flowchart of NSA Censor phase.




Source: Authors.

In this article we use the partial match criterion proposed by (Bradley & Tyrrell, 2002). The rate of affinity is defined in equation (1). Using this equation, the affinity rate is estimated for this problem. This calculation is defined using equation (2), which has a total of 699 samples, of which 444 samples are normal, that is, without cancer:

$$Af = \left(\frac{444}{699}\right) * 100 = 63.52\% \quad (2)$$

The affinity ratio is of 63.52%, and that means that to confirm a match between two cancer samples there must be an affinity/similarity of at least 63.52% of the samples. The rate of casual affinity has a fixed value based on the calculation shown in (2), that is, it is based on the statistical sample. However, in this article the results with other rates of affinity values will be presented to provide a better assessment regarding the quality and performance of the system in the diagnostic process.

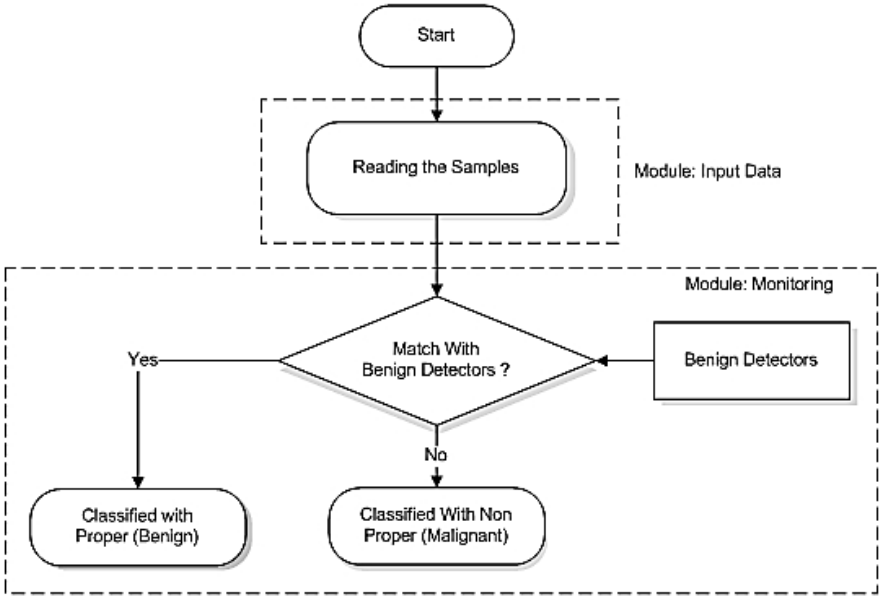
### 2.7 Monitoring

The monitoring phase is divided into two modules, which are responsible for the reading of the data and the self and nonself-discrimination. Figure 5 illustrates the flowchart of the data monitoring phase.

At this stage, the reading of the samples to be analyzed is performed. These samples are compared to the previously assigned benign detectors (self) in the censor phase. With the comparison of the sample using the set of detectors, a match between the sample and the set of detectors is verified. If the rate of affinity is satisfied, or if there is similarity between the samples analyzed and considered a match, then the sample is diagnosed as “benign” for having its own characteristics. Otherwise, there is no match among the samples being analyzed, thus the sample is diagnosed as “malignant” for not having characteristics known by the system, in other words, the sample is unknown. This process is repeated for each sample, and that

is how the data monitoring is performed.

**Figure 5:** Flowchart of NSA Monitoring Phase.




Source: Authors.

## 3. Results

In this section we present the results obtained with the proposed method implemented based on the tests. All tests were performed using a PC Intel Core 2 Duo 1.9 GHz, 2 GB of RAM, and operating system Windows 7 Ultimate 32-bit. The algorithm was developed in MATLAB (Matlab, 2011).

The test base used to demonstrate the performance of the proposed algorithm was the Wisconsin Breast Cancer Diagnosis Database (Wbcd, 2021) presented in section 2.

As previously reported, the WBCD base has 10 attributes. However, an improved number of attributes to perform the diagnosis was used in this article. To choose the attributes to be used, a selection process based on the calculation of the standard deviation of the samples was done. The five attributes that showed the lowest levels of standard deviation were chosen. When the standard deviation is low, it means that the data are more homogeneous. When the standard deviation is high, it means that the data is varied. This choice was used in order to provide the system greater reliability. The standard deviation was calculated by the following equation:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

The variables chosen that exhibit the lowest standard deviation are: the cell mass thickness (CT), cell size uniformity (CS), the cell shape uniformity (CH), the empty nucleolus (BN), and normal nucleolus (NN). During the separation of these variables, it was observed that some data are not usable. Table 2 presents the data and characteristics of the WBCD database.

