

ASN-MAMA: aplicativo para diagnóstico clínico de câncer de mama utilizando sistemas imunológicos artificiais

ASN-MAMA: application for clinical breast cancer diagnosis using artificial immune systems

DOI:10.34117/bjdv8n4-631

Recebimento dos originais: 21/02/2022

Aceitação para publicação: 31/03/2022

Fernando Parra dos Anjos Lima

Doutor em Engenharia Elétrica pela Universidade Estadual Paulista
Instituição: Instituto Federal do Mato Grosso, Campus Avançado Tangará da Serra
Endereço: Rua 28, 980N – Vila Horizonte, Tangará da Serra – MT, Brasil
E-mail: fernando.lima@ifmt.edu.br

Simone Silva Frutuoso de Souza

Doutora em Engenharia Elétrica pela Universidade Estadual Paulista
Instituição: Universidade do Estado de Mato Grosso, Campus Tangará da Serra
Endereço: Rodovia MT 358, Km 07 (s/n) - Jardim Aeroporto, Tangará da Serra – MT
Brasil
E-mail: simonefrutuoso.mat@gmail.com

Fábio Roberto Chavarette

Doutor em Engenharia Mecânica pela Universidade Estadual de Campinas
Instituição: Universidade Estadual Paulista, Instituto de Química de Araraquara
Departamento de Engenharia, Física e Matemática
Endereço: Prof. Francisco Degni, 55 - Quitandinha, Araraquara – SP, Brasil
E-mail: fabio.chavarette@unesp.br

RESUMO

Neste artigo apresenta-se um aplicativo para diagnóstico clínico de amostras de câncer de mama, utilizando uma abordagem baseada nos sistemas imunológicos artificiais. Tomando-se como base um processo imunológico, utiliza-se o Algoritmo de Seleção Negativa para discriminar as amostras, obtendo uma classificação em casos benignos ou malignos. A principal aplicação deste sistema é auxiliar profissionais no processo de diagnóstico de câncer de mama em ambiente hospitalar, proporcionando rapidez na tomada de decisão, eficiência no planejamento de tratamentos, confiabilidade e a assistência necessária para salvar vidas. O aplicativo também pode ser utilizado para treinamento de novos profissionais. Para calibrar e validar este aplicativo utilizou-se a base de dados Wisconsin Breast Cancer Diagnosis, trata-se de uma base de dados real de câncer de mama. O Aplicativo foi desenvolvido na linguagem C++, em modo visual, apresentando uma interface prática e de fácil utilização. Foram realizados testes com o sistema, e os resultados foram comparados com a literatura especializada, apresentando bom desempenho, precisão, robustez e eficiência no processo de diagnóstico de câncer de mama.

Palavras-chave: diagnóstico de câncer de mama, sistemas imunológicos artificiais, algoritmo de seleção negativa.

ABSTRACT

This paper presents an application for clinical diagnosis of breast cancer samples, using an approach based on artificial immune systems. Based on an immunological process, the Negative Selection Algorithm is used to discriminate the samples, obtaining a classification into benign or malignant cases. The main application of this system is to help professionals in the process of breast cancer diagnosis in a hospital environment, providing fast decision making, efficient treatment planning, reliability and the necessary assistance to save lives. The application can also be used for training new professionals. To calibrate and validate this application we used the Wisconsin Breast Cancer Diagnosis database, a real breast cancer database. The application was developed in C++ language, in visual mode, presenting a practical and easy-to-use interface. Tests were performed with the system, and the results were compared with the specialized literature, presenting good performance, precision, robustness and efficiency in the process of breast cancer diagnosis.

Keywords: breast cancer diagnosis, artificial immune systems, negative selection algorithm.

1 INTRODUÇÃO

O câncer é uma doença crônica que atinge milhões de pessoas em todo o mundo. Com o envelhecimento da população e maior expectativa de vida, observa-se um grande aumento de casos de câncer, em especial o câncer de mama, que é um dos tipos de câncer que tem maior ocorrência. Segundo o Instituto Nacional de Câncer (INCA), em um levantamento realizado no início do ano de 2012, evidenciou-se que o câncer de mama representa 27,9% dos casos ocorridos nas mulheres no ano de 2011, ou seja, entre todos os tipos de cânceres existentes, o que mais atinge as mulheres é o câncer de mama [11].

No mundo todo, a taxa de mortalidade por câncer de mama é muito elevada, provavelmente porque a doença é diagnosticada em estágios avançados. Através de estatísticas realizadas pela Organização Mundial de Saúde (OMS), observa-se que aproximadamente 39% das mulheres que lutam contra o câncer de mama, vão a óbito [18].

Para reduzir estas taxas, torna-se importantíssimo realizar campanhas de conscientização, prevenção e, principalmente, buscar realizar o diagnóstico do câncer de mama em estágio inicial. O diagnóstico correto em um estado prematuro da doença, proporciona decisões rápidas no planejamento de ações e, evidentemente, a eficiência no tratamento. Porém, diagnosticar corretamente o câncer é um processo complexo e muito difícil, pois envolve muitas variáveis. Para um diagnóstico correto é necessária muita experiência por parte do profissional e, principalmente, que a classificação do

estadiamento clínico do tumor (estágio do câncer) esteja certa.

Para auxiliar os profissionais no diagnóstico existem sistemas computacionais, que a partir de receber informações da amostra em análise, realiza o diagnóstico, facilitando e proporcionando maior segurança para este processo. No entanto os tradicionais sistemas de classificação utilizados são detalhados e complexos, e normalmente de difícil utilização, oferecendo limitações aos patologistas, não permitindo rapidez na tomada de decisão [14]. Assim, torna-se necessário o desenvolvimento de sistemas integrados que tenham capacidade de trabalhar com técnicas de processamento e análise de dados, com interface visual e de fácil utilização, e que, combinados com a experiência dos profissionais, proporcione a assistência necessária para realizar o diagnóstico e planejar o tratamento.

Neste sentido, o desenvolvimento de um sistema computacional com interface gráfica, em conjunto com técnicas de Inteligência Artificial (IA) se torna uma possível solução para o problema de diagnóstico. Sistemas com interface gráfica podem proporcionar facilidade, iteratividade, e um ambiente intuitivo para o usuário final. Os métodos inteligentes são capazes de extrair informações e conhecimento de problemas complexos, e são de fácil aplicação. Existem diversos métodos baseados neste conceito, que estão sendo utilizados para auxiliar profissionais a realizar o diagnóstico de doenças, em especial para profissionais com pouca experiência, estes métodos proporcionam segurança, confiabilidade e rapidez do diagnóstico.

