Master's Thesis Proposal

Proposta e Avaliação de um Modelo Híbrido de Seleção de Características para o Prognóstico do Câncer de Mama

Maxwell Esdra Acioli Silva
meas@ic.ufal.br

Advisers:

Dr. Rafael de Amorim Silva
Dr. Bruno Pimentel

Maceió
3 de Março, 2021

Maxwell Esdra Acioli Silva

Proposta e Avaliação de um Modelo Híbrido de Seleção de Características para o Prognóstico do Câncer de Mama

Tese apresentada por Maxwell Esdra Acioli Silva em cumprimento parcial dos requisitos para o grau de Mestre em Ciências da Informática da Universidade Federal de Alagoas, Instituto de Computação.

Advisers:

Dr. Rafael de Amorim Silva

Dr. Bruno Pimentel

Maceió
3 de Março, 2021

Tese apresentada por Maxwell Esdra Acioli Silva em cumprimento parcial dos requisitos para o grau de Mestre em Ciências da Informática da Universidade Federal de Alagoas, Instituto de Informática, aprovada pela banca examinadora que assina abaixo.

Dr. Rafael de Amorim Silva - Adviser
Computing Institute
Federal University of Alagoas

Dr. Bruno Pimentel - Adviser
Computing Institute
Federal University of Alagoas

Dr. A definir - Examiner
Computing Institute
Federal University of Alagoas

Dr. A definir - Examiner
Computing Institute
Federal University of Alagoas

Maceió
3 de Março, 2021

Acknowledgement

Primeiramente, a Deus que permitiu que tudo isso acontecesse, ao longo de minha vida,sem Ele nada disso seria possível. Obrigado, meu Senhor, por colocar amor, fé e esperança no meu coração.

Aos meus pais, Aleniude e Dorgival, e aos meus irmãos, Miquéias e Priscilla, por todo amor, estímulo e apoio incondicional.

À minha amada esposa Helynne, que deu forças para eu vencer essa etapa da minha vida.

Ao Profº. Dr. Rafael Amorim, pela orientação, suporte e confiança.

Resumo

A tecnologia de Inteligência Artificial tem sido fundamental no papel do cuidado à saúde da sociedade. Ela vem sendo amplamente utilizada nos diversos ramos da medicina. Uma de suas principais aplicações é no contexto do prognóstico da doença de câncer de mama. O câncer é considerado como a segunda maior causa de mortes decorrentes de doenças no mundo. Neste contexto, destaca-se o câncer de mama, que é considerado a maior ocorrência de câncer entre as mulheres no mundo. Um dos principais desafios neste cenário é identificar quais são as características mais relevantes no desenvolvimento deste tipo de neoplasia por um paciente. Este tipo de filtro é realizado pelos métodos de seleção de características. Este trabalho apresenta um modelo híbrido de seleção de características que deve ser utilizado por clínicos de um paciente, a fim de realizar uma predição de recorrência do câncer de mama.

Palavras-chave: Aprendizagem de Máquina, Prognóstico, Câncer de Mama, Recorrência.

Abstract

Artificial Intelligence technology has been instrumental in the role of health care in society. It has been used in several branches of medicine. One of its main applications is in the context of the prognosis of breast cancer disease. Cancer is considered to be the second leading cause of death from disease in the world. In this context, breast cancer stands out, which is considered the highest occurrence of cancer among women in the world. One of the main challenges in this scenario is to identify which are the most relevant characteristics in the development of this type of neoplasia by a patient. This type of filter is carried out using the characteristic selection methods. This work presents a hybrid model of selection of characteristics, which is used by the patient's clinicians, in order to make a prediction of breast cancer recurrence.

Keywords: Machine Learning, Prognosis, Breast Cancer, Recurrence.

Contents

List of Figures

List of Tables

List of Algorithms

1

Introdução

No cenário de Machine Learning (ML), dados podem ser definidos como um fato, texto, imagem ou som que não foi processado. São considerados partes essenciais no contexto de ML, pois sem estes não é possível treinar o modelo, e consequentemente inferir alguma informação do objeto de estudo. Um modelo de aprendizagem de máquina é caracterizado como um algoritmo que tem a capacidade de reconhecer padrões sobre um conjunto de dados aplicados a este.

Com os sucessivos avanços que houveram na Inteligência Artificial (IA) e ML, estas tecnologias passaram a ser consideradas fundamentais no papel do cuidado à saúde da sociedade. Sendo amplamente utilizadas nos diversos ramos da medicina. Nesse universo, os dados do paciente utilizados em modelos de ML, podem ser divididos em dois tipos: (i) dados clínicos e (ii) dados moleculares. Dados clínicos são aqueles coletados a partir de diagnóstico, testes laboratoriais e dados hereditários do paciente. Já os dados moleculares, também chamados de microarranjos ou dados genômicos, podem ser definidos como um conjunto de dados que contêm informações das células do indivíduo, por exemplo, a sequência de RNA.

Uma das principais aplicações das tecnologias acima citadas, se dá no contexto de prognóstico e predição da neoplasia câncer. O câncer é considerado como a segunda maior causa de mortes decorrentes de doenças no mundo, de acordo com Organização Mundial de Saúde (OMS) WHO, segundo senso realizado no ano de 2018. Neste contexto, destaca-se o câncer de mama. Este, por sua vez, é considerado a maior ocorrência de câncer entre as mulheres em todo o mundo WCRF.

Um dos principais desafios da medicina neste cenário, é utilizar as tecnologias disponíveis para conseguir fazer o prognóstico ou predição do câncer de mama com a melhor assertividade possível. Sendo assim, já existem diversos modelos apresentados na literatura com o objetivo de ajudar neste tipo de predição. Alguns destes utilizam dados clínicos dos pacientes, outros utilizam dados moleculares dos mesmos.

Quando trata-se da utilização de dados moleculares dos pacientes, a utilização de uma ferramenta capaz de reduzir a dimensionalidade dos dados passou a ser fundamental devido ao

tamanho das características contidas nestes tipos de dados. Como mencionado anteriormente, os avanços tecnológicos possibilitaram obter informações cada vez mais detalhadas acerca das células dos indivíduos, de tal forma que a quantidade de informações disponíveis chegam à casa das milhares. Portanto, é necessário utilizar algum mecanismo apto a reduzir a quantidade de informações, pois nem todas elas são relevantes para a predição da doença em questão, ou até mesmo algumas delas podem ser duplicadas. Neste cenário, surge o conceito de modelos de seleção de características, que pode ser denominada como um processamento dos dados analisados, a fim de reduzir a quantidade de características que serão utilizadas em um modelo de aprendizado de máquina, com o principal objetivo de encontrar quais são as características mais significativas dentre todas as disponíveis na análise [3].

Segundo levantamento feito na literatura acerca de quais são os métodos de seleção de características mais utilizados neste tipo de prognóstico, destacam-se os métodos de seleção de característica: (i) ReliefF; (ii) Information Gain; e (iii) Gain Ratio [Adicionar referência da nossa revisão].

