UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Daniela Fernanda Taino

Um modelo baseado em algoritmo genético para seleção de características e classificação de padrões em imagens médicas

São José do Rio Preto
2020

Daniela Fernanda Taino

Um modelo baseado em algoritmo genético para seleção de características e classificação de padrões em imagens médicas

Dissertação apresentada como parte dos requisitos para obtenção do título do Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Orientador:

Prof. Dr. Leandro Alves Neves

São José do Rio Preto
2020

T134m

Taino, Daniela Fernanda
Um modelo baseado em algoritmo genético para seleção de características e classificação de padrões em imagens médicas / Daniela Fernanda Taino. -- São José do Rio Preto, 2020
102 f.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Leandro Alves Neves

1. Algoritmos genéticos. 2. Reconhecimento de padrões. 3. Inteligência artificial. 4. Imagens médicas. 5. Câncer. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Daniela Fernanda Taino

Um modelo baseado em algoritmo genético para seleção de características e classificação de padrões em imagens médicas

Dissertação apresentada como parte dos requisitos para obtenção do título do Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Comissão Examinadora

Prof. Dr. Leandro Alves Neves
Unesp - Câmpus São José do Rio Preto
Orientador

Profa. Dra. Rogéria Cristiane Gratão de Souza
Unesp - Câmpus São José do Rio Preto

Prof. Dr. Marcelo Costa Oliveira
UFAL - Universidade Federal de Alagoas

São José do Rio Preto
08 de Julho de 2020

Agradecimentos

Primeiramente, agradeço aos meus pais, Aparecida e Lúcio, e à minha irmã, Denise, pelo apoio incondicional durante toda a minha vida. Agradeço também minha família, principalmente minhas primas Tamiris e Tainá. Tive muita sorte de ser tão próxima de vocês.

Agradeço ao Prof. Dr. Leandro Alves Neves pela orientação, paciência e pelo apoio desde a graduação. Sem a sua ajuda, este trabalho não teria saído do papel.

Agradeço às minhas amigas Isabella, Larissa e Daniela Pereira por sempre estarem ao meu lado. Nunca poderei agradecer suficientemente por tudo o que vocês já fizeram por mim.

Agradeço também aos amigos que o Ibilce me deu: Guilherme Morais, Lais, Maycon, Matheus, André, João, Leandro, Hugo e João Paulo. Conhecer vocês foi uma das melhores coisas que já me aconteceu. Tive e, felizmente, ainda tenho o privilégio de acompanhar a trajetória de cada um de vocês. Aqueles jovens que sofriam com provas e tomavam café juntos quase todos os dias se tornaram adultos incríveis, bem-sucedidos e com um futuro brilhante pela frente. Talvez as nossas conversas na cantina deixaram uma certa dúvida sobre o quão promissor o futuro seria, mas *juro* que não duvidei de vocês por mais que um instante.

Por fim, agradeço também aos meus colegas de laboratório, Guilherme Freire, Guilherme Rozendo, Álvaro, Eduardo Pavarino, e novamente, ao Matheus. Nossas trocas de experiências e trabalhos foram fundamentais para que este trabalho fosse concluído.

Resumo

A análise de imagens histológicas é baseada na avaliação visual dos tecidos por especialistas, utilizando um microscópio óptico. Essa tarefa pode ser demorada e desafiadora, principalmente devido à complexidade das estruturas e doenças sob investigação. Estes fatos motivaram o desenvolvimento de métodos computacionais para apoiar especialistas em pesquisas e tomadas de decisões. Apesar das diferentes estratégias computacionais disponíveis na literatura, as soluções baseadas em programação genética não foram totalmente exploradas para fornecer a melhor combinação de recursos, algoritmos de seleção e classificadores. Uma abordagem baseada em algoritmo genético, capaz de avaliar um número significativo de características, métodos de seleção e classificadores, é descrita para fornecer uma combinação aceitável para o diagnóstico e reconhecimento de padrões de linfomas não-Hodgkin e câncer colorretal. A estrutura cromossômica foi representada por quatro genes. A avaliação e seleção dos indivíduos, bem como os processos de cruzamento e mutação, foram definidos para distinguir os grupos investigados, com o maior valor da AUC e o menor número de características. Os testes foram realizados considerando 1.512 características de imagens histológicas, diferentes tamanhos de população, taxas de mutação e número de iterações. Uma população inicial, de 50 indivíduos e 50 iterações, forneceu o melhor resultado para o câncer colorretal. Para os linfomas não-Hodgkin, a população inicial foi composta de 500 indivíduos e também foram necessárias 50 iterações para fornecer melhor resultado. Os valores da AUC foram de 0,984 e 0,947 para o câncer colorretal e os linfomas não-Hodgkin, respectivamente. A metodologia proposta, com informações detalhadas sobre os métodos, características e melhores combinações, é uma contribuição relevante para a comunidade interessada no estudo do reconhecimento de padrões em imagens histológicas de linfomas e câncer colorretal.

Palavras-chave: algoritmo genético, seleção de características, associação de técnicas, classificação de caraterísticas, imagens médicas.

Abstract

The analysis of histological images is based on visual assessment of tissues by specialists using an optical microscopy. This task can be time-consuming and challenging, mainly due to the complexity of the structures and diseases under investigation. These facts have motivated the development of computational methods to support specialists in research and decision-making. Despite the different computational strategies available in the literature, the solutions based on genetic programming have not been fully explored to provide the best combination of features, selection algorithms and classifiers. An approach based on genetic algorithm able to evaluate a significant number of features, selection methods and classifiers is described in order to provide an acceptable association for the diagnosis and pattern recognition of Non-Hodgkin lymphomas and colorectal cancer. The chromosomal structure was represented with four genes. The evaluation and selection of individuals, as well as the crossover and mutation processes were defined to distinguish the groups under investigation, with the highest AUC value and the smallest number of features. The tests were performed considering 1,512 features from histological images, different population sizes, mutation rate and number of iterations. An initial population of 50 individuals and 50 iterations provided the best result for colorectal cancer. For non-Hodgkin's lymphomas, the initial population consisted of 500 individuals and 50 iterations was also necessary to provide the best result. The AUC values were 0.984 and 0.947 for the colorectal cancer and the Non-Hodgkin lymphomas, respectively. The proposed methodology with detailed information regarding the methods, features and best combinations are relevant contributions for the community interested into the study of pattern recognition of colorectal cancer and lymphomas.

Keywords: genetic algorithm, feature selection, combination of techniques, feature classification, medical images.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

G_{clf} Índice do Método de Classificação

G_{id} Identificador Único do Indivíduo

G_{num} Número de Características na Iteração

G_{sel} Índice do Método de Seleção

$Iter$ Iterações

$MaxF$ Número máximo de características a serem consideradas

$MeanAuc$ AUC média

Acc Acurácia

ADABOOST *Adaptive Boosting*

AG Algorítmo Genético

AGFS *Adaptive Genetic Fuzzy System*

AUC Área sob a Curva ROC

BPNN *Back Propagation Neural Network*

CLL *Chronic Lymphocytic Leukemia*

CNN *Convolutional Neural Network*

CSSVM *Cost Sensitive Support Vector Machine*

Curva ROC *Receiver Operating Characteristic Curve*

DL *Deep Learning*

DT *Decision Tree*

FDT *Functional Decision Tree*

FL *Follicular Lymphona*

FP Falsos Positivos

KNN *K-Nearest Neighbors*

LOO *Leave-One-Out*

LWL *Locally Weighted Learning*

MCL *Mantle Cell Lynphoma*

MI *Mutual Information*

MLP *Multilayer Perceptron*

NHL *Non-Hodkin Lymphoma*

PID *Pima Indian Diabetes*

PSO *Particle Swarm Optimization*

SGD *Stochastic Gradient Descent*

SL *Simple Logistics*

SVM *Support Vector Machine*

VP Verdadeiros Positivos

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

As áreas de radiologia e patologia são especialidades médicas que compreendem o uso de tecnologias de imagens para a investigação de doenças e indicações de diagnósticos, principalmente para identificar doenças com alta mortalidade. Essas especialidades se tornaram cada vez mais comuns, especialmente por possibilitarem diagnósticos via técnicas menos invasivas (PARIKH; TICKMAN, 2005).

A principal dificuldade para um diagnóstico médico está na avaliação da gravidade de achados anormais quando diferentes opiniões podem ocorrer interobservador e intraobservadores. Além disso, o processo de análise de imagens médicas para o diagnóstico consome tempo e pode ser influenciada pela experiência do especialista (VAN GINNEKEN; ROMENY; VIERGEVER, 2001; CHAN et al., 1990; DOI; HUANG, 2007). Esses fatos têm motivado o desenvolvimento e a aplicação de sistemas conhecidos como *Computer-Aided Diagnosis* (CAD) para apoiar os especialistas no estudo e nas tomadas de decisão (YU; GUAN, 2000). Esses sistemas consideram diferentes etapas, tais como extração, análise e seleção de características das imagens, bem como a interpretação e o reconhecimento de possíveis padrões. Um desafio comumente observado em propostas de sistemas CAD é o de indicar as melhores combinações entre algoritmos de seleção e classificação para alcançar as maiores taxas de acerto com o menor número de características (GURCAN et al., 2002; DOI; HUANG, 2007).

Os sistemas CAD estão amplamente relacionados com a área do conhecimento definida

como aprendizado de máquina, do inglês, *machine learning*) (ESTEVA et al., 2017; KOOI et al., 2017; BEJNORDI et al., 2017; ZEZOS et al., 2018). O aprendizado de máquina tem como objetivo a construção de estratégias computacionais que podem aprender a partir de seus erros e definir possíveis previsões sobre dados, aplicando, por exemplo, o raciocínio indutivo (AMORIM; BARONE; MANSUR, 2008). Os sistemas pautados em aprendizado de máquina podem ser observados a partir das etapas de levantamento de dados para treinamento e testes (usados para treinar e medir o desempenho do método), seleção dos características, escolha de classificadores, execução do treinamento e avaliação dos resultados obtidos (AMORIM; BARONE; MANSUR, 2008).

Segundo Janssen, Mourão-Miranda e Schnack ((2018)), a eficácia de técnicas baseadas em aprendizado de máquina tem sido alcançada com o desenvolvimento de algoritmos para atender demandas específicas, bem como pela maior disponibilidade de recursos computacionais e conjuntos de dados necessários para testes e validações. Além disso, o sucesso das abordagens baseadas em aprendizado de máquina deve-se ao enfoque na generalização, que é a manutenção do desempenho do modelo a partir da inserção progressiva de novos dados. É importante destacar que a capacidade de generalização pode ser afetada pelos fatores nomeados como: *underfitting*, no qual o modelo não está generalizado o suficiente; *overfitting*, no qual o modelo fornece resultados relevantes sobre o conjunto de dados utilizado para sua construção, porém ineficaz para prever novos resultados; e, por fim, o erro irredutível, quando há presença de ruído nos dados e que pode determinar o limite teórico do desempenho do algoritmo (JANSSEN; MOURÃO-MIRANDA; SCHNACK, 2018; DIETTERICH, 1995). Portanto, o desenvolvimento de sistemas CAD combinando as melhores estratégias para contornar esses problemas é um tema desafiador, principalmente para assegurar que a capacidade de generalização não seja afetada (SAHINER; CHAN; HADJIISKI, 2008; JANSSEN; MOURÃO-MIRANDA; SCHNACK, 2018).

Nesse contexto, desde o desenvolvimento das primeiras técnicas de seleção de características e classificação de dados, muito se fez na área do aprendizado de máquina aplicada em medicina. Por exemplo, existem estudos direcionados para a utilização de sistemas CAD em conjunto com os exames comumente encontrados na prática clínica, tais como raios X

(MELENDEZ et al., 2015; WANG, X. et al., 2017; RAJPURKAR et al., 2017; APOSTOLOPOULOS; MPESIANA, 2020), tomografia computadorizada (GURCAN et al., 2002; HUA et al., 2015; CHENG et al., 2016; JALALIAN et al., 2017; EL-ANWAR et al., 2020) e ressonância magnética (ARIMURA et al., 2009; LEITE et al., 2015; SINGH et al., 2015; GIANNINI et al., 2017; LEE et al., 2020), todos com o objetivo de fornecer uma segunda opinião e apurar ainda mais o diagnóstico. As doenças de maior interesse são definidas a partir das suas taxas de mortalidade, letalidade e incidência, além do fato de o diagnóstico precoce contribuir significativamente para a melhoria do prognóstico. Alguns exemplos são o câncer de mama (HO; LAM, 2003; JALALIAN et al., 2017; KOOI et al., 2017; BEJNORDI et al., 2017; WANG, P. et al., 2020), o câncer de pulmão (TAGUCHI et al., 1999; HUA et al., 2015; HUANG; PARK et al., 2017; EL-REGAILY et al., 2018; SILVA et al., 2020), os tumores cerebrais (VRJI; JAYAKUMARI, 2011; PRAVEEN; AGRAWAL, 2015; ANITHA; MURUGAVALLI, 2016; TIWARI; SACHDEVA et al., 2017; TIWARI; SRIVASTAVA; PANT, 2020), o câncer de próstata (DOYLE et al., 2006; KOIZUMI et al., 2017; AFEF et al., 2018; ALDOJ et al., 2020), as doenças cardíacas (ANBARASI; ANUPRIYA; IYENGAR, 2010; SHARMA et al., 2015; PATIDAR; PACHORI; ACHARYA, 2015; EL-ANWAR et al., 2020), os linfomas (HERMINE et al., 2016; CODELLA et al., 2016; ABAS et al., 2017; TCHRAKIAN et al., 2018; RIBEIRO, Matheus Gonçalves et al., 2019; ROBERTO et al., 2017; SWERDLOW; COOK, 2020) e o câncer colorretal (KATHER et al., 2016; KOMINAMI et al., 2016; MISAWA et al., 2016; RIBEIRO, Matheus Gonçalves et al., 2019; JAYACHANDRAN; GHOSH, 2020).

Para investigar as melhores combinações entre características, algoritmos de seleção e classificadores, estudos disponíveis na literatura foram desenvolvidos para maximizar a acurácia de classificações por meio de novas metodologias para a seleção de características, muitas vezes influenciadas por estratégias baseadas em modelos metaheurísticos e *deep learning* (DL). Como exemplo de métodos baseados em DL, podemos citar As Redes Neurais Convolucionais (BRANCATI et al., 2019) e as Redes Neurais Recorrentes (LI et al., 2018). As técnicas inspiradas em analogias encontradas na natureza ou em processos evolutivos (EIBEN; SMITH et al., 2003) merecem destaque, tais como os métodos baseados em colônia

de abelhas (*Bee Colony*) (ZHANG et al., 2015), colônia de formigas (*Ant Colony*) (KARNAN; LOGHESHWARI, 2010) e *Particle Swarm Optimization* ou PSO (REMAMANY et al., 2015) para estudos de tumores cerebrais, bem como o uso de *Differential Evolution* para investigação de nódulos pulmonares (JAFFAR; SIDDIQUI; MUSHTAQ, 2018) e os modelos baseados em Algoritmos Genéticos (AG) (MUNI; PAL; DAS, 2006; WELIKALA et al., 2015; PAUL et al., 2017; JIANG et al., 2017; KEČO; SUBASI; KEVRIC, 2018; FARID; SELIM; KHATER, 2020). É importante mencionar que as metaheurísticas e os métodos baseados em DL apresentam limitações bem-conhecidas. Por exemplo, a colônia de formigas requer mais tempo de execução para convergência e esse modelo apresenta desempenho ruim quando aplicado a problemas caracterizados por grandes espaços de busca (AB WAHAB; NEFTI-MEZIANI; ATYABI, 2015). O método PSO apresenta uma tendência de convergência prematura em pontos médios ótimos, assim como uma fraca busca local e com convergência lenta (AB WAHAB; NEFTI-MEZIANI; ATYABI, 2015). O fraco desempenho em relação à capacidade de busca local também é observado na abordagem *Differential Evolution*, com problemas de convergência prematura. Além disso, nessa estratégia, o desempenho pode diminuir quando a dimensão do espaço de busca cresce expressivamente (MOHAMED; SABRY; KHORSHID, 2012). O método *Cuckoo Search* também tem o problema de facilmente encontrar soluções locais ótimas e possui uma lenta taxa de convergência (LIU et al., 2018; YANG; DEB, 2009). As desvantagens da estratégia *Bee Colony* estão relacionadas à necessidade de um número expressivo de avaliações na função objetiva e novos testes de aptidão para cada conjunto de novos parâmetros, a fim de melhorar o desempenho (AB WAHAB; NEFTI-MEZIANI; ATYABI, 2015). Os métodos baseados em DL estão associados a um custo computacional extremamente alto quando as imagens de entrada são de alta resolução. Uma solução comumente usada é a redução dessas imagens, mas informações relevantes acabam sendo perdidas no processo (LOPES, 2017). Outras desvantagens são que esses métodos precisam de um número expressivo de dados rotulados para treinamento e teste (RODRIGUES, 2018), além de não ser possível obter um entendimento detalhado sobre as relevâncias das características utilizadas para definir a melhor solução.

A principal desvantagem de um AG é a convergência não rápida, já que os processos

de cruzamento e mutação são aleatórios (AB WAHAB; NEFTI-MEZIANI; ATYABI, 2015). Mesmo considerando essas limitações, a principal vantagem de um AG, em comparação a outras estratégias descritas anteriormente, é ter uma estrutura que permita representar novas formas organizacionais (indivíduos) aceitáveis a partir de uma população anterior bem-sucedida (cruzamento) (BRUDERER; SINGH, 1996) sem perder informações críticas do problema (MUNI; PAL; DAS, 2006) (WHITLEY, 1994). Apresentados por John Holland em 1975 (HOLLAND, 1992), o AG é uma técnica heurística de otimização global, desenvolvida a partir de princípios evolutivos, ou o chamado darwinismo. Essa técnica é empregada na resolução de problemas que geralmente demandam a avaliação de um número significativo de possíveis soluções ou combinações, ou seja, aqueles considerados como problemas de otimização complexos, como o do "caixeiro viajante", por exemplo (MIRANDA, 2007).

Diversos são os trabalhos que propuseram modelos CAD combinando AG com diferentes classificadores (ANBARASI; ANUPRIYA; IYENGAR, 2010; ÖZÇIFT; GÜLTEN, 2013; LU; ZHU; GU, 2014). Anbarasi, Anupriya e Iyengar ((2010)) exploraram AG com os classificadores *Naive Bayes*, *Clustering* e árvore de decisão para auxiliar no diagnóstico de doenças cardíacas. Özçift e Gülten ((2013)) propuseram em um novo método para o diagnóstico diferencial de doenças eritêmato-escamosas. O método, baseado em um AG e uma *Bayesian Network*, apresentou acurácia de 99,2% e foi superior quando comparado aos desempenhos obtidos com outros quatro classificadores conhecidos: *Support Vector Machine* (98,36%), *Multi-Layer Perceptron* (97,00%), *Simple Logistic* (98,36%) e *Functional Decision Tree* (97.81%). Um sistema inteligente para o diagnóstico de câncer de pulmão, baseado em AG e em *Mutual Information*, foi proposto com a finalidade de selecionar os características mais relevantes em um conjunto e, posteriormente, classificá-los (LU; ZHU; GU, 2014). Os classificadores aplicados foram *Support Vector Machine* (SVM), *Back Propagation Neural Network* (BPNN) e o *K-Nearest Neighbors* (KNN).

As abordagens baseadas em AG também foram definidas para reduzir o número de características. Por exemplo, Velu e Kashwan ((2013)) apresentou uma proposta para analisar características de pacientes diabéticos. AG, *Expectation-maximization Algorithm* e *h-means+* foram combinados e utilizados para definir agrupamentos e encontrar similaridades entre as

características. A combinação forneceu um coeficiente de correlação de 0,96. No trabalho proposto por Kumar, Ramachandra e Nagamani ((2014)), foi descrita uma combinação entre os AGs, como método para selecionar as características, e SVM como classificador. O método foi capaz de indicar um ótimo subconjunto de características e fornecer as melhores taxas de desempenho quando comparadas aos trabalhos disponíveis na literatura. Outro exemplo foi a proposta descrita por Gokulnath e Shantharajah ((2019)), em que o AG foi utilizado como meio de encontrar as características mais relevantes e reduzir o conjunto inicial para obter melhores resultados na classificação de doenças cardíacas. Os classificadores utilizados foram BPNN, KNN e *Cost Sensitive Support Vector Machine* (CSSVM). Os autores concluíram que o método superou os desempenhos de trabalhos disponíveis na literatura.

