UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE ELETRICIDADE

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

BRUNO RODRIGUES FROZ

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VIDAS ARTIFICIAIS, MVS E MEDIDAS DIRECIONAIS DE TEXTURA

São Luís

2015

UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE ELETRICIDADE

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

BRUNO RODRIGUES FROZ

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VIDAS ARTIFICIAIS, MVS E MEDIDAS DIRECIONAIS DE TEXTURA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão. Como um dos requisitos para obtenção do título de mestre.

Orientador: Aristófanes Corrêa Silva
Co-orientador: Anselmo Cardoso de Paiva

São Luís

2015

Froz, Bruno Rodrigues.

Classificação de nódulos pulmonares utilizando vidas artificiais, mvs e medidas direcionais de textura / Bruno Rodrigues Froz. – São Luís, 2015.

77 f.

Impresso por computador (fotocópia).

Orientador: Aristófanes Corrêa Silva.

Co-orientador: Anselmo Cardoso de Paiva.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2015.

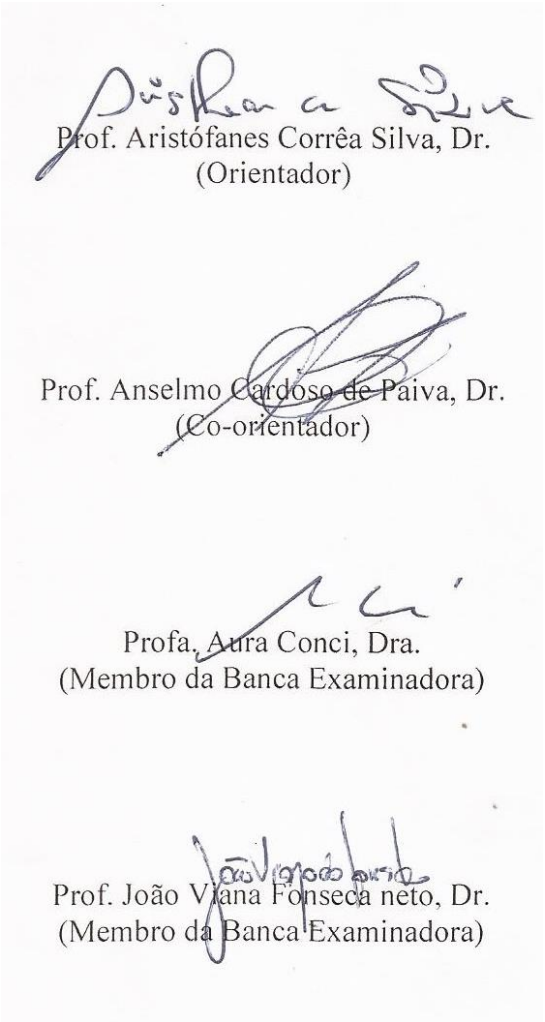
1. Processamento de imagens – Nódulo pulmonar. 2. Classificação de nódulo pulmonar. 3. Reconhecimento de padrões. 4. Vidas artificiais. I. Título.

CDU 621.397.46:616.24-006

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VIDAS ARTIFICIAIS, MVS E MEDIDAS DIRECIONAIS DE TEXTURA

BRUNO RODRIGUES FROZ

Dissertação aprovada em 02 de fevereiro de 2015



“Nós só podemos ver uma curta distância a frente, mas podemos ver muito lá que precisa ser feito.”

Alan Turing

À minha família e amigos.

AGRADECIMENTOS

A Deus, por sua benção e inspiração concedidas nos momentos mais difíceis.

Aos meus pais Antônio José Cutrim Froz e Maria de Fátima Silva Rodrigues e minha irmã Glauce da Conceição Rodrigues Froz por todo apoio e amor incondicionais a mim dados.

Ao meu orientador Prof. Aristófanes Silva por acreditar no meu potencial e ter me oferecido diversas oportunidades de crescimento acadêmico, tanto nas inúmeras discussões nas reuniões semanais, quanto facilitando o caminho para meu mestrado sanduíche na PUC-RJ, além das inúmeras vezes que ajudou com ótimas ideias.

Ao meu co-orientador Prof. Anselmo Paiva pelos ótimos conselhos e ensinamentos, que levarei para sempre comigo.

A Profa. Deane Roehl e ao Prof. Marcelo Gattass pela excelente oportunidade de trabalho no TecGraf e pelos ensinamentos na PUC-RJ.

Aos meus companheiros de apartamento durante a minha estadia no Rio de Janeiro, Wallas e Pedro, pelo companheirismo, aprendizado, compreensão e diversão ao longo e após esse período.

Aos meus colegas do LABPAI, que sempre estiveram aptos a ajudar em todos os momentos necessários, desde os conselhos nas reuniões até os sorrisos diários.

Ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão e os professores relacionados, pelo aprendizado.

Ao CNPQ pelo apoio financeiro durante o mestrado.

Aos meus amigos, em especial Arthur, Daniela, Héber, Júlio, Luciano, Paulo, Priscilla, Suellen, Tássio e tantos outros que não foram citados, mas não foram esquecidos, por fazerem parte da minha rotina, proporcionando ótimos momentos.

Muito obrigado!

RESUMO

O câncer de pulmão é conhecido por apresentar a maior taxa de mortalidade e uma das menores taxas de sobrevida após o diagnóstico, o que é causado principalmente pela detecção e tratamento tardios. Para o auxílio dos especialistas em câncer pulmonar, são desenvolvidos sistemas de diagnósticos auxiliados por computador com o objetivo de automatizar a detecção e diagnóstico dessa doença. Este trabalho propõe uma metodologia para a classificação, através de imagens de tomografias computadorizadas, de candidatos a nódulos pulmonares e candidatos a não-nódulos. O banco de imagens *Lung Image Database Consortium* (LIDC) é utilizado para a criação de uma base de imagens de candidatos a nódulos e uma base de imagens de candidatos a não-nódulos. Três técnicas são utilizadas para a extração de medidas de textura. A primeira delas é o algoritmo de vidas artificiais *Artificial Crawlers*. A segunda técnica é a utilização do *Rose Diagram* para a extração de medidas direcionais. A terceira e última técnica é um modelo híbrido que une as medidas do *Artificial Crawlers* e do *Rose Diagram*. Para a classificação é utilizado o classificador Máquina de Vetor de Suporte (MVS), com o *kernel* de base radial. Os resultados alcançados são muito promissores. Utilizando 833 exames do LIDC divididos em 60% para treino e 40% para teste, alcançou-se uma média de acurácia de 94,30%, média de sensibilidade de 91,86%, média de especificidade de 94,78%, coeficiente de variância da acurácia de 1,61% e área média das curvas ROC de 0,922.

Palavras chave: Processamento de Imagens, Reconhecimento de Padrões, Câncer, Nódulo Pulmonar, Classificação de Nódulo Pulmonar, Vidas Artificiais, Artificial Crawlers, Rose Diagram.

ABSTRACT

The lung cancer is known for presenting the highest mortality rate and one of the lowest survival rate after diagnosis, which is mainly caused by the late detection and treatment. With the goal of assist the lung cancer specialists, computed aided diagnosis systems are developed to automate the detection and diagnosis of this disease. This work proposes a methodology to classify, with computed tomography images, lung nodules candidates and non-nodules candidates. The Lung Image Database Consortium (LIDC) image database is used to create an image database with nodules candidates and an image database with non-nodule candidates. Three techniques are utilized to extract texture measurements. The first one is the artificial life algorithm Artificial Crawlers. The second one is the use of Rose Diagram to extract directional measurements. The third and last one is an hybrid model to join the Artificial Crawlers and Rose Diagram texture measurements. In the classification, que Support Vector Machine classifier is used, with its radial basis kernel. The archived results are very promising. With 833 LIDC exams, divided in 60% for train and 40% for test, we reached na accuracy mean of 94,30%, sensitivity mean of 91,86%, specificity mean of 94,78%, variance coefficient of accuracy of 1,61% and ROC curves mean área of 0,922.

Keywords: Image Processing, Pattern Recognition, Cancer, Lung Nodule, Lung Nodule Classification, Artificial Life, Artificial Crawlers, Rose Diagram.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AE	**Algoritmos Evolutivos**
AC	**Artificial Crawler**
Acc	**Acurácia**
AHSN	**Angular Histograms of Surface Normals**
AL	**Artificial Life**
CAD	**Computed Aided Diagnosis**
CAR	**Colégio Americano de Radiologia**
CV	**Coeficiente de Variância**
DICOM	**Digital Imaging Communications in Medicine**
ELCAP	**Early Lung Cancer Action Program**
Esp	**Especificidade**
FN	**Falso Negativo**
FP	**Falso Positivo**
GS	**Grid Search**
INCA	**Instituto Nacional do Câncer**
IRDI	**Image Database Resource Initiative**
LIDC	**Lung Image Database Consortium**
mACC	**Média da Acurácia**
mESP	**Média da Especificidade**
mROC	**Média da Área sob a curva ROC**
mrMR	**Minimum Redundancy Maximum Relevance**
mSEN	**Média da Sensibilidade**
MVS	**Máquina de Vetor de Suporte**
NEMA	**National Electrical Manufacturers Association**

NPS	**Nódulo Pulmonar Solitário**
PCA	**Principal Component Analysis**
QT	**Quality Threshold**
RD	**Rose Diagram**
ROC	**Receiver Operationg Characteristic**
ROI	**Regions of Interest**
SCA	**Sistemas Complexos Adaptativos**
Sen	**Sensibilidade**
TC	**Tomografia Computadorizada**
UH	**Unidades de Hounsfield**
VN	**Verdadeiro Negativo**
VP	**Verdadeiro Positivo**

1. INTRODUÇÃO

O câncer é um crescimento descontrolado das células em um determinado local do corpo. As células dividem-se rapidamente e tendem a ser muito agressivas e incontroláveis, o que determina a formação de tumores ou nódulos. O nódulo é o acúmulo de células cancerosas ou neoplasias malignas (Educação, 2014).

O câncer de pulmão é o líder na mortalidade relacionada a cânceres no mundo (Rotter, 2012). O fator crucial de ocorrência desse tipo de câncer é a grande exposição ao tabagismo. Na maioria das populações, os casos de câncer de pulmão que possuem o tabaco como principal causa representa 80% ou mais dos casos desse tipo de câncer. São apontados pelo Instituto Nacional do Câncer (INCA) como importantes fatores de risco para o câncer de pulmão, além do tabagismo: exposição ao ar de um ambiente fechado onde ocorre o cozimento de defumados em fogões de carvão; a exposição a elementos cancerígenos como amianto, arsênico, radônio 38 e hidrocarbonetos aromáticos policíclicos; e por fim, histórico familiar de câncer de pulmão (INCA, 2014).

A sobrevida média cumulativa total em cinco anos varia entre 13% e 21% em países desenvolvidos e 7 e 10% em países em desenvolvimento (INCA, 2014). Nos Estados Unidos, a taxa de sobrevivência de uma pessoa com um câncer de pulmão depois de cinco anos do desenvolvimento é em média de 16,8%, de acordo com o *National Cancer Institute* (STATISTICS, 2014), analisando casos de pacientes diagnosticados entre 2004 e 2010. Grande parte da causa dessa taxa de mortalidade alta é devido ao diagnóstico inicial do câncer de pulmão normalmente apresentar um estágio de desenvolvimento avançado da doença.

A tomografia computadorizada (TC) é um exame bastante eficiente na detecção e acompanhamento pós-tratamento do câncer pulmonar. Quanto mais cedo for detectado o nódulo, maior a chance do paciente obter um tratamento adequado, em relação a gravidade da doença (CDB, 2013). O trabalho proposto por MacMahon *et al.* 2005 estuda a importância dos exames em TC para a detecção e diagnóstico de nódulos pulmonares de tamanhos variados.

A análise das TC na busca do câncer de pulmão normalmente é feita visualmente por um profissional especializado. O diagnóstico pode conter erros, pois a avaliação humana é sujeita a falhas. O estudo de Processamento de Imagens Digitais e Reconhecimento de Padrões pode auxiliar na detecção e diagnóstico de nódulos através

de sistemas de auxílio computadorizado, ou *Computer Aided Diagnosis* (CAD) (Polakowski, Cournoyer, Rogers, & DeSimio, 1997).

Para facilitar o diagnóstico de nódulos pulmonares em exames TC há várias formas de utilizar as técnicas de Processamento de Imagens Digitais e Reconhecimento de Padrões. Uma delas é a detecção automática de nódulos pulmonares nas imagens TC. A detecção consiste em encontrar os nódulos nas imagens de pulmão passando por várias etapas, como a segmentação da região de interesse – nesse caso, são os nódulos pulmonares – e diagnostica-los, em relação a sua presença e/ou malignidade.

Este trabalho tem o objetivo de fazer a etapa de diagnóstico da detecção, em relação a presença de nódulo em exames, através da classificação do candidato em nódulo e não-nódulo. Três formas de extração de medidas de textura são analisadas e avaliadas individualmente e juntas: uso do modelo de vidas artificiais *Artificial Crawlers*; uso do modelo *Rose Diagram* para extração de medidas estatísticas; e uso de um modelo híbrido, que une os modelos *Artificial Crawlers* e *Rose Diagram*. Para a classificação baseada nas características extraídas, é utilizado o classificador Máquina de Vetor de Suporte (MVS).

A trabalho inicia com a aquisição de imagens de TC da base *Lung Image Database Consortium* (LIDC) e a criação de duas bases de imagens, uma contendo candidatos a nódulos e uma contendo candidatos a não-nódulos. Após isso, os três métodos propostos são utilizados para a extração de características. Primeiramente utiliza-se o modelo *Artificial Crawlers* para fazer a análise da textura dos candidatos, que terá como resposta curvas de evolução, que serão utilizadas posteriormente como formas de extração de medidas de textura. Depois, os candidatos serão analisados através do modelo *Rose Diagram*, que utiliza os gradientes de Sobel para a extração de medidas estatísticas de um *rose diagram*, um histograma circular. Por fim, o modelo híbrido utiliza as características extraídas dos modelos anteriores e as une, com o objetivo de aumentar o escopo de medidas para a classificação.

Algumas contribuições deste trabalho não foram encontradas no levantamento bibliográfico. Uma delas é a utilização de um modelo de vida artificial para a classificação de nódulos pulmonares. Há algumas mudanças no algoritmo evolutivo de vida artificial *Artificial Crawlers*, que é utilizado em imagens de duas dimensões, o trabalho propõe uma visão em três dimensões, para adaptar-se às imagens TC dos nódulos. Outra contribuição é o uso consistente de técnicas para uma boa classificação utilizando medidas de textura, que são medidas menos utilizadas – ou mais utilizadas de formas não

efetivas – na comunidade científica. Além disso, o uso extenso de uma base pública de imagens de exames de pulmão facilita a validação do trabalho por terceiros, além de mostrar a consistência da metodologia de forma mais clara e válida.

1.1. Trabalhos Relacionados

Nesta seção são apresentados sucintamente trabalhos relacionados com o problema de classificação.

Mousa *et al.* (2002) aplicou o MVS para a classificação de nódulos pulmonares em imagens de tomografia computadorizada. O MVS é treinado com características extraídas de 30 imagens de nódulos e 20 imagens de não-nódulos e é testado com 16 imagens de nódulos e não-nódulos. A sensitividade do classificador MVS alcançou 87,5%.

Jing *et al.* (2010) apresenta uma metodologia baseada em regras geométricas e classificação de nódulos pulmonares utilizando Máquina de Vetor de Suporte. Com essas duas técnicas, os autores comparam três propostas de abordagens: baseada somente em regras geométricas, baseada somente no MVS e uma combinação das duas abordagens anteriores. Após a extração das Regiões de Interesse, ou *Regions of Interest* (ROI), inicialmente é feita a extração de características geométricas (maior e menor eixo da ROI, *ellipticity*, área, circularidade, *slenderness* e grau retangular). A segunda e a terceira abordagem foram as que tiveram maiores acurácias (85,44% e 84,39%).

Lee *et al*. 2010 propõe uma metodologia de classificação assistida por clusterização. Esse método oferece uma estrutura para o desenvolvimento de um algoritmo de floresta aleatória híbrido para classificação de nódulo pulmonar. A base de dados utilizada foi a LIDC mas somente 32 exames de pacientes diferentes foram utilizados. O resultado alcançado foi de 98,33% de sensibilidade e 97,11% de especificidade.

Namin *et al.* (2010) apresenta uma metodologia de segmentação e classificação de nódulos pulmonares em exames de TC utilizando o *Fuzzy K-Nearest Neighbor* (FKNN), tanto para a etapa de segmentação dos candidatos, quanto para a etapa de seleção dos candidatos entre nódulos e não-nódulos. Na etapa de seleção de características para a etapa de classificação dos nódulos, são utilizadas as medidas baseadas na

intensidade dos *voxels* e na geometria do candidato. Os autores não utilizaram nenhuma medida de textura, somente medidas baseadas na intensidade em unidade de *Hounsfield* e baseadas na geometria do objeto, entretanto o método proposto obteve 88% de acurácia e uma média de 10,3 falso positivos por exame.

Tartar *et al*. 2013 propôs uma metodologia com uma abordagem de classificação de nódulos pulmonares utilizando imagens de tomografia computadorizada e características híbridas. Utilizando exames de 63 pacientes diferentes, a metodologia testa três métodos diferentes para a extração de características: Análise de Componente Principal, ou *Principal Component Analysis* (PCA), em duas dimensões sem os resultados estatísticos juntamente com o método *Minimum Redundancy Maximum Relevance* (mrMR); Resultados estatísticos do PCA em duas dimensões e mrMR; Por fim, características de forma. A proposta alcançou um resultado de 90,7% de acurácia, 89,6% de sensibilidade e 87,5% de especificidade.

Zhang *et al*. 2013 apresenta uma abordagem de classificação baseada em características para classificar nódulos pulmonares em imagens de tomografia computadorizadas de baixa dose. A classificação se dá em três categorias: bem-circunscrita, vascularizada, justa-pleural e cauda-pleural. As características focadas descrevem tanto o nódulo quanto seus arredores e utilizam dois estágios: rotulamento de *superpixel* e cálculo de curvas de contexto. Esse último transforma o resultado do rotulamento de *superpixels* em um vetor de características. A análise dos resultados é feita utilizando a base de imagens pública *Early Lung Cancer Action Program* (ELCAP), que contém 50 exames de baixa dose e alcançou o resultado de 82,5% de acurácia.