Portanto, este trabalho apresenta um método híbrido de seleção de características, utilizando dados clínicos e moleculares dos pacientes com susceptibilidade a desenvolverem o câncer de mama. Analisando o desempenho dos dados resultantes da seleção sobre os modelos de aprendizagem de máquina aplicados no contexto de predição da neoplasia em análise.

1.1 Trabalhos Relacionados

A utilização de técnicas de seleção de características na detecção de propriedades determinantes no prognóstico do câncer vêm sendo estudadas ao longo dos anos, muitos trabalhos sobre este tema foram publicados. Neste contexto, foi realizada uma revisão sistemática de literatura, com o objetivo de identificar trabalhos que tratam da utilização de técnicas de seleção de características no contexto do prognóstico da doença de câncer, durante a execução do protocolo definido na revisão, foi possível identificar que não existem muitos trabalhos que tratam deste tema, no final da revisão foram selecionados 21 trabalhos para ser feita uma análise estatística sobre eles a fim de identificar aspectos relevantes sobre o tema proposto Silva et al. (2020).

Um revisão sistemática é uma revisão da literatura realizada a partir de uma pergunta de pesquisa definida, por meio da qual se busca identificar, avaliar, selecionar e sintetizar evidências de estudos empíricos que atendam a critérios de elegibilidade predefinidos Garcia (2014).

Através da revisão foi possível identificar que o principal tipo de dados utilizados no prognóstico de câncer são dados moleculares da doença. Um dado molecular pode ser definido como um dado obtido a partir da utilização de técnicas de biotecnologia a fim de extrair informações mais detalhadas acerca da doença analisada, por exemplo, uma sequência de DNA ou um conjunto de microarranjos das células. Também foi identificado que as técnicas mais utilizadas são:

ReliefF, Information Gain, Gain Ratio, Random Forest e T-Test Silva et al. (2020).

1.2 Objetivos

O nosso trabalho é voltado para solução do problema de seleção de características biológicas aplicadas em modelos de aprendizagem de máquina no prognóstico da doença de câncer de mama. As atividades da nossa proposta podem ser descritas da seguinte maneira:

- Realizadas:
  - (i) Revisão sistemática da literatura para identificar quais são as técnicas mais utilizadas neste contexto;
  - (ii) Definição de um novo modelo híbrido utilizando as técnicas de seleção de características já existentes Relief-F e Gain Ration;
  - (iii) Obtenção de um dataset contendo informações relevantes para predição da recorrência do câncer de mama;
- Sendo executadas atualmente:
  - (i) Adaptação dos dois modelos de selecionados para construção do modelo híbrido proposto;
  - (ii) Teste do modelo híbrido proposto;
- A fazer:
  - (i) Extração de características mais relevantes no prognóstico da recorrência do câncer de mama;
  - (ii) Comparação dos resultados obtidos utilizando o método híbrido com os métodos existentes na literatura.

1.3 Estrutura do Trabalho

Este trabalho está estruturado da seguinte forma: capítulo 2 apresenta elementos importantes que são bastantes relevantes no entendimento teórico, tais como fatores prognósticos para a doença de câncer de mama. Além de trazer uma breve explicação sobre algoritmos de seleção de características. Já o capítulo 3 traz uma breve visão geral sobre os algoritmos utilizados para a construção do algoritmo proposto neste. No capítulo 4 apresentamos o conjunto de dados utilizados no experimento, bem como os resultados obtidos em um pré-experimento realizado utilizando o algoritmo proposto. Por fim, o capítulo 5 traz as considerações finais da proposta, bem como apresenta o cronograma do planejamento para o desenvolvimento das atividades até a defesa da tese proposta.

2

Background

Segundo a Organização Mundial de Saúde (OMS), de acordo com senso realizado no ano de 2020, o câncer é considerada com a segunda maior causa de mortes decorrentes de doença no mundo WHO. Diante deste cenário, o câncer de mama, por sua vez, é considerado como o de maior ocorrência entre as mulheres em todo o mundo. Estima-se a ocorrência de aproximadamente 9,9 milhões de mortes no mundo em decorrência da doença de câncer, dentre este total de mortes, cerca de 6,9% destas mortes são de câncer de mama Sung et al..

2.1 Câncer de mama

Durante o decorrer do último século o conhecimento médico acerca da doença de câncer evoluiu consideravelmente, juntamente com esta evolução foi possível obter uma maior compreensão sobre o mesmo. Isso foi possível graças ao surgimento de novas tecnologias terapêuticas e diagnósticas, as quais possibilitaram descobrir de uma forma mais antecipada a existência de células cancerígenas nos pacientes, bem como, fornecer qual a terapia específica para o tipo de câncer em questão. No Brasil, as primeiras preocupações acerca da doença de câncer surgiram em meados de 1920. Porém, só a partir de 1940 é que surgiram as primeiras instituições especializadas no estudo acerca da doença de câncer. Além disso, foi a partir deste período que iniciaram-se as primeiras campanhas educativas, destacando a importância do diagnóstico como a forma mais efetiva no tratamento da doença, pois quanto mais cedo descobrir-se a doença, maior são as chances de cura Teixeira and AraÃ Neto (2020).

A possibilidade de obtenção do diagnóstico do câncer de mama no Brasil, ganhou outro fator importante a partir da inclusão de imagens de exames que possibilitavam a visualização das primeiras lesões mamárias. A partir daí houve uma maior mobilização para a atenção acerca da saúde da mulher, consequentemente houve um maior rastreamento do câncer de mama Teixeira and AraÃ Neto (2020). O câncer de mama é causado pela multiplicação desordenada das células mamárias, gerando células anormais que se multiplicam, gerando um tumor INCA.

Existem diversos tipos de exames que permitem o diagnóstico de nódulos mamários. Os principais são: a mamografia, ultra-sonografia e ressonância magnética. A mamografia é considerada a mais importante entre os exames, pois trata-se do método mais indicado na avaliação de alterações na mama de pacientes assintomáticas. Este tipo de técnica, atualmente possibilita dois tipos de formação de imagens, a primeira é formada pelo conjunto *filme-écran*, que é considerada a mamografia convencional. Já o segundo tipo de imagem é obtido a partir de um receptor digital, esta também conhecida como mamografia digital Chala and Barros (2007).

