UNIVERSIDADE FEDERAL DO CEARÁ
DEPARTAMETNO DE ENGENHARIA DE TELEINFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

THIAGO MONTEIRO NUNES

CLASSIFICAÇÃO DE ARRITMIAS CARDÍACAS EM ELETROCARDIOGRAMA UTILIZANDO FLORESTA DE CAMINHOS ÓTIMOS

FORTALEZA
2014

THIAGO MONTEIRO NUNES

CLASSIFICAÇÃO DE ARRITMIAS CARDÍACAS EM ELETROCARDIOGRAMA UTILIZANDO FLORESTA DE CAMINHOS ÓTIMOS

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Engenharia de Teleinformática da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Teleinformática.

Aprovada em

BANCA EXAMINADORA

Prof. Dr. Victor Hugo Costa de Albuquerque
(Orientador)
Universidade de Fortaleza (Unifor)
Prof. Colaborador

Prof^{a}. Dr^{a}. Fátima Nelsizeuma Sombra de Medeiros
Universidade Federal do Ceará (UFC)

Prof. Dr. Auzuir Ripardo de Alexandria
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

FORTALEZA
2014

THIAGO MONTEIRO NUNES

CLASSIFICAÇÃO DE ARRITMIAS CARDÍACAS EM ELETROCARDIOGRAMA UTILIZANDO FLORESTA DE CAMINHOS ÓTIMOS

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Engenharia de Teleinformática da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Teleinformática.

Orientador: Prof. Dr. Victor Hugo Albuquerque

FORTALEZA
2014

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca de Ciências e Tecnologia

N929c Nunes, Thiago Monteiro.
Classificação de arritmias cardíacas em eletrocardiograma utilizando floresta de caminhos ótimos / Thiago Monteiro Nunes. – 2014.
66 f. : il. color., enc. ; 30 cm.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Departamento de Engenharia de Teleinformática, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2014.
Área de Concentração: Sinais e Sistemas.
Orientação: Prof. Dr. Victor Hugo Costa de Albuquerque.

1. Sistemas de reconhecimento de padrões. 2. Eletrofisiologia. 3. Processamento de sinais. 4. Teleinformática. I. Título.

CDD 621.38

Este trabalho é dedicado ao meu filho Gabriel.

AGRADECIMENTOS

Agradeço a Deus, pois sem ele, nada seria possível. Ao meu Orientador por toda a dedicação e apoio incansável. A minha esposa Ingrid pela paciência e compreensão. A meus pais, Paulo e Graça e ao meu irmão Paulo pela base de valores que me foi concedida. Aos professores Fátima e Auzuir, por aceitarem avaliar este trabalho.

RESUMO

Anualmente no mundo, milhões de pessoas morrem vítimas de cardiopatias que em grande parte, podem ser detectadas através de sinais em eletrocardiograma. Essa análise envolve o estudo do sinal, correspondendo às arritmias estudadas, processo que pode ser automatizado através do aprendizado de máquinas. Esse trabalho compara os classificadores Floresta de Caminhos Ótimos(OPF), utilizando 6 métricas de distâncias, Máquinas de Vetores de Suporte com núcleo de função de base radial (SVM-RBF) e Classificador Bayesiano aplicados problema da classificação de arritmias em eletrocardiogramas, usando 6 técnicas de extração de atributos e uma metodologia de separação de conjuntos para evitar a interferência das informações de pacientes na classificação. A base de dados utilizada foi a MIT-BIH *Arrhythmia Database* e foram avaliados desempenho em termos de taxa de acerto, generalização, através de sensibilidade e especificidade, e custo computacional. Foram consideradas classificações em 5 e 3 classes de arritmias. O OPF mostrou o melhor desempenho em termos de generalização, enquanto o SVM-RBF obteve as maiores taxas de acerto. Os tempos de treino do OPF foram os menores entre os classificadores. No teste, o SVM-RBF foi o classificador que apresentou o menor custo computacional.

Palavras-chaves: Sistema de Reconhecimento de Padrões, Eletrofisiologia, Processamento de sinais, Teleinformática.

ABSTRACT

Currently in the world, millions of people die, victims of heart diseases, which in large part can be detected by analyzing signals of the electrocardiogram. This analysis involves the study of the signal corresponding to the arrhythmia studied and can be automated through machine learning. This work compares the Optimum Path Forest (OPF) classifier using 6 distance metrics, the Support Vector Machines classifier with radial basis function kernel (RBF-SVM) and the Bayesian classifier, applied to the problem of ECG arrhythmias classification. This is done using 6 feature extraction techniques and a methodology for separating sets, to avoid the interference of patient information in classification. The performance is evaluated in terms of accuracy, generalization, through specificity and sensitivity, and computational cost. Classification was done using 5 and 3 classes of arrhythmias. The OPF showed the best performance in terms of generalization, while the SVM-RBF had the highest accuracy rates. The training times of OPF were the lowest among the classifiers. In the test, the RBF-SVM classifier presented best computational cost. **Palavras-chaves**: machine learning, eletro, electrophysiological signals, pattern classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Diagnósticos e investigações clínicas dependem criticamente da capacidade de registro e análise de dados fisiológicos (GOLDBERGER et al., 2000). Nesse aspecto, diversas ferramentas computacionais já auxiliam o diagnóstico médico. Aplicações são desenvolvidas para áreas como eletroencefalograma (NUNES et al., 2014), eletrocardiograma (ABAWAJY; KELAREV; CHOWDHURY, 2013), eletromiograma (PARK; KIM; OH, 2011), detecção e segmentação em tomografias (HEIMANN; MEINZER, 2009), imagens de ultrassom (NOBLE; BOUKERROUI, 2006) entre outras.

A detecção e classificação automática de arritmias em eletrocardiogramas é amplamente estudada e empregada. São vários os métodos computacionais de extração de atributos, utilizados com o auxílio de classificadores para a correta separação em classes de arritmias, auxiliando no diagnóstico de cardiopatias. Entre as ferramentas de aprendizado de máquinas utilizadas atualmente para solucionar problemas de classificação, destaca-se o classificador chamado Floresta de Caminhos Ótimos (OPF)(PAPA; FALCÃO; SUZUKI, 2009). O OPF vem ganhando destaque, superando por muitas vezes, técnicas tradicionais como Máquinas de vetores de suporte e Redes Neurais Artificiais.

Neste sentido, o presente trabalho analisa o desempenho de técnicas baseadas em aprendizado de máquinas, sobretudo do classificador de Floresta de Caminhos Ótimos, como ferramenta de auxílio ao diagnóstico médico, capaz de detectar arritmias cardíacas em eletrocardiograma, facilitando a um especialista, identificar condições de risco a pacientes.

1.1 Justificativa

Doenças cardiovasculares foram a maior causa de mortes ocasionadas por doenças não infecciosas e não transmissíveis em todo o mundo no ano de 2008 (WHO, 2011), matando mais de 14 milhões de pessoas. Esse número representa mais óbitos que os associados a problemas de saúde como câncer, diabetes e problemas respiratórios. No diagnóstico de doenças do coração, o eletrocardiograma é o mais importante sinal biológico, e a detecção de sinais anormais é de fundamental importância para ministrar o tratamento correto ao paciente (GAO et al., 2005). O custo de realização, abrangência e sua característica não invasiva faz do ECG um método de grande disseminação e fácil aplicabilidade em todo o mundo. Uma das maneiras de realizar esse tipo de teste, é registrar por um longo tempo a atividade cardíaca de um indivíduo em sua rotina normal. Com isso obtém-se o registro de uma grande quantidade de batimentos cardíacos a ser analisados. Essa análise, se realizada manualmente, é cansativa e sujeita a erros de interpretação, devido ao grande número de eventos a serem classificados. Nesse contexto o presente trabalho pretende contribuir com a sociedade, analisando sistemas especialistas de classificação para o auxílio do diagnóstico de arritmias em batimentos cardíacos obtidos por ECG, bem como

contribuir com a ciência, avaliando o comportamento do classificador de Floresta de Caminhos Ótimos como ferramenta para classificação de arritmias cardíacas, utilizando diferentes métricas de distâncias, estudando sua influência em conjunto com técnicas de extração de atributos.

1.2 Objetivos

Esse trabalho tem como objetivo estudar o comportamento de algoritmos de classificação Floresta de Caminhos Ótimos (OPF), no domínio da classificação de arritmias em sinais de ECG.

Como objetivos específicos têm-se:

- Avaliar o uso de 6 métricas de distâncias aplicadas ao OPF, no problema de ECG;
- Avaliar o desempenho 6 métodos diferentes de extração de atributos;
- Comparar o desempenho do OPF com algoritmos de classificação tradicionais como Máquinas de Vetores de Suporte e classificador Bayesiano;

1.3 Organização do Trabalho

Esta dissertação é organizada da seguinte maneira: No Capítulo 2, são apresentados aspectos em eletrocardiograma e os trabalhos que já foram realizados com o classificador OPF, e na área de classificação de sinais de ECG.

No Capítulo 3, são descritas a base de dados, as técnicas de extracão de atributos, e os classificadores utilizados para a comparação com o OPF. Também são definidos os parâmetros considerados para avaliar o desempenho e custo computacional do OPF.

Os resultados são apresentados no Capítulo 4, onde os parâmetros definidos no capítulo anterior são mostrados. É realizada uma análise de desempenho e custo computacional do OPF, utilizando 6 métricas de distância, e dos outros algoritmos de classificação, considerando 3 e 5 classes.

Por fim, no Capítulo 5, são apresentadas as conclusões do trabalho e sugeridos temas de trabalhos futuros que possam aprofundar o conteúdo.

1.4 Publicações

1.4.1 Congressos

- NUNES, T. M.; SILVA, R. C. D.; ALBUQUERQUE, V. H. C.; CORTEZ, P. C. *Desempenho preliminar de uma abordagem vetorial sobre métodos de contornos ativos.* **XXIII Congresso Brasileiro de Engenharia Biomédica**, Porto de Galinhas - PE, 2012.
- SILVA, R. C. D.; NUNES, T. M.; PINHEIRO G. J. B.; ALBUQUERQUE, V. H. C. *Customização da Transformada Imagem-Floresta para grafos densos utilizando o algoritmo*

de Floyd-Warshall - um estudo inicial. **XXIII Congresso Brasileiro de Engenharia Biomédica**, Porto de Galinhas - PE, 2012.

- MORYA, E.; NUNES, T. M.; ALBUQUERQUE, V. H. C; SAMESHIMA, K.; NICOLELIS, M. *Spontaneous spatiotemporal single unit activity pattern in rat primary motor sensory cortex.* **XXVIII Reunião Anual da FeSBE**, Caxambú - MG, 2013.

1.4.2 Periódicos

- NUNES, T. M.; Albuquerque V. H. C.; PAPA, J. P.; SILVA, C. C.; NORMANDO, P. G.; MOURA, E. P.; TAVARES, M. R. S. *Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals.* **Expert Systems with Applications**, v. 40, p. 3096-3105, 2013 - Qualis A2 para Engenharias IV.

- LUZ, E. J.; NUNES, T. M.; ALBUQUERQUE V. H. C.; PAPA, J. P.; MENOTTI, D. *ECG arrhythmia classification based on optimum-path forest.* **Expert Systems with Applications**, v. 40, p. 3561-3573, 2013 - Qualis A2 para Engenharias IV.

- NUNES, T. M.; COELHO, A. L. V.; LIMA, C. A. M.; PAPA, J. P.; ALBUQUERQUE, V. H. C. *EEG signal classification for epilepsy diagnosis via optimum path forest–A systematic assessment.* **Neurocomputing**, v. 136, p. 103-123, 2014 - Qualis - B1 para Engenharias IV.

2 FUNDAMENTAÇÃO TEÓRICA

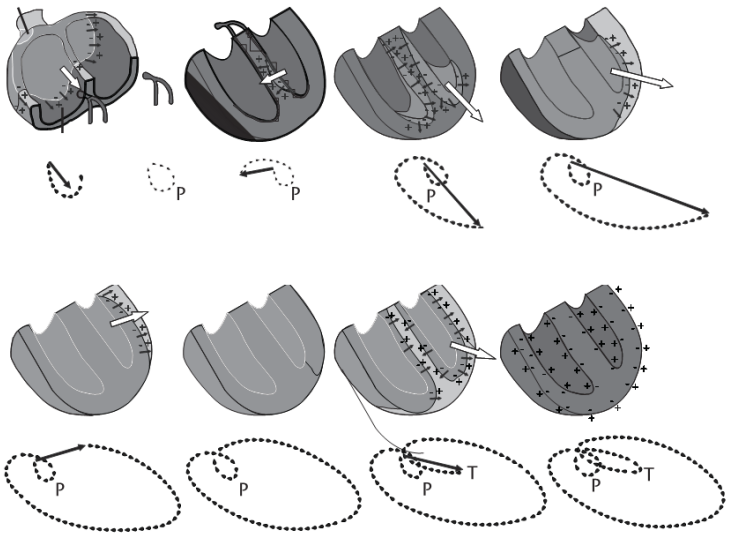
Este capítulo está dividido em três seções. Na primeira seção são descritos aspectos de eletrocardiograma (ECG). Em seguida é apresentado o classificador OPF com suas aplicações e características. Por fim é abordada a classificação em sinais de ECG.

2.1 Aspectos de Eletrocardiograma

O coração é um complexo muscular que, quando contraído, em uma sequencia rítmica, bombeia o sangue, permitindo a sua circulação no organismo. O batimento cardíaco ocorre, devido à contração coordenada desse complexo, estimulado por uma corrente elétrica que, em um batimento normal segue um padrão de propagação repetitivo. Esse padrão de propagação resulta em uma variação de potencial elétrico que pode ser captado através de eletrodos fixados na superfície do corpo humano. Esses sinais, amplificados e filtrados, são conhecidos como eletrocardiograma ou simplesmente ECG (CLIFFORD; AZUAJE; MCSHARRY, 2006).

A atividade elétrica do coração pode ser sintetizada e representada por uma distribuição de polos de corrente, e podem ser representados por um único vetor cardíaco no espaço tridimensional que varia sua magnitude e direção em função do tempo durante a atividade cardíaca. Uma representação da trajetória do vetor cardíaco durante um batimento normal é mostrada na Figura 1.

Figura 1 – Representação da trajetória do vetor cardíaco.



Fonte: Adaptado de (CLIFFORD; AZUAJE; MCSHARRY, 2006)

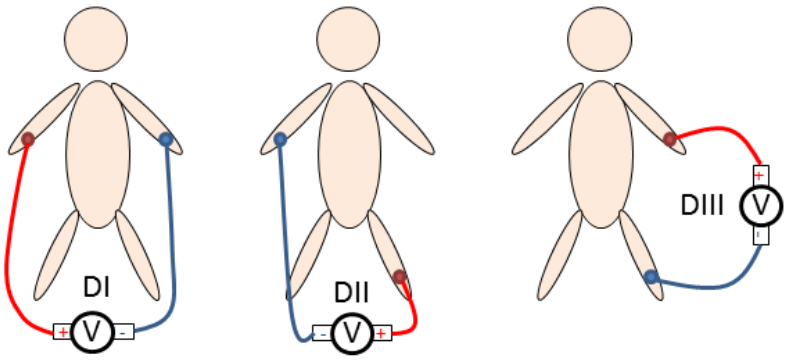
A captação e representação desse vetor, pode ser realizada através da distribuição de ele-

trodos em partes da superfície do corpo. A diferença de potencial entre cada eletrodo e uma referência representa a captação do vetor em uma dimensão, formando um gráfico relacionando potencial e tempo. Cada ponto onde o sinal é captado é chamado de derivação. São doze os tipos de derivações mais comumente utilizadas, divididas em três grupos: derivações periféricas bipolares, derivações periféricas unipolares e derivações pericordiais.

As derivações periféricas bipolares são obtidas através da medição de potencial elétrico entre os pontos do triângulo de Einthoven (CONOVER, 2002), que é composto pelo braço esquerdo, braço direito e perna esquerda. Essas derivações, mostradas na Figura 2, são:

- DI - captada entre o braço direito (+) e o braço esquerdo (-);
- DII - captada entre a perna esquerda (+) e o braço direito (-);
- DIII - captada entre o braço esquerdo (+) e a perna esquerda (-);

Figura 2 – Configuração bipolar do posicionamento dos eletrodos para captação de ECG.



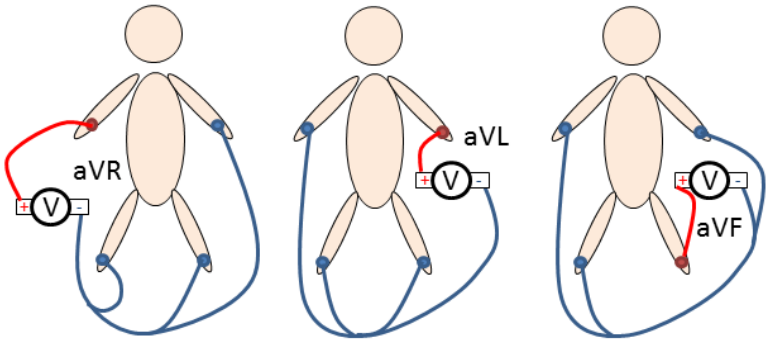
Fonte: Autor

As derivações periféricas unipolares medem o potencial elétrico entre um dos pontos do triângulo de Einthoven e um ponto de potencial zero, chamado de terminal de Wilson (CONOVER, 2002), que é obtido pela união dos pontos dos outros 3 membros em um polo negativo, exemplificado na Figura 3. As três derivações periféricas unipolares são:

- aVR - captada entre o braço direito e o terminal de Wilson;
- aVL - captada entre o braço esquerdo e o terminal de Wilson;
- aVF - captada entre o pé esquerdo e o terminal de Wilson.

As derivações pericordiais medem o potencial entre 6 pontos específicos na superfície do peito do paciente, dados na Figura 4, e um ponto de potencial zero, determinado pelo terminal de Wilson, considerando o braço esquerdo, o braço direito e a perna esquerda. Os pontos onde os eletrodos de medição podem ser colocados são:

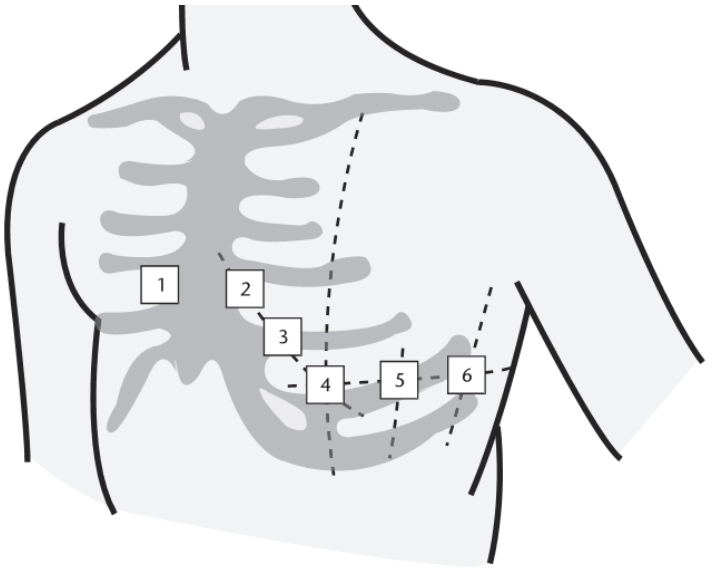
Figura 3 – Configuração unipolar do posicionamento dos eletrodos para captação de ECG.