Choi *et al.* (2014) mostra uma metodologia de segmentação e classificação de nódulos pulmonares em tomografia computadorizada. Na etapa de classificação, os autores utilizaram a mesma matriz Hessiana, utilizada na segmentação das estruturas internas ao parênquima, para calcular seus dois histogramas angulares. Com esses histogramas, são calculadas as características dos *Angular Histograms of Surface Normals* (AHSN). A partir desse conjunto de características é feita uma etapa de eliminação de estruturas ligadas à parede do parênquima. Em seguida e as estruturas restantes são classificadas em nódulos e não-nódulos. Mesmo alcançando 97,4% acurácia com 97,2% de sensibilidade e 97,7% de especificidade, os autores utilizaram somente as medidas de forma.

Filho *et al.* 2014 propõe uma metodologia de classificação de nódulos e não-nódulos utilizando índices de diversidade taxonômicos. O cálculo desses índices é baseado em árvores filogenéticas que são aplicadas para a caracterização dos candidatos. O MVS é utilizado para a classificação e a base de imagens públicas LIDC, com 833 exames. O trabalho alcançou o resultado de média de acurácia de 97,55%, média de sensibilidade de 85,91% e média de especificidade de 97,70%.

Franco *et al*. 2014 apresenta uma metodologia que utiliza descritores geométricos e atributos de Haralick extraídos de tomografias computadorizadas para fazer a classificação de nódulos pulmonares utilizando redes neurais artificiais. Utilizando base proprietária, o trabalho alcançou até 99,17% de taxa de acerto em uma das redes neurais dos testes.

A Tabela 1 mostra um resumo dos resultados dos trabalhos relacionados.

Tabela 1: Resultados dos trabalhos relacionados.

Trabalho	Base	Quantidade de exames	Sensibilidade	Especificidade	Acurácia
Mousa *et al.*, 2002	**Proprietária**	**50**	**87,5%**	**-**	**-**
Jing *et al.*, 2010	**LIDC**	**-**	**-**	**-**	**85,44%**
Lee *et al.* 2010	**LIDC**	**32**	**98,33%**	**97,11%**	**-**
Namin *et al.*, 2010	**LIDC**	**-**	**-**	**-**	**88%**
Tartar *et al.* 2013	**Proprietária**	**63**	**89,6%**	**87,5%**	**90,7%**
Zhang *et al.* 2013	**ELCAP**	**50**	**-**	**-**	**82,5%**
Choi *et al.*, 2014	**LIDC**	**84**	**97,2%**	**97,7%**	**97,4%**
Filho *et al.*, 2014	**LIDC**	**833**	**85,91%**	**97,7%**	**97,55%**

Franco *et al*. 2014	**Proprietária**	**156**	**-**	**-**	**95,70%**

Conforme observado na Tabela 1 e nos resumos dos trabalhos relacionados, alguns problemas são observados. Primeiro pode ser citado a quantidade pequena de exames utilizados para testes. Depois podemos observar que alguns dos trabalhos utilizam bases proprietárias. Também podemos observar que a maioria dos trabalhos utilizou medidas de geometria para fazer a classificação. Aqueles que utilizaram medidas de textura tiveram resultados abaixo da média dos outros trabalhos. Além disso, os testes propostos na maioria dos trabalhos não são suficientes para mostrar a consistência das metodologias utilizadas.

O trabalho proposto tenta resolver esses problemas, começando pela base de imagens ser grande e pública. Além desse problema, neste trabalho serão utilizadas somente medidas de textura, onde nos trabalhos relacionados mostraram o pior desempenho de classificação, e neste os resultados são melhores. Outra vantagem nesse trabalho são os testes propostos, utilizando divisão da base em treino e teste em grupos de 80/20%, 60/40%, 40/60% e 20/80%, sendo treino e teste respectivamente.

1.2. Organização do Trabalho

Este trabalho está dividido em quatro capítulos, que começa dos conhecimentos básicos necessários para o entendimento claro da metodologia proposta e finaliza mostrando os resultados obtidos.

O Capítulo 2 é a fundamentação teórica necessária para entender o problema e a metodologia proposta.

O Capítulo 3 explica de forma bem detalhada a metodologia proposta, todos os artifícios estudados no capítulo anterior sendo utilizados na prática, além das dificuldades e formas de resolvê-las que foram encontradas ao longo do tempo.

O Capítulo 4 discute os testes propostos para validar a metodologia e os resultados obtidos utilizando a mesma.

Por fim, o Capítulo 5 contém considerações finais sobre o trabalho proposto e os resultados obtidos, além de mostrar o caminho para futuros trabalhos que podem ser realizados a partir deste.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo a teoria necessária para o entendimento deste trabalho é explicada. Todo o conhecimento necessário para facilitar a leitura da metodologia de classificação de nódulos é demonstrado de forma sucinta.

2.1. Nódulo Pulmonar

Um nódulo é uma pequena massa de tecido que forma dentro ou sobre o corpo, normalmente em resposta a lesões. Na sua maioria, os nódulos são benignos, e não requer nenhuma ação médica, mas por vezes, eles podem interferir na função do corpo ou podem ser malignos (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

Um nódulo benigno é um tumor não se espalha para outras partes do corpo. Os nódulos benignos tendem a crescer mais lentamente do que os malignos e são menos susceptíveis de causar problemas de saúde. Apesar de seu estado menos agressivo, os nódulos benignos precisam de acompanhamento médico, se possível, em estado inicial de desenvolvimento (Sousa, 2007).

Os nódulos malignos são também denominados câncer, uma doença que é o crescimento anormal de células com a propriedade de invadir tanto os tecidos adjacentes quanto os distantes.

Boa parte dos nódulos pulmonares surge nas paredes dos brônquios, o que dá ao câncer pulmonar também o nome de broncogênico. Eles também podem ocorrer com frequência nas paredes dos pulmões e podem levar vários anos para se desenvolverem.

A fase inicial de desenvolvimento é assintomática e é formada por uma área pré-cancerosa não identificável por exames de imagem. Com o passar do tempo, entretanto, ocorre o desenvolvimento do nódulo e, possivelmente, o espalhamento de suas células alteradas pela corrente sanguínea, configurando um processo chamado metástase (ACS, 2014).

Nódulo Pulmonar Solitário (NPS) é definido como uma lesão esférica ou oval, menor ou igual a 3 centímetros de diâmetro rodeado pelo parênquima pulmonar (Ost, Fein, & Feinsilver, 2003). Um nódulo pulmonar solitário pode ser considerado benigno

caso seu crescimento permaneça estagnado por um período de dois anos. As lesões malignas, ao contrário, apresentam tempo de duplicação que pode variar entre 30 e 400 dias, ou seja, seu volume, geralmente, torna-se duas vezes maior nesse intervalo de tempo (Uehara, Jamnik, & Santoro, 1998).

A presença de calcificações é um fator sugestivo de benignidade pois são encontradas mais frequentemente nos granulomas que nos tumores cancerosos. Nódulos com maior absorção, com valores superiores a 150 ou 200 unidades de Hounsfield (UH) também são, em geral, benignos. Esse aumento na densidade média deve-se a finas calcificações internas à lesão. As calcificações em nódulos benignos geralmente apresentam posição central, enquanto que calcificações excêntricas também podem ser encontradas nas lesões malignas (Sousa, 2007).

2.2. Imagens DICOM

A tomografia computadorizada foi desenvolvida por Godfrey Hounsfield e Allan Cormack em 1972 e o seu funcionamento consiste basicamente em uma máquina que emite Raios X que gira em volta do corpo do paciente e à medida que se move emite Raios X em 360º, ou seja, fazendo uma circunferência completa em torno do paciente (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

Tomografia Computadorizada (TC) é um exame que permite a obtenção de imagens de cortes do corpo do paciente, sendo bastante utilizada como um exame médico de diagnóstico por imagem. Comparada à radiografia tradicional, a TC apresenta maior precisão e sensibilidade, além de não apresentar uma imagem com sobreposição de tecidos (Santos, 2011).

O Colégio Americano de Radiologia (CAR) e a *National Electrical Manufacturers Association* (NEMA) reconheceram a emergente necessidade de um método padrão para a transferência de imagens e informações associadas entre dispositivos fabricados por diversos fornecedores (NEMA, 2014).

O CAR e a NEMA formaram uma comissão conjunta em 1983 para desenvolver o padrão *Digital Imaging Communications in Medicine* (DICOM), que tem como objetivo promover a comunicação de informação de imagem digital, independentemente do

fabricante do dispositivo, facilitando o desenvolvimento e a expansão de arquivamento de imagens e sistemas de comunicação e permitir a criação de bases de informações de dados de diagnóstico que podem ser acessados por uma variedade de dispositivos distribuídos geograficamente (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

2.3. Base LIDC-IDRI

A coleção de imagens *Lung Image Database Consortium* (LIDC) consiste de uma coleção de diagnósticos de câncer pulmonares em tomografias computadorizadas torácicas com marcações de lesões anotadas. É um recurso internacional e de fácil acesso pela internet para o desenvolvimento, treino e validação métodos de sistemas de diagnósticos computadorizados para a detecção e diagnóstico de câncer pulmonar.

Com a parceria da *Image Database Resource Initiative* (IRDI), houve um aumento substancial da base LIDC (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

Todas as imagens estão no formato DICOM e possuem 16 bits por *voxel*. Possuem dimensões de 512 x 512 (altura e largura) com quantidades variadas de imagens para cada exame. A base conta com um arquivo que contém informações sobre marcações de nódulos realizadas por quatro especialistas. Esses arquivos contêm informações dos nódulos como: taxa de malignidade, dificuldade de detecção, textura, nível de calcificação, esfericidade, entre outras.

Essa base dispõe de um arquivo individual para cada exame que contém informações sobre cada nódulo presente no exame. Nessa base encontram-se diversos tipos de nódulos, dentre eles temos: pequenos (diâmetro maior que 3 e menor que 10 mm), conectados a vasos sanguíneos, justa pleurais e nódulos grandes (diâmetro maior que 10 e menor que 30 mm). Existe também a presença de massas, que são estruturas com diâmetro maior que 30 mm.

As imagens DICOM dessa base pública serão utilizadas para validar o trabalho proposto, já que possuem as informações necessárias para a realização de testes confiáveis.

2.4. Evolução e Modelos de Vidas Artificias

Evolução é um processo multinível de alta complexidade. Modelar matematicamente a evolução requer uma boa abstração das características fundamentais e essenciais (Schuster, 2011). São exemplos de características fundamentais: seleção natural, leis de herança de Mendel (McClean, 2000), otimização por mutação e seleção e evolução natural.

Algoritmos Evolutivos (AE) são programas de computador que tentam resolver problemas de alta complexidade reproduzindo o processo da evolução Darwiniana (Darwin, 1859) . Nos AE são criados indivíduos artificiais que percorrem um espaço de um problema específico. Eles competem entre si e descobrem formas de otimizar o espaço de procura. Ao longo do tempo – ciclo ou iteração – o indivíduo mais bem sucedido vai evoluir e descobrir uma solução otimizada (Jones, 1998).

Os indivíduos dos AE são tipicamente representados por uma string ou vetor de tamanho fixo. Cada um guarda uma possível solução para o problema específico. Os AE afetam populações de indivíduos. Inicialmente os AE tem uma população inicial de tamanho μ contendo indivíduos aleatórios, ou seja, cada valor em cada string ou vetor é criado usando um gerador de números aleatórios. Cada indivíduo possui também um valor de fitness, que é uma espécie de qualificador da solução do indivíduo para o problema. Quanto maior o valor do fitness do indivíduo, melhor é a solução que o indivíduo tem para o problema. Esse valor de fitness é calculado através da função de fitness, que utiliza como parâmetro a possível solução do indivíduo para o problema proposto.

Seguindo essa fase inicial, os ciclos iterativos do algoritmo iniciam. Usando mutação e recombinação, os μ indivíduos na população atual produzem λ filhos. Os λ filhos são relacionados com novos valores de fitness. Uma nova população de μ indivíduos é formada a partir dos μ indivíduos e dos λ filhos. A nova população se torna a população atual e o ciclo ocorre novamente. Em alguns momentos, a estratégia de evolução Darwiniana é aplicada e os indivíduos devem competir entre si. A seleção baseada em fitness resolve esse conflito, onde o indivíduo com melhor fitness é mais provável sobreviver.

Com essa definição de evolução, é possível compreender a ideia por trás dos modelos de Vidas Artificiais, onde um modelo chamado *Artificial Crawlers* será utilizado na metodologia proposta.

2.4.1. Vidas Artificiais

A vida natural é organizada em pelo menos quatro níveis de estruturas: nível molecular, nível celular, nível de organismo e nível de população-ecossistema (Taylor & Jefferson, 1995). Em qualquer um desses níveis, um indivíduo é composto por um sistema complexo e adaptativo que exibe comportamentos que surge da interação de indivíduos do nível diretamente inferior. Por exemplo, no corpo humano, um órgão é composto de um sistema de células que tem uma função específica e essa funcionalidade só aparece quando as células formam o sistema e se adaptam a situações onde estão expostas.

Para poder lidar com essa complexidade multinível, a abordagem de modelo de Vidas Artificiais foi desenvolvida na década de 80 (Langton, 1989). O maior desafio desse modelo é formalizar regras e leis da evolução da vida por meio de um modelo computadorizado de entidades que parecem simular a vida.

O termo Vidas Artificiais, ou *Artificial Life* (AL), pode ser interpretado como "vida feita por humanos e não pela natureza". Para Langton (1989), a principal hipótese feita no modelo de Vidas Artificiais é a de que a 'forma lógica' de um organismo pode ser extraída de seu material. Organismos de AL são feitos por homens e são entidades imaginárias, mas são baseadas em organismos vivos na natureza e suas ações são baseadas na forma mais lógica possível, para possibilitar a modelagem de um ser parecido com um ser vivo e analisar seu comportamento em um ambiente.

Alguns modelos de AL conhecidos e bem difundidos na comunidade científica são relatados abaixo:

1. PolyWorld: é um modelo computacional desenvolvido para explorar pontos importantes e problemas das AL. Os organismos simulados se reproduzem, lutam, matam ou aliam-se entre si, comem alimentos que crescem no "mundo" e desenvolvem estratégias bem sucedidas para sobreviver ou morrem. O comportamento do organismo (mover, atacar, comer, aliar, etc.) é controlado por sua rede neural, ou "cérebro". A arquitetura de cada "cérebro" é definida pelo seu código genético, em termo de número, tamanho e composição dos seus clusters neurais (neurônios excitatórios e inibidores) e o tipo de conexões entre esses clusters (mapeamento topológico e conexão por densidade).

A eficácia sináptica é modulada via aprendizado Hebbiano, permitindo que os organismos tenham a habilidade de aprender durante o curso das suas vidas. Os organismos percebem o mundo através do sentido da visão, provido pela renderização de computação gráfica do mundo a partir do ponto de vista de cada organismo. Suas fisiologias são também codificadas geneticamente, então ambos o corpo, cérebro e comportamento evoluirem ao passar das gerações (Yaeger, 1994).

2. Tierra: o Tierra é um modelo de evolução de programas auto-reproduzidos desenvolvido por Tom Ray (1989). A ideia era criar um programa auto replicável que simulasse o comportamento das primeiras criaturas vivas do planeta Terra. Cada vez que o programa se reproduz, há uma pequena chance de uma mutação ocorrer, produzindo uma mudança no programa replicado. Os organismos do Tierra incluem fitas de genomas que determinam as instruções do programa executado. Os organismos são aptos a mudar segmentos dos códigos dos programas. As iterações entre os organismos resultam em uma emergência evolucionária de biodiversidade complexa dos programas autorreplicados. Um grande problema nessa abordagem são as fragilidades das linguagens de programação: quanto mais alto o nível da linguagem, menos variação na evolução. Na prática, o Tierra descreve programas que competem por poder de processamento na CPU e acesso a memória (Charrel, 1995).

3. Avida: é um modelo baseado no Tierra e é designado para o estudo evolucionário da biologia de auto replicação e evolução de programas de computador. No Avida, um modelo matemático descreve a distribuição das espécies na população em evolução. Diferente do Tierra, o Avida aloca para todos os organismos suas próprias regiões da memória e executa-os com seus próprios CPU's virtuais. Por padrão, outros organismos não têm acesso a esses espaços reservados de memória e nem podem executar códigos neles. O Avida garante o paralelismo com iterações locais, provendo tempos de processamento menores que o Tierra, além de fatias de tempo constantes para cada organismo (Adami & Brown, 1994).

4. Echo: esse modelo descreve a evolução de agentes simples, os quais podem interagir aliando-se, lutando e trocando informações entre si. A interação entre os agentes resulta em uma ecologia complexa. Echo é uma ferramenta de simulação desenvolvida para investigar mecanismos que regulam a diversidade e processamento de informações em sistemas compreendidos de muitas agentes adaptativos interagindo, conhecidos como Sistemas Complexos Adaptativos (SCA). Os agentes interagem através de combate, troca e alianças, desenvolvendo estratégias para garantir a sobrevivência em um ambiente de

recursos limitados. Em uma simulação típica, a população desenvolve uma rede de interações que regulam o fluxo de recursos. A rede resultante assemelha-se a comunidade de espécie em sistemas ecológicos. A grande vantagem nesse modelo é a flexibilidade de definição dos parâmetros e condições iniciais, o que facilita o trabalho dos pesquisadores ao tentar simular um ambiente (Holland, The Echo Model, 1992).

5. Modelo de Co-Evolução de Hospedeiro e Parasitas: Esse modelo descreve uma população de indivíduos hospedeiros como algoritmos que tem a intenção de resolver algum problema específico, como por exemplo, problema de ordenação. Além da população de hospedeiros há uma população de parasitas, que são representados como tarefas a serem resolvidas. A população hospedeira evolui para procurar boas soluções do problema específico, enquanto a população de parasitas evolui para tentar tornar esse problema mais difícil. A competição hospedeiro-parasita co-evolui para assegurar a descoberta de soluções significativamente melhores do que as soluções encontradas somente utilizando a população de hospedeiros sozinha. Em outras palavras, esse modelo tem como objetivo encontrar soluções otimizadas e flexíveis para a solução de problemas (Hillis, 1990).