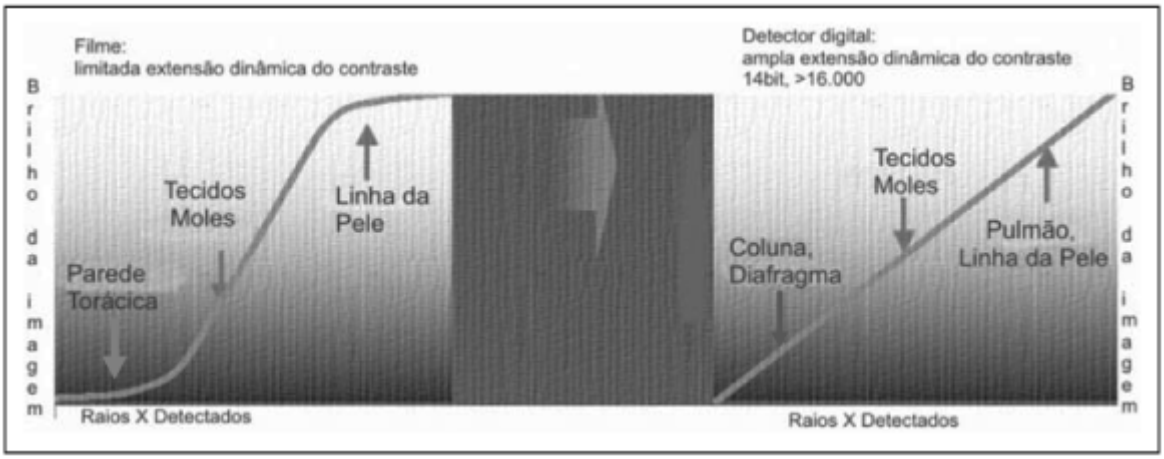

Figure 2.1: Comparação da extensão dinâmica do contraste na aquisição digital e de filme da imagem da mama

(Freitas et al., 2006)

Na mamografia convencional, o meio de aquisição da imagem, de exposição e armazenamento da é representado pelo filme, este por sua vez não deixa uma margem para melhoria da imagem gerada, apesar de fornecer um bom contraste e um boa resolução espacial. Já na mamografia digital, a aquisição, exposição e armazenamento das imagens são feitos em processos distintos. O que possibilita um melhor aperfeiçoamento das imagens, além disso o processo de análise da imagem gerada é feita através da utilização de um monitor de alta resolução, o que permite melhorar o contraste das imagens, a Figura 2.1 demonstra os níveis de contraste nos dois tipos de imagens. Porém a capacidade de detecção do câncer de mama através de imagens de mamografia varia de paciente para paciente, o mais importante destes fatores é a densidade radiológica da mama Chala and Barros (2007). A Figura 2.2 demonstra a comparação entre imagem da mama obtida a partir de uma mamografia convencional e outra a partir do método digital.

O principal método adjacente a mamografia na detecção do câncer de mama é a ultrasonografia. Que é um método que obtém uma boa imagem dos tecidos mamários. Este tipo de método é feito através de um aparelho que emite ondas sonoras de alta frequência. A vibração dos tecidos produz um eco, que por sua vez é lido pelo aparelho e consequentemente convertido em imagem. Este tipo de imagem é mais indicado quando o objetivo é diferenciar e caracterizar nódulos sólidos e cistos previamente identificados pela mamografia. Por conta

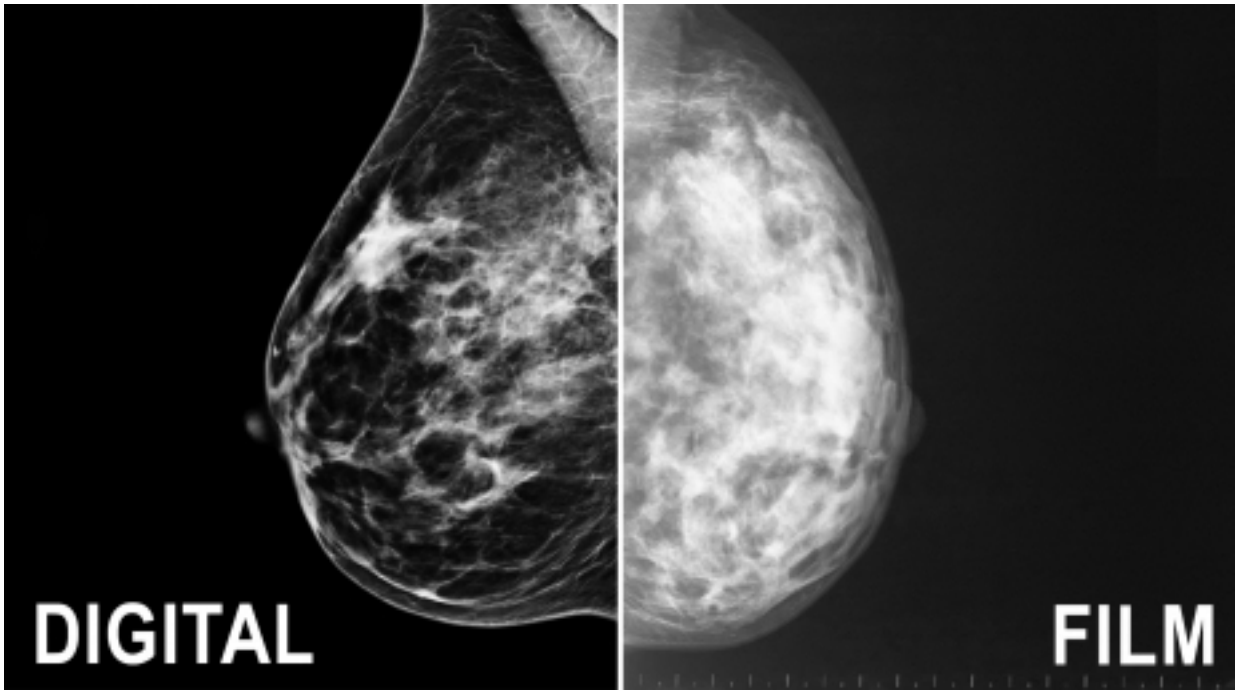

Figure 2.2: Comparação entre imagens de mamografia convencional e digital (BedfordBreastCancer)

disto, esta técnica é tida como um método suplementar à mamografia. A Figura 2.3 demonstra aspecto característico de cisto simples à esquerda (seta branca), que é bem circunscrito com margem imperceptível, anecóica, e demonstra por transmissão. Isso está em contraste com o câncer de mama à direita (seta branca), que é mal margeado, de forma irregular, hipoecóico e não apresenta transmissão por via.

Por fim, a ressonância magnética é outra forma diagnosticar o câncer de mama. Este tipo de exame também é considerado uma forma auxiliar no diagnóstico da doença juntamente com a mamografia e a ultrassonografia. Nele é utilizado um aparelho de ressonância magnética, que faz o uso de ímãs para gerar as imagens. Ele vem sendo bastante utilizado em pacientes com um alto risco de desenvolvimento desta neoplasia Chala and Barros (2007). Na Figura 2.4 podemos verificar algumas imagens de uma ressonância magnética da mama.

Diante disto, verifica-se que é necessário aplicar o tipo de exame mais apropriado para o paciente em questão, para fazer o diagnóstico precoce do câncer de mama. Uma vez que foi feito o diagnóstico, é possível aplicar as medidas terapêuticas mais acertivas para o caso da paciente, bem como utilizar alguma técnica disponível afim de obter o quadro prognóstico da doença.

