UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

HUGO ARAÚJO SOUZA

Classificação da Marcha em Parkinsonianos: Análise dos Algoritmos de Aprendizagem Supervisionada

Maceió
2017

Hugo Araújo Souza

Classificação da Marcha em Parkinsonianos: Análise dos Algoritmos de Aprendizagem Supervisionada

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof.Dr. Marcelo Costa Oliveira
Coorientador: Prof.Dr. Leonardo Melo de Medeiros

Maceió
2017

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário Responsável: Valter dos Santos Andrade

S729c Souza, Hugo Araújo.
Classificação da marcha em pakinsonianos: análise dos algoritmos de aprendizagem supervisionada / Hugo Araújo Souza. – 2017.
86 f.: il.

Orientador: Marcelo Costa Oliveira.
Coorientador: Leonardo Melo de Medeiros.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió, 2017.

Bibliografia: f. 68-76.
Apêndices: f. 77-86.

1. Aprendizagem supervisionada - Algoritmos. 2. Classificação de dados. 3. Seleção de atributos. 4. Marcha humana. 5. Doença de Parkinson. I. Título.

CDU: 004.421:616.858

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Mestrado de Hugo Araújo Souza, intitulada: "Classificação da Marcha em Parkinsonianos: Análise no Desempenho de Algoritmos de Aprendizagem Supervisionada", apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 12 de abril de 2017, às 09h00min, na Sala de Videoconferência do CEPETEC, no Instituto de Computação da UFAL.

COMISSÃO JULGADORA

Prof. Dr. Marcelo Costa Oliveira
UFAL – Instituto de Computação
Orientador

Prof. Dr. Leandro Dias da Silva
UFAL – Instituto de Computação
Examinador

Prof. Dr. Marcelo Zanchetta do Nascimento
UFU – Universidade Federal de Uberlândia
Examinador

Dedico este trabalho aos meus pais: Júlio César Bandeira de Souza e Arlenilda de Abreu Araújo Souza, à minha irmã Halina Araújo Souza.

AGRADECIMENTOS

Após o final deste trabalho, esse grande passo em minha vida e conclusão de um sonho, tenho que agradecer às pessoas que direta ou indiretamente contribuíram para sua conclusão. Primeiramente, aos meus familiares, em especial aos meus pais Arlenilda de Abreu Araújo Souza e Júlio César Bandeira de Souza, minha irmã Halina Araújo Souza, que em todo o tempo mostraram seu afeto, amor e atenção. É um agradecimento sem tamanho aos meus pais que sempre fizeram de tudo para propiciar o melhor da educação para seus filhos.

A Universidade Federal de Alagoas e ao Instituto de Computação que ofereceram infraestrutura de qualidade e um ambiente de excelência acadêmica para desenvolvimento do trabalho. Aos magníficos professores da UFAL, em especial aos do IC, pelos conhecimentos compartilhados. Em particular, ao meu orientador e professor Marcelo Costa Oliveira que sempre me atendeu quando necessitei.

Ao professor Leonardo Medeiros que compartilhou seus conhecimentos e me ajudou como co-orientador no desenvolvimento e conclusão desta dissertação, além de se mostrar um parceiro, amigo e grande incentivador durante o desenvolvimento da pesquisa e escrita.

A todo o pessoal da Secretaria de Estado da Saúde de Alagoas, em especial aos que formam a Gerência Executiva de Tecnologia da Informação, que me apoiaram constantemente na conclusão desta etapa: André Lins, Celyrio Accioly, José Alexandre Ferreira, Marcelo Dias, Marcílio Ferreira, entre muitos grandes amigos que fiz nessa instituição. E a todos aqueles que, direta ou indiretamente, colaboraram para que este trabalho chegasse a atingir aos objetivos propostos.

RESUMO

A Doença de Parkinson é a segunda doença neurodegenerativa mais prevalente em idosos, embora seu domínio e incidência variem de acordo com a idade, sexo e raça/etnia. Estudos apontam que a prevalência aumenta com a idade, tendo estimativa de 5 a 26 casos a cada 100 mil pessoas por ano, sendo de aproximadamente 1% entre os indivíduos de 65 a 69 anos e, variando de 3% a 14,3% entre os idosos acima de 85 anos. Os sinais clínicos mais comuns no processo inflamatório incluem a presença de tremor em repouso, rigidez muscular, bradicinesia e instabilidade postural. O diagnóstico da doença não é uma tarefa simples, pois sabe-se que há padrões de estágios no avanço da doença no organismo humano. Porém, muitos pacientes não seguem esse progresso devido a heterogeneidade de manifestações que podem surgir. A análise da marcha tornou-se um mecanismo quantitativo atrativo e não invasivo que pode auxiliar na detecção e monitoramento de portadores de DP. A extração de características é uma tarefa de suma importância para a qualidade dos dados a serem empregados pelos algoritmos de AM, visando como principal objetivo a redução na dimensionalidade dos dados em um processo de classificação. A partir da redução da dimensionalidade é possível identificar, principalmente, quais atributos são importantes e facilitar a visualização dos dados. Para dados relacionados à marcha humana, o propósito é detectar relevantes atributos que possam ajudar na identificação das fases do ciclo da marcha, como as fases de apoio e swing, cadência, comprimento da passada, velocidade, entre outras. Para tal, é preciso identificar e selecionar quais atributos são mais relevantes, assim como o método de classificação. Este trabalho avalia o desempenho de algoritmos de aprendizagem supervisionada na classificação das características da marcha humana em uma base de dados aberta, também identifica quais atributos são mais relevantes para o desempenho dos classificadores no auxílio à identificação de características da marcha em portadores da DP.

Palavras-chaves: Aprendizagem Supervisionada; Classificação de Dados; Seleção de Atributos; Marcha Humana; Doença de Parkinson.

ABSTRACT

Parkinson's disease is the second most prevalent neurodegenerative disease in the elderly, although its dominance and incidence vary according to age, gender and race/ethnicity. Studies indicate that the prevalence increases with age, with an estimate of 5 to 26 cases per 100,000 people per year, being approximately 1% among individuals aged 65-69 and ranging from 3% to 14.3% among the elderly over 85 years. The most common clinical signs in the inflammatory process include the presence of resting tremor, muscle stiffness, bradykinesia and postural instability. The diagnosis of the disease is not a simple task, as it is known that there are stages patterns of disease progression in the human organism. However, many patients do not follow this progress because of the heterogeneity of manifestations that may arise. The gait analysis has become an attractive and non-invasive quantitative mechanism that can aid in the detection and monitoring of PD patients. Feature extraction is a very important task for quality of the data to be used by the algorithms, aiming as main objective the reduction in the dimensionality of the data in a classification process. From the reduction of dimensionality it is possible to identify which attributes are important and to facilitate the visualization of the data. For data related to human gait, the purpose is to detect relevant attributes that may help in identifying gait cycle phases, such as support and swing phases, cadence, stride length, velocity, etc. To do this, it is necessary to identify and select which attributes are most relevant, as well as the classification method. This work evaluates the performance of supervised learning algorithms in the classification of human gait characteristics in an open database, also identifies which attributes are most relevant to the performance of the classifiers in aiding the identification of gait characteristics in PD patients.

Keywords: Machine Learning; Data Classification; Feature Selection; Human Gait; Parkinson Disease.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AM Aprendizagem de Máquina

AUC do inglês *Area Under the Courve*

CV Coeficiente de Variação

DP Doença de Parkinson

FN Falso Negativo

FP Falso Positivo

FS do inglês *Feature Selection*

IA Inteligência Artificial

k-NN do inglês *k-Nearest Neighbors*

MDF Mediana da Frequência do Sinal

MNF Média da Frequência do Sinal

PDS Poder de Densidade do Sinal

RMS Raiz do Valor Quadrático Médio

RNA Redes Neurais Artificiais

ROC do inglês *Receiver Operator Characteristic*

RPP Relação de Potência Pico-a-Média

RSS Rais Quadrada da Soma

SVM do inglês *Support Vector Machine*

VGRF do inglês *Vertical Ground Reaction Forces*

VN Verdadeiro Negativo

VP Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A Doença de Parkinson (DP) é a segunda doença neurodegenerativa mais prevalente em idosos, embora seu domínio e incidência variem de acordo com a idade, sexo e raça/etnia (TAN, 2013). É uma doença crônica e degenerativa do sistema nervoso central, sendo caracterizada, principalmente, por distúrbios motores e disfunções posturais (POSTUMA et al., 2012).

Estudos apontam que a prevalência aumenta com a idade, tendo estimativa de 5 a 26 casos a cada 100 mil pessoas por ano, sendo de aproximadamente 1% entre os indivíduos de 65 a 69 anos e, variando de 3% a 14,3% entre os idosos acima de 85 anos (WILLIS et al., 2010; MUTHANE; RAGOTHAMAN; GURURAJ, 2007; LAU; BRETELER, 2006).

Os sinais cardinais (sinais clínicos mais comuns no processo inflamatório) mais comuns na DP incluem a presença de tremor em repouso, rigidez muscular, bradicinesia e instabilidade postural. O termo bradicinesia refere-se mais especificamente à lentidão no planejamento, iniciação e execução de atos motores voluntários e automáticos, associada à dificuldade na mudança de padrões motores, comprometendo diretamente o desempenho das atividades de vida diária destes pacientes (KIM et al., 2013). Durante a locomoção o paciente com DP apresenta alterações hipocinéticas [1] caracterizadas pela redução do comprimento dos passos, da velocidade e aumento do tempo de duplo apoio [2] (STEGEMÖLLER et al., 2012). Este sintoma pode ocorrer devido ao comprometimento na programação dos movimentos ou em sua execução.

O diagnóstico de DP não é uma tarefa simples, pois sabe-se que há padrões de estágios do avanço da doença no organismo humano. Porém, muitos pacientes não seguem esse progresso devido a heterogeneidade de manifestações que podem surgir (POSTUMA et al., 2012). Nos últimos anos, o diagnóstico da doença vem avançando graças aos investimentos realizados em pesquisas e a evolução dos aparatos computacionais. Utilizando a computação, pesquisadores estão ampliando seus conhecimentos sobre a DP, influenciados por informações cada vez mais precisas, que vão desde o processamento de imagens neurológicas (SALVATORE et al., 2014; HOLZINGER; DEHMER; JURISICA, 2014), detecção de modificações genéticas (STEFL et al., 2013) e análise de anomalias da marcha (ALVAREZ-ALVAREZ; TRIVINO, 2013). A análise da marcha tornou-se um mecanismo quantitativo atrativo e não-invasivo que pode auxiliar na detecção e monitoramento de portadores de DP. O diagnóstico está sendo ampliado por meio de recursos providos por áreas da computação, como a Inteligência Artificial (IA) (DUBEY; WADHWANI;

[1] Atividade funcional diminuída.

[2] Fase terminal da impulsão pela ponta do pé detrás, e ao mesmo tempo, início do contato com o calcanhar do pé à frente.

WADHWANI, 2013; CHANG; ALBAN-HIDALGO; HSU, 2014), e na área de Processamento de Sinais (DALIRI, 2013; ZHANG et al., 2013).

As bases de dados abertas permitem pesquisas relacionadas à saúde, a exemplo da RCMD (RCMD, 2016), NAHDAP (NAHDAP, 2016) e, a utilizada neste trabalho, Physionet (PHYSIONET, 2015), enriquecendo descobertas de mais características de doenças, a exemplo da DP, apoiando aos profissionais da área no diagnóstico. Essas bases contem relevantes informações de pesquisas que vão desde sinais fisiológicos extraídos de partes do corpo humano à dados fisiológicos de doenças incuráveis como o Alzheimer. A Physionet é um grande repositório aberto disponível na internet que abrange dados complexos de estudos relacionados a fenômenos fisiológicos e patológicos do corpo humano (PHYSIONET, 2015). Porém, em muitos casos os dados não possuem informações explícitas, sendo um conjunto de variáveis que precisam ser correlacionadas, necessitando assim de técnicas capazes de realizar descobertas específicas. O desafio não é apenas extrair informações significativas a partir desses dados, mas também adquirir conhecimento, descobrir previamente manifestações desconhecidas, procurar padrões e dar sentido aos dados (MANAP; TAHIR; YASSIN, 2011).

A Aprendizagem de Máquina (AM) é uma subárea da IA que consiste na união de conceitos que concentra o desenvolvimento de algoritmos afim de quantificar relações existentes em dados e, para isso, aborda a identificação de padrões para fazer predições baseadas no passado (ZHANG; MA, 2012). As descobertas a partir de experiências do passado podem ser realizadas por meio de hipóteses na forma de uma regra ou conjunto de regras (LUGER, 2014). Para ser mais específico, uma hipótese é um conjunto de valores em um determinado domínio, quando se tratar de um conjunto de valores nominais, tem-se um problema de classificação, caso contrário, se o domínio for um conjunto infinito e ordenador de valores, tem-se um problema de regressão (FACELI, 2011).

Os dados sobre a marcha humana de parkinsonianos já foram classificados por intermédio de algoritmos de AM, a exemplo de Redes Bayesianas no trabalho de JIA et al.(2015), Máquinas de Vetor de Suporte (SVM) como classificador na pesquisa de ZHANG et al.(2013) e ALKHATIB et al.(2015) que aplicaram *k-NN*.

A extração de características é uma tarefa de suma importância para a qualidade dos dados a serem empregados pelo algoritmo de AM, visando como principal objetivo a redução na dimensionalidade dos dados em um processo de classificação (HARRINGTON, 2012). A partir da redução da dimensionalidade é possível identificar, principalmente, quais atributos são importantes e facilitar a visualização dos dados (FACELI, 2011). Para os dados relacionados à marcha humana, o propósito é detectar relevantes atributos que possam ajudar na identificação de fases do ciclo da marcha, como as fases de apoio e swing, cadência, comprimento da passada, velocidade, entre outras. Para tal, é preciso identificar e selecionar quais atributos são mais relevantes, assim como o método de classificação.

1.2 Motivação

Visto que as técnicas de seleção de atributos já foram aplicadas em algoritmos de aprendizagem supervisionada na classificação de dados e mostraram um significativo desempenho na precisão (superior a 90%) no diagnóstico de portadores de DP (OZCIFT, 2012). A SVM (PANT; KRISHNAN, 2014), o *k-NN* (ALKHATIB et al., 2015) e as Redes Neurais Artificiais (LEE; LIM, 2012), já foram utilizados em pesquisas com pacientes parkinsonianos aproveitando dados da marcha disponibilizados na base Physionet.

Contudo, nenhum dos trabalhos publicados aplica mais de um algoritmo de AM manipulando atributos da base Physionet. A combinação de diferentes algoritmos pode propiciar uma visão mais ampla das possibilidades das ferramentas de AM para a predição de mais características relacionadas à marcha de parkinsonianos (WAHID et al., 2015; TAHIR; MANAP, 2012). Para isso, existem métodos de avaliação de algoritmos de AM que podem garantir a qualidade na classificação, através de resultados como precisão, sensitividade e especificidade (FACELI, 2011).

A perspectiva é que os resultados sirvam como base para ampliar os conhecimentos sobre as limitações e deficiências nas fases do ciclo de marcha por intermédio da extração da Força Vertical de Reação do Solo (VGRF) em parkinsonianos. A DP caracteriza-se por tremor dos membros em repouso, rigidez muscular e bradicinesia (lentidão anormal dos movimentos). Com a progressão da doença, o paciente tem limitação gradativa dessa independência com prejuízos para sua autonomia (LANGA; LEVINE, 2014). Vale ressaltar que há uma expectativa que até o ano de 2030 mais de oito milhões de pessoas venham a ser portadoras de DP nas dez nações mais populosas do mundo, como Brasil, Estados Unidos e China (TAN, 2013).

Diante do cenário já exposto das aplicações já realizadas utilizando algoritmos de aprendizagem supervisionada em pesquisas na Informática em Saúde, e apesar do recente crescimento das pesquisas, a extração de características e seleção de atributos adequados é uma tarefa de alta relevância na identificação de padrões e classificação de dados, influenciando nos resultados de precisão (DUBEY; WADHWANI; WADHWANI, 2013; LEE, 2015; PANT; KRISHNAN, 2014).

1.3 Objetivos

O objetivo principal deste trabalho é através de uma análise dos algoritmos de aprendizagem supervisionada na classificação dos padrões da marcha humana, sendo extraídas características como o total da VGRF sob os pés e o *timestamp* das fases de Apoio e Swing em uma base de dados aberta, contribuir para a literatura relacionada à parkinsonianos através de uma solução que propicie a identificação dos melhores classificadores e permita o diagnóstico de anomalias na marcha.

1.3.1 Objetivo secundário

Como objetivo secundário serão avaliados atributos de estatística descritiva e processamento de sinais na classificação de padrões da marcha em portadores de DP em uma base de dados aberta. A análise destes atributos permitirá identificar quais influenciam no desempenho dos algoritmos de aprendizagem supervisionada para classificação de dados.

1.4 Organização da dissertação

Neste capítulo, foram apresentadas as considerações iniciais e os objetivos a serem alcançados com o algoritmo proposto. O restante do texto segue a seguinte organização:

- **Capítulo 2 - Fundamentação teórica:** traz os principais conceitos que envolvem a problemática da DP e seus efeitos colaterais a marcha humana, seu diagnóstico e a respeito da tecnologia para fornecer uma base teórica necessária para o entendimento dos algoritmos propostos. Fundamentos dos algoritmos de AM que serão empregados neste trabalho, assim como também as técnicas de seleção de atributos. Também, serão demonstrados quais são os trabalhos relacionados;

- **Capítulo 3 - Materiais e métodos:** descreve os algoritmos propostos, os recursos necessários para sua construção e a forma pela qual serão avaliados;

- **Capítulo 4 - Resultados e discussão:** apresenta os resultados obtidos através da aplicação dos algoritmos propostos no contexto da Doença de Parkinson e discute os resultados alcançados fazendo comparações com outros trabalhos encontrados na literatura corrente;

- **Capítulo 5 - Conclusão:** apresenta as conclusões do trabalho, as limitações e os trabalhos futuros envolvendo os algoritmos propostos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será feita uma revisão da literatura com o objetivo de contextualizar a problemática da Doença de Parkinson no mundo, assim como também o entendimento da Marcha Humana e fornecer uma base teórica suficiente para a compreensão do trabalho proposto. Ele está organizado da seguinte forma: a Seção 2.1 apresenta a problemática envolvendo os parkinsonianos e os efeitos da doença; a Seção 2.2 traz os conceitos necessários para o entendimento dos fundamentos da marcha humana e uma breve explanação sobre seus ciclo; a Seção 2.3 apresenta uma visão geral a respeitos dos conceitos de Aprendizagem de Máquina e o paradigma de aprendizado supervisionado, também descreve a importância da extração de características e como selecionar os atributos mais relevantes para a classificação; Na Seção 2.4 mostra detalhes sobre a base de dados que foi aproveitada; e, por fim, a Seção 2.5 faz uma breve revisão da literatura referenciando alguns trabalhos relacionados ao tema aqui tratado.

2.1 Doença de Parkinson

A Doença de Parkinson foi descrita a primeira vez há quase 200 anos (em 1817), pelo médico inglês James Parkinson, e está entre as doenças neurodegenerativas mais prevalentes, principalmente em idosos (CORTI; LESAGE; BRICE, 2011; TAN, 2013).

A principal característica da DP é a degeneração dos neurônios dopaminérgicos, da parte compacta da substância negra no corpo estriado (núcleo caudado e putâmen), pertencentes aos Gânglios da Base. Dessa forma, há uma redução do neurotransmissor chamado de dopamina, cuja função tanto é inibitória quanto excitatória no controle central dos movimentos. Alterações clínicas começam a surgir na progressão da doença, principalmente, quando ocorre redução dos neurônios e desregulação da dopamina, um neurotransmissor monoaminérgico (STEFL et al., 2013; DUBEY; WADHWANI; WADHWANI, 2013; POSTUMA et al., 2012).

A degeneração dopaminérgica pode comprometer a função motora, modulação de movimentos e equilíbrio. Os pacientes com DP apresentam rigidez, definida como aumento na resistência ao movimento passivo, afetando atividades motoras contralaterais ou desempenho de tarefas cognitivas (KIM et al., 2013). Esse sintoma aumenta durante o movimento, sendo originado pelo aumento do tônus muscular, podendo levar a uma alteração postural, como uma flexão do tronco, influenciando no equilíbrio do paciente. Um dos sintomas evidentes na doença é o tremor, quando presente pode ser acentuado em apenas um dos membros nos primeiros anos da doença e bilateralmente em estágios mais avançados. O tremor pode ser observado também na mandíbula, pescoço, cabeça e face, porém 30% dos sujeitos diagnosticados com a síndrome podem não apresentar (WORTH, 2013; LANGA;

LEVINE, 2014).

Além das manifestações acima referidas, existem as alterações que podem ser observadas durante a marcha destes indivíduos, tais como a diminuição da velocidade, comprimento do passo reduzido, diminuição da cadência (passos/min) (KIM et al., 2013). A idade pode ser um fator agravante na doença, pois pode acelerar os efeitos da doença e alterar funções cognitivas que podem agravar a marcha, inclusive aumentando o risco de quedas e, consequentemente, ocasionando fraturas (SHINE et al., 2013; STEGEMÖLLER et al., 2012). Há redução da amplitude de movimento das articulações do quadril, joelho, tornozelo e da rotação do tronco, assim como redução ou ausência de movimentos membros superiores durante todo o ciclo de marcha (NIEUWBOER; GILADI, 2013).

Apesar de inúmeras pesquisas sobre os mecanismos envolvidos na Doença de Parkinson, não há evidências de tratamento curativo. O tratamento é sintomático e pode ser baseado em uma abordagem farmacológica, não-farmacológica e/ou cirúrgica (SALVATORE et al., 2014; POSTUMA et al., 2012).

A seguir será descrito como funciona a marcha humana, seu ciclo e a forma de análise, em que este trabalho tem por finalidade, visando às explanações que foram apresentadas acima sobre manifestações que podem ser detectadas.

2.2 A Marcha Humana

A marcha humana pode ser definida como um processo de locomoção, quando o corpo se desloca de uma posição para outra pelo movimento rítmico e alternado do tronco e extremidades (ACKERMANN; BOGERT, 2010).

A marcha emprega uma sequência de movimentos repetitivos dos membros inferiores para movimentar o corpo à frente, enquanto mantém-se estável utilizando um dos membros como apoio. Na medida em que o corpo move-se à frente, um dos membros serve de apoio para que o outro avance pelo ar (ALVAREZ-ALVAREZ; TRIVINO, 2013; ACKERMANN; BOGERT, 2010). Dentre as habilidades fundamentais, como o correr e o saltar, a marcha humana se destaca pela sua importância e participação nas mais variadas formas do movimento humano e por meio da análise do seu comportamento dinâmico, pode-se obter dados importantes acerca desta habilidade (ALVAREZ-ALVAREZ; TRIVINO; CORDÓN, 2012).

É uma tarefa complexa e integrada que requer uma coordenação precisa dos sistema músculo-esquelético para assegurar a dinâmica esquelética correta (FEDEROLF; BOYER; ANDRIACCHI, 2013). Portanto, sua análise pode auxiliar no diagnóstico e tratamento de distúrbios de caminhada e movimento, identificação de fatores no equilíbrio e avaliação de intervenções clínicas da marcha e programas de reabilitação (ALVAREZ-ALVAREZ; TRIVINO; CORDÓN, 2012).

2.2.1 Análise da Marcha

A análise da marcha é o estudo sistemático da locomoção humana, este tipo de análise envolve a mensuração, descrição e avaliação de características quantitativas da locomoção humana. Por intermédio desta análise, as fases podem ser identificadas, assim como também os parâmetros e eventos humanos cinemáticos da marcha podem ser determinados, e as funções músculo-esqueléticas podem ser avaliadas quantitativamente (TAO et al., 2012).

O estudo das características da marcha é algo que vem sendo trabalho há certo tempo por muitos pesquisadores. Com o desenvolvimento de softwares específicos, sensores inerciais corpo-fixados, tais como acelerômetros e giroscópios, houve uma melhora substancial no monitoramento da atividade física humana. Durante a década passada uma série de ferramentas baseadas nestes sensores foram propostas para avaliar vários aspectos dos padrões de movimento (YONEYAMA et al., 2014; KE et al., 2013).