1.1 Motivação e Justificativas

Nos trabalhos disponíveis na literatura (ANBARASI; ANUPRIYA; IYENGAR, 2010; ÖZÇIFT; GÜLTEN, 2013; VELU; KASHWAN, 2013; LU; ZHU; GU, 2014; KUMAR; RAMACHANDRA; NAGAMANI, 2014; GOKULNATH; SHANTHARAJAH, 2019), foram propostas diferentes metodologias para identificar as melhores combinações de técnicas e características. Um ponto em comum entre esses trabalhos foi a utilização de um AG para alcançar os melhores resultados. Além disso, o AG foi associado com diferentes classificadores de padrões, tais como SVM (LU; ZHU; GU, 2014) (GOKULNATH; SHANTHARAJAH, 2019), *Decision Tree* (DT) (ANBARASI; ANUPRIYA; IYENGAR, 2010), KNN, BPNN (LU; ZHU; GU, 2014), *Clustering* (ANBARASI; ANUPRIYA; IYENGAR, 2010), *Multilayer Perceptron* (MLP) (ÖZÇIFT; GÜLTEN, 2013), *Simple Logistics* (SL) (ÖZÇIFT; GÜLTEN, 2013), FDT (ÖZÇIFT; GÜLTEN, 2013) e *Naive Bayes* (ANBARASI; ANUPRIYA; IYENGAR, 2010). Nesses trabalhos, os autores também identificaram as características mais relevantes para cada contexto de aplicação, com um número reduzido de características.

Contudo, os estudos encontrados na literatura não envolveram uma combinação mais abrangente de técnicas (características, métodos de seleção e classificadores) com a flexibi-

lidade necessária para diferentes tipos de características e contextos. Por exemplo, Anbarasi, Anupriya e Yengar ((2010)) utilizaram o AG como seletor único para reduzir o conjunto de características e melhorar o diagnóstico em doenças cardíacas, assim como Gokulnath e Shantharajah ((2019)). Neste, os autores utilizaram três métodos de classificação para avaliar o desempenho: BPNN, KNN e CSSVM. Özçift E Gülten ((2013)) e Lu, Zhu e Gu ((2014)) associaram o AG a somente outro método, uma rede Bayesiana e MI (*Mutual Information*), respectivamente. Assim, não é possível afirmar que os bons resultados obtidos podem ser alcançados em outros contextos de interesse. Portanto, uma investigação direcionada para determinar um método capaz de combinar um número significativo de características, métodos de seleção e de classificação, usando-os como parâmetros para um AG, pode contribuir significativamente com o desenvolvimento de métodos de aprendizado de máquina e, consequentemente, sistemas CAD.

É importante destacar que são vários os contextos de imagens médicas que podem ser explorados para avaliar os sistemas computacionais que apoiam o diagnóstico médico, tais como os de imagens de linfomas e de câncer colorretal. A importância do estudo dessas doenças deve-se ao número de casos e à mortalidade. O método mais comum para o diagnóstico de ambos os tipos de câncer é a partir de biópsias dos tecidos que são corados com hematoxilina e eosina (H&E) e avaliados microscopicamente por médicos especialistas, uma tarefa que, além de desafiadora, pode consumir muito tempo do patologista (ORLOV et al., 2010).

O câncer colorretal é um tumor maligno que se desenvolve na parede interna do intestino grosso (cólon) ou reto (ALTERI; KRAMER; S, 2014). As principais motivações para o estudo são o número de casos e a mortalidade. Esse é o terceiro câncer mais comum em homens (746 mil casos) e o segundo mais comum em mulheres (614 mil casos) (SIEGEL; MILLER; JEMAL, 2020). O número de mortes foi 694 mil, com a maior incidência nas regiões menos desenvolvidas do mundo, a uma taxa de 52%. No Brasil, estima-se que, para cada ano do triênio de 2020-2022, haverá 20.520 casos de câncer de cólon e reto em homens e 20.470 em mulheres (INCA, 2019). Em 2020, projeta-se que 78.300 novos casos para homens e 69.650 novos para mulheres ocorram nos Estados Unidos: as mortes estimadas

são 28.630 para homens e 24.570 para mulheres (SIEGEL; MILLER; JEMAL, 2020).

O linfoma é um tipo de câncer que se origina no sistema linfático e, a partir do padrão de crescimento e das características citológicas das células, pode ser classificado em linfomas de Hodgkin (HL) e linfomas não-Hodgkin (NHL) (SIEGEL; MILLER; JEMAL, 2020; XERRI et al., 2016). Em 2020, estima-se que 8.848 novos casos do tipo HL e 77.240 casos do NHL sejam diagnosticados nos Estados Unidos, com aproximadamente 970 e 19.940 mortes, respectivamente (SIEGEL; MILLER; JEMAL, 2020). No Brasil, são esperados, para cada ano do triênio 2020-2020, 6.580 casos em homens e 5.450 em mulheres (INCA, 2019). Considerando sua maior incidência, o NHL é o principal foco de pesquisa. Os linfomas NHL são divididos em três grupos: MCL (*mantle cell lynphoma*), FL (*follicular lymphona*) e CLL (*chronic lymphocytic leukemia*).

1.2 Objetivos

O objetivo principal deste trabalho foi explorar técnicas de aprendizado de máquina para desenvolver um método capaz de fornecer as melhores combinações de técnicas e características, a partir de um AG, para auxiliar na classificação e no reconhecimento de padrões de imagens médicas H&E. Especificamente, buscou-se definir:

1. Um modelo baseado em AG capaz de definir combinações entre características, métodos de seleção e classificadores para o reconhecimento de padrões em linfomas Não-Hodgkin e câncer colorretal;

2. Associações de métodos capazes de distinguir os grupos sob investigação, com desempenho relevante e o menor número de características, além do detalhamento sobre as caraterísticas selecionadas;

3. Aplicar o modelo proposto em imagens H&E de câncer colorretal e linfomas, que apresentam altas taxas de mortalidade.

4. Diferentes testes com os parâmetros mais relevantes, tais como tamanhos de populações

e número de iterações, a fim de classificar os grupos benignos e malignos de imagens colorretais, bem como os grupos MCL, FL e CLL (linfomas Não-Hodgkin).

1.3 Organização do Trabalho

O presente trabalho foi estruturado em cinco capítulos, além da Introdução. No Capítulo 2, é apresentada a Fundamentação Teórica sobre alguns assuntos relacionados ao tema, tais como informações sobre os algoritmos genéticos, técnicas de seleção de características, classificadores e métricas para avaliação de desempenho. No Capítulo 3, encontra-se um detalhamento sobre os trabalhos relacionados, visando obter as informações necessárias para a estruturação do modelo. No Capítulo 4, estão descritas as etapas que embasam o método proposto. No Capítulo 5, os resultados são apresentados. Por fim, no Capítulo 6, as conclusões são apresentadas, assim como uma perspectiva sobre os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são apresentados os conceitos necessários para o desenvolvimento deste trabalho. Na subseção 2.1, a teoria dos Algoritmos Genéticos (AG) é apresentada para mostrar como essas técnicas foram criadas e no que foram fundamentadas, explicitando, assim, o porquê de os AGs serem um método de busca e otimização com vasta aplicação. Na subseção 2.2, os três tipos de técnicas de seleção de características (filtros, *wrappers* e *embedded*) são apresentados e exemplificados. A subseção 2.3 tem como propósito detalhar aspectos importantes sobre os tipos de classificação supervisionada e não supervisionada, apresentar métricas de desempenho (acurácia e curva ROC) e técnicas de validação cruzada. Por fim, na subseção 2.4, encontram-se as considerações finais sobre o capítulo e como os conceitos apresentados foram aplicados.

2.1 Algoritmos Genéticos (AG)

A teoria da evolução das espécies ou darwinismo é considerada uma das maiores descobertas da Ciência. John Holland publicou, em 1975, o livro "*Adaptation in Natural and Artificial Systems*", considerado como a Bíblia dos Algoritmos Genéticos (AGs) (HOLLAND, 1992). Holland utilizou como base os princípios do darwinismo para encontrar a melhor solução para um problema (LUCAS, 2002). Alguns exemplos destes princípios são:

- Disputa: indivíduos de espécies iguais ou diferentes concorrem por recur-

sos limitados em um ambiente;

- Adaptação: alguns indivíduos possuem maior probabilidade de sobrevivência e reprodução;

- Reprodução: indivíduos que mais se reproduzem propagam suas características boas para as gerações seguintes.

Os AGs são técnicas heurísticas de otimização global empregadas na resolução de problemas que geralmente demandam a avaliação de um número considerável de possíveis soluções. Estas técnicas utilizam modelos computacionais baseados no processo biológico de evolução das espécies e permitem que uma população (amostra) de indivíduos heterogêneos (cromossomos) evolua, iterativamente, por meio da aplicação de operadores genéticos. O resultado esperado é uma solução aceitável, talvez a melhor possível (LINDEN, 2006). Os AGs são compostos por: indivíduos, que representam possíveis soluções para o problema; função de avaliação, que mensura a qualidade de cada solução; operadores genéticos, como reprodução e mutação, que garantem a variabilidade genética; e seleção, que mimetiza a seleção natural das espécies.

Os indivíduos são considerados o principal componente de um AG pelo fato de permitir a codificação de uma solução para o problema a ser tratado. Um indivíduo se resume ao seu genótipo, conjunto de genes que ele possui, e ao seu fenótipo, que sintetiza as características daquele indivíduo a partir da decodificação de seus genes, sejam elas morfológicas, fisiológicas ou comportamentais (LUCAS, 2002). Na Tabela 2.1, são exemplificados dois genótipos e fenótipos de problemas comumente abordados na literatura.

Genótipo	Fenótipo	Problema
BCDAE	Ponto B, C, D, A e por fim chego em E	Problema do Caixeiro Viajante
0011	3	Otimização Numérica

Tabela 2.1: Genótipos e Fenótipos (Fonte: Adaptado de Linden, (2006)).

Um conjunto de indivíduos constitui uma população. A evolução como se conhece só é possível graças à dinâmica populacional, responsável pela propagação de características

desejáveis por gerações (LINDEN, 2006). Uma população agrega quatro características: a geração, que diz respeito ao número de vezes que ela passou pelo processo completo de evolução (seleção, reprodução e mutação); a média de adaptação; o grau de convergência, que representa quão próxima a média de adaptação de um valor é considerada ótima; a diversidade, que indica a variedade dos genes em uma população e a manutenção da amplitude de busca; e a elite, composta pelos indivíduos mais adaptados de uma população. Uma técnica comum nos AGs é o elitismo, em que os *m* indivíduos com maior grau de adaptação são mantidos a cada geração (LINDEN, 2006).

É importante ressaltar que, a cada iteração do algoritmo, a população é submetida a uma função de avaliação, denominada *fitness*, responsável por quantificar a qualidade de cada indivíduo (combinação genética) como solução do problema e gerar um ranking de aptidão dos indivíduos no ambiente em que se encontram. Esse processo permite identificar os indivíduos que sobreviverão à seleção natural. Existe uma proporção entre os indivíduos selecionados e aqueles eliminados ao final de cada iteração do algoritmo, um valor imutável que indica quantos indivíduos devem sobreviver e se reproduzir, bem como quantos devem ser descartados (LINDEN, 2006). Este valor imutável denomina-se *threshold* e deve ser definido na inicialização do algoritmo. O *threshold* expressa objetivamente a operação de seleção natural aplicada sobre a população ao final de cada iteração. Para exemplificar, considere um AG inicializado com uma população formada por P indivíduos e com valor de *threshold* $t = 0.7$. Portanto, ao final de cada iteração, os $P(1 - t)$ indivíduos menos aptos serão eliminados. A quantidade de indivíduos mais aptos e que devem sobreviver para a próxima iteração é dado pelo produto entre a quantidade de indivíduos e a taxa da população que sobreviverá para a próxima iteração. A escolha dos indivíduos pais pode ser realizada a partir de diferentes estratégias, compatíveis com os problemas que se busca resolver. Assim, por exemplo, dois indivíduos pais são submetidos ao operador genético de *crossover* e concebem dois novos indivíduos (filhos). Seguindo o exemplo dado, $P(1 - t)/2$ pares de indivíduos mais aptos dentre os que sobreviveram são tomados como pais e conceberão $P(1 - t)$ novos indivíduos filhos a cada iteração.

Antes de serem integrados à população da próxima geração, esses novos indivíduos são

submetidos, individualmente, ao operador de mutação. Dado um valor m (previamente definido) que representa as taxas de mutação consideradas no GA, tal que $m \in \mathbb{R}$ e $m \in (0, 1)$ e j (sorteado aleatoriamente, tal que $j \in \mathbb{R}$ e $j \in (0, 1]$), então o indivíduo avaliado sofrerá mutação se $j \leq m$. Aplicado esse procedimento em todos os indivíduos da população, o algoritmo estará pronto para ser executado por mais uma iteração. Um resumo da estrutura descrita está no Algoritmo 1.

Algoritmo 1 Estrutura Básica de um Algoritmo Genético (Fonte: Adaptado de Furtado e Lorena, (1998)).

```
1: geração ← 0
2: População ← Inicia População
3: enquanto geração menor ou igual a Iter faça
4:     Avalia População(População(geração))
5:     Indivíduos Pais ← Seleciona Aptos(População(geração))
6:     Indivíduos Filhos ← Cruzamento(Indivíduos Pais)
7:     Indivíduos Filhos ← Mutação(Indivíduos Filhos)
8:     População(geração+1) ← Mesclar(Indivíduos Pais,Indivíduos Filhos)
9:     geração ← geração+1
10: fim enquanto
```

2.2 Seleção de Características

Para o correto entendimento do processo de seleção de características, é importante definir o que é uma característica. Uma característica é uma propriedade individual mensurável do processo em observação (CHANDRASHEKAR; SAHIN, 2014). É algo que se pode extrair e enumerar e que define, mesmo que parcialmente, um objeto pertencente a algum domínio. A quantidade de características que podem ser extraídas de um objeto depende apenas do número de técnicas de quantificação disponíveis, da forma e da perspectiva sob as quais o processo é observado. Assim, dependendo da técnica ou método empregado no processo de extração das informações de um domínio, é provável que existam informações repetidas, redundantes e irrelevantes, que não contribuem efetivamente para a sua definição, e informações que, se consideradas, podem prejudicar as taxas obtidas em classificações. Portanto, a remoção de tais informações torna-se benéfica para o processo de descrição efetiva de um domínio e, consequentemente, na diferenciação de classes de dados.

O papel de técnicas de seleção é identificar um subconjunto de variáveis (características) de um domínio dado como entrada, por exemplo, imagens médicas, de forma que possam descrevê-lo eficientemente, reduzindo efeitos de ruído e o número de variáveis irrelevantes, melhorando os desempenhos dos algoritmos de classificação. Para isso, as técnicas de seleção precisam de um critério bem-definido que indique a relevância de cada característica para diferenciar classes de dados distintas (GUYON; ELISSEEFF, 2003) (GNANA; APPAVU; LEAVLINE, 2016). Por exemplo, na área de bioinformática, a motivação para a aplicação de técnicas de seleção de características passou de um exemplo ilustrativo para se tornar um pré-requisito real para o modelo de construção (SAEYS; INZA; LARRAÑAGA, 2007) visando: a) evitar o sobreajuste (*overfitting*) e melhorar o desempenho do modelo; (b) fornecer modelos mais rápidos e econômicos; e (c) obter uma visão mais profunda dos processos subjacentes que geraram os dados.

Existem três tipos de métodos de seleção de características que podem ser utilizados: filtros, *wrappers* e *embedded* (SAEYS; INZA; LARRAÑAGA, 2007). Os filtros avaliam a relevância das características ao analisar apenas as propriedades intrínsecas dos dados, atuando como métodos de pré-processamento, independentemente do método de classificação. Esse processo produz um ranking de relevância das características para diferenciação entre classes, mas ignoram a interação com o classificador e a maioria das técnicas propostas é invariável (cada característica é considerada separadamente, o que pode levar a uma classificação ruim em alguns casos) (SAEYS; INZA; LARRAÑAGA, 2007).

Os *wrappers* (ou "envelopamento") incorporam a pesquisa de hipóteses do modelo dentro da pesquisa do subconjunto de recursos, combinando técnicas de classificação com algoritmos de busca para encontrar o subconjunto de características que maximizam o desempenho de predição. Nesta configuração, um procedimento de busca no espaço de possíveis subconjuntos de características é definido e vários subconjuntos de recursos são gerados e avaliados (SAEYS; INZA; LARRAÑAGA, 2007). A avaliação de um subconjunto é obtida via treinamento e teste de um modelo de classificação específico, tornando essa abordagem adaptada a um algoritmo de classificação específico. No entanto, à medida que o espaço dos subconjuntos de recursos cresce exponencialmente, métodos de pesquisa heurística são usados para

auxiliar na busca de um subconjunto ótimo. A desvantagem desse tipo de técnica é o risco de *sobreajuste*, ainda mais se o classificador associado possuir custo computacional elevado (INZA et al., 2004).

E, por fim, os *embedded* (ou métodos incorporados) incluem a seleção como parte do processo de treinamento do classificador, não sendo necessário, portanto, separar dados entre conjuntos de treinamento e teste. A vantagem desse tipo de método consiste na inclusão da interação com o método de classificação, demandando um custo computacional menor que os *wrappers* (SAEYS; INZA; LARRAÑAGA, 2007).

2.3 Processo de Classificação de Dados

O processo de classificação consiste na discriminação de dados entre as diferentes classes avaliadas por meio do conhecimento obtido a partir de instâncias (amostras) de treinamento fornecidas previamente (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Esse tipo de técnica é parte fundamental do aprendizado de máquina supervisionado e o seu desempenho depende diretamente da qualidade das características consideradas e da quantidade das instâncias de treinamento fornecidas. Nas etapas de treinamento e teste de um classificador, é possível considerar todos os características disponíveis das instâncias ou apenas aqueles indicados como mais relevantes por métodos aplicados de seleção de características. Existem diversos tipos de classificadores que podem ser utilizados e avaliados, os quais estão descritos na próxima subseção.

2.3.1 Tipos de Classificadores

As técnicas de classificação devem ser capazes de diferenciar os objetos e tal diferenciação pode se provar eficiente quando uma imagem com n objetos, cada um pertencendo a n classes diferentes, são identificados corretamente (WEISS; KAPOULEAS, 1990). Há duas abordagens básicas para as técnicas de classificação: a supervisionada e a não supervisionada (CONCI; AZEVEDO; LETA, 2008). Na classificação supervisionada, um conjunto de objetos conhecidos e pertencentes a diferentes classes é analisado com base em

parâmetros ideais para separar as classes. Nesse tipo de classificação, um conjunto de objetos conhecidos e representativos de diferentes classes é analisado por meio de funções discriminantes (CONCI; AZEVEDO; LETA, 2008). As etapas do processo de classificação supervisionada consistem em: escolher um conjunto de treinamento representativo da população; escolher os parâmetros relevantes a serem medidos; obter a função discriminante via método não estatístico ou estatístico; eliminar parâmetros não relevantes; realizar testes com objetos fora do conjunto de treinamento; e, por fim, atualizar os dados. No que diz respeito ao processo de classificação supervisionada, diferentes abordagens disponíveis na literatura podem ser utilizadas, tais como as baseadas em lógica (ou simbólicos), como árvores de decisão (J48 e *Random Forest*); as baseadas em *perceptrons*, como *Multilayer Perceptron*; as baseadas em instâncias, como KNN e *KStar*; as abordagens estatísticas, como *Naive Bayes* e Lasso; *Support Vector Machines*; entre outras (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

As técnicas baseadas em lógica (ou simbologia) buscam descrever os dados de treinamento por meio de um conjunto de símbolos. Nas árvores de decisão, o *dataset* de treinamento é representado via uma árvore (grafo conexo e acíclico) em que cada nó representa uma característica considerada e cada ramo um valor que o nó pode assumir. A classificação é efetuada percorrendo a árvore da classe a partir da raiz até os nós-folhas (ramos) (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). A análise de cada nó é realizada obtendo uma resposta do tipo “Sim” ou “Não”, indicando se pertence ou não à classe da árvore analisada. Por outro lado, os métodos baseados em perceptrons utilizam os princípios apresentados por Rosenblatt ((1962)). Nesse modelo, cada instância possui um vetor de pesos atribuído em que a combinação linear produz um valor que é avaliado e utilizado para indicar se pertence ou não a uma dada classe (MITCHELL et al., 1997).