Na literatura especializada destacam-se alguns artigos que visam o diagnóstico de câncer de mama. Na referência [19], os autores apresentam um método para diagnóstico de amostras de câncer utilizando a lógica fuzzy e o algoritmo genético. Em [22] foi usada uma rede ANFIS [12] para realizar o diagnóstico de câncer de mama, os resultados foram considerados satisfatórios (constar o índice de acerto). No artigo [16] é proposto um sistema híbrido que utiliza uma rede neural em conjunto com um sistema especialista fuzzy. O sistema teve um bom desempenho quando aplicado na base de dados Wisconsin [24]. Em [21] é apresentado um método utilizando uma rede Neuro-Fuzzy (ANFIS) para realizar o diagnóstico de amostras de câncer de mama. Na referência [20] é apresentando um sistema de reconhecimento imunológico artificial (AIRS) para classificação de amostras de câncer de mama. Este método é baseado no sistema de metáforas imunológico, e tenta reproduzir mecanismos inspirados pela imunologia, como a concorrência de recursos, a seleção clonal, a maturação de afinidade e a formação de células de memória. O mecanismo de alocação de recursos é baseado na lógica fuzzy. Um

sistema especialista para o diagnóstico de câncer de mama foi proposto em [13]. No artigo [17], os autores propõem uma nova abordagem utilizando uma rede neural fuzzy hierárquica para o diagnóstico. Já em [10] é apresentado um modelo Fuzzy TSK-type para o diagnóstico do câncer de mama. Em [27] os autores apresentam um método baseado no algoritmo de seleção clonal, onde é adicionada uma modificação através da utilização de um modelo de base radial do tipo mínimos quadráticos parciais para auxiliar no processo de geração de anticorpos. Esta modificação permite realizar a transferência da amostra de um estado original para um estado final através de um processo de regressão linear. A chave para uma boa classificação refere-se à largura do centro da base radial. Os autores fizeram uma série de testes e obtiveram bons resultados, com um índice de acerto de 99,58%. Porém, dependendo do valor dado à largura da base radial o processo de classificação se comporta de maneira desordenada.

O SIA é uma técnica promissora no campo da AI. Sua concepção é inspirada nos Sistemas Imunológicos Biológicos. Visa reproduzir, computacionalmente, suas principais características, propriedades e habilidades [5]. O SIA constitui-se numa ferramenta adequada para a realização do diagnóstico de doenças, como consequência da sua habilidade de detecção de mudanças de comportamento em padrões [7]. O SIA permite incluir novos padrões de doenças no processo de diagnóstico, sem a necessidade de reiniciar a memória do sistema, isto é, permite uma aprendizagem contínua, ou seja, o sistema pode ser tornar mais eficiente à medida que novos padrões sejam disponibilizados.

Neste artigo, apresenta-se um aplicativo comercial para diagnóstico de câncer de mama baseado nos sistemas imunológicos artificiais. A partir das amostras de lesões cancerígenas, aplica-se o algoritmo de seleção negativa (ASN) para diferenciar as amostras entre próprias (benignos) e não-próprias (onde há evidência de malignidade). As amostras classificadas como próprias não representam riscos, isto é, não são prejudiciais ao organismo. As classificadas como não-próprias são amostras que necessitam de uma maior atenção, pois apresentam evidências de malignidade. A análise dos dados é realizada através da comparação entre os detectores previamente criados e as amostras, avaliando a afinidade entre eles. Caso a afinidade entre as amostras ultrapasse um limiar preestabelecido pelo profissional, é encontrado um casamento, e a amostra é classificada.

Para avaliar o desempenho do sistema, foram realizados testes com uma base de dados real bem explorada na literatura. Sendo a base de dados Wisconsin Breast Cancer

Diagnosis (WBCD) [24]. Este sistema apresenta uma interface bem simples e de fácil utilização, e proporcionou resultados satisfatórios, com uma capacidade de generalização elevada, confiabilidade e baixo esforço computacional.

2 BASE DE DADOS WISCONSIN BREAST CANCER DIAGNOSIS (WBCD)

A base de dados WBCD foi criada pelo Dr. William H. Wolberg. Trata-se de um médico dos hospitais da Universidade de Wisconsin Madison, na cidade de Wisconsin nos Estados Unidos [24]. O Dr. Wolberg, em seu artigo, entre os anos de 1989 e 1991, recebeu diversos casos de tumores na mama para serem analisados. Nas análises realizadas, os tumores foram diagnosticados em benignos e malignos. Com referência a estas informações, foi montada uma base de dados com 9 instâncias representando as características do tumor e, evidentemente, a classificação para estas instâncias, totalizando 10 variáveis [15], [25].

As características armazenadas na base de dados são as seguintes: espessura da massa celular (CT); uniformidade do tamanho da célula (CS); uniformidade do formato da célula (CH); adesão marginal (AD); tamanho de uma célula epitelial (EP); núcleo vazio (BN); cromatina branda (CO); nucléolo normal (NN); mitose (MM); classificação (“benigno” ou “maligno”);

Esta base de dados possui 699 amostras, sendo que 65% representam tumores benignos e 35% representam tumores malignos [1]. Na tabela 1, apresenta-se uma pequena amostra dos dados contidos nesta base. Neste problema, a classe 2 corresponde a um padrão normal (“benigno”) e a classe 4 corresponde a um padrão anormal (“maligno”).

A base WBCD [26] possui 10 atributos. No entanto, para este artigo optou-se por utilizar uma quantidade melhor de atributos para realizar o diagnóstico. Para escolher quais atributos utilizar foi feito um processo de escolha com base no cálculo do desvio padrão das amostras. Foram escolhidos os cinco atributos que apresentaram os índices mais baixos de desvio padrão. Quando o desvio padrão é baixo, significa que os dados são mais homogêneos. Quando o desvio padrão é alto significa que há dados variados. Foi adotada esta escolha visando proporcionar maior confiabilidade para o sistema.

Tabela 1 – Amostra dos dados da Wisconsin Breast Cancer Diagnosis [24].

ID	CT	CS	CH	AD	EP	BN	CO	NN	MM	CLASSE
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
21	7	3	2	10	5	10	5	4	4	4
52	5	5	5	8	10	8	7	3	7	4
99	10	3	5	1	10	5	3	10	2	4
162	10	8	10	10	6	1	3	1	10	4
186	7	5	10	10	10	10	4	10	3	4

O desvio padrão foi calculado pela seguinte equação:

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

As variáveis escolhidas que apresentam o desvio padrão mais baixo são: a espessura da massa celular (CT), a uniformidade do tamanho da célula (CS), a uniformidade do formato da célula (CH), o núcleo vazio (BN) e o nucléolo normal (NN). Durante a separação destas variáveis observou-se que alguns dados não são aproveitáveis. Na tabela 2 são apresentados os dados e características da base de dados WBCD.

Tabela 2 – Dados sobre a base WBCD.

Base de dados	UCI Wisconsin Breast Cancer Data
Tipo	Classificação
Número de dados	699
Número de dados aproveitáveis	683
Dados da classe “benigno”	444
Dados da classe “maligno”	239
Número de atributos	10

3 SISTEMA IMUNOLÓGICO BIOLÓGICO (SIB)

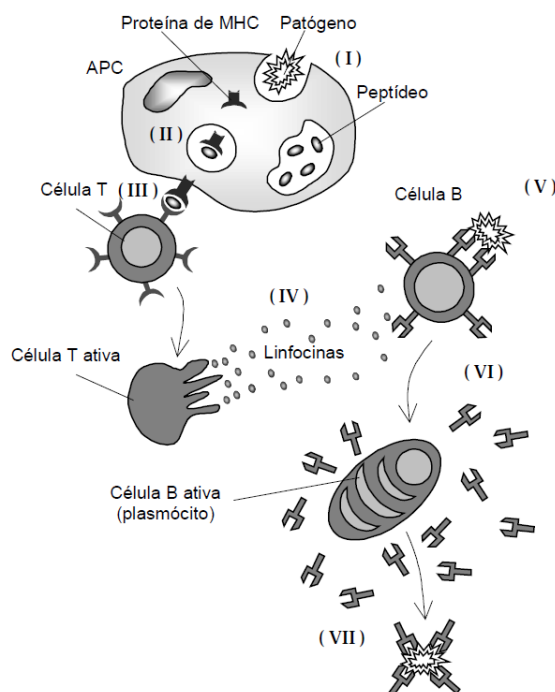
O SIB é a principal defesa do organismo contra diversos agentes infecções que invadem o corpo biológico. Neste caso, o sistema imunológico deve agir instantaneamente respondendo, de forma efetiva, contra os patógenos invasores e identificá-los, visando proteger o corpo humano da ameaça. Existem dois tipos de respostas: a resposta efetuada pelo sistema imunológico inato e a resposta efetuada pelo sistema adaptativo.