6. AntFarm: é um programa de computador que simula a evolução de estratégias de forrageamento de colônias de organismos artificiais que lembram formigas. Esse modelo é usado para investigar problemas ao redor da evolução simulada de comportamentos complexos em ambientes complexos. Além disso, a evolução da cooperação entre indivíduos brevemente semelhantes e a evolução de comunicações químicas são exploradas utilizando o AntFarm (Collins & Jefferson, 1991).

7. Sistema Classificatório: o modelo de Sistema Classificatório induz um esquema de inferência que é baseado em um conjunto de regras lógicas. Cada regra tem a seguinte forma: "Se <condição> Então <ação>". O sistema de regras é otimizado pelo aprendizado e pela pesquisa evolucionária. O aprendizado define a prioridade de cada regra particularmente através de uma força que é modicada por um algoritmo chamado Suporte de Brigada (*Backed-Brigade*). A pesquisa evolucionária é a descoberta de novas regras por meio de um algoritmo genético (Holland, Holyoak, Nisbett, & Thagard, 1989).

2.4.2. Modelo Artificial Crawlers

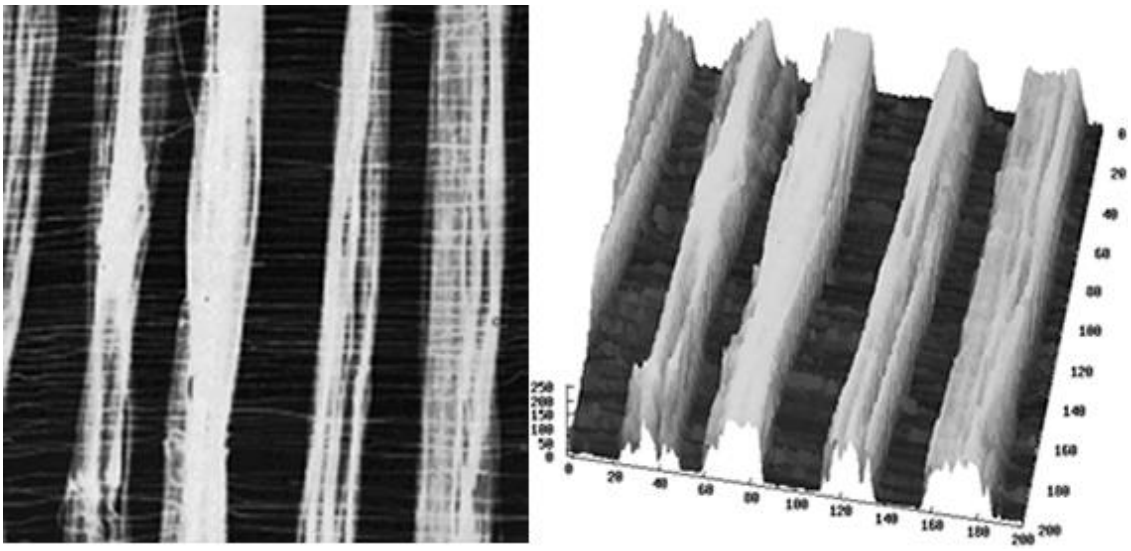
Baseando-se nos modelos de AL, o modelo *Artificial Crawler* (AC) conjectura que cada organismo individual, também chamados de AC, vive em um pixel de uma imagem. Um certo número de AC nasce com o valor de energia e características iguais

para todos. A energia de cada AC pode incrementar ou decrementar devido a influência do ambiente e consumo de energia e eventualmente podem morrer devido ao fim da energia ou competição com outros indivíduos (Zhang & Chen, 2004).

A interação entre os AC com o ambiente vai depender de um equilíbrio. Utilizando a textura da imagem como ambiente, o valor dos pixels vão simular diferentes altitudes que suprem alimentos para os AC. Quanto maior a intensidade do pixel na textura da imagem, maior a altitude do ambiente dos AC.

O processo da evolução remete a um histórico de competição, movimento, assentamento, sobrevivência e morte dos AC. Ao longo da evolução, certas colônias vão surgir, representando características da textura da imagem. Essas características vão variar dependendo propriamente da imagem (Zhang & Chen, 2005).

A Figura 1 mostra um exemplo de textura e seu mapa em três dimensões de altitudes em relação ao valor de cada pixel.



(a) (b)

Figura 1: Textura e mapa em três dimensões. (a) Representa a textura analisada. (b) Representa o mapa da textura em relação ao valor dos pixels Fonte: (Gonçalves, Machado, & Martinez, 2014).

Cada AC possui três características fundamentais:

8. $e_i(t)$: a energia de um AC no tempo t;
9. $\delta_i(t)$: a colônia a que um AC pertence no tempo t;
10. $\beta_i(t)$: a localização de um AC no tempo t.

onde t é o tempo relativo a partir do primeiro ciclo do algoritmo e i é a identidade do AC. A cada novo ciclo, t é incrementado.

Durante a evolução, cada AC deve seguir um conjunto de regras que vai definir seu estado durante cada geração. As regras são:

1. $\forall_i$, $e_i(0) = e^0$: Na primeira geração, todos os AC nascem com a energia e^0, arbitrária;
2. $\forall t,_i$, SE $e_i(t) \leq e_{min}$, AC morre : Para cada ciclo e para cada indivíduo, se a energia desse indivíduo é menor do que e_{min}, o indivíduo morre. e_{min} é um limiar que indica a energia mínima que o indivíduo precisa para sobreviver.
3. $\forall_i$: $e_i(t) > e_{min}$, $\beta_i(t + 1) = f(\beta_i(t))$: Em cada um dos indivíduos vivos no tempo t, eles seguem a regra de movimento f(β). Os possíveis valores de f(β) são mostrados da Equação 1.

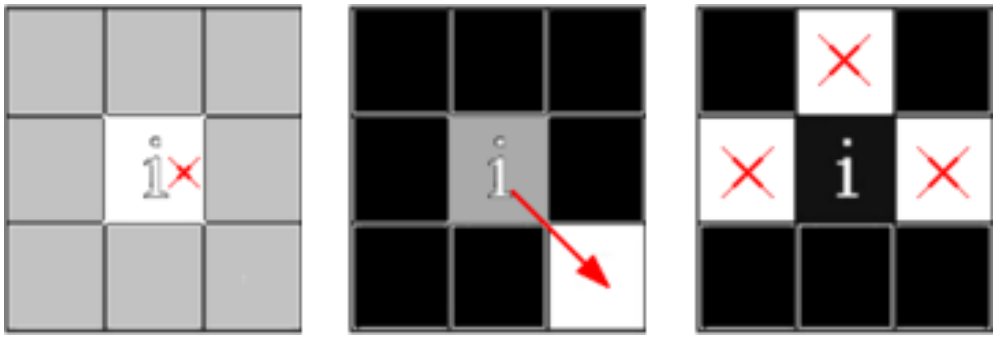
$$f(\beta) = \begin{cases} \beta, se\ (a) \text{é } satisfeito \\ \beta\text{max}, se\ (b) \text{ é } satisfeito \\ \beta 0, \text{ se (c) é satisfeito} \end{cases} \quad (1)$$

(a) Se a intensidade máxima dos 8-vizinhos do AC_i é menor que a intensidade da localização $\beta_i(t)$, ele não se move;

(b) Se a intensidade máxima dos 8-vizinhos do AC_i é maior que a intensidade do pixel dele e é única (ou seja, há apenas um pixel de valor maior do que o dele, entre os vizinhos), ele se move para o pixel de maior intensidade;

(c) Se a intensidade máxima dos 8-vizinhos do AC_i é maior que a intensidade do pixel dele e não é única (ou seja, há mais de um pixel com valor maior do que o dele, entre os vizinhos), ele se move para o pixel que já foi ocupado por outro AC. Do contrário, ele se move para qualquer um de maior intensidade.

A Figura 2 mostra como se comporta o movimento dos AC na regra 3.



(a) (b) (c)

Figura 2: Movimento dos AC na regra 3. (a), (b) e (c) representam os movimentos descritos. Fonte: (Gonçalves, Machado, & Martinez, 2014).

4. $\forall_i : ei(t) > e_{min}, ei(t + 1) = ei(t) – e_{unit}$: cada movimento do AC consome um unidade de energia, definida em e_{unit};
5. $\forall_{i,j} : \beta_i(t + 1) = \beta_j(t + 1), e_{sig\ max\{ei,ej\}} = max\{ei,ej\}$ & sig $min\{ei,ej\}$ morre: quando um indivíduo AC move para um pixel ocupado por outro indivíduo, a Lei da Selva é aplicada. O AC que possui energia superior absorve o AC que possui energia inferior;
6. $\forall_i, e_i(t + 1) = e_i(t) + \lambda I(x,y)$: quando um AC move para um pixel de intensidade I(x,y), ele absorve uma porcentagem de energia do ambiente determinada pelo fator λ, mas não modifica a energia do ambiente;
7. $\forall_{t,i} : e_i(t) \geq e_{max}, e_i(t + 1) = e_{max}$: A energia máxima que o AC pode absorver/incrementar é determinada pelo limiar e_{max};

As regras são aplicadas para todos os AC. O que pode ser observado ao passar dos ciclos é que os AC que ficam mais próximos de áreas de valores altos de pixels provavelmente sobreviverão. Os AC que nascem longe dessas áreas de máximos locais, normalmente morrem no processo, ou por causa do fim de suas energias ou por serem engolidos por outro AC em algum dos ciclos.

O que se pode observar nesse modelo é a convergência para o equilíbrio e a emergência dos AC. Ao passar dos ciclos, um momento será alcançado, onde nenhum AC morrerá ou nenhuma nova colônia será reagrupada, o que caracteriza a emergência. Alguns AC tentarão mover-se para áreas de maior altitude para absorver nutrientes (regra 6) e ao mesmo tempo perderão uma unidade de energia por causa desse movimento (regra 4). Esses dois efeitos opostos demonstram o equilíbrio.

2.4.2.1. Classificação de Textura com AC

A textura é uma característica importante utilizada na interpretação visual de imagens e o uso de medidas de textura pode aumentar o desempenho de classificadores digitais (Rennó & Soares, 1996). Ao contrário das medidas de forma, que extraem informações espectrais baseadas nas variações do nível de cinza de um pixel em uma imagem, a textura contém informações de distribuição espacial dos níveis de cinza de uma região de pixels de uma imagem (Marceau, 1989).

Alguns autores tentam descrever medidas de texturas, pois não há um consenso na definição da textura em uma imagem e nem uma formulação matemática precisa (Rennó & Soares, 1996). Alguns descritores e análise de textura famosos são: padrões de frequência (He, 1990), estatísticas de primeira ordem (Hsu, 1977) e estatísticas de segunda ordem (Haralick, Shanmugam, & Dinstein, 1973).

Muitos desses descritores de texturas de imagens foram utilizados ao longo dos anos, inclusive para a detecção de nódulos pulmonares, como visto nos trabalhos de Sousa (2009), Netto (2010) e Filho (2013).

O modelo AC pode ser utilizado para classificação de textura. Empiricamente, é observado que o número de iterações (ciclos) para ocorrer o equilíbrio no modelo varia de acordo com a imagem, ou seja, imagens diferentes e geram ciclos diferentes – o que caracteriza texturas diferentes. Além disso, alguns desvios em relação a convergência também são observados.

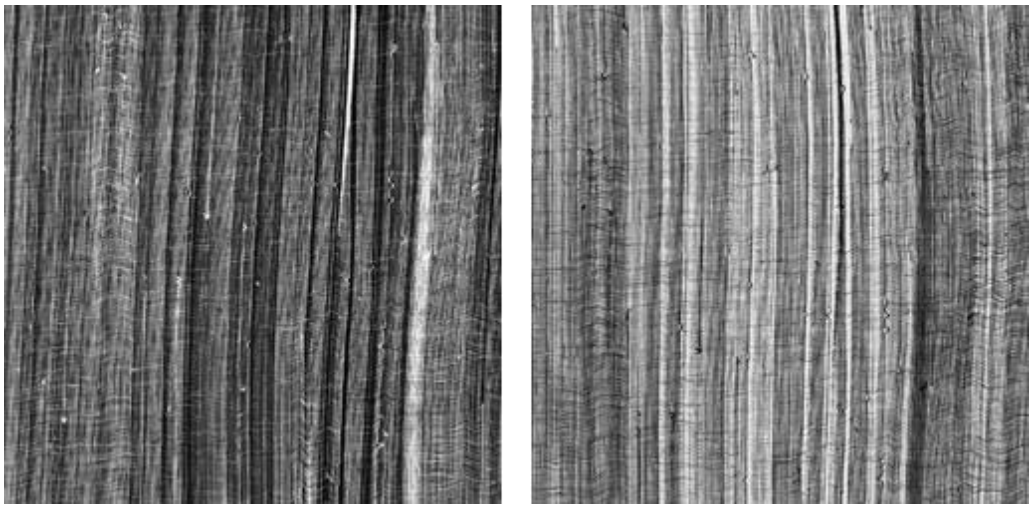
Para iniciar o ciclo de vida, AC são gerados em localizações aleatórias da imagem. Ao iniciar o processo, todos os AC são submetidos às sete regras, em cada época do ciclo de vida. Ao longo do tempo, a evolução dos AC mostra que diferentes colônias surgirão. Em alguma época será alcançada o equilíbrio.

Nesse processo, as propriedades individuais e coloniais dos AC são guardadas a cada iteração. Uma série de curvas de evolução são geradas para representar características de textura. Cada imagem corresponde a quatro curvas específicas: evolução de agentes, assentamento de habitantes, formação de colônias e distribuição escalar. Quanto mais parecidas as texturas são, mais similar é a distribuição dos AC no espaço da imagem.

2.4.2.2. Curvas de Evolução

Cada AC individualmente só pode perceber o ambiente nas suas proximidades. Ao longo da evolução, uma percepção global da textura emerge, sendo um mapa cognitivo da tribo de AC. Cada curva representa essa percepção global.

Cada curva será descrita e exemplificada com duas texturas de Brodatz (Brodatz, 1968), mostradas na Figura 3.



(a) (b)

Figura 3: Texturas de Brodatz. (a) Textura D105. (b) Textura D106. Fonte: (Gonçalves, Machado, & Martinez, 2014).

- Evolução de Agentes

A característica mais relevante na curva de Evolução de Agentes consiste na existência de pontos de inflexão. A quantidade de AC nos pontos de inflexão cai drasticamente a cada iteração, devido ao fim da energia. Em outras palavras, a quantidade de AC nos pontos de inflexão é a quantidade de AC sobreviventes a cada ciclo. Diferentes texturas têm diferentes pontos de inflexão, o que faz com que essa curva seja uma característica importante na classificação. A Figura 4 mostra um exemplo da curva de Evolução de Agentes nas texturas de Brodatz D105 e D106.

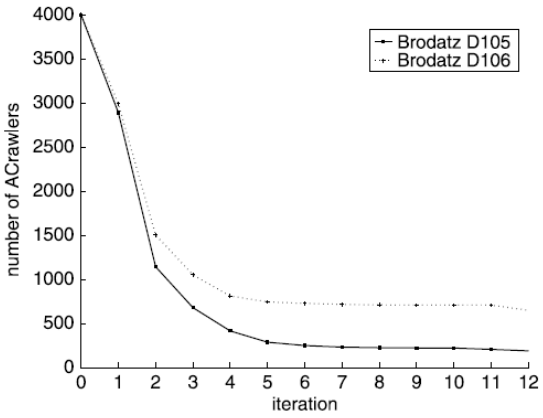


Figura 4: Curva de Evolução de Agentes. Fonte: (Zhang & Chen, Artificial Life: A new approach to texture classification, 2005).

- Assentamento de Habitantes

A curva de Assentamento de Habitantes indica o número de AC que foram eliminados a cada ciclo. Analogamente a curva de Evolução de Agentes, essa curva representa a quantidade de AC em pontos de inflexão temporários, que ao se deslocarem perderam energia suficiente para a sobrevivência ou foram engolidos por outros. A Figura 5 mostra um exemplo da curva de Assentamento de Habitantes nas texturas de Brodatz D105 e D106.

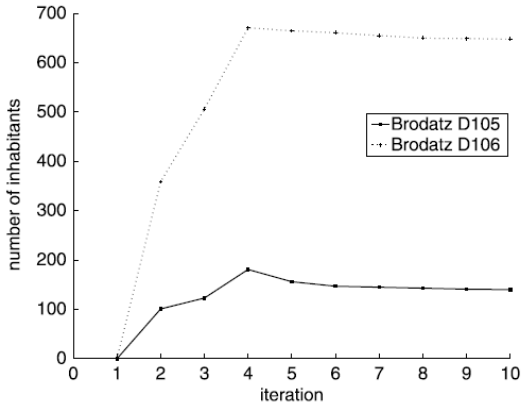


Figura 5: Curva de Assentamento de Habitantes. Fonte: (Zhang & Chen, Artificial Life: A new approach to texture classification, 2005).

- Formação de Colônias

AC que estão dentro de um certo raio vão ser inclusos na mesma colônia. A curva de Formação de Colônias mostra a quantidade de colônias formadas em relação a um raio específico. A curva representa no eixo horizontal o tamanho do raio incrementado e no eixo vertical a quantidade de colônias formadas com esse raio. A Figura 6 mostra um exemplo da curva de Formação de Colônias nas texturas de Brodatz D105 e D106.

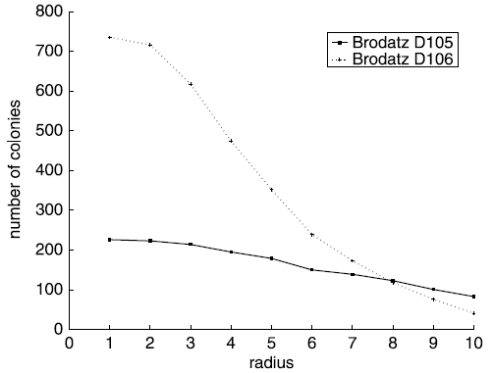


Figura 6: Curva de Formação de Colônias. Fonte: (Zhang & Chen, Artificial Life: A new approach to texture classification, 2005).