Fonte: Autor

- V1 - Quarto espaço intercostal, à direita do externo;
- V2 - Quarto espaço intercostal, à esquerda do externo;
- V3 - Ponto entre V2 e V4;
- V4 - Quinto espaço intercostal, na linha clavicular média;
- V5 - Quinto espaço intercostal, na linha axilar anterior;
- V6 - Quinto espaço intercostal, na linha axilar média;

Figura 4 – Configuração pericordial do posicionamento dos eletrodos para captação de ECG.

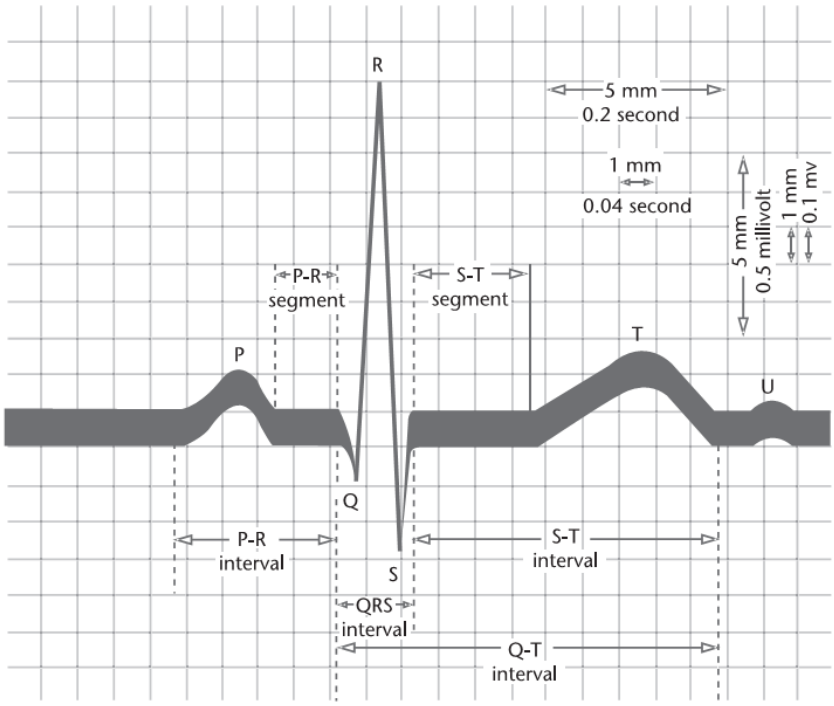


Fonte: (CLIFFORD; AZUAJE; MCSHARRY, 2006)

O sinal de ECG representa uma série de eventos, resultados da polarização e despolarização do tecido cardíaco. Esses eventos, chamados de ondas são nomeadas por letras, sendo as mais importantes, as ondas P, Q, R, S e T. Além das ondas, também são considerados intervalos que as contém, como conjunto das ondas Q, R e S, chamado complexo QRS o intervalo S-T que compreende as ondas S e T e o segmento S-T que compreende o intervalo entre o fim da onda S e o início da onda T.

Na Figura 5, é mostrada a representação de um sinal de ECG com suas ondas e alguns intervalos de interesse.

Figura 5 – Eventos importantes na análise de ECG.



Fonte: Adaptado de (CLIFFORD; AZUAJE; MCSHARRY, 2006)

Diversos motivos podem afetar o funcionamento do coração, alterando o padrão de despolarização e repolarização e, como consequência, refletindo no formato e ocorrência das ondas no eletrocardiograma. Estes batimentos atípicos, também chamados de arritmias, estão fortemente ligados a patologias cardíacas, como por exemplo, síndrome da morte súbita. Assim, é de fundamental importância a detecção e caracterização das arritmias para o diagnóstico cardíaco.

2.2 O classificador OPF

Este trabalho é focado em um classificador, referido como *Optimum Path Forest* (OPF) (PAPA; FALCÃO; SUZUKI, 2009), que vem ganhando crescente atenção nos últimos anos por ter algumas vantagens sobre os classificadores tradicionais. Entre essas vantagens, pode-se citar: a não exigência de parâmetros que possam afetar o desempenho, a não presunção

de formato/separabilidade do espaço de características, o rápido processamento nas fases de treinamentos, mesmo em conjuntos com alta dimensionalidade e decisões baseadas em um critério global.

Além disso, o OPF não interpreta o processo de separação de classes como um conjunto de hiperplanos ótimos, em vez disso, desenvolve caminhos ótimos baseados em amostras chaves, chamadas de protótipos, propagando-se nas outras amostras e penetrando em classes, mesmo que não sejam linearmente separáveis. As amostras se conectam aos caminhos que levam aos protótipos, de maneira a obterem o menor custo, formando assim árvores de caminhos ótimos. Esse processo define uma partição ótima discreta, ou região de influência do espaço de características.

Diversos trabalhos mostram a eficiência e robustez do OPF em dados não lineares e as aplicações OPF vêm sendo estudadas nas mais diversas áreas como detecção de invasão em redes de computadores (PEREIRA et al., 2012a), monitoramento de perfuração de poços de petróleo (GUILHERME et al., 2011), perdas em redes de distribuição elétrica (RAMOS et al., 2011), ocorrência de precipitações pluviométricas (FREITAS et al., 2010) e em grandes bases de dados (PAPA et al., 2012), apresentando bom desempenho em termos de custo computacional.

Na área de processamento de imagens, o desempenho do OPF foi avaliado na classificação de imagens de tecido cerebral obtidas por ressonância magnética (CAPPABIANCO et al., 2012), identificação de parasitas intestinais (SUZUKI et al., 2013), caracterização de partículas de grafite em metalografia (PAPA et al., 2013), classificação de plantas aquáticas (PEREIRA et al., 2012b), imagens oftalmológicas (PAGNIN; ARTIOLI; PAPA, 2011), imagens de satélite (PISANI et al., 2014) e reconhecimento de deslizamento de terra (PISANI et al., 2012).

Na classificação de sinais, o OPF foi utilizado em sinais de ultrassom para estimação de microestruturas em ligas metálicas (NUNES et al., 2013), detecção de falhas em sistemas subterrâneos (SOUZA et al., 2012) e sinais obtidos pela glote para reconhecimento de emoções na fala (ILIEV et al., 2010). Aplicado a sinais fisiológicos, foi utilizado na classificação de sinais de eletroencefalograma para a diagnóstico de epilepsia (NUNES et al., 2014).

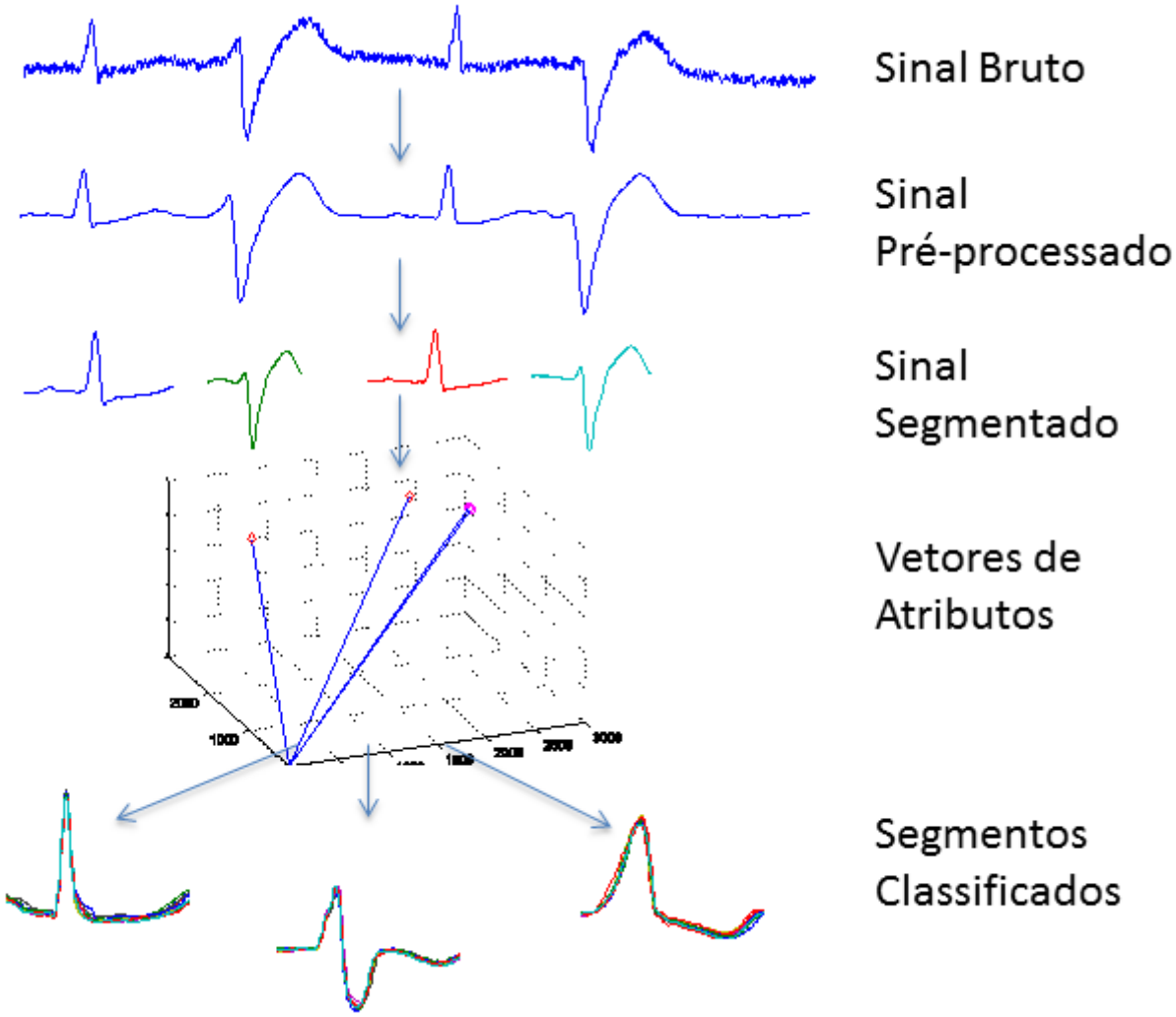
Com estas características de robustez, independência de parâmetros e aplicação anterior em sinais fisiológicos, o OPF pode ser considerado uma alternativa viável para a classificação de arritmias cardíacas em eletrocardiograma.

2.3 Classificação de sinais de ECG

A classificação de arritmias em sinais de ECG pode ser realizada de várias maneiras e muitas abordagens já foram estudadas (KORÜREK; DOGAN, 2010; MAR et al., 2011; ABAWAJY; KELAREV; CHOWDHURY, 2013). Independentemente do algoritmo de classificação utilizado, algumas etapas do processo são de fundamental importância.

Uma representação do processo de classificação de sinais de ECG é mostrada na Figura 6.

Figura 6 – Etapas da classificação de sinais em eletrocardiograma.



Fonte: Autor

2.3.1 Pré-processamento e filtragem

A qualidade da classificação de sinais de ECG depende diretamente da etapa de pré-processamento onde são filtradas as frequências de ruído, e de artefatos que possam interferir na banda de frequência do ECG (MARTIS; ACHARYA; ADELI, 2014).

Junto aos sinais provenientes da passagem de corrente elétrica nos tecidos musculares do coração, outros sinais são captados no ECG. Esses sinais, misturam-se e atrapalham o registro, inserindo ruídos e informações inúteis no sinal que representa o batimento cardíaco. Entre os vários tipos de ruídos que podem interferir no ECG, pode-se destacar (FRIESEN et al., 1990):

- interferência da rede de alimentação elétrica;
- ruído de contato do eletrodo;
- ruído por contrações de outros músculos;

- movimentação da linha de base, devido à movimentação dos eletrodos e mudanças na impedância do corpo;
- modulação devido à movimento de respiração;
- artefatos de movimento.

São várias as técnicas de pré-processamento de sinais em ECG. Pode-se citar o uso de *Empirical Mode Decomposition* (EMD) (KABIR; SHAHNAZ, 2012), Transformada *Wavelet* discreta (DWT) (ADDISON, 2005), filtros digitais (LUO; JOHNSTON, 2010) e Análise de Componentes Independentes (ICA) (HE; GARI; LIONEL, 2006).

2.3.2 Detecção QRS

Após o pré-processamento, é necessário detectar e segmentar cada batimento do sinal de ECG. Para realizar essa tarefa, uma importante etapa é a detecção do complexo QRS, mais especificamente da onda R. Boa parte das técnicas de detecção e segmentação dos batimentos cardíacos se baseiam na localização do complexo QRS como referência. O complexo QRS é composto pelas ondas Q, R e S, mostradas na Figura 5, apresentada no Capítulo 1.

Devido ao acentuado coeficiente angular e amplitude da onda R, o complexo QRS torna-se mais evidente que qualquer outra parte do sinal de ECG, sendo mais fácil de detectado para posterior segmentação. A partir dos complexos QRS, mais especificamente da onda R, obtém-se o intervalo RR que é o período entre duas ondas R consecutivas. Esse intervalo é utilizado na análise da frequência cardíaca, sendo um dos mais importantes parâmetros da análise e detecção de outros eventos do sinal de ECG.

Entre as técnicas de detecção do complexo QRS pode-se citar desde técnicas mais triviais como derivadas de sinais (AHLSTROM; TOMPKINS, 1983) , bancos de filtros (AFONSO et al., 1999) e morfologia matemática (TRAHANIAS, 1993), como métodos mais elaborados como redes neurais (HU; PALREDDY; TOMPKINS, 1997), algoritmos genéticos (POLI; CAGNONI; VALLI, 1995) e transformadas Wavelet e de Hilbert (MADEIRO et al., 2012).

2.3.3 Extração de Atributos

Boa parte da informação, contida no sinal de ECG, está relacionada com amplitudes, duração e local de ocorrência dos picos P, QRS e T. Especialistas interpretam mudanças nesses parâmetros para a interpretação dos sinais de ECG (MARTIS; ACHARYA; ADELI, 2014). Muitas informações, que podem ser clinicamente relevantes não estão evidente no sinal de ECG (GOLDBERGER et al., 2000), sendo necessário o emprego de extratores de atributos.

Atributos de sinais em ECG podem incluir dados de frequência, tempo, morfologia, energia e características do batimento em relação a outros batimentos, como o intervalo RR. Essas características do sinal são responsáveis por caracterizar os batimentos e diferenciá-los, de maneira a serem separados por uma ferramenta de classificação.

O método de extração depende muito do classificador que irá utilizar os atributos, sendo muito dependente dele. Além de isolar e realçar atributos que permitam uma melhor classificação dos sinais de ECG, a técnica de extração de atributos também tem a função de reduzir a dimensionalidade do sinal para o espaço de atributos. Alguns algoritmos de classificação, como redes neurais, têm limitações em trabalhar com uma grande dimensionalidade do espaço de atributos, sendo altamente dependente das técnicas de extração de características.

Entre as várias técnicas de extração propostas, boa parte utiliza-se da Transformada Wavelet Discreta (GÜLER; ÜBEYLI, 2005; SONG et al., 2005; YU; CHEN, 2007; YU; CHOU, 2008; YE; COIMBRA; KUMAR, 2010; CHEN, 2012; RAI; TRIVEDI; SHUKLA, 2013), sendo que em muitos casos combina-se com outras técnicas computacionais como Análise de Componentes Independentes (YE; COIMBRA; KUMAR, 2010), Análise de Discriminantes Lineares (SONG et al., 2005) e características de intervalo RR (YU; CHOU, 2008; YE; COIMBRA; KUMAR, 2010; YU; CHEN, 2007). São também utilizados como extratores de atributos para sinais de ECG, Análise de Componentes Principais (MARTIS et al., 2012; MARTIS et al., 2013; ABAWAJY; KELAREV; CHOWDHURY, 2013), correlação cruzada (DUTTA; CHATTERJEE; MUNSHI, 2010), espectro de potência (KHAZAEE; EBRAHIMZADEH, 2010) e atributos geométricos baseados em imagem (HOMAEINEZHAD et al., 2012).

2.3.4 Classificação

Na etapa de classificação, os vetores de atributos são agrupados por classe, de acordo com suas características em comum. Esta tarefa é realizada, em muitos casos, através de técnicas de aprendizado de máquinas. A partir de um conjunto vetores rotulados quanto ao tipo, chamado conjunto de treinamento, pode ser gerado um modelo de classificador com o objetivo de prever os rótulos de um conjunto de batimentos desconhecidos em um conjunto de teste.

As técnicas mais utilizadas em classificação de sinais de ECG são classificador baseado em Máquinas de Vetores de Suporte (SONG et al., 2005; YE; COIMBRA; KUMAR, 2010; ABAWAJY; KELAREV; CHOWDHURY, 2013; DUTTA; CHATTERJEE; MUNSHI, 2010; KHAZAEE; EBRAHIMZADEH, 2010; DAAMOUCHEA et al., 2012) e Redes Neurais Artificiais (GÜLER; ÜBEYLI, 2005; YU; CHEN, 2007; YU; CHOU, 2008; KORÜEK; DOĜAN, 2010; CHEN, 2012; NEJADGHOLI; MOHAMMAD; ABDOLALI, 2011; RAI; TRIVEDI; SHUKLA, 2013; MARTIS et al., 2012; WANG et al., 2013), e em alguns trabalhos, os dois métodos são comparados na avaliação de técnicas de extração de atributos (MARTIS et al., 2013; MOAVENIAN; KHORRAMI, 2010). Outras técnicas como Discriminantes Lineares (CHAZAL; O'DWYER; REILLY, 2004) e a hibridização entre Máquinas de Vetores de Suporte e Redes neurais Artificiais (HOMAEINEZHAD et al., 2012) também são aplicadas na classificação desses sinais.

2.3.4.1 Formação dos conjuntos de treino e teste

No trabalho de Chazal, O'Dwyer e Reilly (2004), é proposta uma metodologia para agrupar os sinais de ECG de forma a evitar a influência da discriminação do paciente na classificação dos sinais de ECG. Esta metodologia consiste na divisão dos registros de ECG, de maneira que nenhum paciente utilizado no conjunto de treino é usado no conjunto de teste, evitando dessa maneira, a influência do paciente na classificação do sinal. Luz e Menotti (2011) mostrou que os resultados relatados por alguns trabalhos, nos quais a escolha dos sinais nos conjuntos era aleatória, diminuíram significativamente quando submetidos ao modelo de separação proposto. Essa dependência do paciente na classificação também é mostrada em Nejadgholi, Mohammad e Abdolali (2011).

3 METODOLOGIA

Neste capítulo são apresentadas as técnicas computacionais utilizadas para classificar os sinais referente aos batimento cardíacos dos ECG. Primeiramente, a base de dados MIT-BIH (Massashusetts Institute of Technology - Boston Beth Israel Hospital) *Arrhythmia Database* (MARK et al., 1982) é descrita abordando as considerações da norma ANSI/AAMI EC57 (ANSI/AAMI, 2008), que padroniza a avaliação de ferramentas computacionais para classificação de bases de arritmias cardíacas. Em seguida, são descritas as técnicas de extração de atributos utilizadas para a geração dos vetores de atributos a serem classificados. Após a descrição dos métodos de extração, são apresentadas as técnicas de classificação, em especial, o classificador de Floresta de caminhos ótimos. Por fim, são descritos os parâmetros estatísticos utilizados na avaliação de desempenho dos classificadores e os recursos computacionais utilizados.