A análise da marcha pode ser amplamente dividida em duas categorias (ZENG; WANG; LI, 2014): baseada em aparência e baseada em modelos de abordagens. Os baseados em aparência usualmente referem-se sobre as sequências da marcha, sem qualquer modelo específico. Os padrões de marcha são reflexos implícitos sobre a aparência holística do andar indivíduo. Contudo, as abordagens não estão diretamente relacionadas a mecânica do andar, dinâmica da marcha ou estrutura corpora (HU et al., 2013; LEE; TAN; TAN, 2013). Em abordagens baseadas em modelos, as trajetórias em movimentos de corpos móveis são as características mais abordadas na análise da marcha. As informações obtidas refletem as características cinemáticas da maneira de andar, e é então usada para construir um modelo de reconhecimento (ZENG; WANG; LI, 2014). Além disso, o método baseado em modelos usa parâmetros de atividades específicas para identificar os padrões no ciclo da marcha, como tamanho da passada, força, velocidade, e extrair, por exemplo, característica como o duplo suporte (PREIS et al., 2012).

Há instrumentos clínicos de avaliação práticos e de fácil acesso, no entanto eles não analisam todas as variáveis da marcha, tais como as têmporo-espaciais (velocidade, cadencia, tempo de ciclo), de forma descritiva e concisa, dessa forma, diversos tipos de instrumentos de avaliação têm sido propostos com o intuito de quantificar, por exemplo, variáveis da marcha em portadores de DP e, assim, caracterizá-las tanto no pré e pós-tratamento quanto na comparação com pessoas saudáveis (TAO et al., 2012; ROIZ et al., 2011; ACKERMANN; BOGERT, 2010; WAHID et al., 2015).

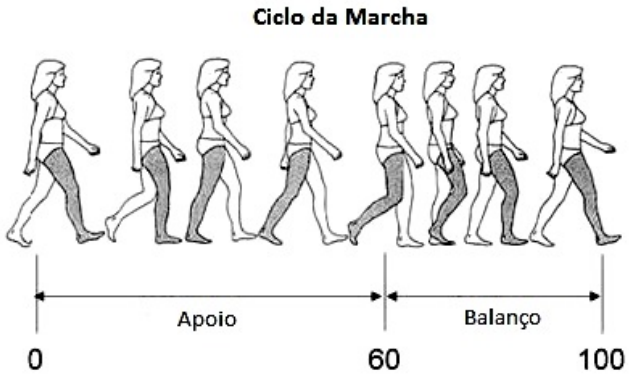
2.2.2 Ciclo da Marcha

Para o andar normal, as sequências da marcha são repetitivas e o processo quase periódico, ou seja, o sinal evolui com tempo aproximadamente repetindo sua forma e período (ABERNETHY, 2013). A maior unidade empregada na descrição da marcha é

denominada ciclo de marcha, que inicia quando o pé do membro de referência contacta a superfície de sustentação e termina quando o mesmo pé toca novamente o solo (UMBERGER, 2010; MUNIZ et al., 2010).

Cada ciclo envolve uma mudança no alinhamento entre o corpo e a base de suporte do pé durante o apoio e o deslocamento do membro no balanço. Essas reações resultam de movimentos executados pelo quadril, joelho e tornozelo, que se movem em velocidades diferentes e em arcos assincrônicos. O ciclo normal de marcha é dividido basicamente em duas fases: apoio e balanço (Figura 1). A fase de apoio inicia-se no instante em que uma extremidade entra em contato com o solo e termina quando o pé deixa o solo. Esta fase corresponde a aproximadamente 60% do ciclo de locomoção. A fase de balanço (*Swing*) inicia quando o membro inferior descola do solo e termina antes de o mesmo ter contato novamente com o solo, constituindo cerca de 40% do ciclo de marcha (MUMMOLO; MANGIALARDI; KIM, 2013; ABERNETHY, 2013).

Figura 1 – Ciclo da Marcha Humana e suas duas fases



Fonte – Adaptado de Abernethy (2013)

Quaisquer doenças neurodegenerativas e relacionadas à progressão etária, tais como DP, estão ligadas a parâmetros que permitem diagnosticar e conhecer a evolução do doente (HERRAN; GARCIA-ZAPIRAIN; MENDEZ-ZORRILLA, 2014). Especialistas avaliam a saúde dos pacientes, usando vários métodos que mensuram parâmetros que representam a marcha humana mais claramente. Herran, Garcia-Zapirain e Mendez-Zorrilla (2014) relatam que alguns destes métodos, são:

- Velocidade;
- Comprimento do passo curto (distância linear entre dois posicionamentos sucessivos do mesmo pé);
- Passo longo ou comprimento da passada (distância linear entre os posicionamentos de ambos os pés);
- Cadência (número de passos por unidade de tempo);

- Largura do passo (distância linear entre dois pontos equivalentes a ambos os pés);
- Tempo de oscilação para cada pé (tempo a partir do momento em que o pé se levanta do chão até que ele o toque novamente, para cada pé);
- Fases de marcha;
- Tempo de apoio (tempo a partir do momento em que o calcanhar toca o chão até que os dedos são levantados, para cada pé);
- Forças de Reação Vertical ao Solo.

O método que será abordado neste trabalho é a Força de Reação Vertical ao Solo, e seu funcionamento será mais bem explicada na seção a seguir.

2.2.3 Força de Reação Vertical ao Solo

O corpo humano requer um contato contínuo com o solo durante a caminhada e, assim, as Forças de Reação ao Solo são formadas como reflexos de várias forças que o corpo inteiro emite durante a marcha. A VGRF é uma unidade que possui a grande abrangência e, portanto, tem sido um tópico de interesse para muitos cientistas, certamente, quando lida com uma grande quantidade de dados (HOUCK et al., 2011; ALKHATIB et al., 2015).

A VGRF está altamente correlacionada ao crescimento ósseo e força. Seu objetivo também é examinar o poder da mecânica da marcha do exoesqueleto-assistido em caminhadas e reflexos para a quantidade de carga dada aos sujeitos sob diferentes níveis de assistência e com diferentes pesos e cadências (HOUCK et al., 2011). Por meio da VGRF é possível capturar vários parâmetros sem a necessidade primária de mensurá-los. Por exemplo, a força de reação vertical do pico vertical mostra uma relação linear com a altura de queda (POULIOT-LAFORTE et al., 2014). Além disso, são usados para diagnosticar a eficácia de cirurgias no joelho e quadril, deficiências neuromusculares, análises de risco de lesão, avaliação do risco de queda, biomecânica e assim por diante (MUNIZ et al., 2010).

Portanto, a VGRF pode gerar dados que auxiliam na investigação de padrões, anomalias e patologias relacionadas à marcha, podendo ser um importante instrumento para a sua análise (ALKHATIB et al., 2015).

Pesquisadores têm abordado vários tipos de classificadores para reduzir a dimensionalidade dos dados da marcha, além da extração de informações sobre padrões de um individuo ou de uma população. A maioria desses classificadores pode ser categorizados em técnicas de parametrização ou que analisam todo o sinal (CHESTER; TINGLEY; BIDEN, 2007; SCAFETTA; MARCHI; WEST, 2009). A seguir será apresentado o conceito de Aprendizagem de Máquina e quais algoritmos podem ajudar na classificação de dados.

2.3 Aprendizagem de Máquina

A área de Aprendizagem de Máquina lida com o estudo de métodos computacionais que permitem a programas de computadores ganhem uma melhoria, de forma autônoma, em uma determinada tarefa mediante experiências (HARRINGTON, 2012). Para isso, baseia-se em ideias de um conjunto diversificado de disciplinas incluindo IA, probabilidade e estatística, complexidade computacional, teoria da informação, psicologia, neurobiologia, teoria de controle e filosofia, sendo aplicada nas mais diversas áreas do conhecimento (TAHIR; MANAP, 2012).

Diferente das metodologias computacionais tradicionalmente abordadas, AM lida com o problemas de modo que a própria máquina irá encontrá-lo, após um processo de aprendizagem, uma hipótese que melhor o define. Para tal, empregam um princípio de inferência denominado indução, na qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos. Assim, os algoritmos de AM aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido (FACELI, 2011; HARRINGTON, 2012).

Visualizando a área de AM de uma forma bem ampla, é possível resumi-la em dois paradigmas de aprendizagem: supervisionado e não-supervisionado. Na primeira, busca-se a criação de um modelo preciso em relação à predição de valores para novos dados, enquanto que na segunda o objetivo é encontrar características que podem resumir os dados. Em ambos os casos existe uma busca por um modelo capaz de generalizar dados desconhecidos, sendo diferenciados basicamente pela existência de um rótulo (resposta) presente nos dados aproveitados na aprendizagem supervisionada (BARBER, 2012). No aprendizado não-supervisionado não existem exemplos já rotulados. O algoritmo de AM busca, a partir dos dados de entrada, criar alguma compreensão dos dados e gerar uma representação interna capaz de codificar as características de entrada em novas classes e agrupá-las corretamente (MICHALSKI; CARBONELL; MITCHELL, 2013). Adicionalmente, há outra abordagem – não utilizada neste trabalho – conhecida como aprendizado semi-supervisionado na qual existe uma tentativa de aprimorar um classificador criado a partir de dados rotulados com o uso de amostras não-rotuladas.

O aprendizado supervisionado, que é o tipo que está sendo trabalhado aqui, é realizado por intermédio de um supervisor externo, que fornece ao sistema as entradas juntamente com os valores de saída desejados (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). A ideia é que, a partir de amostras apresentadas, o sistema seja capaz de construir um classificador para rotular novos dados, desconhecidos até então. A próxima seção explicará melhor os conceitos de aprendizagem supervisionada.

2.3.1 Algoritmos de Aprendizagem Supervisionada

A aprendizagem supervisionada é composta basicamente por duas fases. A primeira é a treinamento, onde exemplos são aproveitados pelo sistema para aprendizagem e geração de um classificador. A segunda fase está relacionada a teste, quando novos exemplos são rotulados a partir do classificador existente. O conjunto de treinamento deve ser estatisticamente representativo, para que a máquina consiga reconhecer os exemplos de teste, propriedade conhecida como generalização (CARVALHO, 2001; FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

Em ocasiões onde a base de treinamento é reduzida o bastante para não permitir a generalização, ocorre o problema conhecido como *underfitting*. Em outras ocasiões, quando há ruído nos dados de treinamento, ou quando estes dados não são adequadamente representados no espaço inteiro de dados possíveis, ou até quando critérios de parada no treinamento não estão bem ajustados, acontece o *overfitting* (SIMON, 2001).

Em problemas que usam o aprendizado supervisionado, cada exemplo é descrito por um vetor de atributos de valores de características e por um especial que descreve uma característica de interessados em criar um modelo (HARRINGTON, 2012). Tais atributos podem ser discreto, ordinal ou contínuo. No caso do discreto, a problemática é conhecida como classificação, e o objetivo é identificar futuros casos em cada uma das classes pré-estabelecidas. Na hipótese em que o atributo seja contínuo, o problema é geralmente conhecido como regressão, e a finalidade é prever o valor desse atributo com base nas características dos exemplos. Já sendo do tipo ordinal, o problema é conhecido como ordenação ou regressão logística, e o propósito é ordenar um conjunto de casos de acordo com uma característica de interesse (PRATI; MONARD, 2006).

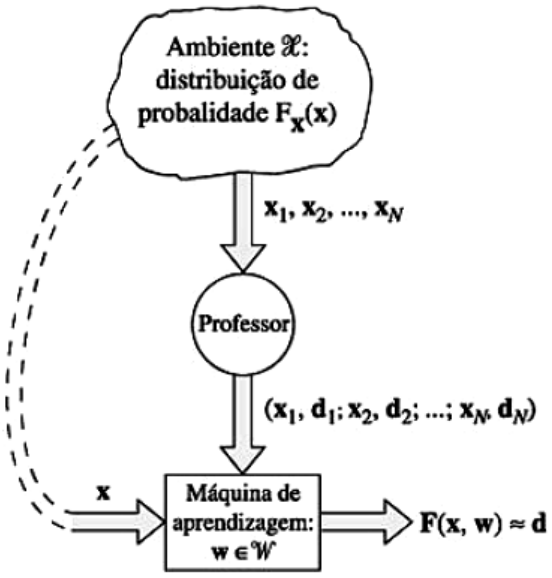
A classificação é uma tarefa constituída em análise de dados e reconhecimento dos padrões que requer construção de um classificador, ou seja, uma função que atribua uma classe a instâncias descritas por um conjunto de atributos (MICHALSKI; CARBONELL; MITCHELL, 2013). Um classificador é construído a partir da execução de um algoritmo de aprendizado sobre um determinado número de exemplos de treinamento, para os quais o rótulo de classe associado é conhecido (HAYKIN, 2000).

O aprendizado aqui empregado é do tipo induzido. A indução de algoritmos de classificação a partir de bases de dados é um problema muito recorrente no campo da AM. Trata-se do aprendizado que, baseado em alguns exemplos do conceito que está sendo estudado, extrapola um modelo para outros exemplos do mesmo conceito. Por este motivo, o algoritmo de aprendizado pode ser chamado também de indutor (FACELI, 2011).

Os conceitos referentes à geração de um classificador são representados de forma simplificada na Figura 2. Tem-se nessa figura um ambiente estacionário, na qual fornece um vetor x com uma função de probabilidade cumulativa fixa, mas desconhecida $F_x(x)$. Uma espécie de "professor" fornece uma resposta desejada d para cada vetor de entrada x recebido do ambiente, de acordo com a função de distribuição cumulativa condicional. A

máquina (algoritmo) de aprendizagem implementa um conjunto de funções para mapeamento de entrada-saída descritas por $y=F(x, w)$, onde y é a resposta produzida pelo algoritmo (HAYKIN, 2000).

Figura 2 – Modelo do processo de aprendizagem supervisionada



Fonte – Haykin (2000)

Uma vez construído um classificador, ele pode ser aplicado para predizer a classe de exemplos que siga a mesma representação e distribuição utilizada no treinamento, inclusive daqueles nunca apresentados ao indutor. Dependendo do indutor usado, o classificador pode ser simbólico, sendo um modelo cuja linguagem de descrição é equivalente a um conjunto de regras, ou seja, o classificador pode ser representado em uma linguagem proposicional ou relacional (MUSA, 2013).

Os algoritmos de aprendizagem supervisionada que foram aplicados para classificação neste trabalho, são (FRIEDMAN; HASTIE; TIBSHIRANI, 2001; FACELI, 2011; HARRINGTON, 2012; TRIPOLITI et al., 2013):

- Árvores de Decisão;
- *k-NN*;
- Redes Neurais Artificiais (RNA);
- Máquina de Vetor de Suporte (SVM).

As técnicas citadas já possuem um amplo estudo e abordagem em muitas áreas, que vão das ciências biológicas às exatas. Com a DP os algoritmos de aprendizagem supervisionada foram aplicados sob diversos sinais cardinas, dentre alguns exemplos por classificadores

podemos citar: (i) A Árvore de Decisão em pesquisas relacionadas a disfunções e atrofia do sistema nervoso central (NAIR et al., 2013); (ii) O *k-NN* aplicada à identificação de disfunções da voz em portadores da doença (SHIRVAN; TAHAMI, 2011); (iii) Já a RNA sendo aplicada para identificar possíveis padrões em disfunções do metabolismo em parkinsonianos (AHMED et al., 2009); (iv) E a SVM no uso para detecção dos sintomas de tremor e bradicinesia (COLE; OZDEMIR; NAWAB, 2012).

A seguir serão apresentados os conceitos e como funcionam os algoritmos de aprendizagem supervisionada que serão usados para classificação da marcha sob a base de dados selecionada.

2.3.1.1 Árvore de Decisão

Uma Árvore de Decisão usa a estratégia "dividir para conquistar" na resolução de um problema, a qual dividem um nó em outros nós que virão a ser seus filhos (FACELI, 2011). Sua estrutura é similar as regras do "se-então", aplicada na implementação de sistemas especialistas e em problemas de classificação. O processo é composto basicamente pela repetição de maneira recursiva em cada nó derivado até que a divisão não seja mais viável ou que a discriminação perfeita tenha sido atingida (JEGADEESHWARAN; SUGUMARAN, 2013).

É um tipo de classificador que recebe como entrada uma situação destacada por um conjunto de atributos e retorna um resultado de decisão. Seus atributos de entrada podem ser discretos ou contínuos. O uso de valores discretos na saída caracteriza um problema de classificação e o uso de contínuos trata de regressão (CHEN et al., 2014).

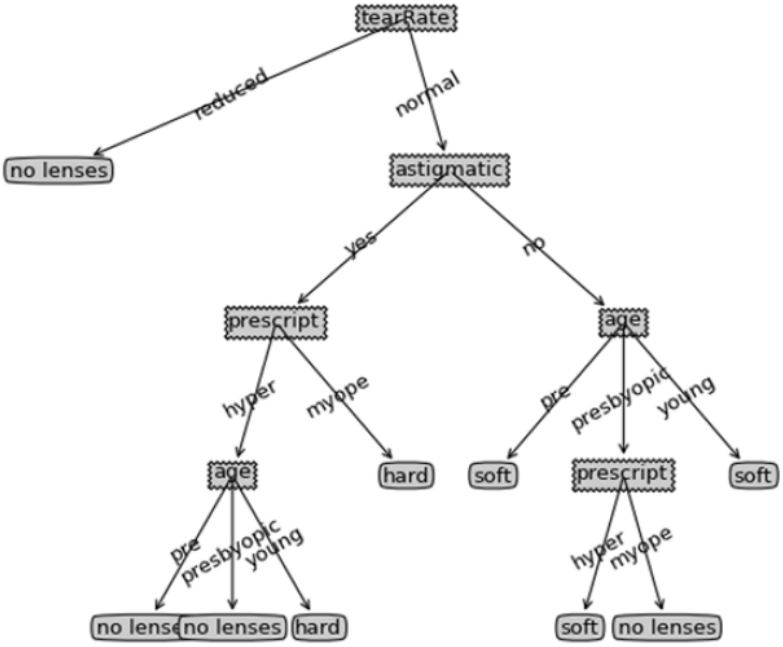
Métodos de classificação por Árvore de Decisão modelam o relacionamento entre uma entrada x_i e uma saída y_i. Cada nó interno da árvore corresponde à um atributo e cada conexão entre esse é um nó subsequente (ou nó filho), representando um valor possível para o atributo correspondente. Um nó folha representa uma saída predita para uma entrada cujos valores dos atributos são os representados pelo caminho que liga a raiz até a folha em questão. Dessa forma, um caminho na árvore pode ser interpretado como um conjunto de pares atributo-valor que leva a uma predição (FRIEDMAN; HASTIE; TIBSHIRANI, 2001; ZHANG et al., 2014).

Dentre as Árvores de Decisão mais abordadas pela comunidade para classificação existe a C4.5 que foi criada por Ross Quinlan (QUINLAN, 1996), também, como uma extensão do trabalho anterior de Quinlan, foi desenvolvida a ID3 (QUINLAN, 1986) e a CART (BURROWS et al., 1995).

O algoritmo adotado por este trabalho é o ID3, por ser o mais amplo conceitualmente, simples de aplicar e pode ser usado tanto em problemas simples de classificação como mais complexos. Para o viés indutivo, esse tipo de árvore realiza uma do tipo subida de encosta através do espaço de possíveis árvores. Em cada estágio da busca, ele examina todos os testes que poderiam ser usados para estender a árvore e escolhe o teste que ganha

a maior informação. Na Figura 3 há um exemplo de uma árvore criada para representar o número de observações baseadas nas condições dos olhos dos pacientes e no tipo de lentes de contato que o médico prescreveu (HARRINGTON, 2012).

Figura 3 – Exemplo de árvore ID3



Fonte – Harrington (2012)

Essa heurística permite que o ID3 busque eficientemente o espaço nas árvores de decisão e aborda, também, o problema da escolha de generalizações plausíveis com base em dados limitados (ROKACH; MAIMON, 2014).

Pode ocorrer casos em que a árvore combina os dados bem demais. Então, ocorre um problema é conhecido como *overfitting*. A fim de reduzir o problema de *overfitting*, pode ser realizada a ação de podar a árvore, remover algumas folhas. Se um nó da folha adiciona somente pouca informação, será cortado e fundido com outra folha (HARRINGTON, 2012).

2.3.1.2 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são modelos computacionais com capacidade de aprendizado e adaptação, podendo ser aplicadas para reconhecimento, classificação e organização de dados complexos, multivariados e não lineares (HAYKIN, 2000).

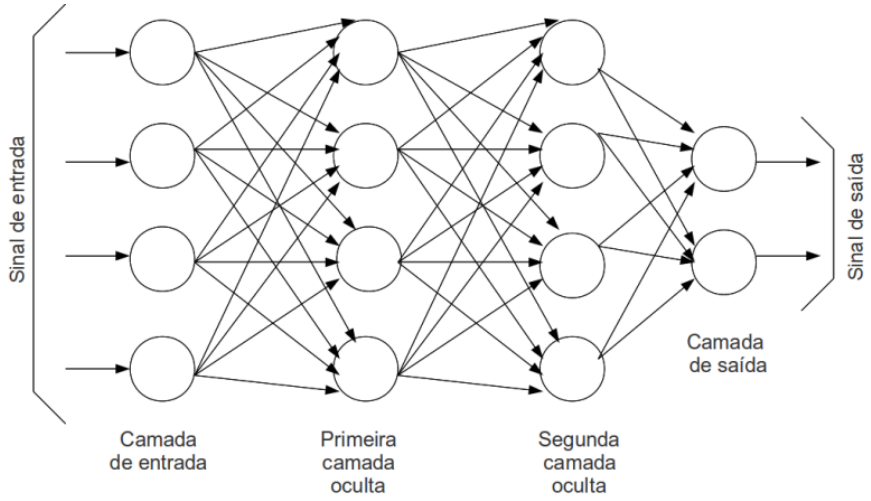
As RNAs mimetizam o comportamento do sistema nervoso, tanto que a unidade de processamento da rede é denominada neurônio (SAMANTA; AL-BALUSHI, 2003). Assim como no cérebro, os neurônios são agrupados em redes e cada neurônio recebe entradas que correspondem à excitação ou inibição de outros neurônios. Quando a rede excitada alcança um nível programado, o neurônio "dispara" (disparar, neste contexto significa propagar

estímulos). O neurônio é binário no modo de funcionamento; portanto, ou ele dispara ou não dispara (HAYKIN, 2000). Em uma RNA, os neurônios podem estar dispostos em uma ou mais camadas. Quando duas ou mais camadas são empregadas, um neurônio pode receber em seus terminais de entrada de valores de saída neurônios da camada anterior e/ou enviar seu valor de saída para terminais de entrada de neurônios da camada seguinte (FACELI, 2011).

Segundo Friedman, Hastie e Tibshirani (2001), existem inúmeros modelos de RNAs que se diferenciam, principalmente, pela estrutura da rede, pelo algoritmo de aprendizagem, pelas funções de ativação empregadas, pela recorrência, dentre outras características. Podemos citar aqui os modelos *Perceptron*, *feed-forward*, GMDH (*Group Method of Data Handling*), NSRBN (*Non Linear Sigmoidal Regression Blocks Networks*), dentre outros.

Uma Rede Neural *Feed-forward*(FF) é formada por unidades de processamento interligadas conhecidas como neurônios (ou nós), e tem a tendência natural para armazenar conhecimento experimental e torná-lo disponível para uso (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Dentre os principais tipos de redes FF, destaca-se o *multilayer perceptron* (MLP). Redes MLP consistem basicamente em múltiplas camadas, sendo uma camada de entrada, uma ou mais escondidas e uma camada de saída (Figura 4). Cada uma possui nós e cada nó está totalmente interligado por pesos com todos os nós da camada subsequente (HAYKIN, 2000).

Figura 4 – Exemplo de RNA multicamada



Fonte – Adaptada de Haykin (2000)

Cada neurônio da camada de saída está associado a uma das classes presentes no conjunto de dados. Os valores gerados pelos neurônios de saída para um dado objeto de entrada podem ser representados por um vetor $y = [y_1, y_2, ..., y_k]^t$, em que k é o número de neurônios da camada de saída (e o número de classes do problema). A rede classifica

corretamente um objeto quando o valor de saída mais elevado produzido pela rede é aquele gerado pelo neurônio de saída que corresponde à classe correta do objeto (FACELI, 2011).