As técnicas baseadas em instâncias são aquelas que efetuam a classificação de uma instância comparando-a com um banco de dados de exemplos pré-classificados, assumindo que instâncias similares pertencerão a classes similares (CLEARY; TRIGG, 1995). Assim, cada vez que uma nova solicitação de classificação é realizada, a relação com todos os outros exemplos anteriores é analisada para se obter um novo valor a partir da função de avaliação.

Por conta desse *delay* de processamento, tais técnicas também se encontram referenciadas na literatura como técnicas de *lazy learning* (MITCHELL et al., 1997). Adicionalmente, as abordagens estatísticas são caracterizadas por utilizarem explicitamente modelos probabilísticos que fornecem n valores, dado por $P(i, C)$, e ($i \in (0, 1)$) representando a probabilidade da instância i pertencer à classe C.

Os classificadores *Support Vector Machines* (SVM) constituem as técnicas mais atuais no ramo de aprendizado de máquina (VAPNIK, V., 2013). O modelo SVM utiliza o princípio de valor marginal; o objetivo é encontrar um valor que separa as instâncias de duas classes distintas (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). O classificador SVM é capaz de separar um conjunto de dados binários a partir de um hiperplano maximamente distante deles. Quando uma classificação linear não pode ser aplicada, as máquinas podem trabalhar em conjunto com uma técnica baseada em *kernels* (núcleos), que automaticamente realiza um mapeamento não linear para um espaço de características (FUREY et al., 2000). Por tentar minimizar o limite superior do erro generalizado por meio da maximização da margem entre o hiperplano de separação e o conjunto de dados treinados, uma SVM pode superar estimativas estatísticas pobres, geralmente associadas a métodos tradicionais que tentam minimizar erros empíricos de treinamento (KUO et al., 2014).

Por fim, a classificação não supervisionada é aplicada em contextos nos quais as informações sobre as classes não estão disponíveis (CONCI; AZEVEDO; LETA, 2008). Nesse tipo de classificação, por exemplo, os agrupamentos (*clusters*) formados por *pixels* com características semelhantes são identificados e servem de base para a análise realizada pelo algoritmo (DAINESE, 2001). Segundo Picoli ((2013)), esse tipo de algoritmo tem um custo de processamento maior, mas permite identificar padrões não conhecidos inicialmente, diferentemente do que é observado na classificação supervisionada, em que os algoritmos precisam da rotulagem prévia dos padrões.

		Previsto	
		Positivo	Negativo
Real	Positivo	a	b
	Negativo	c	d

Tabela 2.2: Estrutura de uma matriz de confusão (Fonte: Elaborado pela autora).

2.3.2 Métricas para Avaliação de Desempenho

A qualidade dos resultados obtidos a partir de um método de classificação pode ser avaliada segundo algumas métricas. Dentre as mais utilizadas na literatura estão a taxa de acurácia e a área sob a curva ROC (*Receiver Operating Characteristic*).

A taxa de acurácia (Acc) é utilizada para indicar o desempenho dos classificadores na diferenciação de grupos na área de sistemas de diagnósticos, por exemplo (HUANG; LING, 2005). Os valores de acurácia são comumente representados no intervalo de 0 a 100%. Assim, caso o classificador tenha aferido corretamente a classe de todas as amostras, o valor de acurácia obtido será de 100%. Esse método avalia a proporção existente entre o número de instâncias classificadas corretamente (verdadeiros positivos) e o número total de instâncias, incluindo aquelas classificadas erroneamente. Para mensurar a acurácia, uma matriz de confusão é utilizada, na qual cada coluna representa as instâncias da classe prevista e cada linha representa as instâncias reais da classe, conforme ilustrado na Tabela 2.2. O valor a refere-se ao número de predições realizadas corretamente de uma instância positiva, o b refere-se ao número de predições realizadas incorretamente de uma instância negativa. O valor c consiste no número de predições incorretas de uma instância positiva e, por fim, d representa o número de predições corretas de uma instância negativa.

Com base na Tabela 2.2, é possível calcular as taxas de verdadeiros positivos (VP) e de falsos positivos (FP). A taxa *VP* é encontrada ao dividir a taxa de verdadeiros positivos a pela soma dos falsos positivos com os verdadeiros positivos, conforme apresentado na Equação 2.1.

$$VP = \frac{a}{a+b}. \tag{2.1}$$

A taxa de falsos positivos (FP) é encontrada ao dividir a taxa de falsos positivos c pela soma da taxa de falsos positivos e falsos negativos, conforme apresentado na Equação 2.2.

$$FP = \frac{c}{c+d}. \tag{2.2}$$

Por fim, a relação que define e permite calcular a Acc é dada pela Equação 2.3.

$$Acc = \frac{VP}{VP+FP}. \tag{2.3}$$

Na Medicina, a confirmação da eficiência de uma hipótese se baseia em testes com alta sensibilidade e especificidade. A sensibilidade é "a positividade na doença", indicando a capacidade em determinar quais são os pacientes efetivamente doentes dentre os suspeitos. A especificidade é a capacidade do teste em determinar que certos pacientes não estão doentes (GUIMARÃES, 1985). Testes com sensibilidade e especificidade produzem resultados claros e um ponto de corte é buscado para classificar indivíduos como doentes e não doentes, por exemplo (MARTINEZ; LOUZADA-NETO; PEREIRA, 2003). As taxas para a sensibilidade (s) e especificidade (e) estão representadas pelas Equações 2.4 e 2.5 (CRISTIANO, 2017).

$$s = \frac{VP}{VP+FN}. \tag{2.4}$$

$$e = \frac{VN}{VN+FP}. \tag{2.5}$$

Um gráfico ROC é definido pela sensibilidade (eixo y) contra $1 - especificidade$ do sistema. Um exemplo de gráfico ROC é indicado na Figura 2.1. É útil para visualizar, organizar e selecionar classificadores baseando-se em seu desempenho (FAWCETT, 2006).

A área sob a curva ROC (AUC) representa a probabilidade de um classificador ranquear uma instância positiva acima de uma instância negativa, ambas aleatoriamente escolhidas. Por conta de uma AUC ser uma porção da área de um quadrado unitário, seu valor varia no intervalo $[0, 1]$. Contudo, em nenhuma situação real um classificador pode obter desempenho

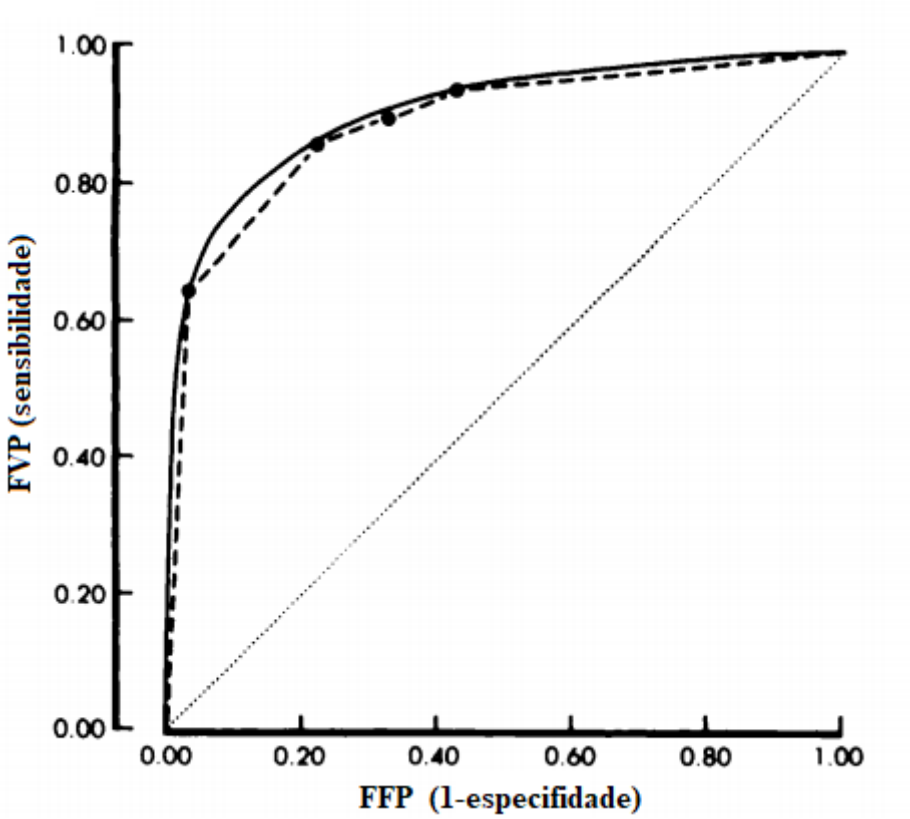


Figura 2.1: Exemplo de uma curva ROC (Fonte: Adaptado de Hanley e McNeil, (1982)).

pior do que o de um classificador aleatório (aquele em que se obtém, no mínimo, 50% de acertos). Uma área de 1 representa um teste perfeito e uma área de 0,5 representa um teste sem relevância (TAPE, 2006). Os níveis de relevância propostos por Tape ((2006)) associam letras às áreas do gráfico e estão representados na Tabela 2.3. Os resultados desejáveis são associados às letras (A), (B) e (C).

Classificação	Área sob a curva ROC (AUC)
Excelente (A)	0.9 a 1
Muito bom (B)	0,8 a 0,9
Bom (C)	0,7 a 0,8
Pobre (D)	0,6 a 0,7
Falha (E)	0,5 a 0,6

Tabela 2.3: Níveis de relevância de um teste medidos pela AUC (Fonte: Elaborado pela autora).

Além disso, para testar e generalizar o desempenho do método, os dados podem ser organizados a partir de três métodos de validação cruzada (*cross-validation*): *Hold-Out*, *K-*

folds e *Leave-One-Out* (LOO). No método *Hold-Out* (KIM, 2009), quando nenhuma amostra independente é fornecida, uma parte do conjunto de treinamento é selecionada e "mantida" (*hold-out*). O classificador é construído somente com o conjunto remanescente. Geralmente, o conjunto de treino é definido com uma proporção de 70% e o de teste em 30% (KIM, 2009). A taxa real de erro será estimada e o processo é repetido até que, por fim, a média de todos os erros seja calculada e utilizada como resposta. No segundo método (*K-folds*), as amostras são distribuídas em k subconjuntos exclusivos, sendo $k-1$ subconjuntos utilizados para treinamento e o subconjunto remanescente definido para teste. Esse procedimento é repetido k vezes e alterna-se o subconjunto de teste, resultando em vários valores de acurácia (ARLOT; CELISSE et al., 2010). A estimativa de erro será o valor médio dos erros em todos os *folds*. Assim, esse tipo de método será dependente de dois fatores: o conjunto de teste e a divisão dos *folds*. Por fim, o método *Leave-One-Out* (GUYON; ELISSEEFF, 2003) é uma especificação da estratégia *K-folds*, consistindo em remover uma única amostra do conjunto de treinamento, testar somente os dados de treinamento remanescentes e, em seguida, a amostra é removida. Com alto custo computacional, esse tipo de estratégia é recomendado quando são poucos os dados disponíveis.

2.4 Considerações sobre a Fundamentação Teórica

Neste capítulo, foi apresentado o embasamento teórico necessário para o desenvolvimento da proposta do trabalho. Por exemplo, foram detalhados conceitos importantes sobre a teoria dos AGs, como o princípio de Darwin, a definição de indivíduos e populações. Foram também definidos os métodos de seleção, cruzamento e *crossover*. A definição do conceito de características e a apresentação dos métodos que podem ser aplicados para sua avaliação também foram descritas em detalhes, tais como filtros, *wrappers* e *embedded*. Os principais conceitos sobre algoritmos para classificação de dados também foram discutidos, destacando as diferenças entre os métodos supervisionados e os não supervisionados. Neste contexto, os conceitos apresentados são importantes, visto que o foco será observar métodos supervisionados, buscando a melhor combinação entre características, métodos de seleção

e classificadores, via um AG, e analisando uma das principais métricas para avaliação de desempenhos de sistemas: a AUC.

Capítulo 3

Trabalhos Relacionados

Na literatura, o Algoritmo Genético (AG) foi amplamente explorado para identificar as melhores características e favorecer a obtenção de desempenhos mais relevantes no processo de classificação.

Neste capítulo, é apresentada uma visão detalhada sobre os trabalhos relacionados com esse contexto. Na subseção 3.1, encontram-se os trabalhos que aplicaram o AG para analisar características representativas de contextos diversificados (conjuntos de imagens variados). Na subseção 3.2, são descritos os trabalhos em que o AG foi aplicado somente em conjuntos de imagens médicas. Por fim, na subseção 3.3, estão algumas considerações sobre os trabalhos coletados.

3.1 Aplicação do AG em Contextos Diversificados

Kuncheva e Jain ((2000)) propuseram duas abordagens baseadas em um AG: a primeira versão do AG visou indicar subconjuntos de características disjuntas; a segunda selecionou subconjuntos de características com possíveis sobreposições. Os conjuntos foram dados como entradas para os classificadores *Linear discrinant*, *Quadratic discriminant* e *Logistic*. A função *fitness* foi construída baseada na acurácia de todos os classificadores. As estratégias foram testadas em conjuntos de dados definidos como *Heart*, *Satimage*, *Letters* e *Forensic*. Os autores concluíram que a segunda versão produziu a menor taxa de erro e, na maioria dos

casos testados, as duas versões superaram outros métodos disponíveis na literatura.

Um método baseado em AG e classificador ADABOOST como função *fitness* foi explorado por Chouaib et al. ((2008)). O método foi dividido em duas etapas: treinamento do classificador ADABOOST para cada característica e uso do AG para selecionar o melhor classificador ou conjunto de características. O AG foi definido com 200 cromossomos e 50 iterações. A proposta foi testada em um conjunto de dados de pequenas imagens binárias manuscritas, com números de 0 a 9. Os métodos de classificação foram SVM, ADABOOST e KNN. Os autores concluíram que os resultados são similares aos observados na literatura, porém utilizando cerca de 35% menos características em comparações envolvendo mais de duas classes. Para comparações com duas classes, os autores destacaram que a estratégia utilizou 75% do total de características.

Um AG binário para a seleção de características foi descrito em Babatunde et al. ((2014)). O modelo analisou 100 características com uma população de 100 indivíduos, probabilidade de *crossover* em 0,8 e probabilidade de mutação em 0,1. O algoritmo foi testado com 300 iterações. Os resultados foram comparados aos fornecidos pelos algoritmos de seleção *Information Gain Ranking Filter* e *CFS Subset Evaluator*. Os conjuntos foram classificados com os algoritmos *Multi-Layer Perceptron*, *Random Forest*, J48, *Naive Bayes* e *Classification Using Regression*. Os autores verificaram que o método reduziu os conjuntos de características em até 89%, com uma acurácia máxima de 94,35%, obtida via classificador *Naive Bayes*.

Dennis e Muthukrishnan ((2014)) apresentaram um método baseado em AG e lógica *fuzzy*, o qual foi nomeado como *Adaptive Genetic Fuzzy System* (AGFS). O objetivo foi otimizar regras e funções de associação para o processo de classificação de dados. O processo geral do AGFS foi dividido em três etapas: 1) geração de regras otimizadas utilizando AG, 2) definição automática da função de associação *fuzzy* e 3) classificação via sistema *fuzzy*. Os autores utilizaram probabilidade de cruzamento de 0,5 e a taxa de mutação foi adaptada automaticamente para obter os melhores desempenhos. O método foi testado em conjuntos de dados nomeados como *Glass*, com onze características; *Wine*, com 13 características; *Pima Indian Diabetes*, com nove características; e *Iris*, com quatro características. O método

foi comparado com outras estratégias e forneceu resultados melhores nos conjuntos *Glass* (acurácia de 86,05%) e *Pima Indian Diabetes* (taxa de acurácia de 89,80%).

Por fim, um método foi proposto a partir da abordagem *Principal Component Analysis* para reduzir a dimensão do conjunto de características, AG como uma estratégia para definir as melhores combinações de características e SVM como classificador (ZHI; LIU, 2019). O método foi aplicado em um banco de imagens de faces do *Institute of Technology of Chinese Academy of Sciences*, composto por 99.450 imagens de 595 homens e 445 mulheres. Os autores concluíram que os resultados foram relevantes, sendo que a maior taxa de acurácia foi de 99%.

3.2 Aplicação do AG em Contextos de Imagens Médicas

Yang e Honavar ((1998)) apresentaram uma abordagem multicritério baseada em *wrappers*, visando selecionar subconjuntos de características a partir de uma associação entre AG e uma rede neural. Os autores indicaram que o desempenho do método estava diretamente associado ao número de neurônios escolhidos para compor a rede. Os autores utilizaram uma população de 50 indivíduos, probabilidade de cruzamento igual a 0,6 e de mutação definida como 0,0001. A seleção por ranqueamento foi associada a uma probabilidade de 0,6. O método foi testado em 21 conjuntos de imagens, sendo oito de imagens médicas, disponíveis no repositório de *Machine Learning* da Universidade da Califórnia. Os autores concluíram que a associação proposta foi capaz de identificar apropriadamente as características com taxas de acurácias superiores a 90%.

A combinação entre programação genética, AG e o classificador *Decision Tree* (J48) foi explorada por Smith e Bull ((2005)), com o objetivo de indicar as características mais relevantes em um conjunto dado como entrada. A programação genética foi aplicada com uma população de 101 genótipos. Cada um dos genótipos foi constituído por árvores, em que cada uma representou o número de características de um subconjunto. Cada um dos genótipos foi avaliado via classificador *Decision Tree* (J48). O processo de evolução considerou seleção, *crossover*, mutação e avaliação da população. O melhor genótipo foi passado para a próxima

geração, considerando a estratégia de seleção por torneio. Os testes foram realizados em 10 bases de dados conhecidas do repositório Universidade da Califórnia, sendo 4 delas de imagens médicas: *Liver*, *Glass*, *Ionosphere*, *New Thyoid*, *Sonar*, *Vehicle*, *Wine*, *Pima Indian Diabetes*, *Wisconsin Breast Cancer – New* e *Wisconsin Breast Cancer – Original*. A maior acurácia obtida foi de 96,27% em comparação aos 92,31% fornecidos a partir do conjunto inicial de características.

Um novo modelo tridimensional de fusão, classificação e seleção foi proposto por Gabrys e Ruta ((2006)) a partir de um AG. O método seguiu dois princípios: o primeiro fez uma análise do potencial dos pesos das características para determinar a melhor combinação com os classificadores; o segundo envolveu um modelo mais geral, no qual o AG foi aplicado no processo para a fusão de classificadores (dados, características e combinação de classificadores). Os autores dividiram o método em seis etapas principais: 1) coletar e fixar o espaço de seleção com características, classificadores e possíveis combinações; 2) inicializar a população aleatória com um número fixo de cromossomos; 3) aplicar a mutação e *crossover* de duas etapas; 4) unir os indivíduos pais aos filhos para o cálculo da função *fitness*; 5) selecionar os melhores cromossomos para a próxima geração; 6) finalizar se ocorrer convergência. Para os experimentos, os autores utilizaram 15 diferentes classificadores, baseados em SVM, modelos estatísticos e algoritmos específicos, tais como *Mean Combiner*, *Minimum Rule Combiner*, *Maximum Rule Combiner*, *Product Rule Combiner* e *Majority Voting Combiner*. O método foi aplicado em dois conjuntos de imagens (íris e fígado). Os autores concluíram que a proposta foi capaz de fornecer resultados superiores aos disponibilizados na literatura.

Um AG também foi aplicado na seleção de características para o diagnóstico de doenças cardíacas (ANBARASI; ANUPRIYA; IYENGAR, 2010). Três classificadores foram utilizados para detectar e eliminar características irrelevantes: *Naive Bayes*, classificação por *Clustering* e árvore de decisão. O método reduziu o conjunto de 13 para seis características. A implementação do método considerou uma população inicial com 20 instâncias, geradas aleatoriamente, e dois tipos de métodos de seleção: a) por ranking, na qual as características foram classificadas via métricas baseadas no ganho de informação; e b) por seleção de sub-

conjuntos, avaliado a partir de métricas baseadas em distâncias. Os indivíduos descendentes foram gerados aplicando as estratégias de *crossover* e mutação. O processo de geração foi executado até obter uma população que atendesse a função *fitness*. Segundo os autores, a melhor acurácia (99,2%) foi obtida via um algoritmo árvore de decisão, enquanto os algoritmos *Naive Bayes* e *Clustering* apresentaram acurácias de 96,5% e 88,3%, respectivamente.

