Jessica Paloma Sousa Cardoso

Classificação de pacientes em saudáveis e não saudáveis em termografia infravermelha dinâmica de mamas usando séries temporais

São Luís - MA

2018

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Sousa Cardoso, Jessica Paloma.
Classificação de pacientes em saudáveis e não saudáveis em termografia infravermelha dinâmica de mamas usando séries temporais / Jessica Paloma Sousa Cardoso. - 2018.
62 p.

Orientador(a): Aristófanes Corrêa Silva.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - MA, 2018.

1. Câncer de Mama. 2. Detecção Precoce. 3. Termografia Infravermelha Dinâmica. I. Corrêa Silva, Aristófanes. II. Título.

Jessica Paloma Sousa Cardoso

Classificação de pacientes em saudáveis e não saudáveis em termografia infravermelha dinâmica de mamas usando séries temporais

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Aristófanes Corrêa Silva

São Luís - MA

2018

Jessica Paloma Sousa Cardoso

Classificação de pacientes em saudáveis e não saudáveis em termografia infravermelha dinâmica de mamas usando séries temporais

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho Aprovado em 23/01/2018

Prof. Dr. Aristófanes Corrêa Silva
Orientador
Universidade Federal do Maranhão

Prof. Dr. João Dallyson Sousa de Almeida
Examinador
Universidade Federal do Maranhão

Prof. Dr. Stelmo Magalhães Barros Netto
Examinador
Universidade Federal do Maranhão

São Luís - MA
2018

Aos meus meus pais e meu padrasto.

Agradecimentos

Inicialmente, agradeço aos meus pais e meu padrasto, por todo incentivo, motivação e educação que me proporcionaram. Um agradecimento especial ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) pelo auxílio recebido.

Quero também agradecer aos colegas pela ajuda, apoio e companheirismo prestados nessa etapa da minha vida, em especial aos meus colegas *Ivan*, Thales Levi e Flávio pela amizade.

Ao meu orientador Prof. Dr. Aristófanes Corrêa Silva por toda a paciência e apoio que me ofereceu. Também agradeço aos meus colegas de laboratório pela ajuda oferecida nos momentos de dúvidas.

Por fim, agradeço aos professores, funcionários, colegas da Universidade Federal do Maranhão (UFMA) que contribuíram direta e indiretamente para a minha formação.

"Não há saber mais ou saber menos: Há saberes diferentes."
(Paulo Freire)

Resumo

O câncer de mama é o segundo tipo de câncer mais comum entre as mulheres no país e no mundo, além de estar associado a maior taxa de mortalidade feminina. No Maranhão, segundo o INCA, esse é o segundo tipo de câncer mais presente na população. A detecção pode ser realizada através de exames de imagens, tais como a mamografia, ressonância magnética, termografia, etc. Uma das características dos cânceres está no fato que as suas células tumorais necessitam de nutrientes que são fornecidos através da corrente sanguínea alterando o fluxo de sangue. Assim, a termografia infravermelha consegue detectar doenças que afetam o fluxo sanguíneo. As imagens termográficas dinâmicas foram adquiridas do Hospital Universitário Antônio Pedro da Universidade Federal Fluminense (HUAPE-UFF), constituindo 70 exames, sendo 35 saudáveis e 35 doentes. As imagens termográficas são convertidas para escala de cinza conforme a sua temperatura. Temperaturas que possuem valores mais altos irão ser representadas com valores de maior intensidade na escala de cinza e valores baixos de corresponderão a menores intensidades. Sobre as imagens em escala de cinza, é aplicado o registro deformável. A região das mamas é extraída e aplicado um janelamento, no qual para cada região da janela é extraída uma série temporal. O resultado consiste em um conjunto de séries temporais as quais são concatenadas gerando uma super-série. Sobre essa super-série, são extraídos *motifs* e *discords* e um conjunto de características sobre eles, a classificação dessas características foi realizada por meio do *Support Vector Machine*. Através dessa modelagem, a metodologia proposta conseguiu alcançar 75% de acurácia.

Palavras-chaves: Termografia Infravermelha Dinâmica, Câncer de Mama, Detecção Precoce.

Abstract

Breast cancer is the second most common type of cancer among women in the country and the world, in addition to being associated with a higher rate of female mortality. In Maranhão, according to INCA, this is the second most common type of cancer in the population. Detection can be performed through imaging tests, such as mammography, MRI, thermography, etc. One of the characteristics present in cancers is in the fact that their tumor cells need nutrients that are supplied through the bloodstream by altering the flow of blood. Thus, infrared thermography can detect diseases that affect physiological or anatomical parameters of blood supply. The thermographic images were acquired from the Antônio Pedro University Hospital of the Federal Fluminense University (HUAPE-UFF), comprising 70 exams, 35 healthy and 35 unhealthy. The thermographic images are converted to grayscale according to their temperature. Temperatures that have higher values will be represented with values of greater intensity in the gray scale and low values will correspond to lower intensities. On grayscale images, the deformable register is applied. The region of the breasts is extracted and a windowing is applied, in which for each region a time series is extracted and the result consists of a set of time series which are concatenated generating a super-series. On this super-series, motifs and discords are extracted and a set of features on them, the classification of these characteristics was performed by means of the Support Vector Machine. The proposed methodology was able to reach 75 % accuracy.
Keywords: Dynamic Infrared Thermography, Breast Cancer, Early Detection.

Lista de ilustrações

Figura 1 – Comparação entre pacientes saudável e doente na termografia. 16
Figura 2 – Séries temporais extraídas da TID 17
Figura 3 – Ciclo celular 24
Figura 4 – Crescimento tumoral de uma célula 25
Figura 5 – Exame de mamografia 27
Figura 6 – Ilustração do exame de ultrassom das mamas. 28
Figura 7 – Ilustração do exame de ressonância magnética sendo aplicada em uma paciente. 28
Figura 8 – Comparação entre termografias. 30
Figura 9 – Aplicação de negativo sobre uma imagem. 32
Figura 10 – Filtro da média. 32
Figura 11 – Aplicação de registro de imagens. 33
Figura 12 – Registro baseado em intensidade 34
Figura 13 – Ilustração de uma série temporal extraída de uma paciente. 35
Figura 14 – Pré-processamento de uma série temporal para remoção de ruídos . . . 37
Figura 15 – Normalização de séries temporais 38
Figura 16 – Representação de séries temporais 39
Figura 17 – Exemplo de série temporal contendo um *motif* com três ocorrências. . 41
Figura 18 – Separação não linear no SVM. 45
Figura 19 – Fluxo da metodologia para classificação das TID das mamas. 47
Figura 20 – Aquisição das termografias 48
Figura 21 – Resultado da aplicação do registro de imagens 49
Figura 22 – Delimitação da região de interesse. 49
Figura 23 – Ilustração da construção das séries temporais para um exame. 50
Figura 24 – Ilustração da construção das séries reduzidas a partir de uma super-série. 51
Figura 25 – Extração de características de uma série temporal. 52
Figura 26 – Ilustração do *Matrix Profile* 56

Lista de tabelas

Tabela 1 – Matriz de confusão 45
Tabela 2 – Resultados obtidos pela metodologia sem *motifs* e *discords*. 54
Tabela 3 – Resultados obtidos pela metodologia considerando apenas os *discords*. . 54
Tabela 4 – Resultados obtidos pela metodologia com *discords* e *motifs*. 55

Lista de abreviaturas e siglas

AC	Acurácia
BIRADS	*Breast Imaging Reporting and Data System*
CAD	*Computer-Aided Diagnosis*
DFT	*Discrete Fourier Transform*
ES	Especificidade
FFT	*Fast Fourier Transform*
FN	Falso Negativo
FP	Falso Positivo
INCA	Instituto Nacional do Câncer
MP	*Matrix Profile*
MPI	*Matrix Profile Index*
MST	Mineração de Series Temporais
PR	Precisão
ROI	*Region Of Interest*
SVM	*Support Vector Machine*
TID	Termografia Infravermelha Dinâmica
TIE	Termografia Infravermelha Estática
VN	Verdadeiro Positivo
VP	Verdadeiro Positivo

Sumário

1 INTRODUÇÃO 14
1.1 Justificativa 15
1.2 Objetivos 17
1.2.1 Objetivo Geral 18
1.2.2 Objetivos Específicos 18
1.3 Organização do trabalho 18

2 TRABALHOS RELACIONADOS 20
2.1 Termografia 20
2.2 Séries temporais 21

3 FUNDAMENTAÇÃO TEÓRICA 23
3.1 O câncer 23
3.1.1 Câncer de mama 24
3.1.2 Exames para o diagnóstico de patologias nas mamas 25
3.1.2.1 Mamografia 26
3.1.2.2 Ultrassom das mamas 27
3.1.2.3 Ressonância magnética 28
3.2 Termografia 29
3.3 Processamento de imagens 30
3.4 Série temporal 35
3.4.1 Pré-processamento 36
3.4.2 Similaridade 39
3.4.3 Transformada discreta de fourier *(Discrete Fourier Transform - DFT)* . . . 39
3.4.4 Descobertas de *motifs* 41
3.4.5 Detecção de anomalia 42
3.4.6 *Matrix Profile* 42
3.5 *Support Vector Machine* 44
3.6 Métricas para avaliar a classificação 45

4 METODOLOGIA PROPOSTA 47
4.1 Aquisição de imagens 47
4.2 Pré-processamento 48
4.3 Construção das séries temporais 49
4.4 Extração de características 51
4.5 Classificação 52

5 RESULTADOS E DISCUSSÃO 53
5.1 Extração de características sobre o conjunto 53
5.2 Extração de características sobre o *motifs* e *discords* 54

6 CONCLUSÃO 57

REFERÊNCIAS 58

1 Introdução

Desde a descoberta experimental do raio-x por Wilhelm Röntgen em 1895 (NUNES, 2006), a medicina diagnóstica sofreu uma grande revolução. E essa técnica possibilitou a visualização das estruturas internas do corpo humano sem a necessidade de qualquer tipo de incisão. Além disso, permitiu a descoberta de diversas outras técnicas de imagens como ultra-sonografia, tomografia e ressonância magnética. Nesse sentido, avanços no processo de aquisição e análise dessas imagens vieram com o objetivo de aperfeiçoar no processo de tratamento e diagnóstico de diversas patologias. Essas imagens contém um conjunto de informações agregadas que podem ser analisadas através de sistemas computacionais denominados *Computer-Aided Diagnosis* (CAD). Esses sistemas analisam esses dados, evidenciando informações importantes que podem ser úteis no diagnóstico de doenças, dando subsídios aos especialistas.

Várias técnicas e métodos computacionais têm sido propostos e aplicados com o objetivo de evidenciar as informações contidas nessas imagens. Dessa forma, técnicas de processamento de imagens, com suporte das áreas de aprendizado de máquina e reconhecimento de padrões, geralmente estão presentes no processo de análise e classificação desses dados. Uma aplicação prática dessas técnicas é no auxílio ao diagnóstico de doenças, como por exemplo, detecção de nódulos em mamografia ou ainda realizar a classificação desses nódulos em benigno ou maligno.

Para o biênio de 2016-2017, foram estimados 600 mil novos casos de cânceres, com os cânceres de pele não melanoma, próstata, pulmão, cólon, mama, colo do útero e glândula da tireoide os que possuem maiores incidências (SILVESTRE, 2016). No Maranhão, o câncer de mama é o segundo tipo de câncer que mais acomete a população feminina (INCA, 2016a). Estima-se que o Maranhão seja um dos estados mais afetados por esta patologia. Isso provavelmente se deve ao fato do estado ser um dos mais pobres da federação, tendo como consequência poucos locais para efetuar a detecção e diagnóstico dessa doença. Logo, metodologias computacionais, que apresentam resultados rápidos e totalmente automáticos, podem auxiliar e facilitar a tarefa de médicos onde os recursos ainda são reduzidos.

Apesar do câncer de mama ser o câncer que possui a maior taxa de mortalidade feminina no mundo, as chances de sobrevida da paciente aumentam caso esta doença seja detectada em seus estágios iniciais (INCA, 2017). Assim, a detecção nos estágios iniciais é importante para a eficiência do tratamento. O diagnóstico precoce e o rastreamento são estratégias para a identificação do câncer de mama. No diagnóstico precoce, exames são solicitados em consultas rotineiras. Dentre esses exames, estão os de imagens como a mamografia, que é recomendado a mulheres a partir dos 40 anos, junto ao exame clínico

com periodicidade anual. Além desse exame há outros, como a ultrassom e ressonância e a termografia que servem de suporte à mamografia.

Diante desse cenário, houve um aumento nos interesses dos pesquisadores na área de ciência da computação em desenvolver métodos e técnicas que auxiliem nesse processo de prevenção do câncer de mama. Técnicas essas que visam extrair tais informações e exibi-las de forma legível à compreensão humana. Assim, este trabalho tem como objetivo utilizar técnicas computacionais com o propósito de classificar exames de termografias em com anomalia e sem anomalia.

1.1 Justificativa

A mamografia é um exame de diagnóstico por imagem que faz uso de radiação. Esse exame é bastante eficaz no processo de detecção de anomalias no tecido das mamas, tais como a existência de nódulos, microcalcificações, dentre outros tipos de lesões que ainda não são perceptíveis através do toque. Logo, devido a sua eficiência em detectar o câncer de mama em estágios iniciais, a mamografia é o mais indicado para o diagnóstico dessa patologia. Entretanto, segundo o INCA (INCA, 2016b), apesar da eficácia na detecção do câncer no começo, o exame não é indicado em mamas densas[1] pois a identificação de lesões fica prejudicada por se assemelhar bastante as regiões de não massas.

O tecido mamário passa a ser substituído por tecido adiposo até a mulher alcançar a menopausa. E isso faz com que a densidade das mamas diminua, permitindo melhor visualização de lesões na imagem. No entanto, mulheres jovens não estão imunes ao câncer mamário. Por isso, outras técnicas de diagnóstico são propostas, tais como a ressonância magnética e mamografia por emissão de pósitrons, que não são afetadas pelo problema de densidade da mama. Entretanto, esses exames são de alto custo e, diante disso, a termografia é mostrada como uma boa opção de detecção por apresentar um custo baixo se comparado aos demais exames (BORCHARTT, 2013).

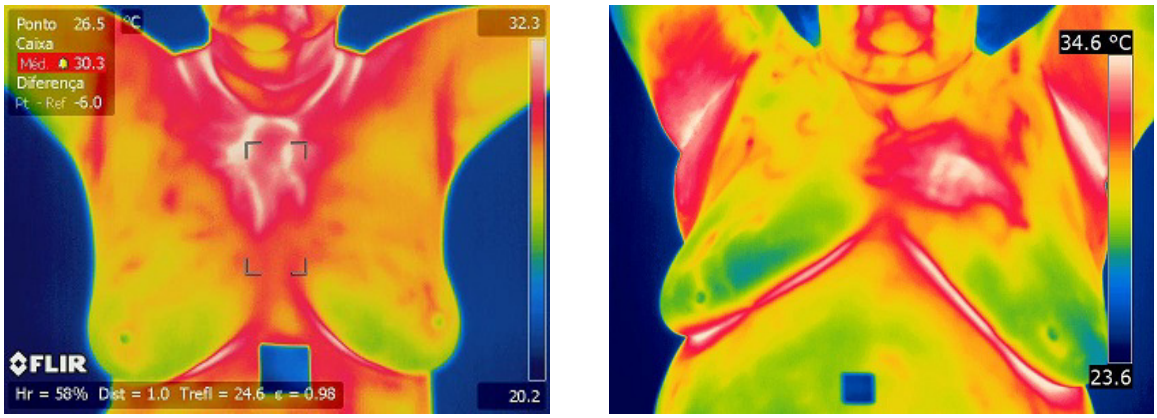
A termografia infravermelha é um exame que não oferece quaisquer riscos ou danos a pacientes, visto que não faz uso de qualquer método invasivo como radiações ou acessos venosos. Esse exame registra as temperaturas emitidas pela superfície corpórea, permitindo observar padrões sobre a distribuição de temperatura.

Tumores geralmente necessitam de um aumento no suprimento de nutrientes para se desenvolverem, como consequência naquela região geralmente há um aumento nos gradientes de temperatura se comparada a um tecido normal (WOLFF et al., 2012). Assim, a detecção de câncer de mama através da termografia demonstra caráter promissor pela sua capacidade em gerar imagens contendo informações sobre a distribuição de temperatura.

[1] Mama que possui grande quantidade de tecido glandular, responsável pela produção de leite(BOYD et al., 2011)

A Figura 1 ilustra uma comparação entre um paciente saudável (a) e um paciente que apresenta câncer de mama (b). No primeiro caso, a variação nos gradientes de temperatura ocorre de forma mais suave, além da similaridade presente entre a região esquerda e direita, preservando um aspecto mais homogêneo. Em contrapartida, em (b) é mostrado um caso de carcinoma ductal infiltrante localizado na mama esquerda, o qual provoca uma desordem nos níveis de temperaturas locais.

Figura 1 – Comparação de termografias entre dois casos, sendo um sadio (a) e outro com câncer (b).