A abordagem genérica para minimizar a dimensão de neurônios é por descida gradiente, chamada *Back-propagation*. Ele é constituído pela iteração de duas fases, uma fase para frente (*forward*) e uma para trás (*backward*) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). O objeto é inicialmente recebido por cada um dos neurônios da primeira camada intermediária da rede, quando é ponderado pelo peso associado a suas conexões de entrada correspondentes. Cada neurônio nessa camada aplica a função de ativação à sua entrada total e produz um valor de saída, que é usado como valor de entrada pelos neurônios seguintes (SAK; SENIOR; BEAUFAYS, 2014). Esse processo continua até que os neurônios na camada de saída produzam cada um seu valor, que é então comparado ao valor desejado para saída desse neurônio (HAYKIN, 2000). A saída de um neurônio é definida por meio de uma aplicação de uma função de ativação à entrada total. Neste trabalho foi aplicada a função sigmoidal, na qual representa uma aproximação e diferentes inclinações podem ser empregadas (YAO et al., 2012).

2.3.1.3 k-NN

O *k-Nearest Neighbors* é um tipo de classificador "lazy" onde não é computada uma função de classificação, aplicando a chamada "aprendizagem baseada em instância". Esse método de aprendizagem funciona com o armazenamento das amostras do conjunto de treino, sendo esse conjunto a chamada "instância" do problema. Quando uma nova amostra é submetida ao classificador, ele irá gerar uma resposta baseada no relacionamento da nova amostra com o conjunto de treino (MICHALSKI; CARBONELL; MITCHELL, 2013).

Segundo Theodoridis (2006), o *k-NN* é um dos algoritmos mais simples dentre as técnicas existentes em AM. Seu funcionamento pode ser descrito por meio dos seguintes passos (SOUSA et al., 2013; HARRINGTON, 2012): 1) Identifica o valor de k, ou seja, o número de vizinhos mais próximos; 2) Calcula-se a distância da nova amostra a ser classificada entre todas as amostras de treinamento; 3) São identificados os k vizinhos mais próximos, independentemente do rótulo das classes; 4) O número de vizinhos mais próximos que pertencem a cada classe do problema é contabilizado; 5) Classifica-se a nova amostra, atribuindo-lhe a classe mais constante na vizinhança.

Sendo um algoritmo baseado em instância, ele assume que todas as amostram correspondem a pontos em um espaço n-dimensional, onde n é o número de descritores usados para representar as amostras. Para classificar é reconhecida a chamada ”vizinhança” da nova amostra, isso quer dizer, os pontos já conhecidos mais próximos (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). A definição de k é trivial para estimar a proximidade entre classes. Em problemas de classificação utiliza-se geralmente a moda ponderada (FACELI, 2011): $y_t = argmax_{c \epsilon Y} \sum_{i=1}^{k} w_i I(c, y_i)$, com $w_i = \frac{1}{d(x_t, x_i)}$ e I(a, b), é uma função que

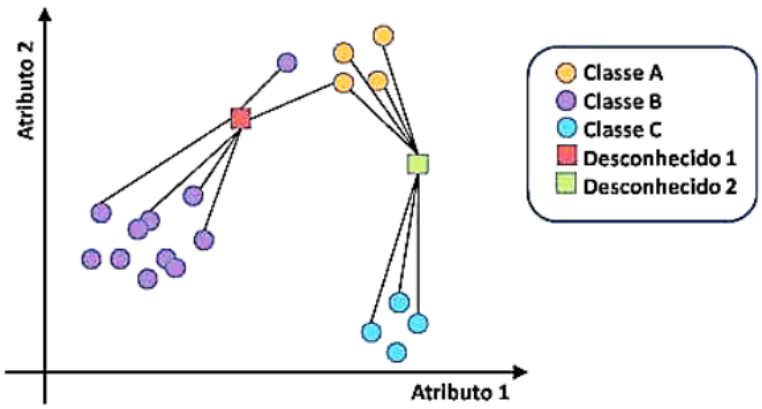
retorna 1 se a = b.

Para calcular a proximidade das amostra é possível usar medidas de distância, dentre as principais (THEODORIDIS, 2006; BACCOUR; JOHN, 2014): distância euclidiana, distância de Canberra e a distância de Chebychev. A distância de Canberra foi a função escolhida para o corrente trabalho. Trata-se de uma medida numérica da distância entre pares de pontos em um espaço vetorial, introduzido em 1966 (LANCE; WILLIAMS, 1966) e refinado em 1967 por Lance e Williams (1967). A distância de Canberra se dá entre os vetores p e q num espaço vetorial n-dimensional. A partir de $p = (p_1, p_2, ..., p_n)$ e $q = (q_1, q_2, ..., q_n)$ a equação d é assim representada (BACCOUR; JOHN, 2014):

$$d(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|} \tag{2.1}$$

Nesse contexto, por exemplo, considerando a classificação de um padrão onde o parâmetro k é igual a 1 (NN), e se o k for maior que 1, por exemplo, k= 3, sendo considerados três vizinhos do novo padrão, são analisadas as três menores distâncias do novo padrão para os padrões de treinamento. A classe que obtiver o maior número de padrões dentre essas distâncias será a classe determinante do novo padrão. A Figura 5 representa esse processo de classificação, onde a partir de um ponto desconhecido, primeiramente se tomam os k-vizinhos mais próximos dele e, dentro desse conjunto, encontra-se a classe mais significativa.

Figura 5 – Exemplo de classificação pelo método *k-NN*



Fonte – Carvalho (2001)

O *k-NN* é eficiente computacionalmente, mas bastante sensível ao número de dimensões presentes nos vetores de características, podendo aumentar de forma significativa o tempo de execução. A escolha do k é o maior influenciador nesta problemática, para otimizar basta uma busca pelo melhor valor (LUGER, 2014).

2.3.1.4 Máquina de Vetor de Suporte

A SVM foi primeiramente apresentada por Vapnik (2013) com o objetivo de resolver problemas de classificação binária de padrões. Vetores de Suporte nada mais são do que os pontos em classes que estão mais próximos do separador de classes. A determinação destes Vetores de Suporte é crucial para o estabelecimento da função separadora de classes, pois o algoritmo faz uso destes dados para gerar classificação. O número de Vetores de Suporte é, consequentemente, menor que a quantidade total de amostras de cada classe (VAPNIK, 2013).

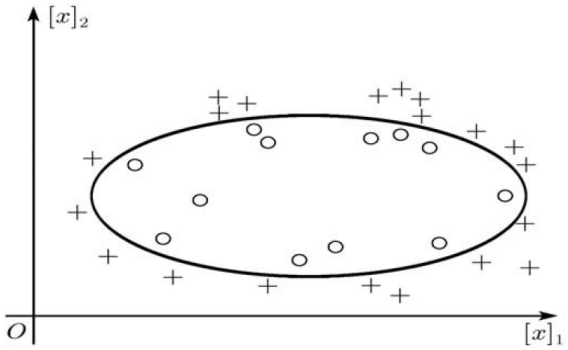
As SVMs podem ser aplicadas na resolução de problemas de classificação e regressão. Segundo Diederich (2008), uma SVM consiste basicamente em três componentes: (i) Ambiente - conjunto de vetores de entrada x; (ii) Professor - fornece para a máquina as entradas, juntamente com as saídas associadas a cada uma delas, ou seja, fornece a resposta d para cada vetor de entrada x recebido de acordo com uma função *f(x)* desconhecida; (iii) AM - capaz de implementar funções de entrada-saída da forma $y = f(x, r)$, onde y é a resposta produzida pela máquina e r é um conjunto de parâmetros usados como pesos aos valores do vetor x.

A partir da entrada de (vários pares) dados em um sistema, são realizadas classificações, onde x dados de entrada são divididos de acordo com os possíveis números de saídas y. Então, a partir de grandes amostras de dados e funções $d = f(x, r)$ que tenham comportamento determinístico, ou seja, para um certo conjunto de entrada x, e um conjunto de parâmetros r, a saída deve ser sempre a mesma. O algoritmo SVM precisa escolher uma função *f(x, r)* que seja capaz de mapear a relação de x e y, onde r são os parâmetros desta relação. As funções usadas para aprender este mapeamento são conhecidas como funções indicadoras em problemas de classificação e de funções de aproximação em problemas de regressão (FACELI, 2011).

Algoritmos SVM têm a capacidade de lidar tanto com problemas de classificação linearmente separáveis quanto não-linearmente separáveis (exemplo da Figura 3 (HARRINGTON, 2012)). No caso de problemas linearmente separáveis e em um contexto de classificação binária, por exemplo, por intermédio de SVM é possível a construção de um hiperplano ótimo entre exemplos positivos e negativos, de modo que a separação entre os exemplos seja máxima. Já os não-lineares lidam também com a classificação de conjuntos de dados linearmente separáveis ou que possuam distribuição aproximadamente linear, mas a versão de margens tolera a presença de alguns ruídos e *outliers* (DENG; TIAN; ZHANG, 2012).

Segundo Harrington (2012), a SVM seleciona o hiperplano que maximiza a margem, ou seja, maximiza a distância da margem para os dados de treinamento, de modo que para o hiperplano de separação ótimo a distância da margem para a fronteira da classe positiva é igual a distância da margem para a fronteira da classe negativa. Então, os Vetores de Suporte podem ser alcançados e extraídos.

Figura 6 – Problema não-linear



Fonte – Deng, Tian e Zhang (2012)

Contudo, o espaço de características pode ter uma dimensão muito alta (até mesmo infinita), tornando o processo altamente custoso a nível computacional (HAYKIN, 2000). Por meio de cálculos de produtos escalares entre objetos no espaço de características é possível extrair informações, e isso é obtido usando funções denominadas *kernels* (FACELI, 2011). Um função *kernel* K é uma função que recebe dois pontos x_i e x_j no espaço de entradas e calcula o produto escalar desses objetos no espaço de características:

$$K(x_i, x_j) = (x_i.x_j)^2 \tag{2.2}$$

Um *kernel* precisa satisfazer matrizes positivas semidefinidas K, em que cada elemento K_{ij} é definido por $K_{ij} = K(x_i, x_j)$, para todo $ij = 1, ..., n$. Para isso, utiliza-se *kernels* polinomiais, dentre os mais referenciados na literatura estão a função base radial (RBF), que será aplicado neste trabalho (FACELI, 2011). A função é definida da seguinte forma (FACELI, 2011):

$$exp(-\sigma||x_i - xj||^2) \tag{2.3}$$

O parâmetro σ deve ser determinado pelo usuário. Segundo Haykin (2000), a escolha de uma boa função *kernel* e seus parâmetros é muito importante para o desempenho do classificador obtido.

2.3.2 Pré-processamento

Não é incomum na etapa de coleta de dados encontrar inconsistências na base. É possível que nos conjuntos de dados haja ruídos e imperfeições, com valores incorretos, inconsistentes, duplicados ou ausentes (LUGER, 2014).

O objetivo do pré-processamento é transformar a representação dos dados afim de superar quaisquer limitações existentes nos algoritmos que serão empregados para a extração de padrões. Essa fase é de fundamental importância para a classificação de dados

tendo em vista a má qualidade presente, por vezes, na coleta e posterior geração de informação (TAN; STEINBACH; KUMAR, 2009; HOLZINGER; DEHMER; JURISICA, 2014).

De uma forma geral, pré-processamento de dados é um processo semi-automático, sendo assim entende-se que essa fase depende da capacidade de um responsável pela análise de dados com intuito de identificar os problemas presentes nos dados, além da natureza desses problemas, e usando as técnicas mais apropriados para solucionar cada um dos problemas (REDDY; REDDY; SITARAMULU, 2013).

Técnicas de pré-processamento são úteis não apenas porque podem minimizar ou eliminar problemas existentes em um conjunto de dados, mas também porque podem tornar os dados mais adequados para uso por um determinado algoritmo de AM (SALVATORE et al., 2014). Segundo Jiawei e Kamber (2001), as principais etapas envolvidas no pré-processamento, são:

- **Limpeza dos dados**: essa etapa visa eliminar os problemas como registros incompletos, valores errados e dados inconsistentes, de modo que eles não influenciem no resultado dos algoritmos aplicados. As técnicas usadas nessa etapa vão desde a remoção do registro com problemas, passando pela atribuição de valores padrões, até a aplicação de técnicas de agrupamento para auxiliar na descoberta dos melhores valores;

- **Integração dos dados**: banco de dados possuem uma heterogeneidade de fontes como arquivos textos, planilhas, *data warehouses*, vídeos e imagens, surgindo assim, a necessidade da integração destes dados de forma a termos um repositório único e consistente. Para isto, é necessária uma análise aprofundada dos dados observando redundâncias, dependências entre as variáveis e valores conflitantes como, por exemplo, categorias diferentes para os mesmos valores, chaves divergentes, regras diferentes para os mesmos dados, entre outros;

- **Redução dos dados**: as técnicas que podem ser aplicadas para que a massa de dados original seja convertida em uma menor, porém sem perder a representatividade dos dados originais. Isto permite que os algoritmos de classificação sejam executados com mais eficiência, mantendo a qualidade do resultado. As estratégias adotadas nesta etapa são a criação de estruturas otimizadas para os dados (cubos de dados), a seleção de um subconjunto dos atributos, a redução da dimensionalidade e a discretização;

- **Transformação dos dados**: alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores nominais, e nestes casos é necessário transformar os valores numéricos em nominais ou os nominais em valores numéricos. Não existe um critério único para transformação dos dados e diversas técnicas podem

ser usadas de acordo com os objetivos pretendidos. Dentre as comumente técnicas empregadas nesta etapa são a suavização, o agrupamento, a generalização, a normalização e a criação de novos atributos a partir de outros já preexistente.

Então, a partir da prática das técnicas apresentadas é possível uma melhor extração, seleção e aproveitamento dos atributos a serem aplicados posteriormente para classificação por meio dos algoritmos de AM.

2.3.2.1 Extração de Características

A extração de características consiste em técnicas para a reconhecimento de padrões em que atributos são identificados em variáveis da base de dados (FACELI, 2011). As técnicas que envolvem essa fase são, além de tudo, aplicadas com a finalidade de reduzir a dimensionalidade dos dados. Se as caraterísticas forem adequadamente escolhidas se espera que resulte em um relevante conjunto de dados a serem explorados através dos algoritmos de AM (DASGUPTA, 2015).

Para o reconhecimento de padrões de dados, mais especificamente aos que se referem a este trabalho que são os da marcha humana, são encontradas na literatura muitas técnicas, dentre as principais genéricas mais abordadas são Análise de Componentes Principais (PCA) (ZHANG et al., 2013) e Análise Linear Discriminante (LDA) (DALIRI, 2013). O uso de técnicas específicas, geralmente aplicadas na área de processamento de sinais, também pode ser uma alternativa para extração de características, a exemplo da Média da Frequência do Sinal (PHINYOMARK et al., 2012), Poder da Densidade do Sinal (ZAKNICH, 2006), Relação de Potência Pico-a-Média (GANGWAR; BHARDWAJ, 2012), transformação de Fourier (DASGUPTA, 2015) e Wavelets (LEE; LIM, 2012). Porém, o uso de simples funções matemáticas e estatísticas, como média, mediana, desvio padrão, também podem ajudar no redimensionamento de dados a serem inseridos em algoritmos de AM como classificadores (KAMATH, 2015; ELLIS; CITI; BARBIERI, 2011; ALKHATIB et al., 2015).

Além do mais, ao aplicar as técnicas descritas logo acima em dados da marcha humana é possível identificar características como as fases de swing e apoio, velocidade, tempo e tamanho da passada (ELLIS; CITI; BARBIERI, 2011). Detectando estas características da marcha, por intermédio de um classificador, é possível conhecer padrões e, por exemplo, diferenciar doentes e sadios (ZHANG et al., 2013). No entanto, podem existir bases de dados com uma grande quantidade de atributos, o que dificulta a qualidade na extração e posterior classificação. Para isso, existem técnicas que selecionam os atributos mais relevantes para a classificação.

2.3.2.2 Seleção de Atributos

A seleção de atributos (do inglês *Feature Selection* (FS)) exerce um importante papel na preparação dos dados, principalmente, em áreas como reconhecimento de padrões, AM e mineração de dados. Mediante essa técnica, por exemplo, é possível a ordenação de atributos segundo o critério de importância, redução da dimensionalidade do espaço de busca de atributos e remoção de ruídos ou outras características indesejadas (GUYON; ELISSEEFF, 2003).

Uma das principais razões para o uso das técnicas de seleção em algoritmos de AM está na manipulação de quantidades reduzidas de atributos, assim tornando o processo menos dispendioso, além da qualidade do processo de aprendizado pode ser aumentada quando se trabalha apenas com o subconjunto de atributos mais relevantes de um conjunto de dados. A identificação do subconjunto mais importante pode auxiliar no entendimento e no estudo do problema que está sendo analisado (CHANDRASHEKAR; SAHIN, 2014).

Além disso, quando um conjunto de dados armazenados tem uma grande quantidade de atributos, isto é, possui alta dimensionalidade, ocorre o aumento do processamento computacional e a diminuição da precisão, ou seja, do grau de acerto dos modelos de classificação (SAEYS; INZA; LARRAÑAGA, 2007). Para evitar os inconvenientes citados, a FS visa redimensionar a quantidade a ser processada, pois deve indicar um subconjunto de atributos mais importante para o processamento (ZAKI; JR, 2011; ZHENG et al., 2009). As consequências esperadas nessas técnicas são a geração de modelos de classificação mais precisos e a redução do processamento computacional (OZCIFT, 2012).

Os métodos de FS podem ser analisados de acordo com alguns critérios e, em particular com relação à tarefa de classificação, eles são agrupados em três tipos de abordagens, que são: *Embendedd*, *Wrapper* e *Filter* (TAN; STEINBACH; KUMAR, 2009; GUYON; ELISSEEFF, 2003).

Na técnica denominada *Embedded*, o método de seleção de atributos é incorporado e dedicado a um algoritmo de classificação específico. Em geral, o subconjunto de atributos é selecionado durante a fase de treinamento, ao longo do processo de construção do modelo de classificação. Os métodos *Embedded* usam uma medida independente para decidir quais são os melhores subconjuntos de atributos. Em seguida, usando o próprio algoritmo de classificação, seleciona o melhor subconjunto entre os melhores. Ocorrem várias iterações até que a qualidade dos resultados advindos do algoritmo classificador possa proporcionar um critério de parada natural (TAN; STEINBACH; KUMAR, 2009; GUYON; ELISSEEFF, 2003).

No tipo *Wrapper*, o método de seleção usa o algoritmo de classificação como uma "caixa-preta" para avaliar subconjuntos de atributos, de acordo com a sua capacidade preditiva. Se um subconjunto de atributos é gerado, então dois modelos de classificação são gerados: um modelo possui o conjunto original de atributos e o outro modelo usa o subconjunto selecionado de atributos. Ambos os modelos resultantes são comparados

e avaliados. Essa abordagem pode ter um custo computacional bastante elevado devido as possíveis interações do algoritmo supervisionado sob os atributos (ZAKI; JR, 2011; ZHANG et al., 2014).

Abordagem *Filter* utiliza um processo à parte, que é executado antes da aplicação do algoritmo de AM escolhido para a classificação dos dados. Essa técnica não usufrui do resultado de um algoritmo de aprendizado para definir o melhor subconjunto de atributos (SAEYS; INZA; LARRAÑAGA, 2007). Segundo Caby et al. (2011), esse método é mais rápido do que um do tipo *Wrapper* ou um do tipo *Embedded*.

A abordagem que será empregada neste trabalho é a *Wrapper*. Ela consiste basicamente em dois processos: *Forward* e *Backward*. Explicando de forma resumida (BENIWAL; ARORA, 2012): em *Forward* é escolhido primeiramente x_3 e depois um dos outros dois atributos, produzindo ordens x_3, x_1, x_2 ou x_3, x_2, x_1; no *Backward* elimina x_3 primeiro e depois uma das outras duas características, obtendo a ordens como x_1, x_2, x_3 ou x_2, x_1, x_3.

Outro fator que caracteriza um método de seleção é a forma como é gerenciada a relação entre atributos. Ela pode avaliar os atributos individualmente, de forma univariada, ou considerar relacionamentos em um subconjunto de atributos capazes de reconhecer amostras da mesma classe e distinguir amostras de classes em diferentes formas multivariadas (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013).

Os chamados métodos "multivariados" levam em conta as dependências dos atributos. Métodos multivariados conseguem resultados potencialmente melhores porque eles não simplificam suposições de independência variável/característica (GUYON; ELISSEEFF, 2003). Segundo Guyon e Elisseeff (2003), alguns dos principais, são:

- Ranking de relevância individual: provê uma visão de qual característica é relevante individualmente e quais não ajudam a proporcionar uma separação de classe melhor. Para determinadas situações a individual funciona bem quando, por exemplo, o atributo fornece uma boa separação de classe por si mesmo e assim será escolhido como prioritário;

- Atributos relevantes que são individualmente irrelevantes: esse método é avaliado o poder preditivo dos atributos que são relevantes conjuntamente e não independentemente. Um recurso útil pode ser irrelevante por si só. Duas características individualmente irrelevantes podem tornar-se relevantes quando usados em combinação;

- Atributos redundantes: esse método tem como objetivo produzir subconjuntos mais compactos. A detecção de redundâncias não pode ser feita analisando apenas as projeções de atributos, como os métodos univariados fazem;

Para a maioria dos algoritmos de AM as técnicas de FS geram três tipos de saída (resultados) (ALELYANI; TANG; LIU, 2013): (1) seleção de subconjuntos, que retorna

um subconjunto de atributos identificados pelo índice; (2) Peso do atributo, que retorna o peso correspondente a cada atributo; e o (3) híbrido, que é uma junção de subconjuntos e peso, que retorna um subconjunto classificado de atributos.

A pesagem de atributos (do inglês *Feature Weighting*) é considerada uma generalização da FS. Nesta abordagem, de forma simplificada é atribuído um valor binário a um atributo, onde 1 significa que o atributo está selecionado e 0 caso contrário. No entanto, a função de pesagem atribui um valor, geralmente no intervalo [0,1] ou [-1,1], a cada atributo. Quanto maior esse valor, mais relevante será o atributo. Contudo, para essa estimativa ser precisa foram desenvolvidos algoritmos específicos para o calculo dos pesos, dentre alguns principais: Fisher, Chi-square e Relif-F (ALELYANI; TANG; LIU, 2013).

Será usado neste trabalho é o Relif-F, por ser o mais simples e utilizado pela comunidade. Introduzido por Kira e Rendell (1992), Relief-F é uma evolução do algoritmo Relif que classifica atributos de acordo com sua maior correlação com a classe observada, levando em consideração as distâncias entre classes opostas. A ideia principal é estimar a qualidade dos atributos de acordo com o quão bem seus valores distinguem entre observações que estão próximas umas das outras (LATKOWSKI; OSOWSKI, 2015).

O algoritmo 1 é o pseudo-código do Relief-F. O Relief-F seleciona aleatoriamente uma instância R_i de observação e então procura por k de seus vizinhos mais próximos da mesma classe, chamados *hits* mais próximos H_j e também k vizinhos mais próximos de cada uma das diferentes classes, chamadas erros (*misses*) mais próximas M_j(C). Ele atualiza a estimativa de qualidade W(A) para todos os atributos A dependendo de seus valores para R_i, *hits* H_j e erros M_j(C). Se as instâncias R_i e H_j tiverem valores diferentes do atributo A, então o atributo A separa duas instâncias com a mesma classe que não é desejável. Assim, a estimativa de qualidade W(A) é diminuída. Se as instâncias R_i e M_j tiverem valores diferentes do atributo A, então este atributo separa duas instâncias de diferentes valores de classe que é desejável. Assim, a estimativa de qualidade W(A) é aumentada. O algoritmo faz a média da contribuição de todos os *hits* e erros (ROBNIK-ŠIKONJA; KONONENKO, 2003)

A maioria dos algoritmos de pesagem de atributos atribui um peso unificado (global) a cada atributo em todas as instâncias. Porém, a importância é relativa, relevância e ruído nas diferentes dimensões podem variar significativamente com a localidade dos dados. O ideal é que a seleção de atributo ou pesagem seja feita no momento da classificação (e não no treinamento), porque o conhecimento do exemplo de teste melhora a capacidade

de seleção (TANG; ALELYANI; LIU, 2014).

Algorithm 1: Pseudo-código do algoritmo Relif-F.