2.2 Fatores prognóstico para o câncer de mama

O curso clínico da doença de câncer de mama, bem como sua sobrevida, podem variar de paciente para paciente, segundo a história natural da neoplasia. Tal variação pode ser de-

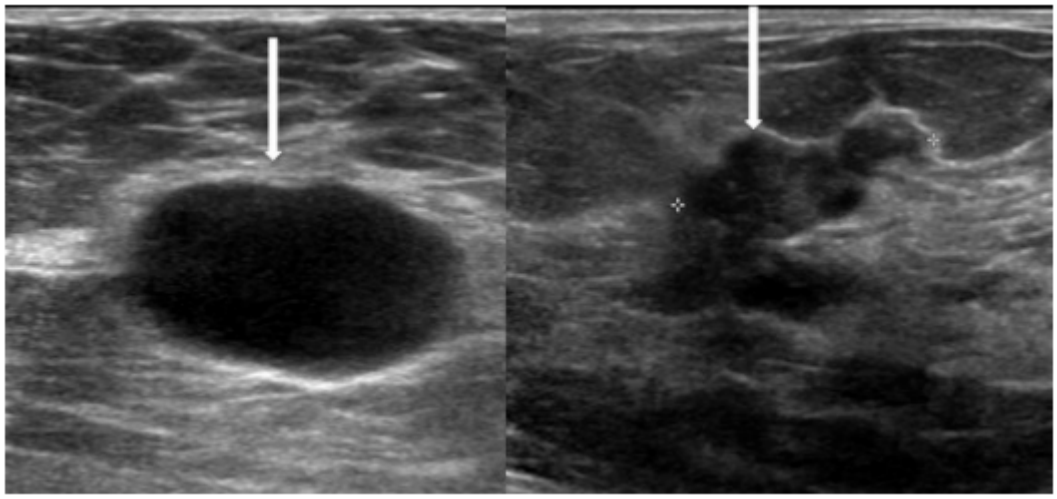

Figure 2.3: Demonstração de um cisto identificado na mama através de uma imagem de ultrassom

(Dunne et al., 2017)

terminada por uma séria de fatores complexos, por exemplo, a diferença na rapidez da duplicação tumoral, a capacidade de metastização do nódulo, ou outros fatores que, por ora, ainda não são totalmente compreendidos, que estão relacionados à condição hormonal, nutricional e imunológica do paciente. Porém, alguns fatores anatômicos continuam sendo de fundamental importância na avaliação prognóstica da doença, tal como o tamanho do nódulo primário e a condição dos linfonodos. Além disso, fatores relacionados às características biológicas e histológicas do tumor também são determinantes para o prognóstico da evolução do câncer Freitas JÃet al. (2017).

Os fatores prognósticos do câncer de mama podem ser úteis em três situações segundo Clark Clark et al. (1994): a primeira diz respeito a identificação de pacientes onde o prognóstico é suficiente, e nenhum tratamento adjuntivo ao tratamento cirúrgico agregará algum benefício ao paciente; a segunda diz respeito a obtenção de um prognóstico ruim em relação ao tratamento convencional, que qualquer outra forma de tratamento mais intensa deveria ser aplicado; a terceira, é aquela cujo o prognóstico é capaz de indicar uma terapia específica para um paciente específico. Um fator prognóstico pode ser definido como um parâmetro possível de ser mensurado no momento do diagnóstico e que serviria como preditor da sobrevida ou do tempo livre de doença.

2.3 Algoritmos para seleção de características

Podemos definir características, no contexto de aprendizagem de máquina (ou do inglês "machine learning" ou ML), podem ser definidas como àquelas propriedades de fenômeno observado. Geralmente são utilizadas em algum processo de identificação sistemática de padrões, através do uso de modelos de ML. Denominamos seleção de características como um processo

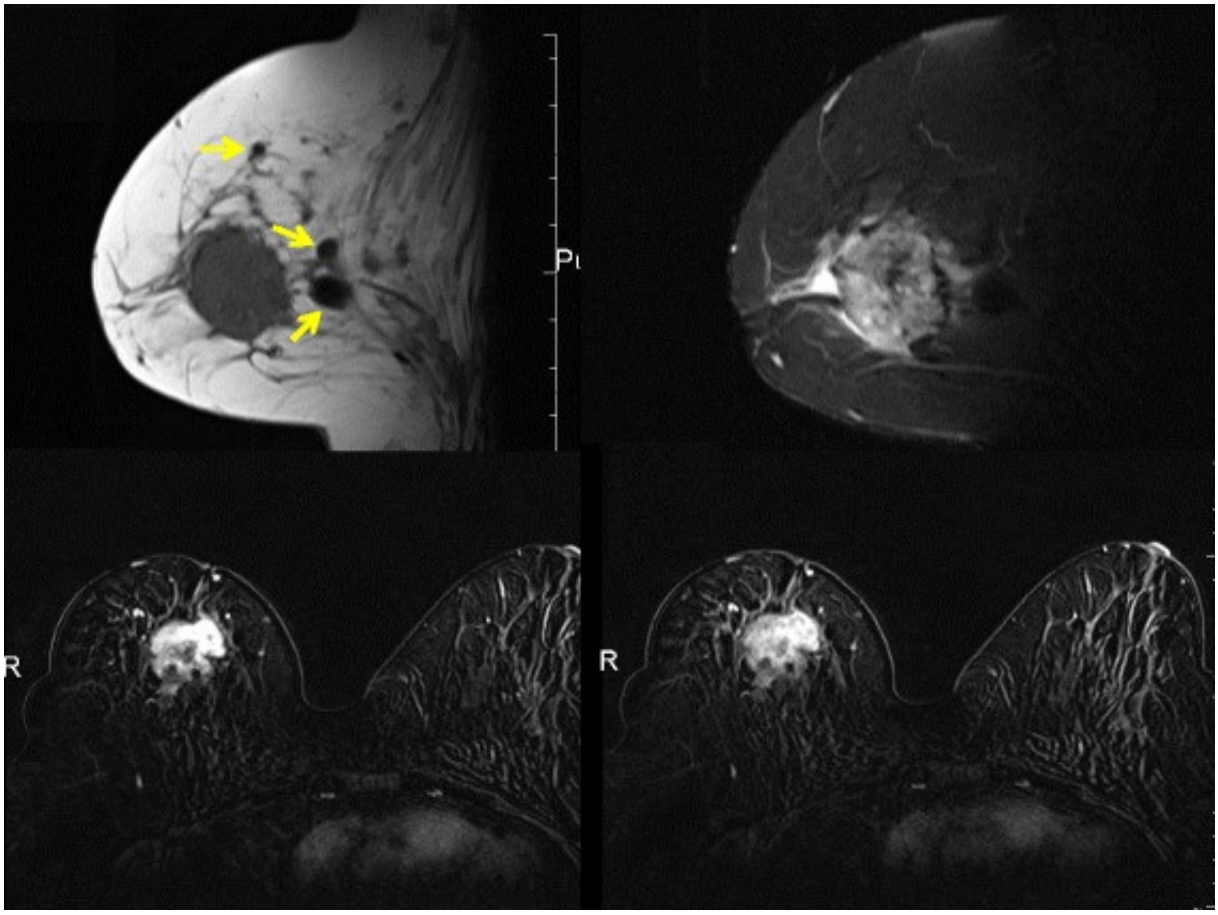

Figure 2.4: Imagens de uma ressonância magnética da mama
(Nakahori et al., 2015)

que tem o objetivo de reduzir a quantidade de propriedades que são aplicadas em um modelo de ML. A etapa de redução de características pode ser feita através da análise e processamento das características, com a finalidade de reconhecer quais destas são as mais relevantes dentre todas as existentes Sherer (2018). Um exemplo de aplicação de seleção de características na área médica é na análise de genes de um microarranjo de amostras de célula de um determinado organismo de um paciente. Este tipo de dado contém milhares de variáveis, que apresentam uma certa correlação entre elas. Este tipo de dependência entre as propriedades, não fornece informação extra sobre a variável alvo, sendo assim geram ruídos para o modelo de aprendizagem de máquina utilizado. Sendo assim, eliminar este tipo de variável, pode resultar em uma melhoria no desempenho do resultado do modelo utilizado.