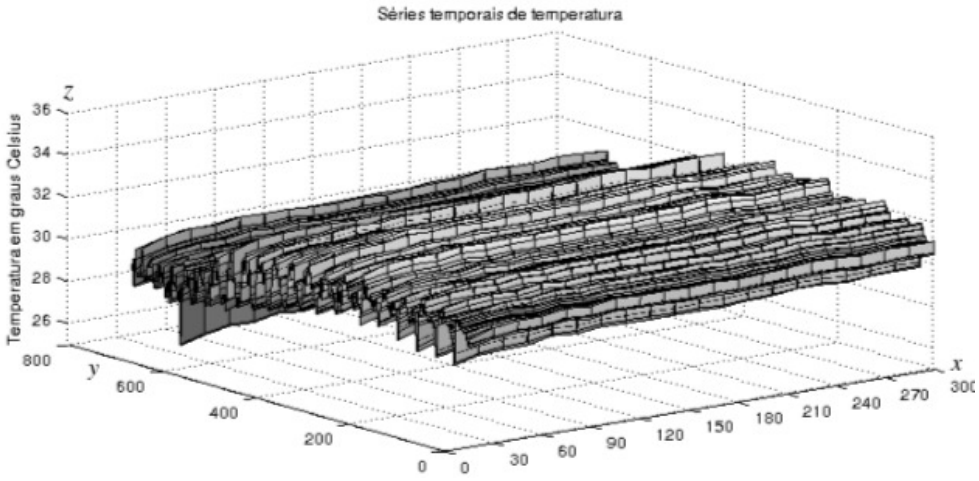
(a) (b)

Nesse contexto, a termografia é capaz de identificar anomalias que provocam pertubações nos níveis de temperatura. Todavia, o processo de análise dessas imagens é uma tarefa exaustiva e repetitiva, que requer muita atenção por parte dos especialistas. Além disso, regiões lesionadas podem apresentar temperaturas semelhantes à regiões sadias, o que pode dificultar a análise. Além da termografia estática apresentada na Figura 1, existe a termografia dinâmica que realiza a captura de uma sequência de imagens com um intervalo de tempo estipulado entre elas de acordo com algum protocolo de aquisição.

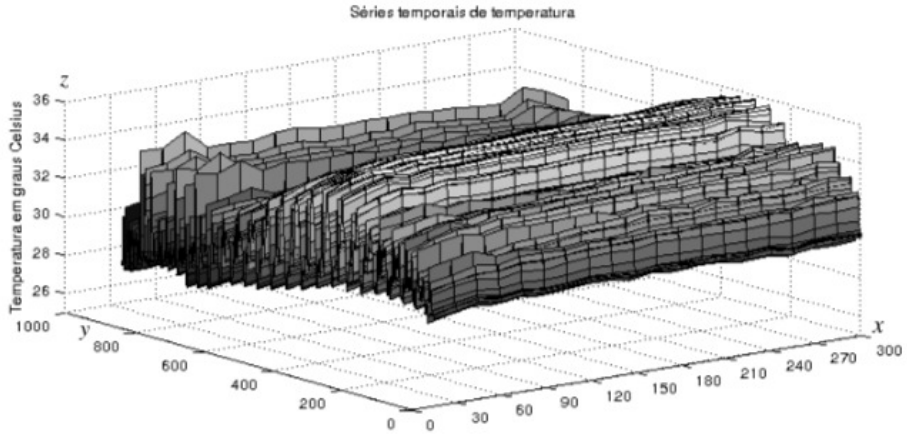
A termografia infravermelha estática pode levar a falsos positivos em imagens que contenham a presença de vascularizações complexas, entretanto, esses padrões desaparecem após sofrerem um estresse térmico e retornam gradualmente, esse retorno pode ser melhor observado em imagens sequenciais (OHASHI; UCHIDA, 2000). Baseado na premissa que regiões com processo de formação de tumor apresentam maior variação de temperatura e que a recuperação dos vasos acontecem de forma gradual, decidiu-se fazer o uso da Termografia Infravermelha Dinâmica (TID) com a modelagem do problema usando a ideia de séries temporais. Visto que os métodos de análise de séries temporais buscam compreender estruturas fundamentais que possam servir para descrever esses tipos de dados sequenciais, tais como tendência, padrões, previsões, dentre outras. Além disso, regiões sadias apresentam um comportamento mais homogêneo (Figura 2a) se comparada a tecidos com anomalia (Figura 2b). Onde o eixo x representa o tempo em segundos, y o

número da série da série temporal e z a temperatura.

Figura 2 – Representação da TID como Série Temporal de dois pacientes distintos.



(a) Saudável



(b) Doente

Fonte – (SILVA, 2015a)

Para realização deste trabalho, é indispensável estudar e utilizar técnicas de processamento de imagens e reconhecimento de padrões. Pretende-se criar um modelo de classificação que realize a distinção entre pacientes saudáveis e não saudáveis.

1.2 Objetivos

Diante da justificativa acima do uso da termografia dinâmica como exame para diagnóstico do câncer mama, os objetivos deste trabalho são descritos a seguir:

1.2.1 Objetivo Geral

Propor uma metodologia para distinguir exames de pacientes em saudáveis e não saudáveis em imagens de termografia infravermelha dinâmica, realizando a modelagem do problema como série temporal.

1.2.2 Objetivos Específicos

Para atingir o objetivo principal deste trabalho de discriminar exames em sadios e com anomalias, faz-se necessário atingir os seguintes objetivos específicos:

- Realizar o levantamento bibliográfico de trabalhos que utilizam a termografia para diagnósticos de patologias na medicina;
- Compreender como a TID pode ser útil no processo de classificar se uma paciente é saudável ou não;
- Estudar conceitos fundamentais de séries temporais, tais como sazonalidade, tendência e ciclos, e tentar fazer uma relação com a problemática do trabalho;
- Investigar trabalhos que utilizam séries temporais com o propósito de classificação ou detecção de elementos que possam servir para descrever essa série;
- Realizar a modelagem do problema como série temporal e geração do modelo de classificação.

1.3 Organização do trabalho

A organização desse trabalho está dividida em cinco capítulos. São estes:

Capítulo 2 — Trabalhos Relacionados capítulo apresenta e discuti os trabalhos presentes na literatura. Esses trabalhos são divididos em duas seções. A primeira seção descreve técnicas aplicadas a exames de imagens térmicas para diagnóstico do câncer de mama, e a segunda seção discute técnicas existentes na literatura voltadas para Mineração de Séries Temporais (MST).

Capítulo 3 — Fundamentação Teórica neste capítulo, são abordados os conceitos necessários para a compreensão desse trabalho. Nele é explicado os conceitos sobre o câncer e como se dá o processo de carcinogênese. São abordados os principais exames de imagens (ultrassom, ressonância magnética e mamografia) utilizados para o prognóstico dessa doença, além da termografia que é o exame proposto nesse trabalho. Também são abordados conceitos sobre séries temporais e técnicas de MST.

Capítulo 4 — Metodologia Proposta neste capítulo são descritos a construção da base, bem como o protocolo de aquisição dos exames utilizados neste trabalho. Além disso, são descritas cada uma das etapas da metodologia proposta, desde o pré-processamento dos exames até a etapa de classificação das séries temporais extraídas dos mesmos.

Capítulo 5 — Resultados e Discussão neste capítulo, são apresentados os resultados obtidos com a aplicação da metodologia proposta sobre a base de exames apresentadas.

Capítulo 6 — Conclusão o último capítulo deste trabalho apresenta as conclusões e considerações finais sobre o estudo proposto, além de descrever possíveis trabalhos futuros.

2 Trabalhos Relacionados

Neste capítulo são discutidos alguns trabalhos encontrados na literatura que serviram de suporte para o desenvolvimento deste trabalho. Na Seção 2.1, são apresentados trabalhos voltados para termografias e, na Seção 2.2, são apresentados os trabalhos voltados para séries temporais.

2.1 Termografia

Nesta seção são apresentados alguns trabalhos encontrados na literatura que fizeram uso da termografia com o objetivo de detectar malignidades nas mamas. Um deles foi o trabalho de Sterns et al. (1996), que utilizaram a termografia para traçar uma relação entre sobrevivência e proliferação da doença em pacientes que possuíam diagnóstico de carcinoma ductal invasivo[1]. Em seus estudos, chegaram a conclusão que o termograma com anormalidades está associado ao tumor, mas não a taxa de proliferação ou densidade dos microvasos.

Ohashi e Uchida (2000) propuseram um protocolo de aquisição da TID. Nesse trabalho, o processo de classificação de pacientes foi realizado considerando um conjunto de regras com o objetivo de definir se um paciente apresenta ou não anormalidades nas mamas. Os critérios definidos levam em consideração a assimilaridade de pontos quentes, anormalidade assimétrica nos padrões vasculares e diferença significante de temperatura. Em seu trabalho, os autores coletaram exames de 728 pacientes no período de 1989-1994. Esses exames correspondem a uma sequência de termogramas obtidas na posição frontal, durante um período de 20 minutos. Após estarem de posse das imagens, realizaram a análise sobre o conjunto de regras definidas por eles, conseguindo uma boa taxa de verdadeiro positivos que superaram 80%. Em contrapartida os falsos negativos passaram de 40%.

Tan et al. (2007) propuseram uma metodologia baseada em redes neurais *fuzzy* para analisar os termogramas. Foram coletadas três termogramas por paciente, um frontal e dois laterais. Sua base possuía termogramas de 78 pacientes, nos quais 28 se configuravam saudáveis, 43 apresentavam tumores benignos e 7 pacientes apresentavam câncer. Nessa pesquisa, os autores antigiram uma acurácia de 93%.

Resmini et al. (2012) realizaram um trabalho que tinha como objetivo classificar as imagens de pacientes saudáveis ou com portadoras de alguma patologia da mama. Foram extraídas 712 características extraídas, sendo elas baseadas em estatísticas simples,

[1] Carcinoma Ductal Invasivo - é iniciado no duto mamário, mas não se limita ao duto crescendo também na parte do tecido adiposo da mama (ONCOGUIA, 2017c)

geometria fractal e características de fundamentação geoestatística. Os autores dividiram o trabalho em três abordagens de acordo com a construção da base: - na primeira, o qual foi denominada "base de dados antiga", era composta por 28 pacientes (24 sadias e 4 doentes); na segunda, chamada de "base nova", foi realizada o acréscimo de 6 pacientes com patologias e 10 sadias na base antiga; e a última abordagem, nomeada "base de subtração", corresponde a subtração das características das mamas esquerda e direita da base aumentada. Para a classificação de pacientes em doentes e sadias foram utilizados três classificadores: IbK (kNN), Naïve Bayes e *Suport Vector Machine* (SVM). O melhor resultado atingindo pelos autores foi com SVM sobre a base nova, o qual atingiu uma acurácia de 82,35%.

Ainda nesse trabalho, os autores realizaram uma seleção de atributos através do *Principal Component Analysis* (PCA) sobre a a base nova e subtraída. O classificador SVM se mostrou superior aos demais ao atingir uma acurácia de 88,23% em ambos os testes e uma área abaixo da curva ROC de 0,800.

2.2 Séries temporais

A área de mineração de dados é uma área que lida com um grande volume de dados, cujo objetivo está pautado na busca por padrões, relações e anomalias que sirvam no processo de tomada de decisões e conhecimento desses dados. Diversas pesquisas têm sido desenvolvidas no meio de séries temporais com o propósito de encontrar elementos que tornem possível a descrição desses dados (MALETZKE, 2009).

Keogh et al. (2001) propuseram um método chamado *Piecewise Aggregate Approximation* que realizava a redução de dimensionalidade da série de um espaço n para m. Essa técnica foi utilizada no trabalho de Lin et al. (2003), que desenvolveu um algoritmo que realizava transformação da séries temporais em símbolos chamada de *Symbolic Aggregate approXimation* (SAX).

Chiu, Keogh e Lonardi (2003) propuseram uma metodologia baseada no algoritmo SAX, onde cada cadeia de símbolos gerada representava um padrão. O objetivo era encontrar os conjuntos de padrões similares que apresentavam o maior número de correspondências, também conhecido como *Motif*. Também houveram outros trabalhos na literatura que buscavam por padrões similares, como Berndt e Clifford (1994), Keogh e Smyth (1997) e Perng et al. (2000).

Em Keogh, Lin e Fu (2004) e Wei, Keogh e Xi (2006), o objetivo foi desenvolver técnicas que buscavam pela subsequência mais dissimilar da série temporal, no qual em ambos trabalhos é chamado de *Discord*. Izakian e Pedrycz (2013) aplicaram o *fuzzy c-means* para buscar anomalias em amplitude e coeficiente de autocorrelação para encontrar anomalias na forma da série.

Maletzke et al. (2013) combinaram características de estatística descritiva e *motifs* para classificar séries temporais de eletrocardiograma de três bases distintas (DB1, DB2 e DB3) usando o classificador j48. A proposta apresentou baixo erro médio nas três bases sendo: 9,07; 3,66; 3,44. Comparada ao *Naïve Method* que atingiu nas três bases, respectivamente os valores de 26,05; 4,30 e 4,99.

Em pesquisas mais recentes, Silva (2015a) e Silva (2015b), a termografia infravermelha dinâmica tem sido abordada sob a ótica de séries temporais. Nesses trabalhos, foi demonstrada uma fronteira visível das características entre pacientes sadias e doentes. O trabalho aqui proposto seguirá a abordagem da modelagem do problema como série temporal, visto que esses trabalhos demonstraram bons resultados sob essa modelagem.

3 Fundamentação Teórica

Neste capítulo são abordados conceitos importantes para a compreensão da metodologia proposta neste trabalho. São abordados conceitos referentes a oncologia, com ênfase ao câncer de mama (métodos de prevenção, diagnóstico e surgimento), métodos de processamento de imagens e conceitos básicos de séries temporais.

3.1 O câncer

Registros dessa doença são datados desde os antigos egípcios e civilizações subsequentes. Até meados do século XIX haviam poucos casos de câncer, pois muitas pessoas morriam de doenças comuns da infância e doenças infecciosas. Com o aumento da expectativa de vida nos últimos anos, mais pessoas tem sobrevivido até a "idade do câncer", tendo por consequência um aumento no número de casos (FRANKS, 1990).

O câncer é o nome dado a um conjunto de mais de 200 doenças que tem como característica principal a produção desordenada de células (ONCOGUIA, 2017b). O processo responsável pelo aumento de células no organismo é denominado de ciclo celular, que tem como objetivo gerar um par de células-filhas com mesmo material genético aos da célula-mãe. Por sua vez esse ciclo, ilustrado na Figura 3, é dividido em cinco fases (FRANKS, 1990) (PIARDI et al., 2016):

G0: também denominado fase de repouso do ciclo, corresponde a fase em que as células não realizam suas divisões. Por exemplo, as células como os neurônios permanecem nessa fase até o momento em que ocorra um estímulo externo.

G1: nessa fase há a síntese de proteínas, produção de RNA [1] e consequentemente aumento celular;

S: corresponde a fase de síntese de DNA[2], tendo como consequência uma cópia do material genético da célula precursora;

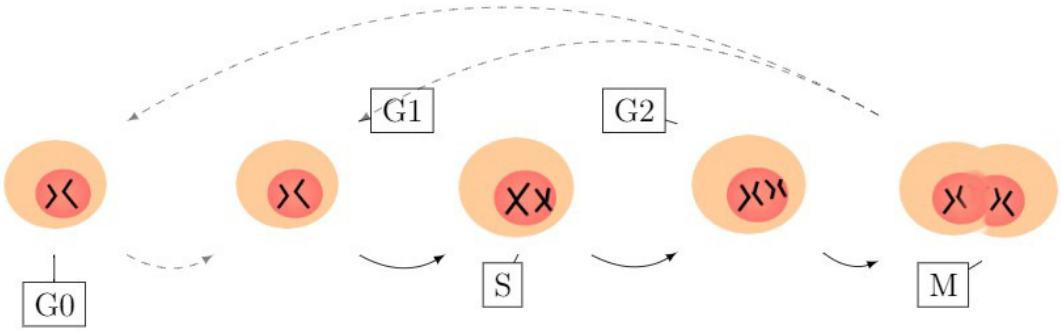
G2: nessa fase, é decidido se a célula passará por um processo de divisão celular, também conhecido como mitose.

[1] As moléculas de RNA estão envolvidas na síntese de proteínas e às vezes na transmissão de informações genéticas (MANDAL, 2016).

[2] O DNA é definido como um ácido nucleico que contém as instruções genéticas utilizadas no desenvolvimento e funcionamento de todos os organismos vivos conhecidos (MANDAL, 2016).

M: por fim, nessa fase ocorre a mitose, que gera dois grupos idênticos a célula-mãe. Essas novas células podem recomeçar a partir da fase G1 ou se tornarem adormecidas ao entrar em G0.

Figura 3 – Ilustração da fases do ciclo celular.