O sistema imunológico inato é a primeira linha de defesa com resposta bem rápida, sendo caracterizado por células fagocitárias (Granulócitos, Macrófagos, etc.), responsáveis pela ingestão de partículas estranhas ao organismo, e por outros tipos de defesas como, barreiras físicas (pele) e químicas. Já o sistema imunológico adaptativo está em segundo nível, sendo capaz de reconhecer microrganismos como vírus, bactérias, fungos, protozoários, helmintos e alguns tipos de vermes. O sistema imunológico adaptativo é a chamada memória imunológica. Ela é capaz de gravar informações dos agentes patógenos infecciosos, em uma primeira detecção, a fim de acelerar uma resposta a este mesmo tipo de agente que possa ocorrer futuramente [7], [28], [29].

3.1 MECANISMOS BÁSICOS DE DEFESA

O corpo biológico, em especial, o humano é protegido por diversas células e moléculas que trabalham em harmonia, objetivando respostas a substâncias estranhas ao organismo, os chamados antígenos. Todo este processo complexo está apresentado de forma simplificada na Figura 1.

Figura 1 – Mecanismos de reconhecimento e ativação do SIB



No passo (I), o processo é iniciado quando algum patógeno (agente infeccioso) é ingerido por uma molécula apresentadora de antígeno especializado (APCs). Nesta fase, os patógenos são digeridos, fragmentados em peptídeos antigênicos. No passo (II), os

fragmentos de peptídeos se ligam a moléculas MHC (*major histocompatibility complex*) e são apresentados à superfície da molécula APC. Em seguida, no passo (III), as células T, as quais possuem moléculas receptoras em sua superfície, são capazes de reconhecer diferentes antígenos MHC/peptídeos processados pela APC, ou seja, o reconhecimento faz com que o estado seja de ativação. O terceiro passo representa a discriminação próprio/não-próprio realizada pelo organismo (Algoritmo de Seleção Negativa), diferenciando as células próprias dos agentes infecciosos [7], [5].

No passo (IV), onde o sistema já está ativado por ter reconhecido algum antígeno MHC/peptídeo, as células T se dividem e secretam sinais químicos (linfocina) que sinalizam para outros componentes do sistema imunológico que um antígeno foi encontrado. Após esta sinalização, no passo (V) as células B, que possuem moléculas receptoras de especificidade única em sua superfície, são capazes de reconhecer os antígenos livres no organismo, sem a necessidade da ingestão e digestão das células apresentadoras (APC), e assim são ativadas. Quando ativadas, as células B, no passo (VI), elas se dividem e se transformam em plasmócitos, secretando anticorpos em altas taxas. No passo (VII), estes anticorpos gerados são ligados aos antígenos encontrados. Assim, o patógeno é neutralizado, levando a destruição da ameaça. Algumas das células T e B se transformam em células de memória, e continuam circulando no sistema garantindo uma resposta eficiente e rápida a uma futura exposição ao mesmo tipo de patógeno infeccioso. Note que todo o processo é realizado com a cooperação entre o conjunto de células formadoras do sistema imunológico, sendo cada uma responsável por uma função relativamente simples, e, no conjunto, realiza um trabalho extremamente complexo [7], [5].

3.2 ALGORITMO DE SELEÇÃO NEGATIVA (ASN)

O ASN [8] é uma técnica que se baseiam no processo de reconhecimento de padrões exercido pelo sistema imunológico biológico, sendo elaborado como um modelo computacional. O ANS proposto por [8], para detecção de mudanças em sistemas, é baseado na seleção negativa de linfócitos T dentro do timo. Este processo representa a discriminação das células, entre próprias e não-próprias realizado pelo organismo. O algoritmo é executado em duas fases, como é mostrado a seguir [7], [30], [31]:

1. Censoriamento
 - a) Definir o conjunto de cadeias próprias (S) que se deseja proteger;

b) Gerar cadeias aleatórias e avaliar a afinidade (Match) entre cada uma delas e as cadeias próprias. Caso a afinidade seja superior a um limiar estipulado, rejeitar a cadeia. Caso contrário, armazene-a em um conjunto de detectores (R).

2. Monitoramento

a) Dado o conjunto de cadeias que se deseja proteger (cadeias protegidas), avaliar a afinidade entre cada uma delas e o conjunto de detectores. Se a afinidade for superior a um limiar preestabelecido, então um elemento não-próprio é identificado.

As Figuras 2 e 3 ilustram os fluxogramas da fase de sensoriamento e monitoramento do algoritmo de seleção negativa.

Na fase de sensoriamento do ASN, são definidos, inicialmente, os de detectores próprios, que representam uma condição normal do organismo, sendo conhecidos como cadeias próprias (S). O objetivo desta fase é gerar um conjunto de padrões detectores (R), que tenham a capacidade de reconhecer algum padrão não-próprio, na fase de monitoramento dos dados. Então, a partir da leitura dos dados, escolhem-se cadeias aleatoriamente e verifica-se a afinidade comparando estas cadeias ao conjunto de cadeias próprias (S). Supondo que a afinidade seja superior a um limiar preestabelecido, rejeita-se a cadeia. Caso contrário, está cadeia é aceita no conjunto de detectores (R), e será utilizada para fazer as classificações durante o monitoramento dos dados. Os detectores são análogos às células do tipo T maturadas capazes de reconhecer agentes patogênicos, isto é, detectar praticamente qualquer elemento não-próprio, uma modificação ou erro nos dados que se deseja monitorar.

Figura 2 – Fluxograma da fase de Censoriamento do ASN.

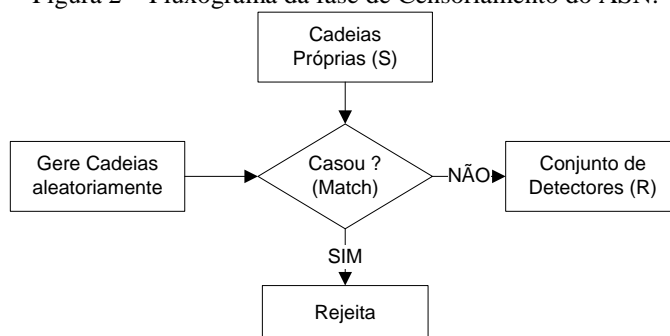
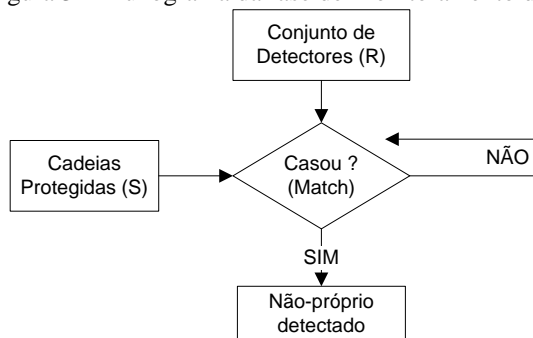


Figura 3 – Fluxograma da fase de Monitoramento do ASN.