Portanto, faz-se necessário a definição de um critério para remoção de características que não tem um nível de relevância aceitável com a variável alvo. Por outro lado, vale a pena ressaltar que a remoção de características irrelevantes não tem relação com métodos de redução de dimensionalidade de um conjunto de dados, tais como o Principal Component Analysis (PCA). A eliminação de propriedades através de técnicas de seleção de características não cria novas propriedades. Pois uma vez que é selecionado um critério de seleção de característi-

cas, é iniciado um procedimento para encontrar um subconjunto de características úteis dentro do conjunto analisado Chandrashekar and Sahin (2014).

Os modelos de seleção de características podem ser distinguidos de acordo com sua relação com o modelo de ML, e geralmente são classificados em três tipos: a) filters; b) wrappers; e c) embedded.

2.3.1 Filters

Métodos do tipo filter Figura 2.5 estabelecem um rank das características utilizando algum critério pré-estabelecido. São executados em uma etapa anterior a aplicação do uso do modelo de ML, e são independentes do modelo aplicado no estudo. Este tipo de técnica pode ser classificado de acordo com os parâmetros de filtragem que empregam, tais como nível de dependência entre as propriedades ou grau de similidade. Alguns exemplos deste tipo de algoritmo são o Information Gain, Chi-Square, Gain Ration e ReliefF Urbanowicz et al. (2018).

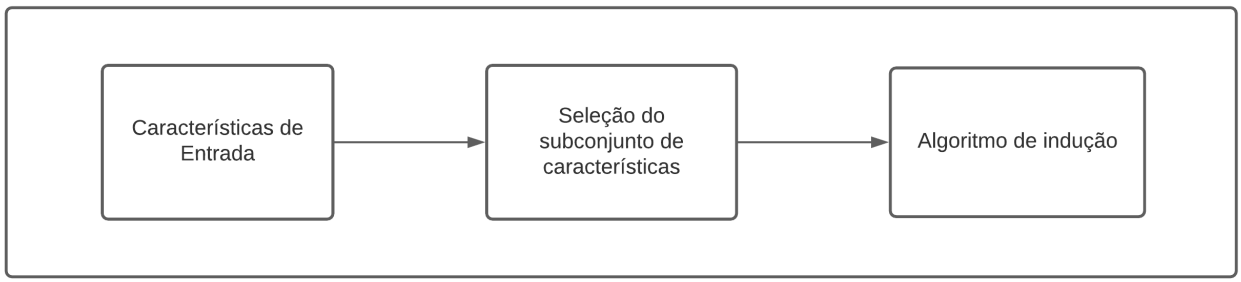

Figure 2.5: Modelo de seleção de características do tipo *filter*

2.3.2 Wrappers

Métodos wrappers Figura 2.6 funcionam criando uma quantidade de subconjuntos pré-definido de acordo com a quantidade total de características, em seguida calcula um score para cada subconjunto utilizando uma função objetiva específica. O subconjunto com a maior performance é considerado em detrimento dos demais subconjuntos. Geralmente este tipo de modelo demanda um maior esforço computacional em relação aos algoritmos filters Singh (2019).

Diferentemente de métodos do tipo filter, métodos wrappers estão relacionados ao modelo e aprendizagem de máquina aplicado, e é executado na fase de treinamento do modelo, caso seja utilizado um novo modelo de aprendizagem, deve ser refeito o processo de definição do subconjunto de características.

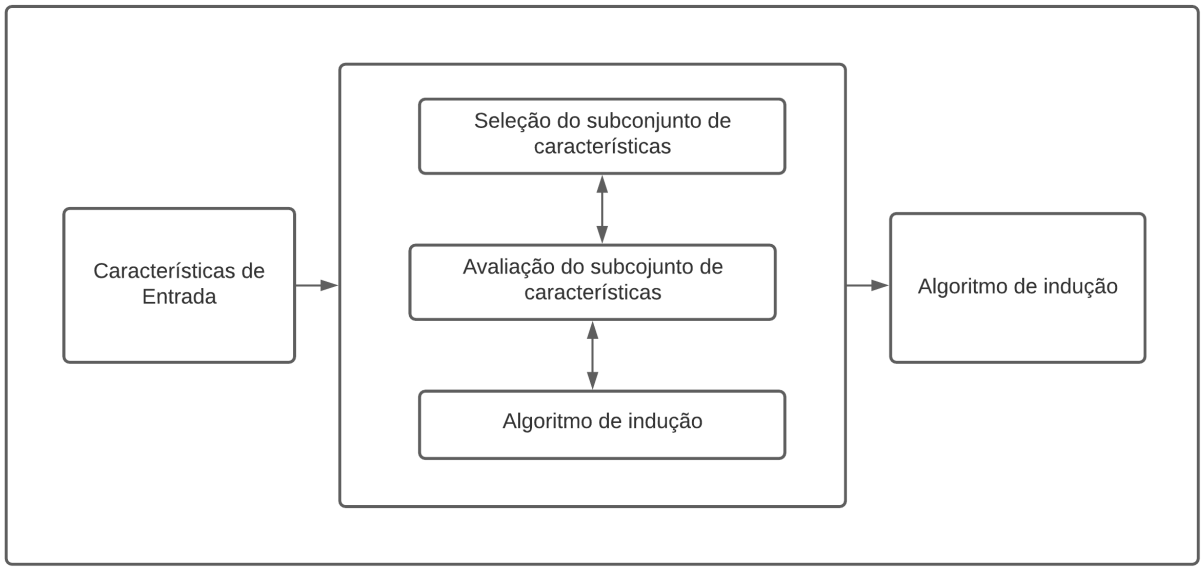

Figure 2.6: Modelo de seleção de características do tipo *wrapper*

2.3.3 Embedded

Os métodos embedded podem ser definidos como a combinação dos métodos citados anterior. Com a vantagem de ser mais performático que os métodos wrappers, pois realiza a integração de forma paralela entre o modelo e as características selecionadas. Isto pode ser realizada, por exemplo, através de uma função objetiva dividida em duas partes, primeiramente definindo um termo de adequação e, estabelecendo uma penalidade para um subconjunto com um número elevado de características.

3

Algoritmos Propostos

Neste capítulo apresentaremos os algoritmos propostos para solucionar o problema da identificação de características biológicas mais relevantes no prognóstico da doença de câncer de mama. Primeiramente, apresentaremos uma visão geral do algoritmo Relief-F. Em seguida, abordaremos o funcionamento do algoritmo Gain Ratio. Por fim, apresentaremos a abordagem utilizada para criação do algoritmo híbrido proposto neste trabalho.