Fonte – Acervo da autora

A característica do tumor está no fato deste apresentar crescimento anormal, ou seja, a célula tumoral diferencia-se da normal pelo fato de não possuir mais o controle de crescimento normal. Os tumores podem ser divididos em três grupos: (1) Tumores benignos - não se espalham para lugares distantes e podem surgir em qualquer tecido; (2) Tumores *in situ* - são similares às células tumorais, mas estão restritos ao epitélio; e (3) Cânceres, que são tumores malignos com capacidade de se espalharem. Geralmente as células tumorais necessitam de nutrientes. Essas células produzem o fator de angiogênese tumoral que realiza a criação de novos vasos ao redor do tumor. Os vasos servem para levar nutrientes para essa região, entretanto, podem transportar células cancerígenas para outros órgãos, gerando metástases (FRANKS, 1990).

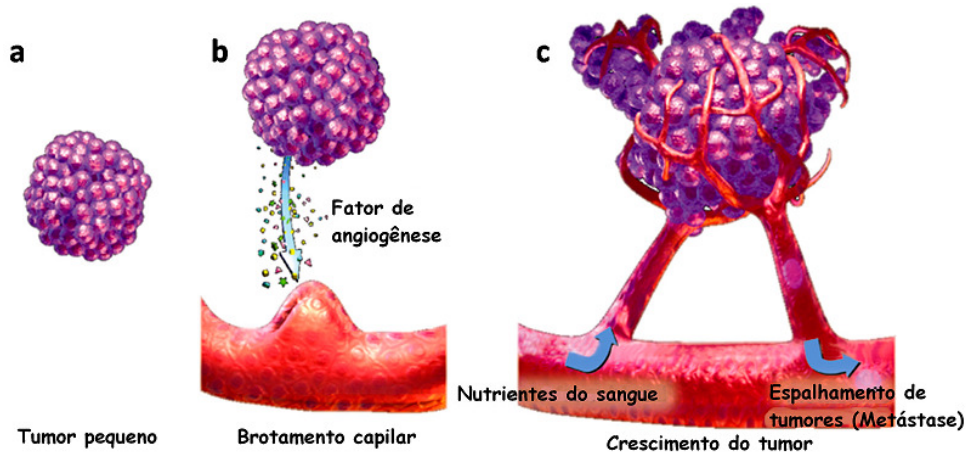
A Figura 4 ilustra o processo de formação de vasos ao redor do tumor, o qual é ativado através do fator de angiogênese tumoral. Percebe-se que além do transporte de nutrientes, esses vasos podem espalhar as células tumorais através da corrente sanguínea.

3.1.1 Câncer de mama

É a produção anormal de células localizadas na região do tecido mamário que tem como consequência a produção de um tumor. Há diversos tipos de câncer de mama, os mais comuns são: *Carcinoma Ductal In Situ* - é um câncer de mama não-invasivo; *Carcinoma Lobular In Situ* - não se desenvolve através das paredes dos lóbulos[3]; *Carcinoma Ductal Invasivo* - é iniciado no duto mamário, podendo surgir também na parte do tecido adiposo

[3] Os lobos são glândulas que possuem pequenos lóbulos que tem como finalidade a produção de leite (ONCOGUIA, 2017a) (BORCHARTT, 2013).

Figura 4 – (a) tumor em estágio inicial; (b) processo de formação nos vasos através do fator de angiogênese; (c) vasos formados ao redor do tumor.



Fonte – Adaptado de (LOIZZI et al., 2017)

da mama; e *Carcinoma Lobular Invasivo* - similar ao Ductal Invasivo, mas com risco de metástase (ONCOGUIA, 2017c).

Quando detectado precocemente, as chances de cura do câncer de mama aumentam substancialmente. Por essa razão, é necessário que as mulheres estejam atentas à saúde de suas mamas, através de consultas e exames regulares. Na seção a seguir, são descritos os exames utilizados pelos especialistas para o prognóstico do câncer de mama.

3.1.2 Exames para o diagnóstico de patologias nas mamas

Um dos exames mais simples que pode ser realizado pelas mulheres para a detecção de patologias nas mamas é o autoexame. Este exame consiste no apalpamento e observação das mamas em busca de mudanças, como saída de secreção nos mamilos, nódulos mamários ou quaisquer alterações na textura ou aspecto das mamas. Caso seja encontrado algo anômalo, deve-se procurar ajuda profissional (INCA, 2017). Outro exame é o exame clínico, que também é baseado na palpação das mamas, mas dessa vez realizado pelo ginecologista/mastologista. Por não ser um exame muito preciso, geralmente o médico solicita exames através de imagens (SEDICIAS, 2017).

Existem dois tipos de exames realizados através de imagens que servem para diagnóstico de doenças mamárias: (1) estruturais, que permitem a visualização interna das estruturas mamárias, e (2) funcionais, que mostram a visão detalhada de funcionamento dos órgãos, bem como o fluxo dos líquidos (BORCHARTT, 2013). Ao primeiro grupo, pertencem a mamografia, ultrassom e ressonância magnética. Já ao segundo grupo pertencem a ultrassom, a ressonância com contraste e a termografia.

O Relatório de Imagens de Mamas e Sistema de Dados (do inglês BIRADS - *Breast*

Imaging Reporting and Data System) é uma padronização internacional utilizadas com o objetivo de avaliar os achados mamográficos de acordo com determinados critérios estabelecidos pelo mesmo. Essa padronização está presente em diversos tipos de exames de imagens (ANDRADE, 2015), e classifica as alterações encontradas de acordo com uma escala de valores entre 0 a 6 (PINHEIRO, 2017):

BIRADS-0: nessa categoria de classificação, o profissional de saúde não pôde realizar uma avaliação completa, seja devido a movimentação da paciente no processo de aquisição ou mesmo dúvida a respeito de um diagnóstico considerado inconclusivo, o que torna indispensável a solicitação de exames adicionais.

BIRADS-1: Significa que é um exame com resultado normal ou negativo para malignidade, ou seja, não foram encontradas lesões ou alterações nas mamas;

BIRADS-2: existe uma alteração nas mamas, mas devido as suas características é classificada como benigna. Como exemplos, tem-se os fibroadenomas calcificados, calcificações vasculares, implantes de silicone e cicatriz cirúrgica;

BIRADS-3: há presença de alterações no exame, provavelmente benignas. Neste caso, é recomendado um controle semestral durante um período. Caso a anomalia permaneça igual, ela passa a ser considerada BIRADS-2;

BIRADS-4: quando o médico classifica nesta categoria, significa que foi encontrada uma lesão suspeita de câncer e é necessário realizar uma biópsia para descartar ou confirmar o diagnóstico de câncer;

BIRADS-5: lesão com suspeitas acima de 95% de chance de ser câncer de mama foi encontrada. É necessária a realização da biópsia;

BIRADS-6: é utilizada em pacientes que já possuam diagnóstico de câncer e estejam em tratamento oncológico;

3.1.2.1 Mamografia

A mamografia é uma modalidade de exame que produz imagens das estruturas internas das mamas por meio de raio-x, permitindo a detecção de calcificações, cistos e pequenos nódulos que não são perceptíveis através da palpação. Ele é indicado para rastreio do câncer de mama geralmente em mulheres a partir dos 40 anos, exceto se houverem casos na família de parentes de primeiro grau, em que o acompanhamento deverá ocorrer mais cedo (SERPEJANTE, 2014). As imagens são capturadas de ambas as mamas em dois ângulos diferentes e o processo é realizado após a compressão das mamas entre duas placas, como ilustrado na Figura 5.

Figura 5 – Ilustração do exame de mamografia.

Fonte – (VILLELA, 2014)

3.1.2.2 Ultrassom das mamas

É um exame auxiliar utilizado para analisar as alterações presente nas mamas, bem como a identificação de alguma anormalidade em mamas densas que não puderam ser visualizadas na mamografia. Este exame também permite a distinção entre massas sólidas e cistos com líquidos. Diferente de outros exames de imagens, não há o emprego de nenhum tipo de radiação ionizante, apenas fazendo uso do som para formação da imagem. Para a realização desse exame a paciente é deitada em posição dorsal, com as mãos na nuca e nas mamas é aplicado um gel que serve para auxiliar no processo. O transdutor (aparelho que emite ondas sonoras nas mamas) é deslizado sobre as regiões mamárias e as imagens são desenhadas na tela em tempo real (ONCOGUIA, 2017d), conforme ilustrado na Figura 6.

Figura 6 – Ilustração do exame de ultrassom das mamas.

Fonte – (HOLDORF, 2017)

3.1.2.3 Ressonância magnética

É um exame indicado a pacientes com mamas densas, que possuem alto risco de terem câncer de mama ou que estejam fazendo tratamento oncológico. O exame de ressonância magnética, ilustrado na Figura 7, possui sensibilidade acima de 95%, o que significa que é capaz de encontrar uma quantidade variada de anomalias nas mamas. No entanto, nem todas as anomalias encontradas são malignas e, também, seu uso não é recomendado como exame de rastreio, como a mamografia. Para sua realização, é necessário que seja aplicada uma injeção contendo contraste na paciente (GIANNOTTI, 2016).

Figura 7 – Ilustração do exame de ressonância magnética sendo aplicada em uma paciente.

Fonte – (GIANNOTTI, 2016)

3.2 Termografia

A partir de teoria de Hipócrates, que define que quando uma parte do corpo está mais quente ou mais fria que o normal é indício de doença. William Herschel descobriu em 1800 que cada cor do arco-íris permitia passar diferentes quantidades de calor, o que acarretou na descoberta do infravermelho (CÔRTE; HERNANDEZ, 2016). Inicialmente, a termografia ou imagem por infravermelho foi projetada pelo exército americano nos anos de 1950 com o objetivo de auxiliar na visão noturna (WOLFF et al., 2012) (BRIOSCHI, 2017).

Dessa forma, o processo de criação das imagens termográficas é feito através de câmeras térmicas, tais câmeras possuem a capacidade de realizar a captura da radiação infravermelha que é emitida por qualquer corpo que tenha temperatura acima do zero absoluto (0K ou -273ºC) (ALTOÉ; FILHO, 2012). Essa radiação é definida pela lei de *Stefan-Boltzmann*, que relaciona a emissão de radiação com a temperatura do corpo. Tal lei é descrita pela Equação 3.1

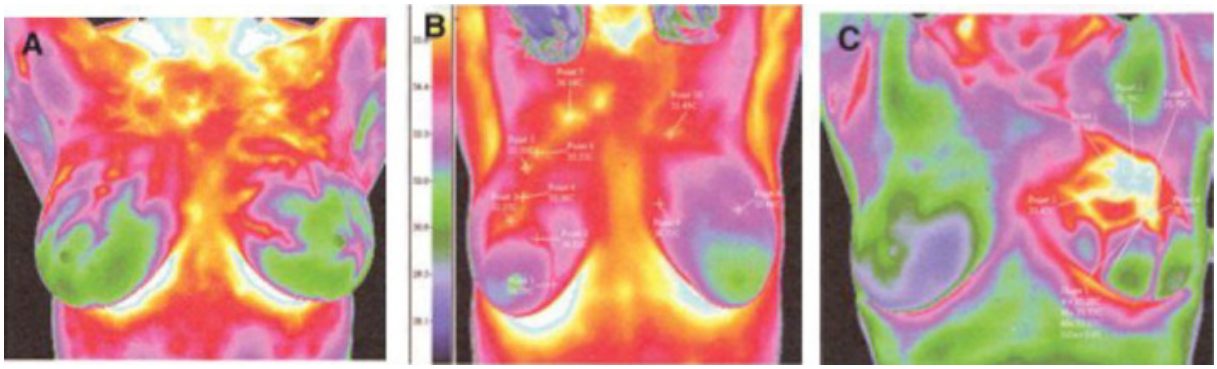
$$SB = \epsilon \cdot B \cdot T^4 \qquad [W/m^2] \quad (3.1)$$

onde ϵ representa a emissividade que o corpo possui, B é a constante de *Stefan-Boltzmann*, equivalente a aproximadamente $5,7 \times 10^{-8} W m^{-2} K^{-4}$, e T corresponde a temperatura absoluta do corpo dada em Kelvin.

O ser humano apresenta um mecanismo que realiza o controle da temperatura do corpo, deixando-a em uma mesma faixa independentemente do ambiente o qual se encontre. Esse mecanismo é denominado termorregulação. A literatura aponta que mudanças de temperatura no corpo humano que sejam anômalas a termorregulação podem estar associadas a alguma doença (CÔRTE; HERNANDEZ, 2016). O uso na medicina começou a partir da descoberta, em 1957 por Ray Lawson, sobre a pertubação dos níveis de temperatura consequentes da formação de tumores malignos nas mamas (BRIOSCHI, 2017).

Na Figura 8, são ilustrados um caso sadio e dois casos que possuem câncer. Percebe-se que em (A), que representa uma paciente saudável, há uma similaridade na distribuição de temperatura entre as mamas esquerda e direita. Em contrapartida (B) e (C), por apresentarem câncer, é visível a presença de uma perturbação nos níveis de temperatura na região a qual se encontra o tumor. Além disso, outra característica interessante é a detecção, pela termografia, da perturbação causada por um nódulo em estágios iniciais, conforme pode ser visto em (B).

Figura 8 – Uso da termografia para diagnóstico de doenças nas mamas. Em (A) está uma paciente saudável, (B) câncer nos estágios iniciais na mama direita, e (C) avanço de câncer na mama esquerda.



Fonte – (KANDLIKAR et al., 2017)

Como descrito na Seção 3.1, no processo de formação do tumor há a criação de vasos e consequentemente aumento do fluxo sanguíneo no local. Esse processo perturba os níveis de gradientes de temperatura locais em comparação ao tecido normal. Assim, a análise dessas regiões pode servir de auxílio ao diagnóstico dessa patologia. Além disso, a termografia é um exame não invasivo, pois não utiliza qualquer radiação ionizante, nem requer acesso intravenoso, dessa forma, é um exame que não oferece riscos às pacientes.

A termografia infravermelha estática voltada para mamas segue um protocolo de aquisição que é dividido em quatro etapas: (1) recomendações a pacientes, que consiste em evitar consumo de bebidas alcoólicas, cigarro e cafeína, e aplicação de produtos na região momentos antes do período do exame; (2) organização do ambiente, para evitar influências externas de temperatura; (3) preparo da paciente; e (4) ajuste das configurações da câmera, considerando variáveis como a distância da paciente, temperatura ambiente e outros elementos que possam influenciar na temperatura. Outra modalidade de termografia é a dinâmica, que realiza o monitoramento das mudanças de temperatura e depende menos das condições de temperatura da sala se comparada a estática (SILVA, 2015a).

3.3 Processamento de imagens

Uma imagem pode ser definida como uma função de duas variáveis $f(x, y)$, onde x e y representam as coordenadas do plano cartesiano e o valor da função f em quaisquer coordenadas corresponde a intensidade desse ponto na imagem (GONZALEZ; WOODS, 2008). Ela é dita digital quando os valores de x, y e f são todos finitos. Métodos que aplicam operações com o intuito de extrair informação útil e/ou realizar o melhoramento desse dado são conhecidos como técnicas de processamento de imagens. Assim, uma imagem

pode ser representada matricialmente como:

$$F = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,m) \\ f(1,0) & f(1,1) & \dots & f(1,m) \\ \vdots & \vdots & \dots & \vdots \\ f(n,0) & f(n,1) & \dots & f(n,m) \end{bmatrix} \tag{3.2}$$

Cada ponto, pertencente a matriz definida como imagem, é denominado *pixel* ou *elemento da imagem*. Esse elemento pode ter sua intensidade luminosa separada em componente de iluminação e componente de reflectância (QUEIROZ, 2012). Onde o produto entre a reflectância e iluminação é expresso por:

$$f(x,y) = i(x,y).r(x,y) \tag{3.3}$$

As técnicas de processamento e análise de imagens podem ser classificadas, de acordo com o resultado das operações produzidas, em três níveis: baixo, médio e alto (FACON, 2006). No primeiro, as operações sobre os *pixels* têm como resultado valores numéricos associados a cada um deles. No médio, a saída é uma lista de características. E no último, essas características são analisadas com o objetivo de realizar uma interpretação sobre imagem.

Segundo Gonzalez e Woods (2008), podem ser realizadas três operações espaciais sobre os *pixels*, são elas: operação de um único pixel; operações vizinhança; e transformações do espaço geométrico.