O estudo da classificação de pacientes diabéticos foi o foco principal do trabalho apresentado por Velu e Kashwan ((2013)), a partir do conjunto de dados *Pima Indian Diabetes* (PID). Esta pesquisa foi baseada em três técnicas: *Expectation-Maximization Algorithm*, *h-means+* e AG. As técnicas foram empregadas para formar *clusters* e encontrar semelhanças entre características. A análise dos resultados provou que a combinação *h-means+* e o AG com *crossover* duplo forneceu um desempenho superior nas comparações de desempenhos. Os testes foram realizados aplicando três distintos classificadores. Por fim, quatro características (dos sete investigados) foram associados com a condição de diabetes.

Özçift e Gülten ((2013)) apresentaram um método baseado em AG para o diagnóstico diferencial de doenças eritêmato-escamosas. As doenças analisadas foram psoríase, dermatite seborreica, líquen plano, pitiríase rosa, dermatite crônica e pitiríase rubro pilar. Na implementação, os autores consideraram cada indivíduo como um cromossomo composto por 34 genes, que foi a quantidade de características no conjunto de dados. A população inicial foi gerada aleatoriamente e os métodos de roleta e *crossover* de dois pontos foram utilizados para a seleção e o cruzamento. O método *Bayesian Network* foi aplicado como função *fitness*, ranqueando os genes do cromossomo a cada iteração. O resultado foi uma acurácia de 99,2%, valor que superou os desempenhos fornecidos por classificadores conhecidos, tais como SVM, *Multi-Layer Perceptron*, *Simple Logistice* e *Functional Decision Tree*.

Lu, Zhu e Gu ((2014)) propuseram um sistema para o diagnóstico de câncer de pulmão. O sistema visou selecionar as melhores características a partir de um AG com uma estratégia nomeada *Mutual Information* (MI). A abordagem MI foi aplicada para capturar as informações de correlações não lineares entre as características. As soluções fornecidas pelo método foram usadas para treinar três classificadores: SVM, *Back Propagation Neu-*

ral Network e o *K-Nearest Neighbors*. O melhor resultado foi definido com o classificador SVM.

Uma combinação envolvendo um AG e o classificador SVM, nomeado de "SVM genético", foi descrito em Kumar, Ramachandra e Nagamanii ((2014)). O classificador SVM foi treinado com as características selecionadas pelo AG. O método foi testado em cinco conjuntos de dados do repositório de aprendizado de máquina da Universidade da Califórnia: câncer de mama, *Pima Indian Diabetes*, *mammographic mass*, dermatológica e de cirurgia torácica. Os resultados obtidos mostraram que o método proposto foi capaz de encontrar um subconjunto de características ótimo, com reduções importantes no número de características utilizadas em cada teste.

Um AG híbrido foi proposto por Nagarajan et al. ((2016)) para extrair e selecionar características em imagens médicas. O método foi dividido em três fases. Na primeira, três algoritmos (*Texton based contour gradient extraction algorithm*, *Intrinsic pattern extraction algorithm* e *modified shift invariant feature transformation algorithm*) foram usados para extrair as características das imagens. Na segunda, os vetores de características foram analisados para identificar os conjuntos mais relevantes para os contextos de câncer de mama, tumor cerebral e imagens da tiroide, considerando uma abordagem híbrida entre os algoritmos *Branch and Bound* e *Artificial Bee Colony Algorithm*. O *Branch and Bound* foi aplicado para definir um conjunto de características menor ou igual ao original. As características mais relevantes foram obtidas a partir da união entre os dois conjuntos. Na terceira fase, o algoritmo *Diverse Density* foi aplicado para melhorar o desempenho do método. Os autores concluíram que, apesar do método identificar a maioria das características relevantes, a estratégia falhou na seleção de características previamente identificados como importantes.

Al-Rajab, Lu e Xu ((2017)) apresentaram uma análise de algoritmos de seleção e classificação de características para o diagnóstico de câncer de cólon. Os autores avaliaram 24 algoritmos. A primeira etapa visou identificar os desempenhos dos algoritmos para seleção. O resultado foi que o algoritmo *Particle Swarm Optimization* (PSO) superou o AG. Os autores indicaram que a melhor combinação foi entre PSO e SVM, fornecendo uma acurácia média de aproximadamente 87%.

Um método baseado em programação genética foi composto por dois estágios, um destinado para a extração e outro para a seleção de características, foi apresentado por Ain et al. ((2018)). O modelo foi aplicado no estudo de imagens de câncer de pele (melanoma). O padrão local binário foi a estratégia definida para a extração de características. Os resultados foram organizados em vetores de características para treino e teste do modelo. O vetor de treino foi submetido ao processo completo de evolução, com avaliação, seleção, *crossover* e mutação. Os melhores indivíduos do conjunto de treino foram os selecionados. Os testes foram realizados em um conjunto de 200 imagens de dermatoscopia fornecidos pelo hospital Pedro Hispano (Portugal), sendo 160 do grupo benigno e 40 do grupo maligno. A acurácia máxima fornecida pelo modelo foi de 90.64%.

Bhardwaj et al. ((2018)) indicaram que, apesar do crescimento dos métodos baseados em aprendizado de máquina, os desempenhos desses métodos são afetados quando existem características irrelevantes ou redundantes no conjunto sob investigação. Nesse contexto, os autores propuseram um método para a seleção e classificação de características a partir da programação genética. O método foi aplicado para o diagnóstico de câncer de mama. O programa genético foi descrito por uma árvore para cada um dos indivíduos. Cada indivíduo foi submetido ao processo de evolução e os que forneceram os melhores desempenhos foram selecionados para uma nova geração. O número de gerações foi definido como 100. Os testes foram realizados em dois conjuntos de dados do repositório de aprendizado de máquina da Universidade da Califórnia, WBC (*Winconsin Breast Cancer*) e WDBC (*Winconsin Diagnostic Breast Cancer*). Os resultados também foram comparados aos fornecidos pelo método descrito em Muni, Pal e Das ((2006)). Os autores relataram que a acurácia foi de 100% para o conjunto WBC e 98.24% para o conjunto WDBC.

Para tentar diminuir diagnósticos negativos de câncer de mama em casos positivos (falsos negativos), Gokulnath e Shantharajah ((2019)) apresentaram um método baseado em AG e *Information Gain* para a seleção dos características mais relevantes. Foram utilizados três métodos de classificação para avaliar o desempenho: BPNN, KNN e CSSVM. Para o AG, foi definida uma população inicial de 50 indivíduos, 200 iterações máximas e uma taxa de mutação de $0,1$. Os testes foram realizados em dois conjuntos de dados do repositório de

aprendizado de máquina da Universidade da Califórnia, WBC e WDBC. Considerando as seis comparações de métodos de seleção e classificação apresentadas, os resultados mostraram que, além de conseguir avaliar o custo de classificações incorretas (falsos negativos), o método proposto também conseguiu desempenhos superiores quando comparado a outros métodos da literatura.

3.3 Considerações sobre os Trabalhos Relacionados

Neste capítulo, foi apresentada uma visão detalhada sobre os principais trabalhos que associaram técnicas de seleção e classificação de características a partir de AGs. Os exemplos apresentados estão diretamente relacionados com a proposta deste trabalho, principalmente os apresentados na seção 3.2.

Além disso, é possível constatar que um objetivo comum na maioria dos trabalhos foi o de maximizar os desempenhos com um conjunto reduzido de características (YANG; HONAVAR, 1998; LU; ZHU; GU, 2014; KUMAR; RAMACHANDRA; NAGAMANI, 2014; NAGARAJAN et al., 2016; AIN et al., 2018; GOKULNATH; SHANTHARAJAH, 2019).

Porém, independente do contexto explorado e das valiosas contribuições que cada trabalho proporcionou, não foi possível observar um método que analisasse vários métodos de seleção, classificadores e influências das características sobre os resultados, como pretendido na proposta deste trabalho. O AG foi aplicado como um método de seleção de características, não como um meio para indicar as melhores relações entre algoritmos e características, de maneira integrada, em diferentes contextos de imagens médicas (YANG; HONAVAR, 1998; LU; ZHU; GU, 2014; KUMAR; RAMACHANDRA; NAGAMANI, 2014; NAGARAJAN et al., 2016; AIN et al., 2018).

Os exemplos de modelos que consideraram relações entre mais de um método de seleção e mais de um classificador não incluíram a combinação integrada com características (GABRYS; RUTA, 2006; AL-RAJAB; LU; XU, 2017; GOKULNATH; SHANTHARAJAH, 2019). Por outro lado, o modelo que incluiu os classificadores e os características não abordou os métodos de seleção (BHARDWAJ et al., 2018).

Portanto, um método capaz de realizar uma verificação mais ampla, relacionando diferentes métodos de seleção, classificação e características, pode contribuir significativamente com a literatura especializada no tema. Essa versatilidade é possível via um AG amplamente explorado na seleção de características, mas não como uma estratégia descrita neste trabalho.

Capítulo 4

Metodologia

O método proposto foi organizado em etapas. Na primeira, foi descrita a estrutura cromossômica que representa o núcleo do Algoritmo Genético (AG) e a população inicial. A segunda etapa (Avaliação e seleção da população) possibilita calcular a Área sob a Curva ROC (AUC) produzida por cada indivíduo. Além disso, nessa etapa foram definidos quais indivíduos serão cruzados na próxima etapa, aplicando o cruzamento de dois pontos. Finalmente, na terceira etapa (operadores de *crossover* e mutação) são avaliados quais indivíduos recém-gerados serão mutados antes de serem inseridos na população para avaliação. A construção de cada uma das etapas tem como objetivo indicar a melhor combinação de características, métodos de seleção e classificadores com a maior taxa de AUC utilizando o menor número possível de características.

4.1 Estrutura cromossômica e população inicial

O núcleo do método foi baseado em estrutura de AG. A estrutura considerada de cada indivíduo foi definida da seguinte forma: cada cromossomo é constituído por quatro genes que são representados por números inteiros. A carga genética ou as informações armazenadas em cada genes são: identificador único do indivíduo (G_{id}), índice de um método de seleção (G_{sel}), índice de um método de classificação (G_{clf}) e número de características consideradas (G_{num}). Os valores iniciais para os parâmetros G_{sel}, G_{clf} e G_{num} são definidos

por sorteio (aleatoriamente), mas respeitando os intervalos estabelecidos em cada um. Uma ilustração da estrutura descrita previamente está na Figura 4.1. Essa estrutura permite avaliar o desempenho de uma combinação específica de um método de seleção, um classificador e um número de características. Assim, uma população (P) é definida por um conjunto de indivíduos e determinada por diferentes associações de métodos e características. Cada combinação é única e é identificada pelo gene G_{id} para definir uma solução aceitável.

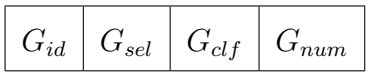


Figura 4.1: Estrutura cromossomial adotada no algoritmo proposto para representar os indivíduos da população, considerando a identificação do indivíduo, índice do método de seleção, índice de método de classificação e número de características.

É importante destacar que cada gene G_{sel} é usado para definir um método capaz de fornecer um ranking das características mais significativas para distinguir as classes sob investigação. Portanto, as técnicas escolhidas foram: *TStatistics*, *Information Gain*, *Relief*, *Gain Ratio*, *Chi-Squared* e *Probabilistic Relevance*. As características foram avaliadas por cada classificador indicado no gene (G_{clf}). Os classificadores aplicados nesta proposta foram: *Decision Tree* (SAFAVIAN; LANDGREBE, 1991), *Random Tree* (BREIMAN, 2001), *Random Forest* (BREIMAN, 2001), *Lasso* (ABINASH; VASUDEVAN, 2019), *Gaussian Naive Bayes* (JOHN; LANGLEY, 1995), *Multinomial Naive Bayes* (JOHN; LANGLEY, 1995), *Bernoulli Naive Bayes* (JOHN; LANGLEY, 1995), *Multilayer Perceptron* (GARDNER; DORLING, 1998), *Support Vector Machine* (SVM) (VAPNIK, V. N., 1999), *K-Nearest Neighbors* (KNN) (GOU et al., 2019), *KStar* (CLEARY; TRIGG, 1995), *Locally Weighted Learning* (LWL) (ATKESON; MOORE; SCHAAL, 1997) e *Stochastic Gradient Descent* (SGD) (ZHUO et al., 2019). As técnicas foram aplicadas em cada um dos conjuntos de treino e teste, usando a validação cruzada *k-folds*, com $k = 10$. O valor de *k* escolhido indica que, a partir do conjunto completo, serão gerados 10 conjuntos para treinamento e 10 conjuntos para teste (KOHAVI et al., 1995). Esse valor de *k* foi utilizado nas propostas descritas por (ROBERTO et al., 2017) (linfomas) e (RIBEIRO, Matheus Gonçalves et al., 2019) (câncer colorretal).

A partir da estrutura cromossomial apresentada previamente, os parâmetros *de entrada associados* ao método proposto são: tamanho da população (P), fornecido pelo usuário; o número máximo de gerações ou iterações ($Iter$), fornecido pelo usuário; um *threshold* (t) para indicar os indivíduos mais aptos para a reprodução, fornecido pelo usuário; a probabilidade de mutação genética (m), fornecido pelo usuário; e, o número de características G_{num} que são obtidas a partir do conjunto inicial de características. O parâmetro G_{num} é definido aleatoriamente, mas respeita o total de características ($MaxF$) que está sob investigação. Um resumo dos parâmetros está na Tabela 4.1.

Parâmetro	Definição	Valor Inicial
P	Tamanho da população	Usuário
$Iter$	Número de iterações do algoritmo	Usuário
t	Limiar de seleção dos indivíduos mais aptos	Usuário
m	Probabilidade de mutação genética	Usuário
$MaxF$	Número máximo de características selecionadas	Sorteio
G_{sel}	Método de seleção considerado na iteração	Sorteio
G_{clf}	Método de classificação considerado na iteração	Sorteio
G_{num}	Número de características considerado na iteração	$\in (0, MaxF)$

Tabela 4.1: Parâmetros necessários para a aplicação do método proposto (Fonte: Elaborado pela autora).

4.2 Avaliação e Seleção da População

A avaliação da população consiste no cálculo da AUC média associada a indivíduo a partir das técnicas de seleção e classificação indicadas nos seus genes. Portanto, para cada

indivíduo da população:

- Aplica-se o método de seleção indicado no gene G_{sel} sobre cada um dos L arquivos de treinamento, obtendo-se as N melhores características (N definido a partir de G_{num});

- O classificador indicado no gene G_{clf} é treinado com as N características selecionadas. Esse processo é realizado para cada conjunto de treinamento construído com a validação cruzada da estratégia (*k*=10). Cada conjunto de teste é definido pelas características selecionadas no conjunto de treinamento correspondente. A classificação é realizada em cada conjunto de teste correspondente. Portanto, para cada indivíduo G_{id}, k valores de AUC são obtidos, um para cada conjunto de teste. É feita a média desses valores e, logo em seguida, é calculada uma nova média para as L iterações realizadas até o momento, representando o valor desejado. A AUC média ($MeanAuc(G_{id})$) considerando a estratégia validação cruzada *k*=10) é calculada aplicando a equação 4.1.

$$MeanAuc(G_{id}) = \frac{\sum_{i=1}^{L} Auc(i)}{L}. \tag{4.1}$$

Nesta proposta, para que o método tenha o comportamento de seleção natural definida por Darwin, os indivíduos com a maior taxa de AUC são selecionados a partir da população ordenada pelo critério $MeanAUC$, ou seja, os indivíduos com as maiores taxas de AUC são os selecionados. Para tanto, considerando a população inicial P e o *threshold* de seleção $t = 0,7$, o número de indivíduos que serão escolhidos é encontrado pelo produto entre P e t. Essa abordagem e valor de t foram definidos a partir da proposta descrita por Yang e Honavar ((1998)). É importante ressaltar que os indivíduos escolhidos são definidos como os pais na próxima geração do algoritmo por reunir genes que indicam as melhores combinações entre técnicas de seleção e classificação.

4.3 Operações de *Crossover* e Mutação

A reprodução é a responsável por complementar a população da geração atual com novos indivíduos produzidos a partir dos selecionados como mais aptos na etapa anterior. Esse tipo de abordagem visa simular a reprodução sexuada encontrada em diversas espécies da natureza com o intuito de perpetuar genes e, ao mesmo tempo, garantir a variabilidade genética de toda a população, uma vez que cada indivíduo filho é gerado com as informações de dois indivíduos pais ao invés de se tornarem cópias genéticas de apenas um preceptor. No algoritmo proposto, a operação genética de *crossover* foi implementada utilizando a abordagem de dois pontos, com o objetivo de buscar a melhor solução via troca dos métodos de seleção de classificação. O *crossover* de dois pontos consiste na escolha de dois locus de um cromossomo como pontos de troca (ou pivôs) e efetuar, alternadamente, a cópia dos genes dos pais para os dois filhos gerados. No caso da estrutura cromossomial considerada neste trabalho, com exceção do gene G_{id}, os demais genes são utilizados nas próximas gerações. Além disso, os genes G_{sel} e G_{clf} são sempre escolhidos como pontos de troca. Esse processo é ilustrado na Figura 4.2, na qual os genes dos filhos C e D foram obtidos a partir do cruzamento entre os genes dos pais A e B.

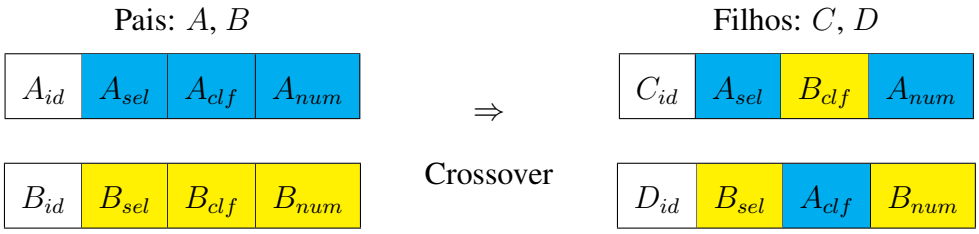


Figura 4.2: Ilustração da formação de dois novos indivíduos C e D por meio da estratégia de *crossover* de dois pontos entre os pais A e B.

Em seguida, o operador de mutação é aplicado sobre os indivíduos filhos para que seja possível definir a população da próxima iteração do algoritmo. A operação de mutação foi definida seguindo alguns procedimentos:

- para cada indivíduo recém-nascido, um número aleatório $\alpha \in \mathbb{R} \mid 0 \leq \alpha \leq$ 1 α é sorteado para indicar se o indivíduo filho sofrerá mutação em algum

dos seus genes (MIRANDA, 2007);

- Se $\alpha > m$, sendo *m* a probabilidade de mutação, o indivíduo não é mutado.
  Caso contrário, $\alpha \leq m$, o indivíduo é submetido ao processo de mutação.

Quando um indivíduo é submetido à mutação, um novo número aleatório β é sorteado para indicar qual índice do gene deve ser mutado. A variável β pode assumir os índices 1, 2 ou 3, associados aos genes G_{sel}, G_{clf} e G_{num}, respectivamente. Esse procedimento foi possível aplicando as mutações *flip* sobre os genes que representam listas (G_{sel} e G_{clf}) e *creep* para o gene que indica um número (G_{num}). Na mutação *flip*, uma função é substituída por outra do mesmo tipo. Na mutação do tipo *creep*, um valor é substituído ou somado ao gene (LUCAS, 2002). Considerando esses processos de mutação, os seguintes passos são seguidos:

- caso β = 1 ou 2, a estratégia de "pegue-o-próximo" foi aplicada sobre as listas de métodos de seleção e classificação;

- se β = 3, o gene escolhido é G_{num}. Neste caso, um número aleatório é γ sorteado, considerando γ = 0 ou 1. Se γ = 1, G_{num} é incrementado em uma unidade. Caso contrário (γ = 0), G_{num} é decrementado em uma unidade. O valor máximo para G_{num} é delimitado por $MaxF$ (número máximo de características).