Na fase de monitoramento, faz-se um monitoramento nos dados visando identificar mudanças no comportamento das amostras e, então, classificar estas mudanças utilizando o conjunto de detectores criados na fase de sensoriamento. Assim, analisando-se as cadeias protegidas (S) e comparando-as com o conjunto de detectores (R), avalia-se a afinidade entre cada uma das cadeias. Caso a afinidade seja superior a um determinado limiar, então, o elemento não-próprio é detectado e classificado. Ressalta-se que as fases de sensoriamento e de monitoramento são realizadas de modo *off-line* e em tempo real, respectivamente [7].

3.3 CRITÉRIO DE CASAMENTO

Para avaliar a afinidade entre as cadeias e afirmar que elas são semelhantes, utiliza-se um critério conhecido como casamento. O casamento pode ser perfeito ou parcial.

O casamento perfeito é quando as duas cadeias, que estão sendo analisadas, têm os mesmos valores em todas as posições, ou seja, as duas cadeias necessariamente devem ser iguais. Já no casamento parcial não existe a necessidade de que todas as posições das cadeias tenham o mesmo valor. No casamento parcial, apenas uma quantidade de posições entre as cadeias deve ter o mesmo valor para afirmar o casamento, sendo esta quantidade definida previamente. Esta quantidade é conhecida como a taxa de afinidade.

Neste artigo, optou-se por utilizar o casamento parcial proposto em [2], onde a taxa de afinidade representa o grau de semelhança que deve ocorrer entre as duas cadeias em análise para que o casamento seja confirmado.

A taxa de afinidade é definida através da seguinte equação:

$$Af = \left(\frac{An}{At} \right) * 100 \quad 2)$$

sendo:

Af : taxa de afinidade;

An : número de amostras normais;

At : número total de amostras.

4 MÉTODO PROPOSTO

The hybrid system for diagnosis of structural faults proposed in this paper consists of an offline phase (learning) and an online phase (monitoring), as shown in figure 4. In the offline phase is the training process of ANNs and get the knowledge to be used in the monitoring phase for detection and fault classification process. Also, there is the phase of acquiring and processing data. After performing signal acquisition using the wavelet transform to decompose the signals into 3 levels of resolution in the wavelet domain.

O sistema de diagnóstico de câncer de mama apresentado nesta seção é baseado nos sistemas imunológicos artificiais, em especial no algoritmo de seleção negativa, que foi apresentado na seção 3 deste artigo.

O algoritmo de seleção negativa é um método computacional que visa reproduzir o processo de seleção negativa realizado pelas células T maturadas dentro do organismo humano. Em resumo, este processo faz a discriminação próprio/não-próprio, que representa o principal mecanismo de diagnóstico do corpo humano, mecanismo o qual é responsável por identificar agentes patógenos desconhecidos pelo sistema imunológico (vírus, bactérias, fungos, etc.).

Tendo como base este algoritmo, o sistema de diagnóstico proposto neste artigo realizará a discriminação próprio/não-próprio, sendo próprio um tumor benigno e não-próprio onde foi identificada evidência de malignidade.

O método proposto consiste-se em duas fases: sensoriamento e o monitoramento dos dados. Na fase de sensoriamento, realiza-se um censo nos dados, a fim de passar conhecimento, criando cadeias de detectores para identificação de anormalidades (não-próprio) no processo de monitoramento. Na fase de monitoramento os dados são analisados, sendo comparados com os detectores criados na fase de sensoriamento, visando apresentar um diagnóstico através da discriminação próprio/não-próprio. A seguir apresentam-se as fases de sensoriamento e de monitoramento do sistema de diagnóstico de câncer de mama.

4.1 CENSORIAMENTO

Nesta fase, são gerados os detectores para serem utilizados pelo SIA durante o processo de monitoramento. Por causa do problema em questão ter duas classes de padrões, é necessário somente que o algoritmo tenha conhecimento das amostras normais (próprios), para com base nestas informações realizar a discriminação do que é próprio (benigno) e do que é não-próprio (maligno). Sendo assim a fase de censoriamento é apresentada nos passos a seguir:

1º Passo: Realizar a leitura de todas as amostras próprias (Classe “Benigna”);

2º Passo: Defina a quantidade de amostras que vão ser escolhidas como detectores;

2º Passo: Inicie um ciclo de repetição, e escolha a cada iteração uma Amostra Aleatoriamente. O ciclo termina quando o número de detectores definido pelo usuário for armazenado;

3º Passo: Verifique a cada iteração se a amostra escolhida já foi analisada, caso já tenha sido analisada rejeite a amostra, caso contrário, armazene como um detector benigno;

Neste artigo utiliza-se o critério de casamento parcial proposto por [2]. A taxa de afinidade é definida na equação (2). Porém para este sistema o usuário final é quem define a taxa de afinidade a ser utilizada no processo de diagnóstico. Esta escolha pode influenciar diretamente na precisão do diagnóstico. Partindo dos dados originais da base de dados, calcula-se a taxa de afinidade para este problema. Este cálculo é definido usando-se a equação (3), onde se tem um total de 699 amostras, sendo 444 amostras normais, isto é, sem câncer:

$$Af = \left(\frac{444}{699} \right) * 100 = 63,52\% \quad 3)$$

O valor da taxa de afinidade é de 63,52%, isto significa que para confirmar um casamento entre duas amostras de câncer é necessário que exista uma afinidade/semelhança de no mínimo 63,52% entre as amostras analisadas. Este valor encontrado com o cálculo representa a zona de segurança do sistema, isto é, se a taxa de afinidade tiver valores próximos ao valor encontrado pelo cálculo, o sistema terá melhor desempenho.

4.2 MONITORAMENTO

A fase de monitoramento é dividida em dois módulos, os quais são responsáveis por fazer a leitura dos dados e realizar a discriminação próprio/não-próprio. Esta fase pode ser descrita nos passos a seguir:

1º Passo: Realizar a leitura das amostras que se deseja monitorar;

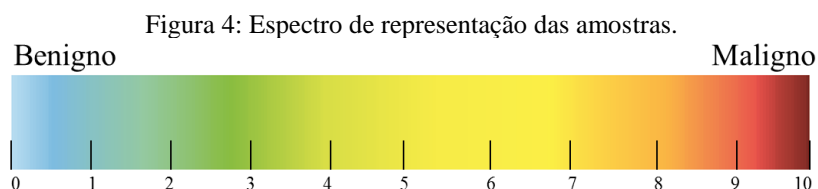
2º Passo: Verificar se existe casamento entre a amostra e o conjunto de detectores benignos (definidos na fase de sensoriamento). O casamento representa a semelhança entre os valores das variáveis, isto é, se elas têm os mesmos valores, elas se casam, garantindo uma porcentagem de afinidade;

3º Passo: Caso exista afinidade entre a amostra e o conjunto de detectores, classifica-se a amostra como próprio (classe “benigno”), caso contrário, classifica-se a amostra como não-próprio (classe “maligno”);

Nesta fase, faz-se a leitura das amostras que se deseja analisar. Essas amostras são comparadas aos detectores benignos (próprios) atribuídos previamente na fase de sensoriamento. Realizando a comparação da amostra com o conjunto de detectores, verifica-se um casamento entre a amostra e o conjunto de detectores. Caso a taxa de afinidade seja satisfeita, ou seja, se existir semelhança entre as amostras analisadas é considerado um casamento e, então, a amostra é diagnosticada como “benigna”, por ter características próprias. Caso contrário, não existe o casamento entre as amostras em análise, assim a amostra é diagnosticada como “maligna”, por não ter características conhecidas pelo sistema, isto é, a amostra é desconhecida. Este processo é repetido em cada amostra, e assim é feito o monitoramento dos dados.