- Distribuição Escalar

A Distribuição Escalar das colônias é uma representação estatística das colônias formadas em diferentes escalas de tamanho. Texturas diferentes contém distribuição escalar diferentes. O número de AC em cada colônia formada ao fim do ciclo de vida deve ser contado. Toda colônia com λ indivíduos é chamada de Colônia-λ. O número de Colônias-λ é representada por gi(λ), onde $\{(\lambda, gi(\lambda))\epsilon\, R^2 :\ \lambda\, \epsilon\, R\}$. Por exemplo, se λ é o valor 2, gi(2) representa a quantidade de colônias com o rótulo Colônia-2, que são colônias formadas por 2 AC. A Figura 7 mostra um exemplo da curva de Formação de Colônias nas texturas de Brodatz D105 e D106.

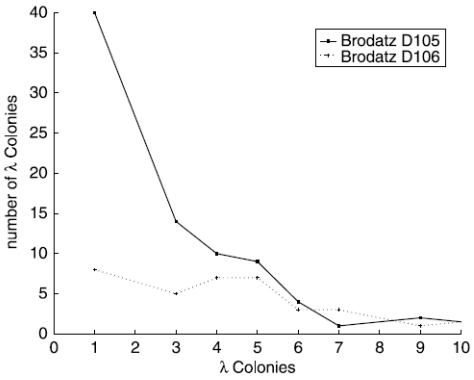


Figura 7: Distribuição Escalar. Fonte: (Zhang & Chen, Artificial Life: A new approach to texture classification, 2005).

2.5. Distâncias

Uma definição informal de distância pode dizer que ela é uma medida de separação entre dois pontos. Em Física, a distância percorrida por um corpo ao longo do seu movimento é a medida da linha de trajetória do corpo. A Distância Percorrida é uma grandeza escalar, que só pode tomar valores positivos ou nulos (Machado, 2014).

A distância também pode ser generalizada como o grau de dissimilaridade entre dois conjuntos. Em Matemática, sendo C um conjunto, uma função d : C x C -> R é chamada distância (ou dissimilaridade), em C se, para todo x,y ϵ C, ela contém as seguintes propriedades:

1. d(x,y) $\geq$ 0 (não negativo);

2. d(x,y) = d(y,x) (simetria);
3. d(x,y) = 0 (Deza & Deza, 2006).

Há várias formas de calcular distâncias e várias formas de interpretação. A seguir, algumas das distâncias mais conhecidas serão explicadas.

2.5.1. Distância Euclidiana

A distância Euclidiana é a distância mais utilizada comumente. Na maioria dos casos, quando as pessoas se referem a distância, elas estão se referindo a distância Euclidiana (Teknomo, 2006). Para entender a distância Euclidiana é necessário entender um pouco sobre o Teorema de Pitágoras.

O Teorema de Pitágoras é um dos teoremas mais antigos conhecido pelas civilizações antigas. O matemático e filósofo Pitágoras, que nomeou e criou o teorema, estabeleceu que "A área do quadrado construído em cima da hipotenusa de um triângulo é igual à soma das áreas dos quadrados construídos acima dos lados remanescentes" (Morris, 1997). A Figura 8 mostra o teorema de forma visual.

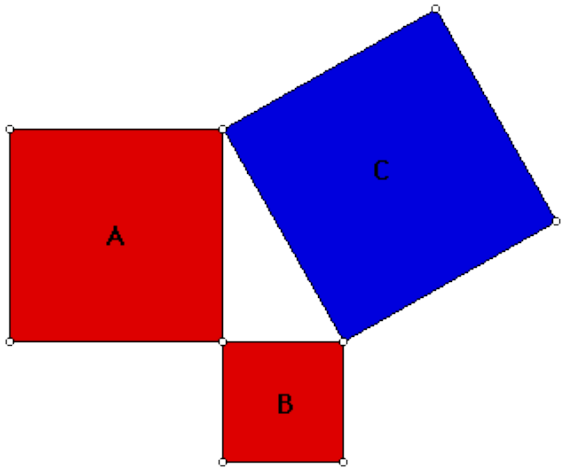


Figura 8: Teorema de Pitágoras. Fonte: (Morris, 1997).

Utilizando o Teorema de Pitágoras como base na Figura 8, implica-se que a soma das áreas dos quadrados vermelhos A e B é igual a área do quadrado azul C. Algebricamente, tendo a^2 = área de A, b^2 = área de B e c^2 = área de C, o Teorema de Pitágoras pode ser definido como

$$a^2 + b^2 = c^2 \quad (2)$$

sendo c o comprimento da hipotenusa do triângulo formado pelos quadrados e a e b os comprimentos dos catetos.

A consequência imediata do teorema é de que o quadrado do comprimento de um vetor $x = [x_1, x_2]$ é a soma dos quadrados de suas coordenadas e a distância quadrada entre dois vetores x e $y = [y_1, y_2]$ é a soma da diferença quadrática de suas coordenadas. Em outras palavras, a distância entre os vetores x e y pode ser denotada como $d_{x,y}$, que pode ser escrito como

$$d_{x,y}{}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2$$
$$= d_{x,y} \quad = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3)$$

De forma generalizada, a distância entre vetores j-dimensionais pode ser denotada como

$$d_{x,y} = \sqrt{\sum_{i=0}^{j} (x_i - y_i)^2} \quad (4)$$

Essa medida de comprimento é a distância generalizada que nos traz a noção de distância em duas ou três dimensões no espaço multidimensional é chamada de Distância Euclidiana (Greenacre & Primicerio, 2013).

2.5.2. Distância de Jaccard

A distância de Jaccard mensura a dissimilaridade entre dois conjuntos de dados. Paralelamente, o coeficiente de Jaccard mensura a similaridade entre dois conjuntos. Quanto maior a distância, menos esses conjuntos são similiares, e o inverso acontece com o coeficiente de Jaccard. Considerando dois conjuntos A e B, o coeficiente de Jaccard pode ser calculado como

$$J(A, B) = \frac{A \cap B}{A \; U \; B} \quad (5)$$

O valor do coeficiente está entre 0 e 1, onde quanto mais próximo a 1, mais similar os dois conjuntos são, e quanto mais próximo a 0, menos os dois conjuntos são similares.

A distância de Jaccard pode ser calculada como o complemento do coeficiente de Jaccard, ou seja,

$$d_J(A,B) = \frac{A\ U\ B - A \cap B}{A\ U\ B} = 1 - J(A,B) \quad (6)$$

onde $d_j(A,B)$ é a distância de Jaccard (Phillips, 2012).

2.5.3. Distância *Simple Matching*

Simple Matching é uma distância que, analogamente a distância de Jaccard, também mensura a dissimilaridade entre dois conjuntos de dados (Dekhtyar, 2009). Matematicamente, tendo X e Y como sendo conjuntos de dados com n elementos. A distância d_{sim} entre X e Y é definida como

$$d_{sim}(X,Y) = \sum_{j=1}^{n} \delta(x_j, y_j), \quad (7)$$

onde x_j é o j-ésimo elemento de X e y_j é o j-ésimo elemento de Y e

$$\delta(x_j, y_j) = \begin{cases} 0\ se\ x_j = y_j, \\ 1\ se\ x_j \neq y_j \end{cases} (\text{Gan}, 2011). \quad (8)$$

2.5.4. Distância de Chebychev

A distância de Chebyshev é definida em um espaço vetorial onde a distância entre dois vetores é a maior das diferenças de suas coordenadas. Ela é também conhecida como distância de valor máximo, pois ela descobre a magnitude absoluta da diferença entre as coordenadas de um par de objetos (Klove, Bergen, Lin, Tsai, & Tzeng, 2010).

Seja a distância de Chebyshev entre os vetores V e W ser $d_c(V,W)$, e suas coordenadas respetivas sendo V_i e W_i, $d_c(V,W)$ é calculada através da equação

$$d_c(V,W) = max_i(|V_i - W_i|) \quad (9)$$

2.5.5. Distância de Manhattan

Também conhecida como Geometria do Táxi, a distância de Manhattan computa a distância que deve ser percorrida entre um ponto e outro seguindo um caminho de "grade" (Black, 2006). Em outras palavras, a distância de Manhattan percorre um espaço em duas dimensões somente em linhas retas na vertical e na horizontal. A Figura 9 mostra a diferença entre a distância de Manhattan e a distância Euclidiana.

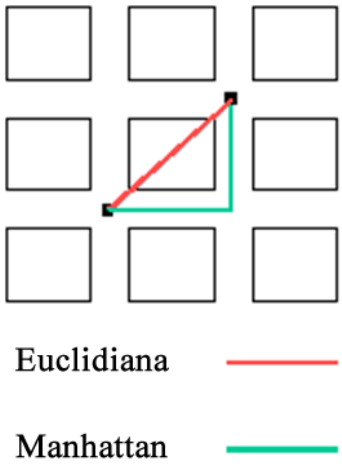


Figura 9: Distância de Manhattan vs Distância Euclidiana. Fonte: (The Best-Run Businesses Run SAP, 2014).

A distância de Manhattan entre dois itens é a soma das diferenças de seus correspondentes componentes. Para calcular a distância de Manhattan d_m entre os pontos X e Y com n dimensões é utilizada a fórmula

$$d_m = \sum_{i=1}^{n} |X_i - Y_i| \qquad (10)$$

2.6. *Rose Diagram*

Grupos de dados podem ser representados por histogramas lineares ou circulares. A diferença entre eles é que nos histogramas circulares, cada barra é iniciada no ponto central de um círculo e são correspondidos por grupos de ângulos. O tamanho de cada barra depende da frequência dos correspondentes ângulos. Já no histograma linear, as barras representam grupos de dados lineares e suas incidências.

O *Rose Diagram* é uma importante variante do histograma circular, onde as barras do histograma circular são substituídas por setores (Mardia & Jupp, 2006). A área de cada setor é proporcional a frequência do correspondente grupo. Para conseguir essa proporcionalidade quando grupos tem tamanho iguais, o raio de cada setor deve ser proporcional a raiz quadrada da frequência mais relevante. Ou seja, o maior setor vai determinar o tamanho dos setores menores.

A aparência do *Rose Diagram* assemelha-se com uma flor e seus setores com pétalas de rosas com tamanhos variados. Naturalmente, o número e setores presentes no gráfico do *Rose Diagram* é igual ao número de grupos determinado (Falta, 2011). A Figura 10 mostra um exemplo de dados representados em um *Rose Diagram*.

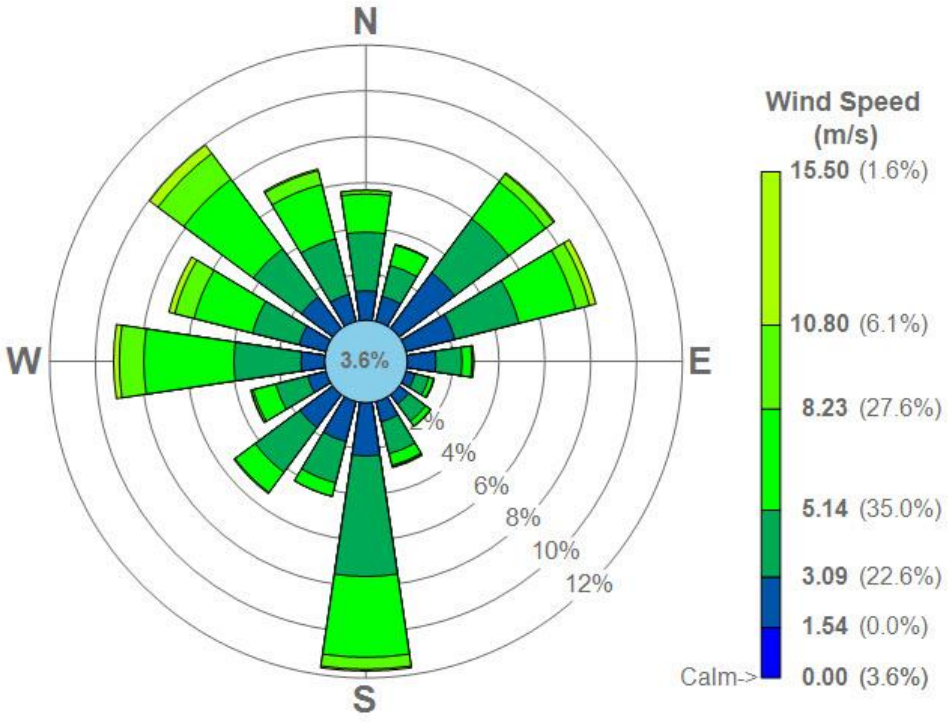


Figura 10: Rose Diagram representando as direções do vento. Fonte: (Wind rose plot for LaGuardia Airport, 2010).

Na Figura 10, cada setor representa ângulos de ocorrência de ventos em grupos de 22,5 graus. O maior setor representa a maior frequência de ventos no grupo de ângulos representado, e o menor setor mostra que houveram poucas ocorrências de ventos naqueles ângulos. As cores representam a velocidade dos ventos.

O conceito do gradiente de Sobel será descrito a seguir, pois será fundamental para entender o modelo proposto do *Rose Diagram* na metodologia. Além dele, algumas

medidas podem ser extraídas do *Rose Diagram*, e serão também descritas ao longo da seção.

2.6.1. Gradiente de Sobel

O operador de Sobel mensura o gradiente espacial 2D de uma imagem e enfatiza regiões de alta frequência espacial, que normalmente corresponde a bordas. Normalmente, ele é usado para encontrar a magnitude absoluta do gradiente a cada ponto de uma imagem em escala de cinza (Fisher, Perkins, A., & E., 2003).

Esse operador consiste de um par de *kernels* de convolução 3x3, onde um *kernel* é igual ao outro, mas rotacionado em 90 graus. A Figura 11 mostra os dois *kernels*.

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figura 11: Kernels do operador de Sobel. Gx é o *kernel* relacionado com o eixo x. Gy é o *kernel* relacionado com o eixo y. Fonte: (Fisher, Perkins, A., & E., 2003).

Esses *kernels* são feitos para maximizarem as bordas verticalmente (Gy) e horizontalmente (Gx). Eles podem ser aplicados separadamente na imagem de entrada, o que resultará em duas componentes de gradiente, uma em cada orientação. Quando combinados, é possível encontrar a magnitude do gradiente em cada ponto e sua orientação.

A magnitude do gradiente é dada por

$$|G| = \sqrt{Gx^2 + Gy^2} \tag{11}$$

onde $|G|$ corresponde a magnitude. O ângulo de orientação da borda, é calculado através da equação

$$\theta = arctan\left(\frac{Gy}{Gx}\right) \tag{12}$$

onde θ é o ângulo do gradiente espacial. A orientação com o valor 0 significa que a direção de máximo contraste, no caso do *kernel* Gx, ocorre da esquerda para a direita. No

caso do *kernel* Gy, caso a orientação do gradiente espacial seja 0, o máximo contraste ocorre de cima para baixo.

O valor da Equação 12 é o valor do gradiente de Sobel, que será utilizado nessa metodologia.

2.6.2. Direção Média

Uma medida que pode ser extraída do *Rose Diagram* é a direção média, que pode ser calculada através da equação

$$X = arctan\left(\frac{\sum_{i=1}^{n} sin\,\theta i}{\sum_{i=1}^{n} cos\,\theta i}\right) + \lambda \tag{13}$$

onde X é a direção média, θ é o ângulo representativo do setor, n é quantidade de ângulos representativos e λ é

$$\lambda = \begin{cases} \pi\ se\ \sum_{i=1}^{n} cos\,\theta\ i > 0 \\ 0\ se\ \sum_{i=1}^{n} sin\,\theta\ i > 0\ e\ \sum_{i=1}^{n} cos\,\theta\ i > 0 \end{cases} \tag{14}$$

2.6.3. Variância Circular

Outra medida que pode ser extraída é a variância circular, que é o complemento da direção média e pode ser calculada através da equação

$$S_0 = 1 - X \tag{15}$$

onde S_0 é a variância circular.

2.6.4. Desvio-Padrão Circular

Tendo a variância circular, o desvio-padrão circular s_0 pode ser obtido através da equação

$$s_0 = \sqrt{-2 * log_e(1 - S_0)} \tag{16}$$

2.6.5. Força do Vetor Resultante

A medida de força do vetor resultante *F* e calculada através da equação

$$F = \sqrt{\frac{\sum_{i=1}^{n} sin\,\theta i^2 + \sum_{i=1}^{n} cos\,\theta i^2}{n}} \quad (17)$$

2.6.6. Assimetria e Curtose

Por fim, pode-se extrair as medidas curtose e assimetria. A curtose *K* pode ser calculada através da equação

$$K = \frac{1}{n} * \sum_{i=1}^{n} cos\,2 * (\theta i - X) \quad (18)$$

e a assimetria *A* pode ser calculada através da equação

$$A = \frac{1}{n} * \sum_{i=1}^{n} sin\,2 * (\theta i - X) \quad (19)$$

2.7. Máquina de Vetor de Suporte

O classificador Máquina de Vetor de Suporte (MVS) é um grupo de métodos de aprendizado supervisionado que podem ser usados para classificação ou regressão. O algoritmo do MVS é baseado na teoria de aprendizado estatístico e na dimensão de Vapnik-Chervonenkis (Ivanciuc, 2005).

A principal característica no grupo de métodos do MVS consiste no fato deles minimizarem o erro empírico de classificação, maximizando ao mesmo tempo a margem geométrica de erro (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

A técnica consiste no mapeamento de vetores de entrada em um espaço amostral com dimensão superior, por meio de construção do hiperplano máximo de separação. Para isso, dois hiperplanos auxiliares são construídos de uma forma que o hiperplano de separação resultante maximize a distância entre eles. É assumido que, quanto maior for a margem de distância entre os planos, menor é o erro de generalização do classificador. A Figura 11 ilustra os hiperplanos.