3.1 Descrição da Base MIT-BIH *Arrhythmia Database*

Para a avaliação dos classificadores, foi utilizada a base de dados de arritmias cardíacas MIT-BIH *Arrhythmia Database* (MARK et al., 1982) e disponibilizada *online* por Goldberger et al. (2000), recomendada na norma da ANSI/AAMI EC57 (ANSI/AAMI, 2008).

A base de dados MIT-BIH *Arrhythmia Database*, aqui denominada simplesmente de MIT-BIH, é uma base de sinais provenientes de eletrocardiograma (ECG) amplamente utilizada para avaliação do desempenho de algoritmos para a detecção de arritmias (MOODY; MARK, 2001). Os dados consistem em 48 trechos de 30 minutos de registro de sinais de ECG, retirados de gravações de 24 horas de duração.

Os sinais de ECG foram amostrados em dois canais, onde um é obtido através de eletrodos posicionados no peito do paciente, em uma derivação modificada DII localizada entre o braço direito e a perna esquerda, e o outro canal, geralmente no ponto pericordial V1, podendo ter sido captado também nos pontos V2, V4 ou V5, dependendo do paciente.

Os registros foram adquiridos através de equipamento *Holter*, (Delmar Avionics, modelo 455, Estados Unidos). Os sinais são referentes a 47 pacientes, captados entre os anos de 1975 e 1979, no Laboratório de Arritmia do *Boston's Beth Israel Hospital*.

Cada trecho é identificado por um número. Do total de registros, 23 foram escolhidos aleatoriamente de um banco de 4000 registros de pacientes. A identificação desses registros é dada por um número menor que 200. Outros 25 registros são de pacientes escolhidos, contendo batimentos patológicos, identificados por um número maior que 199. Os pacientes tinham idade entre 23 e 89 anos sendo 22 do sexo feminino e 25 do sexo masculino.

Os registros analógicos das fitas Holter foram digitalizados a uma taxa de amostragem de $360Hz$ e os batimentos marcados e classificados manualmente por especialistas em 15 classes, quanto ao tipo de arritmia. Os tipos de arritmia identificadas na base são mostradas na Tabela

1. As indicações de especialistas permitem a utilização desta base na avaliação de algoritmos para classificação de arritmias em ECG.

Tabela 1 – Descrição das classes consideradas nas anotações da base MIT-BIH e na norma AAMI.

Classe AAMI	Classe MIT-BIH	Tipo de batimento MIT-BIH
Normal (N)	N	Batimento normal
	L	Bloqueio de ramo esquerdo
	R	Bloqueio de ramo direito
	e	Escape atrial
	j	Escape nodal
Batimento supraventricular ectópico (SVEB)	A	Batimento atrial prematuro
	a	Batimento atrial prematuro aberrado
	J	Batimento prematuro nodal
	S	Batimento supraventricular prematuro
Batimento ventricular ectópico (VEB)	V	Contração ventricular prematura
	E	Escape ventricular
Batimento de fusão (F)	F	Fusão de batimento ventricular e normal
Batimento desconhecido (Q)	/	Batimento de marca-passo
	f	Fusão de batimentos normal e de marca-passo
	Q	Batimento inclassificável

Fonte: Autor

Apesar de importante, as etapas de detecção e segmentação dos batimentos em sinais de ECG não são objetivos deste trabalho, e por isso, são utilizadas as marcações informadas na base para localizar os batimentos.

Na base de dados, existem 4 registros oriundos de pacientes que utilizavam marca-passo. Estes registros não foram considerados por recomendação da norma ANSI/AAMI EC57 (ANSI/AAMI, 2008), que também recomenda que as 15 classes informadas nas anotações da base sejam agrupadas em 5 classes, conforme apresentado na Tabela 1.

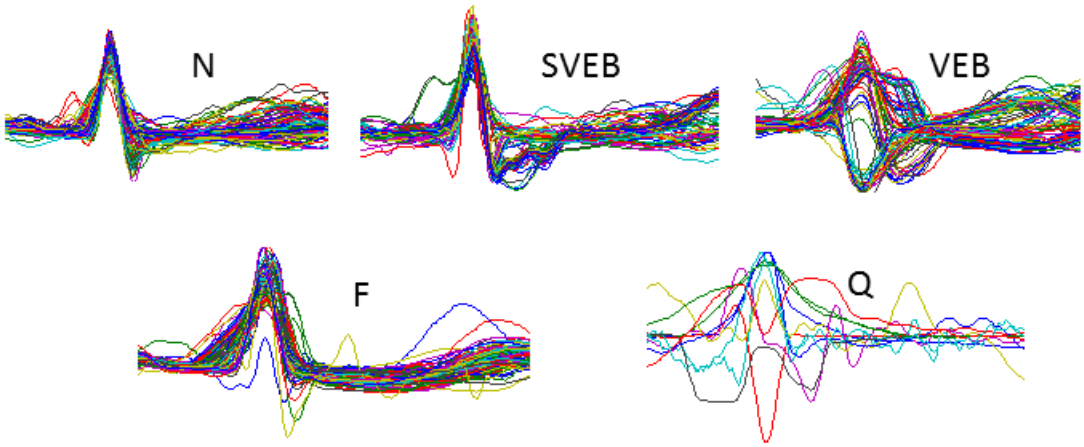
Com a exclusão dos pacientes portadores de marca-passo, os 44 registros considerados não contém batimentos dos tipos $/$ e f.

Na Figura 7, são ilustrados os batimentos correspondentes aos sinais de ECG, sendo representados 10 para da classe Q e 100 das outras classes. Os sinais foram escolhidos de maneira aleatória na base.

3.2 Composição dos conjuntos de treino e teste

A base foi dividida em dois conjuntos de sinais separando os pacientes, utilizados para treinamento e para teste dos algoritmos de classificação. Assim, é possível avaliar o desempenho dos classificadores evitando que a classificação ocorra em função do paciente e não do tipo de sinal (LUZ; MENOTTI, 2011). A formação dos conjuntos, com os respectivos pacientes, é baseada no trabalho de Chazal, O'Dwyer e Reilly (2004), que propõe uma separação com o balanceamento entre batimentos de treino e teste para cada classe. Esta separação é apresentada

Figura 7 – Sinais de batimentos cardíacos da base MIT-BIH das classes consideradas em ANSI/AAMI (2008).



fonte: Autor

na Tabela 2, onde os pacientes com identificação iniciando em 100 representam os registros escolhidos aleatoriamente na formação da base e os que iniciam com 200 são os pacientes escolhidos por conterem batimentos patológicos relevantes (MOODY; MARK, 1989).

Tabela 2 – Divisão dos registros nos conjuntos de treino e teste.

Conjunto	Registros
Treino	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223 e 230
Teste	100, 103, 105, 11, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233 e 234

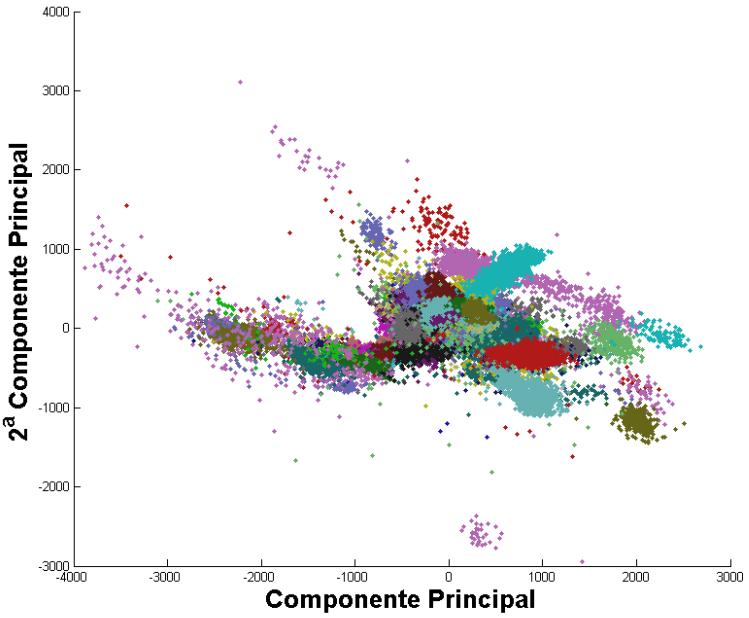
Fonte: Autor

Na Figura 8 é mostrada a representação das duas principais componentes dos sinais de batimentos, obtidas por análise de componentes principas (PCA) onde cada paciente é plotado utilizando-se uma cor diferente. Isto ilustra a dependência dos sinais em relação aos pacientes.

É possível observar o particionamento dos dados em função dos pacientes, sendo fundamental a separação destes nos conjuntos de treino e teste. Uma abordagem detalhada deste assunto é mostrada no trabalho de Luz e Menotti (2011), que relata a influência da não separação dos pacientes entre treino e teste nas taxas de acerto.

Além da divisão dos batimentos em 5 classes, definidas em ANSI/AAMI (2008), é considerada também a classificação de batimentos proposta por Llamedo e Martínez (2011), que considera a divisão em três classes principais de (ANSI/AAMI, 2008), sendo essas N, SVEB

Figura 8 – PCA dos sinais de batimentos de ECG em função dos pacientes.



fonte: Autor

e VEB. As classes F e Q, menos significativas, foram agregadas à classe VEB, aumentando sensivelmente sua representatividade.

3.3 Métodos de extração de atributos

Os métodos de extração de atributos, utilizados para a formação dos vetores característicos, de forma a representar os batimentos de ECG nas classificação foram escolhidos com base em Luz e Menotti (2011), por terem sido estudadas utilizando os critérios de separação. Os métodos são baseados em Transformada *Wavelet* Discreta (DWT), Análise de Componentes Independentes (ICA), Análise de Componentes Principais (PCA) e informações sobre o intervalo RR, que é a distância entre os picos de duas ondas R consecutivas em um sinal de ECG.

Os métodos considerados são:

- Chazal, O'Dwyer e Reilly (2004) - morfologia do sinal e intervalos RR;
- Güler e Übeyli (2005) - DWT;
- Song et al. (2005) - DWT;
- Yu e Chen (2007) - DWT, intervalo RR e energia do sinal;
- Yu e Chou (2008) - DWT, ICA e intervalo RR;
- Ye, Coimbra e Kumar (2010) - DWT, ICA, PCA e intervalo RR.

A distribuição dos batimentos cardíacos, por classe e método de extração, nos conjuntos de treino e teste, para as 5 classes referidas em ANSI/AAMI (2008) é mostrada na Tabela 3, e na Tabela 4 é mostrada a distribuição em 3 classes proposta por Llamedo e Martínez (2011).

T_b representa o total de batimentos do conjunto e n_f o número de atributos extraídos com cada técnica. Para fins de simplificação, foi adotado uma letra de A a E para representar cada conjunto extraído, ou seja:

- conjunto A - (CHAZAL; O'DWYER; REILLY, 2004);
- conjunto B - (GÜLER; ÜBEYLI, 2005);
- conjunto C - (SONG et al., 2005);
- conjunto D - (YU; CHEN, 2007);
- conjunto E - (YU; CHOU, 2008);
- conjunto F - (YE; COIMBRA; KUMAR, 2010).

Tabela 3 – Descrição das bases considerando 5 classes (ANSI/AAMI, 2008)

				Classe de batimentos					
	Conjunto	Técnica de Extração	n_f	N	$SVEB$	VEB	F	Q	T_b
Treino	A	Chazal, O'Dwyer e Reilly (2004)	155	45747	940	3777	415	8	50887
	B	Güler e Übeyli (2005)	19	45845	943	3788	415	8	50999
	C	Song et al. (2005)	21	45825	943	3788	414	8	50978
	D	Yu e Chen (2007)	13	45844	943	3788	415	8	50998
	E	Yu e Chou (2008)	31	45511	929	3770	412	8	50630
	F	Ye, Coimbra e Kumar (2010)	100	45844	943	3788	415	8	50998
Teste	A	Chazal, O'Dwyer e Reilly (2004)	155	44181	1786	3218	388	7	49580
	B	Güler e Übeyli (2005)	19	44238	1836	3221	388	7	49690
	C	Song et al. (2005)	21	44218	1836	3219	388	7	49668
	D	Yu e Chen (2007)	13	44238	1836	3221	388	7	49690
	E	Yu e Chou (2008)	31	43905	1823	3197	388	7	49320
	F	Ye, Coimbra e Kumar (2010)	100	44238	1836	3221	388	7	49690

Fonte: Autor

Deve-se ressaltar que a variação no número de batimentos entre os métodos é consequência das técnicas de extração, que, muitas vezes, não permitem o uso de toda a base. As amostras das extremidades não contêm alguns segmentos ou amostras vizinhas suficientes para, em alguns casos, realizar a correta extração do atributo, por exemplo: média dos sinais RR nos 10 sinais da vizinhança. Sendo assim, é necessário desconsiderar algumas amostras iniciais e finais de cada paciente para extrair os atributos corretamente.

As técnicas de extração de atributos implementadas são descritas detalhadamente a seguir.

Tabela 4 – Descrição das bases considerando 3 classes (LLAMEDO; MARTÍNEZ, 2011)

				Classe de batimentos			
	Conjunto	Método	n_f	N	$SVEB$	VEB	T_b
Treino	A	Chazal, O'Dwyer e Reilly (2004)	155	45747	940	4200	50887
	B	Güler e Übeyli (2005)	19	45845	943	4211	50999
	C	Song et al. (2005)	21	45825	943	4210	50978
	D	Yu e Chen (2007)	13	45844	943	4211	50998
	E	Yu e Chou (2008)	31	45511	929	4190	50630
	F	Ye, Coimbra e Kumar (2010)	100	45844	943	4211	50998
Teste	A	Chazal, O'Dwyer e Reilly (2004)	155	44181	1786	3613	49580
	B	Güler e Übeyli (2005)	19	44238	1836	3616	49690
	C	Song et al. (2005)	21	44218	1836	3614	49668
	D	Yu e Chen (2007)	13	44238	1836	3616	49690
	E	Yu e Chou (2008)	31	43905	1823	3592	49320
	F	Ye, Coimbra e Kumar (2010)	100	44238	1836	3616	49690

Fonte: Autor

3.3.1 Chazal, O'Dwyer e Reilly (2004)

O método de extração de atributos, proposto por Chazal, O'Dwyer e Reilly (2004) para a classificação de batimentos de ECG é baseado nas características de intervalos RR, intervalo entre ondas no sinal e na morfologia do sinal de ECG, totalizando 155 atributos.

Os 7 atributos relativos aos eventos que ocorrem nos batimentos são:

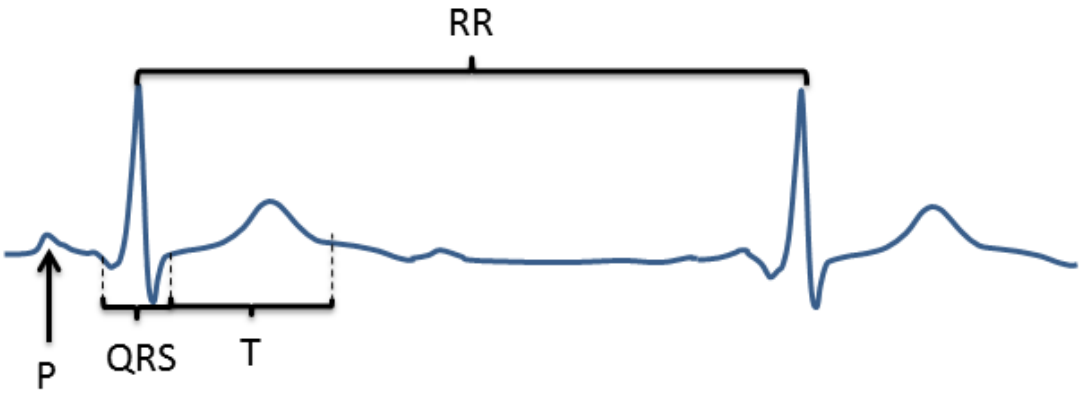
- Intervalo pré-RR - intervalo RR em relação ao batimento anterior;
- Intervalo pós-RR - intervalo RR em relação ao batimento posterior;
- Intervalo RR médio - média dos intervalos RR do paciente;
- Intervalo RR local - média dos intervalos RR na vizinhança do batimento analisado;
- Duração QRS - intervalo entre o início e fim do complexo QRS;
- Duração da onda T - intervalo entre o fim do complexo QRS e o fim da onda T;
- Presença da Onda P - atributo booleano indicando presença ou não da onda P.

Para ilustrar esses atributos, os principais eventos são mostrados na Figura 9.

Os 148 atributos relativos à morfologia do sinal de ECG, sendo 74 em cada canal captado durante o ECG, foram:

- Morfologia QRS: 10 amostras, uniformemente espaçadas, durante o complexo QRS;
- Morfologia T: 9 amostras, uniformemente espaçadas, entre o fim do complexo QRS e fim da onda T;

Figura 9 – Ondas no sinal de ECG relacionados aos atributos extraídos por Chazal, O'Dwyer e Reilly (2004).



Fonte: Autor.

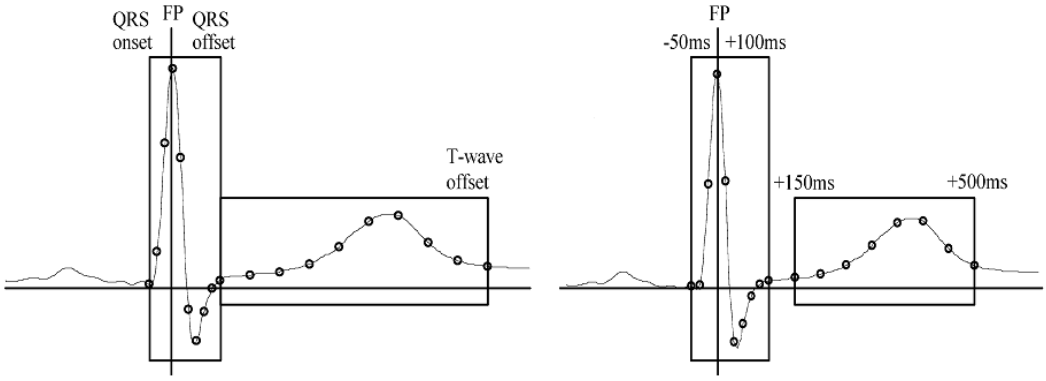
- Morfologia normalizada QRS: mesmo método aplicado à morfologia QRS, porém em sinal normalizado, de maneira a tornar o desvio padrão da amostra unitário;
- Morfologia normalizada T: mesmo método aplicado à morfologia ST, porém em sinal normalizado;
- Morfologia PF-QRS: 10 amostras, uniformemente espaçadas, coletadas entre -50 milissegundos e +100 milissegundos do pico R que é considerado o ponto fiducial (PF);
- Morfologia PF-T: 8 amostras, uniformemente espaçadas, coletadas entre +150 milissegundos e +500 pico R, uniformemente espaçadas;
- Morfologia normalizada PF-QRS: mesmo método aplicado à morfologia PF-QRS, porém em sinal normalizado;
- Morfologia normalizada PF-T: mesmo método aplicado à morfologia PF-ST, porém em sinal normalizado;

Uma representação dos atributos relativos à morfologia do sinal de ECG é mostrada na Figura 10.