4.3 REPRESENTAÇÃO VISUAL DAS AMOSTRAS DE CÂNCER

Para representar visualmente as amostras de câncer foi proposto um espectrograma de cores (espectro de calor), como apresentado na Figura 4.



Para representar a amostra visualmente, foram utilizados os valores dos atributos, onde foi calculada a média aritmética, obtendo um valor entre o intervalo [0,10]. Para exemplificar será utilizada uma amostra, como apresentado a seguir:

ID	CT	CS	CH	BN	NN	CLASSE
30	3	1	1	1	1	2

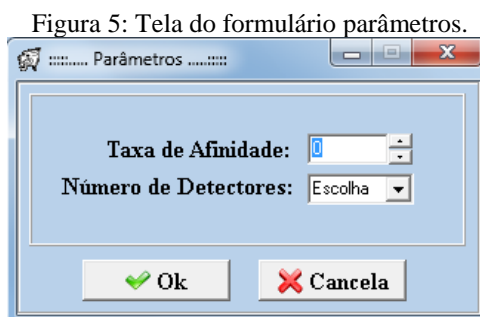
Como exemplo foi escolhido à amostra 30 da base de dados WBCD [WBCD]. Calculando a média dos atributos desta amostra temos:

$$Média = (3 + 1 + 1 + 1 + 1) / 5 = 1,4 \quad 4)$$

Portanto o valor da média para esta amostra representa visualmente uma coloração entre o intervalo [1,2], isto é, o valor da média é de 1,4 e representa uma cor dentro do 2º intervalo de cores no espectro. Vale ressaltar que visivelmente pode-se confirmar que a classe desta amostra é benigna, como apresentado nos atributos, ou seja, a tonalidade de cor, facilita ao usuário visualizar em que escala de espectro o câncer está proporcionando um diagnóstico de forma implícita visualmente.

4.4 ASN-MAMA: SISTEMA DE DIAGNÓSTICO DE CÂNCER DE MAMA

O sistema foi desenvolvido na linguagem C++ 6.0 [4]. Este sistema tem 4 processos principais, que são: a definição dos parâmetros (figura 5), o processamento do ASN, que é a fase de calibração do sistema, (figuras 6 e 7), o relatório do processo de calibração (figura 8) e o diagnóstico (figura 9).



No formulário parâmetros (figura 5) o usuário define os valores da taxa de afinidade e a quantidade de padrões detectores que vão ser utilizados no sistema. A taxa de afinidade pode variar de 0 a 100%, não necessariamente sendo valores inteiros, já o

número de detectores é escolhido pelo usuário em uma das opções apresentadas no campo, sendo as opções de 10, 20 e 30 detectores.

Após configurar os parâmetros do sistema é necessário realizar um processo de calibração, o qual é composto pela fase de sensoriamento e monitoramento do algoritmo de seleção negativa. Neste processo é utilizada a base de dados WBCD [24] para gerar um conjunto de detectores, e verificar qual o desempenho quando aplicado a amostras reais desta mesma base, ou seja, na fase de sensoriamento é definido um conjunto de detectores e na fase de monitoramento o mesmo é aplicado no diagnóstico das amostras da base WBCD, e ao final deste processo obtemos o índice de precisão (acerto) para o conjunto de detectores no processo de diagnóstico. Este índice de acerto representa a precisão do conjunto de detectores no diagnóstico das amostras reais da base de dados, o usuário final pode realizar este processo e avaliar se o conjunto de detectores escolhido inicialmente tem potencial para ser utilizado no processo de diagnóstico final. Caso o usuário queira o conjunto de detectores pode ser descartado, e então pode-se gerar novos conjuntos, repetindo o mesmo processo até que seja encontrado um conjunto de detectores com um índice de precisão que o usuário final desejar.

Na figura 6 apresenta-se a tela da fase de sensoriamento, onde vai ser gerado aleatoriamente um conjunto de detectores, respeitando o parâmetro da quantidade fornecido pelo usuário inicialmente. Nesta fase o usuário ao clicar no botão executar, faz com que o sistema escolha aleatoriamente na base de dados WBCD (somente nas amostras da classe “Benigno”) a quantidade de amostras que o usuário definiu nos parâmetros. Assim a tela é preenchida apresentando os respectivos detectores escolhidos, representando-os visualmente através do espectro de cores. Caso o conjunto de detectores gerados não seja agradável ao usuário, o usuário pode refazer a escolha, clicando no botão limpar, e repetir o processo para criar um novo conjunto de detectores. Ao final da tela é apresentado um campo chamado análise espectral da amostra, onde o usuário pode avaliar o conjunto de detectores, e observar qual o nível das tonalidades de cores, e caso algum padrão detector não agrade o usuário, o conjunto pode ser gerado novamente, até que sejam encontrados detectores com uma coloração mais próxima da classe “benigna”.

Na figura 7 apresenta-se a tela da fase de monitoramento do sistema, onde o conjunto de detectores gerado na fase de sensoriamento é avaliado, sendo aplicado no diagnóstico da base de dados WBCD. Nesta tela os parâmetros são carregados, a partir da tela de parâmetros onde o usuário os definiu. O campo “Amostra” irá representar as

683 amostras da base de dados, que serão comparadas com o conjunto de detectores (canto direito da tela) representados pelo campo “Detectores”.

Figura 6: Tela da fase de sensoriamento.