As etapas descritas são repetidas até que número máximo de iterações seja atingida ou caso 99% ou mais dos indivíduos forneça AUC média igual a 1. Um resumo do método proposto está representado na Figura 4.3, com as indicações dos parâmetros descritos previamente e resumidos na Tabela 4.1.

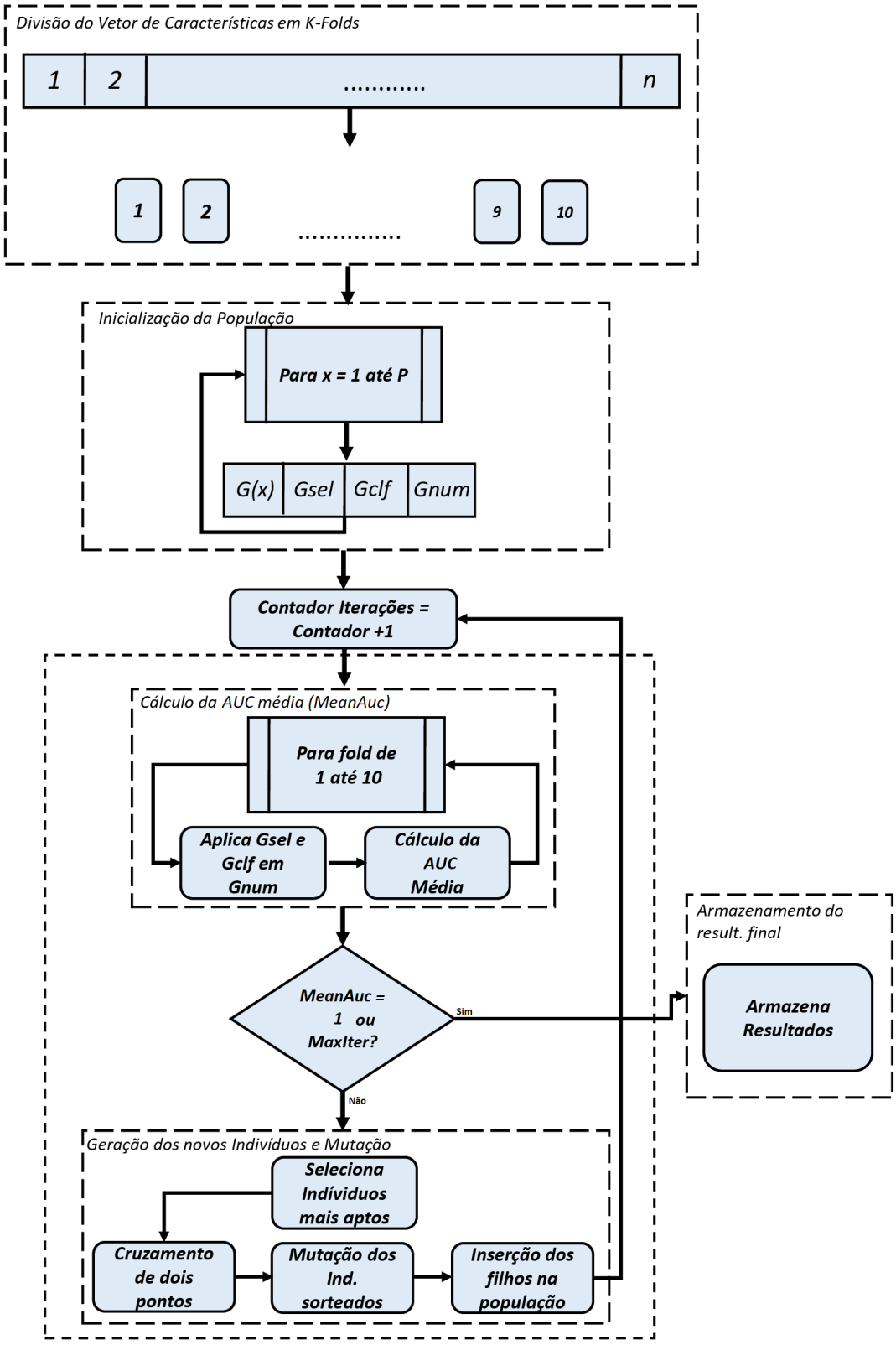


Figura 4.3: Ilustração do método proposto (Fonte: Elaborado pela autora).

4.4 Especificações técnicas para implementação do modelo

A escolha das tecnologias utilizadas na concepção do método proposto foi realizada considerando a possibilidade de agregação e integração de diferentes tecnologias. Portanto, o modelo proposto foi escrito em linguagem Python (VAN ROSSUM; DRAKE, 2003), na versão 2.7. Os métodos de seleção e classificação escolhidos, apresentados na subseção 4.1, possuem origens diversas. Com exceção do algoritmo de seleção *TStatistics*, o qual foi implementado baseando-se na formulação apresentada por ELTOUKHY, FAYE e SAMIR ((2012)), os demais algoritmos foram obtidos em pacotes disponibilizados gratuitamente, tais como WEKA (versão 3.8.1) (HALL et al., 2009) e *Scikit-Learn* (versão 0,18.1) (PEDREGOSA et al., 2011). Esses pacotes são coleções de algoritmos de aprendizado de máquina que podem ser utilizados para mineração e visualização de dados, sob as licenças GNU (*General Public License*) (LICENSE, 1989) e *BSD License*, respectivamente. O primeiro, o pacote WEKA, fornece implementações escritas em linguagem Java. O segundo, o pacote *Scikit-Learn*, possui códigos escritos na linguagem Python.

Os testes foram realizados em três ambientes. O primeiro foi o GridUNESP (IOPE; LEMKE; WINCKLER, 2010), uma infraestrutura localizada na Barra Funda (São Paulo) no Núcleo de Computação Científica da Unesp (NCC/Unesp), composta por 3.104 núcleos de processamento capazes de atingir 77 *TeraFlops*. Devido à alta usabilidade do GridUNESP, as instabilidades são comuns, o que demandou o uso de duas outras máquinas: uma com processador Intel(R) Core (TM) i5-6200U CPU 2.30GHz, 8 *gigabytes* de memória RAM e Windows 1 de 64 bits e uma máquina virtual com 12 *gigabytes* de RAM, processador AMD-V e sistema operacional Ubuntu 17.04, também de 64 bits.

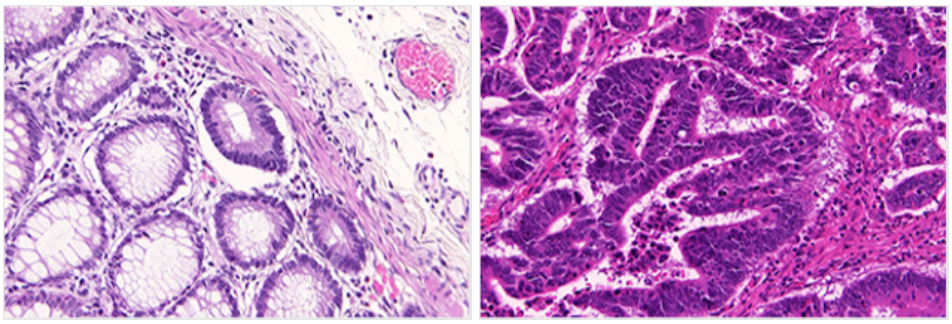
4.5 Aplicação: Casos de Estudo

Os testes foram realizados a partir de características extraídas de dois conjuntos de dados de imagens histológicas: colorretal e linfomas. As características foram definidas aplicando a abordagem descrita por Ribeiro et al. ((2019)). As imagens são derivadas de 16 seções

coradas em H&E, com indicações de estágios T3 ou T4 de câncer colorretal. Cada seção pertencia a um paciente, das quais foram extraídas áreas com diferentes arquiteturas histológicas. A digitalização das seções histológicas em *whole-slide images* (WSI) foi realizada com um escâner deslizante Zeiss MIRAX MIDI com um resolução de *pixel* $0,465\mu m$. As WSIs foram reescalonadas para uma resolução de pixel de $0,620\mu m$, sendo equivalente a uma magnificação de 20x. As imagens foram rotuladas por um especialista na área de patologia, categorizadas nos grupos benigno ou maligno (SIRINUKUNWATTANA et al., 2017). A base é composta por 151 imagens com dimensões de 775 x 522 *pixels*, divididas em 67 casos benignos e 84 casos malignos. Na Figura 4.4, são apresentados exemplos de imagens dos grupos benigno e maligno, respectivamente.

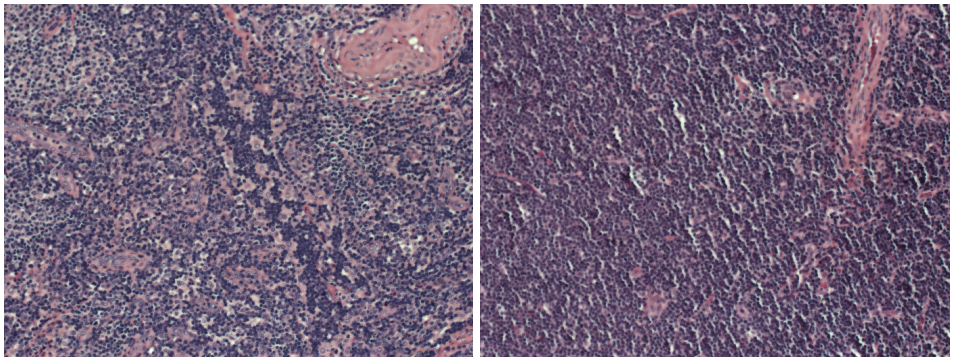
Outro conjunto de dados utilizado para testar o modelo é de imagens de linfomas não-Hodgkin (SHAMIR et al., 2008). O banco de imagens de linfomas utilizado neste trabalho é público, proveniente de estudos realizados por pesquisadores do *National Cancer Institute* (INSTITUTE, 2020) e do *National Institute on Aging* (AGING, 2020), ambos localizados nos Estados Unidos. Esse banco foi constituído a partir de 30 lâminas histológicas de linfonodos, contendo dez casos de três tipos de linfomas: *Follicular Lymphoma* (FL), *Chronic Lymphocytic Leukemia* (CLL) e *Mantle Cell Lymphoma* (MCL). A partir das características de cada grupo, para representar melhor a prática clínica, em vez do ambiente rigidamente controlado de laboratórios, as lâminas foram obtidas com significativas variações de corte e coloração. As imagens microscópicas foram adquiridas digitalmente, por meio de um microscópio de luz (Zeiss Axioscope) com objetiva de 20x e câmera digital colorida (AXio Cam MR5) acoplada. Regiões de interesse de cada lâmina foram selecionadas por especialistas, fotografadas digitalmente e gravadas sem compressão, padrão de cor RGB, resolução espacial de 1388x1040 *pixels*, com taxa de quantização de 24 bits. No total, foram geradas 375 imagens, sendo 122 de regiões com MCL, 140 de regiões com FL e 113 com regiões de CLL. Na Figura 4.5, são apresentados exemplos de imagens dos grupos MCL, FL e CLL, respectivamente.

Figura 4.4: Exemplos de imagens histológicas colorretais. Em (a), exemplo do grupo benigno e em (b), exemplo do grupo maligno. (Fonte: Sirinukunwattana et al., (2017)).

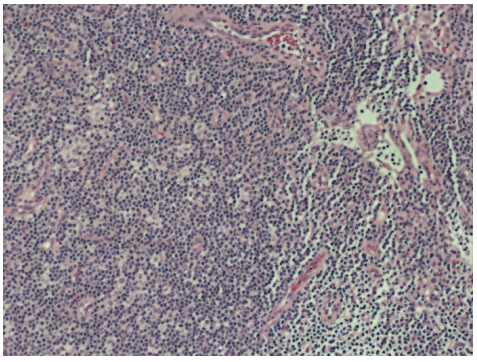


(a) (b)

Figura 4.5: Exemplos de imagens histológicas dos grupos de linfomas. Em (a), exemplo do grupo MCL. Em (b), do grupo FL e por fim, em (c), exemplo do grupo CLL. (Fontes: Aging, (2020); Institute, (2020)).



(a) (b)



(c)

4.5.1 Composição do Vetor de Características

As características foram obtidas a partir de diferentes combinações, como descritores fractais baseados em dimensão de probabilidade (DF_p) e abordagem *box merging* (DF_f), bem como percolação ($Perc$), lacunaridade (Lac) e descritores de Haralick (Har). Cada conjunto de características foi definido através do modelo descrito por (RIBEIRO, Matheus Gonçalves et al., 2019).

As abordagens de dimensão fractal, percolação e lacunaridade permitem quantificações baseadas em propriedades multiescala e multidimensionais (**cualiman2012psoriasis**; IVANOVICI; RICHARD; DECEAN, 2009; ROBERTO et al., 2017). O cálculo de DF_p foi baseado no método probabilístico apresentado por (IVANOVICI; RICHARD; DECEAN, 2009), que considera uma caixa varrendo todos os *pixels* de uma imagem RGB dada como entrada. Esse processo cria uma matriz de probabilidade $P(m, L)$, que representa a probabilidade de m pontos pertencerem a uma caixa L. O descritor DF_f foi baseado na abordagem de combinação de caixas (NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011). Essa característica é calculada a partir da tabela de partição composta por cinco dimensões, a quantidade de linhas não repetidas presentes na tabela de partição e o tamanho da partição observada (NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011).

O cálculo de $Perc$ foi baseado na existência de um caminho de *pixels* conectados (*cluster*) de uma extremidade à outra da imagem. A abordagem probabilística descrita por (IVANOVICI; RICHARD; DECEAN, 2009) também foi aplicada para obter a percolação multiescala e multidimensional (ROBERTO et al., 2017). Os resultados associados a percolação foram representados por três funções: *cluster average* C, *percolating box ratio* Q e *average coverage ratio of the largest cluster* Γ. O comportamento das funções C, Q e Γ foi analisado considerando cinco métricas diferentes: área sob a curva (ARC), *skewness* (SKW), proporção da área (AR), ponto máximo (MP) e escala do ponto máximo (SMP) (ROBERTO et al., 2017). A lacunaridade foi definida a partir do modelo descrito por (IVANOVICI; RICHARD; DECEAN, 2009), que utiliza os momentos de primeira e segunda ordem da matriz de uma matriz de probabilidade para obtenção da característica. A curva de lacu-

naridade também foi analisada pelas mesmas métricas aplicadas à percolação: ARC, SKW, AR, MP e SMP (ROBERTO et al., 2017; RIBEIRO, Matheus Gonçalves et al., 2019).

As medidas Haralick (Har) foram obtidas calculando a matriz de co-ocorrência no nível de cinza de cada imagem e aplicando as métricas de texturas estatísticas descritas em (HARALICK, 1979). As matrizes de co-ocorrência foram calculadas com os valores de $d = 1$, $\theta = 0^c irc$, $\theta = 45°$, $\theta = 90°$ e $\theta = 135°$. Assim, consideramos os valores médios das matrizes de co-ocorrência para definir as características de Haralick (**samanta2014haralick**; **chakraborty2016new**). As medidas foram: (1) segundo momento angular, (2) contraste, (3) correlação, (4) soma dos quadrados (variância), (5) diferença de momento inverso, (6) média da soma, (7) soma da variância, (8) soma da entropia, (9) entropia, (10) diferença de variância, (11) diferença de entropia, (12) e (13) referem-se a primeira e a segunda medidas de correlação de informação, (14) máximo coeficiente de correlação (RIBEIRO, Matheus Gonçalves et al., 2019).

A estrutura do vetor considerou: duas características de Dimensão Fractal, DF_p (IVANOVICI; RICHARD; DECEAN, 2009) e DF_f (NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011); cinco características de Lacunaridade (Lac), 14 características de Haralick (Har) (HARALICK, 1979), representadas por Har(1) a Har(14) e 15 características de percolação, de $Perc$(1) a $Perc$(15) (ROBERTO et al., 2017). Assim, cada imagem foi quantificada por 36 características distintas, conforme ilustrado na Tabela 4.2. Essas características também foram calculadas para sub-imagens curvelet (CANDÈS; DONOHO, 2004). As sub-imagens curvelet foram calculadas aplicando quatro níveis de resolução e uma sequência de oito ângulos de rotação. Essa abordagem resultou em 41 sub-imagens curvelet (1 + 8 + 16 + 16) para cada imagem histológica fornecida como entrada (RIBEIRO, Matheus Gonçalves et al., 2019). Portanto, cada conjunto de características foi definido com 1.512 valores (36 + 1.476).

Tabela 4.2: Ilustração do vetor de características (Fonte: Adaptado de Ribeiro et al., (2019)).

DF_p	DF_f	Lac(1)	...	Lac(5)	Har (1)	...	Har (14)	Perc(1)	...	Perc(15)

Capítulo 5

Resultados e Discussão

Neste capítulo, são discutidos os resultados obtidos a partir dos testes realizados com os conjuntos de características extraídos das bases de linfomas NHL e câncer colorretal, considerando variações nos valores dos parâmetros requeridos pelo método. Os detalhes estão na subseção 5.1. Também é apresentada uma visão geral sobre o desempenho de classificação do método, com comparações em relação aos resultados fornecidos por importantes métodos disponíveis para classificar e reconhecer padrões em imagens H&E de linfomas e câncer colorretal, bem como uma discussão das características que forneceram os melhores resultados. Os destalhes estão na subseção 5.2.

5.1 Parâmetros e Variações

As características foram extraídas das imagens H&E de linfomas e câncer colorretal, conforme descrição apresentada na subseção 4.3 da Metodologia. Em seguida, os vetores de características foram fornecidos como entradas para o método e os testes foram realizados com diferentes valores de P (número de indivíduos), $Iter$ (iterações ou gerações) e taxa de mutação (m). Os valores utilizados nos testes estão na Tabela 5.1. Esses valores foram definidos considerando os estudos apresentados por Bruderer e Singh ((1996)), bem como Al-Rajab, Lu e Xu e ((2017)), com o objetivo de verificar o comportamento do método proposto em diferentes situações, indicar as melhores combinações em cada conjunto de dados

e identificar possíveis padrões no contexto das imagens H&E de linfomas NHL e câncer colorretal. Portanto, os resultados representam soluções aceitáveis envolvendo a melhor associação entre métodos de seleção, classificadores, número de características ($NumF$) e maior área sob a curva ROC (AUC) média ($AverageAUC$).

Tabela 5.1: Parâmetros e valores explorados no método proposto.

Parâmetros	Valores Testados
P	5, 25, 50 e 500
$Iter$	5, 25, 50, 100 e 500
m	0,3, 0,5 e 0,7

Os resultados foram divididos e apresentados a partir das melhores taxas de AUC para classificar as classes envolvidas nas imagens H&E de linfomas NHL e câncer colorretal. Em ambos os casos, as discussões foram pautadas nos resultados obtidos com uma taxa de mutação m = 0,7, valor que forneceu os melhores desempenhos de classificações. Os resultados obtidos com as taxas m = 0,3 e m = 0,5 estão disponíveis no apêndice 6.2. Uma taxa de mutação m = 0,7 pode ter sido a mais bem sucedida pelo fato de que um número maior de indivíduos é submetido ao processo de mutação, o que aumenta a variabilidade genética da população e, consequentemente, das combinações analisadas. Nas Tabelas 5.2, 5.3, 5.4 e 5.5, são exibidos os resultados obtidos para classificar o câncer colorretal, considerando m = 0,7 e variações nas populações P e interações $Iter$. As melhores combinações estão destacadas em negrito.

Tabela 5.2: Imagens colorretais: resultados obtidos com P = 5 e iterações definidas como 5, 25, 50, 100 e 500, sendo que a melhor associação está em negrito.