3.3.2 Güler e Übeyli (2005)

O método de extração de atributos proposto por Güler e Übeyli (2005) é baseado na análise espectral dos sinais de ECG através de Transformada Wavelet Discreta (DWT). A Transformada *Wavelet* é utilizada por ser eficiente na separação componentes em um sinal, que pode ser considerado uma superposição de estruturas ocorrendo em tempos e frequências diferentes (GÜLER; ÜBEYLI, 2005). Como a DWT faz a decomposição do sinal, considerando os domínios do tempo e da frequência, é aplicada satisfatoriamente para a discriminação do sinal

Figura 10 – Atributos morfológicos do sinal de ECG (CHAZAL; O'DWYER; REILLY, 2004).



Fonte: Chazal, O'Dwyer e Reilly (2004).

proveniente de ECG. A técnica utiliza a *wavelet Daubechies* de ordem 2 (db2). O método utiliza os coeficientes de detalhes dos 4 primeiros níveis e os de aproximação do quarto nível.

Uma vez obtidos os conjuntos de coeficientes, são extraídos parâmetros estatísticos, afim de obter um vetor de atributos de pequena dimensionalidade. Estes parâmetros, obtidos através da Transformada *Wavelet*, são:

- média dos coeficientes para cada sub-banda (nível) (5 atributos);
- energia média dos coeficientes para cada sub-banda (5 atributos);
- desvio padrão dos coeficientes para cada sub-banda (5 atributos);
- proporção da média absoluta entre duas bandas adjacentes (4 atributos).

Os dois primeiros grupos de atributos são relativos à distribuição de frequência no sinal e os dois últimos relativos às variações nessa distribuição. No total são extraídos 19 atributos representando as características de frequência do ECG.

3.3.3 Song et al. (2005)

A técnica empregada em Song et al. (2005) é baseada na DWT.

O ruído do sinal é tratado, utilizando a DWT empregando a *wavelet Haar*. A banda de frequência de aproximação de oitavo nível é subtraída da banda de aproximação de segundo nível. Com isso mantém-se as frequências entre 0,7 e 45 Hz para frequência de amostragem da base MIT-BIH de 360 Hz.

Após a filtragem, um segmento do sinal de cada batimento com intervalo entre -200 ms e +200 ms do pico da onda R, composto por 144 amostras, é submetido novamente à DWT, utilizando a *wavelet Haar*. Considerando os coeficientes de detalhe dos níveis 4, 5, 6 e 7, contendo 9, 5, 3 e 2 coeficientes, respectivamente, obtém-se um subconjunto de 19 atributos. No vetor de

características são consideradas também os resultados das divisões entre uma constante, K, e as durações dos intervalos RR do batimento em relação às ondas R e dos batimentos anterior e posterior.

A constante K, que representa um intervalo RR, esperado para um indivíduo normal, de 300 amostras para a frequência de amostragem de 360 Hz (SONG et al., 2005). Esses dois atributos do intervalo RR, somando-se aos 19 obtidos por DWT, compõem um vetor de 21 atributos.

3.3.4 Yu e Chen (2007)

A técnica empregada por Yu e Chen (2007), também utiliza a DWT aplicando a *wavelet Haar* sob a argumentação de ser a menor e a mais simples das *wavelets*, atendendo satisfatoriamente a discriminação em sinais de ECG. Só são utilizados os coeficientes de detalhe de primeiro e segundo nível e os coeficientes de aproximação de segundo nível.

Os coeficientes em cada banda geram os seguintes atributos:

- energia média (representada pela variância);
- coerência do sinal (representada pela variância da autocorrelação);
- morfologia do sinal (representada pela amplitude relativa do sinal).

Os 3 atributos aplicados às três conjuntos de coeficientes wavelets, no sinal de ECG compõem um conjunto de 9 atributos. Além destes, são considerados outros atributos, tais como:

- variância QRS (variância calculada no sinal durante o complexo QRS);
- média dos intervalos RR em torno da amostra;
- RR posterior (duração do intervalo RR entre o batimento e o batimento posterior);
- RR anterior (duração do intervalo RR entre o batimento e o batimento anterior).

Estes 4 atributos relativos ao sinal e os 9, derivados da DWT, formam um vetor de atributos de dimensionalidade 13.

3.3.5 Yu e Chou (2008)

Na metodologia proposta por Yu e Chou (2008), são utilizadas informações sobre o intervalo RR e Análise de Componentes Independentes (ICA), que consiste em uma técnica de análise de sinais cujo o objetivo é representar um conjunto de variáveis aleatórias como combinações lineares de componentes de variáveis estatisticamente independentes (YU; CHOU, 2008).

Inicialmente, cada sinal de batimento é segmentado em 200 amostras em torno do pico R e subtraído do valor médio para retirar a componente de frequência zero. Em seguida, cada

segmento é normalizado, dividindo-se o sinal por seu desvio padrão, obtendo-se assim desvio padrão 1.

Para estimar os atributos relacionados à técnica de ICA, são utilizados 2 segmentos extraídos de cada registro, formando uma matriz de ordem $n \times m$, em que n é duas vezes o número de registros de ECG (2 segmentos por registro) e m o tamanho do segmento do batimento (200 amostras). Uma vez montada a matriz, as componentes independentes são calculadas utilizando o algoritmo *fast-ICA* (HYVÄRINEN, 1999) e, em seguida, os sinais são projetados nas bases geradas. Com as componentes projetadas ordenadas, em função do peso de ponderação, são preservadas somente as 30 primeiras, sendo estas as mais significativas que com o valor da duração do intervalo RR em relação ao batimento anterior forma um conjunto de 31 atributos.

3.3.6 Ye, Coimbra e Kumar (2010)

Na metodologia descrita em Ye, Coimbra e Kumar (2010), que considera a *wavelet daubechies* de oitava ordem (db8), são extraídos 64 coeficientes de detalhe dos níveis de decomposição 3 e 4 bem como 50 coeficientes de aproximação do nível 4. Os coeficientes são calculados com base em um segmento de 300 amostras do batimento.

Além dos coeficientes da DWT, 9753 batimentos normais são utilizados para criar 18 bases de ICA, que geram, para cada batimento, 18 coeficientes. Estes , concatenados com os coeficientes resultantes da DWT são concatenados e têm sua dimensionalidade reduzida por PCA, para 48 atributos, afim de conter a maior parte da variância. Esse processo é realizado nos dois canais dos registros, compondo 96 atributos.

São considerados ainda 4 atributos relativos aos intervalos RR, sendo eles:

- intervalo RR do batimento em relação ao batimento anterior;
- intervalo RR do batimento em relação ao batimento posterior;
- média de intervalo RR entre os 10 batimentos que cercam o batimento analisado;
- média de intervalo RR entre as batidas contidas nos 5 minutos de amostragem que cercam o batimento analisado.

Considerando todos os atributos, têm-se um vetor de atributos de dimensionalidade 100.

3.4 Algoritmos de classificação

Nessa seção serão apresentados, os classificadores Floresta de Caminhos Ótimos (OPF), Máquinas de Vetores de Suporte (SVM) e Bayesiano (BC), utilizados para comparação de desempenho com o OPF.

3.4.1 Floresta de Caminhos Ótimos

O classificador Floresta de Caminhos Ótimos, ou *Optimum Path Forest* (OPF), modela problema de reconhecimento de padrões como partições de um grafo em um dado espaço de características, onde nós representam vetores de atributos e todos os pares de nós são conectados por arcos, definindo assim um grafo completo (NUNES et al., 2014). Os arcos entre as amostras contêm pesos que variam com as distâncias estabelecidas entre seus respectivos vetores de atributos.

O OPF estabelece um processo de competição entre algumas amostras de referência, determinadas durante o treinamento, chamadas de protótipos, na busca de particionar o grafo em uma floresta de caminhos ótimos (PAPA; FALCÃO; SUZUKI, 2009). Amostras de uma mesma árvore, estão mais fortemente conectadas a um protótipo do que a outras árvores. Essa ligação é definida pelo custo do arco que conecta os dois pontos no grafo, custo esse definido por uma função de custo pré-definida. Cada amostra conectada carrega consigo o seu custo. Amostras a se conectarem à árvore, terão esse custo herdado, ou um custo de distância computado, o que for maior. Assim, durante o processo de competição, cada nó que não é protótipo, será conquistado pela árvore que apresentar o menor custo de ligação, assumindo assim a identidade da árvore, que por sua vez é a identidade da classe definida pelo protótipo.

Considerando Z_{TR} e Z_{TS} os conjuntos de vetores de atributos das amostras de treinamento e teste, respectivamente, em que as classes de Z_{TR} são conhecidas, e sendo o subconjunto $S \subseteq Z_{TR}$ o conjunto de protótipos definidos em Z_{TR}, o OPF mapeia uma floresta de caminhos ótimos à partir de S, tendo como base seu espaço de características, alocando os outros elementos de Z_{TR}, em que cada elemento $s \in Z_{TS}$ pode ser associado à uma classe de uma amostra de Z_{TR}. A floresta de caminhos ótimos é calculada utilizando o algoritmo da transformada imagem floresta (Image Forest Transform - IFT) (FALCÃO; STOLFI; LOTUFO, 2004). A seguir são detalhados os algoritmos de treino e classificação do OPF.

Para processar as classificações nesse trabalho foi utilizada a versão 2.0 da biblioteca LibOPF (PAPA; FALCÃO; SUZUKI, 2009).

3.4.1.1 Treinamento com o OPF

Durante o treinamento supervisionado, utilizando amostras do conjunto de treinamento, Z_{TR}, são definidas inicialmente as amostras de protótipo, ou raízes. Essas amostras são obtidas através do algoritmo da árvore de menor espalhamento (*Minimum Spanning Tree* - MST). Com isso, computa-se um grafo cíclico, em que cada amostra em Z_{TR}, representada por um nó, pertence ao grafo e está conectada a apenas dois outros nós, de maneira a obter o caminho no com o menor custo total. Uma vez computada a MST, onde ocorrerem dois nós de classes diferentes, esses são considerados protótipos pertencentes ao subconjunto de protótipos, S. Às amostras do conjunto S é atribuído custo zero e às amostras não pertencentes à S é atribuído custo infinito.

Uma vez que os protótipos são escolhidos, é calculado o custo de ligação entre as amostras $s \notin S$ e todas as amostras do conjunto de treinamento, através de uma função de conectividade. Após a conquista das amostras pelas ligações que representem menor custo, são formados os caminhos ótimos. O caminho é considerado ótimo se seu custo for menor que o custo para qualquer outro caminho. O OPF utiliza como função de conectividade a função f_{max}, mostrada na equação 3.1, para custo do caminho entre a amostra $\mathbf{s}$ e $\mathbf{t}$.

$$f_{max}(\pi \cdot \langle \mathbf{s}, \mathbf{t} \rangle) \quad = \quad \max\{f_{max}(\pi), d(\mathbf{s}, \mathbf{t})\}, \tag{3.1}$$

em que, $f_{max}(\pi)$ representa a maior distância entre duas amostras adjacentes ao longo do caminho π, e $d(\mathbf{s}, \mathbf{t})$ é a medida da distância entre os vetores $\mathbf{s}$ e $\mathbf{t}$. Outras funções de custo também podem ser utilizadas como função de conectividade, desde que sejam uma funções suaves.

Por padrão o OPF utiliza a norma euclideana como métrica de distância, apresentada na equação 3.2.

$$d(\vec{s}, \vec{t}) = ||\vec{s} - \vec{t}||. \tag{3.2}$$

Outras funções de distância podem ser utilizadas pelo OPF, sendo pré-computadas em uma matriz de distâncias entre todas as amostras, antes do treinamento.

Durante o processo de competição, entre as amostras, a elas são atribuídos os custos do menor caminho, calculado pela equação 3.3, para o custo de uma amostra t.

$$C(t) = min\{f_{max}(\pi_t) \forall \pi_t \in Z_{TR}\}. \tag{3.3}$$

Ao calcular o caminho ótimo para cada amostra t, é atribuído um predecessor $P(t)$, que consiste na identificação da outra amostra com quem está conectada pelo caminho ótimo. Isso resulta em uma floresta de caminhos entre as amostras de treinamento de mesmo rótulo, contendo um mapa de custos C de todas as amostras de treinamento, um mapa de classes ou rótulos L e uma lista ordenada, Z' de amostras do conjunto de treinamento em função do seu custo.

A fase de treino do OPF é apresentada no Algoritmo 3.1.

3.4.1.2 Classificação com o OPF

Durante a classificação, cada amostra t do conjunto de teste Z_{TS} é classificada da seguinte maneira: uma amostra $t \in Z_{TS}$, de classe $L(t)$ desconhecida, é inserida como um nó no grafo e são calculados os custos de ligação com cada amostra s do conjunto de treinamento Z_{TR}. A amostra de teste t herdará a classe da amostra de treino s com a qual obtiver o menor custo de ligação, $c(t, s)$, classificando a primeira como pertencente àquela classe, fazendo $L(t) = L(s)$.

Algoritmo 3.1: Etapa de treino do OPF. Adaptado de Papa, Falcão e Suzuki (2009).

Entrada: Conjunto de treinamento Z_{TR}, conjunto de protótipos rotulados S.

Saída: Floresta de Caminhos Ótimos P contendo Mapa de custos C, Mapa de Rótulos L, lista ordenada em função de custo Z'.

Auxiliares: Fila de Prioridade Q, e variável de custo cst

```
início
    para s ∈ Z_TR faça
        C(s) ← +∞
    fim
    para s ∈ S faça
        C(s) ← 0
        P(s) ← nil
        L(s) ← λ_s
        insira s em Q
    fim
    enquanto Q ≠ ∅ faça
        Remover de Q a amostra s com menor custo
        para cada amostra t ≠ s, tal que, C(t) < C(s) faça
            cst ← max{C(s), d(s,t)}
            se cst < C(t) então
                se C(t) ≠ +∞ então
                    Remova t de Q
                fim
                P(t) ← s
                L(t) ← L(s)
                C(t) ← cst
                Insira Q na fila
            fim
        fim
    fim
fim
```

O custo de ligação é calculado em função do custo da amostra de treino, $C(s)$, e pela distância entre as duas, $d(s,t)$ pela equação 3.4.

$$c(t,s) = min\{max\{C(s), d(s,t)\}\}. \tag{3.4}$$

Cada amostra é processada independentemente das outras, sendo comparada somente com as amostras de treino. A fase de teste do OPF é mostrada no Algoritmo 3.2

A condição de interrupção de laço mostrada no algoritmo é responsável por um ganho no custo computacional, pois uma vez que o custo é comparado com amostras de treino em uma lista ordenada, uma vez que o custo atual da amostra de teste se torna menor que o custo de uma amostra listada em ordem crescente de custo, não há a necessidade de computar os custos do restante das amostras da lista.

Algoritmo 3.2: Etapa de classificação do OPF. Adaptado de Papa, Falcão e Suzuki (2009).

Entrada: Conjunto de amostras em teste Z_{TS}, Floresta de caminhos ótimos P do conjunto de treinamento contendo conjuntos de rótulos (L), Custos (C), lista de amostras de treinamento (Z'), ordenada em função do custo.

Saída: Rótulos λ atribuídos as amostras de Z_{TR}

Auxiliares: Variável de custo cst

início
  para *cada amostra* $t \in Z_{TS}$ **faça**
    $cst \leftarrow +\infty$
    para *cada amostra* $s \in Z'$ *em ordem de custo* **faça**
      se $cst > max\{C(s), d(s,t)\}$ **então**
        $cst \leftarrow max\{C(s), d(s,t)\}$
        $\lambda(t) \leftarrow L(s)$
      fim
      se $C(s) > cst$ **então**
        interromper laço
      fim
    fim
  fim
fim

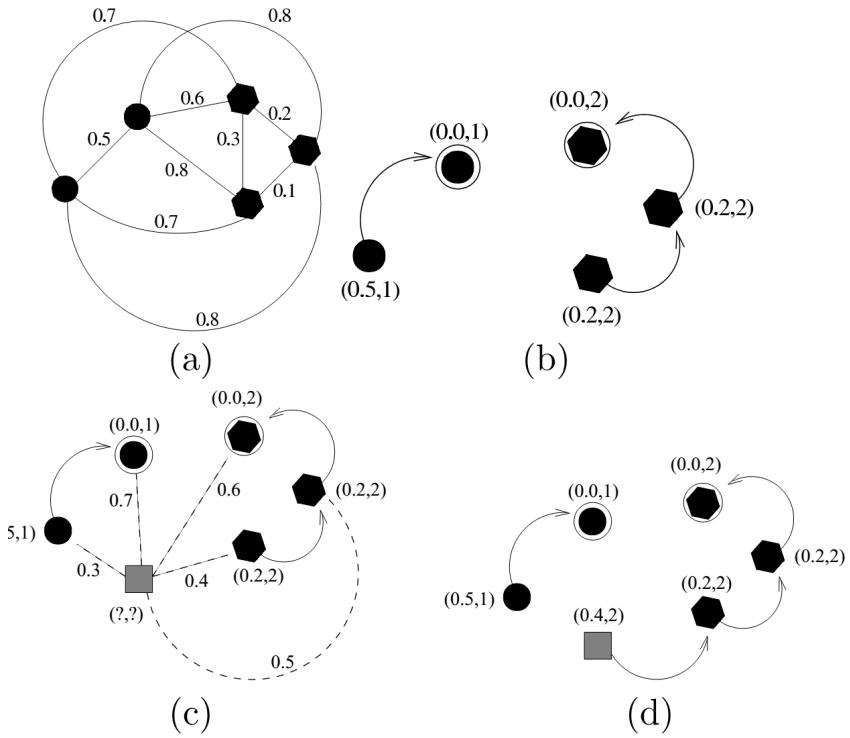
Na Figura 11 são ilustradas as etapas de funcionamento do OPF. Na Figura 11-a, é ilustrado o início da fase de treinamento onde são calculadas as distâncias entre todos os nós do grafo, segundo uma função de distância $d(s,t)$. Na Figura 11-b são ilustrados dois protótipos, escolhidos através do algoritmo da árvore de espalhamento mínimo (MST). As amostras se ligam entre si, em função da distância ou do custo dos predecessores, o que for maior. Uma amostra de teste, ilustrada na Figura 11-c na cor cinza , é inserida a distância calculada com as amostras de treino. A Figura 11-d, mostra a ligação da amostra de teste à amostra que apresentou menor custo, segundo a equação 3.4.

3.4.1.3 Métricas de distância

Além da norma euclideana, o OPF pode ser utilizado com outras distâncias já implementadas no código. Para trabalhar com distâncias, o OPF, antes da etapa de treinamento, já computa as distâncias entre todas as amostras, tanto de treino como de teste. As distâncias são pré-computadas e armazenadas em uma tabela de distâncias, em função do número total de amostras. As distâncias, consideradas na LibOPF (PAPA; FALCÃO; SUZUKI, 2009) e utilizadas neste trabalho foram: Euclidean, Chi-Squared, Manhattan, Canberra, Square Chi-Square e Bray-Curtis.