**Table 2:** Data on the WBCD base.

| Database | UCI Wisconsin Breast Cancer Data |
|---|---|
| Type | Classification |
| Number of Data | 699 |
| Number of usable data | 683 |
| “benign” class data | 444 |
| “malignant” class data | 239 |
| Number of Attributes | 10 |

Source: Authors.

Several tests were carried out to evaluate the performance of the diagnostic system. Next, the tests and their results are described.

### 3.1 Test I

In test I, the goal is to execute the proposed method with unmodified normal parameters and settings.

For test I, the affinity rate with a fixed value calculated in (2) was used, which is of 63.52%. For this test, in the censor stage, three sets of detector patterns were generated and the test was performed on each one of them. Sets I, II and III have 5, 10 and 20 detector patterns, respectively. The detectors generated use 1.12%, 2.25% and 4.50% of the benign samples, which have a total of 444 samples. Table 3 shows the results for this test.

**Table 3:** Results for Test I.

| Diagnostic | Samples tested | Benign | Malign | Execution time (*ms*) | Accuracy (%) | Error (%) |
|---|---|---|---|---|---|---|
| *Detector set I* | 683 | 413 | 270 | 190,00 | 93,01% | 6,99% |
| *Detector set II* | 683 | 434 | 249 | 70,00 | 97,74% | 2,26% |
| *Detector set III* | 683 | 443 | 240 | 50,00 | 99,77% | 0,23% |

Source: Authors.

Test I did not show good performance (accuracy level above 93%) of the diagnostic system, or that the number of benign detectors directly influences the final diagnosis. Up to 30% of the information in the database is commonly used to generate detectors, and in this case the proposal is to use up to 5% in order to provide robustness to the diagnosis. As the number of detectors increases, the diagnosis becomes more accurate, and this is because the amount of knowledge provides higher efficiency to the diagnosis.

### 3.2 Test II

In test II, the objective is to verify the sensitivity of the proposed method with modifications in the rate of affinity.

For test II, four affinity rates were used, with values of 40%, 60%, 80% and 90%. For this test, the set of pattern detectors I was used, illustrated in Test I. Table 4 shows the results for this test.

**Table 4:** Results for Test II.

| Diagnostic | Samples tested | Benign | Malign | Execution time (millisecond) | Accuracy (%) | Error (%) |
|---|---|---|---|---|---|---|
| *Rate I* | 683 | 509 | 174 | 60,00 | 85,37% | 14,63% |
| *Rate II* | 683 | 444 | 239 | 70,00 | 100,00% | 0,00% |
| *Rate III* | 683 | 396 | 287 | 65,00 | 89,18% | 10,82% |
| *Rate IV* | 683 | 46 | 637 | 50,00 | 10,36% | 89,64% |

Source: Authors.

Test II shows that the rate of affinity of the samples directly influences the final diagnosis. A statistical calculation is performed using equation (1) by relating the number of normal and total samples of the database. When a variation is performed to a lower rate of affinity, a very good performance is observed, which may result in 100% of accuracy. However, as the rate of affinity decreases, the match criterion is less precise, resulting in non-self-samples to be classified as self. However, when a higher rate of affinity is executed, the performance greatly worsens, and this is because the match criterion is too rigorous, in order to validate samples that have almost a perfect match, which rarely happens. Given the reduced size of the set of detectors, many wrong diagnoses can occur. To solve this problem a greater number of benign detectors would have to be used, more than 30% of the information in the database.

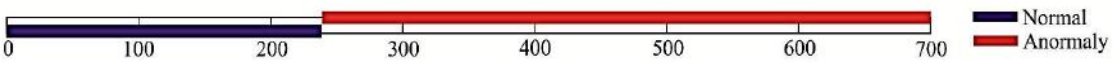
It should be noted that the rate of affinity must be precisely calculated for the system to perform well in the diagnostic process, and is commonly calculated as expressed in (1). But this requirement is not a prerequisite during the calculation, there is no need to use equation (1) to calculate the affinity rate. This rate can change – the problem is to find the value that provides accuracy to the diagnostic.

### 3.3 Test III

In test III, the objective is to verify the sensitivity of the proposed method in relation to the position of the samples in the set of tests.

For test III, the configuration that showed the best performance for safety and reliability during the diagnosis was used. The rate of affinity is set to a value of 63.52% and a set of detectors with 20 samples is used. Three sets of tests were prepared, and in the first one all the normal data is grouped, followed by the abnormal data. In the second test the normal data are grouped interspersed with the abnormal data, and in the third test the data are randomly grouped. Figures 6 and 7 illustrate the data arrangement for tests 1 and 2.