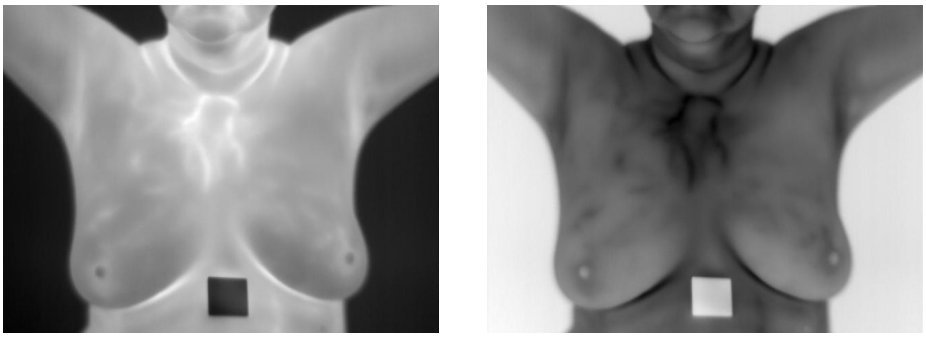
Operações de único pixel

Essa operação realiza a aplicação de uma função que considera apenas os valores de intensidade e independe da posição do pixel. Geralmente essas operações são irreversíveis, pois dois valores de cinzas podem ser mapeados para um ponto (JAHNE, 1997). Ela pode ser representada pela Equação 3.4

$$s = T(z) \tag{3.4}$$

onde s corresponde ao resultado da função de transformação T sobre o valor de intensidade z do pixel. Um exemplo de sua aplicação é na obtenção da negativa de uma imagem de 8-bit (Figura 9) que é definida por $s = L - z$, sendo L o valor máximo de intensidade que o pixel pode assumir, no caso de uma imagem de $8 - bits$, L equivale a $2^8 - 1 = 255$.

Figura 9 – Ilustração da aplicação da operação negativa sobre uma imagem 8 bits. (a) Imagem original em nível e cinza. (b) Imagem negativa.



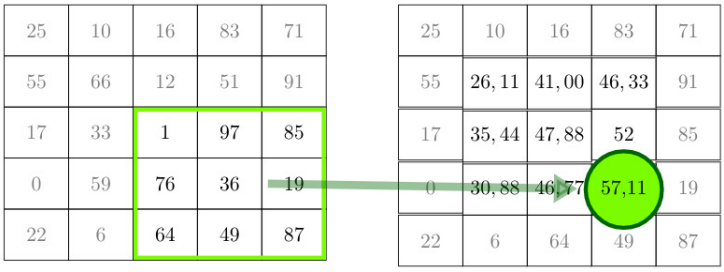
(a) (b)

Operações de vizinhança

A operação de vizinhança gera um novo valor de pixel, nas mesmas coordenadas, considerando os valores em torno da sua vizinhança para a realização dessa ação (GONZALEZ; WOODS, 2008). Os vizinhos de um pixel $p(x, y)$ localizado an posição x,y correspondem ao conjunto de todos os pontos vizinhos a p(x,y).

Na Figura 10, é ilustrada uma operação de vizinhança, que é definida pela expressão $g(x, y) = \frac{\sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} f(x+i,y+j)}{nm}$, onde n e m representam as dimensões da janela. Nesta imagem, os valores de n e m são considerados iguais a 3, ou seja, o novo pixel $g(x, y)$ é gerado pela média entre o pixel $f(x, y)$ e seus 8-vizinhos. Além disso, essa operação não está considerando os valores de borda e decidiu-se manter os valores iguais aos da imagem original.

Figura 10 – Nova imagem, g, construída a partir da operação da média sobre a vizinhança do pixel (x, y) efetuada sobre uma imagem f qualquer. Os valores das bordas foram mantidos iguais à imagem original.



Transformações geométricas espaciais

As transformações geométricas são definidas como o relacionamento entre pontos de duas imagens diferentes. Esta relação pode ser expressa de duas maneiras: as coordenadas da imagem de saída, $\boldsymbol{x}'$, são expressas como uma função sobre as coordenadas de uma

imagem entrada, $\boldsymbol{x}$, ou o contrário (JAHNE, 1997) (JÄHNE; HAUSSECKER; GEISSLER, 1999):

$$\boldsymbol{x'} = T(\boldsymbol{x}) \text{ ou } \boldsymbol{x} = T^{-1}(\boldsymbol{x'}) \tag{3.5}$$

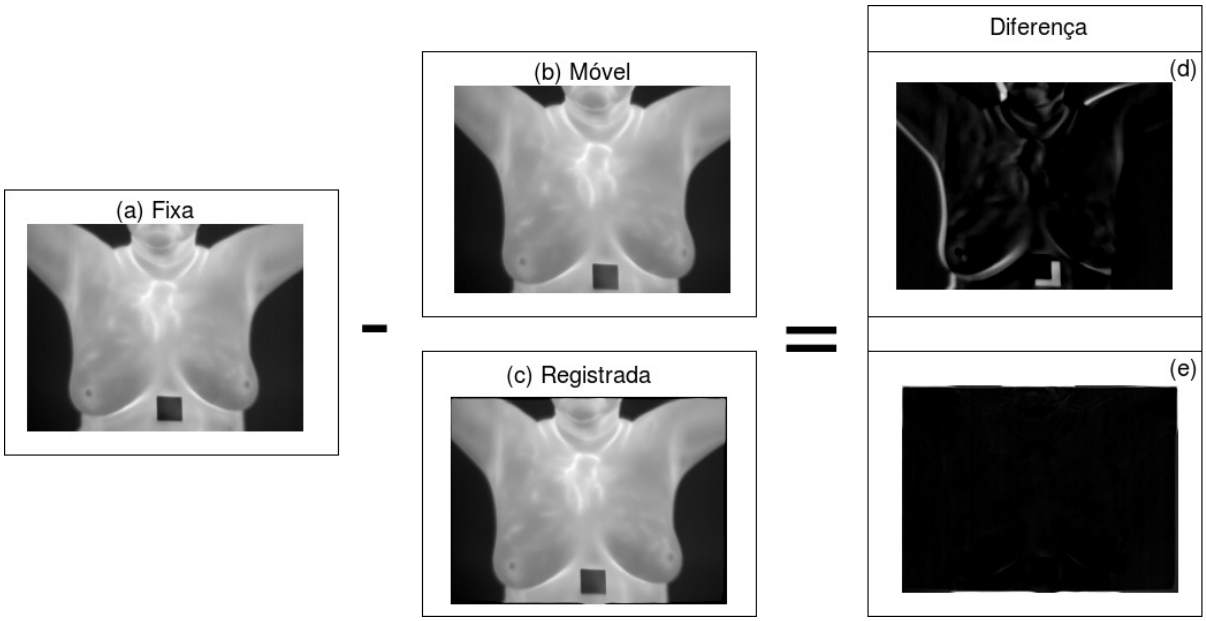
onde, o T representa a transformação e T^{-1} a sua inversa. Essas duas operações são chamadas de *mapeamento direto* e *mapeamento inverso*. A primeira atribui o pixel de entrada a imagem de saída e na segunda, as coordenadas de um ponto na imagem de saída são mapeados de volta para a imagem de entrada (JÄHNE; HAUSSECKER; GEISSLER, 1999).

Registro de imagens

As operações geométricas descritas anteriormente são aplicadas no processo de registro de imagens. Esse processo consiste no alinhamento entre duas ou mais imagens pertencentes a uma mesma cena obtidas por diferentes sensores ou tiradas em instantes distintos, fazendo o uso de um mesmo sensor. Em outras palavras, o objetivo é encontrar a função de transformação T que realize o alinhamento de uma imagem com base em uma imagem referência (Figura 11)(GONZALEZ; WOODS, 2008).

A Figura 11 ilustra um exemplo de registro de imagens, onde (a) representa a imagem fixa e (b) a imagem a ser registrada, também conhecida como móvel. O resultado do registro em (b) aparece em (c). Percebe-se que a diferença entre (a) e (c) é menor que (a) e (b), pois em (c) a imagem foi ajustada para se assemelhar com a imagem original.

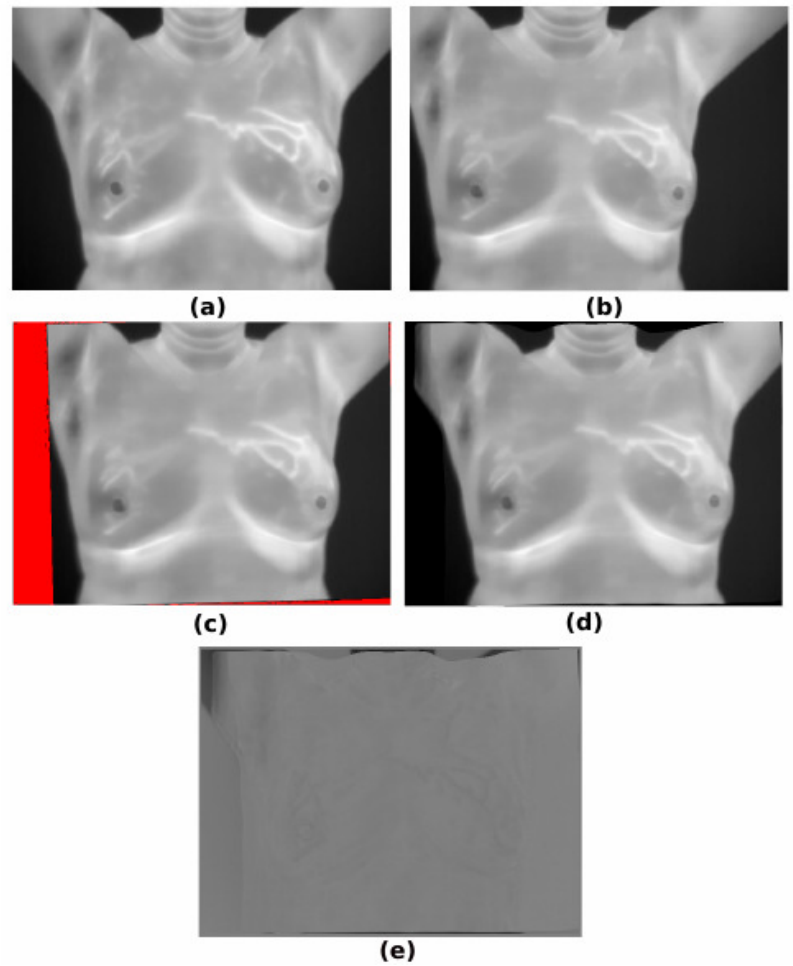
Figura 11 – Aplicação de registro de imagens.



Por sua vez, o registro pode ser dividido em três categorias: (1) baseados por pontos - onde é obtida uma transformação que alinha os pontos; (2) baseado por superfícies - a transformação que melhor alinha as superfícies em diferentes imagens é computada; e (3) por intensidade - calcula uma transformação entre duas imagens utilizando apenas os valores de *pixels* (FITZPATRICK; HILL; R.MAURER, 2000). Neste trabalho, foi utilizado o registro por intensidade descrito por Silva (2015a).

Segundo Silva (2015a), o processo de registro das imagens é executado em dois estágios: (1) o primeiro utiliza informação mútua que serve para verificar o quanto uma imagem é similar a outra, com base na intensidade dos *pixels*, e a função gerada realiza transformações de translação, rotação e escala; (2) a segunda usa uma medida local de similaridade que verifica as distorções no domínio espacial da imagem. A Figura 12 ilustra o processo de aplicação do registro sobre as imagens.

Figura 12 – (a) representa a imagem fixa; (b) imagem a ser registrada; (c) registro implementado com o primeiro estágio; (d) registro do segundo estágio sobre (c); (e) diferença entre a imagem (d) e (a).

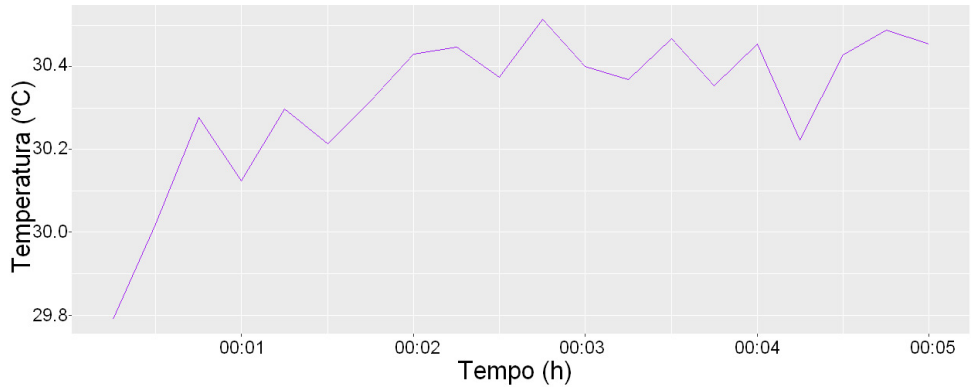


Fonte – (SILVA, 2015a)

3.4 Série temporal

Uma série temporal (ST) corresponde a um conjunto de observações feitas em diferentes instantes de tempo que apresentam dependência temporal, sendo a ordem dos dados uma característica importante para o processo de análise. Assim, uma série temporal, denotada como T, de comprimento n, pode ser representada como $T = (t_1, t_2, t_3, \ldots, tn)$, onde t_i representa a observação de um dado no instante de tempo i. A Figura 13 ilustra uma série temporal extraída de uma paciente, onde o eixo x corresponde ao tempo da observação, medido em horas (h), e o eixo y corresponde a temperatura, medida em graus Celsius (°C).

Figura 13 – Ilustração de uma série temporal extraída de uma paciente.



Em seus estudos, Chatfield (1986) define que o processo de análise de séries temporais envolve, principalmente, 4 objetivos. São eles:

Descrição: Normalmente, um dos primeiros passos feitos no processo de análise de séries temporais é a visualização do dados, que tem por objetivo obter medidas descritivas que possam descrevê-las (CHATFIELD, 1986). Por exemplo, na Figura 13, algumas das características que podem ser observadas são os valores mínimo e máximo da série. O menor valor está presente no início da coleta dos dados, e o maior valor no meio do processo.

Explicação: Quando as observações consideram duas ou mais variáveis, pode ser utilizado a variação de uma série para explicar outra. Isso pode servir para compreender os processos que geram as séries dadas.

Predição: A predição tem como objetivo, obter futuros valores das séries temporais, conhecendo os dados das séries observadas.

Controle: Uma série temporal gerada por medidas de qualidade de fabricação, o objetivo da análise pode ser o controle do processo. Um controle de qualidade estatístico, as

observações são colocadas em gráficos de controle e ações são tomadas de acordo com o resultado desses gráficos.

Muitas abordagens utilizadas para análise de séries temporais tem como base o uso de modelos estatísticos, como técnicas de regressão polinomiais, ou modelos auto-regressivos, como o modelo auto-regressivo integrado de médias móveis (ARIMA). Um grande desafio no processo de mineração de séries temporais está no fato que na análise de um grande conjunto de séries, as técnicas tradicionais tornam-se complexas (MALETZKE, 2009).

Nesta seção, são apresentados detalhes sobre técnicas empregadas no processo de mineração de séries temporais utilizadas neste trabalho. Isso inclui métodos que realizam a conversão entre os modelos de dados, como, por exemplo a modelagem de uma série temporal no domínio da frequência.

3.4.1 Pré-processamento

Séries temporais podem apresentar problemas que negativamente impactam a qualidade dos resultados dos algoritmos de mineração de dados, como ruído e valores faltantes. Por essa razão, faz-se necessário uma etapa de pré-processamento, ou preparação desses dados. A literatura aponta que mais de 80% do processo de mineração de séries temporais está concentrada nessa etapa (MALETZKE, 2009), que se dá, basicamente, através das seguintes técnicas principais: tratamento de dados faltantes, remoção de ruídos e normalização.

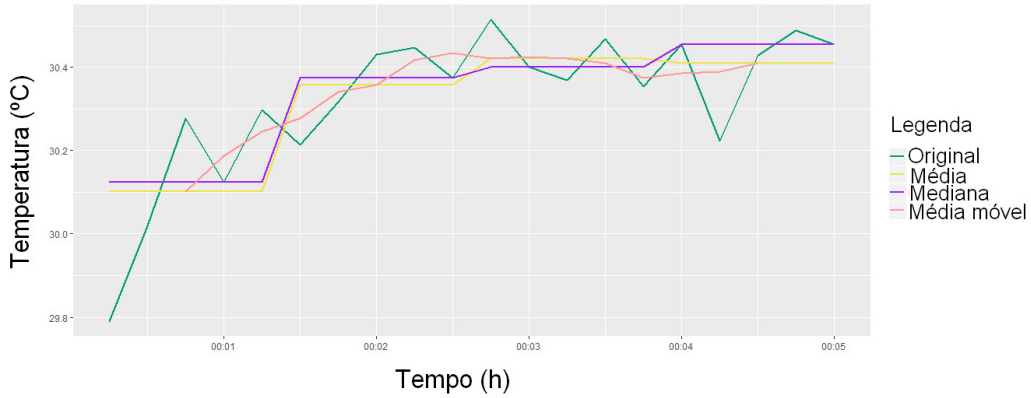
Duas abordagens podem ser utilizadas para lidar com dados faltantes: (1) ignorar dados faltantes: neste caso, em tarefas que utilizam técnicas de comparações por similaridades, por exemplo, os cálculos são realizados ignorando as regiões que não contém valores; (2) preencher dados faltantes: os valores omissos podem ser preenchidos através de técnicas específicas, como a interpolação dos dados com base nos valores vizinhos àquele que se deseja estimar (CHINO, 2014).

Para tratar ruídos em séries temporais, duas técnicas apontadas pela literatura foram utilizadas neste trabalho: *binning* e suavização por média móvel. Na primeira, a série é dividida em segmentos ou baldes, contendo a mesma quantidade de elementos. Esses elementos são suavizados utilizando a média, a mediana ou os valores extremos do segmento (CHINO, 2014).