A vantagem de pré-computar as distâncias, é a redução da quantidade de cálculos realizados durante as etapas de treino e teste, melhorando o custo computacional. Por outro lado, como

Figura 11 – Etapas do funcionamento do OPF: a) Cálculo das distâncias entre amostras de treino; b) escolha dos protótipos e cálculo dos caminhos ótimos; c) insersão de uma amostra de teste e cálculo do custo de ligação; d) ligação da amostra de teste com a amostra de menor custo.



Fonte: Papa, Falcão e Suzuki (2009).

o armazenamento é realizado em formato de tabelas, bases de dados com muitas amostras, requerem grande capacidade de armazenamento, processamento, bem como memória de acesso randômico (RAM) durante a computação e armazenamento das distâncias.

Como a base MIT-BIH é composta por aproximadamente 100 mil amostras, seria necessário armazenar cerca de 10 bilhões de valores de distância. Assim, é necessário modificar código da LibOPF para computar as distâncias somente quando necessário, durante as etapas de treino e teste, sem armazená-la em forma de tabela.

3.4.2 Máquinas de Vetores de Suporte

Dentro da teoria do aprendizado de máquinas, uma problemática que pode ser levantada é a classificação de uma amostra entre duas classes. Assim, o reconhecimento de padrões entre as duas classes pode ser definido por uma função definida a partir do conjunto de dados de treinamentos na forma (SCHÖLKOPF; SMOLA, 2002),

$$f : \chi \rightarrow \{1\}, \tag{3.5}$$

em que um conjunto l de observações contendo um par de vetores de amostra $x_i \in \mathbb{R}^n$, e a verdade "associada" $y = \pm 1$, que é dada por uma fonte confiável, indica se x_i pertence ou não

a uma classe.

Baseado no princípio de minimização de risco estrutural (VAPNIK, 1999), o processo de otimização das Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM), objetiva estabelecer uma função de separação equilibrando as considerações entre a generalização e o *over-fitting* durante uma classificação.

Considerando a classe de hiperplanos como um espaço $\mathcal{H}$ de produto escalar,

$$\langle \mathbf{w}, \mathbf{x} \rangle - b = 0, \tag{3.6}$$

correspondendo à função de decisão,

$$f(x) = y_i(\langle \mathbf{w}, \mathbf{x} \rangle - b), \tag{3.7}$$

é proposto um algoritmo de aprendizado em (VAPNIK, 1999), para problemas separáveis por hiperplanos, baseado nos seguintes argumentos:

- entre todos os hiperplanos separando os dados, existe um único hiperplano ótimo, caracterizado pela maior margem de separação entre quaisquer pontos no hiperplano;
- o aumento da margem diminui os problemas de *over-fitting*.

Assim, para construir um hiperplano ótimo, é necessário resolver o seguinte problema de minimização:

$$\min \tau(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2, \tag{3.8}$$

sujeito a,

$$y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle - b) \geq 1 \forall i = 1, ..., m, \tag{3.9}$$

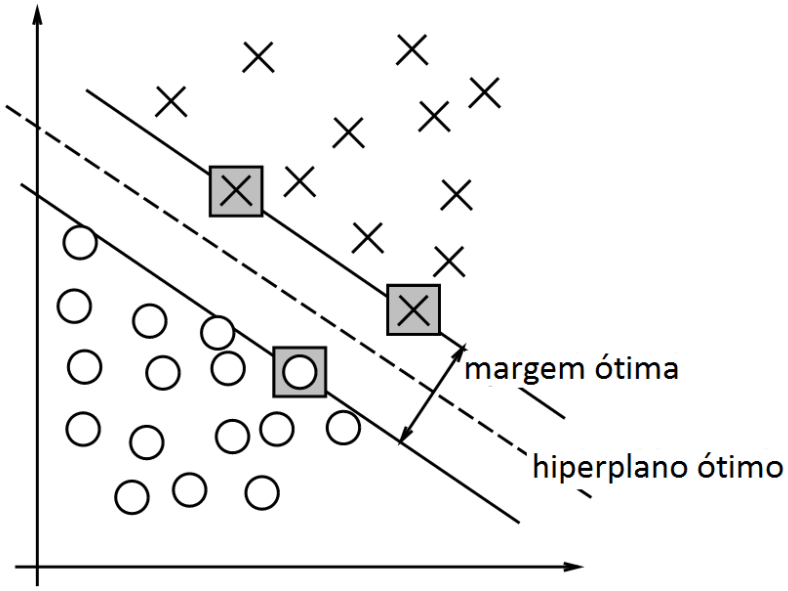
que garante que $f(x_i)$ será +1 somente para $y_i = +1$, e -1 para $y_i = -1$, fixando uma escala para w. A função τ, na equação 3.8 , chama-se função objetiva. A equação 3.9 representa os limites impostos ao conjunto de treinamento.

Assim, a função de separação de classes pode ser definida por uma combinação ponderada de alguns vetores do conjunto de treinamento, que, são chamados de vetores de suporte e caracterizam a superfície de decisão ótima com a máxima margem entre duas classes. Uma ilustração mostrando uma separação e os vetores de suporte utilizados é mostrada na Figura 12, em que a superfície de decisão corresponde à linha tracejada entre as duas classes.

A margem da superfície corresponde às linhas contínuas, e os vetores de suporte são identificados pelos quadrados ao redor das amostras de treinamento no limite da margem.

Para tentar solucionar problemas de não linearidade na superfície de decisão, pode-se substituir o produto interno de pesos e vetores por uma função de núcleo (CORTES; VAPNIK, 1995), que é utilizada para estender o conceito de classificadores baseados em hiperplano para máquinas de vetores de suporte não lineares.

Figura 12 – Representação dos vetores de suporte do SVM



Fonte: Adaptado de (CORTES; VAPNIK, 1995)

Entre as várias funções de núcleo (ou *kernel*), a função denominada *Radian Basis Function* (RBF) é uma das mais utilizadas para resolver os problemas relacionados à não linearidade e é dada pela equação 3.10.

$$K(x_1, x_2) = e^{-||x_1||^2/2\sigma^2}, \tag{3.10}$$

em que σ^2 denota a variância da largura de banda no núcleo e é definida pelo parâmetro γ que representa $1/\sigma$.

Mesmo determinando o melhor hiperplano, uma perfeita separação de classes pode não existir, assim, um erro de treinamento pode ser introduzido. Para manter o balanço entre generalização e *over-fitting*, pode-se introduzir, na função objetiva um termo que envolve a soma dos erros de treinamento multiplicados por uma constante C, compreendendo, juntamente com o valor de γ, importantes parâmetros a serem estabelecidos pelo usuário (CORTES; VAPNIK, 1995).

No caso de problemas envolvendo mais de uma classe, uma das estratégias que pode ser utilizada é o método One-Against-One. Esse foi o método adotado nos experimentos realizados neste trabalho e consiste na comparação entre cada classe, sendo o problema de classificação dividido em vários outros problemas de classificação binários.

Os valores dos parâmetros C e γ, utilizados nas classificações, são obtidos na fase de treino variando-se ambos buscando maior taxa de acerto. O conjunto de treino é dividido em duas partes, contendo aproximadamente o mesmo número de amostras de cada classe. C e γ são definidos como logaritmos de n na base dois, conforme equação 3.11,

$$\log_2 n, \tag{3.11}$$

variando-se n de -5 a 15 para C e -15 a 3 para γ. Os valores com maior taxa de acerto são considerados para treino. Foi utilizada a biblioteca LibSVM, versão 3.18 (CHANG; LIN, 2011).

3.4.3 Classificador Bayesiano

O teorema de Bayes (JAYNES, 2003), define que a probabilidade $p(\omega_i|x_i)$ de uma amostra x_i, pertencer uma classe ω_i, é dada por:

$$p(\omega_i, x_i) = \frac{p(x_i|\omega_i)P(\omega_i)}{p(x_i).} \tag{3.12}$$

Assim, a função de decisão de um classificador Bayesiano (Bayesian Classifier - BC), definida pelo discriminante para a classe ω_i, é dada por:

$$d_i(x) = p(x|\omega_i)P(\omega_i). \tag{3.13}$$

Essa função de decisão mostra-se dependente do conhecimento à priori de $p(x|\omega_i)$ e $P(\omega_i)$, obtido pelo cálculo de histogramas de cada classe, para obter-se as respectivas funções, de densidade de probabilidade e de probabilidade acumulada.

Uma alternativa para facilitar a aquisição dessa função de densidade de probabilidade é assumir a distribuição como sendo Gaussiana e dependente apenas do conhecimento da matriz de covariância, C_i, e do vetor médio, ou centroide μ_i, de cada classe i através da equação:

$$p(x|\omega_i) = \frac{1}{2\pi^{n/2}|C_i|^{1/2}} e^{[-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)]}. \tag{3.14}$$

Desta forma, o problema de classificação pode ser resolvido, somente com os dados de μ e C, sem a necessidade da definição de parâmetros, o que torna o classificador Bayesiano, rápido e de simples utilização.

3.5 Avaliação estatística

A análise dos resultados foi realizada considerando dois aspectos: desempenho e custo computacional.

A análise de desempenho corresponde à capacidade do algoritmo em realizar corretamente a tarefa de classificação. Um importante parâmetro, considerado neste trabalho, para avaliação de desempenho, é a taxa de acerto (Ac), de cada classificador utilizando cada extrator de atributos, definido pela equação 3.15.

$$Ac = \frac{\text{\# Amostras classificadas corretamente}}{\text{\# Total de amostras.}} \tag{3.15}$$

Em conjuntos balanceados, nos quais todas as classes têm a mesma representatividade, essa medida é fundamental, porém, em conjuntos desbalanceados, como o conjunto objeto desse estudo, parâmetros adicionais são de suma importância. Como no conjunto de dados, MIT-BIH, há uma classe predominante N, com cerca de 90% das amostras, a taxa de acerto é muito

influenciada por esta classe. As outras classes, apesar de menor peso na taxa de acerto são fundamentais por serem clinicamente relevantes, correspondendo às classes patológicas (de maior interesse na detecção), ou seja, contendo os batimentos cardíacos que possam indicar risco ao paciente.

Além da taxa de acerto, outros métodos estatísticos de avaliação devem ser utilizados para avaliação da generalização dos algoritmos de classificação. São adotados neste trabalho sensibilidade (Se) e especificidade (Sp) para esse tipo de avaliação. Através dessas medidas é possível avaliar como é o comportamento da classificação para cada tipo de arritmia, individualmente.

A sensibilidade (Se) de uma classe *i* é dada por,

$$Se_i = \frac{VP_i}{VP_i + FN_i}, \tag{3.16}$$

em que, VP_i (Verdadeiros Positivos) representa os batimentos da classe i, corretamente classificados na classe i, e FN_i (Falsos Negativos) são os batimentos da classe i classificados como outro tipo de batimento, ou seja, em outra classe. A sensibilidade se comporta como uma taxa de acerto específica da classe, ou seja, qual a taxa de amostras corretamente classificadas entre todas pertencentes a uma classe.

Já a especificidade (Sp) de uma classe *i*, é dada por,

$$Sp_i = \frac{VN_i}{VN_i + FP_i}, \tag{3.17}$$

em que, VN_i (Verdadeiros Negativos) são as amostras de outras classes, diferentes de i, que não foram classificadas na classe i, e FP (Falsos Positivos) são as amostras de outras classes, classificadas na classe i. Desta forma, a especificidade avalia a taxa de acerto na classificação de amostras negativas, ou seja, não pertencente à classe i.

Para sintetizar estes dois parâmetros, é calculado o parâmetro H (equação 3.18, que consiste na média harmônica entre sensibilidade e especificidade para cada classe.

$$H = 2 \times \frac{Se \times Sp}{Se + Sp}. \tag{3.18}$$

Finalmente, para calcular o desempenho, considerando todas as classes, foi calculada a média aritmética de H (M_H), considerando todas as classes, obtendo-se um parâmetro com pouca influência do desbalanceamento das amostras em relação à classe N.

$$M_H = \frac{\sum_{i=1}^{n} H_i}{n}, \tag{3.19}$$

na qual, n é o número de classes e H_i a média harmônica entre sensibilidade e especificidade para a classe i. Um classificador é menos generalista para uma classe i se o valor de H se aproxima de um para esta classe.

Para avaliar o custo computacional de cada classificador, são computados os tempos de treino e teste utilizando um computador com processador Intel i7 de terceira geração com frequência de *clock* de 3 GHz e 16 Gb de memória RAM. É utilizado o software Matlab 2010

com sistema operacional Windows 7 para a extração de atributos. Para computar os algoritmos de classificação, as bibliotecas são compiladas em sistema Linux Ubuntu 12.04 no mesmo computador.

4 RESULTADOS

Nesse capítulo, são apresentados os resultados referentes ao custo e computacional e desempenho de cada método de extração e técnica de classificação. Inicialmente, o classificador OPF é avaliado considerando 6 métricas de distâncias: *Euclidean*, *Chi-Square*, *Manhattan*, *Squared Chi-Squared* (SCS), e *Bray-Curtis*. Em seguida, é realizada uma comparação das melhores métricas do OPF com as outras técnicas de classificação *Support Vector Machines*, utilizando função de núcleo *Radian Basis Function* (SVM-RBF) e o Classificador Bayesiano (BC). Na última seção resultados são discutidos.

Para efeito de simplificação os métodos de extração são identificados de acordo com a Tabela 3, apresentada na seção 3.3 do capítulo 3.

4.1 Análise da eficiência e eficácia do OPF

Nesta seção é realizada uma avaliação do desempenho e do custo computacional do OPF utilizando 6 métricas de distâncias, já implementadas na biblioteca LibOPF(PAPA; FALCÃO; SUZUKI, 2009). A avaliação é feita considerando a classificação em 5 classes (ANSI/AAMI, 2008) e em 3 classes (LLAMEDO; MARTÍNEZ, 2011).

4.1.1 Avaliação do OPF considerando 5 classes

O desempenho do OPF é avaliado através das taxas de acerto para cada distância classificando cada método de extração. Os resultados de taxa de acerto são mostrados na Tabela 5.

A maior acurácia, considerando 5 classes (ANSI/AAMI, 2008), é obtida utilizando a distância *Manhattan* (91,21%), destacado em negrito na tabela. Este valor é aproximadamente 0,35% maior que o segundo melhor resultado, obtido com a métrica de distância *Canberra* (90,88%) e 0,5% maior que o resultado obtido com a métrica *Squared Chi-Squared* (90,75%), todos classificando o conjunto D. Os resultados utilizando o conjunto D foram os melhores para todas as distâncias utilizadas, mostrando que o método de Yu e Chen (2007) é um bom extrator de atributos para uso com o OPF.

Conforme explicado no capítulo anterior, os conjuntos de sinais de batimentos de ECG da base MIT-BIH são desbalanceados. A Figura 13, mostra a análise de componentes principais que ilustra a distribuição das amostras, por classe, de todas as amostras da base de dados.

Além das taxas de acerto, foram realizados cálculos de sensibilidade e especificidade, bem como a média harmônica desses dois parâmetros. Os parâmetros de Se, Sp e H são mostrados na Tabela 6. Em negrito são destacadas os maiores valores de H para cada classe.

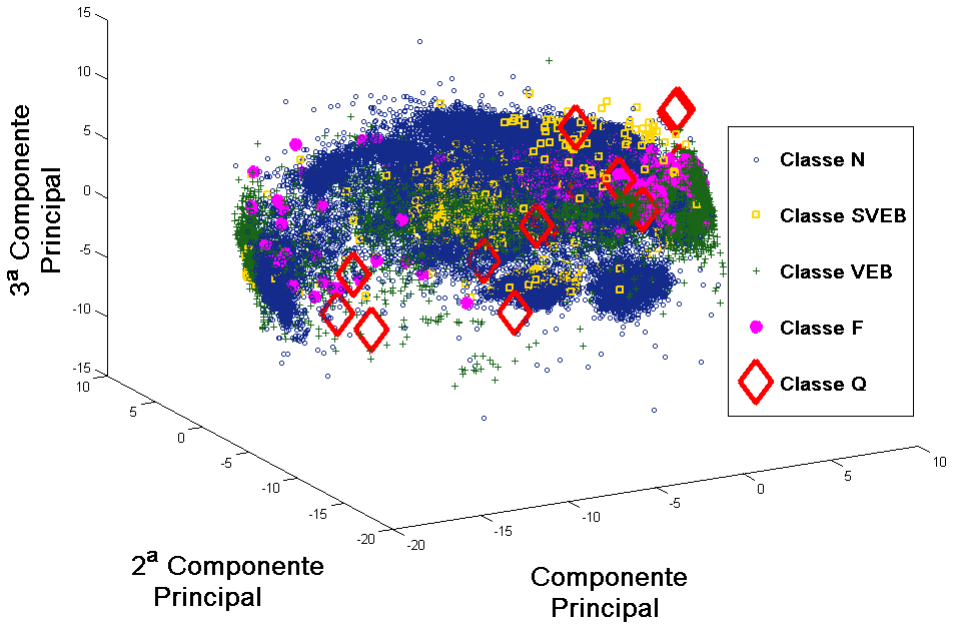
Os melhores valores de H para a classe N são obtidos utilizando as distâncias *Canberra* (0,78) e *Squared Chi-Squared* (0,78), ambos utilizando o extrator C. A combinação da distância

Tabela 5 – Taxas de acerto do OPF considerando 5 classes.

	Métrica de Distância		
Conjunto	Euclidean [%]	Chi-Square [%]	Manhattan [%]
A	80,68	83,26	77,57
B	79,63	88,80	79,43
C	81,25	87,60	84,46
D	90,70	89,12	**91,21**
E	86,54	89,05	86,47
F	89,12	85,28	90,39
	Métrica de Distância		
Conjunto	Canberra [%]	Squared Chi-Squared [%]	Bray-Curtis [%]
A	77,93	76,14	79,81
B	80,51	80,61	87,69
C	84,90	82,63	76,55
D	90,88	90,75	88,90
E	86,53	86,62	81,79
F	86,60	85,70	78,41

Fonte: Autor

Figura 13 – Distribuição das amostras da base de dados, nas 5 classes, utilizando PCA, considerando as 3 componentes principais.



Squared Chi-Squared e o extrator C, também resultaram no maior valor de H para a classe $SVEB$ (0,60). Já para as classes VEB e F, a distância que apresenta os melhores resultados é a *Euclidean*, classificando os conjuntos F (0,91) e A (0,55), respectivamente.

Em relação à classe Q, o OPF não classifica nenhuma amostra corretamente devido a dois fatores principais: a distribuição, por não ser concentrada, e a baixa representatividade das amostras tanto no conjunto de treinamento como no conjunto de teste, com aproximadamente

Tabela 6 – H, sensibilidade, especificidade para as distâncias do OPF considerando 5 classes. Valores multiplicados por 100.