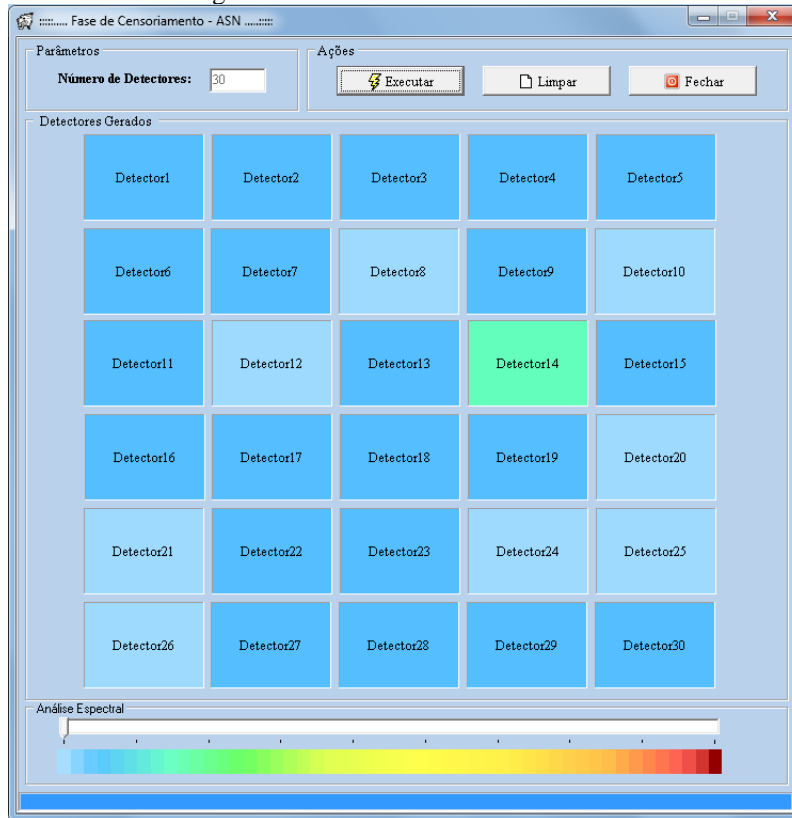
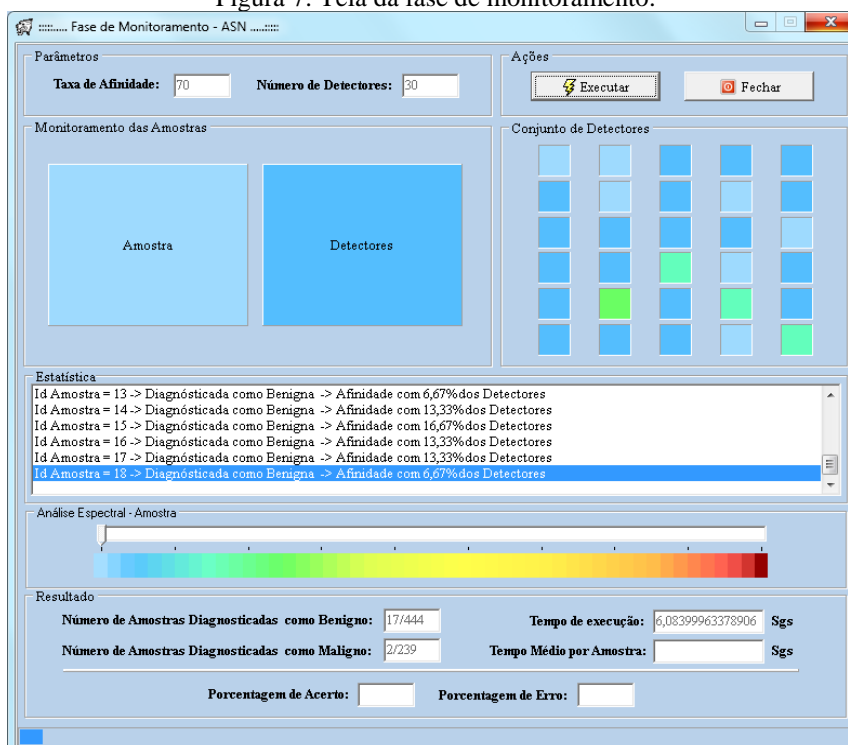


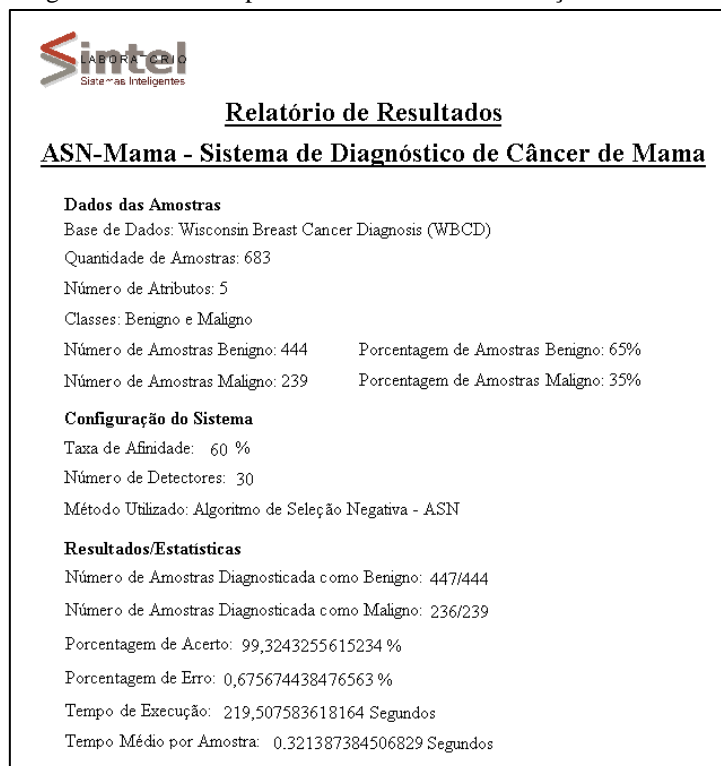
Figura 7: Tela da fase de monitoramento.



O processo de monitoramento inicia quando o usuário clica no botão executar, e assim o sistema apresenta uma comparação visual da amostra com o conjunto de detectores, indicando no campo estatística o diagnóstico para a amostra, a quantidade de detectores que a amostra se casou (afinidade em porcentagem). No campo análise espectral da amostra é indicada a coloração da amostra através do cursor que se movimenta, se posicionando na cor a qual a amostra se representa visualmente. E por fim no campo resultado é apresentado o número de amostras diagnosticadas como benigno e maligno, o tempo de execução total e o tempo médio por amostra, e a porcentagem de acerto e erro.

Ao Final do monitoramento o sistema exibe um relatório (figura 8), apresentando os dados das amostras utilizadas, a configuração do sistema, e os resultados, indicando a precisão do conjunto de detectores quando aplicado a esta base de dados.

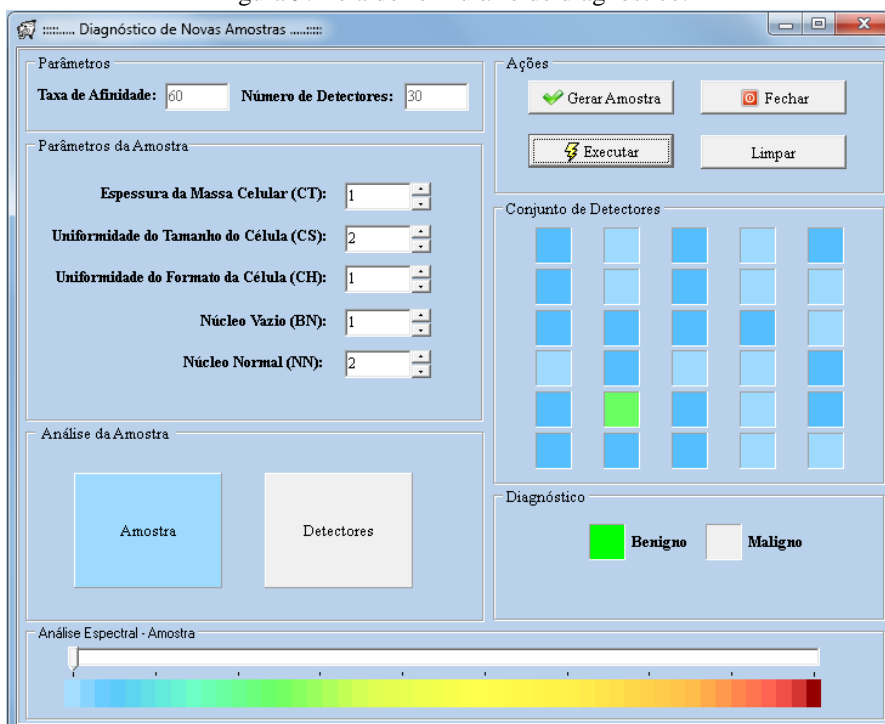
Figura 8: Relatório apresentado na fase de calibração do sistema.