Definir todos os pesos $W[A] := 0.0$;
for $i = 1$ ***to*** m **do**
  Selecione aleatoriamente uma instância Ri;
  Ache k vizinhos aos hits Hj;
  for *cada classe* $C \neq classeRi$ **do**
    a partir da classe C ache k vizinhos erros $Mj(C)$;
    for $A := 1$ ***to*** a **do**
      $W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)l(m.k) +$
      $\sum_{C \neq class(R_i)}[\frac{P(C)}{1-P(class(R_i)} \sum_{j=1}^{k} diff(A, R_i, M_j(C))]/(m.k)$;
    end
  end
end

A função $diff(A, I_1, I_2)$ calcula a diferença entre os valores do atributo A para duas instâncias I_1 e I_2. Para atributos nominais ele foi originalmente definido como (ROBNIK-ŠIKONJA; KONONENKO, 2003):

$$diff(A, I_1, I_2) = \begin{cases} 0; valor(A, I_1) = valor(A, I_2) \\ 1; \text{senão} \end{cases} \tag{2.4}$$

E para atributos numéricos:

$$diff(A, I_1, I_2) = \frac{|valor(A, I_1) - valor(A, I_2)|}{max(A) - min(A)} \tag{2.5}$$

2.3.3 Validação

Neste trabalho, a técnica adotada para validação dos algoritmos de classificação foi a validação cruzada, que é usada para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). É amplamente empregada em problemas onde o objetivo da modelagem é a predição.

Basicamente busca estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados (ARLOT; CELISSE et al., 2010). O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutualmente exclusivos e, posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo (REFAEILZADEH; TANG; LIU, 2009).

Neste trabalho foram abordadas as três principais formas de particionamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

- ***Holdout*** - é reservada um subconjunto para treinamento e outro, geralmente menor, para teste.
- ***K-fold*** - subconjuntos são divididos exclusivamente sob um mesmo tamanho e um subconjunto para teste;
- ***Leave-one-out*** - compara N cálculos de erros para cada dado.

No processo de validação dos classificadores são aproveitados dados das características extraídas. Para cada característica, é possível obter desempenho com base nos valores extraídos na matriz de confusão. A matriz de confusão (Figura 7) quantifica exemplos na base de dados selecionada que são classificados bem pelo modelo construído (representado na diagonal principal), assim como também outros que são mal classificados (FACELI, 2011). É um tipo especial de tabela de contingência, com duas dimensões ("verdadeiro" e "previsto"), e conjuntos idênticos de "classes" em ambas as dimensões (cada combinação de dimensão e de classe é uma variável na matriz de confusão) (ARLOT; CELISSE et al., 2010; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). As variáveis entre as duas linhas e duas colunas informam o número de falsos positivos (FP) que corresponde ao número de exemplos cuja classe verdadeira é negativa mas que foram classificados incorretamente como pertencendo à classe positiva, falsos negativos (FN) é o número de exemplos pertencentes originalmente à classe positiva que foram incorretamente preditos como pertencentes como da classe negativa, verdadeiros positivos (VP) é o número de exemplos da classe positiva classificados corretamente e verdadeiros negativos (VN) o número de exemplos da classe negativa classificados corretamente (FACELI, 2011; WEISS, 2013).

Figura 7 – Matriz de confusão para um problema com duas classes

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

Fonte – Faceli (2011)

A partir dos resultados das formulas apresentadas é possível avaliar o sucesso na classificação do algoritmo ou técnica aplicada aos dados. A avaliação mais comum para avaliar o desempenho dos classificadores é a taxa de erros (ou acertos). As medidas de desempenho que serão utilizadas neste trabalho são (i) Precisão que é a proporção de exemplos positivos classificados corretamente entre todos aqueles preditos como positivos; (ii) Sensitividade corresponde à taxa de acerto na classe positiva; (iii) Especificidade que está relacionada à taxa de acerto na classe negativa; e a (iv) Acurácia é a proporção de acertos, ou seja, o total de verdadeiros positivos e verdadeiros negativos, em relação a n

amostra avaliada (FACELI, 2011). As fórmulas das quatro medidas de desempenho são as seguintes (LÓPEZ et al., 2012; LIN; LI, 2012):

i. **Precisão**:
$$\frac{VP}{VP + FP} \tag{2.6}$$

ii. **Sensitividade**:
$$\frac{VP}{VP + VN} \tag{2.7}$$

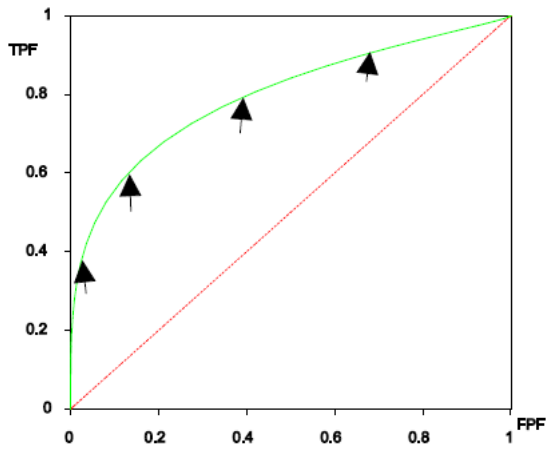
iii. **Especificidade**:
$$\frac{VN}{FP + VN} \tag{2.8}$$

iv. **Acurácia**:
$$\frac{VP + VN}{n} \tag{2.9}$$

O método mais empregado é o *k-fold*. Além de ser bem referenciada, o poder da divisão dos dados em subconjuntos e ter um custo computacional variado (FACELI, 2011; ARLOT; CELISSE et al., 2010). Contudo, outro método muito empregado pela comunidade para avaliar o desempenho de classificadores de forma visual é a curva de ROC.

As curvas ROC (*Receiver Operator Characteristic*) baseiam-se na noção de uma variável "separadora" (ou decisão). As frequências dos resultados positivos e negativos do teste de diagnóstico variam se alterar o "critério" ou "*cut-off*" para a positividade no eixo de decisão (BOLLE et al., 2013). Quando os resultados de uma classificação são avaliadas com base em julgamento subjetivo, a escala de decisão é apenas "implícita". Essa variável de decisão é freqüentemente chamada de variável "latente" ou não observável (HAJIAN-TILAKI, 2013). O gráfico de TPF (sensitividade) versus FPF (especificidade) através de diferentes *cut-offs* gera uma curva na unidade quadrada chamada curva ROC (Figura 8).

Figura 8 – Curva de ROC derivada da distribuição de duas sobreposições



Fonte – Hajian-Tilaki (2013)

A curva ROC corresponde a capacidade de discriminar progressivamente os testes de diagnóstico situados progressivamente mais perto do canto superior esquerdo no "espaço ROC". Uma curva ROC situada na linha diagonal reflete o desempenho de um teste de diagnóstico que não é melhor do que o nível de acaso, isto é, um teste que produz os resultados positivos ou negativos não relacionados com a doença (BOLLE et al., 2013). A área sob a curva (AUC) resume toda a localização da curva ROC em vez de depender de um ponto de operação específico. A AUC é uma medida eficaz e combinada de sensitividade e especificidade que descreve a validade inerente dos testes diagnósticos (HAJIAN-TILAKI, 2013).

A AUC, como um índice unidimensional, resume a localização global de toda a curva ROC. É de grande interesse, uma vez que tem uma interpretação significativa. Pode ser interpretada como a probabilidade de que um indivíduo doente seja classificado aleatoriamente ou classificado como mais susceptível de estar doente do que um indivíduo não escolhido aleatoriamente escolhido. A outra interpretação é o valor médio de sensitividade para todos os possíveis valores da especificidade (HAJIAN-TILAKI, 2013).

O valor máximo para AUC é 1 e significa que o teste de diagnóstico é perfeito na diferenciação entre os doentes e sadios. AUC = 0,5 significa a chance de discriminação da curva localizada na linha diagonal na curva ROC. A AUC mínima deve ser considerada um nível de chance, isto é, AUC = 0,5, enquanto AUC = 0 significa que o teste classifica incorretamente todos os indivíduos com doença como negativos e todos os indivíduos sadios como positivo, que é extremamente improvável que aconteça na prática clínica (HAJIAN-TILAKI, 2013).

2.4 Descrição da Base de Dados

O projeto PhysioNet é uma grande base de dados pública de gravações de sinais fisiológicos e softwares *opensource* correlacionados. Essa base foi desenvolvida pelo National Institutes of Health, por meio do National Institute of Biomedical Image and Bioengeeniring.

Foi concebido como mecanismo para a disseminação livre e aberta para estudo de sinais biomédicos gravados, assim como também para avaliação de softwares de código aberto específicos na área, propiciando mecanismos para a análise de dados e avaliação de propostas de novos algoritmos de forma colaborativa. Criado em 1999, o projeto foi dividido em três partes estritamente independentes: (i) PhysioToolkit, (ii) PhysioNetWorks e (iii) Physiobank.

i. A PhysioToolkit trata-se de um repositório de softwares para processamento e análise de sinais fisiológicos, detecção de eventos fisiologicamente significativos usando técnicas clássicas e novos métodos baseados em física estatística e dinâmica não-linear, exibição e caracterização interativa de sinais, simulação de Sinais fisioló-

gicos, avaliação quantitativa e comparação de métodos, análise de processos não-equilibrados e não-estacionários.

ii. O PhysioNetWorks é um espaço de trabalho para membros da comunidade PhysioNet que desejam disponibilizar seus projetos em andamento publicamente. Os membros do PhysioNetWorks podem visualizar títulos e resumos de projetos ativos do PhysioNetWorks (espaços de trabalho que incluem armazenamento seguro de arquivos para dados ou software que eventualmente serão contribuídos para o PhysioNet). Eles também podem se candidatar a ser revisores ou colaboradores de projetos que os interessam para acessar o arquivo do projeto em desenvolvimento.

iii. PhysioBank é uma grande e crescente base de gravações digitais bem caracterizadas por sinais fisiológicos, séries de tempo e dados relacionados a pesquisa biomédica para uso pela comunidade. Atualmente, inclui mais de 60 coleções de sinais cardiopulmonares, neurais e outros sinais biomédicos de indivíduos saudáveis e, também, de pacientes com uma variedade de condições com as principais implicações para a saúde pública, incluindo a morte súbita cardíaca, insuficiência cardíaca congestiva, epilepsia, distúrbios da marcha, apneia do sono e envelhecimento.

A PhysioBank possui um repositório para dados clínicos e outro sobre sinais, onde há dez conjuntos de dados organizados de acordo com os tipos de sinais e descrições contidas em cada base, dentre os principais:

- Multi-Parameter Databases: os sinais disponíveis variam, mas podem incluir ECG, pressão arterial invasiva contínua, respiração, saturação de oxigênio e EEG, entre outros;
- ECG Database: há também dados da *Multi-Parameter Databases*, a maioria dos quais inclui sinais de ECG;
- Interbeat (RR) Interval Databases: estes contêm anotações de batidas obtidas de gravações de ECG;
- Gait and Balance Databases: contêm séries temporais de intervalos de passadas (duração do ciclo de marcha) em forma de texto.

O repositório que foi utilizado por este trabalho foi o Gait and Balance Databases é uma coleção de estudos realizados sobre a marcha em portadores da DP (FRENKEL-TOLEDO et al., 2005; YOGEV et al., 2005; HAUSDORFF et al., 2007) (Tabela 1). Na pesquisa desenvolvida por Hausdorff et al. (2007) foi estudado um grupo de pacientes controle e parkinsoniano por meio de uma técnica chamada Estimulação Auditiva Rítmica (RAS), que melhora o comprimento da passada, mas os efeitos sobre a variabilidade de passo a passo, são desconhecidos. Então identificaram informações como a fase de swing

da marcha e avaliaram com e sem RAS, extraindo a média e desvio padrão sobre os totais da VGRF. Yogev et al. (2005) avaliam a dupla tarefa motoro-cognitiva, identificando a fase de swing e mensurando a variação de tempo da passada entre parkinsonianos e controle. Frenkel-Toledo et al. (2005) estabeleceram parâmetros denominados Velocidade de Marcha Confortável (CWS), compararam as variações do tempo na passada e fase de swing em pacientes portadores de DP e controle.

A base inclui os registros da VGRF de amostras como eles caminharam em seu habitual ritmo, quando boa parte foi auto-selecionada por aproximadamente dois minutos em terreno plano. Em cada pé foram instalados oito sensores (Ultraflex Computer Dyno Graphy, Infotronic Inc.), que mensuram força (em Newtons) como uma função relacionada à tempo. A saída de cada um destes 16 sensores foi digitalizado e gravado em 100 amostras por segundo, e os registos também incluem dois sinais que refletem a soma das oito saídas de sensor para cada um dos pés.

Tabela 1 – Detalhes das Amostras por Grupo de Pesquisa.

Grupo de Pesquisa	Parkinson	Controle	Masc.	Fem.
(FRENKEL-TOLEDO et al., 2005)	29	26	28	27
(YOGEV et al., 2005)	35	29	40	24
(HAUSDORFF et al., 2007)	29	18	30	17
Total	93	73	98	68

Fonte – Elaborada pelo autor.

Na próxima seção será demonstrado trabalhos relacionados à base de dados.

2.5 Trabalhos Relacionados

Alkhatib et al. (2015) classificaram pacientes extraindo onze atributos, alguns como média, mediana e desvio padrão, sob o total da VGRF em ambas as pernas dos pacientes. Aplicando o *k-NN* como algoritmo de classificação e foram avaliados pela curva de ROC, atingindo a acurácia máxima de 90,42%.

As Redes Bayesianas foram abordadas por Jia et al. (2015) para classificação das pisadas, extraindo a média e desvio padrão como atributos, conseguiram precisão de 82,40% e acurácia de 88,30%. Dubey, Wadhwani e Wadhwani (2013) usaram as RNAs, extraindo média do coeficiente de variação e a soma da VGRF, conseguiram sensitividade de 98,03%. Ainda, ao extraírem a média máxima e média do desvio padrão da VGRF obtiveram 94,44% de sensitividade. Lee (2015) extraiu o total da VGRF de ambas as pernas e, através da aplicação de Fourier e PCA, classificou os pacientes controle e parkinsoniano.

A SVM foi analisada em três trabalhos publicados na comunidade. Pant e Krishnan (2014) classificaram intervalos de tempo das passadas, sendo extraídas as suas médias e desvio padrão, e obtiveram a margem de erro por característica de cada amostra. Chang,

Alban-Hidalgo e Hsu (2014) usaram o algoritmo para identificar a severidade da DP por meio de quatro características, como variância do centro de pressão do pé e média das coordenadas da pisada, obtendo precisão máxima de 99,30%. Zhang et al. (2013) realizaram a extração de características que combina VGRF interpolado abaixo dos calcanhares e dedos em ambos os pés e o pé inteiro de um único membro, onde cada amostra de um ciclo de marcha é da mesma dimensão. A transformação de Fourier também foi escolhida para remover as variações entre ciclos e extrair características discriminativas. Por conseguinte, aplicaram a SVM para classificação e conseguiram 83,00% de acurácia.

Algumas teorias e técnicas de processamento de sinais foram mais usadas por outros autores, a exemplo de Dasgupta (2015) que desenvolveram um algoritmo baseado em períodos de tempo, e utilizando a técnica de transformação de Fourier para FS, extraíram as fases Apoio e Swing da marcha dos parkinsonianos e controle conseguiram o resultado máximo de 97,14% de precisão. Já Han, Ma e Zhou (2009) aplicaram o modelo Autoregressivo baseado na equação de Yule-Walker para analisar a frequência das amostras nos grupos Ga, Ju e Si da base de dados. Afsar, Tirnakli e Kurths (2016) mediante os dados em todos os dezesseis sensores localizados em ambas as pernas dos pacientes, extraíram o total de variação do tempo e o total de reação da força, sendo identificados através da entropia e alcançando o máximo de 80,00% de sensitividade. Medeiros et al. (2016) a partir dos dados do total da VGRF em ambas as pernas, foram extraídos e normalizados através da PCA e usaram como classificador a Distância Euclidiana, aplicando a validação cruzada obtiveram um total de 81,00%. No trabalho de Lee e Lim (2012) foram obtidas características da marcha, comparando valores máximos e mínimos da VGRF e, a partir da transformação de Wavelets, aplicaram as RNAs com associação lógica da função Fuzzy alcançaram o resultado máximo na precisão de 75,90% e acurácia de 77,33%.

No trabalho de Kamath (2015) foram extraídos e comparados o coeficiente de variação de séries temporais em passos iniciais e o desvio padrão da série temporal da variação de passadas dos pacientes, mas não só de parkinsonianos como também portadores de Esclerose Lateral Amiotrófica e doença de Huntington. Aplicando a curva de ROC como técnica de avaliação dos classificadores alcançou AUC igual a 0,885 e, ainda, a precisão máxima de 92,90%, sensitividade de 92,90% e especificidade de 92,70%.

Por fim, autores abordaram modelos estatísticos específicos como método de classificação. Ellis, Citi e Barbieri (2011) compararam a base selecionada por este trabalho com outra que captou a marcha de jovens (entre dezoito e vinte e nove anos), extraíram a média e o desvio padrão das passadas e cadência em ambas as pernas dos pacientes e conseguiram 88,40% de sucesso na classificação ao aplicar o teorema de Kolmogorov–Smirnov para avaliação da precisão. Geman et al. (2012) usaram ferramentas baseadas em análise linear e não linear para explorar de forma quantitativa a fase de Swing e os tempos de passadas dos pacientes.

Os trabalhos apresentados nesta seção fazem um breve resumo sobre a classificação de

pacientes portadores da DP e a extração de características, em sua maioria, usando pouco mais de dois ou três tipos de atributos. Embora existam trabalhos que conseguiram identificar características específicas da marcha, outros que se apoiaram por meio de técnicas de processamento de sinais ou ainda teoremas estatísticos, nenhum deles apresenta uma proposta onde identifique quais atributos são mais relevantes na classificação de dados por meio de algoritmos de AM. É neste ponto onde o trabalho aqui proposto se localiza, por trazer quais atributos, aplicando técnicas específicas, são mais relevantes na classificação da marcha em parkinsonianos usando algoritmos de aprendizagem supervisionada.

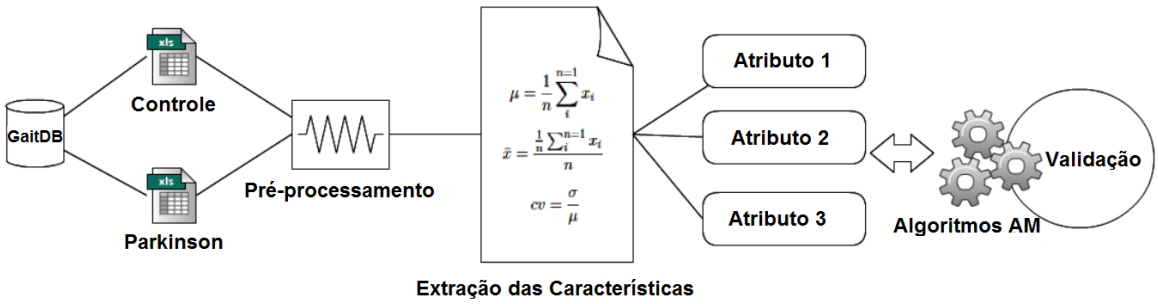
3 MATERIAIS E MÉTODOS

Este capítulo está organizado da seguinte forma: a Seção 3.1 faz uma breve explanação sobre o Processo de Classificação; a Seção 3.2 apresenta como a Base de Dados foi empregada; na Seção 3.3 define quais características do ciclo da marcha foram identificadas; a Seção 3.4 mostra quais tipos de atributos foram trabalhados para reconhecimento dos padrões das fases Apoio e Swing; Seção 3.5 explica a técnica aplicada para seleção dos atributos mais relevantes na classificação; a Seção 3.6 traz uma explicação detalhada de quais e como os algoritmos supervisionados foram usados para classificação; e, por fim, a Seção 3.7 demonstra como os algoritmos e seus resultados foram validados.

3.1 Processo de Classificação

O processo de classificação nesta pesquisa se deu seguindo as principais etapas dos trabalhos relacionados (DASGUPTA, 2015; LEE; LIM, 2012; GEMAN et al., 2012). A base de dados que serviu para este trabalho está disponível na internet através do repositório aberto chamado Physionet (PHYSIONET, 2015). Após a aquisição da base de dados, cinco etapas foram desenvolvidas para a classificação (Figura 9): Pré-processamento, Extração das Características, Seleção dos Atributos, Aplicação dos Algoritmos de AM e validação.

Figura 9 – Etapas do Processo de Classificação.



Fonte – Elaborada pelo autor.

Foi escolhida a ferramenta MATLAB neste trabalho com o intuito de facilitar as etapas de Pré-processamento (Seção 3.3) e Extração de Características (Seção 3.4). MATLAB é um software de computação científica que possui bibliotecas para o processamento de dados numéricos e identificação de padrões.

Para trabalho com as etapas Seleção dos Atributos (Seção 3.5), Aplicação dos Algoritmos de AM (Seção 3.6) e Validação (Seção 3.7) foi utilizado um software chamado Rapidminer. A Rapidminer é uma ferramenta *opensource* bem referenciada pela comu-

nidade para mineração de dados e desenvolvimento de processos analíticos preditivos. A escolha dessa ferramenta foi porque verificou-se que a MATLAB exigia muito computacionalmente, principalmente na aplicação dos algoritmos de AM com o método escolhido de seleção de atributos.

Na seção seguinte será explicada como foram identificadas as variáveis pertinentes à classificação e, em seguida, descritas as etapas do processo de classificação.

3.2 Base dos Dados

A base de dados usada neste trabalho foi descrita em detalhes na Seção 2.4. Primeiramente, foi realizada uma identificação das variáveis mais relevantes, pois a Physiobank possui dezenove tipos diferentes em cada amostra. Apenas três tipos de variáveis foram utilizadas, que foram: Tempo, total da VGRF sob o pé direito e total da VGRF sob o pé esquerdo. O por que da escolha dessas variáveis será melhor explicado na seção seguinte, Pré-processamento 3.3.

As amostras estão disponibilizadas no formato texto (TXT) e as variáveis estão dimensionadas por colunas. A primeira coluna de cada uma refere-se ao tempo (em segundos), e foi necessário extrair apenas uma como referência. Nas duas últimas colunas foram extraídos os dados do total da VGRF sob o pé direito e total da VGRF sob o pé esquerdo. Criadas duas planilhas eletrônicas (em formato XLS), a organização destas se deu em dois grupos: uma parkinsonianos e outra dos pacientes controle. A seleção das amostras foi restringida apenas as que possuem no mínimo 2 minutos de captura dos movimentos e foi realizada de forma manual. Há pacientes com mais de uma amostra, mas só foi aproveitada apenas uma amostra por paciente. Tais critérios utilizados têm como objetivo primário conseguir capturar um quantitativo máximo e uniforme de pisadas por paciente a serem classificados (ELLIS; CITI; BARBIERI, 2011; CHANG; ALBAN-HIDALGO; HSU, 2014).

O repositório Gait and Balance Databases é composto por 166 amostras ao total dos três estudos de medidas da marcha, na qual 93 pacientes são portadores da DP idiopática e 73 controles saudáveis, idade média de 65,7 anos, onde há predominância do sexo masculino, total de 98, e 68 do sexo feminino.

A partir dos critérios e passos já relatados, das 166 amostras disponíveis na base, foi reconhecido um quantitativo de 88, destes 44 são pacientes controle e 44 parkinsonianos, dentre todas por grupo de pesquisa (Tabela 2).

O grupo Hausdorff et al. (2007) teve o maior quantitativo de amostras descartadas para pacientes controle por não atender os critérios adotados na pesquisa (mínimos de 2 mins de captura), um total de 22. Para as amostras de parkinsonianos o grupo Hausdorff et al. (2007) teve apenas quatro amostras aproveitadas de um montante de 29. O maior aproveitamento foi o grupo Yogev et al. (2005), com apenas sete descartes de um total de 29 amostras, e o único que teve todas as amostras de pacientes controle aproveitadas.

Tabela 2 – Amostras utilizadas por Grupo de Pesquisa.

Grupo de Pesquisa	Parkinson	Controle
(FRENKEL-TOLEDO et al., 2005)	18	22
(YOGEV et al., 2005)	22	18
(HAUSDORFF et al., 2007)	4	4
Total	44	44

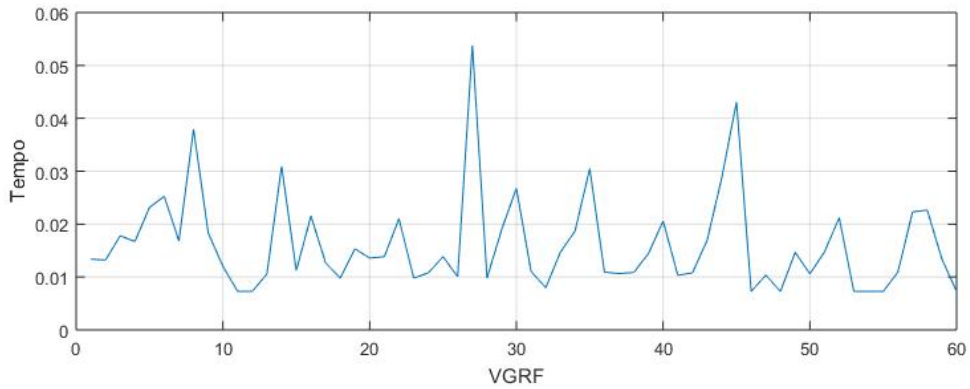
Fonte – Elaborada pelo autor.