Iterações	Método de Seleção	Classificadores	*NumF*	AUC
5	*Chi Squared*	*Random Forest*	650	0,848
25	*Chi Squared*	*Random Forest*	713	0,814
50	***Gain Ratio***	***Random Forest***	**401**	**0,896**
100	*Gain Ratio*	*Random Forest*	469	0,869
500	*Gain Ratio*	*J48*	904	0,862

Tabela 5.3: Imagens colorretais: resultados obtidos com P = 25 e iterações definidas como 5, 25, 50, 100 e 500, sendo que a melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*	*Random Forest*	923	0,838
25	*Relief*	*Random Forest*	904	0,837
50	***Relief***	***J48***	**345**	**0,905**
100	*Info Gain*	*Random Forest*	467	0,864
500	*Gain Ratio*	*J48*	876	0,880

Tabela 5.4: Imagens colorretais: resultados obtidos com P = 50 e iterações definidas como 5, 25, 50, 100 e 500, sendo que a melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*KStar*	480	0,915
25	*Relief*	*KStar*	261	0,901
50	***Relief***	***Random Forest***	**24**	**0,984**
100	*Relief*	*KStar*	148	0,940
500	*Relief*	*J48*	1047	0,900

Tabela 5.5: Imagens colorretais: resultados obtidos com P = 500 e iterações definidas como 5, 25, 50, 100 e 500, sendo que a melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*KStar*	56	0,902
25	*Relief*	*KStar*	32	0,854
50	*Relief*	*KStar*	57	0,899
100	*Relief*	*Star*	356	0,911
500	***Relief***	***Random Forest***	**499**	**0,918**

Os resultados permitem constatar que a melhor taxa de AUC foi obtida com uma população P = 50 e $Iter$ = 50. Essa combinação foi capaz de fornecer uma AUC média de 0,984 e com o menor número de características; somente 24. Um número reduzido de características pode tornar a solução mais compreensível para especialistas, visto que as características mais relevantes foram identificadas. Esse resultado foi definido por meio da associação do método de seleção *Relief* com o classificador *Random Forest*.

Para as imagens H&E de linfomas, os resultados obtidos são apresentados nas Tabelas 5.6, 5.7, 5.8 e 5.9, em que as melhores associações também foram destacadas em negrito. É possível verificar que a melhor combinação foi por meio de uma população P = 500 e $Iter$ = 50. Essa configuração permite destacar que a associação entre o método de seleção *Relief* e o classificador *K-Nearest Neighbors* utilizou apenas 73 características para distinguir as três classes de linfomas NHL. Esse número reduzido de características representa somente 4,82% do conjunto original, composto por 1.512 valores. Além disso, essa configuração representa o menor número de iterações necessárias para indicar uma solução aceitável. O resultado foi uma taxa média de AUC = 0,947.

Tabela 5.6: Imagens de linfomas NHL: resultados obtidos com P = 5 e $Iter$ definido como 5, 25, 50, 100 e 500, A melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Chi Squared*	*Random Forest*	534	0,897
25	*Gain Ratio*	*K-Nearest Neighbors*	389	0,882
50	*Gain Ratio*	*K-Nearest Neighbors*	567	0,901
100	*Gain Ratio*	*Random Forest*	664	0,913
500	***Gain Ratio***	***K-Nearest Neighbors***	**923**	**0,918**

Tabela 5.7: Imagens de linfomas NHL: resultados obtidos com P = 25 e $Iter$ definido como 5, 25, 50, 100 e 500, A melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*K-Nearest Neighbors*	841	0,820
25	*Relief*	*K-Nearest Neighbors*	596	0,818
50	***Relief***	***K-Nearest Neighbors***	**503**	**0,896**
100	*Relief*	*Random Forest*	678	0,840
500	*Gain Ratio*	*Random Forest*	751	0,862

Tabela 5.8: Imagens de linfomas NHL: resultados obtidos com P = 50 e $Iter$ definido como 5, 25, 50, 100 e 500, A melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*Random Forest*	512	0,887
25	*Relief*	*K-Nearest Neighbors*	456	0,900
50	*Relief*	*K-Nearest Neighbors*	221	0,913
100	*Relief*	*K-Nearest Neighbors*	245	0,912
500	***Relief***	***Random Forest***	**105**	**0,935**

Tabela 5.9: Imagens de linfomas NHL: resultados obtidos com P = 500 e $Iter$ definido como 5, 25, 50, 100 e 500, A melhor associação está em negrito.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*J48*	976	0,812
25	*Relief*	*J48*	367	0,901
50	***Relief***	***K-Nearest Neighbors***	**73**	**0,947**
100	*Relief*	*K-Nearest Neighbors*	141	0,947
500	*Relief*	*K-Nearest Neighbors*	199	0,939

5.2 Visão Geral do Desempenho de Classificação

Pesquisas relevantes com diferentes abordagens foram desenvolvidas para classificar linfomas NHL e o câncer colorretal. Alguns exemplos são os métodos descritos por Kalkan et al. ((2012)), Rathore et al. ((2013)), Naiyar, Asim e Shahid ((2015)), Kather et al. ((2016)), Jørgensen et al. ((2017)) e Ribeiro et al. ((2019)) para o câncer colorretal, bem como as técnicas apresentadas por Song ((2017)), Roberto et al. ((2017)), Ribeiro et al. ((2018)) e Bai et al. ((2019)) para classes de linfomas (CLLxFLxMCL). Assim, uma visão geral entre os melhores resultados fornecidos pelo método proposto e os trabalhos correlatos está na Tabela 5.10 para o câncer colorretal e na Tabela 5.11 para os linfomas NHL.

Tabela 5.10: Câncer colorretal: taxas de AUC fornecidas por trabalhos correlatos e pela melhor associação do método proposto.

Método	Classificadores	Número de Características	AUC
KALKAN et al. ((2012))	*Logistic Regression*	120	0,950
RATHORE et al. ((2013))	*Rotation Boost*	26	0,970
NAIYAR, ASIM e SHAHID,. ((2015))	SVM	56	0,960
KATHER et al. ((2016))	KNN, SVM e J48	74	0,976
JORGENSEN et al. ((2017))	RaF	9	0,960
RIBEIRO et al. ((2019))	*Polynomial Classifier*	43	0,994
Método proposto	*Random Forest*	24	0,984

Tabela 5.11: Linfomas NHL: taxas de AUC de trabalhos correlatos e da melhor combinação fornecida pelo método proposto.

Método	Classificadores	Número de Características	AUC
SONG et al. ((2017))	SVM	374	0,993
ROBERTO et al. ((2017))	DECORATE	6	0,943
RIBEIRO et al. ((2018))	*Random Forest*	49	0,963
BAI et al. ((2019))	*Random Forest*	-	0,998
Método proposto	*K-Nearest Neighbors*	73	0,947

Os valores apresentados nas Tabelas 5.10 e 5.11 indicam que a proposta é interessante e capaz de fornecer resultados compatíveis com os observados na literatura, principalmente em relação aos métodos descritos por Kalkan et al. ((2012)), Rathore et al. ((2013)), Naiyar, Asim e Shahid ((2015)), Kather et al. ((2016)) e Jørgensen et al. ((2017)) (para o conjunto de dados colorretal) e a abordagem apresentada por Roberto et al. ((2017)) (para o conjunto de dados de linfomas NHL). Embora as condições de testes e os conjuntos de dados de imagens não sejam os mesmos, o que limita esse tipo de comparação, essa visão geral pode mostrar

como o método é capaz de apresentar resultados importantes no estudo do câncer colorretal e dos linfomas NHL, com um número reduzido de características.

Para obter uma comparação mais apropriada, algumas técnicas disponíveis na literatura foram testadas nos mesmos conjuntos de imagens, aplicando modelos baseados em DL, ou com o mesmo conjunto de características, utilizando a abordagem descrita por Kotthoff et al. ((2017)). Os modelos considerados nesses testes foram as redes neurais *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e *ResNet-50* (HE et al., 2016), todos explorando os mesmos conjuntos de imagens de linfomas NHL e de câncer colorretal. O método de otimização automático de hiper-parâmetros (KOTTHOFF et al., 2017) foi aplicado sobre os mesmos conjuntos de características. Os detalhes das imagens e das características foram apresentados na subseção 4.3 da Metodologia.

O método *Deep Learning 4J* (GIBSON; NICHOLSON; PATTERSON, 2016) é baseado em uma arquitetura com uma única camada de saída, que aplica *Stochastic Gradient Descent* como algoritmo de aprendizado e *softmax* como função de ativação. A *AlexNet* é uma conhecida rede neural convolucional (CNN), projetada por Alex Krizhevsky. Essa rede é composta por oito camadas, cinco convoluções e três *fully connected*, e sessenta milhões de parâmetros (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A *ResNet-50* é uma rede que utiliza um método de conexão residual a partir de blocos residuais, sendo cada bloco composto de várias camadas convolucionais empilhadas. A *ResNet-50* tem menor perda de convergência, sem *overfitting*, e é cerca de 20 vezes mais profunda que a *AlexNet* (HE et al., 2016; GU et al., 2018). Por fim, o método de otimização automática de hiperparâmetros *Auto-WEKA 2.0* (KOTTHOFF et al., 2017), é capaz de selecionar características e padrões de classificação em um conjunto significativo de classificadores. Esse método utiliza otimização Bayesiana para testar as melhores combinações e indicar uma solução aceitável. Os testes com esse método foram realizados considerando intervalos de tempos de execução, respeitando o tempo máximo disponível no modelo, de 15 minutos (HALL et al., 2009). É importante ressaltar que as redes neurais, assim como o método descrito por Kotthoff et al. ((2017)), utilizam a acurácia como métrica de avaliação de desempenho. Portanto, essa métrica foi a

utilizada para realizar as próximas comparações.

Nas Tabelas 5.12 e 5.13, são mostrados os valores de acurácias fornecidos pelos modelos baseados em DL para o câncer colorretal e para os linfomas NHL, respectivamente. Os testes foram realizados com *epochs* variando de 1 a 100, bem como considerando porcentagem de divisão do conjunto de imagens em 80% para treinamento e 20% para teste. Na Tabela 5.14, são apresentados os resultados obtidos pelo método de otimização automática de hiperparâmetros (KOTTHOFF et al., 2017) para o câncer colorretal e para os linfomas NHL, considerando uma variação de tempo de execução de 1 a 15 minutos.

Tabela 5.12: Câncer Colorretal: Taxas de acurácias obtidas pelos métodos *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e *ResNet-50* (HE et al., 2016).

Epochs	*Deep Learning 4j*	*AlexNet*	*ResNet-50*
1	64,28%	76,67%	86,67%
5	67,85%	93,33%	96,67%
10	71,42%	70,00%	100,00%
25	64,28%	100,00%	93,33%
50	64,28%	96,67%	93,33%
75	64,28%	93,33%	100,00%
100	67,85%	93,33%	100,00%

Tabela 5.13: NHL: Taxas de acurácias obtidas pelos métodos *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e *ResNet-50* (HE et al., 2016).

Epochs	*Deep Learning 4j*	*AlexNet*	*ResNet-50*
1	70,00%	36,23%	60,87%
5	80,00%	62,32%	69,57%
10	85,33%	57,97%	73,91%
25	76,77%	69,57%	76,81%
50	80,00%	82,61%	79,71%
75	80,00%	73,91%	82,61%
100	80,00%	73,91%	66,67%

Tabela 5.14: Taxas de acurácias obtidas pelo método *Auto-WEKA 2.0* (KOTTHOFF et al., 2017), considerando os mesmos conjuntos de características do método proposto.

Tempo de Execução (em minutos)	Câncer Colorretal	Linfomas NHL
1	64.28%	81,11%
2	64.28%	83,45%
5	75,00%	87,32%
7	75,00%	83,45%
10	89,79%	87,61%
13	64,28%	75,00%
15	90,78%	84,76%

É importante mencionar que, para o conjunto de dados colorretais, a melhor combinação fornecida pelo método proposto indicou uma acurácia de 96,04%, que corresponde a uma AUC média de 0,984. Esse valor de acurácia é melhor do que os valores fornecidos pelos métodos *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016) e *Auto-WEKA 2.0* (KOTTHOFF et al., 2017), conforme as Tabelas 5.12 e 5.14. Por outro lado, o método

não conseguiu superar os resultados obtidos via *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e *ResNet-50* (HE et al., 2016), especialmente para os testes com dez e 25 *epochs*, respectivamente. Nestes casos, a vantagem do método proposto está em apresentar uma solução aceitável com a identificação das características mais relevantes, algo que não é possível via métodos baseados em DL. Para o conjunto de dados de linfomas NHL (Tabelas 5.13 e 5.14), a melhor associação apresentada pelo método proposto resultou em uma acurácia de 92,25%, que corresponde a uma AUC de 0,947. Esse valor de acurácia é superior aos fornecidos pelas abordagens *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), *ResNet-50* (HE et al., 2016) e *Auto-WEKA 2.0* (KOTTHOFF et al., 2017) em todas as variações testadas.

Apesar dos pontos destacados previamente, o teste estatístico de *Wilcoxon-Mann-Whitney* (WILCOXON, 1992) foi aplicado para realizar uma comparação mais apurada entre os métodos. Esse tipo de teste é utilizado quando se tem duas populações independentes e busca-se identificar se uma população tende a ter valores maiores do que a outra ou se elas têm a mesma mediana (FAY; PROSCHAN, 2010). Portanto, o teste é capaz de indicar se a diferença entre duas populações é estatisticamente relevante. Assim, a falta de significância estatística indica que os desempenhos são similares. O teste exige que cada população seja constituída de pelo menos cinco amostras (resultados). Os valores utilizados para essas comparações estão listados na Tabelas 5.12, 5.13 e 5.14, totalizando sete testes para cada método. Em relação ao método proposto, nas Tabelas A.17 e A.18, são apresentadas as informações dos valores de acurácias que foram considerados nas comparações estatísticas. Os valores selecionados foram os sete melhores desempenhos obtidos a partir dos resultados indicados na subseção 5.1 dos Resultados. O nível de significância aplicado nos testes estatísticos foi de 0,5.

Nas Tabelas 5.15 e 5.16, estão indicados se os resultados do método proposto *versus* os fornecidos pelas abordagens *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), *ResNet-50* (HE et al., 2016) e *Auto-WEKA 2.0* (KOTTHOFF et al., 2017) são significantes ou não, com as notações

dos valores de p. Considerando os desempenhos para investigar o câncer colorretal (Tabela 5.15), é possível destacar que apesar dos métodos *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e *ResNet-50* (HE et al., 2016) fornecerem valores máximos de acurácias (100,00%) em algumas situações, as diferenças em relação ao método proposto não são estatisticamente relevantes. As diferenças estatisticamente relevantes estão em relação aos resultados obtidos via *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016) e *Auto Weka 2.0* (KOTTHOFF et al., 2017). Em ambos os casos, o desempenho do método proposto foi melhor. Em relação aos resultados para as imagens de linfomas NHL (Tabela 5.16), as diferenças são estatisticamente relevantes em todas as comparações. Nesses casos, o desempenho do método foi superior aos modelos comparados. Outros testes podem ser realizados para identificar se essa superioridade é mantida em outras condições ou até mesmo fazer ajustes nos métodos utilizados para determinar os limites de cada um. No entanto, acredita-se que os testes realizados fornecem uma visão consistente do método proposto. É possível notar que a proposta contribui com a literatura especializada, pois as associações fornecidas foram capazes de atingir resultados, no mínimo, equivalentes aos produzidos por métodos importantes e amplamente explorados na área de aprendizado de máquina, sem realizar ajustes de hiperparâmetros dos classificadores, estratégia comumente utilizada nos métodos apresentados previamente. Mais ainda, é possível indicar que o método proposto pode ser útil como uma etapa de pré-processamento para definir as melhores combinações de técnicas nos cenários investigados e, consequentemente, facilitar a tarefa de ajuste e otimização de hiperparâmetros dos classificadores.

Tabela 5.15: Resultados do teste estatístico de *Wilcoxon-Mann-Whitney* para os desempenhos envolvendo as imagens colorretais: método proposto versus *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), *ResNet-50* (HE et al., 2016) e *Auto-WEKA 2.0* (KOTTHOFF et al., 2017).

Métodos	Significante?	Valor de p
Deep Leaning 4j	Sim	0,00214
AlexNet	Não	0,30772
ResNet-50	Não	0,05486
Auto Weka 2.0	Sim	0,04136

Tabela 5.16: Resultados do teste estatístico de *Wilcoxon-Mann-Whitney* para os desempenhos envolvendo as imagens de linfomas NHL: método proposto versus *Deep Learning 4j* (GIBSON; NICHOLSON; PATTERSON, 2016), *AlexNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), *ResNet-50* (HE et al., 2016) e *Auto-WEKA 2.0* (KOTTHOFF et al., 2017).

Métodos	Significante?	Valor de p
Deep Leaning 4j	Sim	0,00604
AlexNet	Sim	0,00604
ResNet-50	Sim	0,00604
Auto Weka 2.0	Sim	0,00107

Por fim, outra vantagem da proposta é a identificação das características mais relevantes e suas associações. Considerando que o processo de geração da população é aleatório, princípio fundamental do método para fornecer soluções aceitáveis, não se pode afirmar que os resultados obtidos serão sempre associados às mesmas características. Porém, acredita-se que, pela variação considerável do número de indivíduos, existe uma tendência de convergência para bons resultados com as mesmas características. Dessa forma, uma visão geral dessas distribuições para classificar o câncer colorretal e os linfomas NHL estão nas Tabelas 5.17 e 5.18, respectivamente. As distribuições representam as características obtidas a partir dos melhores resultados.

Tabela 5.17: Discriminação das características selecionadas no melhor resultado para distinguir o conjunto colorretal.

DF_p		DF_f		Lac		Har		Perc		Total
Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	
0	1	0	2	0	7	0	11	0	3	24

Para obter o melhor resultado no conjunto de dados colorretais, é possível observar que as características selecionadas foram de associações com sub-imagens *curvelets*. A maioria das características foi selecionada a partir do descritor de Haralick, totalizando onze características. O atributo lacunaridade foi o segundo mais selecionado, com sete características. Além disso, os descritores de percolação contribuíram com três valores e os descritores DF_p e DF_f contribuíram com um e dois valores, respectivamente.

Tabela 5.18: Discriminação das características selecionadas no melhor resultado para distinguir os linfomas NHL.

DF_p		DF_f		Lac		Har		Perc		Total
Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	Imagem H&E	Sub-image	
0	0	0	0	0	12	1	23	1	36	73

Considerando o conjunto de dados de linfomas NHL, é possível observar que a maioria das características foi selecionada a partir da associação do descritor percolação com sub-imagens curvelet: 36 características de um total de 37. Por outro lado, as medidas de Haralick foram o segundo tipo mais selecionado, totalizando 24. Nesse caso, a principal associação também foi Haralick com sub-imagens curvelet. O descritor de lacunaridade contribuiu com 12 valores, todos calculados de sub-imagens curvelet. Os descritores de dimensão fractal multiescala e multidimensionais (DF_p e DF_f) não foram selecionados. Esse fato pode indicar que esses atributos e suas combinações com sub-imagens *curvelet* não são relevantes para o reconhecimento de padrões nas imagens H&E de linfomas NHL. Isso pode ter ocorrido porque os valores para esses atributos são similares, mesmo para classes distintas

das imagens sob investigação.

Nesse contexto, acredita-se que esses resultados podem contribuir significativamente para o desenvolvimento de sistemas *Computer-Aided Diagnosis* (CAD), principalmente com testes representando diferentes condições para classificar e reconhecer padrões em imagens H&E de câncer colorretal e linfomas NHL.

Capítulo 6

Conclusão

Neste trabalho, foi desenvolvido um método baseado em Algoritmo Genético (AG) para encontrar a melhor combinação entre características, métodos de seleção e de classificação. A metodologia foi desenvolvida a partir do principio de evolução apresentado por Charles Darwin, agregando um conjunto de indivíduos (chamado de população) que foi submetido a três etapas: definição da estrutura cromossômica e população inicial, avaliação e seleção da população e, por fim *crossover* e mutação. O método foi testado em imagens histológicas de câncer colorretal e linfomas NHL, considerando diferentes combinações de parâmetros. Foram analisadas variações da taxa de mutação (m), população P e número máximo de iterações $Iter$, permitindo encontrar soluções aceitáveis e com um número reduzido de características.

Para o conjunto de dados colorretal, a melhor solução foi obtida com 50 indivíduos e 50 iterações, com apenas 24 características, algoritmo de seleção *Relief* e classificador *Random Forest*. A área sob a curva ROC (AUC) foi de 0,984 e a taxa de mutação associada foi m = 0,7. Em relação ao conjunto de dados de linfomas, a melhor solução foi determinada com 500 indivíduos e 50 iterações, também com taxa de mutação m = 0,7. Nesse caso, o valor da AUC foi de 0,947. O modelo também conseguiu reduzir o conjunto de dados de 1.512 características para 73, aplicando o algoritmo de seleção *Relief* e o classificador *K-Nearest Neighbors*.