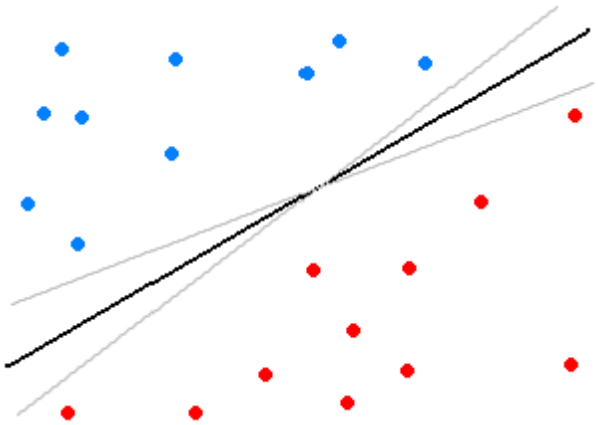


Figura 12: Espaço separado pelos hiperplanos. Os pontos vermelhos e azuis representam classes distintas e as retas hiperplanos. Fonte: (Netto, 2010).

O MVS classifica os padrões de entrada representados por um vetor de pesos ρ-dimensional em diversas classes, com o objetivo de separar tais classes com um hiperplano de dimensão ρ−1. Além disso, a técnica visa encontrar um hiperplano que possibilita a maior separação entre as classes – ou seja, encontrar um, cuja distância até o padrão atual de entrada seja máxima, conhecido como hiperplano de margem máxima.

Considerando $X \in R_0 \subseteq \Re_n$ os vetores de entrada, $y \in \{-1,+1\}$ os eixos, e $\varphi : R_0 \rightarrow F$ a função de mapeamento do espaço de entrada para o espaço de apresentação, o algoritmo visa encontrar um hiperplano (w, b) tal que o valor da Equação 20 seja máximo, sendo γ a margem, vetor w com a mesma dimensão de F, e b um número real.

$$\gamma = min_i y_i \{\langle w, \Phi(X_i)\rangle - b\} \tag{20}$$

A função de decisão *f(x)* correspondente é obtida através da equação

$$f(x) = sign(\langle w, \Phi(X)\rangle - b) \tag{21}$$

O mínimo da função *f(x)* ocorre quando a Equação 20 é satisfeita.

$$w = \sum_i \alpha_i y_i \Phi(X_i) \tag{22}$$

sendo que o parâmetro α_i assume valores reais positivos que maximizam a equação

$$\sum_i \alpha_i - \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \Phi(X_i), \Phi(X_j)\rangle \tag{23}$$

e satisfazem a seguinte Equação 22:

$$\sum \alpha_i y_i = 0, \alpha_i > 0 \quad (24)$$

Assim, é possível colocar a função de decisão da Equação 19 da forma da equação

$$f(X) = sign(\sum_i \alpha_i y_i \langle (X_i), \Phi(X) \rangle - b) \quad (25)$$

É importante observar que somente um subconjunto de pontos é associado a αi não-nulos. Tais pontos são denominados de vetores de suporte e são os pontos mais próximos ao hiperplano de separação.

Dessa forma, a técnica possibilita a classificação de padrões de acesso, separando-os de acordo com os vetores de suporte do hiperplano determinados.

2.8. Validação de Resultados

Uma das maiores dificuldades ao propor técnicas ou modelos para o processamento de imagens digitais é a mensuração dos resultados (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

Fazer reconhecimento de padrão é um processo aproximado que resulta mais em probabilidade de se estar certo do que em certeza. Embora possam existir várias medidas para verificar o desempenho de um classificador qualquer, a medida mais importante é o desempenho do mesmo a partir da classificação de novos casos. Nos problemas ligados à área de saúde, a estrutura básica dos testes de classificação é determinar quão bem um teste discrimina a presença ou ausência de uma doença (Netto, 2010).

Quando se avalia um teste de classificação em relação a presença ou ausência de doenças em pacientes, quatro são os possíveis cenários:

1. O teste é positivo e o paciente tem a doença - Verdadeiro Positivo (VP);
2. O teste é positivo, mas o paciente não tem a doença - Falso Positivo (FP);
3. O teste é negativo e o paciente tem a doença - Falso Negativo (FN) e
4. O teste é negativo e o paciente não tem a doença - Verdadeiro Negativo (VN).

Considere-se doença neste trabalho a presença ou não de nódulo pulmonar. O resultado VP, por exemplo, indica que a classificação resultou em positivo e a estrutura analisada é de fato um nódulo.

Para avaliar o desempenho de modelos de classificação, geralmente utilizam-se algumas estatísticas descritivas, como Sensibilidade (Sen), Especificidade (Esp) e Acurácia (Acc).

A sensibilidade é a proporção de verdadeiros positivos, ou seja, a capacidade da metodologia em predizer corretamente a condição para casos que realmente ocorre ela. A Equação 24 mostra o cálculo da sensibilidade Sen.

$$Sen = \frac{VP}{VP+FN} \tag{26}$$

A especificidade é a proporção de verdadeiros negativos, ou seja, a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não ocorrem ela. A Equação 25 mostra o cálculo da especificidade Esp.

$$Esp = \frac{VN}{VN+FP} \tag{27}$$

A acurácia é a proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é muito suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho da metodologia. A Equação 26 mostra como é calculada a acurácia Acc.

$$Acc = \frac{VP+VN}{VP+VN+FP+FN} \tag{28}$$

2.8.1. Análise de Desempenho através da curva ROC

Uma forma bastante utilizada de avaliar o desempenho quantitativo de uma determinada técnica ou modelo proposto pela comunidade científica são as chamadas curvas *Receiver Operationg Characteristic* (ROC) (van Erkel & Pattynama, 1998). Essa forma de avaliação foi desenvolvida com bases estatísticas onde a principal característica é a relação percentual de acertos. O ideal para um sistema é que a quantidade de acertos

no diagnóstico tanto em fatores VP quanto em VN sejam máximas, conceitos definidos como sensibilidade e especificidade.

Do ponto de vista da análise das curvas ROC, isso significa que o ponto ótimo da curva é definido pelo extremo superior esquerdo que corresponde a máxima sensibilidade e mínimo número de FP, ou seja, máximo índice de VN. Consequentemente, quanto mais próximo de 1 (equivalente a 100%) a área sob a curva, melhor desempenho alcançado (de Carvalho Filho, Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs, 2013).

3. METODOLOGIA

Neste capítulo a metodologia proposta é descrita detalhadamente. Sua finalidade é a classificação de nódulos pulmonares em nódulo ou não nódulo. A metodologia consiste de quatro etapas principais. A primeira é a aquisição de imagens, que mostra como a base LIDC-IRDC será explorada no trabalho. A segunda é a etapa da extração de características, que mostra os métodos do *Rose Diagram* e *Artificial Crawlers* em ação, e como serão extraídas características da base de imagens utilizando esses dois artifícios. A terceira é a classificação, onde o classificador MVS será utilizado para obtenção do resultado utilizando as medidas extraídas na segunda etapa utilizando cada método separadamente e juntos. Por fim, a quarta e última etapa é a validação da classificação da terceira etapa, onde serão utilizados métodos para medir a influência das medidas extraídas na segunda etapa e eficácia da classificação da terceira etapa. A Figura 12 mostra um resumo da metodologia proposta e o fluxo de forma visual.

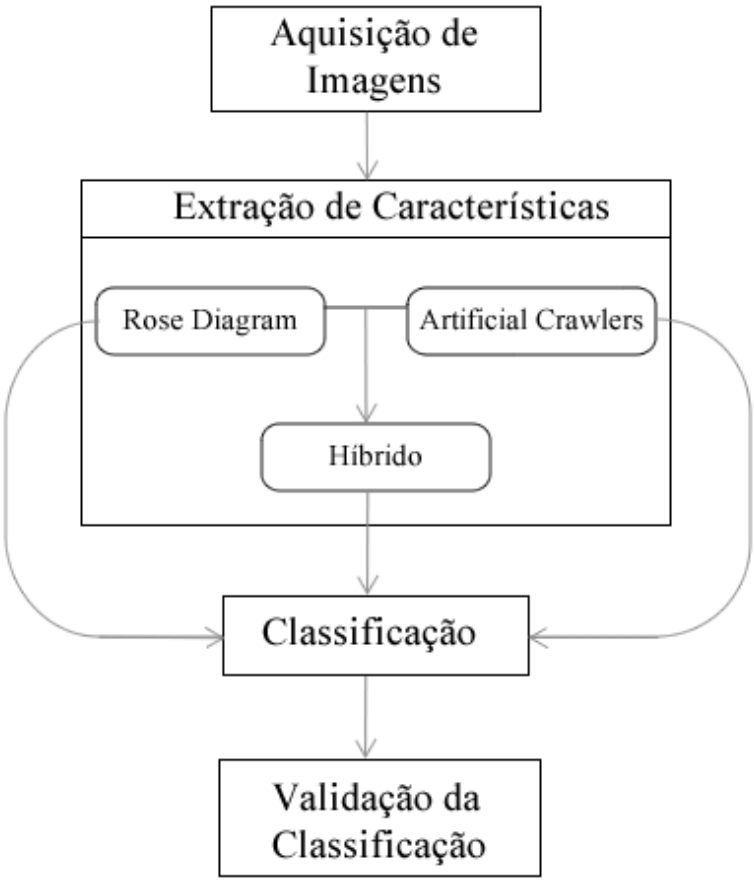


Figura 13: Metodologia proposta.

3.1. Aquisição de Imagens

As imagens utilizadas neste trabalho são provenientes da base pública LIDC-IRDC. O principal motivo da escolha dessa base é o fato dela ser acessível *online* em (LIDC-IDRI, 2012). Além disso, por se tratar de uma base pública, ela segue alguns princípios, como a proteção das identidades dos pacientes, assim como o consentimento dos mesmos para a disponibilização dos exames para pesquisas e treinamento de diagnósticos de câncer de pulmão. Outra vantagem do uso dessa base pública, é a grande quantidade de exames disponíveis. Por fim, as vantagens em ser pública, ser bem acessível e ter uma boa quantidade de exames facilita a validação da metodologia proposta por terceiros.

Esta base contém 1018 séries de exames em imagens de TC e 290 séries de exames em imagens de radiografias computadorizadas e digitais. Cada exame contém informações sobre os nódulos presentes nos pacientes. Encontram-se diversos tipos de nódulos, dentre eles há pequenos (diâmetro maior que 3 e menor que 10 mm), conectados a vasos sanguíneos, justapleurais e nódulos grandes (diâmetro maior que 10 e menor que 30 mm), além da presença de massas, que são estruturas com diâmetro maior que 30 mm.

Cada arquivo contém marcações feitas por quatro especialistas, indicando o tamanho e posição dos nódulos. Essas marcações são encontradas apenas em nódulos que possuem o diâmetro maior que 3 mm e menor que 30 mm. Para nódulos menores é encontrada apenas a marcação referente ao centro de massa dele.

Mesmo contendo 1018 séries de imagem em TC, a base toda não foi utilizada. O primeiro motivo disso é que alguns exames continham nódulos menores que 3mm, e sua marcação era somente o centro de massa. O segundo motivo é que alguns arquivos DICOM possuem informações, em seus cabeçalhos, diferentes das informações cedidas dos exames. Para evitar erros na validação da metodologia, esses arquivos foram excluídos, restando 833 exames em TC.

Depois de escolhidos os arquivos dos exames, é necessário separar as informações para as etapas seguintes. Uma nova base de imagens será criada baseada nas imagens dos exames do LIDC-IRDC: uma base de candidato a nódulos e uma base de candidatos a não-nódulos.

3.1.1. Base de candidatos a nódulos

Cada exame que será utilizado contém marcações de quatro especialistas, referentes as localizações dos nódulos. Um exame de um paciente pode conter vários nódulos na opinião dos especialistas, mesmo quando os mesmos não concordam em relação a quantidade. Por exemplo, o primeiro especialista pode ter diagnosticado dois nódulos a mais do que o segundo especialista. Todas essas informações estão contidas no arquivo de informações dos exames.

Com base nesse arquivo, de cada exame é extraído somente o local onde os especialistas marcaram os nódulos, resultando em diversas novas imagens. Essas imagens contém os candidatos a nódulos baseados nas marcações dos especialistas. Se um exame contém, por exemplo, duas marcações por especialistas, cada marcação será um candidato a nódulo, então desse exame serão usados oito candidatos a nódulos. Assim os nódulos das marcações dos quatro especialistas serão todos utilizados como candidatos.

A Figura 14 mostra exemplos de fatias de candidatos a nódulos que vão compor a base.

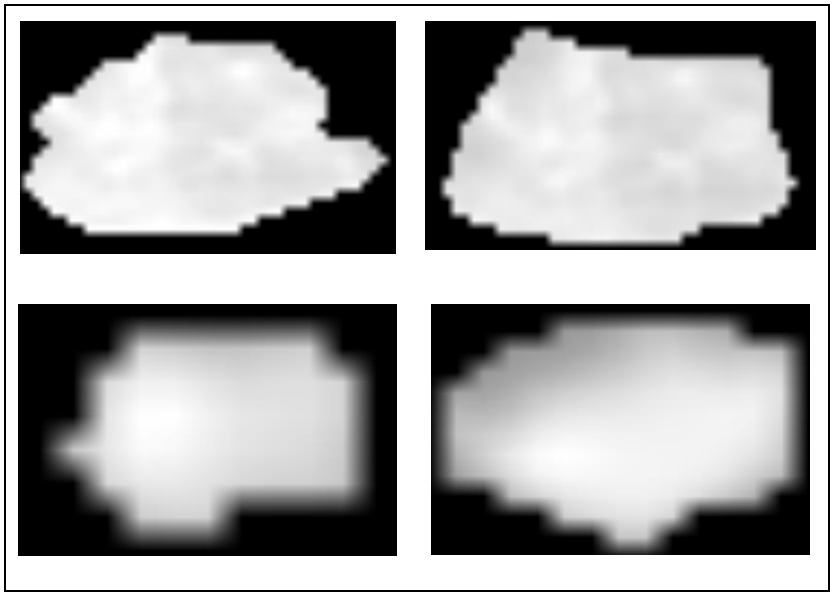


Figura 14: Fatias de Candidatos a Nódulos.

3.1.2. Base de candidatos a não-nódulos

Para a base de candidatos a não-nódulos foi utilizada a metodologia do Filho *et al*. 2013, que utiliza o algoritmo *Quality Threshold* (QT) seguido do algoritmo de crescimento de região para segmentar nos exames do LIDC-IRDC candidatos a nódulos

e a não-nódulos. Posteriormente, os candidatos a não-nódulos são excluídos da metodologia dele, que tem por objetivo a detecção de nódulos. Esses candidatos são utilizados no trabalho proposto como uma nova base de imagens de não-nódulos.

A Figura 15 mostra algumas fatias de candidatos a não-nódulos originadas desse método.

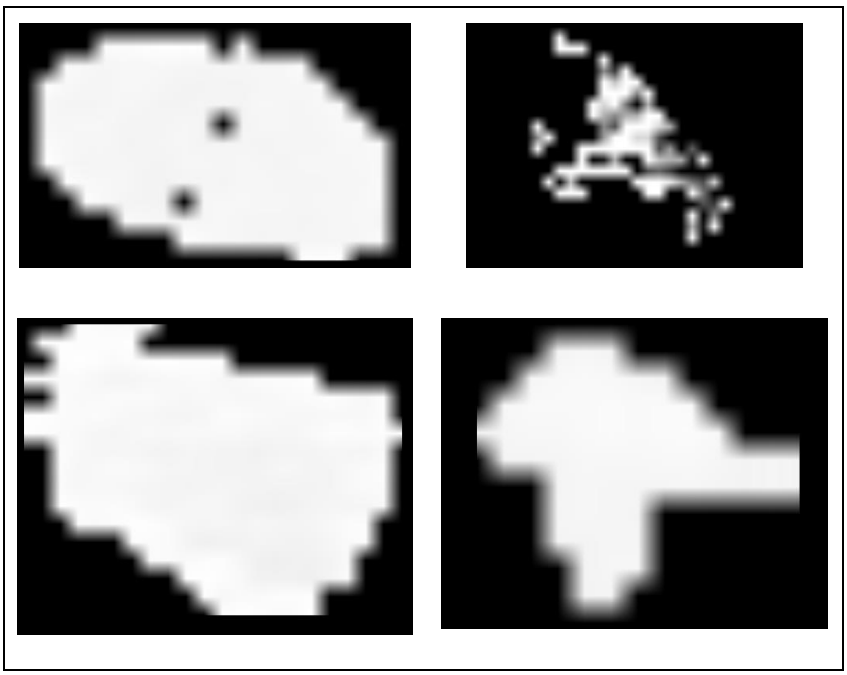


Figura 15: Fatias de Candidatos a Não-Nódulos.

O motivo de não utilizar os nódulos segmentados por Filho *et al*. 2013 como candidatos a nódulos nesse trabalho é que a quantidade de nódulos advindas dessa técnica é menor do que a quantidade de nódulos extraídos das marcações dos especialistas. Uma quantidade muito pequena de candidatos a nódulos pode comprometer a classificação posterior, já que a quantidade de candidatos a não-nódulos tende a ser muito maior na metodologia de Filho *et al*. 2013, desequilibrando a quantidade de amostras e viciando o classificador.

3.2. Extração de Características

Nesta etapa ocorre a extração das características necessárias para a classificação da etapa seguinte, utilizando as bases de imagens da etapa anterior. A metodologia proposta utiliza três modelos de extração de características. O primeiro modelo faz o uso

do algoritmo evolutivo *Artificial Crawler* para analisar a textura das imagens, onde de cada imagem serão extraídas curvas de evolução, mostradas na Seção 2.4.2. O segundo modelo utiliza uma análise estatística do *Rose Diagram* sobre as imagens baseada nos gradientes locais, descrita na Seção 2.6. O terceiro é um modelo híbrido, uma junção das características dos dois modelos anteriores em um só.

A seguir, os modelos são descritos em detalhes, com exemplos baseados nas imagens das bases de nódulos e não-nódulos.

3.2.1. Características no modelo *Artificial Crawlers*

O algoritmo *Artificial Crawlers* (AC) recebe como entrada os candidatos a nódulos e não-nódulos para a análise de textura e extração das curvas de evolução. Para isso, alguns passos devem ser seguidos. A Figura 16 mostra um diagrama de fluxo e os passos do modelo proposto.