A suavização por média móvel, técnica bastante utilizada no mercado de ações, é um tipo especial de filtragem que realiza a transformação de uma série temporal em outra. O termo "média móvel" é usado para descrever este procedimento porque cada média é calculada deixando a observação mais antiga e incluindo a próxima observação (HYNDMAN, 2011).

A Figura 14 ilustra a aplicação da técnica binning utilizando a média e a mediana dos segmentos e a técnica de suavização por média móvel, aplicadas sobre uma série temporal de uma paciente.

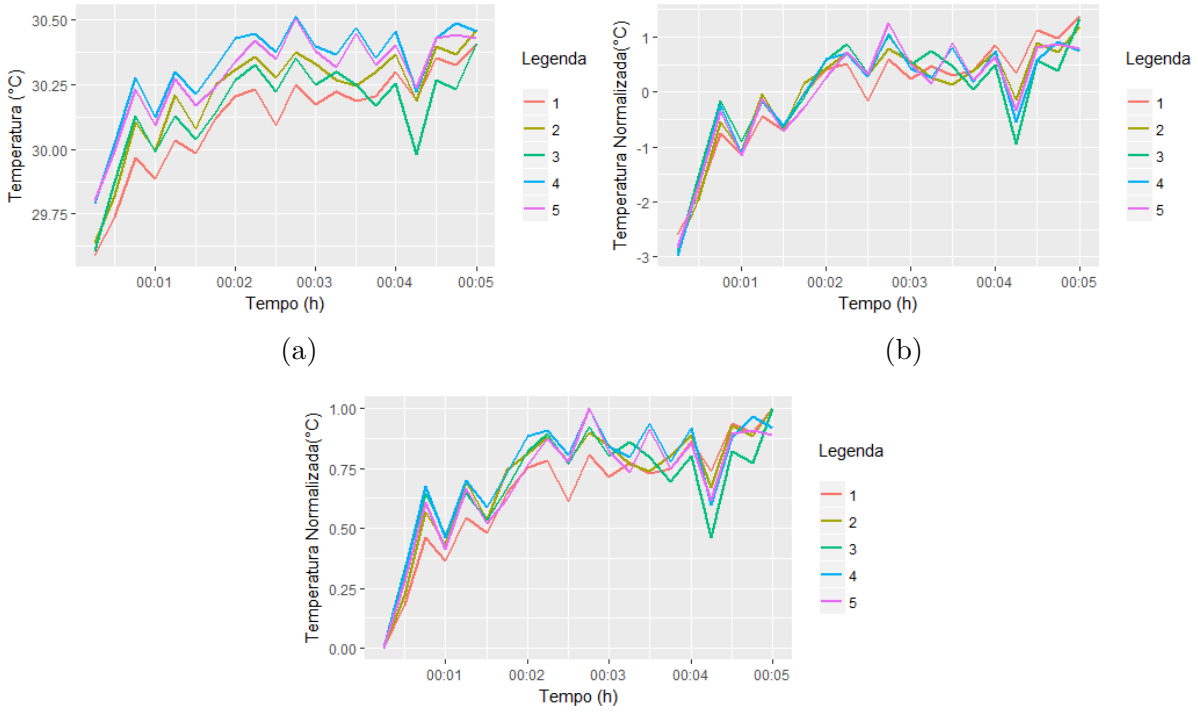
Figura 14 – Pré-processamento de uma série temporal para remoção de ruídos utilizando as técnicas *binning* (média e mediana) e suavização por janela móvel.



Para tratamento de dados faltantes, a técnica *Matrix Profile*, explicada na Seção 3.4.6, utiliza a primeira abordagem. Para a remoção de ruídos, neste trabalho é aplicada a suavização por médias móveis.

As abordagens tradicionais de medidas de similaridades utilizam comparações morfológicas (MALETZKE, 2009). Para realizar o processo de comparação entre duas séries temporais, faz-se necessário realizar ajustes sobre esses dados. Esses ajustes são feitos com o objetivo de evitar que erros de medição dos valores, causados pelos dados estarem em escalas ou localizações diferentes, não sejam introduzidos. Na Figura 15a , é ilustrado um conjunto de cinco séries temporais visualmente similares. Entretanto, por não estarem bem posicionadas, levam a erros de medição pelas métricas discutidas na Subseção 3.4.2, que realizam a análise das séries ponto-a-ponto.

Figura 15 – (a) conjunto de séries temporais sem normalização; (b) normalização em amplitude; (c) normalização em escala



Em Maletzke (2009) são discutidas várias técnicas de normalização de séries temporais, dentre elas:

Normalização de escala: A Equação 3.6 realiza a normalização da série temporal para o intervalo [0,1]. A Figura 15c ilustra o resultado da aplicação desta técnica sobre 5 séries extraídas de uma paciente.

$$Z'(t) = \frac{Z(t) - min(Z)}{max(Z) - min(Z)} \tag{3.6}$$

onde $Z'(t)$ representa o resultado na operação no tempo t e Z a série temporal.

Normalização de amplitude: a série temporal é ajustada para possuir média zero e desvio padrão unitário. A normalização é obtida através da Equação 3.7. A Figura 15b ilustra o resultado da aplicação desta técnica sobre 5 séries extraídas de uma paciente.

$$Z'(t) = \frac{Z(t) - \mu(Z)}{\sigma(Z)} \tag{3.7}$$

onde $Z'(t)$ representa o resultado na operação no tempo t, Z a série temporal, $\mu(Z)$ o valor médio de Z e $\sigma(Z)$ a variância da série.

3.4.2 Similaridade

A busca por padrões similares é muito importante no processo de mineração de séries temporais, pois ajuda em tarefas como predição, teste de hipóteses e detecção de regras (FALOUTSOS; RANGANATHAN; MANOLOPOULOS, 1994).

A distância euclidiana, é uma das medidas de distância mais utilizadas (KEOGH; KASETTY, 2003). Sejam duas séries temporais de comprimento n, $X = (x_1, x_2, \ldots, x_n)$ e $Y = (y_1, y_2, \ldots, y_n)$, o cálculo da distância euclidiana pode ser definido de acordo com Equação 3.8:

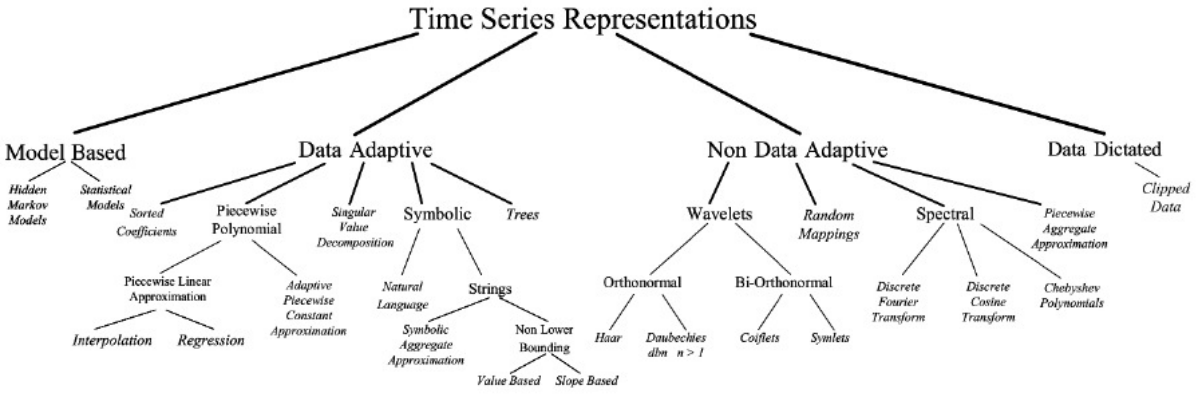
$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sum_{i=0}^{n} \sqrt{(x_i - y_1)^2} \tag{3.8}$$

3.4.3 Transformada discreta de fourier *(Discrete Fourier Transform - DFT)*

As séries temporais apresentam alta complexidade, e a análise em sua forma original se torna uma tarefa bastante custosa. Dessa forma, é necessário o uso de mecanismos que reduzam essa complexidade, representando essas séries de forma simplificada e preservando suas características principais (CHINO, 2014).

Na Figura 16, são ilustradas as representações que podem ser aplicadas nas séries temporais, as quais são divididas em quatro grupos, sendo eles: baseado em modelo; dado adaptativos, dado não adaptativo; e dado ditado. Segundo Chino (2014), na categoria não adaptativa os mais utilizados são a Transformada Discreta de Fourier (*Discrete Fourier Transform* - DFT), Transformada Discreta de Wavelet (*Discrete Wavelet Transform* - DWT) e a *Piecewise Aggregate Approximation* (PAA).

Figura 16 – Árvore contendo as representações de séries temporais.



Fonte – (CHINO, 2014)

Uma série de *Fourier* que possua período T, é representada pela Equação 3.9 que

é definida a seguir:

$$f(t) = \sum_{i=0}^{\infty} a_i sen(\omega_i t) + b_i cos(\omega_i t) + c \tag{3.9}$$

onde a_i e b_i correspondem aos pesos atribuídos aos elementos cossenoidais e senoidais, ω corresponde a frequência angular dada por $\omega_i = \frac{2\pi i}{t}$ e c uma constante. A sua aplicação em sinais não periódicos é denominada transformada de *Fourier* (SILVA, 2014).

Na transformada discreta de *Fourier*, as séries temporais são moldadas no domínio da frequência. Os coeficientes complexos, f_k, gerados a partir da transformação de uma série temporal $T = t_1, t_2, ..., t_n$ são obtidos através da Equação 3.10.

$$f_k = \sum_{j=0}^{n-1} t_j e^{\frac{-i2\pi jk}{n}} \tag{3.10}$$

onde $i = \sqrt{-1}$ e $k = 0, ..., n$.

Propriedades básicas da DFT

A transformada discreta de fourier apresenta algumas propriedades importantes de acordo com o Teorema 1.

Teorema 1 *Suponha que a sequência* $\{h_j\}_{j=0}^{N-1}$ *tenha DFT* $\{H_k\}$ *de N pontos e a sequência* $\{g_j\}_{j=0}^{N-1}$ *tenha DFT* $\{G_k\}$ *de N pontos (PUPIN; SILVA; CARBONE, 2010). Então, têm-se as seguintes propriedades:*

(a) linearidade: Para as todas constantes complexas a e b, a sequência $\{ah_j + bg_j\}_{j=0}^{N-1}$ *tem DFT* $\{aH_k + bG_k\}$ *de N pontos.*

(b) periodicidade: Para todos os inteiros k, tem-se $H_{k+N} = H_k$

(c) inversão: Para $j = 0, 1, \ldots, N$*, tem-se:*

$$h_j = \frac{1}{N} \sum_{k=0}^{N-1} h_j e^{-i2\pi jk/N}$$

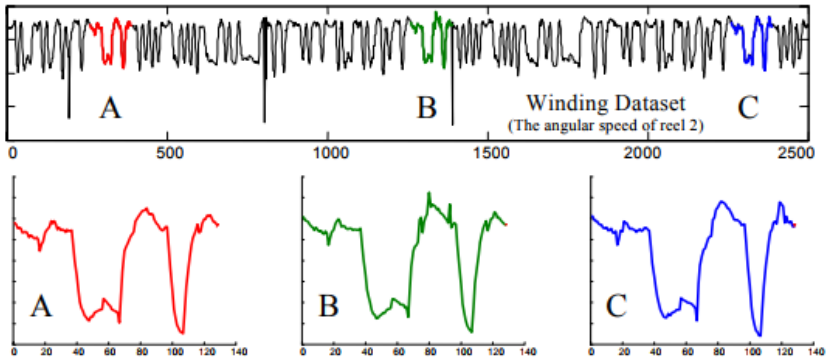
O cálculo da transformada discreta de *Fourier* pode ser feito através do método da transformada rápida de *Fourier* (FFT - do inglês, *Fast Fourier Transform*), o qual possui a complexidade $O(n \log n)$ (CHINO, 2014).

Neste trabalho é utilizado a FFT como parte da técnica *Matrix Profile*, que é descrita na Seção 3.4.6.

3.4.4 Descobertas de *motifs*

Em séries temporais, a presença de padrões desconhecidos e frequentes é denominado *motifs* em analogia com suas contrapartes discretas em biologia computacional (LONARDI; PATEL, 2002). *Motifs* também podem ser definidos como subsequências que se repetem dentro de uma série temporal (MALETZKE et al., 2013). A Figura 17 ilustra a ocorrência de um motif em um *dataset* industrial, onde A, B e C correspondem a subsequências que são mais similares entre si.

Figura 17 – Exemplo de série temporal contendo um *motif* com três ocorrências.



Fonte – (CHIU; KEOGH; LONARDI, 2003)

O processo de identificação de *motifs* leva em consideração a comparação entre subsequências a partir de uma medida de similaridade. A seguir são apresentadas um conjunto de definições presentes na literatura que são importantes na tarefa de busca por *motifs*.

Definição 3.4.1 *(Subsequência)(LIN et al., 2003) (CHIU; KEOGH; LONARDI, 2003) Dada uma série temporal T de comprimento m, uma subsequência C de T é uma amostragem de comprimento $n < m$ de contígua posição de T, isto é, $C = t_p, ..., t_{p+n-1}$ para $1 \leq p \leq m - n + 1$*

Na Figura 17, uma subsequência teve três ocorrências. A definição de subsequência é importante e usada em tarefas que busquem verificar sua similaridade com demais subsequências (LIN et al., 2003). A seguir é explicada a operação de *casamento*, que tem como o objetivo dizer o quanto as subsequências são similares entre si.

Definição 3.4.2 *(Casamento)(LIN et al., 2003) (CHIU; KEOGH; LONARDI, 2003) Dado um número real positivo R, chamado de limiar, e uma série temporal T contendo uma subsequência C começando na posição p e subsequência M começando em q, e D a distância entre duas sequências, se $D(C, M) \leq R$, então M é dito similar a C.*

Apesar da definição de casamento ser algo simples, faz-se necessário definir o que é casamento trivial. Visto que, as melhores correspondências de uma subsequência é consigo mesma, entretanto, o principal interesse é em encontrar os melhores casamentos, além da própria subsequência.

Definição 3.4.3 *(Casamento trivial)(LIN et al., 2003) (CHIU; KEOGH; LONARDI, 2003) Dada uma série temporal T, contendo uma subsequência C começando na posição p e uma subsequência correspondente M começando em q, considera-se M e C um casamento trivial, caso $p = q$ ou se não existir uma subsequência M' começando em q' tal que $D(C, M') > R$, e $q < q' < p$ ou $p < q' < q$.*

Para o processo de identificação de *motifs* existem várias abordagens, dentre elas: a força bruta - consiste em fazer comparações sucessivas de cada subsequência existente dentro da série temporal (MALETZKE, 2009); o *Random Projection* proposto por (CHIU; KEOGH; LONARDI, 2003) - método probabilístico, que utiliza uma heurística para a *motifs* e o *Matrix Profile* (YEH et al., 2016) que usa um algoritmo de busca de similaridade rápido, explorando a sobreposição entre as subsequencias através da transformada rápida de *fourier*. Neste trabalho, é usado o *Matrix Profile*.

3.4.5 Detecção de anomalia

A detecção de anomalia, também conhecida como *novelty detection*, busca verificar se os dados possuem informações incomuns, ou seja, diferente do esperado (MÖRCHEN, 2006). A sua aplicação está presente em diferentes campos que exijam a detecção de anomalias, como detecção de intrusões em redes de computadores, busca por falhas em máquinas. Keogh, Lin e Fu (2004) introduziram um novo termo chamado *discord*, discords de séries temporais são subsequências de uma ST que são mais diferentes em comparação as demais subsequências dessa série.

Para este trabalho, como o objetivo de encontrar subsequências similares e anomalias nas séries temporais, é utilizado o *Matrix Profile*, o qual será abordado na Seção 3.4.6.

3.4.6 *Matrix Profile*

O *Matrix Profile* (MP) é um método recente, robusto e rápido, aplicado em diversas tarefas de mineração de séries temporais, como, por exemplo descoberta de *motifs*, detecção de anomalia e segmentação semântica (YEH et al., 2016) (YEH; HERLE; KEOGH, 2016) (DAU; KEOGH, 2017)(GHARGHABI et al., 2017).

O método utiliza um algoritmo de busca de similaridade rápido sob a distância euclidiana z-normalizada (normalização em amplitude), como uma sub-rotina, explorando a sobreposição entre subsequências usando a FFT (YEH et al., 2016).

Para compreensão da técnica, é necessário definir alguns conceitos importantes, que são descritos a seguir.

Definição 3.4.4 *(Distance Profile - D)(YEH et al., 2016) (YEH et al., 2017) é um vetor das distâncias euclidianas entre uma determinada consulta e cada subsequência em todo conjunto de subsequências.*

Para uma subsequência qualquer, pode ser calculada a sua distância para as demais subsequências e armazenada em uma estrutura ordenada chamada de *distance profile*. O é definido a seguir.