3.3 Pré-processamento

A base Physiobank já possui uma boa qualidade nos dados, então não houve muito trabalho em identificar ruídos ou dados inconsistentes. O trabalho principal nesta fase foi identificar as fases do ciclo da marcha por meio das variáveis escolhidas.

As características extraídas dos ciclos da marcha foram as duas fases primárias: Apoio e Swing (CHANG; ALBAN-HIDALGO; HSU, 2014). Relacionando os dados do tempo e o total da VRGF em cada uma das pernas foi possível realizar o reconhecimento destas fases. Com base em estudo realizado na literatura (PANT; KRISHNAN, 2014), para detectar o momento em que o paciente manteve a perna em apoio verificou-se quando o valor da VGRF é maior do que zero e relaciona com os respectivos tempos. Já para a fase Swing verifica-se quando o valor da VGRF é igual a zero e relaciona com seus tempos. Na Figura 10 é possível visualizar um exemplo de extração da série de ciclos temporais da VGRF em uma perna de uma paciente portador de DP, onde x = VGRF e y = Tempo.

Figura 10 – Ciclo temporal da VGRF em uma perna de uma paciente portador de DP.



Fonte – Elaborada pelo autor.

Na seção seguinte, será explanado quais atributos foram extraídos a partir das características identificadas e, mais a frente, qual técnica de seleção foi adotada para identificar os mais relevantes.

3.4 Extração de Características

A partir do que foi explicado na seção 2.3.2.1, para o reconhecimento dos padrões das características foram aplicadas funções de estatística descritiva e algumas específicas de processamento de sinais para obtenção dos atributos a serem usados na classificação, já abordadas por trabalhos relacionados (ALKHATIB et al., 2015; DUBEY; WADHWANI; WADHWANI, 2013; JIA et al., 2015), são elas:

a) Média:

$$\mu = \frac{1}{n}\sum_{n=1}^{i} x_i \tag{3.1}$$

b) Mediana:

$$\bar{x} = \frac{\frac{1}{n}\sum_{n=1}^{i} x_i}{n} \tag{3.2}$$

c) Variância:

$$\sigma^2 = \frac{\sum f_i(pm_i - \overline{x})^2}{\sum f_i} \tag{3.3}$$

d) Desvio Padrão:

$$\sigma = \sqrt{\frac{\sum f_i(pm_i - \overline{x})^2}{\sum f_i}} \tag{3.4}$$

e) Coeficiente de Variação:

$$cv = \frac{\sigma}{\mu} \tag{3.5}$$

f) Curtose:

$$k = \frac{\mu^4}{\sigma^4} \tag{3.6}$$

g) Obliquidade:

$$v = \frac{(x - (\mu_3))}{\sigma^3} \tag{3.7}$$

h) Interquartil:

$$iqr = q_3 - q_1 \tag{3.8}$$

Ainda, foram aproveitadas algumas funções comumente aplicadas em estudos relacionados a processamento de sinais que utilizaram a mesma base de dados (ALKHATIB et al., 2015), já que melhor contemplam alterações fisiológicas do sinal eletromiográfico permitindo uma melhor análise, são elas:

i) Raiz Quadrada da Soma (RSS):

$$RSS = \sqrt{\sum_{n=1}^{n} xn^2} \tag{3.9}$$

Onde n é um conjunto de valores e X é um vetor (ALKHATIB et al., 2015).

j) Raiz do Valor Quadrático Médio (RMS)

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{3.10}$$

l) Poder de Densidade do Sinal (PDS):

$$PDS = \int_{-\infty}^{+\infty} f_x(\tau) e^{j2\pi\tau} d\tau \tag{3.11}$$

A equação da PDS é baseada na transformada de Fourier, onde τ é o tempo de observação, f é a frequência e dτ é a função impulso (ZAKNICH, 2006). Já a Relação de Potência Pico-a-Média (RPP) nada mais é que a relação entre picos dos sinais sobre RMS (GANGWAR; BHARDWAJ, 2012).

m) Relação de Potência Pico-a-Média (RPP):

$$RPP = \frac{|x|_{peak^2}}{x_{rms^2}} \tag{3.12}$$

Duas funções geralmente usadas para analisar a potência espectral do sinal foram utilizadas (PHINYOMARK et al., 2012). Sendo P_i a i-ésima linha do espectro da potência e M é o comprimento da frequência (essas duas variáveis fornecem algumas informações básicas sobre o espectro o sinal e suas mudanças em função do tempo):

n) Média da Frequência do Sinal (MNF):

$$MNF = \frac{\sum_{i=1}^{M} f_i P_i}{\sum_{i=1}^{M} P_i} \tag{3.13}$$

o) Mediana da Frequência do Sinal (MDF):

$$\sum_{i=1}^{MDF} P_i = \sum_{i=MDF}^{M} P_i = \frac{1}{2}\sum_{i=1}^{M} P_i \tag{3.14}$$

Foram realizadas duas formas de análise neste trabalho. Uma foi a análise das fases Apoio e Swing em apenas uma da pernas dos pacientes (a direita) afim de identificar eventuais padrões unilaterais, assim como visto nos trabalhos relacionados (ZHANG et al., 2014; DUBEY; WADHWANI; WADHWANI, 2013). Em seguida, houve uma análise das duas pernas, assim possibilitando a identificação de padrões quando o paciente usa as pernas direita e esquerda ao mesmo tempo para as fases Apoio e Swing. Por meio das funções já relacionadas foram realizadas sob um conjunto de dados relativos à ambas as pernas, ou seja, a quantidade de tipos de atributos alcançados por perna de cada paciente,

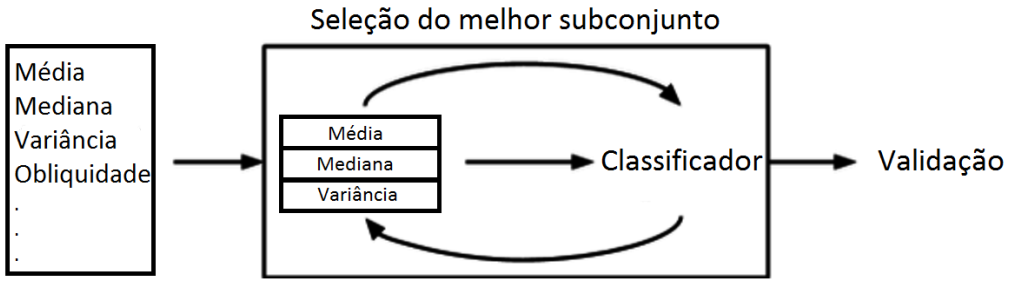
sendo assim foi lançado um quantitativo igual a 14 tipos de atributos para a perna direita e 28 para as pernas direita e esquerda. Baseado nesse quantitativo, houve a necessidade de selecionar os atributos que podem influenciar negativamente ou positivamente no desempenho dos classificadores, assim como também há a possibilidade em não ter qualquer tipo de influência. Na próxima seção será descrita a técnica usada para seleção desses atributos.

3.5 Seleção de Atributos

Reconhecida a necessidade de selecionar atributos mais importantes no contexto dos tipos de características extraídas e a possibilidade de eliminar possíveis redundâncias foi empregada uma técnica específica em FS. A técnica escolhida foi a *Wrapper* por ser de fácil entendimento e aplicação.

Por meio da *Wrapper* é possível avaliar atributos multivariados com o método *Forward /Backward* (ZAKI; JR, 2011). Ao aplicar esse método sob um conjunto vazio são adicionados atributos progressivamente originando subconjuntos, obtendo e avaliando um índice de desempenho para classificação (GUYON; ELISSEEFF, 2003). Como ilustrado de forma simplificada na Figura 11, os atributos extraídos são testados em um ciclo pra frente e/ou pra trás avaliando qual influência no desempenho dos algoritmos.

Figura 11 – Descrição do funcionamento básico do método *Wrapper*.



Fonte – Elaborada pelo autor.

Segundo Guyon e Elisseeff (2003), para *k-NN* a seleção mais adequada é a *Forward*. Neste trabalho o método foi usado da seguinte forma, sendo n a quantidade de tipos de atributos, e levando em conta que no trabalho foram extraídos 14 tipos diferentes, a seleção foi gradativamente aplicando os atributos ao algoritmo *k-NN*, com o escolhido método *Forward*, tendo x_n, a seleção se deu $x_1, x_2...x_{14}$. Lembrando que a seleção foi por subconjunto de atributos por perna sob grupos de parkinsonianos e controle. Para Árvores de Decisão os tributos também foram selecionados desta forma visto que houve sucesso no uso por Chen et al. (2014) e Zhang et al. (2014).

Para os algoritmos SVM e RNA foi aplicado de forma *Backward*, conforme recomendação de Guyon e Elisseeff (2003). A partir de n, que representa a quantidade de tipos de

atributos a seleção, foi reduzindo os atributos de 14 para apenas um quando classificada apenas uma das pernas, e 28 para apenas um quando classificada as duas pernas. Sempre aplicando aos algoritmos de classificação e os validando em seguida. Assim, o referido método com x_n atributos, a seleção se deu $x_{14}, x_{13}...x_1$.

No quesito avaliação dos atributos mais importantes que influenciam no desempenho dos algoritmos, que podem ter mais ou menos "peso" na Fase de Classificação, foi escolhida a técnica *Feature Weight* (explicado na seção 2.3.2.2). Todos os 14 tipos para uma perna e 28 para as duas pernas foram avaliados sob as classificações realizadas. Por meio de gráficos em barras foram apresentadas as incidências dos atributos na classificação de cada perna e sob as respectivas validações, ou seja, por quantas vezes quais atributos tiveram maior peso.

A seguir serão apresentados como os algoritmos de AM foram abordados para classificação e, em seguida, os métodos usados para validação.

3.6 Aplicação dos Algoritmos de AM

Após a definição dos dados a serem usados para a classificação, a etapa seguinte foi aplicar os algoritmos de aprendizagem supervisionada juntamente com a técnica em FS *Wrapper*. Para essa etapa a ferramenta RapidMiner se mostrou uma alternativa muito boa, por ser bem usada em classificação de dados por meio de IA e *Data Mining*, ademais é intuitiva e possui uma grande comunidade para suporte.

Os algoritmos de Aprendizagem Supervisionada presentes nesta pesquisa formam um total de quatro e as implementações estiveram em acordo com as principais referências encontradas. Por dez vezes os algoritmos foram aplicados sob as fases Apoio e Swing, em um primeiro momento, na perna direita e, em seguida, nas pernas direita e esquerda.

(a) **Árvores de Decisão**

Esse algoritmo foi desenvolvido para principalmente ganhar informação sob cada atributo, permitindo a amplitude e uniformidade dos seus valores. A profundidade mínima foi definida em $\frac{1}{8}$ a quantidade total de pacientes, ou seja, onze. Essa definição se deu na tentativa de manter a árvore balanceada (LUGER, 2014). A estratégia para poda teve uma confiança estabelecida em 20%. Segundo Faceli (2011), a poda é o processo mais importante na construção da árvore, pois seu principal objetivo é diminuir o ruído na classificação. O método estabelecido foi o da pré-poda, na qual conta com regras de parada que previnem a construção de ramos que não parecem melhorar a precisão preditiva da árvore (HARRINGTON, 2012). Na pré-poda foi estabelecida que a árvore há de ser constituída por no mínimo duas folhas (classes) e cada nó há de ter no mínimo um atributo.

(b) **k-NN**

O valor de k é o que determina basicamente como se dará a classificação dos dados nessa abordagem. Segundo recomendação da literatura (HARRINGTON, 2012), para um desempenho desejável neste algoritmo é preciso definir um número baixo e ímpar, assim foi determinado o valor 1. K ao receber um valor igual a um (1-NN) os atributos são classificados baseados em uma moda ponderada, como explicado na seção 2.3.1.3. A distância de Canberra (LANCE; WILLIAMS, 1967) foi a função aplicada para calcular a proximidade das amostras. A escolha se deu principalmente pelos dados extraídos serem todos números reais e ser mais sensível do que outras abordagens (BACCOUR; JOHN, 2014).

(c) **RNA**

A escolha das RNAs se deu por duas principais razões, ser algoritmo mais abordado para classificação de padrões, particularmente em grandes problemas (HAYKIN, 2000), e ser orientado a testes dimensionais de classes, o que aumenta o sucesso na classificação (SAMANTA; AL-BALUSHI, 2003). Neste trabalho a quantidade de ciclos de *back-propagation* se deu em 100x ao total de tipos de amostras. Esse múltiplo foi definido porque ao observar o trabalho de Sak, Senior e Beaufays (2014) quanto maior o número de vezes os objetos passam por treinamento maiores são os percentuais de precisão. Conforme recomendado pela literatura (LIAO, 2013), foi estabelecido um percentual em profundidade de granularidade em cerca de 20%. Haja vista, quanto maior for a quantidade de treinamento e granularidade mais é dispendioso o processo a nível computacional, então esse foi o máximo conseguido de acordo com os meios de trabalho na pesquisa. Por fim, por intermédio de uma função sigmoidal a valores dos atributos classificados foram normalizados em uma escala de -1 a +1.

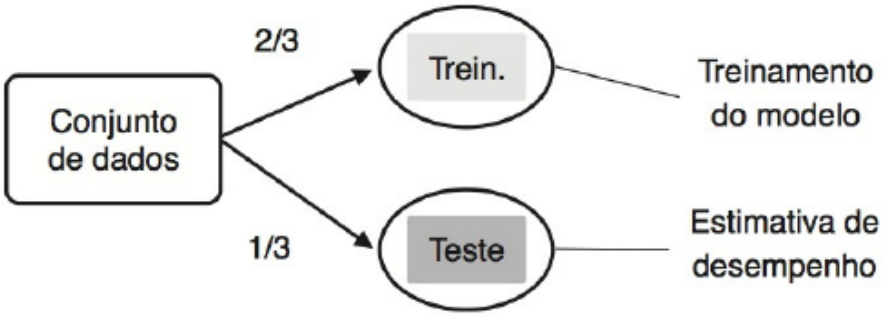
(d) **SVM**

A aplicação do algoritmo SVM exige um trabalho a mais no que se refere especificações de parâmetros do que os demais abordados neste trabalho. A RBF foi o tipo de *kernel* escolhido, como afirmado na seção 2.3.1.4. Portanto, duas variáveis foram escolhidas: o fator de escala σ e o fator de penalidade C. Foi estabelecido o valor 1 para σ e C recebeu o mesmo valor 1 para serem analisados com o conjunto de características extraídas. Foi analisada apenas essa combinação de valor σ e C tanto para o conjunto de dados treinamento como validação. As iterações também, objetivando balancear as classes a serem identificadas, os dados de C foram normalizados entre positivos e negativos.

3.7 Validação

Os métodos escolhidos para a validação dos algoritmos de classificação foram explicados na seção 2.3.3. A metodologia *Holdout* foi uma das aplicadas neste trabalho. Esse método se mostrou o mais simples de ser empregado e, conforme visto na literatura (FACELI, 2011), pouco apresenta problemas tanto em grande quantidade de dados tampouco em menores. Segundo Faceli (2011), nesse método o recomendado é reservar um subconjunto de $\frac{1}{3}$ para validação e o restante $\frac{2}{3}$ para treinamento, como mostra a Figura 12.

Figura 12 – Descrição do método *Holdout*.



Fonte – Faceli (2011)

Outra técnica foi a *k-fold*, na qual foi aplicado em uma proporção de dez subconjuntos de dados (10-*fold*) mutuamente exclusivos. A partir disto, um subconjunto é usado para teste e os k-1 restantes são para estimação dos parâmetros. Este processo foi realizado o número de vezes da quantidade de subconjuntos distintos, ou seja, dez vezes alternando de forma circular o subconjunto de teste. Segundo Friedman, Hastie e Tibshirani (2001), a validação *leave-one-out* é que um caso específico do *k-fold*, com k igual ao número total de dados n.

Ao obter a matriz de confusão, foi possível calcular as seguintes medidas de desempenho: precisão, sensitividade, especificidade e acurácia. A matriz é de extrema relevância ao trabalho, pois com a sensitividade permitiu avaliar a classificação correta dos doentes e a especificidade permitiu avaliar a classificação correta dos sadios (ARLOT; CELISSE et al., 2010). Tendo os resultados das classificações realizadas das fases Apoio e Swing em um das pernas e das duas pernas com cada algoritmo sob as três validações foram obtidas as médias e AUC da acurácia.

Também, a partir da geração da sensitividade e especificidade foram criadas as curvas ROC afim de demonstrar visualmente uma comparação do desempenho dos classificadores. Essa análise utilizando a curva ROC foi realizada sob duas formas: algoritmos de classificação quando aplicados nas fases Apoio e Swing em apenas uma das pernas; e quando aplicados as fases Apoio e Swing das pernas direita e esquerda. Esse critério

se deu para avaliar se os algoritmos possuem melhor desempenho sob dados de Apoio e Swing com uma perna ou as duas pernas.

4 RESULTADOS E DISCUSSÃO

Na classificação realizada sob os dados da perna direita na fase Apoio o algoritmo que melhor classificou foi a SVM sob a validação *k-fold*, quando alcançou 97,77% de precisão e 95,45% na acurácia, como pode ser visto na Tabela 9 no Apêndice A.

O bom desempenho com a SVM neste trabalho foi devido ao processo de indução com base nos dados de treinamento, na qual é constituída pelo princípio de minimização do risco empírico, ou seja, a minimização do erro no treinamento. Entretanto, a minimização do risco empírico nem sempre ocorre, principalmente em conjuntos de dados menores (FACELI, 2011).

Um algoritmo com desempenho semelhante foi o *k-NN* quando alcançou 95,45% em acurácia e 93,62% na precisão. A acurácia teve o mesmo resultado com a SVM, mas a precisão foi 4,15% menor. Contudo, usando a validação *holdout* foi o algoritmo que teve a pior acurácia com 31,82%. Ademais, outro algoritmo que obteve bom desempenho na classificação da fase Apoio sob a perna direita utilizando a técnica de validação *k-fold* foi a RNA. A acurácia foi apenas 1,14% inferior, e a precisão foi igual a 97,77%.

A Árvore de Decisão foi o classificador que apresentou menor precisão com 38,71% sob validação da técnica *holdout*, por conseguinte 59,06% menor do que a obtida pelo algoritmo SVM sob validação *k-fold*. Não há trabalhos relacionados à base que aplicaram esse algoritmo. O resultado na classificação da classe Parkinson com a Árvore de Decisão sob validação *holdout* teve 36,36% de acurácia, inferior se comparado aos demais algoritmos. Para esse desempenho baixo com a Árvore de Decisão em relação aos demais deve-se a um problema denominado *underfitting*. Esse problema pode acontecer quando o conjunto de treinamento é pequeno e não é significativo, ou seja, não apresenta casos que futuramente serão exercitados (ROKACH; MAIMON, 2014). Como é o caso do treinamento dos algoritmos com a técnica de validação *holdout*. Ainda, pode ocorrer também instabilidades devido a ruídos na base, produzindo grandes variações no resultado final. Também, pode acontecer que dois ou mais atributos em nós sejam classificados de forma similar e pequenas variações podem alterar o resultado, modificando todas as subárvores (FACELI, 2011).

O resultado da precisão nessa classificação chega a ser menor do que o publicado por Dasgupta (2015). No entanto, Dasgupta (2015) aplicou a transformação de Fourier como técnica de FS em um algoritmo baseado em períodos de tempo e extraiu as fases de Apoio e Swing com amostras de dados em torno de 1 min, resultando em uma precisão de 97,14% (Tabela 3). No trabalho publicado por Zhang et al. (2013), os autores utilizaram SVM sob validação da técnica transformação de Fourier e alcançaram 99,30% na precisão. Os autores aplicaram esse algoritmo sob todos os dados da pisada, não apenas nos totais da VGRF, e extraíram apenas duas características (média e variância).

Tabela 3 – Comparativo entre o melhor resultado obtido com a fase Apoio sob a perna direita e os trabalhos relacionados.

	Classificador	**Atributos**	**Resultado (Precisão)**
Referência	SVM	14	97,77%
Zhang et al. (2013)	SVM	Média e Variância	97,14%
Jia et al. (2015)	Redes Bayesianas	Média e Desvio Padrão	82,40%

A RNA e a SVM apresentaram uma maior tendência a identificar pacientes Controle, já que a maioria dos resultados da especificidade na classificação com esses algoritmos foi maior que 90,00% (Tabela 9). Os autores Kim et al. (2013) e Langa e Levine (2014) relatam sobre a dificuldade dos pacientes no equilíbrio, comprimento do passo reduzido e tremor até em membros inferiores. Esses sintomas podem ser apresentados na captura dos dados e gerar, por exemplo, *timestamp* com padrão diferente (ou não) para os pacientes parkinsonianos. Apesar de serem fatores que influenciam na geração de dados que podem identificar as duas classes (parkinsonianos e Controle), há algoritmos de aprendizagem supervisionada que não conseguem distinguir instâncias (um ou mais atributos) de cada classe (WEISS, 2013).

Todos os algoritmos apresentaram desempenho tanto na precisão como na acurácia superiores a 60,00% para a fase Swing da perna direita. Dois classificadores alcançaram 100% de precisão sob validação *k-fold*: Árvore de Decisão e SVM. O *k-NN* sob duas diferentes validações (*k-fold* e *leave-one-out*) obteve considerável precisão de 95,56% e 97,73% na acurácia (Tabela 9). Ainda, todos os classificadores apresentaram um desempenho melhor também na classificação de pacientes Controle, a exemplo da Arvore de Decisão e a SVM que alcançaram 100% de especificidade quando validados pela técnica *k-fold*. A classificação das fase Apoio e Swing foi identificado apenas no trabalho de Dasgupta (2015). Porém, o autor não apresentou o resultado por fase, classificou apenas as capturas em torno de 1 min e utilizou um algoritmo não abordado neste trabalho. Alcançou um resultado um pouco inferior na precisão que foi 97,14% (Tabela 4).

Tabela 4 – Comparativo entre o melhor resultado obtido com a fase Swing da perna direita e os trabalhos relacionados.

	Classificador	**Atributos**	**Resultado (Precisão)**
Referência	SVM	14	100%
Dasgupta (2015)	Algoritmo próprio	Tempo	97,14%
Chang,Alban-Hidalgo e Hsu (2014)	SVM	4 baseados em Fourrier	99,30%

Na avaliação de desempenho dos algoritmos na classificação com dados extraídos das fases de Apoio e Swing em apenas um das pernas dos pacientes foi utilizada também a

curva ROC (Figura 13). O *k-NN* foi o algoritmo que obteve melhor a AUC com 0,789 (σ = 0,248), conforme pode ser visto na Tabela 5. Quando a curva ROC foi aplicada a Árvore de Decisão e a SVM, os valores da AUC foram respectivamente 0,677 (σ = 0,222) e 0,667 (σ = 0,162).

Resultado superior ao obtido por Alkhatib et al. (2015) que classificaram pisadas dos paciente aplicando esse mesmo classificador sob dados de uma perna extraídos da Physionet e, o desempenho sendo avaliado pela curva ROC alcançou AUC igual a 0,611. Esse resultado se deu com a soma das variáveis dos oito sensores implantados nos pés dos pacientes, e não apenas nos totais da VGRF utilizadas neste trabalho. Kamath (2015) alcançou resultado inferior quando classificou os dados das passadas de parkinsonianos (extraídos da Physionet). O autor aplicou entropia para classificar dados da perna direita e, na avaliação com a curva ROC, obteve AUC máxima de 0,885. Vale ressaltar que Kamath (2015) teve um maior conjunto de dados, pois usou de todos os sensores implantados nos pacientes que estão disponíveis na base Physionet. A RNA obteve um desempenho inferior entre todos nessa classificação com AUC igual a 0,522 (σ = 0,159).

Tabela 5 – Desempenho na classificação das fases Apoio e Swing da perna direita com AUC e σ de cada algoritmo.