**Figure 6:** Data arrangement of test 1.




Source: Authors.

**Figure 7:** Data arrangement of test 2.


Source: Authors.

Table 5 shows the results for these tests.

**Table 5 –** Results for Test III.

| Diagnostic | Samples tested | Benign | Malign | Execution time (*ms*) | Accuracy (%) | Error (%) |
|---|---|---|---|---|---|---|
| *Test 1* | 683 | 443 | 240 | 70,00 | 99,77% | 0,23% |
| *Test 2* | 683 | 440 | 243 | 75,00 | 99,09% | 0,91% |
| *Random* | 683 | 441 | 242 | 50,00 | 99,32% | 0,68% |

Source: Authors.

Test III shows that the diagnostic system is very sensitive regarding the data arrangement. This does not affect its performance.

### 3.4 Analyses of results

The results obtained for the tests performed in this article are satisfactory (greater accuracy rate) and prove that the negative selection algorithm is effective in the diagnostic process. The parameters used, as well as the number of detectors, directly influence its performance. Thus, a trial test phase should be performed to find the correct parameters. It should be highlighted that for this article only 5% of the information samples was used to generate the benign detectors, which were incidentally generated with 30% of the information. Of a total of 444, the benign samples were randomly chosen, even samples to be defined as detectors, which is a very small amount of information. This shows that the method is robust and effective in the diagnostic process. The execution time is relatively small, which provides a rapid diagnosis.

Table 6 presents a comparative study between the proposed method and the main methods available in the specialized literature.

**Table 6 –** Comparative study.

| Reference | Database | Technique used | Accuracy (%) |
|---|---|---|---|
| (Wang, 2005) | WBCD | *Backpropagation* | 95,16% |
| (Whag & Lee, 2002) | WBCD | *ANFIS* | 96,30% |
| (Cmastra, 2006) | WBCD | *Kohonen* | 96,70% |
| (Meesad & Yen, 2003) | WBCD | *Fuzzy* | 96,71% |
| (Pena-Reyes & Sipper, 1999) | WBCD | *Fuzzy-genetic* | 97,07% |
| (Meesad & Yen, 2003) | WBCD | *ILFN and Fuzzy* | 98,13% |
| (Polat et al., 2007) | WBCD | *SAI-Fuzzy* | 98,51% |
| (Zhao; Davis, 2011) | WBCD | *AIRS and radial basis network* | 99,58% |
| This paper. | WBCD | *Negative Selection Algorithm* | 99,77% |

Source: Authors.

Table 6 shows that the proposed method, in this work, presents a higher success rate than other techniques.

## 4. Conclusion

This paper presents a breast cancer diagnostic method based on artificial immune systems, in particular the negative selection algorithm. The main stages and characteristics of the NSA and its application in the proposed problem were described. As the system's input data, the algorithm needs only five attributes of breast cancer samples. The proposed algorithm showed excellent results, achieving a success rate of 99.77% of accuracy in all samples tested.

The detector generation phase is the one that demands greater computational time, but as it is executed off-line, it does not compromise the algorithm. It should be noted that with a minimal amount of information, that is, with a greatly reduced set of detectors, the method was able to perform a diagnosis with considerable accuracy and safety. The monitoring phase of the

system, with the data reading, is performed rapidly with a time less than 100 milliseconds, which accredits the algorithm as a tool used in real time, given the urgency for quick decision making.

Therefore, it is concluded that artificial immune systems, based on the negative selection algorithm, obtained satisfactory performance in the tests conducted, and the method proved to be quite reliable, secure and robust for breast cancer diagnosis. Following this study, the focus will be on other strategies of immune systems, in order to render the application more competitive (efficiency, reliability and with greatly reduced processing time, among other requirements).

## References

Bennett, K. P., & Mangasarian, O. L. (1992). Robust Linear Programming Discrimination of Two Linearly Inseparable Sets, *Optimization Methods and Software, Gordon & Breach Science Publishers*, 23-34.

Bradley, D. W., & Tyrrell, A.M. (2002). Immunotronics - Novel Finite-State-Machine Architectures with Built-In Self-Test Using Self-Nonself Differentiation. *IEEE Trans. on Evolutionary Computation*. 6 (3), 227-238.

Camastra, F. (2006). Kernel Methods for Clustering. *Lecture Notes in Computer Science*. 3931, 1–9.