3.1 Relief-F

Esta seção apresenta uma simples visão do funcionamento do algoritmo Relief-F.

O algoritmo Relief-F pertence à família de algoritmos Relief. Os algoritmos desta família podem ser divididos em três grupos: o algoritmo básico Relief; o algoritmo Relief-F e o algoritmo RReliefF. O algoritmo básico é limitado a ser utilizado na classificação de problemas com duas classes. Já o algoritmo Relief-F é considerado uma extensão do algirmto básico, e consegue se sair bem no cenário em que temos multiclasses. Por fim, o algoritmo RReliefF é uma adaptação do algoritmo Relief-F para problemas de regressão Robnik-Šikonja and Kononenko (2003).

Uma das vantagens do algoritmo em relação aos outros algoritmos de seleção de características é que, enquanto outros modelos assumem uma independência condicional entre os atributos do conjunto de dados para estimar a qualidade. Já os algoritmos da família ReliefF não fazem esta suposição, e são bastante eficientes na seleção das propriedades, levando em consideração na avaliação da qualidade das propriedades a dependência existente entre estas Robnik-Šikonja and Kononenko (2003).

A ideia original do algoritmo Relief, demonstrada no algoritmo 1, é estimar a qualidade dos atributos de um conjunto de dados de acordo com a proximidade entre as instâncias pertencentes ao mesmo. Para realizar esta distinção, primeiramente é selecionada uma instância aleatória R_i (linha 3), em seguida são procurados os vizinhos mais próximos pertencentes a

mesma classe, chamando o vizinho que tem maior proximidade de H, e outro vizinho de classe diferente, é chamado de M (linha 4). Após isto, atualiza-se a estimativa da qualidade $W[A]$ para todos os atributos A dependendo de seus valores R_i, M e H (linhas 5 e 6). Observando que, se as instâncias R_i H tiverem valores diferentes do atributo A, então este atributo separa duas instâncias pertencentes a mesma classe, então diminui-se a estimava da qualidade de $W[A]$. Por outro lado, se A separa duas instâncias com valores de classes diferentes, então aumenta-se a estimativa da qualidade de $W[A]$, o que é desejável. Dessa forma, todo este processo é repetido m vezes, onde m é um parâmetro definido pelo usuário Robnik-Šikonja and Kononenko (2003).

Algorithm 1: Relief Original

Input : Para cada instância de treinamento, um vetor de valores de atributos e a classe valor

Output: O vetor W de estimativas das qualidades dos atributos

1 set all weights $W[A] := 0.0$;
2 **for** $i := 1$ ***to*** m **do**
3 randomly select an instance R_i;
4 find nearest hit *H* and nearest miss *M*;
5 **for** $A := 1$ ***to*** a **do**
6 $W[A] := W[A] - \mathit{diff}(A, R_i, H)/m + \mathit{diff}(A, R_i, M)/m$;
7 **end**
8 **end**

A função $\mathit{diff}(A, I_1, I_2)$ calcula a diferença existente das duas instâncias I_1 e I_2 para o atributo A. Para atributos nominais ela é definida como:

$$\mathit{diff}(A, I_1, I_2) = \begin{cases} 0, & \mathit{value}(A, I_1) = \mathit{value}(A, I_2). \\ 1, & \text{otherwise.} \end{cases}, \tag{3.1}$$

e para atributos numéricos:

$$\mathit{diff}(A, I_1, I_2) = \frac{|\mathit{value}(A, I_1) - \mathit{value}(A, I_2)|}{max(A) - min(A)} \tag{3.2}$$

O algoritmo original Relief consegue trabalhar bem em conjunto de dados com atributos nominais e numéricos. Porém, ele é limitado a funcionar em conjuntos com duas classes. Para casos de dados com multiclasses, é utilizada sua extensão Relief-F. Cujo algoritmo é descrito a

seguir:

Algorithm 2: Relief-F

Input : Para cada instância de treinamento, um vetor de valores de atributos e o valor da classe

Output: O vetor W de estimativas das qualidades dos atributos

1 set all weights $W[A] := 0.0$;
2 **for** $i := 1$ ***to*** m **do**
3 randomly select an instance R_i;
4 find k nearest hit H_j;
5 **for** *each class* $C \neq$ *class* (R_i) **do**
6 from class C find k nearest misses $M_j(C)$;
7 **end**
8 **for** $A := 1$ ***to*** a **do**
9 $W[A] := W[A] - \sum_{j=1}^{k}$ *diff*(A, R_i, H_j)$/(m.k) + \sum_{C \neq (R_i)}(\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k}$ *diff*(A, $R_i, M_j(C))/(m.k)$);
10 **end**
11 **end**

O algoritmo Relief-F não está limitado a problemas com classes binárias, ele é mais robusto que o original e pode trabalhar bem em dados com ruídos. Seu funcionamento é similar ao algoritmo original, inicialmente é selecionada uma instância R_i aleatória (linha 3), depois ocorre a busca pelo k vizinho mais próximo pertencente a mesma classe, chamado de H_j (linha 4), também é buscado o vizinho mais próximo pertencente a outra classe, este chamado de $MJ_j(C)$ (linhas 5 e 6). A atualização da estimativa $W[A]$ da qualidade de todos atributos A depende dos valores para R_i, H_j e $M_j(C)$ (linhas 7, 8 e 9). Para trabalhar com dados faltantes é necessário mudar a função *diff*. Valores faltantes são tratados utilizando probabilidade. Calcula-se a probabilidade de duas instâncias que apresentam valores diferentes para um determinado atributo condicionado ao valor da classe da seguinte maneira:

se uma instância tem algum valor desconhecido:

$$\textit{diff}(A, I_1, I_2) = 1 - P(value(A, I_2)|class(I_1)) \tag{3.3}$$

se ambas instância têm algum valor desconhecido:

$$\textit{diff}(A, I_1, I_2) = 1 - \sum_{V}^{\#values(A)} (P(V|class(I_1)) \times P(V|class(I_1))) \tag{3.4}$$

Utilizam-se frequências relativas do conjunto de treinamento para realizar a aproximação de probabilidades condicionais. (Referência: Theoretical and Empirical Analysis of ReliefF and RReliefF)