AUC	
k-NN	0,789 $\pm$ 0,248
Árvore Decisão	0,677 $\pm$ 0,222
RNA	0,522 $\pm$ 0,161
SVM	0,667 $\pm$ 0,162

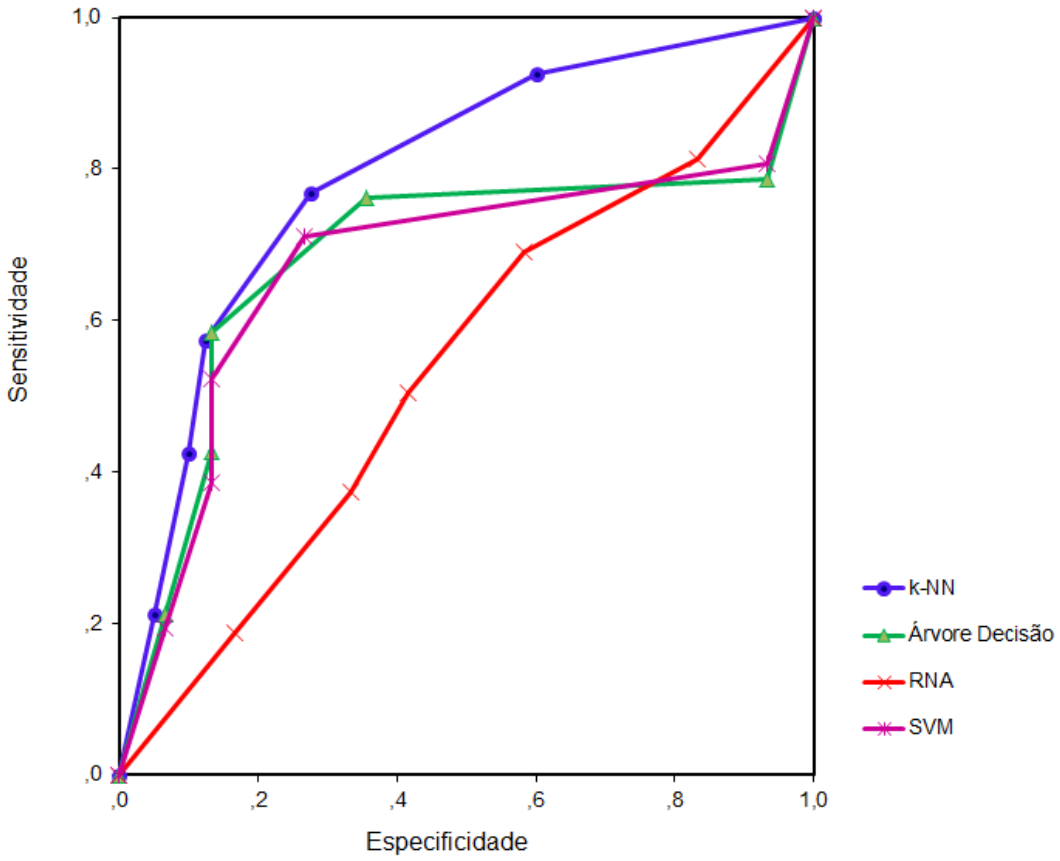
Fonte – Elaborada pelo autor.

A avaliação por peso dos atributos na classificação com algoritmos supervisionados não foi encontrada entre os trabalhos relacionados à base Physionet (seção 2.5), apesar disso alguns tipos de atributos utilizados avaliados já foram aplicados por outros autores. Jia et al. (2015) extraíram média e desvio padrão como atributo para um classificador, no caso as Redes Bayesianas. Pant e Krishnan (2014) também fizeram uso da média e desvio padrão, mas aplicaram o algoritmo SVM.

A mediana foi o atributo que por mais frequência exerceu peso com maior valor na avaliação realizada com os algoritmos e respectivas validações, quando por nove vezes foi importante para os resultados na classificação da fase Apoio sob a perna direita, como pode ser visto na Figura 14. Ao ser aplicado com Árvore de Decisão esse atributo foi relevante em todos os tipos de classificação. A obliquidade foi o segundo mais importante, pois obteve maior peso por sete vezes. Variância, MNF e RPP foram os que menos obtiveram peso para os classificadores nessa fase.

Ainda na fase Apoio de apenas uma das pernas, a média teve maior influência em quatro classificações, e desvio padrão pesou mais em seis. Em alguns casos esses atributos

Figura 13 – Avaliação desempenho dos algoritmos de aprendizagem supervisionada por meio da curva ROC na classificação das fases Apoio e Swing da perna direita.



Fonte – Elaborado pelo autor.

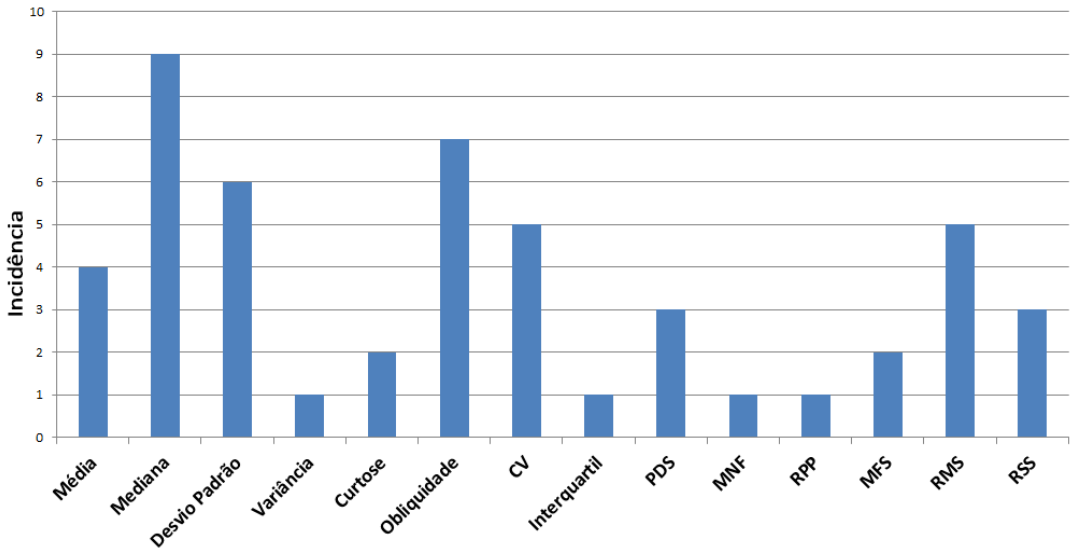
alcançaram peso igual a 1. Todos os valores dos pesos avaliados por classificador, atributo e validação na fase Apoio em apenas uma das pernas estão no Apêndice A, nas tabelas 11 à 14.

O atributo que mais influenciou em todos os classificadores na fase Swing da perna direita foi a obliquidade, quando por 11 vezes obteve o maior peso (Figura 15). Esse, quando utilizado com o *k-NN*, foi o único que exerceu maior peso em todas as classificações. Vale ressaltar que a obliquidade foi o atributo que por mais vezes obteve peso ao somar os resultados da avaliação realizada na fase de Apoio e a Swing da perna direita, um total de 18 vezes. Apesar da importância mostrada da obliquidade na classificação, apenas Alkhatib et al. (2015) fez uso desse atributo com os dados da Physionet.

Variância, curtose, interquartil, MNF e RPP foram os que menos influenciaram na classificação da fase Swing da perna direita. Além disso, esses atributos obtiveram maior peso apenas para a RNA. Os resultados dos pesos para a fase Swing em uma das pernas estão no Apêndice A, nas tabelas 15 à 18.

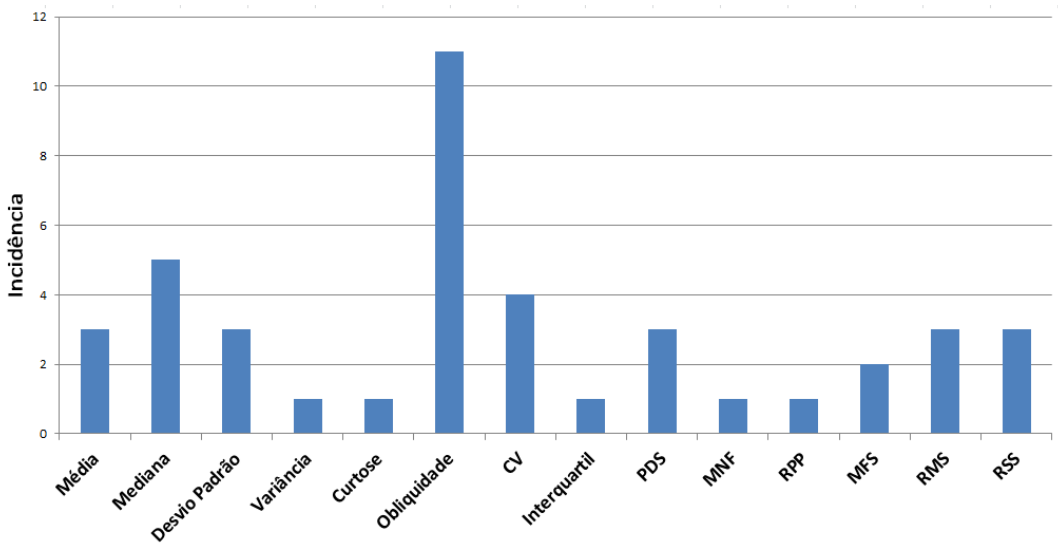
Na classificação da fase Apoio sob as pernas direita e esquerda a RNA obteve o me-

Figura 14 – Incidência dos atributos que exerceram maior peso, utilizando a técnica *Feature Weight*, sob os classificadores na fase Apoio sob a perna direita.



Fonte – Elaborado pelo autor.

Figura 15 – Incidência dos atributos que exerceram maior peso, utilizando a técnica *Feature Weight*, sob os classificadores na fase Swing da perna direita.



Fonte – Elaborado pelo autor.

lhor desempenho 97,78% de precisão e 97,73% na acurácia sob a técnica de validação *k-fold*. Os resultados detalhados da fase Apoio sob as pernas direita e esquerda estão na Tabela 10 no Apêndice A. Enquanto Dubey, Wadhwani e Wadhwani (2013) aplicaram o mesmo algoritmo sob atributos extraídos em dados das pernas direita e esquerda da base Physionet conseguiram um resultado inferior, precisão máxima em 89,47%. O resultado no trabalho de Lee e Lim (2012) também foi inferior ao utilizar a RNA com Lógica Fuzzy sob os dados da Physionet, quando obtiveram 75,90% de precisão. Vale ressaltar que esses autores utilizaram todos os dados dos pacientes (Tabela 6).

Tabela 6 – Comparativo entre o melhor resultado obtido com a fase Apoio sob as pernas direita e esquerda e os trabalhos relacionados.

	Classificador	**Atributos**	**Resultado (Acurácia)**
Referência	RNA	28	97,73%
Dubey, Wadhwani e Wadhwani (2013)	RNA	Média, Dv, Cv e Soma	89,47%
Alkhatib et al. (2015)	*k-NN*	11	90,42%

Dois algoritmos sob validação *holdout* tiveram desempenho inferior quando comparados aos resultados dos classificadores sob as demais validações. A Árvore de Decisão que alcançou 32,26% na precisão e 34,09% de acurácia, e o *k-NN* com 64,52% de precisão e 39,77% em acurácia, mesmo com o conjunto de dados das duas pernas. A razão está na mesma apresentada com a classificação da perna direita, o conjunto menor de dados que é aplicado para treinamento nesse tipo de validação.

Os resultados para a classe Controle foram mais significativos na fase Apoio se comparados a classe Parkinson, pois a maioria dos resultados para a especificidade alcançou percentuais superiores a 90,00%, como a RNA sob a validação *k-fold* com 97,67% e a Árvore de Decisão sob validação *leave-one-out* resultou em 93,33%. O que ocorreu foi uma sobreposição de classes, pois se as regras discriminativas nas classes são construídas de forma a minimizar o número de instâncias classificadas incorretamente, isso pode levar a um mal desempenho para atributos na área de sobreposição para a classe minoritária (parkinsonianos) (WEISS, 2013).

Na classificação da fase Swing das pernas direita e esquerda muitos classificadores apresentaram acurácia maior que 90,00%. O algoritmo RNA alcançou o melhor desempenho com acurácia em 98,86%, tanto sob validação *k-fold* como *leave-one-out*. Esse bom desempenho das RNAs nessa medida de desempenho tem como razão principal a alta capacidade de generalização e considerável tolerância a falhas e ruídos maior que a maioria dos algoritmos supervisionados (HAYKIN, 2000). Resultado superior ao obtido por Zhang et al. (2013) que utilizando a SVM em dados das passadas conseguiu 83,00% de acurácia. E, Lee e Lim (2012) aplicou a RNA mas apenas sob quatro atributos, obteve 77,33% de acurácia (Tabela 7). A Árvore de Decisão e o *k-NN* sob a validação *k-fold* alcançaram mesmo desempenho de 96,59% para a acurácia, apenas 2,27% inferior ao melhor resultado obtido nessa classificação.

Tabela 7 – Comparativo entre o melhor resultado obtido com a fase Swing das pernas direita e esquerda e os trabalhos relacionados.

	Classificador	**Atributos**	**Resultado (Acurácia)**
Referência	RNA	28	98,86%
Zhang et al. (2013)	SVM	Passadas	83,00%
Lee e Lim (2012)	RNA	Soma, máximo e mínimo VGRF	77,33%

Assim como na classificação da fase Swing da perna direita, os algoritmos se mostraram nessa fase com as pernas direita e esquerda mais tendenciosos a identificar pacientes Controle. Os resultados para a especificidade de todos os algoritmos sob todas as técnicas de validação utilizadas foram maior que 90,00%. A RNA e o SVM alcançaram os maiores resultados, 100% de especificidade sob a validação *holdout*. Além dos fatores já citados que justificam as diferenças nos resultados entre sensitividade e especificidade nas classificações anteriores. Outra questão importante para este tipo de problema é a presença de pequenas disjunções no conjunto de dados, quando a maioria dos algoritmos de aprendizagem supervisionada tem dificuldade em detectar essas regiões. Este problema está ligado a conjuntos de dados que não representam a classe positiva, pois pode acontecer que a classe minoritária seja representada por um número de subconceitos, o que significa que seus atributos formam vários "pedaços" de dados (LÓPEZ et al., 2012).

O bom desempenho dos algoritmos para a fase Swing, tanto de uma das pernas como as duas, têm relação com os sintomas da doença. A redução na velocidade, cadência e comprimento do passo influenciam no *timestamp*. A fase Swing constitui cerca de 40,00% do ciclo (Figura 1), pode até ocorrer a redução desse percentual em pacientes parkinsonianos com efeitos colaterais avançados (NIEUWBOER; GILADI, 2013). Portanto, esses fatores influenciam na coleta dos dados, podendo ocorrer fenômenos de baixa frequência (AFSAR; TIRNAKLI; KURTHS, 2016). Assim, houve uma redução e maior linearidade nos dados.

Tabela 8 – Desempenho na classificação das fases Apoio e Swing das pernas direita e esquerda com AUC e σ de cada algoritmo.

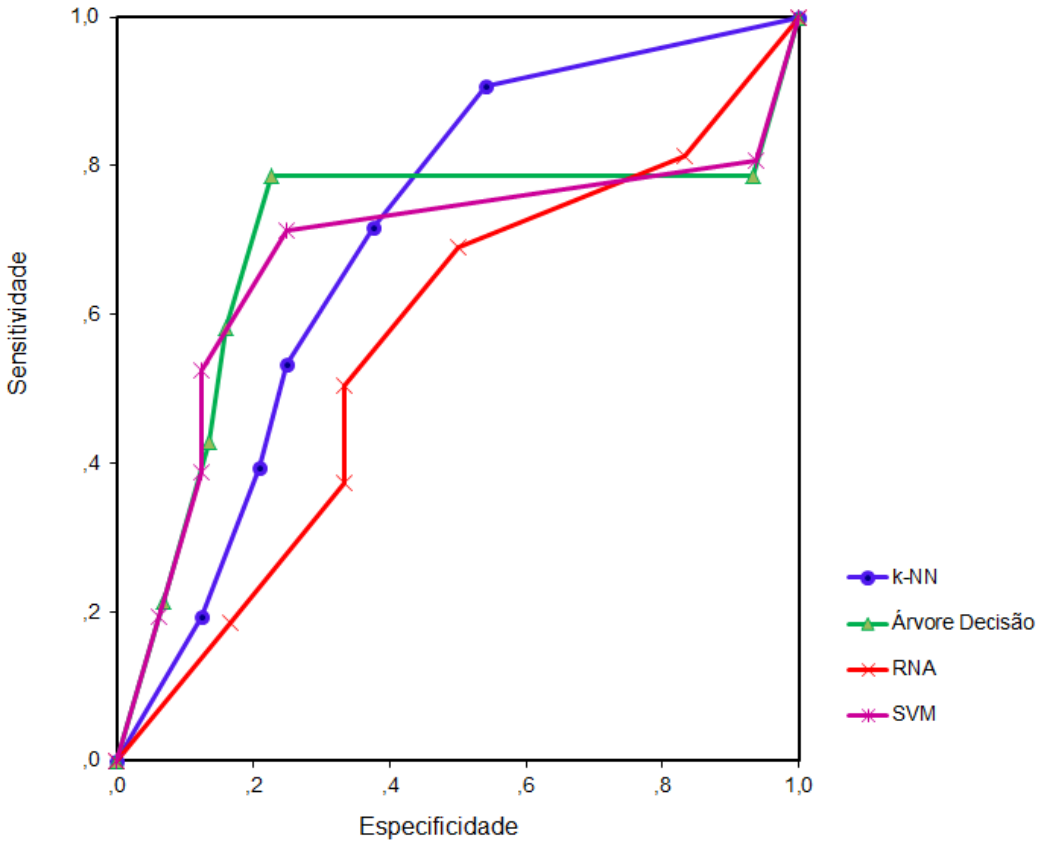
AUC	
k-NN	0,694 $\pm$ 0,228
Árvore Decisão	0,694 $\pm$ 0,242
RNA	0,548 $\pm$ 0,161
SVM	0,674 $\pm$ 0,162

Fonte – Elaborada pelo autor.

Na avaliação do desempenho dos classificadores com a curva ROC para as fase Apoio e Swing nas pernas direita e esquerda, o *k-NN* e a Árvore de Decisão alcançaram os melhores resultados (Figura 16). AUC para o *k-NN* foi 0,694 (σ = 0,228) e para a Árvore de Decisão foi 0,694 (σ = 0,242), como pode ser visto na Tabela 8. Valor superior se comparado com o trabalho publicado por Alkhatib et al. (2015) que classificaram a soma dos dados dos oito sensores com o algoritmo *k-NN* e alcançou AUC máximo de 0,611. A SVM apresentou desempenho próximo AUC igual a 0,674 (σ = 0,154). A RNA teve um resultado inferior se comparado aos demais, alcançou AUC igual a 0,548 (σ = 0,161).

A incidência dos atributos na fase Apoio nas pernas direita e esquerda aumentou em todos os 28 tipos (Figura 17). O que obteve maior peso por oito vezes foi RMS, sendo

Figura 16 – Avaliação desempenho dos algoritmos de aprendizagem supervisionada por meio da curva ROC na classificação das fases Apoio e Swing das pernas direita e esquerda.



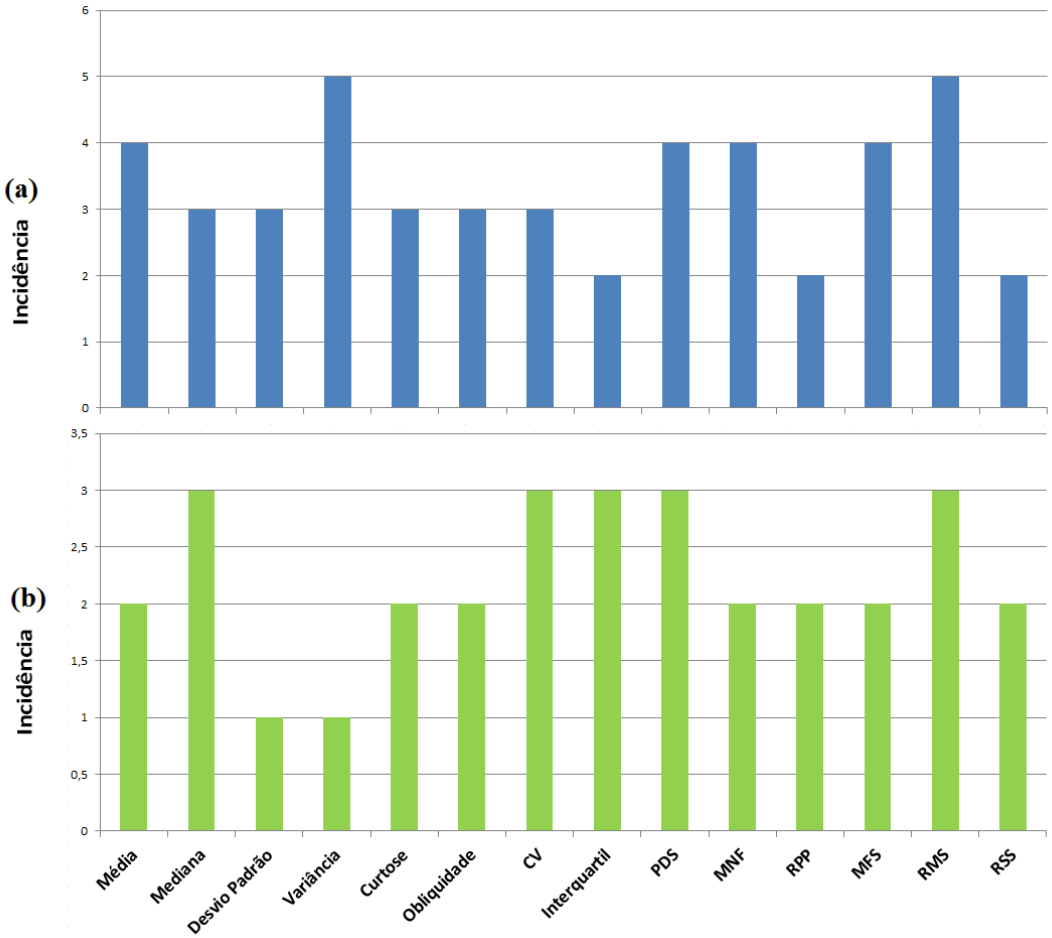
Fonte – Elaborado pelo autor.

cinco classificações com dados da perna direita e três a esquerda. Esse atributo alcançou peso igual a 1 quando utilizado com a SVM sob todas as validações. Vale salientar que esse atributo não foi identificado nos trabalhos relacionados à base. Na classificação da perna direita, além do atributo RMS, a variância obteve mais influência sob os algoritmos. Já com a perna esquerda foram mediana, CV, interquartil, PDS e RMS.

Na classificação com a RNA e a SVM a maioria dos atributos foram importantes para o desempenho em ambos. Ao classificar os dados da fase Apoio sob a perna esquerda, esses dois algoritmos sob validação *holdout* tiveram influência de todos os atributos e alcançaram peso igual a 1. MNF e RMS foram importantes ao classificar as duas pernas e em todos os tipos de validação.

Uma observação para a Árvore de Decisão, ao utilizar os dados da perna esquerda e a validação *holdout* nenhum dos atributos obteve peso. Com a mesma validação, ao classificar a perna direita, apenas a média obteve peso que foi igual a 1. Por isso, o desempenho inferior desse classificador em relação aos demais. Os valores dos pesos para essa fase estão nas Tabelas 19 à 22, Apêndice A.

Figura 17 – Incidência dos atributos que exerceram maior peso, utilizando a técnica *Feature Weight*, sob os classificadores na fase Apoio sob as pernas (a) direita e (b) esquerda.



Fonte – Elaborado pelo autor.

Na avaliação dos pesos dos atributos na fase Swing das pernas direita e esquerda foi observado que, por pelo menos cinco vezes, todos os atributos foram relevantes aos resultados apresentados pelos algoritmos de classificação. A variância por 14 vezes exerceu maior influência, destes, apenas nos dados da perna direita esse atributo obteve maior peso por oito vezes. Ainda que esse atributo tenha influência sob os algoritmos utilizados nessa classificação, o uso da variância com dados da Physionet foi identificada apenas no trabalho de Alkhatib et al. (2015).