Castro, L. N. (2001). *Immune engineering: development and application of computational tools inspired by artificial immune systems*. PhD. Thesis. UNICAMP. Campinas, São Paulo, Brazil. (In Portuguese).

Castro, L. N. & Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer.

Dasgupta, D. (1998). *Artificial Immune Systems and Their Applications*, Springer, New York, USA.

Dasgupta, D. (2006). Advances in Artificial Immune Systems. *IEEE Computational Intelligence Magazine*, 40-49.

Forrest, S., Hofmeyr, S. A., & Somayaji, A. (1997). Computer Immunology. *Communications of the AC*. 88-96.

Forrest, S., Perelson, A., Allen, L., & Cherukuri, R. (1994). Self-Nonself Discrimination in a Computer, *Proc. of IEEE Symposium on Research in Security and Privacy*. 202-212.

Hamdi, R. El., Njah, M., & Chtourou, M. (2010). An Evolutionary Neuro-Fuzzy Approach to Breast Cancer Diagnosis. *IEEE International Conference on Systems, Man and Cybernetics*, 142-146.

INCA – National Institute of Cancer (Brazil). Available at: http://www.inca.gov.br.

Jung, J-S. R. (1993) ANFIS: Adaptive Network-Based Fuzzy Inference System. *IEEE Trans. on Systems, Man, and Cybernetics*. 23 (3), 665-685.

Karabatak, M., Ince, M. C., & Avci, E. (2008). An Expert System for Diagnosis Breast Cancer Based on Principal Component Analyses Method. *IEEE Proceedings on Communication and Applications Conference*, 1-4.

Lima, F. P. A., Lotufo, A. D. P., & Minussi, C. R. (2013). Artificial Immune Systems Applied to Voltage Disturbance Diagnosis in Distribution Electrical Systems, *PowerTech-2013*, Grenoble, France, 1-6.

Mangasarian O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis. *Large-scale Numerical Optimization*, 22-30.

Manikantan, K., Sayed, S.I., Syrigos, K.N., Rhys-Evans, P., Nutting, C.M., Harrington, K.J., & Kazi, R. Challenges for The Future Modifications of The TNM Staging System for Head and Neck Cancer: Case for a New Computational Model. *Cancer Treatment Reviews*, 35 (7), 639-644.

MATLAB 7.8 version, Mathworks Company.

Meesad, P. & Yen, G. G. (2003). Combined Numerical and Linguistic Knowledge Representation and Its Application to Medical Diagnosis. *IEEE Trans. on Systems, Man, and Cybernetics -Part A: Systems and Humans*. 33 (2), 206-222.

Naghibi, S. S., Teshnehlab, M. & Shoorehdeli, M. A. “Breast Cancer Detection by Using Hierarchical Fuzzy Neural System with EKF Trainer”, IEEE Proceedings of the 17th Iranian Conference of Biomedical Engineering - ICBME2010, November-2010, pp. 1-4.

OMS – World Health Organization. http://www.who.int/en/

Pena-Reyes, C. A., & Sipper, M. (1999). Designing Breast Cancer Diagnostic System Via Hybrid Fuzzy-Genetic Methodology. *IEEE International Fuzzy Systems Conference Proceeding*. 135-139.

Polat, K., Sahan, S., Kodaz, H. E., & Gunes, S. (2007). Breast Cancer and Liver Disorders Classifications Using Artificial Immune Recognition System (AIRS) With Performance Evaluation by Fuzzy Resource Allocation Mechanism, *Expert Systems with Applications*, 32 (1), 172–183.

Song, H-J, Lee, S-G., & Park, G-T. (2005). A Methodology of Computer Aided Diagnostic System on Breast Cancer. *Proceedings of the 2005 IEEE Conference on Control Applications Toronto*, 831-836.

Wang, J-S., & Lee, C.S.G. Self-Adaptive Neuro-Fuzzy Inference Systems for Classification Applications. *IEEE Transactions on Fuzzy Systems*, 10 (6), 790-802.

Wang, J.-Y. (2005). Data Mining Analysis (Breast-Cancer Data). http://www.csie.ntu.edu.tw/ p88012/AI-final.pdf.

WBCD – Wisconsin Breast Cancer Data – UCI Machine Learning Repository. www.arquives.ics.uci.edu/ml/

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proceedings of the National Academy of Sciences of USA*, 87 (23), 9193-9196.

Zhao, W. & Davis, C. E. (2011). A Modified Artificial Immune System Based Pattern Recognition Approach – An Application to Clinical Diagnostics. *Artificial Intelligence in Medicine*. 52 (1), 1-9.