3.2 Gain Ratio

Nesta seção apresentaremos uma breve introdução do funcionamento da medida seleção de característica *gain ratio*. Um dos modelos de aprendizagem de máquina mais utilizados é o de árvore de decisão. Ele é caracterizado como uma estrutura simples, onde nós não-terminais representam testes sobre um conjunto de atributos, e os nós terminais representam as saídas do problema. Neste contexto, o conceito de ganho de informação é definido como uma medida utilizada para selecionar a ordem dos atributos testes da árvore de decisão. Esta medida dá preferência à utilização de atributos que têm um grande número de valores. Existem vários tipos de algoritmos de árvore de decisão, o mais popular é o ID3. Porém, ao passar dos anos surgiram novas versões de algoritmo de indução, como por exemplo , o algoritmo C4.5. Este algoritmo, que é tido como o sucessor do ID3, utiliza uma extensão de ganho de informação, denominada de *gain ratio*. Que tem como objetivo diminuir o viés da seleção de atributos com um alto número de valores. As informações necessárias para classificar uma determinada amostra, pode ser definida da seguinte forma:

$$I(S) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{3.5}$$

Onde S é definido é constituído de s amostras de dados com m classes distintas, e p_i é a probabilidade de uma amostra arbitrária pertencer a classe C_i e calcula através de s_i/s. Considerando que um atributo A tem valores distintos v. E s_{ij} seja o número de amostras das classe c_i em um subconjunto s_j. S_j contém as amostras em s que tem o valor a_j de A. Dessa forma, a entropia, ou informação esperada baseada na partição do conjunto A em subconjuntos é dada por:

$$I(S) = -\sum_{i=1}^{m} I(S) \frac{s_{1i} + s_{2i} + ... + s_{mi}}{s} \tag{3.6}$$

O cálculo do ganho de informação a partir da partição do conjunto A é definido como:

$$Gain(A) = I(S) - E(A) \tag{3.7}$$

O algoritmo C4.5, que utiliza o *gain ratio*, aplica uma normalização do ganho de informação utilizando uma informação definida como:

$$SplitInfo_A(S) = -\sum_{i=1}^{v} (|S_i| / |S|) log_2(|S_i| / |S|) \tag{3.8}$$

Este valor representa a informação produzida através da divisão do conjunto de dados S em v partições correspondente a n resultados de testes realizados sobre o atributo A. O *gain ratio* é definido como:

$$GainRatio(A) = Gain(A)/SplitInfo_A(S) \tag{3.9}$$

Sendo assim, o atributo com a maior taxa de ganho é selecionado como atributo teste da árvore de decisão. (Referência: COMPARATIVE STUDY OF ATTRIBUTE SELECTION USING GAIN RATIO AND CORRELATION BASED FEATURE SELECTION)

3.3 Algoritmo Híbrido

Um algoritmo híbrido pode ser definido como a combinação de dois ou mais algoritmos afim de resolver o mesmo problema. Segundo Silva et al. (2020) alguns dos algoritmos mais utilizados para o prognóstico da doença de câncer são os algoritmos Relief-F e Gain Ratio. Levando este fator em consideração, a ideia proposta é realizar a combinação destes algoritmos e criar um novo modelo híbrido. O novo algoritmo consiste de realizar a seleção de característica sobre um conjunto de dados selecionando inicialmente um conjunto de atributos utilizando o algoritmo Relief-F, a partir do resultado desta primeira seleção, aplicar o algoritmo Gain Ratio, afim de refinar a seleção de atributos.

Algorithm 3: Algoritmo Hibrido

Input : Para cada instância de treinamento, um vetor de valores de atributos e o valor da classe
Output: O vetor W de estimativas das qualidades dos atributos

set all weights $W[A] := 0.0$;
for $i := 1$ ***to*** m **do**
  randomly select an instance R_i;
  find k nearest hit H_j;
  for *each class* $C \neq$ *class* (R_i) **do**
    from class C find k nearest misses $M_j(C)$;
  end
  for $A := 1$ ***to*** a **do**
    $W[A] := W[A] - \sum_{j=1}^{k}$ *diff*(A, R_i, H_j)$/(m.k) + \sum_{C \neq (R_i)}(\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k}$ *diff*(A, $R_i, M_j(C)$)$/(m.k)$);
  end
end
apply gain ratio algorithm over outcome from Relief-F algorithm

A implementação do algoritmo proposto foi realiza através a utilização do framework Orange. Orange é um ferramente bastante utilizada para realização de análise de dados por meio de scripts na linguagem de programação Python e também na forma de programação visual Demšar et al. (2013).

4

Results and Discussion

Esta seção apresenta o conjunto de dados utilizados no experimento, bem como traz uma avaliação da eficácia de nosso algoritmo propostos. Também apresentaremos uma comparação das métricas de avaliação dos modelos de aprendizagem de máquina sem aplicar o algoritmo proposto e após a aplicação do mesmo, as métricas utilizadas para avaliar os modelos foram acurácia, f1, precisão e sensibilidade. Para realizar a comparação foram utilizados os seguintes modelos de aprendizagem de máquina: (i) Redes Neurais, (ii) Floresta Aleatória (iii) SMV (iv) Regressão Logística (v) Naive Bayes (vi) KNN

4.1 Dataset

Para realização do experimento foi utilizado o data set "Breat Cancer Wiscon (Prognostic) Data Set" (Referência). Neste conjunto de dados cada instância, são 198 no total, representa um paciente atendido pelo Dr. Wolberg, e incluem apenas casos que tratam de câncer de mama invasivo, e nenhuma evidência de metástase à distância no momento do diagnóstico. Além disto, este conta com um conjunto de 34 atributos, e as primeiras 30 características foram extraídas a partir de imagens digitalizadas de uma punção aspirativa por agulha fina de uma massa mamária. A classe do conjunto de dados, representa a recorrência ou não-recorrência do câncer de mama no paciente. Dentre os atributos contidos no conjunto, temos informações relacionadas às características do tumor detectado no paciente tais como, raio, textura, perímetro, área do tumor, etc.

4.2 Pré-Resultados

A seguir serão apresentados os resultados da análise dos dados com e sem a aplicação do algoritmo proposto neste trabalho.

4.2.1 Análise sem aplicação do modelo proposto

	Métricas de Avaliação			
	Acurácia	F1	Precisão	Sensibilidade
KNN	0,696	0,670	0,653	0,696
SVM	0,763	0,678	0,705	0,763
Floresta Aleatória	0,747	0,706	0,697	0,747
Rede Neural	0,763	0,744	0,737	0,763
Naive Bayes	0,629	0,655	0,716	0,629
Regressão Logística	0,768	0,751	0,745	0,768

4.2.2 Análise com a aplicação do modelo proposto

	Métricas de Avaliação			
	Acurácia	F1	Precisão	Sensibilidade
KNN	0,701	0,677	0,662	0,701
SVM	0,778	0,703	0,783	0,778
Floresta Aleatória	0,763	0,721	0,721	0,763
Rede Neural	0,789	0,773	0,770	0,789
Naive Bayes	0,593	0,623	0,707	0,593
Regressão Logística	0,773	0,758	0,752	0,773

5

Final Considerations and Schedule

5.1 Final Considerations

Neste trabalho, propusemos um algoritmo híbrido para seleção de caractéristicas utilizando os algoritmos já existentes na liteturatura Relief-F e Gain Ratio. Inicialmente os pré-resultados obtidos a partir dos primeiros experimentos realizados, demonstram que o modelo proposto apresenta uma melhora na predição de recorrência do câncer de mama em comparação com a aplicação do mesmo conjunto de dados em alguns dos modelos já existentes. Porém, vale ressaltar que foram utilizados apenas dados clínicos do câncer em análise. Seria interessante, aprofundar os experimentos através da utilização de dados moleculares de pacientes diagnósticos com câncer de mama.