Definição 3.4.5 *(Conjunto de todas as subsequências)(YEH et al., 2016) (YEH et al., 2017) um conjunto de todas as subsequências* $\boldsymbol{A}$ *de uma série temporal* $\boldsymbol{T}$ *é um conjunto ordenado de todas as subsequências possíveis de* $\boldsymbol{T}$ *obtidas pelo deslizamento de uma janela de comprimento m em* $T = \{T_{1,m}, T_{1,m}, ..., T_{n-m+1,m}\}$*, onde m é um comprimento de subsequência definido pelo usuário.* $A[i]$ *é usado para denotar* $T_{i,m}$*.*

O relacionamento entre as subsequências mais similares e que são mais próximas é o objetivo do algoritmo. A seguir são definidos os conceitos de junção e similaridade.

Definição 3.4.6 *(Função de junção 1NN)(YEH et al., 2016) (YEH et al., 2017) dada dois conjuntos de todas as subsequências* $\boldsymbol{A}$ *e* $\boldsymbol{B}$ *e duas subsequências* $A[i]$ *e* $B[j]$*, uma função junção 1NN* $\theta_{1nn}(A[i], B[j])$ *é uma função lógica que retorna "verdadeiro" apenas se* $B[j]$ *é o vizinho próximo de* $A[i]$ *no conjunto* B*.*

Definição 3.4.7 *(Conjunto de junção de similaridade)(YEH et al., 2016) (YEH et al., 2017) dado um conjunto de todas as subsequências* $\boldsymbol{A}$ *e* $\boldsymbol{B}$*, um conjunto de junção de similaridade* $\boldsymbol{J_{AB}} = \boldsymbol{A}$ *de* $\boldsymbol{A}$ *e* $\boldsymbol{B}$ *é o conjunto que contém cada subsequência em* $\boldsymbol{A}$ *com seu vizinho mais próximo em* $\boldsymbol{B}$*:* $\boldsymbol{J_{AB}} = \{\langle \boldsymbol{A[i]}, \boldsymbol{B[j]} \rangle \mid \theta_{1nn}(\boldsymbol{A[i]}, \boldsymbol{B[j]})\}$*, o símbolo é definido como* $\boldsymbol{J_{AB}} = \boldsymbol{A} \bowtie_{\theta_{1nn}} \boldsymbol{B}$*.*

Definição 3.4.8 *(Conjunto de junção de auto-similaridade)(YEH et al., 2016) (YEH et al., 2017) Um conjunto de junção de auto-similaridade* $\boldsymbol{J_{AA}}$ *é o resultado do join de similaridade consigo mesmo. Sendo representado como* $\boldsymbol{J_{AA}} = \boldsymbol{A} \bowtie_{\theta_{1nn}} \boldsymbol{A}$ *e é representado como* P_{AA}*.*

A metodologia proposta por Yeh et al. (2016) realiza a criação de duas séries temporais denominadas *Matrix Profile* e *Matrix Profile Index* (MPI), armazenando a distância e localização dos vizinhos mais próximos de cada subsequência, sejam eles na mesma série ou em uma segunda série temporal.

Definição 3.4.9 *(Matrix profile)(YEH et al., 2016) (YEH et al., 2017)* P_{AB} *é um vetor de distâncias euclidianas entre cada par em* $\boldsymbol{J}_{AB}$ *onde* $P_{AB}[i]$ *contém a distância entre* $\boldsymbol{A}[i]$ *e seu vizinho mais próximo em* $\boldsymbol{B}$*.*

Definição 3.4.10 *(Matrix profile index)(YEH et al., 2016) (YEH et al., 2017) Um índice de matrix profile* I_{AB} *de um join de similaridade do conjunto* $\boldsymbol{J_{AB}}$ *corresponde a um vetor de inteiro, onde* $IAB = j$ *se* $\{A[i], B[j]\} \in \boldsymbol{J_{AB}}$

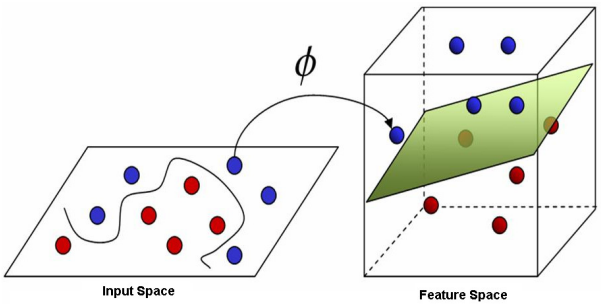
Uma característica importante no *matrix profile* é uma estrutura que armazena a localização do vizinho mais próximo e a sua similaridade. Como o *motif* representa as subsequências mais similares, então, a extração dos *motifs* da série temporal, pode ser feita através da análise da MP, recuperando o endereço das subsequências que possuírem menor distância. Por sua vez, os *discords* correspondem às subsequências mais dissimilares, ou seja, que possuam maiores valores de distância, assim a sua recuperação através da MP é trivial.

3.5 *Support Vector Machine*

Support Vector Machine (SVM) é uma técnica de aprendizagem supervisionada que faz uso da teoria de aprendizado estatístico, proposto por Vladimir Vapnik (VAPNIK; VAPNIK, 1998). O SVM constrói uma fronteira n-dimensional, conhecida como hiperplano, que realiza a separação dos dados em categorias através da maximização da margem de separação entre os exemplos (MOREIRA, 2011). Os pontos que se localizam sobre as margens são chamados de vetores de suporte. Tais pontos são bastante importantes, pois delimitam as margens do conjunto de treinamento.

Para problemas linearmente não separáveis, ou seja, que não é possível criar hiperplanos sem encontrar erros de classificação, é aplicado uma técnica chamada *kernel trick* que maximiza a fronteira do hiperplano, como ilustra a Figura 18 (MOREIRA, 2011). Assim, o *kernel trick* realiza o mapeamento desses dados para um espaço de características em que eles sejam linearmente separáveis.

Figura 18 – Separação não linear no SVM.



Fonte – (MOREIRA, 2011)

Segundo (YEKKEHKHANY et al., 2014), os três *kernels* mais comumente utilizados em casos não lineares são:

- O *kernel* polinomial: $K(x, x_i) = (\langle x, x_i \rangle + 1)^p$
- O *kernel* sigmoidal: $K(x, x_i) = \tanh(\langle x, x_i \rangle + 1)$
- O *kernel* RBF: $K(x, x_i) = exp^{-\frac{|x-x_i|^2}{2\sigma^2}}$

onde x e x_i são vetores de características do espaço de entrada.

3.6 Métricas para avaliar a classificação

A matriz de confusão é uma tabela que permite visualizar a performance do algoritmo de classificação sobre um conjunto de teste em que as classes são conhecidas. Assim, uma matriz dois por dois pode ser construída, representando a disposição do conjunto de instâncias. Esta matriz constitui a base para muitas métricas comuns (FAWCETT, 2006), dentre elas: Acurácia (AC); Sensibilidade(SE), Especificidade (ES) e Precisão (PR). A Tabela 1 contém a matriz de confusão para um classificador contendo duas classes: doente e saudável.

Tabela 1 – Matriz de confusão

	Predição	
Atual	Doente	Saudável
Doente	Verdadeiro Positivo	Falso Negativo
Saudável	Falso Positivo	Verdadeiro Negativo

As métricas Acurácia, Sensibilidade, Especificidade e Precisão são descritas respectivamente pelas Equações 3.11, 3.12, 3.13 e 3.14:

$$AC = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.11}$$

$$SE = \frac{VP}{VP + FN} \tag{3.12}$$

$$ES = \frac{VN}{VP + VN} \tag{3.13}$$

$$PR = \frac{VP}{VP + FP} \tag{3.14}$$

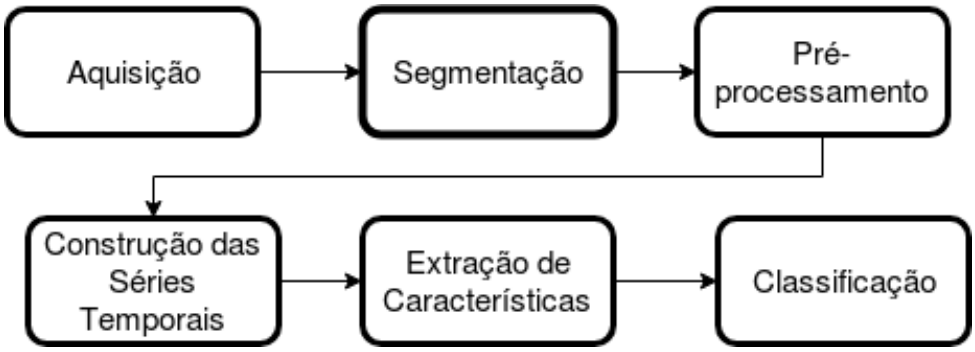
onde:

- VP: número de verdadeiros positivos (séries com anomalias classificadas corretamente);
- VN: número de verdadeiros negativos (séries normais classificadas como normais);
- FP: número de falsos positivos (séries normais classificadas como séries com anomalia);
- FN: número de falsos negativos (séries com anomalias classificadas como normais);

4 Metodologia Proposta

Neste capítulo, é apresentada a metodologia desenvolvida com o objetivo de realizar a classificação de pacientes em saudáveis (sem anomalias) e doente (com anomalias) através da análise das termografias infravermelha dinâmicas. São utilizadas técnicas de processamento de imagens e análise de séries temporais com o objetivo de realizar a classificação. A Figura 19 ilustra as etapas seguidas para desenvolvimento da metodologia proposta. As seções a seguir detalham as etapas dessa metodologia.

Figura 19 – Fluxo da metodologia para classificação das TID das mamas.



Fonte – Acervo da autora

4.1 Aquisição de imagens

Os exames utilizados no processo provém da *Database for Mastology Research with Infrared Image* (DMR-IR)(SILVA, 2015a). As imagens foram adquiridas de pacientes oriundos do Hospital Universitário Antônio Pedro da Universidade Federal Fluminense (HUAPE-UFF), constituindo 70 exames, sendo 35 saudáveis e 35 não saudáveis. O processo de aquisição das imagens é feito através do resfriamento das mamas fazendo uso de um ventilador elétrico. Este resfriamento é realizado sobre a região torácica até atingir a temperatura de $30,5^{o}$C ou após 5 minutos do início do mesmo. Após isso, o equipamento é desligado e inicia-se o processo de captura das imagens, que ocorre em intervalos de 5 segundos. No final do processo, são obtidas 20 imagens por exame. A Figura 20 ilustra o processo de aquisição das imagens.

Figura 20 – Em (a) o equipamento utilizado para resfriamento das mamas, em (b) termo-higrômetro e (c) a posição da paciente.

(a) (b) (c)

Fonte – (SILVA, 2015a)

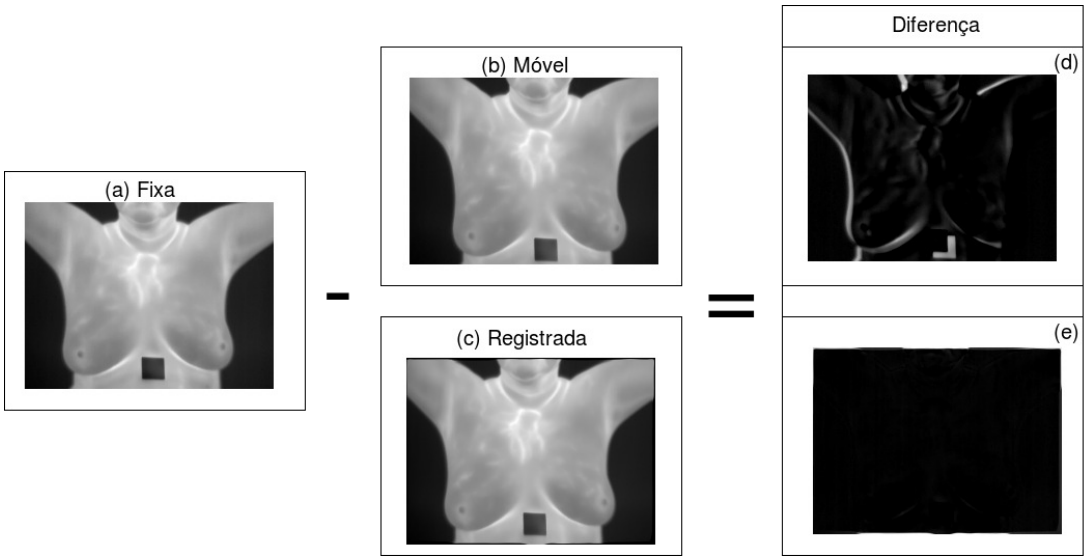
4.2 Pré-processamento

Durante o processo de captura, as pacientes apresentam movimentos involuntários, a fim de realizar essas correções é aplicado um registro baseado em intensidade descrito na Subseção 3.3. Esse registro é aplicado sobre uma imagem em níveis de cinza. Inicialmente, as matrizes de temperatura são convertidas em níveis de cinza de acordo com a Equação 4.1:

$$I_{normalizada} = 255 \times \frac{I - I_{min}}{I_{max} - I_{min}} \tag{4.1}$$

onde I representa a matriz de temperatura da imagem. Assim, as temperaturas que possuem valores mais altos são representadas com valores de maior intensidade e valores baixos irão corresponder a menores intensidades (SILVA, 2015a). Após a conversão da imagem em escala de cinza é aplicado do registro proposto por Silva (2015a). Uma imagem é escolhida como fixa e as demais são registradas com base nessa, a Figura 21 ilustra o resultado do registro.

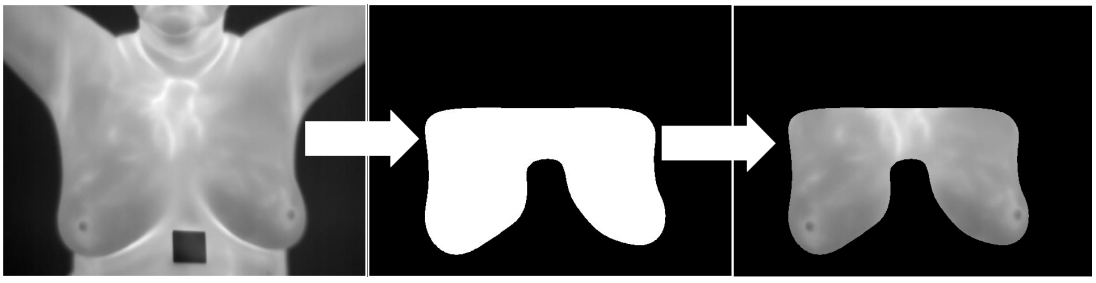
Figura 21 – (a) Imagem fixa, (b) imagem a ser registrada, (c) imagem registrada, (d)(e) diferença da fixa com a imagem (b) e (c) respectivamente.



Fonte – Acervo da autora

Como o objetivo é buscar por anomalias na região das mamas, é realizada uma segmentação da *Region of Interest* (ROI), com o intuito de eliminar informações desnecessárias e manter apenas a área das mamas. Essa operação é feita manualmente, uma vez para cada paciente (SILVA, 2015a). A Figura 22 ilustra o resultado da segmentação da região de interesse das mamas. Onde a imagem intermediária corresponde a delimitação da área, a qual será analisada nas etapas posteriores.

Figura 22 – Delimitação da região de interesse.