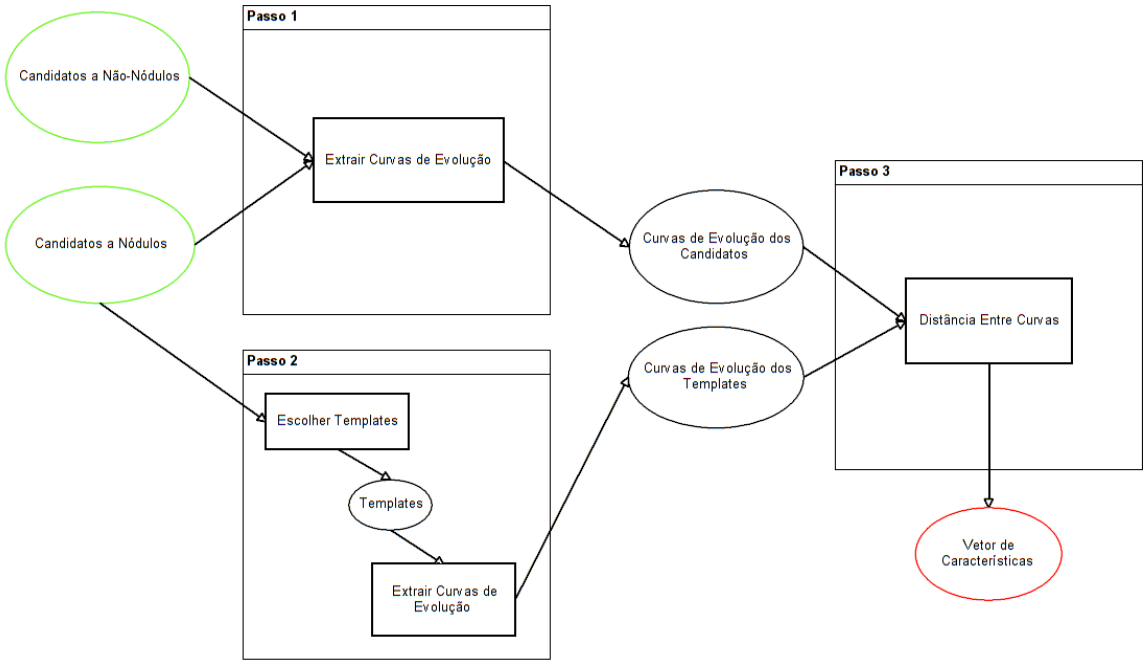


Figura 16: Diagrama de fluxo do modelo AC. Elipses representam dados de entrada/saída e retângulos representam métodos. As entradas do fluxo estão destacadas na cor verde e a saída na cor vermelha.

O primeiro passo (Passo 1) do modelo AC é utilizar as bases de candidatos a nódulos e não-nódulos como ambientes de vida dos organismos do AC na etapa Extrair

Curvas de Evolução. Após essa etapa, as curvas de evolução do ciclo de vida do AC de cada imagem serão adquiridas.

O segundo passo (Passo 2) é a escolha de *templates* que representam a base de imagens de candidatos a nódulos. Esses *templates* são candidatos a nódulos que serão de auxílio posteriormente para realizar a diferença entre curvas. Após essa etapa, os *templates* são escolhidos e passam também pelo processo de extração de curvas de evolução.

O terceiro passo (Passo 3) é utilizar as distâncias estudadas na Seção 2.5 para calcular a distância entre as curvas de evolução dos candidatos com as curvas de evolução dos *templates*. Para cada *template*, o vetor de características da saída desse passo contém pelo menos cinco características para cada uma das quatro curvas de evolução, totalizando vinte características. Adicionalmente, as áreas das quatro curvas das imagens dos candidatos também são adicionadas ao vetor de características e independe dos *templates*. A quantidade de características ao fim dessa etapa é

$$\mathrm{Qt}_{\mathrm{Características}} = \mathrm{Qt}_{\mathrm{Curvas}} * \mathrm{Qt}_{\mathrm{Distâncias}} * \mathrm{Qt}_{\mathrm{Templates}} + \mathrm{Qt}_{\mathrm{Áreas}} \quad (29)$$

onde $\mathrm{Qt}_{\mathrm{Características}}$ é a quantidade de características no vetor de característica que representa a imagem do candidato, $\mathrm{Qt}_{\mathrm{Curvas}}$ é a quantidade de curvas de evolução, $\mathrm{Qt}_{\mathrm{Distâncias}}$ é a quantidade de distâncias usadas na comparação com os *templates*, $\mathrm{Qt}_{\mathrm{Templates}}$ é o número de *templates* escolhidos e por fim $\mathrm{QT}_{\mathrm{Áreas}}$ é a quantidade de áreas. É possível notar que a quantidade de áreas é igual à quantidade de curvas, já que para cada curva de evolução é possível extrair somente sua área. Então é possível afirmar que $\mathrm{QT}_{\mathrm{Áreas}}$ é igual a $\mathrm{QT}_{\mathrm{Curvas}}$.

As etapas serão descritas com mais detalhes nas seções seguintes.

3.2.1.1. Extrair Curvas de Evolução

Essa etapa inicia com uma imagem de candidato e ela é utilizada como o ambiente de vida dos AC. Na metodologia, cada imagem é um nódulo ou um não-nódulo.

Alguns passos do algoritmo AC foram adaptados para receber as imagens dos candidatos. No algoritmo original, descrito na Seção 2.4.2, cada indivíduo AC percebe somente seus 8-vizinhos em um ambiente 2D. Nesse trabalho, a percepção dos indivíduos

é adaptada para o ambiente 3D, onde eles percebem seus 26-vizinhos (nos eixos x, y e z). O indivíduo pode andar nas direções de todos os vizinhos. A Figura 17 mostra a percepção 2D do método original e percepção 3D proposta na metodologia.

AC 2D AC 3D

Figura 17: AC com a percepção 2D e 3D. O cubo preto representa o AC. Os cubos transparentes representam os vizinhos.

Outro passo adaptado é a escolha das posições iniciais dos AC na imagem. No algoritmo original, são escolhidos *pixels* aleatórios em um número arbitrário, escolhido pelo usuário, da quantidade de indivíduos AC iniciais. Na metodologia proposta, a quantidade de indivíduos iniciais é adaptativa: cada *voxel* das imagens de nódulos/não-nódulos são analisados, e se fizerem parte desse nódulo/não-nódulo, ou seja, esse *voxel* representa um valor diferente do fundo da imagem, um indivíduo AC será alocado nesse *voxel*. A vantagem de utilizar os *voxels* que não são do fundo da imagem como locais de nascimento dos indivíduos AC é que isso elimina a possibilidade de um AC nascer em um local fora do nódulo/não-nódulo, o que ocasionaria indivíduos que não se movimentariam, já que estão fora da textura dos nódulos ou não-nódulos. Além dessa vantagem, há o fato de não utilizar posições aleatórias, que caso fossem utilizadas, os resultados das curvas de evoluções ao término do algoritmo seriam sempre diferentes.

Com as modificações concluídas, o ciclo de vida do AC começa. A quantidade de ciclos depende diretamente do tamanho e da textura de cada imagem. Ao fim do ciclo, restam indivíduos AC sobreviventes, exemplificados na Figura 18 em três fatias de um nódulo.

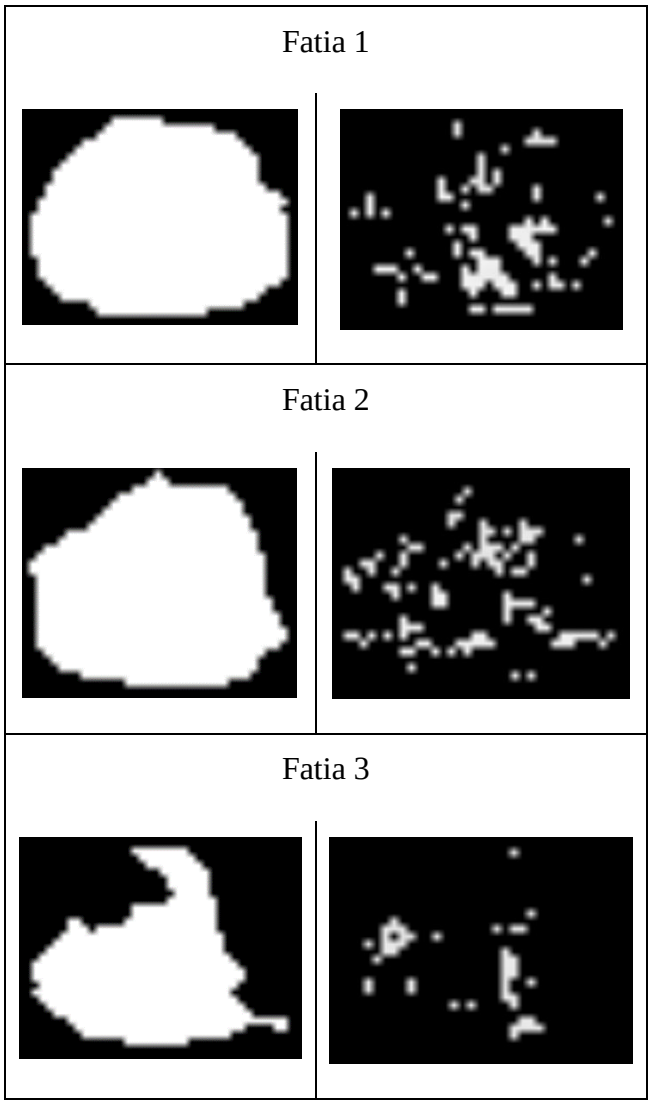


Figura 18: Populações de AC. (a) População inicial. (b) População final.

Após o fim do ciclo, temos as quatro curvas de evolução: curva de Evolução de Agentes, curva de Assentamento de Habitantes, curva de Formação de Colônias e curva de Distribuição Espacial.

3.2.1.2. Escolher *Templates*

Essa etapa é responsável por escolher quais candidatos a nódulos serão utilizados como *templates*. Serão utilizados somente *templates* de nódulos, pois os nódulos são considerados uma classe distinta dos não-nódulos e para realizar a etapa de distância entre curvas é necessário *templates* de uma das classes para diferenciar essa classe da outra

classe. Os nódulos foram escolhidos como representantes por que sua quantidade é bem inferior a quantidade de não-nódulos, o que não ocasionaria nenhum vício na classificação.

A metodologia propõe o uso de mais de um *template* para a extração de características, pois a base de candidatos a nódulos tem nódulos de tamanhos muito variados. Apenas um nódulo não é suficiente para representar a base toda. Com isso foram escolhidos três *templates*: um nódulo pequeno (entre 3mm e 10mm), um nódulo médio (entre 11mm e 20mm) e um nódulo grande (entre 21mm e 30mm).

Após escolhidos os *templates*, eles passam pelo mesmo processo do resto das bases de candidatos, sendo extraídas as curvas de evolução deles também.

3.2.1.3. Distância Entre Curvas

Tendo as curvas de evolução dos candidatos a nódulo e não-nódulos de um lado e as curvas de evolução dos *templates* do outro, essa etapa é responsável pela criação do vetor de características. Essas características são advindas de duas sub-etapas: a diferença entre curvas e a área das curvas de evolução dos candidatos.

Para realizar a diferença entre curvas, os pontos delas são utilizados para calcular as distâncias entre elas. Estudadas na Seção 2.5, são utilizadas: distância Euclidiana, distância de Jaccard, distância *Simple Matching*, distância de Chebychev e distância de Manhattan. Quanto maior a diferença entre as curvas dos candidatos e as curvas dos *templates*, maior a probabilidade desse candidato ser um não-nódulo.

Por convenção, todas as curvas de evolução devem ter o mesmo tamanho no eixo y. Devido ao tamanho variado das imagens dos candidatos, é necessário normalizar as curvas no eixo y entre valores fixo. Essa normalização foi feita entre os valores de 0 e 1. Essa etapa é muito importante para o cálculo da diferença entre as curvas de evolução, já as curvas de uma mesma classe tendem a ter uma forma semelhante, então para fazer a diferença entre elas, elas devem estar na mesma escala.

Após a normalização, são calculadas as cinco distâncias entre as curvas de evolução do candidato e as curvas de evolução de todos os *templates*.

A segunda sub-etapa é o cálculo da área das curvas de evolução dos candidatos. Para isso calcular a área de uma curva em um plano cartesiano é utilizada a Integral da função da curva (Bourne, 2014).

3.2.2. Características no modelo *Rose Diagram*

O modelo proposto do *Rose Diagram* (RD) recebe como entrada os candidatos a nódulo e não-nódulos e utiliza o gradiente de Sobel e o RD para extrair características da textura. A Figura 19 mostra o diagrama do fluxo do modelo RD.

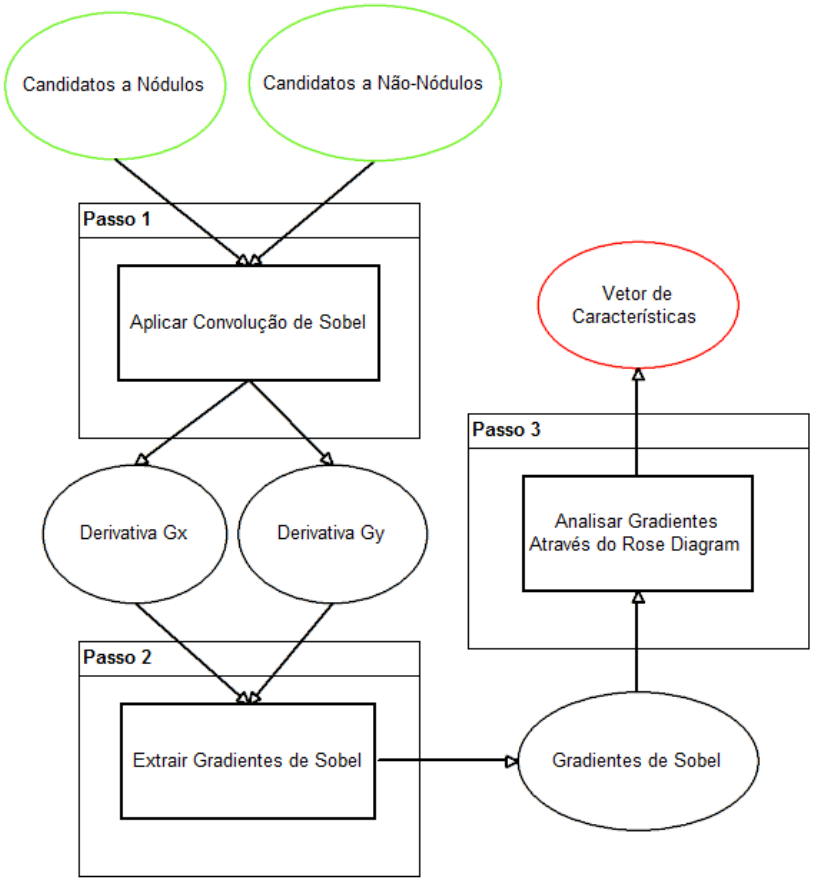


Figura 19: Diagrama de fluxo do modelo RD. Elipses representam dados de entrada/saída e retângulos representam métodos. As entradas do fluxo estão destacadas na cor verde e a saída na cor vermelha.

O primeiro passo (Passo 1) inicia com as imagens dos candidatos a nódulos e não-nódulos sendo convoluídas, utilizando o *kernel* do operador de Sobel. Ao fim dessa etapa,

são obtidas as imagens derivativas dos candidatos, chamadas de Gx (operador na horizontal) e Gy (operador na vertical).

O segundo passo (Passo 2) utiliza as imagens derivativas para extrair os gradientes. Ao fim tem-se os gradientes de Sobel do candidato.

O terceiro passo (Passo 3) utiliza os gradientes de Sobel extraídos dos candidatos e os analisa sobre a perspectiva do RD. Ao finalizar esse passo tem-se as características estatísticas extraídas do RD.

As etapas dos três passos serão detalhadas a seguir.

3.2.2.1. Aplicar Convolução de Sobel

Essa etapa aplica os operadores de Sobel estudados na Seção 2.6.1 em cada imagem de entrada. Ao fazer isso, a imagem convoluída, ou seja, a imagem que sofreu influência dos operadores, serão duas: a derivativa Gx que corresponde ao operador de Sobel horizontal e a derivativa Gy que corresponde ao operador de Sobel vertical. Para cada imagem de um candidato, um par de derivativas D_{par}(Gx, Gy) é gerada. Essas duas imagens resultantes são necessárias para a etapa de extração de gradientes.

3.2.2.2. Extrair Gradientes de Sobel

As imagens derivativas são utilizadas em conjunto nesta etapa. Em outras palavras, um par de derivativas D_{par}(Gx, Gy) é utilizado para se adquirir os gradientes de sobel. Utilizando a Equação 12 é possível calcular os gradientes providos pelas derivativas. Esses gradientes correspondem ao ângulo do gradiente de Sobel em cada um dos pontos da imagem do candidato.

A quantidade de gradientes é equivalente ao tamanho dos nódulos. Para cada fatia do candidato, um grupo de gradientes é extraído. Todos os grupos de todas as fatias serão utilizados no Passo 3.

3.2.2.3. Analisar Gradientes através do *Rose Diagram*

Tendo um conjunto de gradientes de Sobel, ou seja, um conjunto de ângulos, eles são analisados no diagrama circular RD. A Figura 20 mostra um exemplo de uma fatia de um candidato a nódulo tendo seus gradientes colocados em um RD.