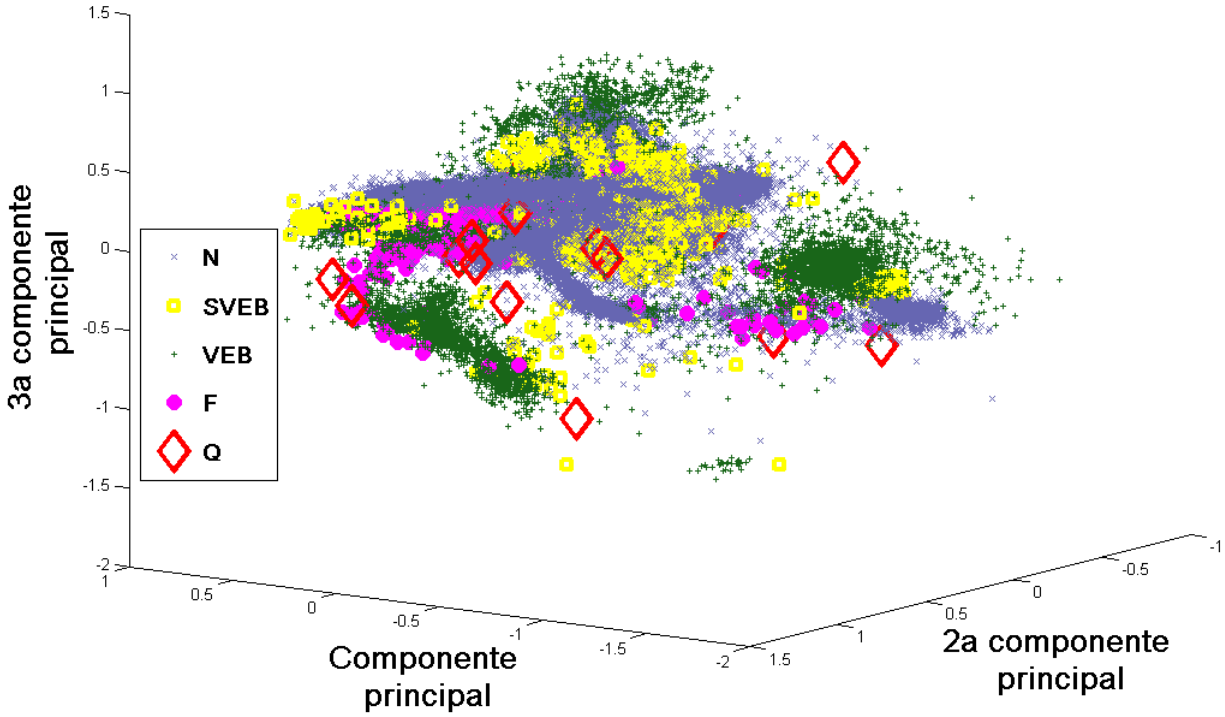
		Classe do batimento				
		N	SVEB	VEB	F	Q
Métrica	Conjunto	H | Se | Sp	H | Se | Sp	H | Se | Sp	H | Se | Sp	H | Se | Sp
Euclidean	A	066 | 085 | 054	002 | 001 | 097	084 | 078 | 091	**055** | 038 | 097	000 | 000 | 100
	B	050 | 086 | 035	005 | 002 | 097	056 | 041 | 090	001 | 001 | 098	000 | 000 | 100
	C	074 | 085 | 066	031 | 018 | 095	084 | 078 | 091	014 | 007 | 097	000 | 000 | 100
	D	073 | 096 | 059	030 | 018 | 099	084 | 075 | 097	007 | 004 | 099	000 | 000 | 100
	E	065 | 092 | 050	006 | 003 | 098	076 | 062 | 097	029 | 017 | 097	000 | 000 | 100
	F	074 | 093 | 062	022 | 012 | 099	**091** | 086 | 097	031 | 018 | 097	000 | 000 | 100
Chi-Square	A	017 | 093 | 009	003 | 001 | 099	015 | 008 | 095	005 | 002 | 099	000 | 000 | 100
	B	002 | 100 | 001	000 | 000 | 100	000 | 000 | 100	001 | 001 | 100	000 | 000 | 100
	C	015 | 098 | 008	002 | 001 | 100	017 | 009 | 098	000 | 000 | 100	000 | 000 | 100
	D	004 | 100 | 002	000 | 000 | 100	004 | 002 | 100	000 | 000 | 100	000 | 000 | 100
	E	004 | 100 | 002	000 | 000 | 100	004 | 002 | 100	000 | 000 | 100	000 | 000 | 100
	F	012 | 095 | 006	002 | 001 | 099	011 | 006 | 097	001 | 001 | 100	000 | 000 | 100
Manhattan	A	066 | 081 | 056	003 | 002 | 095	086 | 082 | 090	051 | 035 | 096	000 | 000 | 100
	B	046 | 087 | 031	006 | 003 | 097	051 | 035 | 090	001 | 000 | 099	000 | 000 | 100
	C	076 | 088 | 066	031 | 018 | 097	085 | 078 | 093	013 | 007 | 098	000 | 000 | 100
	D	071 | 096 | 056	023 | 013 | 099	085 | 075 | 098	008 | 004 | 099	000 | 000 | 100
	E	064 | 092 | 048	006 | 003 | 098	075 | 061 | 097	023 | 013 | 097	000 | 000 | 100
	F	072 | 095 | 058	010 | 006 | 099	090 | 083 | 098	034 | 021 | 098	000 | 000 | 100
Canberra	A	071 | 081 | 063	047 | 031 | 098	080 | 073 | 087	025 | 015 | 096	000 | 000 | 100
	B	042 | 088 | 028	005 | 003 | 098	045 | 029 | 091	002 | 001 | 099	000 | 000 | 100
	C	**078** | 088 | 071	056 | 039 | 096	084 | 077 | 093	016 | 009 | 099	000 | 000 | 100
	D	071 | 096 | 057	026 | 015 | 099	085 | 075 | 098	011 | 006 | 099	000 | 000 | 100
	E	064 | 092 | 049	007 | 004 | 098	076 | 062 | 097	024 | 014 | 097	000 | 000 | 100
	F	064 | 092 | 049	007 | 004 | 099	083 | 071 | 098	009 | 005 | 095	000 | 000 | 100
Squared Chi-Squared	A	065 | 080 | 056	006 | 003 | 097	081 | 076 | 086	037 | 023 | 097	000 | 000 | 100
	B	048 | 088 | 033	004 | 002 | 098	053 | 037 | 091	001 | 001 | 098	000 | 000 | 100
	C	**078** | 085 | 072	**060** | 044 | 095	085 | 078 | 093	022 | 012 | 097	000 | 000 | 100
	D	073 | 096 | 060	032 | 019 | 099	085 | 075 | 097	009 | 005 | 099	000 | 000 | 100
	E	065 | 092 | 050	006 | 003 | 098	076 | 062 | 097	027 | 016 | 097	000 | 000 | 100
	F	069 | 090 | 056	020 | 011 | 099	087 | 079 | 098	008 | 004 | 093	000 | 000 | 100
Bray-curtis	A	051 | 086 | 036	018 | 010 | 098	057 | 041 | 090	016 | 009 | 098	000 | 000 | 100
	B	005 | 098 | 002	000 | 000 | 100	000 | 000 | 098	000 | 000 | 100	000 | 000 | 100
	C	055 | 083 | 041	008 | 004 | 096	057 | 043 | 088	001 | 000 | 099	000 | 000 | 100
	D	002 | 100 | 001	001 | 000 | 100	000 | 000 | 100	000 | 000 | 100	000 | 000 | 100
	E	036 | 090 | 022	005 | 003 | 098	044 | 028 | 096	006 | 003 | 096	000 | 000 | 100
	F	053 | 085 | 039	003 | 002 | 096	054 | 038 | 091	021 | 012 | 098	000 | 000 | 100

Fonte: Autor

0,00015% das amostras em cada conjunto. A distribuição das amostras do conjunto C é apresentada na Figura 14, na qual é utilizada a análise de componentes principais (PCA) para ilustrar a distribuição das amostras considerando as 3 componentes principais, com destaque para a classe Q em vermelho.

Ainda que não apresente baixa taxa de acerto, os resultados obtidos pela métrica de distância

Figura 14 – Distribuição das amostras do conjunto de teste C, utilizando PCA, considerando as 3 componentes principais.



Chi-Square, classificando o conjunto B (88,80%), não se traduz em um desempenho satisfatório para a separação de classes, uma vez que apresenta baixos valores de sensibilidade e especificidade para todas as classes, com exceção dos valores de sensibilidade para a classe N. Isso se deve à classificação da maioria das amostras das classes $SVEB$, VEB, F e Q na classe N, levando a um baixo valor de especificidade (2%). Esta especificidade faz com que o valor de H para a classe N aproxime-se de zero. Para demonstrar esses resultados, uma matriz de confusão da classificação do conjunto B utilizando a métrica *Chi-Square* é mostrada na Tabela 7.

Tabela 7 – Matriz de confusão da classificação do conjunto B utilizando distância Chi-Square.

		Classe verdadeira				
		N	SVEB	VEB	F	Q
Classificado como	N	44115	1834	3209	350	7
	SVEB	27	0	2	4	0
	VEB	76	0	7	32	0
	F	19	2	3	2	0
	Q	1	0	0	0	0

Fonte: Autor

Para a análise das melhores condições de classificação em termos de extração de atributos

e distâncias do OPF, é computada a média aritmética, M_H, considerando todas as classes. O cálculo de M_H é mostrado na equação 3.19, na seção 3.5 do Capítulo 3.

Na Tabela 8 são apresentados os resultados de M_H e em negrito é destacado o maior valor.

Tabela 8 – M_H das distâncias do OPF, por conjunto, considerando 5 classes.

	Métrica de Distância					
Conjunto	Euclidean	Chi-Square	Manhattan	Canberra	Squared Chi-Squared	Bray-Curtis
A	0,41	0,08	0,41	0,45	0,38	0,28
B	0,22	0,01	0,20	0,19	0,21	0,01
C	0,41	0,07	0,41	0,47	**0,49**	0,24
D	0,39	0,02	0,37	0,39	0,40	0,01
E	0,35	0,02	0,33	0,34	0,35	0,18
F	0,43	0,05	0,41	0,33	0,37	0,26

Fonte: Autor

Apesar de apresentar a maior taxa de acerto (91,21%) com a distância *Manhattan* classificando o conjunto extraído com D, este não obteve o maior resultado de M_H (0,37), que foi obtido utilizando a distância *Squared Chi-Squared* classificando o conjunto C (0,49). Duas matrizes de confusão,contendo essas classificações são mostradas na Tabela 9.

Tabela 9 – Matrizes de confusão das classificações utilizando distâncias *Manhattan* e *Squared Chi-Squared*

Manhattan (YU; CHEN, 2007)

		Classe verdadeira				
		N	SVEB	VEB	F	Q
Classificado como	N	42656	1463	624	307	5
	SVEB	392	239	151	1	0
	VEB	888	122	2408	63	2
	F	301	12	38	17	0
	Q	1	0	0	0	0

Squared Chi-Squared - (SONG et al., 2005)

		Classe verdadeira				
		N	SVEB	VEB	F	Q
Classificado como	N	37677.00	612	591	322	2
	SVEB	2495	802	33	2	0
	VEB	2798	412	2514	17	5
	F	1245	9	81	47	0
	Q	3	1	0	0	0

Fonte: Autor

Nota-se que no conjunto classificado com a distância *Manhattan* ocorre maior acerto na classe N e uma confusão nas classes $SVEB$, VEB e F com a classe N. Na segunda matriz de confusão, utilizando a distância *Squared Chi-Squared*, há um menor acerto na classe N, porém menor confusão das outras classes, com exceção da classe VEB com a $SVEB$, mostrando menor generalização. Nenhuma das duas classificações obteve sucesso em classificar a classe Q. Isso se deve à sua baixa representatividade e espalhamento no espaço de características.

O valor de M_H para o conjunto B, classificada utilizando a distância *Chi-Square*, da ordem de 0,01, corrobora com o apresentado na matriz de confusão da Tabela 7, apresentando valor de 0,01.

Os resultados de eficiência do OPF, avaliados em termos de custo computacional, são apresentados na Tabela 10, considerando as fases de treino e teste. Também é considerado o custo computacional de todo o processo.

Tabela 10 – Custo computacional do OPF considerando 5 clsses. Tempo em segundos.

	Métrica de Distância		
	Euclidean	Chi-Square	Manhattan
Conjunto	treino ǀ teste ǀ total	treino ǀ teste ǀ total	treino ǀ teste ǀ total
A	443,24 ǀ 685,25 ǀ 1128,50	3230,71 ǀ 1987,74 ǀ 5218,45	336,86 ǀ 585,06 ǀ 921,92
B	145,75 ǀ 195,85 ǀ 341,60	416,94 ǀ 6,73 ǀ 423,68	54,20 ǀ 105,45 ǀ 159,66
C	142,38 ǀ 200,98 ǀ 343,36	465,81 ǀ 106,89 ǀ 572,70	55,30 ǀ 99,68 ǀ 154,99
D	131,50 ǀ 131,58 ǀ 263,08	300,20 ǀ 2,79 ǀ 302,99	40,10 ǀ 60,35 ǀ 100,45
E	171,92 ǀ 198,41 ǀ 370,33	652,29 ǀ 14,60 ǀ 666,89	80,39 ǀ 118,80 ǀ 199,19
F	315,92 ǀ 450,70 ǀ 766,62	2109,70 ǀ 821,64 ǀ 2931,34	218,32 ǀ 380,72 ǀ 599,04
	Métrica de Distância		
	Canberra	Squared Chi-Squared	Bray-Curtis
Conjunto	treino ǀ teste ǀ total	treino ǀ teste ǀ total	treino ǀ teste ǀ total
A	1478,39 ǀ 1594,39 ǀ 3072,78	1474,88 ǀ 1593,94 ǀ 3068,82	1139,51 ǀ 1175,38 ǀ 2314,89
B	188,69 ǀ 168,36 ǀ 357,06	189,46 ǀ 181,00 ǀ 370,46	43,45 ǀ **0,04** ǀ 43,48
C	204,15 ǀ 251,44 ǀ 455,59	204,61 ǀ 233,42 ǀ 438,04	105,46 ǀ 126,20 ǀ 231,66
D	131,79 ǀ 128,25 ǀ 260,04	131,54 ǀ 126,86 ǀ 258,40	**35,26** ǀ 5,97 ǀ **41,23**
E	298,19 ǀ 204,20 ǀ 502,39	298,85 ǀ 226,00 ǀ 524,85	151,19 ǀ 139,22 ǀ 290,41
F	970,43 ǀ 996,50 ǀ 1966,93	967,41 ǀ 987,13 ǀ 1954,55	428,99 ǀ 533,89 ǀ 962,88

Fonte: Autor

No que se refere ao treino, o melhor tempo foi obtido utilizando a métrica *Bray-Curtis* (35,26 segundos), seguida da métrica *Manhattan* (40,10 segundos), ambas classificando o conjunto D. O terceiro melhor tempo foi obtido pela métrica *Bray-Curtis* treinando com o conjunto B.

O melhor tempo de teste também foi obtido com a distância *Bray-curtis*, classificando o conjuntos B em 0,04 segundos, quase 70 vezes mais rápido que o segundo melhor tempo, obtido classificando o conjunto D, utilizando a métrica de distância *Chi-Square* em 2,79 segundos.

Considerando o tempo total das duas etapas, treino e teste, os melhores resultados foram obtidos através da métrica *Bray-Curtis* classificando D e B obtendo tempos de 41,22 e 43,21 segundos, respectivamente. O terceiro melhor tempo, foi obtido pela métrica de distância *Manhattan*, também classificando o conjunto D em 100,45 segundos, mais de duas vezes o tempo levado utilizando a distância *Bray-Curtis*.

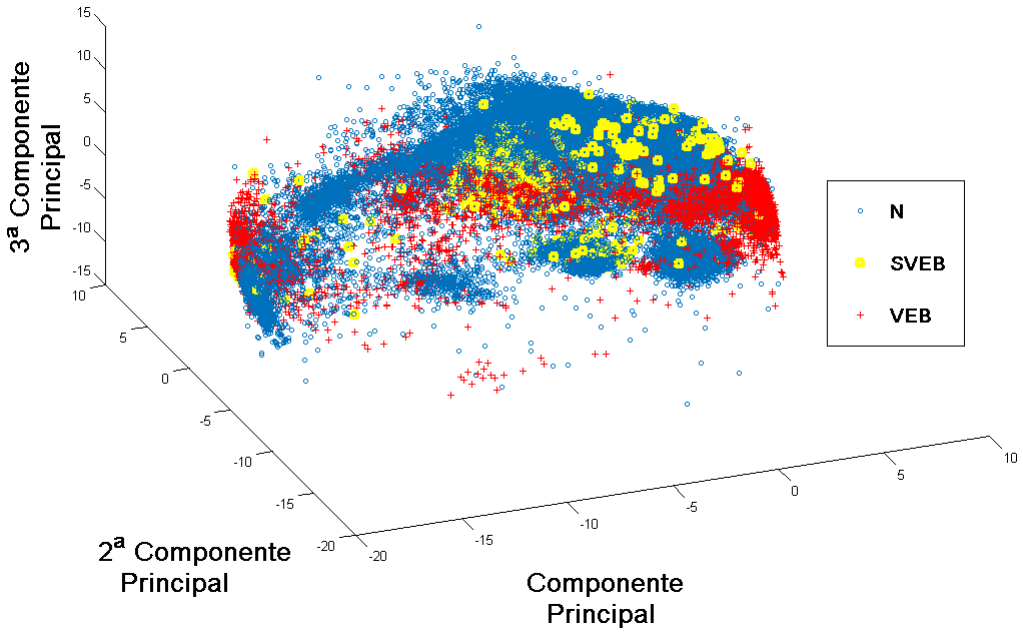
Apesar de menor custo computacional, o uso das métricas de distância *Bray-curtis* e *Chi-Square*, classificando os conjuntos B e D, resultaram em valores de M_H, abaixo de 0,2. Através dos valores de sensibilidade e especificidade, nota-se uma generalização das amostras em relação à classe N, explicando o baixo custo.

4.1.2 Avaliação do OPF considerando 3 classes

Uma avaliação do OPF também é realizada considerando 3 classes conforme (LLAMEDO; MARTÍNEZ, 2011), em que as classes F e Q são agregadas na da classe VEB.

Na Figura 15 é ilustrada a distribuição das amostras após a redução para 4 classes, através de PCA sobre os sinais brutos de ECG.

Figura 15 – Distribuição das amostras da base de dados, nas 3 classes, utilizando PCA, considerando as 3 componentes principais.



Os resultados de acurácia das classificações considerando 3 classes são mostrados na Tabela 11

Assim como os resultados obtidos, considerando 5 classes, as três maiores taxas de acerto também foram obtidas utilizando as distâncias *Manhattan* (91,42%), *Canberra* (91,07%) e *Squared Chi-Squared* (90,94%). Todos esses resultados são obtidos classificando o conjunto de atributos D, que possui as maiores taxas de acerto para todas as distâncias.

O conjunto de sinais divididos em 3 classes, também não é um conjunto balanceado. Com a agregação das classes F e Q à classe VEB, os efeitos do desbalanceamento tendem a ser amenizados para os classificadores, porém a classe N ainda concentra aproximadamente 90% das amostras. Nas avaliações considerando 3 classes, também são analisados os parâmetros de sensibilidade, especificidade e média harmônica, H, apresentados na Tabela 12.