Fonte – Acervo da autora

4.3 Construção das séries temporais

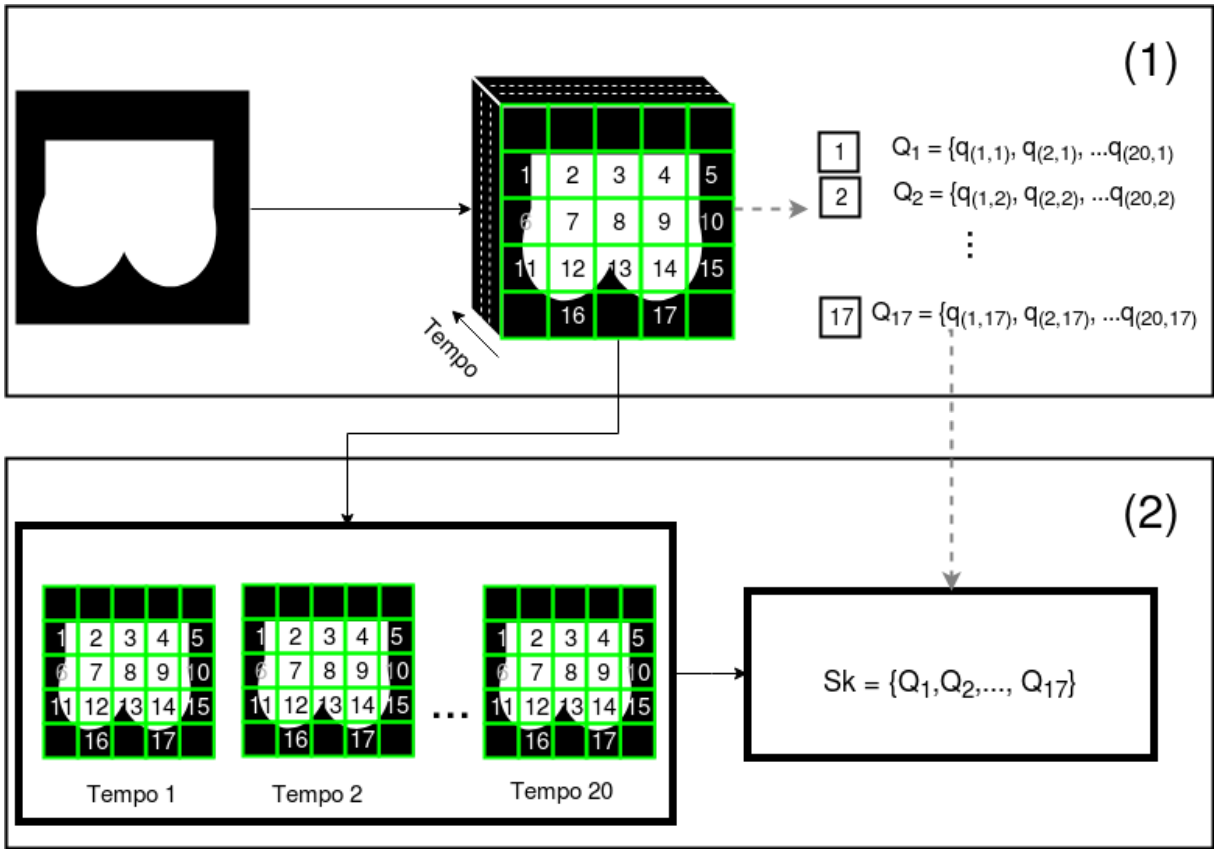
Seja T um exame de termografia da base utilizada, e T_i uma imagem do exame T no tempo i, com $1 \leq i \leq n$, com n sendo o número de imagens por paciente. Considere T_{ij} uma região j da imagem Ti que intercepta a ROI e q_{ij} a temperatura média da região T_{ij}. Define-se um série temporal S_j de uma região T_{ij} como $Sj = q_{1,j}, q_{2,j}, ..., q_{n,j}$. Esse processo de construção das séries temporais, segue a abordagem de Silva (2015a).

São ilustradas na Figura 2, as séries temporais entre duas pacientes distintas. Percebe-se que, visualmente, as séries da paciente sadia (Figura 2a) possuem maior similaridade. Além disso, nota-se que as ST com anomalias se sobressaem na paciente doente, conforme ilustrada em Figura 2b.

Entretanto, as séries temporais de cada região possuem poucas instâncias, 20 tempos, o que dificulta a busca por padrões dentro de séries temporais. Com o objetivo de suprir esse problema, optou-se por gerar super-séries através do concatenamento das séries formadas por essas regiões. Nas próximas etapas da metodologia, para a busca de anomalias e/ou similaridades são consideradas as super-séries.

A Figura 23 ilustra o processo de construção de uma super-série S_k, em uma imagem que contém 17 regiões intersectando a ROI e 20 tempos, onde k indica a *k-ésima* paciente da base de dados. Esse processo é dividido em duas partes: (1) sobre a região das mamas é realizado o janelamento e as temperaturas médias dessas janelas são obtidas. Para cada janela, é construída uma série temporal, essa etapa é semelhante a abordagem de Silva (2015a); (2) cada série gerada em (1) é concatenada sequencialmente, onde, $q_{(i,j)}$ representa a temperatura média da região j no instante i. Por fim, é obtida uma super-série contendo 20×17 valores de temperatura média.

Figura 23 – Ilustração da construção das séries temporais para um exame.



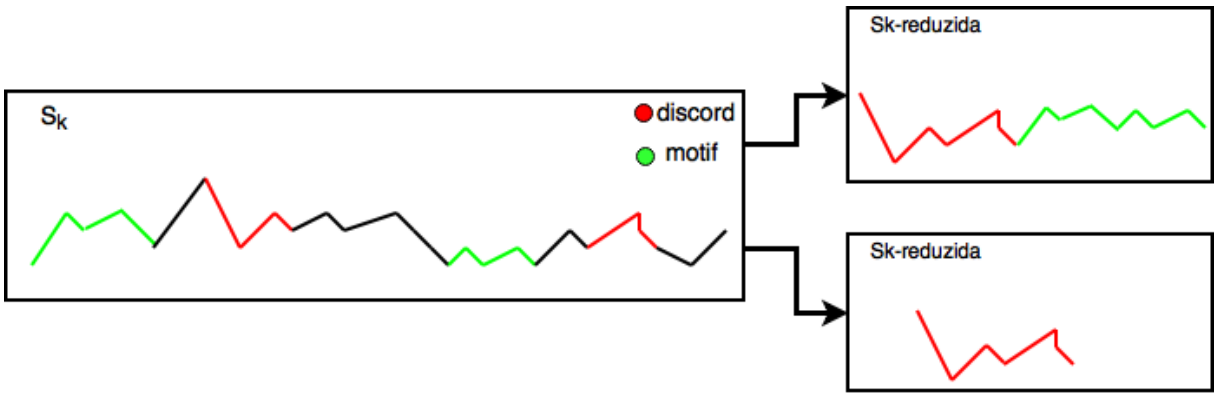
Fonte – Acervo da autora.

4.4 Extração de características

Para cada paciente, é gerada uma super-série temporal de acordo com o processo descrito na Seção 4.3. Sobre as séries construídas, são extraídas as subsequências consideradas *motifs* e *discords* utilizando o *Matrix Profile*, técnica explicada na Seção 3.4.6. O tamanho para a subsequência definido foi 20, que corresponde ao número de imagens por exame/paciente. As subsequências apontadas como *discords* e *motifs* são concatenadas gerando uma nova série temporal.

A Figura 24 ilustra o processo de construção de uma série reduzida, o qual é realizado de duas maneiras: (1) concatenamento considerando apenas os *discords*; e (2) concatenamento considerando *discords* e *motifs*. No restante do trabalho, será usado o termo *série reduzida* para representar as séries geradas por esse concatenamento.

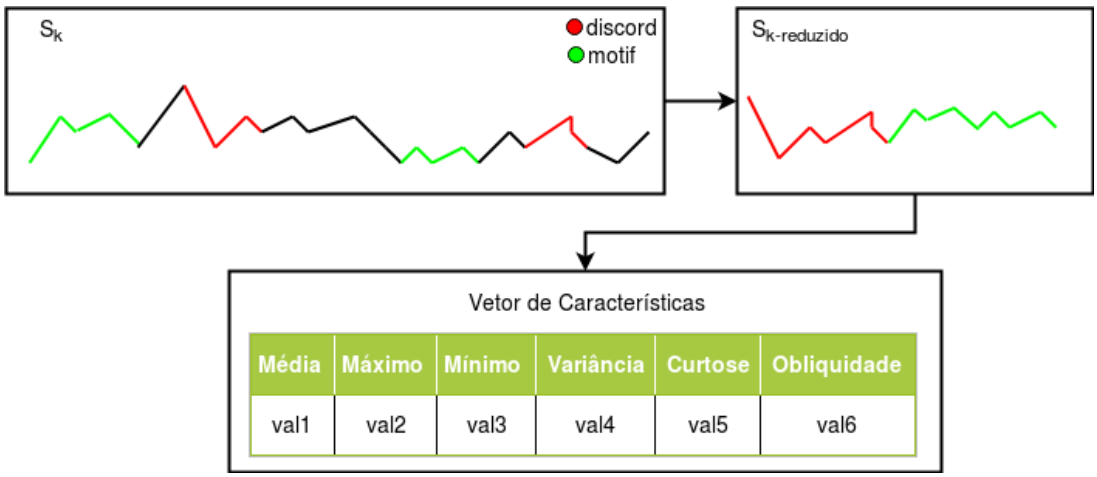
Figura 24 – Ilustração da construção das séries reduzidas a partir de uma super-série.



Fonte – Acervo da autora.

Sobre a *série reduzida*, são extraídas as características estatísticas utilizadas nos trabalhos de Maletzke (2009) e Maletzke et al. (2013). Na Figura 25, é ilustrado esse processo, na qual a primeira parte consiste na descoberta de *motifs* e *discords*, e a segunda realiza a extração das características sobre esses elementos detectados.

Figura 25 – S_k representa uma super-série; $S_{k-reduzida}$ a versão reduzida da serie temporal contendo apenas as subsequências consideradas *motifs* e *discords*; a tabela corresponde as características extraídas sobre a série reduzida.



Fonte – Acervo da autora.

4.5 Classificação

A etapa final consiste em classificar cada padrão obtido através da execução da etapa anterior com o objetivo de verificar se a paciente é doente ou saudável. Para isso, são utilizadas técnicas de reconhecimento de padrões de acordo com a análise realizada através do SVM, discutido na Seção 3.5. Para cada série temporal, são extraídas características estatísticas (média, máximo, mínimo, variância, curtose e obliquidade) sobre as séries reduzidas, conforme descrito na Seção 4.4.

Para a tarefa de classificação, foi utilizada a biblioteca *Scikit-learn* (PEDREGOSA et al., 2011) que contém uma implementação do SVM em *Python*. O objetivo é realizar a classificação de séries temporais em com anomalia e sem anomalia. Para avaliar o método proposto, foram utilizadas as métricas sensibilidade, precisão e acurácia, explicadas na Seção 3.6.

5 Resultados e Discussão

Neste capítulo, são mostrados os resultados atingidos pela metodologia apresentada no Capítulo 4. Os testes foram realizados utilizando três combinações diferentes de séries temporais: (1) série gerada pela concatenação de *discords*; (2) série obtida a partir da união de *motifs* e *discords*; e (3) super-série gerada pela concatenação de séries menores, conforme explicado na Seção 4.3. Neste trabalho, foram utilizadas as características estatísticas citadas na Seção 4.4.

A técnica *Matrix Profile* constrói duas estruturas contendo o endereço (MPI) e a distância (MP) do vizinho mais próximo de cada subsequência, conforme explicado na Subseção 3.4.6. O endereço das subsequências com menores valores na MP correspondem aos *motifs* e os maiores a *discords*. A partir da estrutura MPI, são extraídos os *discords* e *motifs* sobre as séries normalizadas e não normalizadas para os testes. Com o objetivo de verificar o quanto a normalização impacta nos resultados.

Sobre as imagens, foram extraídas as séries temporais rotuladas como saudáveis e não saudáveis utilizando o classificador SVM para validação com configuração padrão. Como a base possui poucos casos, foi utilizada a validação cruzada que avalia a capacidade de generalização de uma técnica (KOHAVI et al., 1995). Para o método de validação cruzada, foram considerados cinco grupos e obtida a média das métricas citadas anteriormente. Nas próximas seções, $AC_{\text{média}}$, $SE_{\text{média}}$ e $PR_{\text{média}}$ serão utilizadas para se referir a média das métricas entre os cinco grupos. Além disso, para testes, foram realizados considerando as séries previamente normalizadas ou não, a fim de verificar o impacto da normalização no resultado.

5.1 Extração de características sobre o conjunto

Esta seção ilustra e discute os resultados obtidos pela metodologia proposta realizando a extração de características apenas sobre a super-série construída, sem utilizar *motifs* e *discord*.

Realizando a normalização em amplitude, o classificador conseguiu classificar corretamente os verdadeiros positivos, conseguindo uma sensibilidade de 73,33%. Entretanto, nesse caso, o número de predições corretas foi baixa, com acurácia de 56,67% e precisão de 56,55%. Já utilizando a normalização em escala, os resultados obtidos foram inferiores à utilização da normalização em amplitude em relação a todas as métricas. Nota-se na Tabela 2 que o classificador atingiu os melhores resultados ao classificar os padrões sem aplicar quaisquer técnicas de normalizações, obtendo 66% de acurácia. Apesar da sensibilidade

estar abaixo em 3% do melhor resultado obtido através dos testes anteriores, houve um aumento considerável na precisão e na acurácia.

Tabela 2 – Resultados obtidos pela metodologia sem *motifs* e *discords*.

	Métricas		
Normalização	$AC_{\text{média}}$	$SE_{\text{média}}$	$PR_{\text{média}}$
Nenhuma	66%	70%	63,78%
Normalização em amplitude	56,67%	73,33%	56,44%
Normalização em escala	48,33%	70%	48,77%

5.2 Extração de características sobre o *motifs* e *discords*

Com o objetivo de melhorar os resultados da Tabela 2, foram extraídos as subsequências consideradas *motifs* e *discords*. Como a super-série é construída a partir do concatenamento de ST obtidas em cada região, espera-se que as séries com anomalia sejam classificadas como *discords*. Assim, foi considerado o tamanho de subsequência igual a 20, que corresponde a quantidade de imagens por exame.

Os *discords* são as subsequências mais diferentes dentro de uma ST e a partir do concatenamento dessas subsequências é esperado que os *discords* de pacientes doentes apresentem maior variação entre eles. Sobre os *discords* encontrados, é realizada a extração de características. A Tabela 3 mostra os resultados atingidos ao usar essa abordagem. O teste sem normalização se sobressaiu comparado aos demais, obtendo acurácia de 75%, sensibilidade de 70% e precisão de $80,17\%$.

Tabela 3 – Resultados obtidos pela metodologia considerando apenas os *discords*.

	Métricas		
Normalização	$AC_{\text{média}}$	$SE_{\text{média}}$	$PR_{\text{média}}$
Nenhuma	75%	70%	80,17%
Normalização em amplitude	50%	43,33%	53,33%
Normalização em escala	50%	33,33%	51,57%

Os *motifs* correspondem às subsequências mais similares dentro de uma série temporal, os *discords* representam o oposto disso. No último teste, foi realizada a concatenação entre essas subsequência, como ilustrado na Figura 24.

A Tabela 4 mostra os resultados obtidos com as características obtidas através do concatenamento entre *discords* e *motifs*. O classificador conseguiu classificar melhor as características obtidas sem a normalização das séries temporais, alcançando uma acurácia média de 71,67%.

Tabela 4 – Resultados obtidos pela metodologia com *discords* e *motifs*.

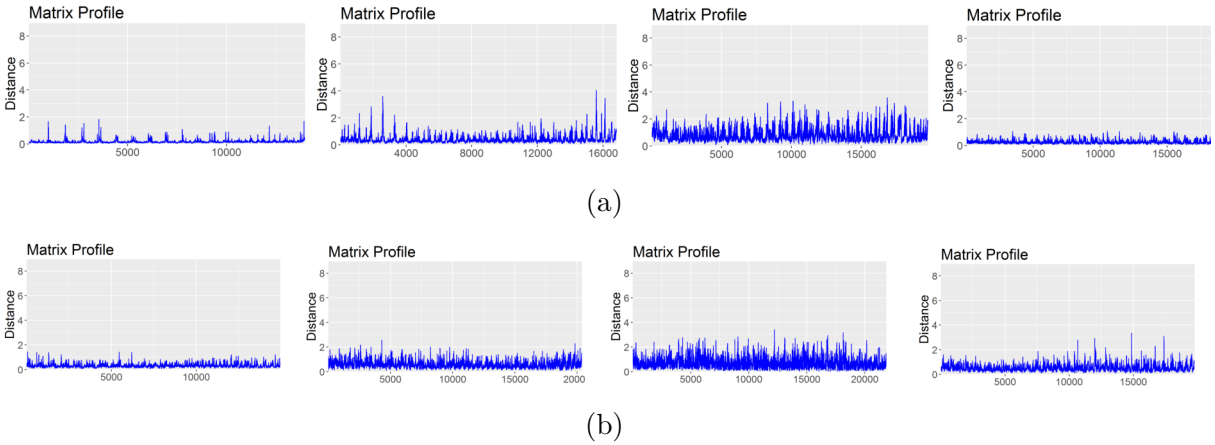
	Métricas		
Normalização	$AC_{\text{média}}$	$SE_{\text{média}}$	$PR_{\text{média}}$
Nenhuma	71,67%	70%	70,76%
Normalização em amplitude	51,67%	73,33%	52,58%
Normalização em escala	58,33%	53,33%	62,67%

A partir da análise dos resultados ilustrados nas Tabelas 2, 3 e 4, percebe-se que os melhores resultados atingidos foram sem a normalização das séries. Além disso, a melhor classificação foi obtida considerando a série formada apenas por *discords* com 75% de acurácia, 70% de sensibilidade e 80,17% de especificidade.

Entretanto, a metodologia proposta não conseguiu alcançar os resultados da literatura para essa base de dados. Silva (2015a) conseguiu uma acurácia média de 100% para as 70 pacientes da base, em contrapartida com 6 novos casos fora da base a acurácia teve uma queda para 83,33%. Silva (2015b), por sua vez, utilizou 64 pacientes da base, sendo 32 doentes e 32 saudáveis, a acurácia média alcançada foi de 100%. O autor ainda realizou testes com 12 pacientes fora da base e conseguiu a mesma porcentagem de acertos. Em seus trabalhos, os autores Silva (2015a) e Silva (2015b) exibem uma fronteira clara entre as características de pacientes doentes e saudáveis.