Após realizar a calibração do sistema o usuário pode utilizar o conjunto de detectores para realizar diagnóstico de novas amostras reais, desconhecidas pelo sistema. Na figura 9 apresenta-se a tela do sistema que permite gerar diagnóstico de novas amostras. Os parâmetros informados inicialmente no sistema são utilizados neste passo, e o conjunto de detectores é o mesmo escolhido pelo usuário. O processo é semelhante

ao monitoramento, onde uma amostra é gerada e comparada com o conjunto de detectores. Para gerar uma nova amostra são utilizados os atributos, então um profissional especialista da área, partindo de uma mamografia por exemplo, identifica o valor dos atributos e cria a nova amostra, assim informando os valores no sistema e clicando no botão gerar amostra. Quando o usuário clica no botão executar o sistema compara a amostra gerada com os detectores e apresenta o diagnóstico no campo “diagnóstico”, informando se a amostra é benigna ou maligna.

Figura 9: Tela do formulário de diagnóstico.



Caso o usuário deseje realizar teste de novas amostras é só clicar no botão limpar, e assim o sistema é zerado, e pode-se repetir o processo novamente informando novos dados e criando novas amostras, para serem diagnosticadas. Este é o funcionamento do sistema, onde tudo se baseia no método apresentado, baseando no algoritmo de seleção negativa.

5 RESULTADOS

Nesta seção apresentam-se os resultados obtidos em testes realizados com a aplicação utilizando a base de dados WBCD, ou seja, quais os índices de precisão que o sistema apresentou. O sistema foi desenvolvido em C++ [4]. Para avaliar o desempenho

do sistema de diagnóstico, foram feitos vários testes. A seguir encontram-se descritos os testes e seus respectivos resultados.

5.1 TESTE I

Para o teste I, utilizou-se a taxa de afinidade com um valor fixo de 60%. Para este teste, na fase de sensoriamento, foram gerados os três conjuntos de padrões detectores, e foi realizado o teste para cada um deles, sendo que os conjuntos I, II e III possuem 10, 20 e 30 padrões detectores, respectivamente. Os detectores gerados utilizam 2,25%, 4,50% e 6,75% das amostras benignas, que tem um total de 444 amostras. Os resultados para este teste são apresentados na tabela 3.

Tabela 3 – Resultados para teste I.

Diagnóstico	Amostras testadas	Benigno	Maligno	Tempo execução (ms)	Acerto (%)	Erro (%)
<i>Conjunto detectores I</i>	683	438	245	92,00	98,64%	1,36%
<i>Conjunto detectores II</i>	683	444	239	98,00	100,00%	0,0%
<i>Conjunto detectores III</i>	683	448	235	103,00	98,32%	1,68%

No teste I, foi possível observar que o sistema de diagnóstico apresentou um bom índice de precisão na base de testes (índice de acerto superior a 98%), e que a quantidade de detectores benignos influencia diretamente no diagnóstico final. Comumente utiliza-se até 30% das informações da base de dados para gerar detectores e, neste caso, foi utilizado até 6,75%, visando proporcionar robustez ao diagnóstico.

5.2 TESTE II

No teste II, o objetivo é verificar a sensibilidade do método proposto com modificações na taxa de afinidade. Para o teste II, utilizou-se cinco taxas de afinidade com valores de 40%, 50%, 60%, 70% e 80%. Para este teste foi utilizado o conjunto de padrões detectores II, apresentado no teste I. Os resultados para este teste são apresentados na tabela 4.

Tabela 4 – Resultados para teste II.

Diagnóstico	Amostras testadas	Benigno	Maligno	Tempo execução (ms)	Acerto (%)	Erro (%)
Taxa I	683	509	174	90,00	72,80%	27,20%
Taxa II	683	483	200	92,00	83,68%	16,32%
Taxa III	683	444	239	93,00	100,00%	0,00%
Taxa IV	683	46	637	101,00	10,36%	89,64%
Taxa V	683	37	646	88,50	8,33%	91,67%

No teste II é possível observar que a taxa de afinidade entre as amostras influencia diretamente no diagnóstico final. Através da equação (2), faz-se um cálculo estatístico, relacionando a quantidade de amostras normais e totais da base de dados. Quando o valor da taxa de afinidade escolhida pelo usuário for próximo do valor obtido pelo cálculo em (3), o resultado é muito bom, já quando realiza-se uma variação na taxa de afinidade, observa-se que o desempenho pode ser bom, ou simplesmente ruim, ou seja, a escolha do valor da taxa de afinidade é muito importante, pois precisa ser bem calibrada para que o sistema tenha um bom desempenho. Nos resultados observa-se que o sistema chega a 100% de acerto, porém, à medida que se varia a taxa de afinidade o critério de casamento se torna mais ou menos rigoroso, fazendo que amostras que não-próprias sejam classificadas como próprias, ou vice-versa.

5.3 ANÁLISE DOS RESULTADOS

Os resultados obtidos para os testes realizados neste artigo são satisfatórios (com configurações que proporcionam um índice de acerto de 100%) e comprovam que o algoritmo de seleção negativa é eficaz no processo de diagnóstico. Os parâmetros utilizados, bem como a quantidade de detectores influencia diretamente seu desempenho. Assim, torna-se necessário realizar uma fase experimental, chamada de fase de calibração, onde vão ser encontrados os parâmetros corretos para o processo de diagnóstico. Vale ressaltar que para este artigo utilizou-se apenas 6,75% da informação das amostras para gerar os detectores benignos, o que casualmente é gerado com 30% de informação. De um total de 444 amostras benignas foram escolhidas aleatoriamente no máximo até 30 amostras para serem definidas como detectores, o que é uma quantidade muito reduzida de informação. Isto evidencia que o sistema é robusto e eficaz no processo de diagnóstico. O tempo de execução é bem reduzido, o que proporciona rapidez no diagnóstico.

Na tabela 6 apresenta-se um estudo comparativo entre o método proposto e os principais métodos disponibilizados na literatura.

Tabela 6 – Estudo comparativo.

Referência	Base de dados	Técnica utilizada	Acerto (%)
[19]	WBCD	<i>Fuzzy-genético</i>	97,07%
[16]	WBCD	<i>Fuzzy</i>	96,71%
[16]	WBCD	<i>ILFN e Fuzzy</i>	98,13%
[22]	WBCD	<i>ANFIS</i>	96,30%
[3]	WBCD	<i>Kohonen</i>	96,70%
[23]	WBCD	<i>Backpropagation</i>	95,16%
[20]	WBCD	<i>SAI-Fuzzy</i>	98,51%
[27]	WBCD	<i>SAI e Rede de base radial</i>	99,58%
Este artigo.	WBCS	<i>Algoritmo de seleção negativa</i>	100,00%

Na tabela 6 pode-se observar que o método proposto, neste trabalho, apresenta índice de acerto superior às demais técnicas, conseguindo atingir 100% de acerto no diagnóstico de câncer de mama.

6 CONCLUSÃO

Neste artigo foi apresentado um sistema comercial para diagnóstico de câncer de mama baseado nos sistemas imunológicos artificiais, em especial, o algoritmo de seleção negativa. Foram descritas as principais etapas e características do ASN e sua aplicação no problema proposto. Como dados de entrada do sistema, o algoritmo precisa apenas cinco de atributos das amostras de câncer de mama. O sistema proposto apresentou excelentes resultados obtendo um índice de acerto de 100% de acerto nos testes realizados com a base de dados WBCD. Vale ressaltar que o sistema tem uma interface simples e de fácil utilização, e pode auxiliar profissionais no diagnóstico clínico de câncer de mama, bem como, ajudar no treinamento de novos profissionais.