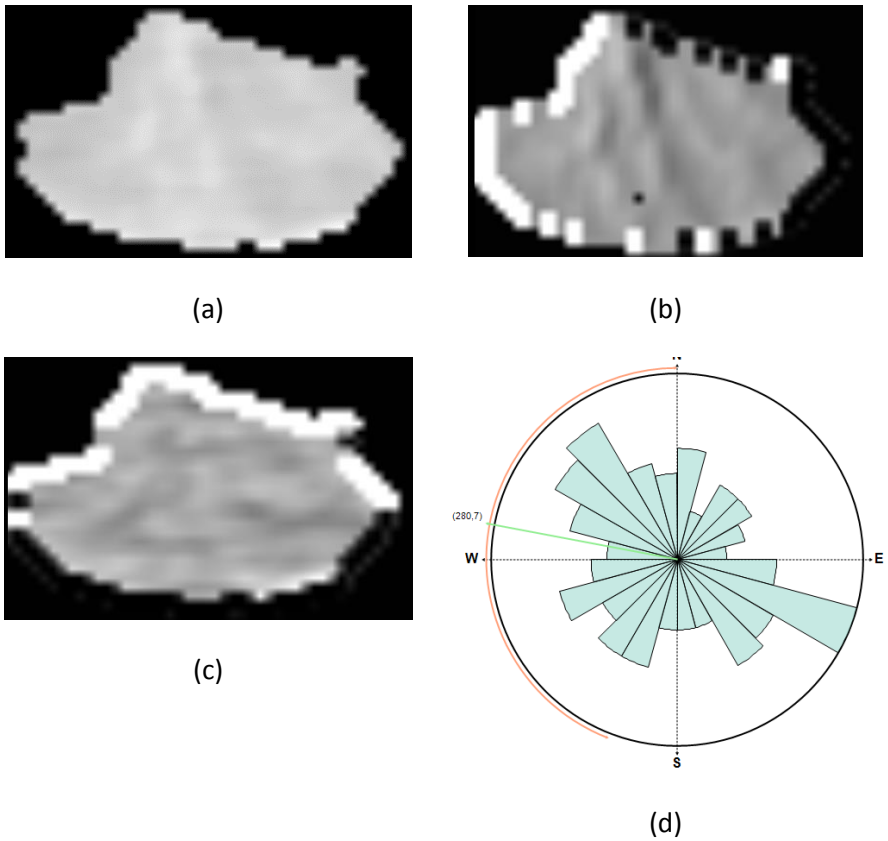


Figura 20: *Rose Diagram* de uma fatia de nódulo. (a) é a fatia original do nódulo (b) é a derivativa Gx (c) é a derivativa Gy e (d) é o *Rose Diagram* com os gradientes.

Na Figura 20, os gradientes são organizados no RD em 24 setores que agrupam ângulos em um alcance de 15 graus. A maior ocorrência dos gradientes acontece entre os ângulos de 105 e 120 graus. A escolha de 24 setores se dá de forma empírica. Dos testes feitos com esse modelo, o intervalo de 15 graus foi o que deu melhores resultados.

Através do RD gerado, é possível extrair as medidas estudadas na Seção 2.6: direção média, variância circular, desvio-padrão circular, força do vetor resultante, assimetria e curtose. Essas medidas serão utilizadas como características e assim, é gerado o vetor de características representante do modelo RD.

3.2.3. Características no modelo Híbrido

O modelo Híbrido é uma representação dos vetores de características originados dos modelos AC e RD e unidos em um. Ou seja, a quantidade de características do vetor resultante é a soma das quantidades dos vetores dos modelos anteriores.

O objetivo deste modelo é avaliar posteriormente se os modelos AC e RD tem melhores resultados juntos ou separadamente.

A Tabela 2 mostra um resumo das medidas que são utilizadas em cada modelo.

Tabela 2: Medidas dos modelos *Artificial Crawlers* e *Rose Diagram*.

	ARTIFICIAL CRAWLERS	**ROSE DIAGRAM**
MEDIDAS	Distâncias entre Curvas de Evolução de Agentes	Direção Média
	Distâncias entre Curvas de Assentamento de Habitantes	Variância Circular
	Distâncias entre Curvas de Formação de Colônias	Desvio-Padrão Circular
	Distâncias entre Curvas de Distribuição Escalar	Força do Vetor Resultante
	Áreas das Curvas de Evolução	Assimetria
		Curtose

A Tabela 2 resume as medidas utilizadas em ambos os modelos. No modelo AC, são utilizadas uma área de cada curva de evolução, totalizando quatro medidas de área. A quantidade de distâncias utilizadas como medidas depende da quantidade de *templates*. Foram adotados três *templates*, como citado na Seção 3.2.1.2, então o total de medidas no modelo AC, utilizando a Equação 29, é de 64 medidas. No modelo RD são utilizadas as 6 medidas presentes na Tabela 2. Então o modelo híbrido utiliza o total de 70 medidas de textura para a classificação.

3.3. Classificação

Esta etapa consiste na classificação dos candidatos a nódulos e não-nódulos. Os nódulos e os não-nódulos são tratados como duas classes de dados distintas. Cada nódulo e não-nódulo corresponde a um vetor de características e uma classe, e serão utilizados na classificação.

O classificador escolhido para esta etapa é o MVS, pela sua robustez e flexibilidade, já que é um classificador baseado em *kernel* e pode classificar dados de forma linear ou não-linear de acordo com o uso de quem o aplica. A vantagem é que ele consegue capturar relacionamentos muito complexos entre os dados sem fazer transformações muito difíceis (Lamp, 2012). Na metodologia proposta, o MVS deve receber os vetores de características das duas classes (nódulo e não-nódulo) e classifica-las.

A ferramenta selecionada para a utilização do MVS foi a biblioteca de licença livre *Weka* 3.6, disponível em (Frank, Hall, Reutemann, & Trigg, 1993). O *kernel* utilizado é o de base radial, o que necessita uma boa estimativa dos parâmetros *c* e *gamma* do MVS. Esses parâmetros são estimados através do algoritmo *Grid Search* (GS) (Bergstra & Bengio, 2012), presente na biblioteca LIBMVS, que está contida no *Weka* 3.6, através do *script* grid.py.

3.4. Validação da Classificação

Essa etapa é importante para a validação da metodologia proposta e discussão de resultados, após a classificação. A metodologia usa medidas comumente aplicadas em sistemas de diagnósticos de processamento de imagens, para a análise de performance. As medidas são a sensibilidade, especificidade e acurácia, descritas na Seção 2.8. Com a intenção de analisar de forma mais aprofundada os resultados da metodologia proposta, é calculado o coeficiente de variação, que calcula a dispersão das amostras de acordo com a sua média (Soliman, Ellah, Abou-Elheggag, & Abd-Elmougod, 2012).

Outra forma que será utilizada para mensurar a performance é utilizando a curva ROC, descrita na Seção 2.8.1. Uma curva ROC indica a taxa de verdadeiro positivo (sensibilidade) em função da taxa de falso positivo (1 - especificidade).

4. RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos nos testes da metodologia proposta para a classificação de nódulos pulmonares. A estratégia para a realização dos testes segue a seguinte forma: No primeiro momento, a aquisição de imagens usadas para treinar e testar a metodologia é realizada; Depois é analisado o processo de extração de características; Em seguida, é feita a validação da classificação para todas as proporções de testes, sendo realizados através da média da acurácia, média da sensibilidade, média da especificidade, variação de coeficiente e curva ROC; Por fim é feita uma análise comparativa com os trabalhos relacionados.

4.1. Aquisição de Imagens

As imagens utilizadas para testar e validar a metodologia são retiradas da base LIDC-IRDC. O total de 833 exames foram utilizados para os testes, sendo essa base dividida em quatro grupos distintos de treino/teste (respectivamente) de 20/80%, 40/60%, 60/40% e 80/20%. Para cada um dos grupos, as amostras foram escolhidas aleatoriamente para treino e teste.

São realizadas cinco classificações para cada grupo, sendo que há um revezamento nas amostras de treino e teste. Por exemplo, no grupo de 20/80% serão realizadas cinco classificações, onde tanto a amostra de treino (20%) quanto a amostra de teste (80%) serão diferentes em cada classificação.

É possível perceber que em alguns grupos pode ocorrer sobreposição a cada teste, ou seja, imagens que estavam em um dos grupos na parte de treino ou teste em uma classificação podem aparecer em outra classificação na mesma condição anterior, representando uma repetição. Porém, este problema não é muito agravante, pois o objetivo de separar os treinos e testes em grupos distintos é mostrar que a metodologia é consistente nos piores e nos melhores casos.

Ao fim dos testes, são extraídas as médias de cada uma das medidas resultante, a acurácia, sensibilidade e especificidade.

4.2. Extração de Características

A extração de características ocorre com base nos modelos descritos no Capítulo 3. É realizada essa etapa três vezes: a primeira utilizando o modelo do *Artificial Crawlers*, a segunda utilizando o modelo do *Rose Diagram* e a terceira utilizando o modelo híbrido.

No modelo do *Artificial Crawlers* são escolhidos os três *templates* (candidatos a nódulos de tamanho pequeno, médio e grande) e são utilizados para fazer a diferença entre curvas de evolução dos candidatos a nódulos e não-nódulos, além da extração das áreas das curvas de evolução.

No modelo *Rose Diagram* todos os candidatos a nódulos e não nódulos tem seus gradientes de Sobel sobre a perspectiva do RD para a extração das características.

No modelo híbrido, ambos os modelos AC e RD são utilizados em combinação.

4.3. Classificação

Nesta seção serão mostrados os resultados da classificação, assim como foram divididos os testes. Cada conjunto de testes representa uma das técnicas abordadas. O primeiro conjunto é relacionado com os testes do AC, o segundo conjunto é relacionado com os testes do RD e o terceiro conjunto é relacionado com os testes da técnica híbrida.

O *kernel* do MVS utilizado foi o de base radial. Os parâmetros foram estimados utilizando a técnica *Grid Search*, que encontrou os valores do parâmetro *c* igual a 512 e *gamma* igual a 2.

Em cada conjunto de testes de cada técnica serão mostradas tabelas com as médias da acurácia (mACC), sensibilidade (mSEN) e especificidade (mESP), coeficiente de variância (CV) da acurácia e a média da área sob a curva ROC (mROC). Os valores são mostrados em porcentagem (%), com exceção da média sob a curva ROC, que pode ter valores entre 0 e 1, sendo que quanto mais próximo de 1, melhor o resultado. Os valores da mACC, mSEN e mESP quanto mais próximo de 100, melhor a classificação, mas o coeficiente de variância da acurácia, ao contrário, quanto mais próximo de 0 mostra um resultado mais consistente, pois mostra uma dispersão menor dos testes.

4.3.1. Testes do Modelo *Artificial Crawlers*

A Tabela 3 mostra os resultados dos testes do modelo AC.

Tabela 3: Resultados dos testes do modelo AC

CASOS DE TREINO/TESTE (%)	mACC (%)	mSEN (%)	mESP (%)	CV (%)	mROC
20/80	93.28	91.48	93.82	1,39	0,901
40/60	93,4	91,82	93,9	1,67	0,909
60/40	93,46	91,74	93,62	2,18	0,905
80/20	93.27	91.34	93.38	2,50	0,908

De acordo com os resultados dessa tabela, o melhor caso de mACC é o treino/teste 60/40. O pior caso (80/20) ainda mostra um resultado bem próximo do melhor caso de mACC. Já a mSEN e mESP tem como melhor caso o treino/teste 40/60 e como pior caso o treino/teste (80/20). O caso onde houve menor dispersão na acurácia foi o treino/teste 20/80. Isso significa que os valores da acurácia tiveram menos diferença para a média em todos os 5 testes. Apesar da maior mROC ser do treino/teste 40/60, não houve muita diferença para o pior caso (20/80).

As médias da acurácia, sensibilidade e especificidade estão acima de 90% em todos os casos de testes. As áreas médias das curvas ROC estão acima de 0,9.

4.3.2. Testes do Modelo *Rose Diagram*

A Tabela 4 mostra os resultados dos testes do modelo *Rose Diagram*.

Tabela 4: Resultados dos testes do modelo *Rose Diagram*.

CASOS DE TREINO/TESTE (%)	mACC (%)	mSEN (%)	mESP (%)	CV (%)	mROC
20/80	64,18	65,24	65,04	7,71	0,651
40/60	65,82	56,38	71,02	5,31	0,637
60/40	66,9	61,82	70,46	4,33	0,661
80/20	66,72	68,26	66,96	2,14	0,676

Os resultados observados na Tabela 4 mostram que a média da sensibilidade, coeficiente de variância da acurácia e área média das curvas ROC tem os melhores valores no caso de teste 80/20. A melhor média de acurácia é obtida no caso 60/40, com o valor de 66,9% e a melhor média de especificidade foi obtida no caso 40/60, com o valor de 71,02%.

A maioria dos casos apresentam os valores das médias de acurácia, sensibilidade e especificidade abaixo de 70%. Além disso os valores do coeficiente de variância são altos na maioria dos casos, e a diferença dos valores do melhor caso para o pior caso é grande, o que mostra que a classificação nesses casos tem uma discrepância bem grande.

4.3.3. Testes do Modelo Híbrido

A Tabela 5 mostra os resultados dos testes do modelo híbrido.

Tabela 5: Resultados dos testes do modelo híbrido

CASOS DE TREINO/TESTE (%)	mACC (%)	mSEN (%)	mESP (%)	CV (%)	mROC
20/80	93.54	90.58	94.46	1,5	0,908
40/60	94,2	91,7	94,84	1,14	0,918
60/40	94,30	91,86	94,78	1,61	0,922
80/20	94.22	92	94.52	1,94	0,921

É possível observar que o conjunto de treino/teste 60/40 obteve a melhor média da acurácia, com o valor de 94,3%. O pior caso da média da acurácia foi o conjunto 20/80, com 93,54%. O melhor caso de sensibilidade foi obtido no caso 80/20 com o valor 92%, enquanto o pior caso de sensibilidade foi o conjunto 20/80 com o valor 90,58%. O melhor caso de especificidade e do coeficiente de variância foi obtido no conjunto de treino/teste 40/60. A maior média de área da curva ROC foi obtido no treino/teste 60/40, com o valor de 0,922.

Com essa técnica é possível observar que os valores das médias da acurácia, sensibilidade e especificidade são acima de 90%. A discrepância entre melhor e o pior caso de testes do coeficiente de variância é bem pequena, o que mostra que os piores casos ainda obtêm resultados bons.

4.3.4. Comparação das Três Técnicas

A Tabela 6 mostra os resultados de todos os testes das três técnicas.

Tabela 6: Comparação dos resultados das três técnicas

MODELO	TREINO/TESTE (%)	mACC (%)	mSEN (%)	mESP (%)	CV (%)	mROC
ARTIFICIAL CRAWLERS	20/80	93.28	91.48	93.82	1,39	0,901
	40/60	93,4	91,82	93,9	1,67	0,909
	60/40	93,46	91,74	93,62	2,18	0,905
	80/20	93.27	91.34	93.38	2,50	0,908
ROSE DIAGRAM	20/80	64,18	65,24	65,04	7,71	0,651
	40/60	65,82	56,38	71,02	5,31	0,637
	60/40	66,9	61,82	70,46	4,33	0,661
	80/20	66,72	68,26	66,96	2,14	0,676
HÍBRIDO	20/80	93.54	90.58	94.46	1,5	0,908
	40/60	94,2	91,7	94,84	1,14	0,918
	60/40	**94,30**	**91,86**	**94,78**	**1,61**	**0,922**
	80/20	94.22	92	94.52	1,94	0,921

Os melhores resultados são apresentados no modelo híbrido, onde o melhor resultado em relação à acurácia foi destacado em negrito. O melhor resultado foi obtido no modelo híbrido, no caso de treino/teste 60/40%.

Os piores resultados estão concentrados no modelo *Rose Diagram*. Isso significa que o modelo *Rose Diagram* não é uma boa técnica para classificação se usada isoladamente, mas ela ajuda a melhorar os resultados obtidos no modelo *Artificial Crawlers*, mesmo que seja uma melhoria modesta.

Dessas métricas, a que teve melhoria com mais expressividade é o coeficiente de variância da acurácia. Em outras palavras, o modelo híbrido é o modelo mais consistente em relação à média da acurácia, o que mostra que os piores e os melhores casos não têm uma diferença expressiva.

4.4. Comparação de Resultados

A Tabela 7 apresenta os resultados dos trabalhos relacionados estudados para ajudar a propor a metodologia deste trabalho. A comparação é aproximativa, já que os trabalhos relacionados utilizam bases e exames diferentes na maioria dos casos, assim como a proporção dos exames.

Tabela 7: Comparação dos resultados dos trabalhos relacionados com a metodologia proposta.

Trabalho	Base	Quantidade de exames	Sensibilidade	Especificidade	Acurácia
Mousa *et al.*, 2002	Proprietária	50	87,5%	-	-
Jing *et al.*, 2010	LIDC	-	-	-	85,44%
Lee *et al.* 2010	LIDC	32	98,33%	97,11%	-
Namin *et al.*, 2010	LIDC	-	-	-	88%
Tartar *et al.* 2013	Proprietária	63	89,6%	87,5%	90,7%

Zhang *et al*. 2013	ELCAP	50	-	-	82,5%
Choi *et al*., 2014	LIDC	84	97,2%	97,7%	97,4%
Filho *et al.*, 2014	LIDC	833	85,91%	97,70%	97,55%
Franco *et al*. 2014	Proprietária	156	-	-	95,70%
Trabalho Proposto	**LIDC**	**833**	**91,86%**	**94,78%**	**94,30%**

A Tabela 7 contém o melhor resultado do trabalho proposto, que foi encontrado no caso de treino/teste 60/40% em relação à métrica da acurácia, pois ela é a métrica que melhor define o potencial da metodologia para diferenciar nódulo de não-nódulo.

Sobre aos trabalhos relacionados, apenas em Filho *et al.*, 2014 a proporção dos exames utilizados é a mesma, e os resultados da acurácia e especificidades são superiores, porém, a sensibilidade é inferior. Isso indica que houve maior taxa de falsos positivos no trabalho de Filho *et al.*, 2014 do que no trabalho aqui proposto. Outra semelhança do trabalho proposto por Filho *et al.*, 2014 e este trabalho é que ambos usam somente medidas de textura, obtendo resultados bons. Porém, essa abordagem baseada em árvores filogenéticas e índices taxonômicos tende a ser bem lentos computacionalmente. Para o cálculo das distâncias taxonômicas, é preciso escolher dois indivíduos (*voxels*) aleatoriamente em toda a amostra com sobreposição, o que aumenta bruscamente o tempo de execução dos treinos. Já na abordagem dos *Artificial Crawlers* e na abordagem do *Rose Diagram*, *os* indivíduos são analisados em relação aos seus vizinhos locais – e no caso do *Artificial Crawlers*, os ciclos tendem a ser menores, já que os indivíduos morrem ao longo do tempo de vida.