Para classe N, as distâncias *Canberra* e *Squared Chi-Squared* apresentam o melhor valor para H (0,78), classificando o conjunto C e para classe $SVEB$ a distância *Squared Chi-Squared* apresenta o melhor resultado de H (0,60). Isto indica que a agregação em 3 classes não influencia no parâmetro H para as classes N e $SVEB$. Em relação à classe VEB, incluindo as classes F e Q, a melhor classificação é obtida através das distâncias *Euclidean* e *Manhattan*, ambas apresentando um valor H de 0,88 classificando o conjunto F.

Tabela 11 – Taxas de acerto do OPF considerando 3 classes.

	Métrica de Distância		
Conjunto [%]	Euclidean [%]	Chi-Square [%]	Manhattan [%]
A	81,00	83,41	77,82
B	80,43	88,89	80,18
C	81,41	87,68	84,61
D	90,92	89,13	**91,42**
E	86,81	89,08	86,80
F	89,46	85,35	90,78
	Métrica de Distância		
Conjunto [%]	Canberra [%]	Squared Chi-Squared	Bray-Curtis [%]
A	78,29	76,43	80,20
B	81,21	81,46	88,48
C	85,18	82,84	76,88
D	91,07	90,94	88,92
E	86,84	86,90	82,11
F	86,87	85,94	78,73

Fonte: Autor

Os valores de H para o conjunto B geralmente são baixos, comparados com os maiores valores para cada classe, para todas as distâncias, enquanto os resultados do conjunto C são os melhores para as classes N e $SVEB$. A distribuição das amostras, por classe, do conjunto C são apresentadas, através de PCA, na Figura 16. As classe N e VEB apresentam partições mais aglutinadas, portanto tendem a ser classificadas de maneira mais eficiente que a classe $SVEB$, que mostra um maior espalhamento. Isto explica os valores de H mais baixos para a classe $SVEB$ em relação às outras classes.

Uma análise considerando o desempenho em todas as classes é realizada através dos valores médios M_H, mostrados na Tabela 13.

O maior resultado de M_H, é obtido utilizando-se a distância *Squared Chi-Squared*, classificando C (0,73), mesmo conjunto destacado na classificação em 5 classes. Esse resultado foi obtido por uma menor generalização da classe N, o que resulta em uma taxa de acerto de 82,84%, que é 10% mais baixa que a maior taxa de acerto.

A matriz de confusão dessa classificação é mostrada na Tabela 14, juntamente com a matriz de confusão da classificação do conjunto D utilizando a métrica de distância *Manhattan*, a qual obtém a maior taxa de acerto.

Através da análise das matrizes de confusão observa-se que, com a redução para três classes, houve pouca alteração nas classificações envolvendo as classes N e $SVEB$, porém tem-se um pequeno ganho de aproximadamente de 5% para a classe VEB. Com esses resultados, pode-se ressaltar que a distância *Squared Chi-Squared*, classificando a base extraída por C confunde-se menos as classes patológicas, principalmente a classe $SVEB$, acertando 3 vezes mais amostras que a classificação da distância *Manhattan* para o conjunto D.

Os resultados de custo computacional para treino, teste e tempo total estão apresentados

Tabela 12 – H, sensibilidade e especificidade para as distâncias do OPF considerando 3 classes. Valores multiplicados por 100.

		Classe do batimento		
		N	SVEB	VEB
Métrica	Conjunto	H|Se|Sp	H|Se|Sp	H|Se|Sp
Euclidean	A	066 | 085 | 054	002 | 001 | 097	082 | 077 | 088
	B	050 | 086 | 035	005 | 002 | 097	062 | 047 | 090
	C	074 | 085 | 066	031 | 018 | 095	080 | 072 | 089
	D	073 | 096 | 059	030 | 018 | 099	081 | 070 | 096
	E	065 | 092 | 050	006 | 003 | 098	074 | 060 | 094
	F	074 | 093 | 062	022 | 012 | 099	**088** | 082 | 094
Chi-Square	A	016 | 093 | 009	003 | 001 | 099	016 | 009 | 094
	B	002 | 100 | 001	000 | 000 | 100	002 | 001 | 100
	C	015 | 098 | 008	002 | 001 | 100	016 | 009 | 098
	D	004 | 100 | 002	000 | 000 | 100	004 | 002 | 100
	E	004 | 100 | 002	000 | 000 | 100	004 | 002 | 100
	F	012 | 095 | 006	002 | 001 | 099	011 | 006 | 097
Manhattan	A	066 | 081 | 056	003 | 002 | 095	083 | 080 | 086
	B	046 | 087 | 031	006 | 003 | 097	056 | 041 | 090
	C	076 | 088 | 066	031 | 018 | 097	081 | 072 | 091
	D	071 | 096 | 056	023 | 013 | 099	081 | 070 | 097
	E	064 | 092 | 048	006 | 003 | 098	073 | 060 | 094
	F	072 | 095 | 058	011 | 006 | 099	**088** | 081 | 096
Canberra	A	071 | 081 | 063	047 | 031 | 098	077 | 072 | 083
	B	042 | 088 | 028	005 | 003 | 098	051 | 035 | 091
	C	**078** | 088 | 071	056 | 039 | 096	081 | 073 | 092
	D	071 | 096 | 057	026 | 015 | 099	081 | 070 | 097
	E	064 | 092 | 049	007 | 004 | 098	074 | 061 | 094
	F	064 | 092 | 049	007 | 004 | 099	078 | 068 | 093
Chi-Squared Squared	A	066 | 080 | 056	006 | 003 | 097	078 | 074 | 083
	B	048 | 088 | 033	004 | 002 | 098	059 | 044 | 090
	C	**078** | 085 | 072	**060** | 044 | 095	081 | 074 | 090
	D	074 | 096 | 060	032 | 019 | 099	081 | 070 | 096
	E	065 | 093 | 050	006 | 003 | 098	074 | 061 | 094
	F	069 | 090 | 056	020 | 011 | 099	082 | 074 | 091
Bray-curtis	A	051 | 086 | 036	018 | 010 | 098	057 | 042 | 088
	B	001 | 099 | 001	000 | 000 | 100	001 | 000 | 099
	C	055 | 083 | 041	008 | 004 | 096	057 | 042 | 087
	D	002 | 100 | 001	001 | 000 | 100	000 | 000 | 100
	E	035 | 090 | 022	005 | 003 | 098	045 | 030 | 092
	F	053 | 085 | 039	003 | 002 | 096	055 | 039 | 089

Fonte: Autor

na Tabela 15, em que os melhores tempos para a etapa de treino são obtidos com as métricas *Bray-Curtis* e *Manhattan*, ambas classificando o conjunto D com 32,22 e 40,09 segundos, respectivamente.

O terceiro melhor tempo de treino é obtido também utilizando a métrica *Bray-Curtis* classificando o conjunto extraído segundo B (43,35). Na etapa de teste, a métrica *Bray-curtis* apresenta

Figura 16 – Distribuição das amostras do conjunto C, nas 3 classes, utilizando PCA, considerando as 3 componentes principais.

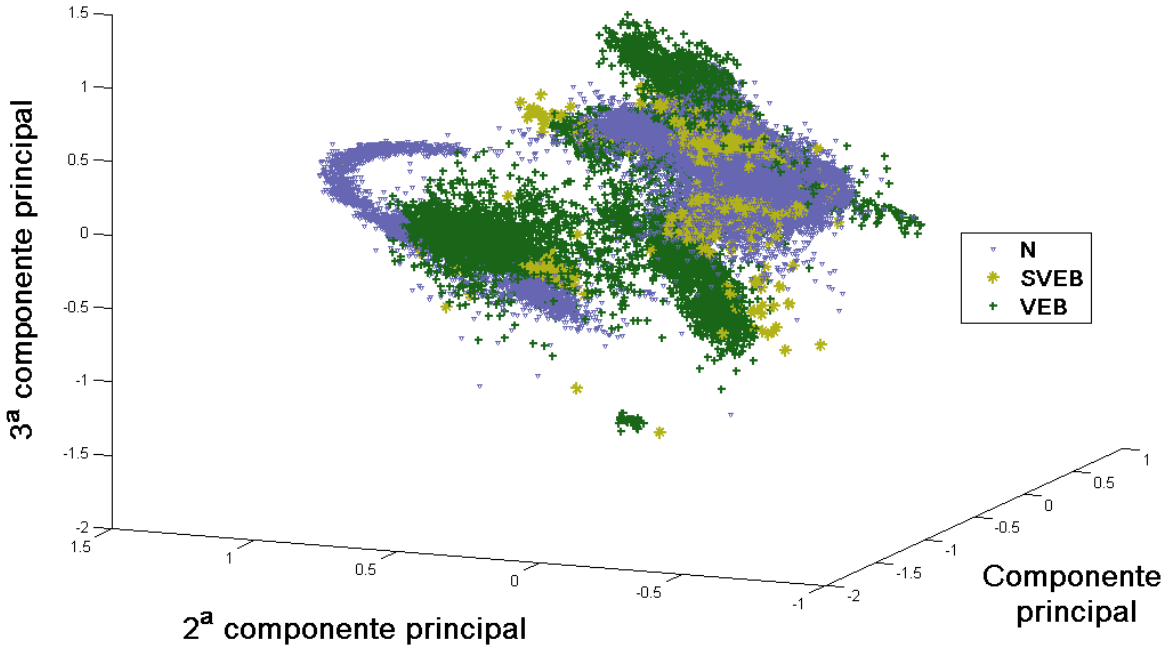


Tabela 13 – M_H das distâncias do OPF considerando 3 classes.

	Métrica de Distância					
Conjunto	Euclidean	Chi-Square	Manhattan	Canberra	Squared Chi-Squared	Bray Curtis
A	0,50	0,12	0,51	0,65	0,50	0,42
B	0,39	0,01	0,36	0,33	0,37	0,01
C	0,61	0,11	0,62	0,72	**0,73**	0,40
D	0,61	0,03	0,58	0,59	0,62	0,01
E	0,48	0,03	0,48	0,48	0,48	0,28
F	0,61	0,08	0,57	0,50	0,57	0,37

fonteAutor

o melhor resultado, considerando o conjunto B, levando 0,06 segundos para executar a classificação. O segundo e terceiro melhores tempos, são obtidos utilizando as métricas *Chi-Square* (2,74 segundos) e *Bray-Curtis* (6,01 segundos), ambos classificando o conjunto extraído por D. É importante ressaltar que esses resultados de custo computacional são acompanhados por uma classificação satisfatória, na qual os conjuntos que apresentaram os três melhores tempos de teste têm valor M_H muito baixo, devido a uma péssima classificação dos sinais patológicos das classes $SVEB$ e VEB.

4.1.3 Análise comparativa entre os classificadores considerando 5 classes

Para comparar a performance do OPF em relação aos classificadores tradicionais (SVM e Bayesiano), são consideradas apenas duas métricas de distância diferentes, *Manhattan* e *Squared Chi-Squared*,sendo a primeira a que obteve melhores taxas de acerto e a segunda os melho-

Tabela 14 – Matrizes de confusão da classificação utilizando distâncias *Manhattan* e *Squared Chi-Squared*

Manhattan - (YU; CHEN, 2007)				
		Classe verdadeira		
		N	SVEB	VEB
Classificado como	N	42657	1463	931
	SVEB	392	239	152
	VEB	1189	134	2526

Squared Chi-Squared - (SONG et al., 2005)				
		Classe verdadeira		
		N	SVEB	VEB
Classificado como	N	37680	612	913
	SVEB	2495	802	35
	VEB	4043	422	2659

Fonte: Autor

Tabela 15 – Custo computacional do OPF considerando 3 classes. Tempo em segundos

	Métrica de Distância		
	Euclidean	Chi-Square	Manhattan
	treino ǀ teste ǀ total Conjunto	treino ǀ teste ǀ total	treino ǀ teste ǀ total
A	445,18 ǀ 682,89 ǀ 1128,07	3230,23 ǀ 1987,14 ǀ 5217,37	335,07 ǀ 584,79 ǀ 919,86
B	145,53 ǀ 173,48 ǀ 319,01	416,65 ǀ 6,57 ǀ 423,21	54,50 ǀ 94,53 ǀ 149,03
C	142,70 ǀ 177,63 ǀ 320,34	464,50 ǀ 102,25 ǀ 566,74	55,19 ǀ 97,24 ǀ 152,43
D	128,90 ǀ 132,96 ǀ 261,86	299,52 ǀ 2,74 ǀ 302,26	40,09 ǀ 53,07 ǀ 93,15
E	172,76 ǀ 181,13 ǀ 353,89	650,34 ǀ 14,66 ǀ 665,00	80,58 ǀ 108,42 ǀ 189,00
F	317,86 ǀ 444,96 ǀ 762,82	2103,07 ǀ 818,28 ǀ 2921,35	219,07 ǀ 360,48 ǀ 579,55
	Métrica de Distância		
	Canberra	Squared Chi-Squared	Bray-Curtis
	treino ǀ teste ǀ total Conjunto	treino ǀ teste ǀ total	treino ǀ teste ǀ total
A	1479,33 ǀ 1588,78 ǀ 3068,11	1478,00 ǀ 1588,67 ǀ 3066,67	1143,30 ǀ 1168,50 ǀ 2311,81
B	187,74 ǀ 161,54 ǀ 349,28	189,63 ǀ 183,48 ǀ 373,11	43,35 ǀ **0,06** ǀ 43,41
C	203,89 ǀ 225,74 ǀ 429,63	204,11 ǀ 230,33 ǀ 434,44	104,88 ǀ 118,07 ǀ 222,96
D	130,45 ǀ 142,55 ǀ 273,00	132,00 ǀ 136,51 ǀ 268,52	**35,22** ǀ 6,01 ǀ **41,22**
E	297,03 ǀ 198,04 ǀ 495,07	299,07 ǀ 227,00 ǀ 526,07	150,78 ǀ 125,34 ǀ 276,12
F	969,35 ǀ 997,24 ǀ 1966,59	968,29 ǀ 976,22 ǀ 1944,50	431,96 ǀ 519,69 ǀ 951,66

Fonte: Autor

res índices M_H. Para efeitos de simplificação, as seguintes abreviaturas são utilizadas:

- **OPF-L1**: OPF utilizando a métrica de distância Manhattan;
- **OPF-SCS**: OPF utilizando a métrica de distância Squared Chi-Squared;
- **SVM-RBF**: Máquinas de Vetores de Suporte utilizando *kernel* RBF;
- **BC**: Classificador Bayesiano.

4.1.3.1 Avaliação dos classificadores considerando 5 classes

A taxa de acerto de cada classificador para os conjuntos considerando 5 classes de batimentos cardíacos é mostrado na Tabela 16. Em parênteses o desvio padrão.

Tabela 16 – Taxas de acerto dos classificadores considerando 5 classes.

	Classificador			
Conjunto	OPF-L1 [%]	OPF-SCS [%]	SVM-RBF [%]	BC [%]
A	77,57 (0,00)	76,14 (0,00)	88,21 (1,04)	80,69 (0,00)
B	79,43 (0,00)	80,61 (0,00)	84,06 (0,65)	79,52 (0,00)
C	84,46 (0,00)	82,63 (0,00)	89,82 (1,02)	81,37 (0,00)
D	91,21 (0,00)	90,75 (0,00)	**94,09** (0,00)	90,95 (0,00)
E	86,47 (0,00)	86,62 (0,00)	87,06 (0,99)	86,82 (0,00)
F	90,39 (0,00)	85,70 (0,00)	87,12 (0,00)	89,14 (0,00)

Fonte: Autor

Comparando os algoritmos de classificação, a maior taxa de acerto é obtida utilizando o SVM-RBF, obtendo 94,09% seguido do OPF-L1, BC e OPF-SCS, obtendo 91,21%, 90,95% e 90,75, respectivamente, todos classificando o conjunto D. Nota-se também que há desvio padrão nos resultados de alguns conjuntos classificados pelo SVM, que se dá pela dependência do SVM quanto aos parâmetros C e γ, definidos por buscas em grade que nem sempre apresentam os mesmo valores, conforme descrito no Capítulo 3, seção 3.5.

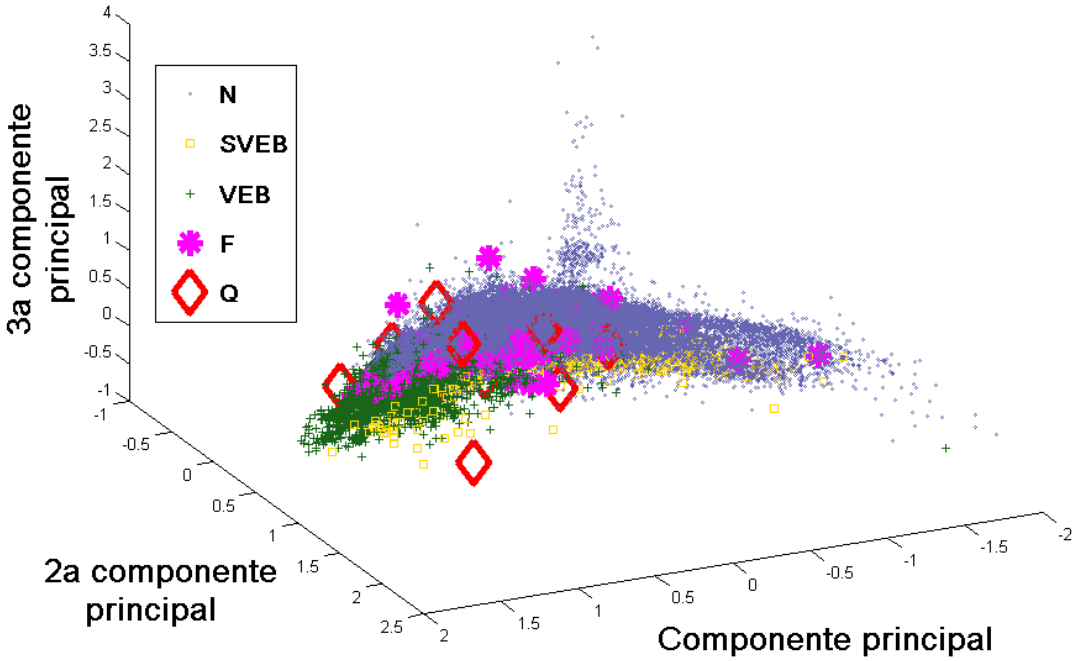
Na Figura 17, é mostrada a distribuição das amostras do conjunto D, que utiliza a técnica descrita por Yu e Chen (2007), a qual apresentou os melhores resultados em termos de taxa de acerto para todos os classificadores avaliados, inclusive para todas as métricas do OPF testadas.

Não há uma separação clara entre as classes, formando uma grande partição, permitindo uma generalização em relação à classe com a maior quantidade de amostras, no caso, a classe N. Essa generalização pode ser comprovada através dos dados de sensibilidade e especificidade e média harmônica H, mostrados na Tabela 17.