Sendo assim, conclui-se que os sistemas imunológicos artificiais, com base no algoritmo de seleção negativa, obtiveram um desempenho satisfatório nos testes realizados, e o sistema desenvolvido é bastante confiável, seguro e robusto para o diagnóstico de câncer de mama em ambiente hospitalar.

REFERÊNCIAS

- [1] BENNETT, K. P. and MANGASARIAN, O. L. (1992) "Robust linear programming discrimination of two linearly inseparable sets", *Optimization Methods and Software* 1, pp. 23-34 (Gordon & Breach Science Publishers).
- [2] BRADLEY, D.W. and TYRRELL, A.M. Immunotronics - Novel Finite-State-Machine Architectures with Built-In Self-Test Using Self-Nonself Differentiation. *IEEE Transactions on Evolutionary Computation*. Vol. 6, pp. 227-238, Jun 2002.
- [3] CAMASTRA, F. (2006). Kernel Methods for Clustering. In WIRN/NAIS, volume 3931 of *Lecture Notes in Computer Science*, pp. 1–9.
- [4] C++ Builder 6.0, Borland company.
- [5] DASGUPTA, D. (1998). "Artificial Immune Systems and Their Applications". Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- [6] DASGUPTA, D. (2006). "Advances in Artificial Immune Systems", *IEEE Computational Intelligence Magazine*, pp. 40-49.
- [7] de CASTRO, L. N. and TIMMIS, J. (2002). "Artificial Immune Systems: A New Computational Intelligence Approach", Springer. 1st edition.
- [8] FORREST, S., A. PERELSON, ALLEN, L. and CHERUKURI, R. (1994), "Self-Nonself Discrimination in a computer", *Proc. do IEEE Symposium on Research in Security and Privacy*, pp. 202-212.
- [9] FORREST, S., HOFMEYR S. A. and SOMAYAJI A. "Computer Immunology". *Communications of the AC*. pgs. 88-96. 1997.
- [10] HAMDI, R. E.; NJAH, M. and CHTOUROU, M. (2010). "An Evolutionary Neuro-Fuzzy approach to breast Cancer Diagnosis" *Proceedings of the Systems Man and Cybernetics - IEEE*.
- [11] INCA - Instituto Nacional do Câncer (Brasil), disponível em: www.inca.gov.br, acessado em: 01/09/2012.
- [12] JUNG, J-S. R. "ANFIS: Adaptive Network-Based Fuzzy Inference System", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 23, No. 3, May/June-1093, pp. 665-685.
- [13] KARABATAK, M.; INCE, M. C. and AVCI, E. (2008) "An Expert System for Diagnosis Breast Cancer Based on Principal Component Analyses Method". *Proceedings on Communication and Applications Conference, IEEE*, pp. 1-4.
- [14] MANIKANTAN, K.; SAYED, S.I.; SYRIGOS, K.N.; RHYS-EVANS, P.; NUTTING, C.M. HARRINGTON, K.J. and KAZI, R. (2009). "Challenges For the Future Modifications of the tnm (?) Staging System for Head and Neck Cancer. case for a new computational model?" *Cancer Treatment Reviews*, pg. 35-44.

- [15] MANGASARIAN OL, SETIONO, R. and WOLBERG, W. H. (1990) "Pattern recognition via linear programming: Theory and application to medical diagnosis ", in: "Large-scale numerical optimization", Thomas F. Coleman e Yuying Li, editores, Publicações SIAM, Philadelphia 1990, pp 22-30.
- [16] NAGHIBI, S. S.; TESHNEHLAB, M. and SHOOREHDELI, M. A. (2010) "Breast Cancer Detection by using Hierarchical Fuzzy Neural System with EKF Trainer". Proceedings of conference of Biomedical Engineering. Pg. 1-4.
- [17] MEESAD, P. and YEN, G. G. (2003). "Combined Numerical and Linguistic Knowledge Representation and Its Application to Medical Diagnosis", IEEE Transactions on Systems, Man, and Cybernetics.
- [18] OMS - Organização Mundial da Saúde, Disponível em: <http://www.who.int/en/>, acessado em: 01/09/2012.
- [19] PENA-REYES, C. A. and SIPPER, M. (1999). "Designing Breast Cancer Diagnostic System via Hybrid Fuzzy-Genetic Methodology", IEEE International Fuzzy Systems Conference Proceeding, 1999.
- [20] POLAT, K.; SAHAN, S.; KODAZ, H. E and 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 whit applications, Elsevier, n° 32, pp. 172-183.
- [21] SONG, H. J; LEE, S. G and PARK, G. T. (2005). "A Methodology of Computer Aided Diagnostic System on Breast Cancer", Proceedings of the Conference on Control applications - IEEE, Toronto, Canada, pp. 831-836.
- [22] WANG, J. and GEORGE LEE, C. S. (2002). "Self-Adaptive Neuro-Fuzzy Inference Systems for Classification Applications", IEEE Transactions on Fuzzy Systems, pp. 790-802.
- [23] WANG, J.-Y. (2005). Data Mining Analysis (Breast-Cancer Data). <http://www.csie.ntu.edu.tw/~p88012/AI-final.pdf>. Acesso em: 07/09/2012.
- [24] WBCD – Wisconsin Breast Cancer Data – UCI Machine Learning Repository, disponível em: www.arquives.ics.uci.edu/ml/.
- [25] WOLBERG, W. H. and MANGASARIAN, O. L. (1990). "Cancer Diagnosis Via Linear Programming", SIAM News, Volume 23, Número 5, setembro de 1990, pp 1 e 18.
- [26] WOLBERG, W. H. and MANGASARIAN, O. L. (1990) "Multisurface Method of Pattern Separation For Medical Diagnosis Applied to Breast Cytology", Proceedings, da Academia Nacional de Ciências dos EUA, Volume 87, dezembro de 1990, pp. 9.193-9.196.
- [27] ZHAO, W. AND DAVIS, C. E. (2011) "A Modified Artificial Immune System Based Pattern Recognition Approach – An Application to Clinical Diagnostics". Artificial Intelligence in medicine, Elsevier, n° 52, pp. 1-9.

[28] CAMPOS, M. B. P.; MACIEL, G. S.; SOUZA, S. S. F.; CHAVARETTE, F. R.; LIMA, F. P. A. Inteligência artificial com aprendizado continuado aplicada ao reconhecimento de padrões. *Brazilian Journal of Development*. v. 6, n. 5, p. 22778-22797, 2020.

[29] OLIVEIRA, D. C.; CHAVARETTE, F. R.; LIMA, F. P. A. Structural Health Monitoring using Artificial Immune System. *Brazilian Journal of Development*. v. 6, n.4, p.16948-16963, 2020.

[30] SOUZA, S. S. F.; CAMPOS, M. B. P.; CHAVARETTE, F. R.; LIMA, F. P. A. A New Approach Experimental to Diagnosis of The Failures in Mechanical Structures Using the Artificial Immune Algorithm with Negative Selection. *Brazilian Journal of Development*. v. 7, n. 7, p. 66372-66392, 2021a.

[31] SOUZA, S. S. F.; CAMPOS, M. B. P.; CHAVARETTE, F. R.; LIMA, F. P. A. Structural Failures Diagnosis using a Hybrid Artificial Intelligence Method. *Brazilian Journal of Development*. *Brazilian Journal of Development*. v. 7, n. 7, p. 66873-66893, 2021b.