As características extraídas neste trabalho estão baseadas nos resultados obtidos pelo cálculo do *matrix profile*. Entretanto, para as séries temporais construídas nesse trabalho, a distinção entre pacientes saudáveis e doentes não são visíveis, como ilustrado na Figura 26. A parte superior da figura contém as séries obtidas de pacientes doentes, a inferior as obtidas com pacientes saudáveis. O eixo y representa a distância e o eixo x o índice de cada elemento da série.

Figura 26 – Ilustração do *matrix profile* sobre pacientes doentes (a) e saudáveis (b).



(a)

(b)

Fonte – Acervo da autora.

Apesar de existirem trabalhos na literatura com ótimos resultados, esses trabalhos foram testados em uma pequena base. Apesar disso, todos destacam a vantagem na modelagem dos exames através de séries temporais. Ademais, a metodologia proposta neste trabalho acrescentou evidências sobre o uso de séries temporais na resolução desse problema.

6 Conclusão

Este trabalho apresentou uma metodologia para classificação das séries, obtidas através de termografia infravermelha dinâmica, em com anomalia e sem anomalia. O método proposto realiza a modelagem do problema como série temporal. A partir dessa representação do problema como série temporal, torna-se possível o uso de técnicas de mineração de séries temporais. Neste trabalho, foi empregada a técnica *Matrix Profile*, o qual é uma técnica recente na literatura que se mostra eficiente na resolução de diversos problemas em mineração de séries temporais, dentre eles a descoberta de *motifs* e *discords*.

Foram considerada três abordagens para a construção de séries temporais: (1) construção da super-série; (2) concatenamento dos *discords*; (3) série gerada pelo concatenamento entre *motifs* e *discords*. O conjunto de características estatísticas é extraído sobre a série de cada paciente. Por fim, essas características são submetidas ao classificador SVM. Dentre essas abordagens, o método que obteve o melhor resultado foi (2) com séries não normalizadas, o qual atingiu a acurácia de 75%.

A metodologia proposta apresentou a ideia de extrair séries temporais de termografias e realizou a análise de uma nova série criada a partir do *Matrix Profile*. Assim, a principal contribuição deste trabalho está em uma nova metodologia automática, que utiliza uma das técnicas mais recentes e robusta em tarefas de mineração de séries temporais, conhecida como *Matrix Profile*.

Como trabalhos futuros, pretende-se testar outras maneiras de realizar a construção das séries temporais como, por exemplo, fazer o uso de algoritmos de agrupamento de regiões similares a fim de evitar a presença de ruídos provocados pela janela deslizante sobre a imagem. Para a extração de características, pode-se realizar uma mescla entre características extraídas da imagem e de série temporais, com o intuito de obter mais informações sobre os dados. Um dos problemas está na base de dados ser escassa, uma solução para isso seria gerar séries sintéticas, como em Forestier et al. (2016), que realiza o aumento dos dados. Futuramente, com a melhoria dos resultados, o estudo proposto poderá ser incorporado em um CAD para auxiliar o especialista no diagnóstico de doenças que possam ser identificadas através de termografias.

Referências

ALTOÉ, L.; FILHO, D. O. Termografia infravermelha aplicada à inspeção de edifícios. *Acta Tecnológica*, v. 7, n. 1, p. 55–59, 2012. Citado na página 29.

ANDRADE, W. P. *Câncer de mama: entenda a classificação BIRADS*. 2015. Disponível em: <http://www.minhavida.com.br/saude/materias/18470-cancer-de-mama-entenda-a-classificacao-birads>. Citado na página 26.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: SEATTLE, WA. *KDD workshop*. [S.l.], 1994. v. 10, n. 16, p. 359–370. Citado na página 21.

BORCHARTT, T. Análise de imagens termográficas para a classificação de alterações na mama. *UFF, Niterói*, 2013. Citado 3 vezes nas páginas 15, 24 e 25.

BOYD, N. F. et al. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Research*, BioMed Central, v. 13, n. 6, p. 223, 2011. Citado na página 15.

BRIOSCHI, M. *Saiba mais sobre a TERMOGRAFIA MÉDICA*. 2017. Disponível em: <https://www.4medic.com.br/saiba-mais-sobre-termografia-medica/>. Citado na página 29.

CHATFIELD, C. *The analysis of time series: an introduction*. [S.l.]: CRC press, 1986. Citado na página 35.

CHINO, D. Y. T. *Mineração de padrões frequentes em séries temporais para apoio à tomada de decisão em agrometereologia*. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado 3 vezes nas páginas 36, 39 e 40.

CHIU, B.; KEOGH, E.; LONARDI, S. Probabilistic discovery of time series motifs. In: ACM. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2003. p. 493–498. Citado 3 vezes nas páginas 21, 41 e 42.

CÔRTE, A. C. R.; HERNANDEZ, A. J. Termografia médica infravermelha aplicada à medicina do esporte. *Revista Brasileira de Medicina do Esporte*, SciELO Brasil, v. 22, n. 4, p. 315–319, 2016. Citado na página 29.

DAU, H. A.; KEOGH, E. Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery. In: ACM. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2017. p. 125–134. Citado na página 42.

FACON, J. *Técnicas de Processamento Digital de Imagens Aplicadas à Área da Saúde*. [S.l.]: ERI, 2006. Citado na página 31.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. *Fast subsequence matching in time-series databases*. [S.l.]: ACM, 1994. v. 23. Citado na página 39.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 45.

FITZPATRICK, J. M.; HILL, D. L. G.; R.MAURER, C. Chapter 8: Image registration. In: SONKA, J. M. F. M. (Ed.). *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis*. Bellingham, Washington USA: Spie Press, 2000. cap. 8, p. 447–513. Citado na página 34.

FORESTIER, G. et al. Generating synthetic time series to augment sparse datasets. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. [S.l.: s.n.], 2016. Citado na página 57.

FRANKS, L. Introdução à biologia celular e molecular do câncer. In: *Introdução à biologia celular e molecular do câncer*. [S.l.]: Roca Ltda, 1990. Citado 2 vezes nas páginas 23 e 24.

GHARGHABI, S. et al. Matrix profile viii: Domain agnostic online semantic segmentation at superhuman performance levels. In: IEEE. *2017 IEEE International Conference on Data Mining (ICDM)*. [S.l.], 2017. p. 117–126. Citado na página 42.

GIANNOTTI, D. G. *Sua Saúde Ressonância magnética das mamas mostra lesões não identificadas em outros exames*. 2016. Disponível em: <https://hospitalsiriolibanes.org.br/sua-saude/Paginas/ressonancia-magnetica-das-mamas-mostra-lesoes-nao-identificadas-outros-exames.aspx>. Citado na página 28.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2008. ISBN 013168728X. Citado 4 vezes nas páginas 30, 31, 32 e 33.

HOLDORF, H. H. *Breast Sonography Lecture 4 Introduction and Instrumentation*. 2017. Disponível em: <https://hholdorf.com/2017/03/06/breast-sonography-lecture-4-introduction-and-instrumentation/>. Citado na página 28.

HYNDMAN, R. J. Moving averages. In: _____. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 866–869. ISBN 978-3-642-04898-2. Disponível em: <https://doi.org/10.1007/978-3-642-04898-2_380>. Citado na página 36.

INCA. *ESTIMATIVA | 2016 Incidência de Câncer no Brasil*. 2016. Disponível em: <http://www.inca.gov.br/estimativa/2016/tabelaestados.asp?UF=MA>. Citado na página 14.

INCA. *INCA ratifica recomendações de faixa etária para início da mamografia*. 2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2016/inca-ratifica-recomendacoes-de-faixa-etaria-para-inicio-da-mamografia>. Citado na página 15.

INCA. *Controle do Câncer de Mama Detecção Precoce*. 2017. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_controle_cancer_mama/deteccao_precoce>. Citado 2 vezes nas páginas 14 e 25.

IZAKIAN, H.; PEDRYCZ, W. Anomaly detection in time series data using a fuzzy c-means clustering. In: IEEE. *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*. [S.l.], 2013. p. 1513–1518. Citado na página 21.

JAHNE, B. *Digital Image Processing: Concepts, Algorithms, and Scientific Applications*. 4th. ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997. ISBN 3540627243. Citado 2 vezes nas páginas 31 e 33.

JÄHNE, B.; HAUSSECKER, H.; GEISSLER, P. *Handbook of computer vision and applications*. [S.l.]: Academic Press San Diego, 1999. v. 2. Citado na página 33.

KANDLIKAR, S. G. et al. Infrared imaging technology for breast cancer detection–current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, Elsevier, v. 108, p. 2303–2320, 2017. Citado na página 30.

KEOGH, E. et al. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, v. 3, n. 3, p. 263–286, Aug 2001. ISSN 0219-1377. Disponível em: <https://doi.org/10.1007/PL00011669>. Citado na página 21.

KEOGH, E.; KASETTY, S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, Springer, v. 7, n. 4, p. 349–371, 2003. Citado na página 39.

KEOGH, E.; LIN, J.; FU, A. Hot sax: Finding the most unusual time series subsequence: Algorithms and applications. In: *Proc. of the 5th IEEE Int'l Conf. on Data Mining*. [S.l.: s.n.], 2004. p. 440–449. Citado 2 vezes nas páginas 21 e 42.

KEOGH, E. J.; SMYTH, P. A probabilistic approach to fast pattern matching in time series databases. In: *Kdd*. [S.l.: s.n.], 1997. v. 1997, p. 24–30. Citado na página 21.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 53.

LIN, J. et al. A symbolic representation of time series, with implications for streaming algorithms. In: ACM. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. [S.l.], 2003. p. 2–11. Citado 3 vezes nas páginas 21, 41 e 42.

LOIZZI, V. et al. Biological pathways involved in tumor angiogenesis and bevacizumab based anti-angiogenic therapy with special references to ovarian cancer. *International journal of molecular sciences*, Multidisciplinary Digital Publishing Institute, v. 18, n. 9, p. 1967, 2017. Citado na página 25.

LONARDI, J.; PATEL, P. Finding motifs in time series. In: *Proc. of the 2nd Workshop on Temporal Data Mining*. [S.l.: s.n.], 2002. p. 53–68. Citado na página 41.

MALETZKE, A. G. *Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características*. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado 6 vezes nas páginas 21, 36, 37, 38, 42 e 51.

MALETZKE, A. G. et al. Time series classification using motifs and characteristics extraction: A case study on ecg databases. In: ATLANTIS PRESS. *Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support*. [S.l.], 2013. Citado 3 vezes nas páginas 22, 41 e 51.

MANDAL, A. *What is RNA?* 2016. Disponível em: <https://www.news-medical.net/life-sciences/What-is-RNA.aspx>. Citado na página 23.

MÖRCHEN, F. *Time series knowledge mining*. [S.l.]: Görich & Weiershäuser, 2006. Citado na página 42.

MOREIRA, C. *Learning to rank academic experts*. Tese (Doutorado) — Master Thesis, Technical University of Lisbon, 2011. Citado 2 vezes nas páginas 44 e 45.

NUNES, F. L. Introdução ao processamento de imagens médicas para auxílio a diagnóstico–uma visão prática. *Livro das Jornadas de Atualizações em Informática*, p. 73–126, 2006. Citado na página 14.

OHASHI, Y.; UCHIDA, I. Applying dynamic thermography in the diagnosis of breast cancer. *IEEE Engineering in Medicine and Biology Magazine*, IEEE, v. 19, n. 3, p. 42–51, 2000. Citado 2 vezes nas páginas 16 e 20.

ONCOGUIA. *A Mama*. 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/a-mama/748/12/>. Citado na página 24.

ONCOGUIA. *O que é Câncer*. 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer/12/1/>. Citado na página 23.

ONCOGUIA. *Tipos de Câncer de Mama*. 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/tipos-de-cancer-de-mama/1382/34/>. Citado 2 vezes nas páginas 20 e 25.

ONCOGUIA. *Ultrassom das Mamas*. 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/ultrassom-das-mamas/1392/264/>. Citado na página 27.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 52.

PERNG, C.-S. et al. Landmarks: a new model for similarity-based pattern querying in time series databases. In: IEEE. *Data Engineering, 2000. Proceedings. 16th International Conference on*. [S.l.], 2000. p. 33–42. Citado na página 21.

PIARDI, L. et al. *Tudo sobre CICLO CELULAR*. 2016. Disponível em: <http://www.ciclocelular.com.br>. Citado na página 23.

PINHEIRO, P. *ENTENDA A CLASSIFICAÇÃO BI-RADS DA MAMOGRAFIA*. 2017. Disponível em: <https://www.mdsaude.com/2016/12/classificacao-bi-rads.html>. Citado na página 26.

PUPIN, J. R.; SILVA, K. S.; CARBONE, V. L. Introdução às séries e transformadas de fourier e aplicações no processamento de sinais e imagens. *Trabalho de conclusão de curso–Centro de Ciências Exatas e de Tecnologia-Universidade Federal de São Carlos*, 2010. Citado na página 40.

QUEIROZ, A. C. S. d. *Extração e Representação de Conhecimento de Séries Temporais de Demanda de Energia Elétrica Usando TSKR*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2012. Citado na página 31.

RESMINI, R. et al. Diagnóstico precoce de doenças mamárias usando imagens térmicas e aprendizado de máquina. *REAVI-Revista Eletrônica do Alto Vale do Itajaí*, v. 1, n. 1, p. 55–67, 2012. Citado na página 20.

SEDICIAS, S. *Exames que detectam e confirmam o câncer de mama*. 2017. Disponível em: <https://www.tuasaude.com/exame-para-cancer-de-mama/>. Citado na página 25.

SERPEJANTE, C. *Mamografia: exame detecta o câncer de mama*. 2014. Disponível em: <http://www.minhavida.com.br/saude/tudo-sobre/16864-mamografia>. Citado na página 26.

SILVA, D. F. *Classificação de séries temporais por similaridade e extração de atributos com aplicação na identificação automática de insetos*. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado na página 40.

SILVA, L. F. da. *Uma Análise Híbrida para Detecção de Anomalias da Mama usando Séries Temporais de Temperatura*. Tese (Doutorado) — Universidade Federal Fluminense, 8 2015. Citado 9 vezes nas páginas 17, 22, 30, 34, 47, 48, 49, 50 e 55.

SILVA, T. A. E. da. *Uma Metodologia de Auxílio ao Diagnóstico de Doenças de Mama a Partir de Termografias Dinâmicas*. Tese (Doutorado) — Universidade Federal Fluminense, 11 2015. Citado 2 vezes nas páginas 22 e 55.

SILVESTRE, J. R. *Especial Capa - Câncer: O imperador de todos os males*. 2016. Disponível em: <http://rsaude.com.br/criciuma/materia/especial-capa-cancer-o-imperador-de-todos-os-males/9581>. Citado na página 14.

STERNS, E. E. et al. Thermography: its relation to pathologic characteristics, vascularity, proliferation rate, and survival of patients with invasive ductal carcinoma of the breast. *Cancer*, Wiley Online Library, v. 77, n. 7, p. 1324–1328, 1996. Citado na página 20.

TAN, T. Z. et al. A novel cognitive interpretation of breast cancer thermography with complementary learning fuzzy neural memory structure. *Expert Systems with Applications*, Elsevier, v. 33, n. 3, p. 652–666, 2007. Citado na página 20.

VAPNIK, V. N.; VAPNIK, V. *Statistical learning theory*. [S.l.]: Wiley New York, 1998. v. 1. Citado na página 44.

VILLELA, F. *Brasileiras fazem menos mamografias do que é recomendado*. 2014. Disponível em: <http://atarde.uol.com.br/brasil/noticias/1649098-brasileiras-fazem-menos-mamografias-do-que-e-recomendado>. Citado na página 27.

WEI, L.; KEOGH, E.; XI, X. Saxually explicit images: Finding unusual shapes. In: IEEE. *Data Mining, 2006. ICDM'06. Sixth International Conference on*. [S.l.], 2006. p. 711–720. Citado na página 21.

WOLFF, M. L. et al. *Avaliação de Tecnologias em Saúde Sumário das Evidências e Recomendações sobre o uso da Termografia no Diagnóstico de Câncer de Mama*. 2012. Disponível em: <https://www.unimed.coop.br/documents/2159147/2162831/2012-termografia-no-diagnostico-do-cancer-de-mama.pdf>. Citado 2 vezes nas páginas 15 e 29.

YEH, C.-C. M.; HERLE, H. V.; KEOGH, E. Matrix profile iii: The matrix profile allows visualization of salient subsequences in massive time series. In: IEEE. *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. [S.l.], 2016. p. 579–588. Citado na página 42.

YEH, C.-C. M. et al. Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In: IEEE. *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. [S.l.], 2016. p. 1317–1322. Citado 3 vezes nas páginas 42, 43 e 44.

YEH, C.-C. M. et al. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, Springer, p. 1–41, 2017. Citado 2 vezes nas páginas 43 e 44.

YEKKEHKHANY, B. et al. A comparison study of different kernel functions for svm-based classification of multi-temporal polarimetry sar data. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, v. 40, n. 2, p. 281, 2014. Citado na página 45.