Ao considerar uma visão geral em relação aos estudos disponíveis na literatura, o método

proposto apresenta um importante diferencial: discrimina e detalha possíveis padrões de características indicadas para a classificação dos grupos de câncer sob investigação. O padrão identificado envolveu os descritores de Haralick, lacunaridade e percolação, todos obtidos de sub-imagens *curvelets*. O uso do método estatístico *Wilcoxon-Mann-Whitney* também foi aplicado para comparar os desempenhos do método proposto e os fornecidos por técnicas importantes e amplamente exploradas na área de aprendizado de máquina. Os resultados obtidos foram comparados com os fornecidos por redes DL (GIBSON; NICHOLSON; PATTERSON, 2016; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; HE et al., 2016) e com o método descrito por Kotthoff et al. ((2017)), que utiliza a otimização Bayesiana para indicar as melhores soluções. Para o conjunto de dados de linfomas, o método proposto foi superior em todas comparações e, *Computer-Aided Diagnosis* (CAD) para o conjunto de dados colorretal, superior ao método (GIBSON; NICHOLSON; PATTERSON, 2016). Uma vantagem do método proposto está em indicar uma solução aceitável com a identificação das características mais relevantes para o processo de classificação. Dessa forma, acredita-se que os resultados são relevantes para o estudo e reconhecimento de padrões de linfomas não Hodgkin e câncer colorretal, além de contribuir para o desenvolvimento de sistemas *Computer-Aided Diagnosis* (CAD).

6.1 Contribuições obtidas

As principais contribuições com o desenvolvimento deste trabalho foram:

1. Um modelo que permite fornecer soluções aceitáveis para o reconhecimento de padrões de linfomas NHL e câncer colorretal, mesmo envolvendo inúmeras combinações de parâmetros, características, métodos de seleção e classificadores;

2. Identificação dos parâmetros mais relevantes, como tamanho da população e número de iterações, para classificar os grupos benigno e maligno de imagens colorretais, bem como os grupos MCL, FL e CLL de linfomas NHL;

3. Identificação de quantas e quais foram as características mais relevantes, bem como

suas associações, para investigar os grupos de câncer colorretal e os de linfomas NHL;

4. Indicações de combinações de técnicas que forneceram classificações com taxas melhores do que as obtidas via modelos *DL* e em otimização Bayesiana para definir as melhores soluções.

6.2 Trabalhos futuros

Em trabalhos futuros, pretende-se: 1) explorar o modelo para o reconhecimento de padrões em imagens de H&E de câncer de mama, com ou sem normalização dos corantes presentes nas lâminas. 2) realizar um levantamento das principais características que constituem as melhores soluções; 3) explorar mais variações dos parâmetros utilizados como entrada para o método, tais como t e $MaxF$, e identificar quais deles afetam significativamente os resultados. O parâmetro t permite aumentar o número de indivíduos que serão selecionados e passarão por cruzamento, criando novos indivíduos para avaliação. O parâmetro $MaxF$ possibilita verificar se um número inicial reduzido de características pode causar uma otimização mais rápida no método; 4) alterar a estrutura cromossômica do método para investigar o ajuste de hiperparâmetros dos classificadores. Essa alteração pode fornecer taxas de distinções ainda mais relevantes.

Referências

AB WAHAB, Mohd Nadhir; NEFTI-MEZIANI, Samia; ATYABI, Adham. A comprehensive review of swarm optimization algorithms. **PloS one**, Public Library of Science, v. 10, n. 5, e0122827, 2015.

ABAS, Fazly S et al. Computer-assisted quantification of CD3+ T cells in follicular lymphoma. **Cytometry Part A**, Wiley Online Library, v. 91, n. 6, p. 609–621, 2017.

ABINASH, MJ; VASUDEVAN, V. A Hybrid Forward Selection Based LASSO Technique for Liver Cancer Classification. In: NANOELECTRONICS, Circuits and Communication Systems. [S.l.]: Springer, 2019. p. 185–193.

AFEF, Latrach et al. Comparison study for computer assisted detection and diagnosis CADsystems dedicated to prostate cancer detection using MRImp modalities. In: IEEE. ADVANCED Technologies for Signal and Image Processing (ATSIP), 2018 4th International Conference on. [S.l.: s.n.], 2018. p. 1–6.

AGING, National Institute on. **National Institute on Aging**. [S.l.: s.n.], 2020. ⟨https://www.nia.nih.gov/⟩. (Accessed on 04/17/2020).

AIN, Qurrat Ul et al. Genetic programming for feature selection and feature construction in skin cancer image classification. In: SPRINGER. PACIFIC Rim International Conference on Artificial Intelligence. [S.l.: s.n.], 2018. p. 732–745.

ALDOJ, Nader et al. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. **European radiology**, Springer, v. 30, n. 2, p. 1243–1253, 2020.

ALTERI, R; KRAMER, J; S, Simpson. Colorectal Cancer Facts and Figures 2014-2016. **Atlanta: American Cancer Society**, p. 1–30, 2014.

AMORIM, Maurıcio JV; BARONE, Dante; MANSUR, André Uebe. Técnicas de aprendizado de máquina aplicadas na previsao de evasao acadêmica. In: 1. BRAZILIAN Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). [S.l.: s.n.], 2008. v. 1, p. 666–674.

ANBARASI, M; ANUPRIYA, E; IYENGAR, NCSN. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. **International Journal of Engineering Science and Technology**, v. 2, n. 10, p. 5370–5376, 2010.

ANITHA, V; MURUGAVALLI, S. Brain tumour classification using two-tier classifier with adaptive segmentation technique. **IET computer vision**, IET, v. 10, n. 1, p. 9–17, 2016.

EL-ANWAR, Mohammad Waheed et al. Anterior Ethmoidal Artery: A Computed Tomography Analysis and New Classifications. **Journal of Neurological Surgery Part B: Skull Base**, Georg Thieme Verlag KG, 2020.

APOSTOLOPOULOS, Ioannis D; MPESIANA, Tzani A. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. **Physical and Engineering Sciences in Medicine**, Nature Publishing Group, p. 1, 2020.

ARIMURA, Hidetaka et al. Computer-aided diagnosis systems for brain diseases in magnetic resonance images. **Algorithms**, Molecular Diversity Preservation International, v. 2, n. 3, p. 925–952, 2009.

ARLOT, Sylvain; CELISSE, Alain et al. A survey of cross-validation procedures for model selection. **Statistics surveys**, The author, under a Creative Commons Attribution License, v. 4, p. 40–79, 2010.

ATKESON, Christopher G; MOORE, Andrew W; SCHAAL, Stefan. Locally weighted learning for control. In: LAZY learning. [S.l.]: Springer, 1997. p. 75–113.

BABATUNDE, Oluleye H et al. A genetic algorithm-based feature selection. IJECCE, 2014.

BAI, Jie et al. NHL Pathological Image Classification Based on Hierarchical Local Information and GoogLeNet-Based Representations. **BioMed research international**, Hindawi, v. 2019, 2019.

BEJNORDI, Babak Ehteshami et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. **Jama**, American Medical Association, v. 318, n. 22, p. 2199–2210, 2017.

BHARDWAJ, Harshit et al. Breast Cancer Diagnosis using Simultaneous Feature Selection and Classification: A Genetic Programming Approach. In: IEEE. 2018 IEEE Symposium Series on Computational Intelligence (SSCI). [S.l.: s.n.], 2018. p. 2186–2192.

BRANCATI, Nadia et al. A deep learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images. **IEEE Access**, IEEE, v. 7, p. 44709–44720, 2019.

BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BRUDERER, Erhard; SINGH, Jitendra V. Organizational evolution, learning, and selection: A genetic-algorithm-based model. **Academy of management journal**, Academy of Management Briarcliff Manor, NY 10510, v. 39, n. 5, p. 1322–1349, 1996.

CANDÈS, Emmanuel J; DONOHO, David L. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. **Communications on pure and applied mathematics**, Wiley Online Library, v. 57, n. 2, p. 219–266, 2004.

CHAN, Heang-Ping et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms. **Arbor**, v. 1001, p. 48109–0326, 1990.

CHANDRASHEKAR, Girish; SAHIN, Ferat. A survey on feature selection methods. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 1, p. 16–28, 2014.

CHENG, Jie-Zhi et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. **Scientific reports**, Nature Publishing Group, v. 6, p. 24454, 2016.

CHOUAIB, Hassan et al. Feature selection combining genetic algorithm and adaboost classifiers. In: IEEE. PATTERN Recognition, 2008. ICPR 2008. 19th International Conference on. [S.l.: s.n.], 2008. p. 1–4.

CLEARY, John G; TRIGG, Leonard E. K*: An instance-based learner using an entropic distance measure. In: MACHINE Learning Proceedings 1995. [S.l.]: Elsevier, 1995. p. 108–114.

CODELLA, Noel et al. Lymphoma diagnosis in histopathology using a multi-stage visual learning approach. In: INTERNATIONAL SOCIETY FOR OPTICS e PHOTONICS. MEDICAL Imaging 2016: Digital Pathology. [S.l.: s.n.], 2016. v. 9791, 97910h.

CONCI, Aura; AZEVEDO, Eduardo; LETA, Fabiana R. **Computação gráfica**. [S.l.]: Elsevier, 2008.

CRISTIANO, Mariana Vitória de Menezes Bordalo. **Sensibilidade e especificidade na curva roc: um caso de estudo**. 2017. Tese (Doutorado).

DAINESE, Renata Cilene. Sensoriamento remoto e geoprocessamento aplicado ao estudo temporal do uso da terra e na comparação entre classificação não supervisionada e análise visual. Universidade Estadual Paulista (UNESP), 2001.

DENNIS, Binu; MUTHUKRISHNAN, Selvi. AGFS: Adaptive Genetic Fuzzy System for medical data classification. **Applied Soft Computing**, Elsevier, v. 25, p. 242–252, 2014.

DIETTERICH, Tom. Overfitting and undercomputing in machine learning. **ACM computing surveys (CSUR)**, ACM, v. 27, n. 3, p. 326–327, 1995.

DOI, Kunio; HUANG, HK. Computer-aided diagnosis (CAD) and image-guided decision support. **Computerized Medical Imaging and Graphics**, v. 4, n. 31, p. 195–197, 2007.

DOYLE, Scott et al. A boosting cascade for automated detection of prostate cancer from digitized histology. In: SPRINGER. INTERNATIONAL Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.: s.n.], 2006. p. 504–511.

EIBEN, Agoston E; SMITH, James E et al. **Introduction to evolutionary computing**. [S.l.]: Springer, 2003. v. 53.

ELTOUKHY, Mohamed Meselhy; FAYE, Ibrahima; SAMIR, Brahim Belhaouari. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. **Computers in biology and medicine**, Elsevier, v. 42, n. 1, p. 123–128, 2012.

ESTEVA, Andre et al. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, Nature Publishing Group, v. 542, n. 7639, p. 115, 2017.

FARID, Ahmed Abdullah; SELIM, Gamal; KHATER, Hatem. A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast Cancer Detection, 2020.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FAY, Michael P; PROSCHAN, Michael A. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. **Statistics surveys**, NIH Public Access, v. 4, p. 1, 2010.

FUREY, Terrence S et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. **Bioinformatics**, Oxford University Press, v. 16, n. 10, p. 906–914, 2000.

FURTADO, Joao Carlos; LORENA, Luiz Antonio Nogueira. Algoritmo genético construtivo na otimizaçao de problemas combinatoriais de agrupamentos. **São José dos Campos. Tese (Doutorado em Computação Aplicada)–Instituto Nacional de Pesquisas Espaciais**, 1998.

GABRYS, Bogdan; RUTA, Dymitr. Genetic algorithms in classifier fusion. **Applied soft computing**, Elsevier, v. 6, n. 4, p. 337–347, 2006.

GARDNER, Matt W; DORLING, SR. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GIANNINI, Valentina et al. Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. **European radiology**, Springer, v. 27, n. 10, p. 4200–4208, 2017.

GIBSON, Adam; NICHOLSON, Chris; PATTERSON, Josh. **Eclipse Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0.** [S.l.: s.n.], 2016. ⟨http://deeplearning4j.org⟩. Accessed: 2019-12-10.

GNANA, D Asir Antony; APPAVU, S; LEAVLINE, E Jebamalar. Literature review on feature selection methods for high-dimensional data. **methods**, v. 136, n. 1, 2016.

GOKULNATH, Chandra Babu; SHANTHARAJAH, SP. An optimized feature selection based on genetic approach and support vector machine for heart disease. **Cluster Computing**, Springer, v. 22, n. 6, p. 14777–14787, 2019.

GOU, Jianping et al. A generalized mean distance-based k-nearest neighbor classifier. **Expert Systems with Applications**, Elsevier, v. 115, p. 356–372, 2019.

GU, Jiuxiang et al. Recent advances in convolutional neural networks. **Pattern Recognition**, Elsevier, v. 77, p. 354–377, 2018.

GUIMARÃES, M Carolina S. Exames de laboratório: sensibilidade, especificidade, valor preditivo positivo. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 18, n. 2, p. 117–120, 1985.

GURCAN, Metin N et al. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system. **Medical Physics**, Wiley Online Library, v. 29, n. 11, p. 2552–2558, 2002.

GUYON, Isabelle; ELISSEEFF, André. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, Mar, p. 1157–1182, 2003.

HALL, Mark et al. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HANLEY, James A; MCNEIL, Barbara J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HARALICK, Robert M. Statistical and structural approaches to texture. **Proceedings of the IEEE**, IEEE, v. 67, n. 5, p. 786–804, 1979.

HE, Kaiming et al. Deep residual learning for image recognition. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016. p. 770–778.

HERMINE, Olivier et al. Addition of high-dose cytarabine to immunochemotherapy before autologous stem-cell transplantation in patients aged 65 years or younger with mantle cell lymphoma (MCL Younger): a randomised, open-label, phase 3 trial of the European Mantle Cell Lymphoma Network. **The Lancet**, Elsevier, v. 388, n. 10044, p. 565–575, 2016.

HO, WT; LAM, PWT. Clinical performance of computer-assisted detection (CAD) system in detecting carcinoma in breasts of different densities. **Clinical Radiology**, Elsevier, v. 58, n. 2, p. 133–136, 2003.

HOLLAND, John H. Genetic algorithms. **Scientific american**, JSTOR, v. 267, n. 1, p. 66–73, 1992.

HUA, Kai-Lung et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. **OncoTargets and therapy**, Dove Press, v. 8, 2015.

HUANG, Jin; LING, Charles X. Using AUC and accuracy in evaluating learning algorithms. **IEEE Transactions on knowledge and Data Engineering**, IEEE, v. 17, n. 3, p. 299–310, 2005.

HUANG, Peng; PARK, Seyoun et al. Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. **Radiology**, Radiological Society of North America, v. 286, n. 1, p. 286–295, 2017.

INCA. **Estimativa — 2020 Incidência de Câncer no Brasil**. Rio de Janeiro, Brazil, 2019.

INSTITUTE, National Cancer. **Comprehensive Cancer Information - National Cancer Institute**. [S.l.: s.n.], 2020. ⟨https://www.cancer.gov/⟩. (Accessed on 04/17/2020).

INZA, Iñaki et al. Filter versus wrapper gene selection approaches in DNA microarray domains. **Artificial intelligence in medicine**, Elsevier, v. 31, n. 2, p. 91–103, 2004.

IOPE, R; LEMKE, Ney; WINCKLER, G von. GridUNESP: a multi-campus Grid infrastructure for scientific computing. In: PROCEEDINGS of the 3rd Latin American Conference on High Performance Computing (CLCAR 2010); Gramado: 25-28 August. [S.l.: s.n.], 2010. p. 76–84.

IVANOVICI, M; RICHARD, N; DECEAN, H. Fractal dimension and lacunarity of psoriatic lesions-A colour approach. **medicine**, v. 6, n. 4, p. 7, 2009.

JAFFAR, M Arfan; SIDDIQUI, Abdul Basit; MUSHTAQ, Mubashar. Ensemble classification of pulmonary nodules using gradient intensity feature descriptor and differential evolution. **Cluster Computing**, Springer, v. 21, n. 1, p. 393–407, 2018.

JALALIAN, Afsaneh et al. Computer-assisted diagnosis system for breast cancer in computed tomography laser mammography (ctlm). **Journal of digital imaging**, Springer, v. 30, n. 6, p. 796–811, 2017.

JANSSEN, Ronald J; MOURÃO-MIRANDA, Janaina; SCHNACK, Hugo G. Making individual prognoses in psychiatry using neuroimaging and machine learning. **Biological Psychiatry: Cognitive Neuroscience and Neuroimaging**, Elsevier, 2018.

JAYACHANDRAN, Srinath; GHOSH, Ashlin. Deep Transfer Learning for Texture Classification in Colorectal Cancer Histology. **arXiv preprint arXiv:2004.01614**, 2020.

JIANG, Shancheng et al. Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. **Expert Systems with Applications**, Elsevier, v. 82, p. 216–230, 2017.

JOHN, George H; LANGLEY, Pat. Estimating continuous distributions in Bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. PROCEEDINGS of the Eleventh conference on Uncertainty in artificial intelligence. [S.l.: s.n.], 1995. p. 338–345.

JØRGENSEN, Alex Skovsbo et al. Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides. **Cytometry Part A**, Wiley Online Library, v. 91, n. 8, p. 785–793, 2017.

KALKAN, Habil et al. Automated classification of local patches in colon histopathology. In: IEEE. PATTERN Recognition (ICPR), 2012 21st International Conference on. [S.l.: s.n.], 2012. p. 61–64.

KARNAN, M; LOGHESHWARI, T. Improved implementation of brain MRI image segmentation using ant colony system. In: IEEE. 2010 IEEE International Conference on Computational Intelligence and Computing Research. [S.l.: s.n.], 2010. p. 1–4.

KATHER, Jakob Nikolas et al. Multi-class texture analysis in colorectal cancer histology. **Scientific reports**, Nature Publishing Group, v. 6, p. 27988, 2016.

KEČO, Dino; SUBASI, Abdulhamit; KEVRIC, Jasmin. Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. **Neural Computing and Applications**, Springer, v. 30, n. 5, p. 1601–1610, 2018.

KIM, Ji-Hyun. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. **Computational statistics & data analysis**, Elsevier, v. 53, n. 11, p. 3735–3745, 2009.

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA, 2. IJCAI. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

KOIZUMI, Mitsuru et al. Diagnostic performance of a computer-assisted diagnosis system for bone scintigraphy of newly developed skeletal metastasis in prostate cancer patients: search for low-sensitivity subgroups. **Annals of nuclear medicine**, Springer, v. 31, n. 7, p. 521–528, 2017.

KOMINAMI, Yoko et al. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. **Gastrointestinal endoscopy**, Elsevier, v. 83, n. 3, p. 643–649, 2016.

KOOI, Thijs et al. Large scale deep learning for computer aided detection of mammographic lesions. **Medical image analysis**, Elsevier, v. 35, p. 303–312, 2017.

KOTSIANTIS, Sotiris B; ZAHARAKIS, I; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, v. 160, p. 3–24, 2007.

KOTTHOFF, Lars et al. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. **The Journal of Machine Learning Research**, JMLR. org, v. 18, n. 1, p. 826–830, 2017.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: ADVANCES in neural information processing systems. [S.l.: s.n.], 2012. p. 1097–1105.

KUMAR, G Ravi; RAMACHANDRA, GA; NAGAMANI, K. An efficient feature selection system to integrating svm with genetic algorithm for large medical datasets. **International Journal**, v. 4, n. 2, 2014.

KUNCHEVA, Ludmila I; JAIN, Lakhmi C. Designing classifier fusion systems by genetic algorithms. **IEEE Transactions on Evolutionary computation**, IEEE, v. 4, n. 4, p. 327–336, 2000.

KUO, Bor-Chen et al. A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, IEEE, v. 7, n. 1, p. 317–326, 2014.

LEE, Jung Hwan et al. Spine Computed Tomography to Magnetic Resonance Image Synthesis Using Generative Adversarial Networks: A Preliminary Study. **Journal of Korean Neurosurgical Society**, Korean Neurosurgical Society, 2020.

LEITE, Mariana et al. Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. **Journal of Medical Imaging**, International Society for Optics e Photonics, v. 2, n. 1, p. 014002, 2015.

LI, Wenyuan et al. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. **IEEE transactions on medical imaging**, IEEE, v. 38, n. 4, p. 945–954, 2018.

LICENSE, GNU General Public. GNU General Public License. **Retrieved December**, v. 25, p. 2014, 1989.

LINDEN, Ricardo. **Algoritmos genéticos: uma importante ferramenta da inteligência computacional**. [S.l.]: Brasport, 2006.