Os outros trabalhos utilizam bases de imagens ou pequena ou proprietária, o que diminui a confiabilidade dos resultados. A metodologia proposta apresenta-se bastante promissora em relação aos resultados alcançados nos trabalhos relacionados, já que uma de suas principais contribuições é a utilização de vidas artificiais na classificação de imagens médicas, algo não proposto na literatura pesquisada.

5. CONCLUSÃO

Esta dissertação apresentou uma metodologia de classificação de nódulos e não-nódulos utilizando três técnicas distintas: vidas artificiais com o modelo *Artificial Crawlers*, *Rose Diagram* para a extração de medidas estatísticas e um modelo híbrido. O modelo híbrido reúne as características extraídas pelo *Artificial Crawlers* e pelo *Rose Diagram* para tentar desenvolver uma técnica com o melhor das duas outras técnicas juntas.

No modelo do *Artificial Crawlers*, há algumas adaptações para a utilização das imagens de TC, como o aumento da percepção dos vizinhos para um ambiente 3D, o que permite a análise de textura nas imagens da base sem adapta-las para um ambiente 2D, que é a proposta inicial da técnica. Outra alteração importante é a eliminação do fator de aleatoriedade das localizações dos indivíduos que representam as vidas artificiais, sendo consideradas as texturas somente dos nódulos e não-nódulos, e não do fundo das imagens. Essas adaptações podem ser consideradas como contribuições importantes.

No modelo do *Rose Diagram* é utilizado o gradiente de Sobel para a análise direcional das texturas dos nódulos e não-nódulos. Dessas direções são extraídas medidas estatísticas direcionais através do *rose diagram*.

Ambas as técnicas não estão muito difundidas no contexto de processamento de imagens digitais e reconhecimento de padrões, então foram pouco utilizadas para o propósito de classificação. A metodologia avalia o uso dessas técnicas, o que pode ser considerado uma grande contribuição para a comunidade científica.

Pelos resultados obtidos nos usos das técnicas, o uso de medidas de textura foi muito consistente e efetivo, algo que é pouco comum de ser observado nos trabalhos com abordagens semelhantes e também pode ser considerado uma boa contribuição.

A diversidade de imagens da base de dados LIDC-IDRI e os 833 exames utilizados foram importantes para os treinos e testes, pois há inúmeros casos diferentes de nódulos pulmonares, além do fato de que os exames foram extraídos de diversos tomógrafos diferentes, o que torna a detecção, classificação e diagnóstico de nódulo mais difícil. A base foi dividida em candidatos a nódulos e candidatos a não-nódulos. Os testes propostos utilizaram grupos de treino/teste nas respectivas proporções: 20/80%, 40/60%,

60/40% e 80/20%. Esses artifícios ajudam a aumentar a confiabilidade dos treinos e testes da metodologia proposta.

Algumas dificuldades foram encontradas ao longo do desenvolvimento da metodologia. Uma dessas dificuldades foi encontrada na base LIDC-IDRI, já que algumas imagens da base tinham problemas de divergência de informações entre os cabeçalhos das imagens e as marcações dos especialistas. Esse problema foi resolvido excluindo essas imagens das bases de candidatos a nódulos e não-nódulos.

Outra dificuldade foi a seleção de candidatos a nódulos para serem *templates* na técnica do A*rtificial Crawlers* para fazer a extração de características por diferença de curvas, de forma que eles representem a base toda. Foi proposto uma seleção de candidatos a nódulos de acordo com seu tamanho: um *template* para nódulos pequenos, um para nódulos médios e um para nódulos grandes.

Os resultados obtidos foram bons no modelo *Artificial Crawlers* e no modelo híbrido, mas não foram bons no modelo *Rose Diagram*, sendo o melhor resultado sendo apresentado pelo modelo híbrido. Mesmo o modelo *Rose Diagram* tendo apresentado desempenho muito abaixo do esperado, ele não deve ser descartado, já que teve participação nos resultados do modelo híbrido.

Os testes utilizando o modelo híbrido para a classificação de nódulos e não-nódulos obteve resultados que alcançaram, no caso de amostra de treino com 60% da base e amostra de teste com os 40% restantes, 94,3% de média da acurácia, 91,86% de média da sensibilidade, 94,78% de média da especificidade, 1,61% de coeficiente de variância da acurácia e área da curva ROC de 0,922.

Por serem técnicas pouco utilizadas e exploradas no contexto de processamento de imagens e reconhecimento de padrões, de acordo com as referências pesquisadas, algumas melhorias podem ser feitas nos modelos propostos e são destacadas como possíveis trabalhos futuros:

- Utilização de técnicas de pré-processamento tanto no modelo *Artificial Crawlers* quanto no modelo *Rose Diagram*, para aumentar a diferença entre as duas bases de candidatos (realce de classe) e diminuir a quantidade de falsos positivos;
- Desenvolver métodos de escolha do tamanho dos setores do *Rose Diagram* de forma adaptativa;

- Melhorar a escolha de *templates* no modelo *Artificial Crawlers*, encontrando-os com mais critérios além do tamanho do nódulo, como por exemplo, a malignidade;
- Utilizar as técnicas abordadas em imagens de outros tipos de nódulos, como os nódulos mamários;
- Utilizar outras vidas artificiais além do *Artificial Crawlers*
- Reduzir a quantidade de candidatos a não-nódulos ou equilibrar as bases de dados.

Com isso, é possível considerar que a metodologia proposta é bem promissora, podendo ser aplicada em casos reais do cotidiano dos radiologistas para a classificação de nódulos em imagens de TC, pois mostra resultados bem consistentes e acima da média dos trabalhos relacionados da mesma área de detecção de cânceres pulmonares.

6. REFERÊNCIAS

ACS, A. C. (2014). *Cancer Reference*. Acesso em 10 de Dezembro de 2014, disponível em http://www.cancer.org/

Adami, C., & Brown, C. T. (1994). Evolutionary Learning in the 2D Artificial Life System "Avida". Em R. A. Brooks, & P. Maes, *Artificial Life IV* (pp. 377-381). Massachusetts: MIT Press.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research 13*, 281-305.

Black, P. E. (2006). *Manhattan distance*. Acesso em 11 de Dezembro de 2014, disponível em Dictionary of Algorithms and Data Structures: http://xlinux.nist.gov/dads/HTML/manhattanDistance.html

Bourne, M. (2014). *The Area under a Curve*. Acesso em 25 de Dezembro de 2014, disponível em http://www.intmath.com/integration/3-area-under-curve.php

Brodatz, P. (1968). *Textures: A Photographic Album for Artists and Designers.* Reinhold: Dover Publications.

CDB. (2013). *Câncer de pulmão*. Acesso em 27 de Dezembro de 2014, disponível em http://www.cdb.com.br/noticias/cancer-de-pulmao-tomografia-computadorizada-poderia-salvar-ate-12-mil-vidas-por-ano-diz-estudo-americano/85

Charrel, A. (1995). Tierra Network Version. *ATR Technical Report TR-H-145*.

Choi, W., & Choi, T. (2014). Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer Methods and Programs in Biomedicine*, 37–54.

Collins, R. J., & Jefferson, D. R. (1991). AntFarm: Towards Simulated Evolution. Em *Artificial Life II* (pp. 579-601). Addison-Wesley.

Darwin, C. (1859). *A Origem das Espécies.* Porto: LELLO & IRMÃO.

de Carvalho Filho, A. O. (2013). Detecção Automática De Nódulos Pulmonares Solitários Usando Quality Threshold Clustering E Mvs. *Dissertação do mestrado do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão*.

de Carvalho Filho, A. O., Sampaio, W. B., Silva, A. C., Paiva, A. C., Nunes, R. A., & Gattass, M. (2014). Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial Intelligence in Medicine*, 165-177.

Dekhtyar, A. (2009). Distance/Similarity Measures. *Knowledge Discovery from Data*.

Deza, M. M., & Deza, E. (2006). *Dictionary of Distances.* Elsevier Science.

Educação, P. (2014). *Biologia Celular: Câncer*. Acesso em 27 de Dezembro de 2014, disponível em http://www.portaleducacao.com.br/biologia/artigos/28204/biologia-celular-cancer#ixzz2rsz10E2k

Falta, K. (2011). Analysis of Orientations of Micro Cracks via Circular Statistics. *Bachelor Thesis.Technische Universität Dortmund*.

Fisher, R., Perkins, S., A., W., & E., W. (2003). *Sobel Edge Detector*. Acesso em 25 de Dezembro de 2014, disponível em http://homepages.inf.ed.ac.uk/rbf/HIPR2/sobel.htm

Franco, M. L., Nunes, L. M., Froner, A. P., Silva, A. M., & Patrocínio, A. C. (2014). Redes Neurais Artificiais Aplicadas Na Classificação De Tumores Pulmonares. *XXIV Congresso Brasileiro de Engenharia Biomédica*.

Frank, E., Hall, M., Reutemann, P., & Trigg, L. (1993). *Machine Learning Group at the University of Waikato*. Acesso em 24 de Dezembro de 2014, disponível em http://www.cs.waikato.ac.nz/ml/weka/

Gan, G. (2011). Simple Matching Distance. Em G. Gan, *Data Clustering in C++: An Object-Oriented Approach* (p. 142). CRC Press.

Gonçalves, W. N., Machado, B. B., & Martinez, B. O. (2014). Texture descriptor combining fractal dimension and artificial crawlers. *Physica A: Statistical Mechanics and its Applications*, 358–370.

Greenacre, M., & Primicerio, R. (2013). *Multivariate Analysis of Ecological Data.* Rubes Editorial.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *Systems, Man and Cybernetics*, 610-621.

He, D. (1990). Texture Unit, Texture Spectrum, And Texture Analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 509 - 512.

Hillis, D. M. (1990). The Co-evolution of Host and Parasite Population. Em C. G. Langton, *Artificial Life II* (pp. 313-324). Addison-Wesley.

Holland, J. H. (1992). The Echo Model. *Proposal for a Research Program in Adaptive Computation*.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1989). *Induction: Processes of Inference, Learning, and Discovery.* Cambridge: MIT Press.

Hsu, S.-Y. (1977). A texture-tone analysis for automated landuse mapping with panchromatic images. *American Society of Photogrammetry, Annual Meeting*, 203-215.

INCA. (2014). *INSTITUTO NACIONAL DE CÂNCER*. Acesso em 27 de Dezembro de 2014, disponível em http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao

Ivanciuc, O. (2005). *The SVM - Support Vector Machines*. Acesso em 14 de Dezembro de 2014, disponível em http://www.support-vector-machines.org/

Jing, Z., Bin, L., & Lianfang, T. (2010). Lung nodule classification combining rule-based and SVM. *Fifth International Conference on Bio-Inspired Computing: Theories and Applications*, 1033-1036.

Jones, G. (1998). Genetic and Evolutionary Algorithms. *Encyclopedia of Computational Chemistry*.

Klove, T., Bergen, N., Lin, T., Tsai, S., & Tzeng, W. (2010). Permutation Arrays Under the Chebyshev Distance. *Information Theory*, 2611-2617.

Lamp, G. (2012). *Why use SVM?* Acesso em 24 de Dezembro de 2014, disponível em http://www.yaksis.com/posts/why-use-svm.html

Langton, C. G. (1989). *Artificial Life: Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems.* Boston: Addison-Wesley Longman Publishing Co., Inc.

Lee, S., Kouzani, A., & Hu, E. (2010). Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics*, 535–542.

LIDC-IDRI, N. C. (2012). *Lung Image Database Consortium image collection*. Acesso em 24 de Dezembro de 2014, disponível em https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

Machado, N. (2014). *Ciências Físico-Químicas*. Acesso em 09 de 12 de 2014, disponível em http://www.aulas-fisica-quimica.com/7f_07.html

MacMahon, H., Austin, J. H., Gamsu, G., Herold, C. J., Jett, J. R., Naidich, D. P., . . . Swensen, S. J. (2005). Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society. *Radiology*, 395–400.

Marceau, D. (1989). Automated Texture Extraction From High Spatial Resolution Satellite Imagery For Land-cover Classification: Concepts And Application. *Geoscience and Remote Sensing Symposium, 1989. IGARSS'89. 12th Canadian Symposium on Remote Sensing., 1989 International*, 2765 - 2768.

Mardia, K. V., & Jupp, P. E. (2006). *Directional Statistics.* John Wiley & Sons.

McClean, P. (2000). *Mendelian Genetics*. Acesso em 09 de 12 de 2014, disponível em Pubweb: http://www.ndsu.edu/pubweb/~mcclean/plsc431/mendel/mendel1.htm

Morris, S. J. (1997). *The Pythagorean Theorem*. Acesso em 09 de Dezembro de 2014, disponível em http://jwilson.coe.uga.edu/emt669/student.folders/morris.stephanie/emt.669/essay.1/pythagorean.html

Mousa, W. A., & Khan, M. A. (2002). Lung nodule classification utilizing support vector machines. *International Conference on Image Processing*, 153-156.

Namin, S. T. (2010). Automated detection and classification of pulmonary nodules in 3D thoracic CT images. *Systems Man and Cybernetics*, 3774-3779.

NEMA. (2014). *Medical Imaging & Technology Alliance*. Acesso em 10 de Dezembro de 2014, disponível em http://medical.nema.org/standard.html

Netto, S. M. (2010). Segmentação Automática de Nódulo Pulmonar com Growing Neural Gas e Máquina de Vetores de Suporte. *Dissertação de Mestrado na área de Ciência da Computação. (Programa de Pós-Graduação em Engenharia de Eletricidade) - Universidade Federal do Maranhão*.

Ost, D., Fein, A. M., & Feinsilver, S. H. (2003). The solitary pulmonary nodule. *The New England Journal of Medicine*.

Phillips, J. M. (2012). Jaccard Similarity and Shingling. University of Utah.

Polakowski, W., Cournoyer, D., Rogers, S., & DeSimio, M. (1997). Computer-Aided Breast Cancer Detection And Diagnosis Of Masses Using Difference Of Gaussians And Derivative-Based Feature Saliency. *IEEE Transactions on Medical Imaging*, 811–819.

Rennó, C. D., & Soares, J. V. (1996). Utilizacao de Medidas Texturais Na Discriminacao de Classes de Uso do Solo do Perimetro Irrigado de Bebedouro, Pernambuco, Brasil, Utilizando-Se Imagens SAR. *Image Processing Techniques, First Latino-American Seminar on Radar Remote Sensing*, 171.

Rotter, A. J., Schabath, M. B., Sequist, L. V., Tong, B. C., Travis, W. D., Unger, M., & Yang, S. C. (2012). Lung Cancer Screening. *Journal of National Comprehensive Cancer Network*, 240-265.

Santos, A. M. (2011). *Detecção de nódulos pulmonares pequenos usando o Modelo de Mistura Gaussiana e Matriz Hessiana*. São Luís: Dissertação de Mestrado na área de Ciência da Computação. (Programa de Pós-Graduação em Engenharia de Eletricidade) - Universidade Federal do Maranhão.

Schuster, P. (2011). Mathematical modeling of evolution. Solved and open problems. *Theory in Biosciences*, 71-89.

Soliman, A. A., Ellah, A. H., Abou-Elheggag, N. A., & Abd-Elmougod, G. A. (2012). A simulation-based approach to the study of coefficient of variation of Gompertz distribution under progressive first-failure censoring. *Indian Journal of Pure and Applied Mathematics*, 335-356.

Sousa, J. R. (2007). Metodologia para detecção automática de nódulos pulmonares. *Dissertação de Mestrado na área de Ciência da Computação. (Programa de Pós-Graduação em Engenharia de Eletricidade) - Universidade Federal do Maranhão*.

STATISTICS, C. (2014). *SEER Stat Fact Sheets: Lung and Bronchus Cancer*. Acesso em 27 de Dezembro de 2014, disponível em http://seer.cancer.gov/statfacts/html/lungb.html

Tartar, A., Kilic, N., & Akan, A. (2013). Classification of Pulmonary Nodules by Using Hybrid Features. *Computational and Mathematical Methods in Medicine*.

Taylor, C., & Jefferson, D. (1995). Artificial Life as a Tool for Biological Inquiry. Em *Artificial Life: An Overview* (pp. 1-13).

Teknomo, K. (2006). *Euclidian Distance*. Acesso em 09 de 12 de 2014, disponível em http://people.revoledu.com/kardi/tutorial/Similarity/EuclideanDistance.html

The Best-Run Businesses Run SAP. (2014). Acesso em 16 de Dezembro de 2014, disponível em http://help.sap.com/saphelp_470/helpdata/pt/fc/6ef90a1b6811d5928d0008c7d94f4b/content.htm

Uehara, C., Jamnik, S., & Santoro, I. L. (1998). Câncer de Pulmão. Em *Medicina, Ribeirão Preto* (pp. 266-276).

van Erkel, A., & Pattynama, P. (1998). Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology 27(2)*, 88-94.

Wind rose plot for LaGuardia Airport. (2010). Acesso em 16 de Dezembro de 2014, disponível em http://en.wikipedia.org/wiki/Wind_rose#mediaviewer/File:Wind_rose_plot.jpg

Yaeger, L. (1994). Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context. Em C. G. Langton, *Proceedings of the Artificial Life III Conference* (pp. 263-298). Addison-Wesley.

Zhang, D., & Chen, Y. Q. (20 de Setembro de 2004). Classifying Image Texture with Artificial Crawlers. *International Conference on Intelligent Agent Technology*, pp. 446 - 449.

Zhang, D., & Chen, Y. Q. (2005). Artificial Life: A new approach to texture classification. Em *International Journal of Pattern Recognition and Artificial Intelligence* (Vol. 19, pp. 355-374). World Scientific Publishing Company.

Zhang, F., Song, Y., Cai, W., Zhou, Y., Shan, S., & Feng, D. (2013). Context Curves for Classification of Lung Nodule. *Digital Image Computing: Techniques and Applications (DICTA)*, 1-7.