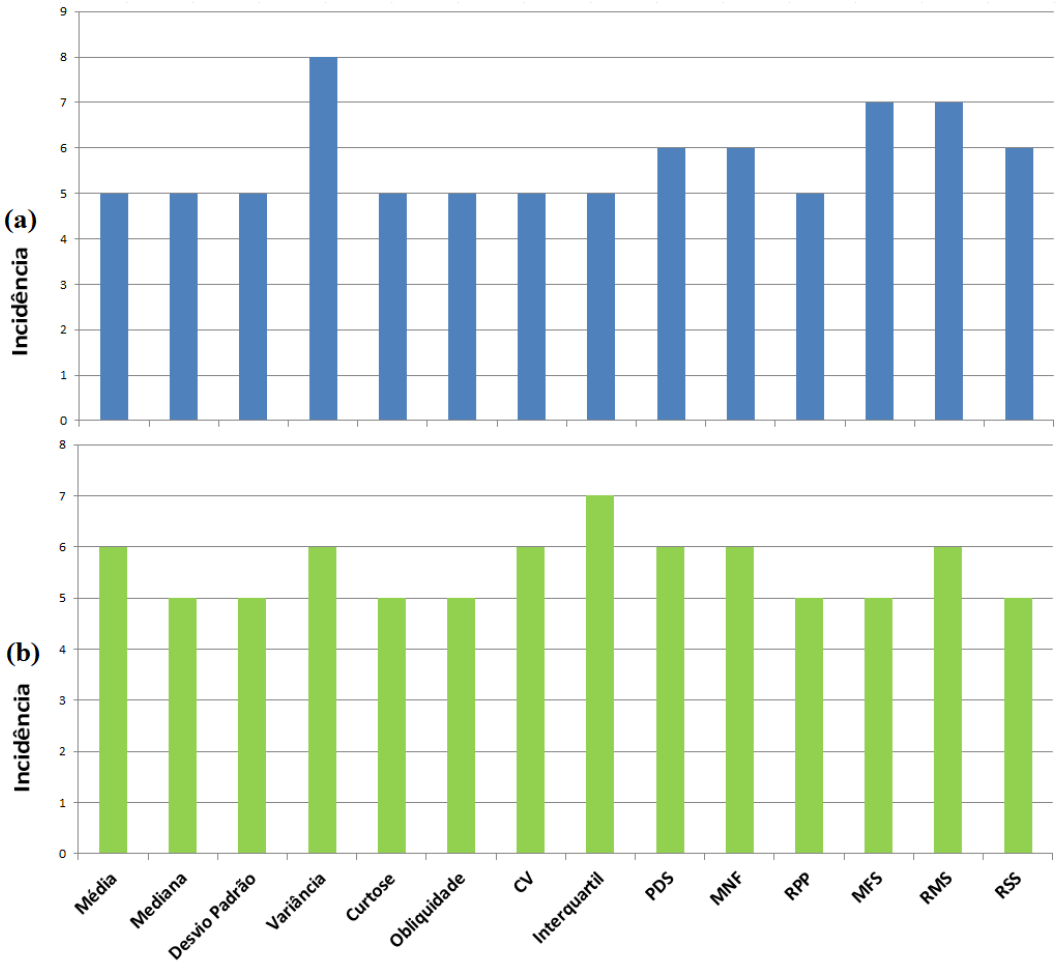
Houve classificação em que todos os atributos obtiveram peso igual a 1, ou seja, o maior, a exemplo da RNA na fase Swing em dados da perna esquerda sob as três técnicas de validação. Todos os atributos obtiveram peso igual a 1 quando aplicados em SVM, sob a validação *holdout*, na classificação da fase Swing da perna direita.

Contudo, nenhum atributo exerceu peso na classificação dessa fase na perna esquerda com a Árvore de Decisão sob validação *holdout*. Com o mesmo algoritmo, e sob a mesma validação, apenas um atributo obteve peso na classificação da perna direita, que foi a

média. Apesar da Árvore de Decisão ter tido um desempenho melhor na fase Swing tanto com uma perna como nas duas pernas, foi possível observar que, por exemplo, sob a validação *holdout* houve atributos que receberam peso igual a 0, ou seja, não exerceram qualquer influência. Isso mostra um importante fator para o desempenho inferior deste algoritmo na classificação dessa fase. A SVM sob a mesma validação *holdout* resultou no menor desempenho e apenas um atributo não foi relevante para o resultado, que foi MNS.

Assim como na classificação da perna direita, nenhum dos atributos influenciou negativamente sob os algoritmos quando utilizados na classificação das pernas direita e esquerda. O resultado da avaliação dos pesos para a fase Swing das duas pernas está nas Tabelas 23 à 26 no Apêndice A.

Figura 18 – Incidência dos atributos que exerceram maior peso, utilizando a técnica *Feature Weight*, sob os classificadores nas fase Swing sob as pernas (a) direita e (b) esquerda.



Fonte – Elaborado pelo autor.

Apesar dos algoritmos de aprendizagem supervisionada apresentarem um considerável desempenho na especificidade, foi possível verificar que para sensitividade a maioria dos resultados foi em torno de 50,00%. O que apresentou o melhor resultado para sensitividade foi a SVM sob validação *holdout* com 62,00% na classificação da fase Swing das pernas

direita e esquerda. Porém, esse resultado é inferior se comparado ao trabalho publicado por Dubey, Wadhwani e Wadhwani (2013), já que conseguiram 98,03%. Também, abaixo da sensitividade em 97,60% alcançada por Kamath (2015).

Esse baixo desempenho para a sensitividade deve-se a mais um problema chamado desequilíbrio de classes. É um problema comum em classificações binárias e em aplicações que envolvem dados do mundo real, como os que são gerados na área da saúde (LIN; LI, 2012). É importante ressaltar que a classe minoritária geralmente representa o conceito de interesse, por exemplo, pacientes doentes em um problema de diagnóstico médico; Enquanto a outra classe representa a contrapartida desse conceito (pacientes saudáveis). Normalmente, algoritmos têm um viés padrão em relação à classe majoritária, uma vez que as regras que predizem o maior número de exemplos são positivamente ponderadas durante o processo de aprendizagem em favor da métrica de acurácia. Consequentemente, as instâncias que pertencem à classe minoritária são mal classificadas com mais frequência do que as pertencentes à classe majoritária (LÓPEZ et al., 2012).

Embora tenha sido percebida uma influência mais recorrente de alguns tipos de atributos nos algorítimos de aprendizagem supervisionada para classificação dos dados da marcha humana, principalmente os relativos à estatística descritiva, não foi possível estabelecer uma correlação entre os atributos extraídos e os classificadores. Mas foi perceptível a influência destes em muitas das classificações e, também, no trabalho por autores que utilizaram a base Physionet.

Não foi possível mensurar suas influências para distinção entre as fases de Apoio e Swing, ou seja, não se identificou quais atributos são "chave" para distinguir os padrões das fases do ciclo. Além disso, não foi possível estabelecer uma diferença ou influência do método de FS escolhido neste trabalho em relação ao desempenho apresentado pelos algoritmos. Tanto os classificadores que tiveram seus atributos selecionados por *Backward* como *Forward* tiveram resultados diferentes tanto para a fase de Apoio como a Swing com dados de uma ou duas pernas. Também, não foi possível observar se esses métodos foram responsáveis pelos diferentes valores nos pesos dos atributos pelos resultados obtidos com os algoritmos aplicados.

5 CONCLUSÃO

Este trabalhou apresentou o desempenho de algoritmos de aprendizagem supervisionada na classificação das fases Apoio e Swing em portadores de DP e Controle. Ainda, foram identificados atributos que mais influenciaram no desempenho desses algoritmos. Com isso, buscou-se alcançar melhores resultados do que os que já foram alcançados em trabalhos relacionados, mas mensurando os pesos dos atributos aqui extraídos.

O algoritmo SVM sob validação *k-fold* foi o classificador que obteve melhor desempenho para a fase Apoio sob a perna direita com 97,77% de precisão e 95,45% na acurácia. A mediana foi o atributo com maior peso na classificação da fase Apoio sob uma das pernas em todos os classificadores. Contudo, os resultados mostraram que a média com uma considerável frequência influenciou nos resultados para essa classificação quando os algoritmos foram validados por meio de *holdout*, alcançando peso máximo igual a 1; Ao aplicar a técnica de validação *k-fold* o atributo que obteve maior peso foi a obliquidade, resultando em 0,900; Na *leave-one-out*, a obliquidade e a RMS por mais vezes exerceram maior peso, valor de até 1, nos algoritmos de classificação abordados.

A maior precisão na fase de Swing da perna direita foi com os classificadores Árvores de Decisão e SVM quando alcançaram 100% ambos sob validação *holdout*. A obliquidade exerceu maior peso em todos os classificadores nessa fase em uma das pernas. Contudo, com o algoritmo SVM esse atributo foi o único que não obteve peso para o resultado da classificação quando validado pela técnica *holdout*. Seis atributos influenciaram, com peso igual a 1, em todas as validações aplicadas a SVM, foram: Interquartil, PDS, MNF, MFS, RMS e RSS. CV teve peso em todas as classificações realizadas com RNA sob todas as validações.

A RNA teve o melhor desempenho para a classificação da fase Apoio sob ambas as pernas e utilizando a validação *k-fold*, pois obteve para precisão 97,78% e 97,73% na acurácia. Os algoritmos sob validação *holdout* tiveram o atributo média com maior peso no resultado para a perna direita, e na perna esquerda foi CV, ambos com peso máximo igual 1; Ao aplicar a validação *k-fold* o atributo que obteve mais peso foi MDF para a perna direita e para a perna esquerda foi interquartil; Na *leave-one-out*, a variância com mais frequência exerceu maior peso nos algoritmos de classificação abordados para a perna direita e para a perna esquerda foram a Curtose, RPP e RSS, obtendo peso com valor máximo de 1.

A melhor precisão na fase Swing das pernas direita e esquerda foi com os classificadores RNA e SVM quando alcançaram 100%, ambos sob validação *holdout*. Ainda, para essa fase nas duas pernas e os algoritmos sob validação *holdout* os atributos média, mediana e PDS foram os que frequentemente influenciaram no resultado para a perna direita, com valor máximo do peso igual 1; Já ao aplicar a validação *k-fold* o atributo que obteve mais

peso foi a MDF para a perna direita, e para a perna esquerda foi interquartil; Na *leave-one-out*, a variância por mais vezes exerceu maior peso com valor de até 1, nos algoritmos de classificação abordados para a perna direita, e para a perna esquerda foram variância, interquartil e RMS, obtendo peso com valor máximo de 1.

Por fim, acreditamos que ainda seja possível melhorar o desempenho dos algoritmos, pois alguns não passaram de aproximadamente 50,00% de sensitividade. Para melhorar o desempenho podem ser incluídos mais atributos a serem mensurados, a exemplo dos valores máximos e mínimos da VGRF, inclusive, com mais opções de algoritmos de aprendizagem supervisionada. O propósito é incluir os dados de todos os sensores implantados nos pés dos pacientes e extrair mais características da marcha, como cadência e passada.

5.1 Limitações

Embora a precisão na classificação dos dados tenha alcançado 100% com os algoritmos de aprendizagem supervisionada acredita-se que é possível aumentar os percentuais de sensitividade, assim como também identificar mais características da marcha com a inserção de dois fatores que não foram utilizados nesta pesquisa: o aproveitamento de todas as variáveis extraídas dos sensores; e uso de todas as amostras disponíveis.

Não foi realizada uma análise estatística para avaliar as diferenças entre as médias dos resultados obtidos com a precisão, acurácia, sensitividade e sensibilidade através de abordagens como teste F e teste T.

Faz-se necessário também um estudo mais amplo sobre outras técnicas FS, *Embedded* e *Filtter*, e mais algoritmos de AM. Com isso, espera-se que através dos resultados de mais classificadores sob mais atributos, analisado pelas respectivas técnicas de seleção, possa determinar resultados ainda melhores.

Também vale ressaltar a importância de buscar utilizar mais aspectos das técnicas de FS que tenham a capacidade de identificar a interdependência de atributos, a exemplo de buscas não-determinísticas.

A classificação da marcha humana neste trabalho esteve restrita à análise das duas principais fases. Não houve uma análise feita por especialistas para verificar se a relação do tempo e VGRF há padrões entre parkinsonianos e controle tanto na Apoio e Swing em uma das pernas, quanto em ambas as pernas.

Outro fator que merece maior análise é a utilização de modelos de regressão e baseados em aprendizagem não supervisionada como, por exemplo, agrupamento (*Clustering*) e Análise de Componentes Principais (PCA). Assim como também realizar uma análise empírica dos algoritmos aplicados, possibilitando assim um estudo da performance dos mesmos.

Por fim, no trabalho apresentado não foi possível a indicação da relação entre os pesos dos atributos e as classes dos dados, ou seja, não foi possível afirmar quais atributos são mais relevantes para determinar padrões nos ciclos da marcha dos parkinsonianos e

controle. Entretanto, determinar esta correlação pode alavancar as pesquisas que utilizam seleção de atributos como técnica de avaliação de desempenho de algoritmos de AM.

5.2 Trabalhos futuros

Serão apresentadas aqui algumas melhorias que podem ser aplicadas como trabalhos futuros para aumentar a capacidade de classificação dos dados:

- aplicação de algoritmos de aprendizagem não supervisionada junto as características já extraídas com o objetivo de identificar mais padrões no ciclo da marcha dos pacientes;
- alteração no tempo mínimo de captura, que inclua todas as amostras e seja possível ampliar o quantitativo de pacientes a serem classificados;
- avaliar outras funções estatísticas e mensurar o seu peso sobre os algoritmos de classificação;
- realizar uma análise estatística para avaliar as diferenças entre as médias dos resultados obtidos com a precisão, acurácia, sensitividade e sensibilidade através de abordagens como teste F e teste T;
- aplicar mais algoritmos de aprendizagem supervisionada com outras funções estatísticas;
- avaliar os atributos sob outras técnicas de FS;
- utilizar outras variáveis presentes na base que possam ser correlacionadas para identificação dos ciclos da marcha e padrões da pisada;
- realizar análise empírica dos algoritmos de aprendizagem supervisionada;
- buscar formas de correlacionar a marcha à sua importância na classificação dos pacientes portadores de DP.

REFERÊNCIAS

ABERNETHY, B. *Biophysical foundations of human movement*. [S.l.]: Human Kinetics, 2013.

ACKERMANN, M.; BOGERT, A. J. Van den. Optimality principles for model-based prediction of human gait. *Journal of biomechanics*, Elsevier, v. 43, n. 6, p. 1055–1060, 2010.

AFSAR, O.; TIRNAKLI, U.; KURTHS, J. Entropy-based complexity measures for gait data of patients with parkinson's disease. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, AIP Publishing, v. 26, n. 2, p. 023115, 2016.

AHMED, S. S. et al. Metabolic profiling of parkinson's disease: evidence of biomarker from gene expression analysis and rapid neural network detection. *Journal of biomedical science*, BioMed Central, v. 16, n. 1, p. 1, 2009.

ALELYANI, S.; TANG, J.; LIU, H. Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications*, v. 29, 2013.

ALKHATIB, R. et al. Gait-ground reaction force sensors selection based on roc curve evaluation. *Journal of Computer and Communications*, Scientific Research Publishing, v. 3, n. 03, p. 13, 2015.

ALVAREZ-ALVAREZ, A.; TRIVINO, G. Linguistic description of the human gait quality. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 26, n. 1, p. 13–23, 2013.

ALVAREZ-ALVAREZ, A.; TRIVINO, G.; CORDÓN, O. Human gait modeling using a genetic fuzzy finite state machine. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 20, n. 2, p. 205–223, 2012.

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

BACCOUR, L.; JOHN, R. I. Experimental analysis of crisp similarity and distance measures. In: *SoCPaR*. [S.l.: s.n.], 2014. p. 96–100.

BARBER, D. *Bayesian reasoning and machine learning*. [S.l.]: Cambridge University Press, 2012.

BENIWAL, S.; ARORA, J. Classification and feature selection techniques in data mining. *International Journal of Engineering Research & Technology (IJERT)*, v. 1, n. 6, 2012.

BOLLE, R. M. et al. *Guide to biometrics*. [S.l.]: Springer Science & Business Media, 2013.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. A review of feature selection methods on synthetic data. *Knowledge and information systems*, Springer, v. 34, n. 3, p. 483–519, 2013.

BURROWS, W. R. et al. Cart decision-tree statistical analysis and prediction of summer season maximum surface ozone for the vancouver, montreal, and atlantic regions of canada. *Journal of applied meteorology*, v. 34, n. 8, p. 1848–1862, 1995.

CABY, B. et al. Feature extraction and selection for objective gait analysis and fall risk assessment by accelerometry. *Biomedical engineering online*, BioMed Central, v. 10, n. 1, p. 1, 2011.

CARVALHO, L. A. V. de. *Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração*. [S.l.]: Érica, 2001.

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

CHANG, D.; ALBAN-HIDALGO, M.; HSU, K. Diagnosing parkinson's disease from gait. Stanford, 2014.

CHEN, K.-H. et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC bioinformatics*, BioMed Central, v. 15, n. 1, p. 1, 2014.

CHESTER, V. L.; TINGLEY, M.; BIDEN, E. N. An extended index to quantify normality of gait in children. *Gait & posture*, Elsevier, v. 25, n. 4, p. 549–554, 2007.

COLE, B. T.; OZDEMIR, P.; NAWAB, S. H. Dynamic svm detection of tremor and dyskinesia during unscripted and unconstrained activities. In: IEEE. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2012. p. 4927–4930.

CORTI, O.; LESAGE, S.; BRICE, A. What genetics tells us about the causes and mechanisms of parkinson's disease. *Physiological reviews*, Am Physiological Soc, v. 91, n. 4, p. 1161–1218, 2011.

DALIRI, M. R. Chi-square distance kernel of the gaits for the diagnosis of parkinson's disease. *Biomedical Signal Processing and Control*, Elsevier, v. 8, n. 1, p. 66–70, 2013.

DASGUPTA, H. An algorithm for stance and swing phase detection of human gait cycle. In: IEEE. *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*. [S.l.], 2015. p. 447–450.

DENG, N.; TIAN, Y.; ZHANG, C. *Support vector machines: optimization based theory, algorithms, and extensions*. [S.l.]: CRC press, 2012.

DIEDERICH, J. *Rule extraction from support vector machines*. [S.l.]: Springer Science & Business Media, 2008. v. 80.

DUBEY, M.; WADHWANI, A.; WADHWANI, S. Gait based vertical ground reaction force analysis for parkinson's disease diagnosis using self organizing map. *International Journal of Advanced Biological and Biomedical Research*, v. 1, n. 6, p. 624–636, 2013.

ELLIS, R. J.; CITI, L.; BARBIERI, R. A point process approach for analyzing gait variability dynamics. In: IEEE. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2011. p. 1648–1651.

FACELI, K. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.]: Grupo Gen-LTC, 2011.

FEDEROLF, P.; BOYER, K.; ANDRIACCHI, T. Application of principal component analysis in clinical gait research: identification of systematic differences between healthy and medial knee-osteoarthritic gait. *Journal of biomechanics*, Elsevier, v. 46, n. 13, p. 2173–2178, 2013.

FRENKEL-TOLEDO, S. et al. Effect of gait speed on gait rhythmicity in parkinson's disease: variability of stride time and swing time respond differently. *Journal of neuroengineering and rehabilitation*, BioMed Central, v. 2, n. 1, p. 1, 2005.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1.

GANGWAR, A.; BHARDWAJ, M. An overview: Peak to average power ratio in ofdm system & its effect. *International Journal of Communication and Computer Technologies*, v. 1, n. 2, p. 22–25, 2012.

GEMAN, O. et al. Gait in parkinson's disease-signal processing and modeling. In: *11th International Conference on Development and Application System*. [S.l.: s.n.], 2012. p. 166–170.

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

HAJIAN-TILAKI, K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, Babol University of Medical Sciences, v. 4, n. 2, p. 627, 2013.

HAN, Y.; MA, Z.; ZHOU, P. A study of gaits in parkinson's patients using autoregressive model. In: IEEE. *Bio-Inspired Computing, 2009. BIC-TA'09. Fourth International Conference on*. [S.l.], 2009. p. 1–4.

HARRINGTON, P. *Machine learning in action*. [S.l.]: Manning, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Unsupervised learning. In: *The elements of statistical learning*. [S.l.]: Springer, 2009. p. 485–585.

HAUSDORFF, J. M. et al. Rhythmic auditory stimulation modulates gait variability in parkinson's disease. *European Journal of Neuroscience*, Wiley Online Library, v. 26, n. 8, p. 2369–2375, 2007.

HAYKIN, S. S. Redes neurais artificiais: princípio e prática. *2ª Edição, Bookman, São Paulo, Brasil*, 2000.

HERRAN, A. Muro-de-la; GARCIA-ZAPIRAIN, B.; MENDEZ-ZORRILLA, A. Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 14, n. 2, p. 3362–3394, 2014.

HOLZINGER, A.; DEHMER, M.; JURISICA, I. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, BioMed Central, v. 15, n. 6, p. 1, 2014.

HOUCK, J. et al. Analysis of vertical ground reaction force variables during a sit to stand task in participants recovering from a hip fracture. *Clinical Biomechanics*, Elsevier, v. 26, n. 5, p. 470–476, 2011.

HU, M. et al. Incremental learning for video-based gait recognition with lbp flow. *IEEE transactions on cybernetics*, IEEE, v. 43, n. 1, p. 77–89, 2013.

JEGADEESHWARAN, R.; SUGUMARAN, V. Comparative study of decision tree classifier and best first tree classifier for fault diagnosis of automobile hydraulic brake system using statistical features. *Measurement*, Elsevier, v. 46, n. 9, p. 3247–3260, 2013.

JIA, J. et al. Intensive analysis of gait in the elderly with parkinson's disease using center of pressure during walking. In: IEEE. *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*. [S.l.], 2015. p. 391–396.

JIAWEI, H.; KAMBER, M. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, v. 5, 2001.

KAMATH, C. A novel approach to unravel gait dynamics using symbolic analysis. *Open Access Library Journal*, Scientific Research Publishing, v. 2, n. 05, p. 1, 2015.

KE, S.-R. et al. A review on video-based human activity recognition. *Computers*, Multidisciplinary Digital Publishing Institute, v. 2, n. 2, p. 88–131, 2013.

KIM, S. D. et al. Postural instability in patients with parkinson's disease. *CNS drugs*, Springer, v. 27, n. 2, p. 97–112, 2013.

KIRA, K.; RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In: *AAAI*. [S.l.: s.n.], 1992. v. 2, p. 129–134.

LANCE, G. N.; WILLIAMS, W. T. Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, Br Computer Soc, v. 9, n. 1, p. 60–64, 1966.

LANCE, G. N.; WILLIAMS, W. T. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, v. 1, n. 1, p. 15–20, 1967.

LANGA, K. M.; LEVINE, D. A. The diagnosis and management of mild cognitive impairment: a clinical review. *Jama*, American Medical Association, v. 312, n. 23, p. 2551–2561, 2014.

LATKOWSKI, T.; OSOWSKI, S. Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*, Elsevier, v. 42, n. 2, p. 864–872, 2015.

LAU, L. M. D.; BRETELER, M. M. Epidemiology of parkinson's disease. *The Lancet Neurology*, Elsevier, v. 5, n. 6, p. 525–535, 2006.

LEE, C. P.; TAN, A. W.; TAN, S. C. Gait recognition via optimally interpolated deformable contours. *Pattern Recognition Letters*, Elsevier, v. 34, n. 6, p. 663–669, 2013.

LEE, S.-H. Identifying people with parkinson's disease using foot pressure data. *Indian Journal of Science and Technology*, v. 8, n. 13, 2015.

LEE, S.-H.; LIM, J. S. Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. *Expert Systems with Applications*, Elsevier, v. 39, n. 8, p. 7338–7344, 2012.

LIAO, H. Speaker adaptation of context dependent deep neural networks. In: IEEE. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.], 2013. p. 7947–7951.

LIN, H.-T.; LI, L. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, MIT Press, v. 24, n. 5, p. 1329–1367, 2012.

LÓPEZ, V. et al. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, Elsevier, v. 39, n. 7, p. 6585–6608, 2012.

LUGER, G. F. *Inteligência Artificial-: Estruturas e estratégias para a solução de problemas complexos*. [S.l.]: Bookman, 2014.

MANAP, H. H.; TAHIR, N. M.; YASSIN, A. I. M. Statistical analysis of parkinson disease gait classification using artificial neural network. In: IEEE. *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. [S.l.], 2011. p. 060–065.

MEDEIROS, L. et al. A gait analysis approach to track parkinson's disease evolution using principal component analysis. In: IEEE. *Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on*. [S.l.], 2016. p. 48–53.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013.