LIU, Liping et al. Modified Cuckoo Search Algorithm with Variational Parameters and Logistic Map. **Algorithms**, Multidisciplinary Digital Publishing Institute, v. 11, n. 3, p. 30, 2018.

LOPES, Uilian Kenedi. Redes neurais convolucionais aplicadas ao diagnóstico de tuberculose por meio de imagens radiológicas. Universidade do Vale do Rio dos Sinos, 2017.

LU, Chunhong; ZHU, Zhaomin; GU, Xiaofeng. An intelligent system for lung cancer diagnosis using a new genetic algorithm based feature selection method. **Journal of medical systems**, Springer, v. 38, n. 9, p. 97, 2014.

LUCAS, Diogo C. Algoritmos genéticos: uma introdução. **Universidade Federal do Rio Grande do Sul**, v. 24, 2002.

MARTINEZ, Edson Zangiacomi; LOUZADA-NETO, Francisco; PEREIRA, Basılio de Bragança. A curva ROC para testes diagnósticos. **Cad. saúde colet.,(Rio J.)**, v. 11, n. 1, p. 7–31, 2003.

MELENDEZ, Jaime et al. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays. **IEEE Trans. Med. Imaging**, v. 34, n. 1, p. 179–192, 2015.

MIRANDA, Márcio Nunes de. **Algoritmos Genéticos: Fundamentos e Aplicações**. [S.l.: s.n.], 2007.

MISAWA, Masashi et al. Characterization of colorectal lesions using a computer-aided diagnostic system for narrow-band imaging endocytoscopy. **Gastroenterology**, Elsevier, v. 150, n. 7, p. 1531–1532, 2016.

MITCHELL, Tom M et al. Machine learning. 1997. **Burr Ridge, IL: McGraw Hill**, v. 45, n. 37, p. 870–877, 1997.

MOHAMED, Ali W; SABRY, Hegazy Z; KHORSHID, Motaz. An alternative differential evolution algorithm for global optimization. **Journal of advanced research**, Elsevier, v. 3, n. 2, p. 149–165, 2012.

MUNI, Durga Prasad; PAL, Nikhil R; DAS, Jyotirmay. Genetic programming for simultaneous feature selection and classifier design, 2006.

NAGARAJAN, G et al. Hybrid genetic algorithm for medical image feature extraction and selection. **Procedia Computer Science**, Elsevier, v. 85, p. 455–462, 2016.

NAIYAR, Madeeha; ASIM, Yousra; SHAHID, Aqsa. Automated colon cancer detection using structural and morphological features. In: IEEE. FRONTIERS of Information Technology (FIT), 2015 13th International Conference on. [S.l.: s.n.], 2015. p. 240–245.

NIKOLAIDIS, NS; NIKOLAIDIS, IN; TSOUROS, CC. A Variation of the Box-Counting Algorithm Applied to Colour Images. **arXiv preprint arXiv:1107.2336**, 2011.

ORLOV, Nikita V et al. Automatic classification of lymphoma images with transform-based global features. **IEEE Transactions on Information Technology in Biomedicine**, IEEE, v. 14, n. 4, p. 1003–1013, 2010.

ÖZÇIFT, Akın; GÜLTEN, Arif. Genetic algorithm wrapped Bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases. **Digital Signal Processing**, Elsevier, v. 23, n. 1, p. 230–237, 2013.

PARIKH, Jay; TICKMAN, Ronald. Image-Guided Tissue Sampling: Where Radiology Meets Pathology. **The breast journal**, Wiley Online Library, v. 11, n. 6, p. 403–409, 2005.

PATIDAR, Shivnarayan; PACHORI, Ram Bilas; ACHARYA, U Rajendra. Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. **Knowledge-Based Systems**, Elsevier, v. 82, p. 1–10, 2015.

PAUL, Desbordes et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. **Computerized Medical Imaging and Graphics**, Elsevier, v. 60, p. 42–49, 2017.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **Journal of machine learning research**, v. 12, Oct, p. 2825–2830, 2011.

PICOLI, Ivan Luiz. Uma abordagem de classificação não supervisionada de cargas de trabalho de sistemas analıticos em Apache Hadoop através de análise de log, 2013.

PRAVEEN, GB; AGRAWAL, Anita. Hybrid approach for brain tumor detection and classification in magnetic resonance images. In: IEEE. COMMUNICATION, Control and Intelligent Systems (CCIS), 2015. [S.l.: s.n.], 2015. p. 162–166.

AL-RAJAB, Murad; LU, Joan; XU, Qiang. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. **Computer methods and programs in biomedicine**, Elsevier, v. 146, p. 11–24, 2017.

RAJPURKAR, Pranav et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. **arXiv preprint arXiv:1711.05225**, 2017.

RATHORE, Saima et al. Classification of colon biopsy images based on novel structural features. In: IEEE. 2013 IEEE 9th International Conference on Emerging Technologies (ICET). [S.l.: s.n.], 2013. p. 1–6.

EL-REGAILY, Salsabil A et al. Survey of Computer Aided Detection Systems for Lung Cancer in Computed Tomography. **Current Medical Imaging Reviews**, Bentham Science Publishers, v. 14, n. 1, p. 3–18, 2018.

REMAMANY, Krishna Priya et al. Brain tumor segmentation in MRI images using integrated modified PSO-fuzzy approach. **Int. Arab J. Inf. Technol.**, v. 12, 6A, p. 797–805, 2015.

RIBEIRO, Matheus Gonçalves et al. Analysis of the Influence of Color Normalization in the Classification of Non-Hodgkin Lymphoma Images. In: IEEE. 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). [S.l.: s.n.], 2018. p. 369–376.

RIBEIRO, Matheus Gonçalves et al. Classification of colorectal cancer based on the association of multidimensional and multiresolution features. **Expert Systems with Applications**, v. 120, p. 262–278, 2019. ISSN 0957-4174. DOI: ⟨https://doi.org/10.1016/j.eswa.2018.11.034⟩. Disponível em: ¡⟨http://www.sciencedirect.com/science/article/pii/S0957417418307541⟩¿.

ROBERTO, Guilherme F et al. Features based on the percolation theory for quantification of non-Hodgkin lymphomas. **Computers in biology and medicine**, Elsevier, v. 91, p. 135–147, 2017.

RODRIGUES, Diego Alves. Deep Learning e redes neurais convolucionais: reconhecimento automático de caracteres em placas de licenciamento automotivo. Universidade Federal da Paraıba, 2018.

ROSENBLATT, Frank. Principles of neurodynamics. Spartan Book, 1962.

SAEYS, Yvan; INZA, Iñaki; LARRAÑAGA, Pedro. A review of feature selection techniques in bioinformatics. **bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2507–2517, 2007.

SAFAVIAN, S Rasoul; LANDGREBE, David. A survey of decision tree classifier methodology. **IEEE transactions on systems, man, and cybernetics**, IEEE, v. 21, n. 3, p. 660–674, 1991.

SAHINER, Berkman; CHAN, Heang-Ping; HADJIISKI, Lubomir. Classifier performance prediction for computer-aided diagnosis using a limited dataset. **Medical Physics**, Wiley Online Library, v. 35, n. 4, p. 1559–1570, 2008.

SHAMIR, Lior et al. IICBU 2008: a proposed benchmark suite for biological image analysis. **Medical & biological engineering & computing**, Springer, v. 46, n. 9, p. 943–947, 2008.

SHARMA, Puneet et al. **Method and system for non-invasive assessment of coronary artery disease**. [S.l.]: Google Patents, set. 2015. US Patent 9,119,540.

SIEGEL, Rebecca L; MILLER, Kimberly D; JEMAL, Ahmedin. Cancer statistics, 2020. **CA: A Cancer Journal for Clinicians**, Wiley Online Library, v. 70, n. 1, p. 7–30, 2020.

SILVA, Giovanni da et al. Classification of malignancy of lung nodules in CT images using Convolutional Neural Network. In: SBC. ANAIS do XVI Workshop de Informática Médica. [S.l.: s.n.], 2020. p. 21–29.

SINGH, Siddharth et al. Diagnostic performance of magnetic resonance elastography in staging liver fibrosis: a systematic review and meta-analysis of individual participant data. **Clinical Gastroenterology and Hepatology**, Elsevier, v. 13, n. 3, p. 440–451, 2015.

SIRINUKUNWATTANA, Korsuk et al. Gland segmentation in colon histology images: The glas challenge contest. **Medical image analysis**, Elsevier, v. 35, p. 489–502, 2017.

SMITH, Matthew G; BULL, Larry. Genetic programming with a genetic algorithm for feature construction and selection. **Genetic Programming and Evolvable Machines**, Springer, v. 6, n. 3, p. 265–281, 2005.

SONG, Yang et al. Low dimensional representation of fisher vectors for microscopy image classification. **IEEE transactions on medical imaging**, IEEE, v. 36, n. 8, p. 1636–1649, 2017.

SWERDLOW, Steven H; COOK, James R. As the world turns, evolving lymphoma classifications–past, present and future. **Human Pathology**, Elsevier, v. 95, p. 55–77, 2020.

TAGUCHI, Hiroshi et al. Lung cancer detection based on helical CT images using curved-surface morphology analysis. In: INTERNATIONAL SOCIETY FOR OPTICS e PHOTONICS. MEDICAL Imaging 1999: Image Processing. [S.l.: s.n.], 1999. v. 3661, p. 1307–1315.

TAPE, Thomas G. The area under an ROC curve. **Interpreting Diagnostic Tests [cited 2016 Nov 20]. Available from: http://gim. unmc. edu/dxtests/roc3. htm**, 2006.

TCHRAKIAN, Nairi et al. Quantitative image analysis of the MIB-1 proliferation index in Non-Hodgkin Lymphoma. **Journal of Histology & Histopathology**, Herbert Publications, v. 5, n. 1, p. 3, 2018.

TIWARI, Arti; SRIVASTAVA, Shilpa; PANT, Millie. Brain tumor segmentation and classification from magnetic resonance images: review of selected methods from 2014 to 2019. **Pattern Recognition Letters**, Elsevier, v. 131, p. 244–260, 2020.

TIWARI, Puneet; SACHDEVA, Jainy et al. Computer aided diagnosis system-a decision support system for clinical diagnosis of brain tumours. **International Journal of Computational Intelligence Systems**, Directory of Open Access Journals, v. 10, n. 1, p. 104, 2017.

VAN GINNEKEN, Bram; ROMENY, BM Ter Haar; VIERGEVER, Max A. Computer-aided diagnosis in chest radiography: a survey. **IEEE Transactions on medical imaging**, IEEE, v. 20, n. 12, p. 1228–1241, 2001.

VAN ROSSUM, Guido; DRAKE, Fred L. **Python language reference manual**. [S.l.]: Network Theory United Kingdom, 2003.

VAPNIK, Vladimir. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013.

VAPNIK, Vladimir N. An overview of statistical learning theory. **IEEE transactions on neural networks**, IEEE, v. 10, n. 5, p. 988–999, 1999.

VELU, CM; KASHWAN, KR. Visual data mining techniques for classification of diabetic patients. In: IEEE. ADVANCE Computing Conference (IACC), 2013 IEEE 3rd International. [S.l.: s.n.], 2013. p. 1070–1075.

VRJI, KS Angel; JAYAKUMARI, J. Automatic detection of brain tumor based on magnetic resonance image using CAD System with watershed segmentation. In: IEEE. SIGNAL Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on. [S.l.: s.n.], 2011. p. 145–150.

WANG, Pin et al. Cross-task extreme learning machine for breast cancer image classification with deep convolutional features. **Biomedical Signal Processing and Control**, Elsevier, v. 57, p. 101789, 2020.

WANG, Xiaosong et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. [S.l.: s.n.], 2017. p. 3462–3471.

WEISS, Sholom M; KAPOULEAS, Ioannis. An empirical comparison of pattern recognition, neural nets and machine learning classification methods. **Readings in machine learning**, p. 177–183, 1990.

WELIKALA, RA et al. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. **Computerized Medical Imaging and Graphics**, Elsevier, v. 43, p. 64–77, 2015.

WHITLEY, Darrell. A genetic algorithm tutorial. **Statistics and computing**, Springer, v. 4, n. 2, p. 65–85, 1994.

WILCOXON, Frank. Individual comparisons by ranking methods. In: BREAKTHROUGHS in statistics. [S.l.]: Springer, 1992. p. 196–202.

XERRI, Luc et al. The heterogeneity of follicular lymphomas: from early development to transformation. **Virchows Archiv**, Springer, v. 468, n. 2, p. 127–139, 2016.

YANG, Jihoon; HONAVAR, Vasant. Feature subset selection using a genetic algorithm. In: FEATURE extraction, construction and selection. [S.l.]: Springer, 1998. p. 117–136.

YANG, Xin-She; DEB, Suash. Cuckoo search via Lévy flights. In: IEEE. 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC). [S.l.: s.n.], 2009. p. 210–214.

YU, Songyang; GUAN, Ling. A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. **IEEE transactions on medical imaging**, IEEE, v. 19, n. 2, p. 115–126, 2000.

ZEZOS, Petros et al. 439-Toward Computer-Based Automated Mayo Score Classification in Ulcerative Colitis through Classical and Deep Machine Learning. **Gastroenterology**, Elsevier, v. 154, n. 6, s–99, 2018.

ZHANG, Xin et al. Design, synthesis and evaluation of genistein-polyamine conjugates as multi-functional anti-Alzheimer agents. **Acta pharmaceutica sinica B**, Elsevier, v. 5, n. 1, p. 67–73, 2015.

ZHI, Hui; LIU, Sanyang. Face recognition based on genetic algorithm. **Journal of Visual Communication and Image Representation**, Elsevier, v. 58, p. 495–502, 2019.

ZHUO, Li'an et al. Calibrated Stochastic Gradient Descent for Convolutional Neural Networks, 2019.

APÊNDICE A

A.1 Resultados considerando variações na taxa de mutação (m)

A.1.1 Taxa de mutação m = 0,3 - Colorretal

Tabela A.1: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 5 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificadores	*NumF*	AUC
5	*Chi Squared*	*Random Forest*	520	0,837
25	*Gain Ratio*	*J48*	685	0,796
50	***Gain Ratio***	**J48**	**1014**	**0,860**
100	*Gain Ratio*	*J48*	1006	0,848
500	*Info Gain*	*J48*	1073	0,850

Tabela A.2: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 25 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*Random Forest*	1050	0,855
25	*Relief*	*J48*	613	0,806
50	***Relief***	***J48***	**197**	**0,869**
100	*Info Gain*	*J48*	302	0,868
500	*Gain Ratio*	*J48*	169 396	0,820

Tabela A.3: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 50 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*	*J48*	400	0,895
25	*Chi Squared*	*J48*	168	0,876
50	***Relief***	167 ***Random Forest***	**175**	**0,928**
100	*Relief*	*Random Forest*	199	0,901
500	*Gain Ration*	*K-Nearest Neighbors*	320	0,887

Tabela A.4: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 500 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*Random Forest*	301	0,887
25	*Relief*	*Random Forest*	1158	0,896
50	***Relief***	***Random Forest***	**58**	**0,941**
100	*Relief*	*Random Forest*	241	0,941
500	elief	*Random Forest*	503	9.939

A.1.2 Taxa de mutação m = 0,5 - Colorretal

Tabela A.5: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 5 e iterações definidas como 5, 25, 50, 100 e 500,

Iterações	Método de Seleção	Classificadores	*NumF*	AUC
5	*Chi Squared*	*J48*	761	0,847
25	*Gain Ratio*	*J48*	1064	0,826
50	*Info Gain*	*J48*	513	0,788
100	***Gain Ratio***	***J48***	**1006**	**0,848**
500	*Info Gain*	*J48*	1073	0,816

Tabela A.6: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 25 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*	*J48*	1058	0,860
25	*Info Gain*	*J48*	470	0,890
50	***Relief***	***J48***	**130**	**0,932**
100	*Info Gain*	*J48*	280	0,901
500	*Gain Ratio*	*J48*	295	0,838

Tabela A.7: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 50 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*	*J48*	461	0,901
25	***Chi Squared***	***J48***	**144**	**0,909**
50	*Relief*	*J48*	146	0,852
100	*Relief*	*Random Forest*	145	0,880
500	*Gain Ratio*	*J48*	506	0,896

Tabela A.8: Colorretal: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 500 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Info Gain*	*J48*	34	0,896
25	*Chi Squared*	*J48*	39	0,902
50	***Relief***	***Random Forest***	**29**	**0,967**
100	*Relief*	*Random Forest*	119	0,963
500	*Relief*	*Random Forest*	387	0,961

A.1.3 Taxa de mutação m = 0,3 - Linfomas

Tabela A.9: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 5 e iterações definidas como 5, 25, 5, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Chi Squared*	*J48*	674	0,834
25	*Chi Squared*	*J48*	1069	0,820
50	*Relief*	*K-Nearest Neighbors*	485	0,894
100	*Relief*	*K-Nearest Neighbors*	372	0,855
500	***Relief***	***K-Nearest Neighbors***	**491**	**0,907**

Tabela A.10: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 25 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*Random Forest*	603	0,863
25	*Relief*	*Random Forest*	451	0,889
50	***Relief***	***K-Nearest Neighbors***	**183**	**0,914**
100	*Relief*	*K-Nearest Neighbors*	223	0,914
500	*Relief*	*K-Nearest Neighbors*	289	0,901

Tabela A.11: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 50 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Chi Squared*	*Random Forest*	414	0,876
25	*Relief*	*Random Forest*	783	0,895
50	***Relief***	***K-Nearest Neighbors***	**244**	**0,915**
100	*Relief*	*K-Nearest Neighbors*	290	0,902
500	*Relief*	*J48*	456	0,901

Tabela A.12: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 500 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*J48*	654	0,881
25	*Relief*	*J48*	1062	0,879
50	***Relief***	***K-Nearest Neighbors***	**538**	**0,902**
100	*Relief*	*K-Nearest Neighbors*	221	0,935
500	*Relief*	*K-Nearest Neighbors*	205	0,935

A.1.4 Taxa de mutação m = 0,5 - Linfomas

Tabela A.13: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 5 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*n	*J48*	875	0,810
25	*Chi Squared*	*J48*	1456	0,819
50	*Chi Squared*	*J48*	407	0,842
100	*Gain Ratio*	*J48*	1456	0,810
500	***Gain Ratio***	***KStar***	**997**	**0,853**

Tabela A.14: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 25 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Gain Ratio*	*Random Forest*	1058	0,860
25	*Info Gain*	*J48*	470	0,890
50	***Relief***	***Random Forest***	**130**	**0,932**
100	*Info Gain*	*J48*	280	0,901
500	*Gain Ratio*	*J48*	295	0,838

Tabela A.15: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 50 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Info Gain*	*Random Forest*	1106	0,823
25	*Info Gain*	*J48*	209	0,892
50	*Relief*	*J48*	145	0,904
100	*Relief*	*Random Forest*	351	0,902
500	***Relief***	***Random Forest***	**220**	**0,935**

Tabela A.16: Linfoma: melhor combinação (método de seleção, classificador, número de características e AUC média) obtida com P = 500 e iterações definidas como 5, 25, 50, 100 e 500.

Iterações	Método de Seleção	Classificador	*NumF*	AUC
5	*Relief*	*J48*	670	0,823
25	*Relief*	*J48*	457	0,888
50	***Relief***	***J48***	**73**	**0,947**
100	*Relief*	*Random Forest*	220	0,914
500	*Relief*	*Random Forest*	220	0,935

A.2 Valores utilizados no teste de *Wilcoxon-Mann-Whitney*

Tabela A.17: Valores de Acurácia utilizados no teste de *Wilcoxon-Mann-Whitney* para o câncer colorretal.

Tamanho da população (p)	Iterações ($Iter$)	AUC	Valores de Acurácia
50	**50**	**0,984**	**96,04%**
500	100	0,947	90,89%
50	5	0,915	90,36%
50	25	0,901	90,36%
50	500	0,900	89,68%
500	500	0,928	89,43%
50	100	0,911	88,42%

Tabela A.18: Valores de Acurácia utilizados no teste de *Wilcoxon-Mann-Whitney* para os linfomas NHL.

Tamanho da população (p)	Iterações ($Iter$)	AUC	Valores de Acurácia
500	**50**	**0,947**	**92,25%**
500	100	0,947	92,25%
500	500	0,939	91,54%
50	500	0,935	91,39%
50	50	0,913	89,55%
50	100	0,912	89,54%
25	50	0,896	87,86%