5.2 Cronograma

Resultados de tradução A execução dessas atividades seguirá o cronograma de trabalho apresentado na Tabela ref tab: agendamento$_act, quesodescritosaseguir.$

- **Revisão da literatura:** Esta etapa consiste na realização de uma revisão da literatura existente. Este processo, por ser contínuo, tende a durar todo o período do mestrado. ;
- **Algoritmo Híbrido**: Testaremos o algoritmo proposto e trabalharemos em melhorias quem possam ser feitas no mesmo;
- **Experimentos**: Nesta etapa compararemos nosso algoritmo com os algoritmos existentes na literatura;
- **Algoritmos da literatura**: A qualquer momento pode subir uma implementação de um novo algoritmo que utilize os mesmos algoritmos utilizados neste trabalho, caso isto aconteça, devemos comparar a nossa versão do algoritmo com a publicada na literatura;

Table 5.1: Cronograma

Atividades	2020										2021	
	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev
Revisão da literatura									X	X	X	X
Experimentos									X	X	X	X
Algoritmos da literatura										X	X	X
Algoritmo Híbrido									X	X		
Publicação de artigos									X	X	X	X
Escrita da tese										X	X	X
Apresentação da tese									X	X		

	2021										2022	
	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev
Revisão da literatura	X											
Algoritmos da literatura	X											
Experimentos	X											
Publicação de artigos	X											
Apresentação da tese	X											

- **Publicação de artigos**: Publicação de artigos nos meios científicos relevantes, como congressos, periódicos e conferências, perdurará ao longo de todo o desenvolvimento deste trabalho e ocorrerá sempre que forem descobertos resultados relevantes;
- **Escrita da tese**: A escrita da tese será feita de forma incremental, e ocorrerá durante toda duração do trabalho;
- **Apresentação da tese**: Para consolidar os resultados e a tese escrita, este trabalho será apresentado em agosto de 2021.

References

BedfordBreastCancer. 3d mammography. URL https://www.bedfordbreastcenter.com/mammogram-los-angeles/.

Luciano Fernandes Chala and Nestor de Barros. AvaliaÃ§Ãdas mamas com mÃde imagem. *Radiologia Brasileira*, 40:4 – 6, 02 2007. ISSN 0100-3984. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-39842007000100001&nrm=iso.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers Electrical Engineering*, 40(1):16–28, 2014. ISSN 0045-7906. **DOI** https://doi.org/10.1016/j.compeleceng.2013.11.024. URL https://www.sciencedirect.com/science/article/pii/S0045790613003066. 40th-year commemorative issue.

Gary M. Clark, Susan G. Hilsenbeck, Peter M. Ravdin, Michele De Laurentiis, and C. Kent Osborne. Prognostic factors: Rationale and methods of analysis and integration. *Breast Cancer Research and Treatment*, 32(1):105–112, Jan 1994. ISSN 1573-7217. **DOI** 10.1007/BF00666211. URL https://doi.org/10.1007/BF00666211.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. URL http://jmlr.org/papers/v14/demsar13a.html.

Ruth M. Dunne, Ailbhe C. O'Neill, and Clare M. Tempany. Chapter 9 - imaging tools in clinical research: Focus on imaging technologies. In David Robertson and Gordon H. Williams, editors, *Clinical and Translational Science (Second Edition)*, pages 157–179. Academic Press, second edition edition, 2017. ISBN 978-0-12-802101-9. **DOI** https://doi.org/10.1016/B978-0-12-802101-9.00009-0. URL https://www.sciencedirect.com/science/article/pii/B9780128021019000090.

AndrÃGonÃ de Freitas, ClÃ¡udio Kemp, Maria Helena Louveira, Sandra Maria Fujiwara, and Leandro Ferracini Campos. Mamografia digital: perspectiva atual e aplicaÃ§Ãfuturas. *Radiologia Brasileira*, 39:287 – 296, 08 2006. ISSN 0100-3984. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-39842006000400012&nrm=iso.

Ruffo de Freitas JÃ, Rodrigo Disconzi Nunes, Edesio Martins, Maria Paula Curado, Nilceana Maya Aires Freitas, Leonardo Ribeiro Soares, and JosÃ© Carlos Oliveira. Prognostic factors and overall survival of breast cancer in the city of Goiania, Brazil: a population-based study. *Revista do ColÃBrasileiro de CirurgiÃ*, 44:435 – 443, 10 2017. ISSN 0100-6991. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-69912017000500435&nrm=iso.

Leila Posenato Garcia. RevisÃsistemÃ¡tica da literatura e integridade na pesquisa. *Epidemiologia e ServiÃde SaÃ*, 23:7 – 8, 03 2014. ISSN 1679-4974. URL http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742014000100001&nrm=iso.

INCA. Câncer de mama. URL https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama.

Ryoichi Nakahori, Ryuji Takahashi, Momoko Akashi, Kana Tsutsui, Shino Harada, Roka Matsubayashi, Shino Nakagawa, Seiya Momosaki, and Yoshito Akagi. Breast carcinoma originating from a silicone granuloma: A case report. *World Journal of Surgical Oncology*, 13:72, 02 2015. **DOI** 10.1186/s12957-015-0509-6.

Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):23–69, Oct 2003. ISSN 1573-0565. **DOI** 10.1023/A:1025667309714. URL https://doi.org/10.1023/A:1025667309714.

Tim Sherer. Feature selection (data mining) [internet], 2018. https://docs.microsoft.com/en-us/analysis-services/data-mining/feature-selection-data-mining?view=asallproducts-allversions#:~:text=In%20this%20article&text=Feature%20selection%20refers%20to%20the,or%20features0from%20existing%20data/, acessado em 20/04/2021.

Maxwell E. A. Silva, Victor G. L. Holanda, Rodrigo S. Silva, Paulo V. L. Severiano, and Rafael A. Silva. Seleção de características biológicas para prognóstico de câncer: Revisão sistemática da literatura, Dez 2020. URL https://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/817.

Bikesh Kumar Singh. Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning

paradigm. *Biocybernetics and Biomedical Engineering*, 39(2):393–409, 2019. ISSN 0208-5216. **DOI** https://doi.org/10.1016/j.bbe.2019.03.001. URL https://www.sciencedirect.com/science/article/pii/S0208521618304261.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, n/a(n/a). **DOI** https://doi.org/10.3322/caac.21660. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

Luiz Antonio Teixeira and Luiz Alves AraÃ Neto. CÃde mama no Brasil: medicina e saÃpÃno sÃXX. *SaÃe Sociedade*, 29, 00 2020. ISSN 0104-1290. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-12902020000300313&nrm=iso.

Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018. ISSN 1532-0464. **DOI** https://doi.org/10.1016/j.jbi.2018.07.014. URL https://www.sciencedirect.com/science/article/pii/S1532046418301400.

World Cancer Research Fund WCRF. Breast cancer statistics. URL https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics.

World Health Organization WHO. Cancer. URL https://www.who.int/news-room/fact-sheets/detail/cancer.