MUMMOLO, C.; MANGIALARDI, L.; KIM, J. H. Quantifying dynamic characteristics of human walking for comprehensive gait cycle. *Journal of biomechanical engineering*, American Society of Mechanical Engineers, v. 135, n. 9, p. 091006, 2013.

MUNIZ, A. et al. Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in parkinson disease on ground reaction force during gait. *Journal of Biomechanics*, Elsevier, v. 43, n. 4, p. 720–726, 2010.

MUSA, A. B. Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, Springer, v. 4, n. 1, p. 13–24, 2013.

MUTHANE, U.; RAGOTHAMAN, M.; GURURAJ, G. Epidemiology of parkinson's disease and movement disorders in india: problems and possibilities. *Japi*, v. 55, p. 719–24, 2007.

NAHDAP. *National Addiction and HIV Data Archive Program*. 2016. Disponível em: <http://www.icpsr.umich.edu/icpsrweb/NAHDAP/>.

NAIR, S. R. et al. A decision tree for differentiating multiple system atrophy from parkinson's disease using 3-t mr imaging. *European radiology*, Springer, v. 23, n. 6, p. 1459–1466, 2013.

NIEUWBOER, A.; GILADI, N. Characterizing freezing of gait in parkinson's disease: models of an episodic phenomenon. *Movement Disorders*, Wiley Online Library, v. 28, n. 11, p. 1509–1519, 2013.

OZCIFT, A. Svm feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of parkinson disease. *Journal of medical systems*, Springer, v. 36, n. 4, p. 2141–2147, 2012.

PANT, J. K.; KRISHNAN, S. Foot gait time series estimation based on support vector machine. In: IEEE. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2014. p. 6410–6413.

PHINYOMARK, A. et al. *The usefulness of mean and median frequencies in electromyography analysis*. [S.l.]: INTECH Open Access Publisher, 2012.

PHYSIONET. *Gait in Parkinson's Disease*. 2015. Disponível em: <https://physionet.org/physiobank/database/gaitpdb/>.

POSTUMA, R. B. et al. Identifying prodromal parkinson's disease: Pre-motor disorders in parkinson's disease. *Movement Disorders*, Wiley Online Library, v. 27, n. 5, p. 617–626, 2012.

POULIOT-LAFORTE, A. et al. Validity of an accelerometer as a vertical ground reaction force measuring device in healthy children and adolescents and in children and adolescents with osteogenesis imperfecta type i. *J Musculoskelet Neuronal Interact*, v. 14, n. 2, p. 155–161, 2014.

PRATI, R. C.; MONARD, M. C. *Novas abordagens em aprendizado de máquina para a geraçao de regras, classes desbalanceadas e ordenaçao de casos*. Tese (Doutorado) — Tese de doutorado, ICMC/USP Sao Carlos, 2006.

PREIS, J. et al. Gait recognition with kinect. In: NEW CASTLE, UK. *1st international workshop on kinect in pervasive computing*. [S.l.], 2012. p. P1–P4.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986.

QUINLAN, J. R. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, v. 4, p. 77–90, 1996.

RCMD. *Resource Center for Minority Data*. 2016. Disponível em: <http://www.icpsr.umich.edu/icpsrweb/RCMD/>.

REDDY, K. S.; REDDY, M. K.; SITARAMULU, V. An effective data preprocessing method for web usage mining. In: IEEE. *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*. [S.l.], 2013. p. 7–10.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: *Encyclopedia of database systems*. [S.l.]: Springer, 2009. p. 532–538.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, Springer, v. 53, n. 1-2, p. 23–69, 2003.

ROIZ, R. de M. et al. Doença de parkinson: análise da marcha e uso de pistas visuais. Campinas, SP, 2011.

ROKACH, L.; MAIMON, O. *Data mining with decision trees: theory and applications*. [S.l.]: World scientific, 2014.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. *bioinformatics*, Oxford Univ Press, v. 23, n. 19, p. 2507–2517, 2007.

SAK, H.; SENIOR, A.; BEAUFAYS, F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

SALVATORE, C. et al. Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy. *Journal of neuroscience methods*, Elsevier, v. 222, p. 230–237, 2014.

SAMANTA, B.; AL-BALUSHI, K. Artificial neural network based fault diagnostics of rolling element bearings using time-domain features. *Mechanical systems and signal processing*, Elsevier, v. 17, n. 2, p. 317–328, 2003.

SCAFETTA, N.; MARCHI, D.; WEST, B. J. Understanding the complexity of human gait dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, AIP Publishing, v. 19, n. 2, p. 026108, 2009.

SHINE, J. M. et al. Freezing of gait in parkinson's disease is associated with functional decoupling between the cognitive control network and the basal ganglia. *Brain*, Oxford Univ Press, p. awt272, 2013.

SHIRVAN, R. A.; TAHAMI, E. Voice analysis for detecting parkinson's disease using genetic algorithm and knn classification method. In: IEEE. *Biomedical Engineering (ICBME), 2011 18th Iranian Conference of*. [S.l.], 2011. p. 278–283.

SIMON, H. *Redes Neurais–Princípios e Prática*. [S.l.]: Porto Alegre: Bookman, 2001.

SOUSA, R. T. et al. *Avaliação de classificadores na classificação de radiografias de tórax para o diagnóstico de pneumonia infantil*. Tese (Doutorado) — Universidade Federal de Goiás, 2013.

STEFL, S. et al. Molecular mechanisms of disease-causing missense mutations. *Journal of molecular biology*, Elsevier, v. 425, n. 21, p. 3919–3936, 2013.

STEGEMÖLLER, E. L. et al. Postural instability and gait impairment during obstacle crossing in parkinson's disease. *Archives of physical medicine and rehabilitation*, Elsevier, v. 93, n. 4, p. 703–709, 2012.

TAHIR, N. M.; MANAP, H. H. Parkinson disease gait classification based on machine learning approach. *Journal of Applied Sciences*, Asian Network for Scientific Information (ANSINET), v. 12, n. 2, p. 180, 2012.

TAN, L. C. Epidemiology of parkinson's disease. *Neurology Asia*, v. 18, n. 3, p. 231–238, 2013.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciencia Moderna, 2009.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, CRC Press, p. 37, 2014.

TAO, W. et al. Gait analysis using wearable sensors. *Sensors*, Molecular Diversity Preservation International, v. 12, n. 2, p. 2255–2283, 2012.

THEODORIDIS, S. Konstantinos koutroumbas. *Pattern Recognition (Third edition), Academic Press, Elsevier*, 2006.

TRIPOLITI, E. E. et al. Automatic detection of freezing of gait events in patients with parkinson's disease. *Computer methods and programs in biomedicine*, Elsevier, v. 110, n. 1, p. 12–26, 2013.

UMBERGER, B. R. Stance and swing phase costs in human walking. *Journal of the Royal Society Interface*, The Royal Society, v. 7, n. 50, p. 1329–1340, 2010.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer Science & Business Media, 2013.

WAHID, F. et al. Classification of parkinson's disease gait using spatial-temporal gait features. *IEEE journal of biomedical and health informatics*, IEEE, v. 19, n. 6, p. 1794–1802, 2015.

WEISS, G. M. Foundations of imbalanced learning. *H. He, & Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications*, Citeseer, p. 13–41, 2013.

WILLIS, A. W. et al. Geographic and ethnic variation in parkinson disease: a population-based study of us medicare beneficiaries. *Neuroepidemiology*, Karger Publishers, v. 34, n. 3, p. 143–151, 2010.

WORTH, P. F. How to treat parkinson's disease in 2013. *Clinical Medicine*, Royal College of Physicians, v. 13, n. 1, p. 93–96, 2013.

YAO, K. et al. Adaptation of context-dependent deep neural networks for automatic speech recognition. In: IEEE. *Spoken Language Technology Workshop (SLT), 2012 IEEE*. [S.l.], 2012. p. 366–369.

YOGEV, G. et al. Dual tasking, gait rhythmicity, and parkinson's disease: which aspects of gait are attention demanding? *European journal of neuroscience*, Wiley Online Library, v. 22, n. 5, p. 1248–1256, 2005.

YONEYAMA, M. et al. Accelerometry-based gait analysis and its application to parkinson's disease assessment—part 1: Detection of stride event. *IEEE transactions on neural systems and rehabilitation engineering*, IEEE, v. 22, n. 3, p. 613–622, 2014.

ZAKI, M. J.; JR, W. M. *Fundamentals of data mining algorithms*. [S.l.]: Cambridge University Press, 2011.

ZAKNICH, A. *Principles of adaptive filters and self-learning systems*. [S.l.]: Springer Science & Business Media, 2006.

ZENG, W.; WANG, C.; LI, Y. Model-based human gait recognition via deterministic learning. *Cognitive Computation*, Springer, v. 6, n. 2, p. 218–229, 2014.

ZHANG, C.; MA, Y. *Ensemble machine learning*. [S.l.]: Springer, 2012.

ZHANG, Y. et al. Pathological gait detection of parkinson's disease using sparse representation. In: IEEE. *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*. [S.l.], 2013. p. 1–8.

ZHANG, Y. et al. Binary pso with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems*, Elsevier, v. 64, p. 22–31, 2014.

ZHENG, H. et al. Machine learning and statistical approaches to support the discrimination of neuro-degenerative diseases based on gait analysis. In: *Intelligent Patient Management*. [S.l.]: Springer, 2009. p. 57–70.

APÊNDICE A – RESULTADO DETALHADO

Tabela 9 – Resultado dos algoritmos na classificação das fases Apoio e Swing da perna direita por validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

	Ciclo	Algoritmo		Precisão	Sensitividade	Especificidade	Acurácia
Perna Direita	Apoio	k-NN	(a)	48,39%	53,57%	44,83%	31,82%
			(b)	93,62%	52,39%	93,02%	95,45%
			(c)	86,67%	51,32%	86,05%	86,36%
		Árvore Decisão	(a)	38,71%	37,50%	51,28%	36,36%
			(b)	79,45%	50,72%	79,06%	78,40%
			(c)	77,78%	50,72%	77,27%	78,41%
		RNA	(a)	90,32%	50,91%	90,00%	62,50%
			(b)	97,77%	53,01%	97,50%	94,31%
			(c)	95,56%	51,81%	95,24%	94,32%
		SVM	(a)	74,19%	45,10%	77,78%	57,95%
			(b)	97,77%	52,38%	95,25%	95,45%
			(c)	91,11%	51,90%	90,48%	89,77%
	Swing	k-NN	(a)	93,54%	49,15%	93,75%	95,16%
			(b)	95,56%	50,00%	95,56%	97,73%
			(c)	95,56%	50,00%	95,56%	97,73%
		Árvore Decisão	(a)	100%	50,82%	100%	69,32%
			(b)	93,33%	49,41%	93,48%	96,59%
			(c)	93,33%	49,41%	93,48%	96,59%
		RNA	(a)	96,77%	50,85%	96,67%	67,05%
			(b)	95,56%	50,00%	95,56%	97,73%
			(c)	91,30%	51,85%	90,69%	94,31%
		SVM	(a)	100%	56,36%	100%	62,50%
			(b)	97,78%	53,01%	97,50%	94,32%
			(c)	97,78%	53,66%	97,44%	93,18%

Tabela 10 – Resultados dos algoritmos na classificação das fases Apoio e Swing das pernas direita e esquerda por validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

	Ciclo	Algoritmo		Precisão	Sensitividade	Especificidade	Acurácia
Pernas Direita e Esquerda	Apoio	k-NN	(a)	64,52%	57,14%	57,69%	39,77%
			(b)	91,11%	51,90%	90,48%	89,77%
			(c)	95,56%	58,11%	93,94%	84,09%
		Árvore Decisão	(a)	32,26%	33,33%	48,78%	34,09%
			(b)	71,11%	47,76%	72,92%	76,14%
			(c)	93,02%	48,78%	93,33%	93,18%
		RNA	(a)	93,55%	50,88%	93,33%	64,77%
			(b)	97,78%	51,16%	97,67%	97,73%
			(c)	97,78%	51,76%	97,62%	96,59%
		SVM	(a)	70,97%	45,83%	74,29%	54,55%
			(b)	93,48%	57,33%	91,43%	85,23%
			(c)	95,56%	58,11%	93,94%	84,09%
	Swing	k-NN	(a)	96,77%	51,72%	96,55%	65,91%
			(b)	95,56%	50,59%	95,45%	96,59%
			(c)	93,33%	50,00%	93,33%	95,45%
		Árvore Decisão	(a)	96,77%	50,85%	96,67%	67,05%
			(b)	93,33%	49,41%	93,48%	96,59%
			(c)	93,33%	50,00%	93,33%	95,45%
		RNA	(a)	100%	50,82%	100%	69,32%
			(b)	97,78%	50,57%	97,73%	98,86%
			(c)	97,78%	50,57%	97,73%	98,86%
		SVM	(a)	100%	62,00%	100%	56,82%
			(b)	97,78%	57,89%	96,97%	86,36%
			(c)	97,78%	57,89%	96,97%	86,36%

Tabela 11 – Pesos dos atributos com o classificador *k-NN* para a fase de Apoio sob a perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,375	0,705
2	Mediana	1	0,420	0,705
3	DV	1	0,341	0,705
4	Variância	0	0,307	0,295
5	Curtose	0	0,375	0,295
6	Obliquidade	0	0,466	0,295
7	CV	0	0,455	0,295
8	Interquartil	0	0,159	0,205
9	PDS	0	0,307	0
10	MDF	0	0,239	0
11	RPP	0	0,170	0
12	MNF	0	0,239	0,034
13	RMS	1	0,250	0,489
14	RSS	0	0,045	0

Tabela 12 – Pesos dos atributos para o classificador Árvores de Decisão para a fase de Apoio sob a perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,300	0,057
2	Mediana	0	0,500	0,886
3	DV	0	0,500	0,955
4	Variância	1	0,200	0
5	Curtose	0	0,500	0,080
6	Obliquidade	0	0,300	0,045
7	CV	0	0,200	0
8	Interquartil	0	0	0,034
9	PDS	0	0,100	0,011
10	MDF	0	0	0
11	RPP	0	0,300	0,045
12	MNF	1	0,400	0,057
13	RMS	0	0,300	0
14	RSS	0	0,100	0

Tabela 13 – Pesos dos atributos utilizados com RNA para a fase de Apoio sob a perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,700	0,784
2	Mediana	1	0,900	0,943
3	DV	1	0,800	0,886
4	Variância	1	0,800	0,807
5	Curtose	0	0,900	0,909
6	Obliquidade	1	0,900	0,875
7	CV	1	1	0,875
8	Interquartil	1	0,600	0,841
9	PDS	1	1	0,977
10	MDF	1	0,900	0,955
11	RPP	1	0,800	0,977
12	MNF	1	1	0,955
13	RMS	1	1	1
14	RSS	1	1	0,977

Tabela 14 – Pesos dos atributos usados com SVM para a fase de Apoio sob a perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,800	0,955
2	Mediana	1	1	1
3	DV	1	0,800	0,989
4	Variância	1	0,800	0,989
5	Curtose	0	0,800	0,977
6	Obliquidade	1	0,900	0,920
7	CV	1	0,900	0,966
8	Interquartil	1	0,500	0,057
9	PDS	1	1	0,989
10	MDF	1	1	1
11	RPP	1	0,900	1
12	MNF	1	1	1
13	RMS	1	1	0,966
14	RSS	1	1	1

Tabela 15 – Pesos dos atributos utilizados com *k-NN* para a fase de Swing da perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	0	0	0
2	Mediana	0	0	0
3	DV	0	0	0
4	Variância	0	0	0
5	Curtose	0	0	0
6	Obliquidade	1	1	1
7	CV	0	0	0
8	Interquartil	0	0	0
9	PDS	0	0,200	0
10	MDF	0	0	0
11	RPP	0	0	0
12	MNF	0	0,100	0
13	RMS	0	0	0
14	RSS	0	0	0

Tabela 16 – Pesos dos atributos usados com Árvores de Decisão para a fase de Swing da perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	0	0,100	0
2	Mediana	0	0	0
3	DV	0	0,100	0
4	Variância	0	0,300	0,011
5	Curtose	0	0	0,080
6	Obliquidade	1	0,800	0,989
7	CV	0	0,200	0
8	Interquartil	0	0	0,034
9	PDS	0	0	0
10	MDF	0	0	0
11	RPP	0	0	0
12	MNF	0	0	0
13	RMS	0	0	0
14	RSS	0	0	0

Tabela 17 – Pesos dos atributos usados com RNA para a fase de Swing da perna direita com validação (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,400	0,977
2	Mediana	1	0,700	0,841
3	DV	0	0,800	0,898
4	Variância	1	1	0,955
5	Curtose	1	1	0,966
6	Obliquidade	1	0,900	0,818
7	CV	1	1	1
8	Interquartil	1	0,900	1
9	PDS	1	1	0,773
10	MDF	1	0,900	0,830
11	RPP	1	1	0,898
12	MNF	1	1	0,864
13	RMS	1	1	0,943
14	RSS	1	1	0,886

Tabela 18 – Pesos dos atributos usados com SVM para a fase de Swing da perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso		
		(a)	(b)	(c)
1	Média	1	0,700	0,295
2	Mediana	1	0,600	0,716
3	DV	1	0,700	0,807
4	Variância	1	1	0,989
5	Curtose	1	0,900	0,989
6	Obliquidade	0	0,900	1
7	CV	1	0,600	0,966
8	Interquartil	1	1	1
9	PDS	1	1	1
10	MDF	1	1	1
11	RPP	1	0,900	0,989
12	MNF	1	1	1
13	RMS	1	1	1
14	RSS	1	1	1

Tabela 19 – Pesos dos atributos classificando com *k-NN* a fase de Apoio sob as pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0,1	0,102	0	0,5	0,193
2	Mediana	1	0,4	0,113	0	0,3	0,022
3	DV	0	0,1	0,795	0	0,2	0
4	Variância	0	0,4	0,659	0	0	0
5	Curtose	0	0	0	0	0	0,01
6	Obliquidade	0	0,4	0,579	0	0	0
7	CV	0	0	0	1	0	0,01
8	Interquartil	0	0	0	0	0	0
9	PDS	1	0,3	0,147	0	0	0,181
10	MDF	0	0,6	0,329	0	0,3	0,102
11	RPP	0	0,1	0,681	0	0,2	0,011
12	MNF	0	0,1	0,681	0	0	0
13	RMS	0	0,1	0,113	0	0	0,568
14	RSS	0	0,5	0,329	0	0,1	0,113

Tabela 20 – Pesos dos atributos usados com Árvores de Decisão para a fase de Apoio sob as pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0	0,09	0	0	0
2	Mediana	0	0,3	0,125	0	0	0,193
3	DV	0	0,3	0,045	0	0,1	0
4	Variância	0	0,6	0,727	0	0	0,011
5	Curtose	0	0	0,011	0	0,1	0
6	Obliquidade	0	0	0	0	0	0,034
7	CV	0	0	0	0	0	0,409
8	Interquartil	0	0,3	0,022	0	0,5	0,954
9	PDS	0	0	0,181	0	0,1	0,090
10	MDF	0	0,1	0,227	0	0	0,011
11	RPP	0	0	0,454	0	0	0
12	MNF	0	0,2	0,113	0	0,2	0
13	RMS	0	0,1	0	0	0,1	0
14	RSS	0	0	0,568	0	0	0,318

Tabela 21 – Pesos dos atributos utilizados com RNA para a fase de Apoio sob as pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0,8	0,886	1	1	0,988
2	Mediana	1	0,7	0,977	1	1	0,988
3	DV	0	1	0,761	1	1	0,977
4	Variância	1	0,3	0,988	1	1	0,977
5	Curtose	1	0,8	0,897	1	1	1
6	Obliquidade	1	0,9	0,920	1	1	0,988
7	CV	1	1	0,977	1	1	0,977
8	Interquartil	1	0,8	0,977	1	1	0,965
9	PDS	1	1	0,931	1	1	0,988
10	MDF	1	1	0,954	1	1	0,965
11	RPP	1	1	0,977	1	1	1
12	MNF	1	1	0,977	1	1	0,965
13	RMS	1	1	0,988	1	1	0,988
14	RSS	1	1	0,931	1	1	1

Tabela 22 – Pesos dos atributos usando SVM para a fase de Apoio sob as pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	0	1	0,897	1	1	0,988
2	Mediana	1	1	1	1	1	1
3	DV	1	0,9	0,784	1	1	0,988
4	Variância	1	1	0,977	1	0,9	0,965
5	Curtose	1	0,9	0,715	1	1	1
6	Obliquidade	1	0,9	0,909	1	1	0,988
7	CV	1	0,9	0,897	1	1	1
8	Interquartil	1	1	0,954	1	0,8	0,988
9	PDS	1	0,9	0,977	1	1	1
10	MDF	1	1	0,977	1	1	1
11	RPP	1	0,6	0,329	1	0,9	1
12	MNF	1	1	1	1	1	1
13	RMS	1	1	1	1	1	1
14	RSS	1	0,9	1	1	0,9	1

Tabela 23 – Pesos dos atributos classificando com *k-NN* a fase de Swing da perna direita com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0,1	0,102	0	0,5	0,193
2	Mediana	1	0,4	0,113	0	0,3	0,022
3	DV	0	0,1	0,795	0	0,2	0
4	Variância	0	0,4	0,659	0	0	0
5	Curtose	0	0	0	0	0	0,01
6	Obliquidade	0	0,4	0,579	0	0	0
7	CV	0	0	0	1	0	0,01
8	Interquartil	0	0	0	0	0	0
9	PDS	1	0,3	0,147	0	0	0,181
10	MDF	0	0,6	0,329	0	0,3	0,102
11	RPP	0	0,1	0,681	0	0,2	0,011
12	MNF	0	0,1	0,681	0	0	0
13	RMS	0	0,1	0,113	0	0	0,568
14	RSS	0	0,5	0,329	0	0,1	0,113

Tabela 24 – Pesos dos atributos classificando com Árvores de Decisão a fase de Swing das pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0	0,09	0	0	0
2	Mediana	0	0,3	0,125	0	0	0,193
3	DV	0	0,3	0,045	0	0,1	0
4	Variância	0	0,6	0,727	0	0	0,011
5	Curtose	0	0	0,011	0	0,1	0
6	Obliquidade	0	0	0	0	0	0,034
7	CV	0	0	0	0	0	0,409
8	Interquartil	0	0,3	0,022	0	0,5	0,954
9	PDS	0	0	0,181	0	0,1	0,090
10	MDF	0	0,1	0,227	0	0	0,011
11	RPP	0	0	0,454	0	0	0
12	MNF	0	0,2	0,113	0	0,2	0
13	RMS	0	0,1	0	0	0,1	0
14	RSS	0	0	0,568	0	0	0,318

Tabela 25 – Pesos dos atributos usados com RNA para a fase de Swing das pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	0	0,2	0	1	1	1
2	Mediana	1	0,8	1	1	1	1
3	DV	1	1	1	1	1	1
4	Variância	1	1	1	1	1	1
5	Curtose	1	1	1	1	1	1
6	Obliquidade	1	1	1	1	1	1
7	CV	1	1	1	1	1	1
8	Interquartil	1	1	1	1	1	1
9	PDS	1	1	1	1	1	1
10	MDF	1	1	1	1	1	1
11	RPP	1	1	1	1	1	1
12	MNF	1	1	1	1	1	1
13	RMS	1	1	1	1	1	1
14	RSS	1	1	1	1	1	1

Tabela 26 – Pesos dos atributos utilizados com SVM para a fase de Swing das pernas direita e esquerda com validações (a) *holdout*, (b) *k-fold* e (c) *leave-one-out*.

Id Atributo	Atributo	Peso					
		Perna Dir.			Perna Esq.		
		(a)	(b)	(c)	(a)	(b)	(c)
1	Média	1	0,9	0,329	1	1	1
2	Mediana	1	1	0,886	1	1	1
3	DV	1	0,8	0,943	1	1	0,988
4	Variância	1	0,9	0,977	1	0,9	1
5	Curtose	1	0,7	0,965	1	1	1
6	Obliquidade	1	1	1	1	1	1
7	CV	1	1	0,852	1	1	1
8	Interquartil	1	1	1	1	1	1
9	PDS	1	0,9	1	1	0,9	1
10	MDF	1	1	1	0	0,9	1
11	RPP	1	1	1	1	1	1
12	MNF	1	1	0,988	1	1	1
13	RMS	1	0,9	0,965	1	1	1
14	RSS	1	0,9	0,977	1	1	1