Os melhores resultados, em função da média harmônica H, para a classe N, são obtidos utilizando SVM-RBF (0,80),classificando o conjunto D. Este valor de H é aproximadamente 2% maior que o segundo melhor resultado, obtido pelo OPF-SCS, classificando o conjunto C (0,78). O melhor valor de H, obtido pelo OPF-L1, é de 0,76 considerando o conjunto C. Em relação ao BC, os melhores resultados são obtidos classificando os conjuntos C e F, com H igual a 0,74).

O melhor desempenho dos classificadores em relação à classe $SVEB$ são obtidos pelo OPF-SCS, classificando o conjunto C obtendo um valor de 0,60 para H. O SVM-RBF apresenta o segundo melhor valor de H, com 0,51 classificando o conjunto D, resultado 15% menor que o valor obtido para o OPF-SCS. Os classificadores OPF-L1 e Bayesiano, obtém o melhor valor de H para a classe $SVEB$ igual a 0,31, considerando o conjunto C.

Figura 17 – Distribuição das amostras do conjunto D, nas 5 classes, utilizando PCA, considerando as 3 componentes principais.



Em relação à classe VEB, o SVM-RBF apresenta H de 0,93, resultado aproximadamente 2% maior que o valor obtido pelo BC (0,91). O OPF-L1 e OPF-SCS apresentam H de 0,90 e 0,87. Todos os melhores resultados para a classe VEB são obtidos considerando o conjunto F, mostrando que a técnica de extração proposta por Ye, Coimbra e Kumar (2010) se destaca na caracterização desta classe.

Os melhores resultados para a classe F são obtidos considerando o conjunto A. O SVM-RBF apresenta o melhor resultado de H , com 0,82. O BC obtém um resultado aproximadamente 37% menor com 0,55, seguido do OPF-L1 (0,51) e OPF-SCS (0,37). Esses resultados mostram que o SVM-RBF apresenta acerto para classe F bem superior aos outros classificadores, quando aplicado ao conjunto A, porém com baixíssima sensibilidade para a classe $SVEB$ (0,06). O método de extração que apresenta os maiores valores de H para a classe F é o proposto por Chazal, O'Dwyer e Reilly (2004).

Em relação à classe Q, o único classificador à apresentar H diferente de zero é o SVM-RBF, classificando o conjunto B, porém com um valor baixo se comparando aos melhores valores de H obtidos para as outras classes.

As médias M_H, considerando todas as classes, são apresentadas na Tabela 18.

O classificador que apresenta melhor valor M_H é o SVM-RBF, com 0,51 classificando o conjunto A. Esse valor se deve à contribuição da classe F, com H de 0,82, porém com um baixa sensibilidade para a classe $SVEB$ (0,03). O segundo maior M_H, é obtido com o OPF-SCS (0,49) classificando o conjunto C, porém com uma sensibilidade para os batimentos da classe $SVEB$ igual a 0,44, aproximadamente 15 vezes maior que o SVM-RBF classificando A,

Tabela 17 – H, sensibilidade, especificidade dos classificadores, considerando 5 classes. Valores multiplicados por 100.

		Classe do batimento				
		N	SVEB	VEB	F	Q
Métrica	Conjunto	H|Se|Sp	H|Se|Sp	H|Se|Sp	H|Se|Sp	H|Se|Sp
OPF-L1	A	066 | 081 | 056	003 | 002 | 095	086 | 082 | 090	051 | 035 | 096	000 | 000 | 100
	B	046 | 087 | 031	006 | 003 | 097	051 | 035 | 090	001 | 000 | 099	000 | 000 | 100
	C	076 | 088 | 066	031 | 018 | 097	085 | 078 | 093	013 | 007 | 098	000 | 000 | 100
	D	071 | 096 | 056	023 | 013 | 099	085 | 075 | 098	008 | 004 | 099	000 | 000 | 100
	E	064 | 092 | 048	006 | 003 | 098	075 | 061 | 097	023 | 013 | 097	000 | 000 | 100
	F	072 | 095 | 058	010 | 006 | 099	090 | 083 | 098	034 | 021 | 098	000 | 000 | 100
OPF-SCS	A	065 | 080 | 056	006 | 003 | 097	081 | 076 | 086	037 | 023 | 097	000 | 000 | 100
	B	048 | 088 | 033	004 | 002 | 098	053 | 037 | 091	001 | 001 | 098	000 | 000 | 100
	C	078 | 085 | 072	**060** | 044 | 095	085 | 078 | 093	022 | 012 | 097	000 | 000 | 100
	D	073 | 096 | 060	032 | 019 | 099	085 | 075 | 097	009 | 005 | 099	000 | 000 | 100
	E	065 | 092 | 050	006 | 003 | 098	076 | 062 | 097	027 | 016 | 097	000 | 000 | 100
	F	069 | 090 | 056	020 | 011 | 099	087 | 079 | 098	008 | 004 | 093	000 | 000 | 100
SVM-RBF	A	074 | 092 | 063	006 | 003 | 099	093 | 091 | 096	**082** | 072 | 097	000 | 000 | 100
	B	049 | 091 | 033	001 | 000 | 100	061 | 045 | 093	000 | 000 | 098	**008** | 005 | 100
	C	070 | 094 | 056	015 | 008 | 098	089 | 083 | 097	003 | 002 | 100	000 | 000 | 100
	D	**080** | 098 | 067	051 | 035 | 099	090 | 082 | 099	004 | 002 | 100	000 | 000 | 100
	E	070 | 092 | 057	012 | 007 | 099	089 | 082 | 098	001 | 001 | 094	000 | 000 | 100
	F	074 | 091 | 063	026 | 015 | 099	**093** | 089 | 096	030 | 018 | 096	000 | 000 | 100
BC	A	066 | 084 | 054	002 | 001 | 097	084 | 078 | 091	055 | 038 | 097	000 | 000 | 100
	B	050 | 086 | 035	005 | 002 | 097	056 | 041 | 090	001 | 001 | 099	000 | 000 | 100
	C	074 | 085 | 066	031 | 018 | 095	084 | 078 | 091	013 | 007 | 097	000 | 000 | 100
	D	073 | 096 | 059	029 | 017 | 099	085 | 076 | 097	007 | 004 | 099	000 | 000 | 100
	E	064 | 093 | 049	005 | 003 | 098	076 | 062 | 097	027 | 016 | 097	000 | 000 | 100
	F	074 | 093 | 062	021 | 012 | 099	091 | 086 | 097	031 | 018 | 097	000 | 000 | 100

Fonte: Autor

mostrando-se um classificador menos generalista. O BC obtém um M_H de 0,44 classificando F. O OPF-L1 apresenta um M_H de 0,41 com os conjuntos A, C e F. As matrizes de confusão das classificações que apresentaram os dois melhores resultados são apresentadas na Tabela 19.

Na matriz de confusão considerando o SVM-RBF classificando o conjunto A (CHAZAL; O'DWYER; REILLY, 2004), existe uma confusão da classe $SVEB$ com a classe N, sendo classificadas apenas 37 amostras corretamente na primeira, ou seja 2%. Já a matriz de confusão relativa ao OPF-SCS classificando o conjunto C (SONG et al., 2005), as amostras classificadas corretamente na mesma classe é de 43%.

Os tempos de treino, teste, e total são apresentados na Tabela 20.

O classificador mais rápido na etapa de treinamento é o BC para todos os conjuntos, seguido do OPF-L1. O OPF-SCS é mais rápido que o SVM-RBF em todos os conjuntos, com exceção do conjunto F, onde o SVM-RBF apresenta o terceiro melhor tempo. Os tempos excessivos do SVM se dão devido à busca em grade, necessária para estabelecer os parâmetros C e γ.

O conjunto que apresenta os menores tempos de treino para todos os classificadores é o

Tabela 18 – M_H dos classificadores considerando 5 classes.

	Classificador			
Conjunto	OPF-L1	OPF-SCS	SVM-RBF	BC
A	0,41	0,38	**0,51**	0,41
B	0,20	0,21	0,24	0,22
C	0,41	0,49	0,36	0,41
D	0,37	0,40	0,45	0,39
E	0,33	0,35	0,35	0,35
F	0,41	0,37	0,45	0,44

Fonte: Autor

Tabela 19 – Matrizes de confusão da classificação utilizando distâncias *Manhattan* e *Squared Chi-Squared*

SVM-RBF - (CHAZAL; O'DWYER; REILLY, 2004)

		Classe verdadeira				
		N	SVEB	VEB	F	Q
Classificado como	N	40099	1705	221	83	3
	SVEB	361	37	3	0	0
	VEB	2285	23	2933	15	4
	F	1436	21	61	290	0
	Q	0	0	0	0	0

OPF-SCS - (SONG et al., 2005)

		Classe verdadeira				
		N	SVEB	VEB	F	Q
Classificado como	N	37677	612	591	322	2
	SVEB	2495	802	33	2	0
	VEB	2798	412	2514	17	5
	F	1245	9	81	47	0
	Q	3	2	0	0	0

Fonte: Autor

conjunto D. Para este conjunto, o BC apresenta tempo de 8,7 segundos para treino. O OPF-L1, treina com o mesmo conjunto em 40,3 segundos, aproximadamente 4 vezes o tempo do BC, enquanto os classificadores OPF-SCS e o SVM-RBF treinam com o mesmo conjunto em 132,4 e 170,7 segundos respectivamente.

Na etapa de teste, o melhor custo computacional é obtido através do SVM-RBF (6,7 segundos), sendo quase 8 vezes mais rápido que o OPF-L1 (53,3 segundos), ambos classificando o conjunto D. O terceiro melhor tempo é obtido com o OPF-SCS (131,3 segundos) enquanto o BC, apesar de ser o mais rápido na etapa de treino leva 173 segundos na fase de teste para classificar as amostras. Para todos os conjuntos, o SVM-RBF é o mais rápido no teste, sempre seguido do OPF-L1, OPF-SCS e BC.

Analisando o tempo total, o classificador que apresenta o menor custo computacional foi o OPF-L1. Os resultados mostram que o conjunto extraído pela técnica proposta por Yu e Chen (2007) tem um custo computacional menor para todas as etapas.

4.1.3.2 Avaliação dos classificadores considerando 3 classes

A seguir, são analisados os parâmetros de desempenho e custo computacional para a metodologia de Llamedo e Martínez (2011), agrupando as classes F e Q na classe $SVEB$. As taxas

Tabela 20 – Custo computacional dos classifcadores considerando 5 classes. Tempo em segundos e desvio padrão entre parênteses.

Conjunto	Classificador											
	OPF-L1						OPF-SCS					
	treino		teste		total		treino		teste		total	
A	337,0	(1,6)	584,2	(0,7)	921,2	(1,8)	1487,2	(10,8)	1604,7	(12,4)	3091,9	(22,3)
B	54,8	(0,7)	102,9	(13,4)	157,7	(13,7)	191,9	(2,4)	181,2	(4,0)	373,1	(3,6)
C	55,5	(0,4)	95,1	(4,0)	150,6	(3,9)	206,9	(2,0)	242,5	(7,9)	449,4	(9,9)
D	40,3	(0,2)	53,3	(7,3)	**93,6**	(7,2)	132,4	(0,8)	131,3	(4,9)	263,7	(5,4)
E	81,1	(0,8)	115,1	(3,5)	196,2	(3,2)	302,2	(3,7)	223,8	(4,9)	525,9	(6,9)
F	220,4	(2,1)	380,3	(6,9)	600,8	(6,1)	974,9	(6,7)	990,0	(3,4)	1964,8	(9,9)
Conjunto	Classificador											
	SVM-RBF						BC					
	treino		teste		total		treino		teste		total	
A	2668,4	(26,2)	32,2	(6,7)	2700,6	(31,3)	62,9	(0,3)	1622,2	(8,6)	1685,2	(8,9)
B	576,0	(474,0)	12,6	(1,7)	588,6	(472,3)	11,1	(0,2)	236,0	(2,7)	247,1	(2,8)
C	195,3	(8,6)	6,9	(2,1)	202,2	(10,5)	11,9	(0,0)	253,8	(3,0)	265,7	(2,9)
D	170,7	(5,7)	**6,7**	(0,0)	177,4	(5,7)	**8,7**	(0,1)	173,0	(1,9)	181,7	(1,9)
E	546,5	(48,2)	9,4	(0,7)	555,9	(47,5)	15,6	(0,1)	354,1	(4,1)	369,6	(4,0)
F	608,7	(9,5)	15,3	(0,1)	624,0	(9,6)	42,0	(0,1)	1058,8	(9,2)	1100,8	(9,2)

Fonte: Autor

de acerto são apresentadas na Tabela 21.

Tabela 21 – Taxas de acerto dos classificadores considerando 3 classes.

	Classificador			
Conjunto	OPF-L1	OPF-SCS	SVM-RBF	BC
A	77,82 (0,00)	76,43 (0,00)	80,01 (2,93)	80,98 (0,00)
B	80,18 (0,00)	81,46 (0,00)	84,29 (0,36)	80,31 (0,00)
C	84,61 (0,00)	82,84 (0,00)	90,01 (0,00)	81,53 (0,00)
D	91,42 (0,00)	90,94 (0,00)	**93,72** (0,00)	91,17 (0,00)
E	86,80 (0,00)	86,90 (0,00)	88,45 (1,84)	87,07 (0,00)
F	90,78 (0,00)	85,94 (0,00)	83,66 (1,50)	89,47 (0,00)

Fonte: Autor

Assim como acontece considerando 5 classes, o SVM-RBF obtém a melhor taxa de acerto, com 93,72%, seguido do OPF-L1 (91,42%), BC (91.17) e OPF-SCS (90,94), todos classificando o conjunto D.

Os valores de sensibilidade, especificidade e média harmônica H são mostrados na Tabela 22.

Para a classe N, o SVM-RBF é o classificador com o maior valor de H com 0,80 para o conjunto D, enquanto para a classe $SVEB$, o OPF-SCS apresenta o maior valores de H, com

Tabela 22 – H, sensibilidade, especificidade dos classificadores, considerando 3 classes. Valores multiplicados por 100.

		Classe do batimento		
		N	SVEB	VEB
Métrica	Conjunto	H|Se|Sp	H|Se|Sp	H|Se|Sp
OPF-L1	A	066 | 081 | 056	003 | 002 | 095	083 | 080 | 086
	B	046 | 087 | 031	006 | 003 | 097	056 | 041 | 090
	C	076 | 088 | 066	031 | 018 | 097	081 | 072 | 091
	D	071 | 096 | 056	023 | 013 | 099	081 | 070 | 097
	E	064 | 092 | 048	006 | 003 | 098	073 | 060 | 094
	F	072 | 095 | 058	011 | 006 | 099	**088** | 081 | 096
OPF-SCS	A	066 | 080 | 056	006 | 003 | 097	078 | 074 | 083
	B	048 | 088 | 033	004 | 002 | 098	059 | 044 | 090
	C	078 | 085 | 072	**060** | 044 | 095	081 | 074 | 090
	D	074 | 096 | 060	032 | 019 | 099	081 | 070 | 096
	E	065 | 093 | 050	006 | 003 | 098	074 | 061 | 094
	F	069 | 090 | 056	020 | 011 | 099	082 | 074 | 091
SVM-RBF	A	072 | 082 | 065	007 | 004 | 098	**088** | 092 | 084
	B	053 | 090 | 038	001 | 000 | 100	069 | 056 | 091
	C	070 | 095 | 056	012 | 007 | 099	083 | 074 | 095
	D	**080** | 098 | 067	052 | 035 | 099	084 | 074 | 098
	E	071 | 093 | 057	012 | 006 | 099	086 | 080 | 093
	F	072 | 087 | 062	023 | 013 | 099	086 | 085 | 088
BC	A	066 | 085 | 054	002 | 001 | 097	082 | 077 | 088
	B	050 | 086 | 035	005 | 002 | 097	062 | 047 | 089
	C	074 | 085 | 066	031 | 018 | 095	080 | 072 | 089
	D	073 | 096 | 059	029 | 017 | 099	082 | 071 | 096
	E	065 | 093 | 049	005 | 003 | 098	074 | 060 | 094
	F	074 | 093 | 062	021 | 012 | 099	**088** | 083 | 094

Fonte: Autor

0,60 considerando o conjunto C. Esses resultados são compatíveis com os resultados obtidos considerando 5 classes. Em relação à classe VEB, três classificadores apresentam melhor H igual a 0,88, sendo o SVM-RBF classificando o conjunto A e o OPF-L1 e o BC classificando o conjunto F. Nos três casos, os valores são acompanhados de baixos valores de sensibilidade para a classe $SVEB$.

Os valores médios M_H considerando a média das 3 classes são mostrados na Tabela 23.

Comparando as médias M_H entre os classificadores, nota-se que o maior valor é apresentado pelo OPF-SCS, com 0,73, classificando o conjunto C, seguido do SVM-RBF com 0,72 classificando o conjunto D. OPF-L1 e BC, apresentaram 0,62, ambos considerando o conjunto C.

Analisando os valores por classe dos conjuntos com os dois maiores valores de M_H, o SVM-RBF classificando o conjunto D obtém valores de H aproximadamente 2,5 e 3,7% maiores nas classes N e VEB, respectivamente, que o OPF-SCS classificando o conjunto C. Por outro

Tabela 23 – M_H dos classificadores considerando 3 classes.

	Classificador			
Conjunto	Manhattan	Squared Chi-Squared	SVM-RBF	BC
A	0,51	0,50	0,56	0,50
B	0,36	0,37	0,41	0,39
C	0,62	**0,73**	0,55	0,62
D	0,58	0,62	0,72	0,61
E	0,48	0,48	0,56	0,48
F	0,57	0,57	0,60	0,61

Fonte: Autor

lado, o OPF-SCS tem um desempenho, 15% maior para classe $SVEB$, apresentando menor generalização.

Na Tabela 24, são apresentados os resultados de custos computacionais em termos de tempo, para a classificação dos conjuntos considerando 3 classes.

Tabela 24 – Custo computacional dos classificadores considerando 3 classes. Tempo em segundos e desvio padrão entre parênteses.

Conjunto	Classificador											
	OPF-L1						OPF-SCS					
	treino		teste		total		treino		teste		total	
A	336,0	(1,7)	575,2	(11,5)	911,2	(9,9)	1486,1	(7,1)	1597,9	(8,1)	3084,1	(15,1)
B	54,5	(0,2)	93,0	(2,8)	147,5	(3,0)	191,2	(1,7)	192,3	(7,8)	383,4	(9,4)
C	55,5	(0,3)	92,3	(9,2)	147,7	(9,3)	206,2	(2,0)	233,5	(12,7)	439,7	(14,1)
D	40,2	(0,2)	52,2	(5,2)	**92,4**	(5,4)	132,6	(0,5)	134,1	(3,8)	266,7	(3,5)
E	81,0	(0,4)	105,9	(4,2)	186,9	(4,0)	301,6	(2,6)	224,8	(4,2)	526,4	(2,2)
F	221,7	(2,7)	368,8	(11,6)	590,5	(14,1)	973,4	(4,7)	977,3	(2,8)	1950,7	(6,8)
Conjunto	Classificador											
	SVM-RBF						BC					
	treino		teste		total		treino		teste		total	
A	2069,7	(23,7)	25,9	(5,4)	2095,6	(28,6)	62,5	(0,1)	971,2	(5,2)	1033,7	(5,3)
B	280,5	(2,9)	11,6	(0,7)	292,2	(3,6)	11,0	(0,1)	141,5	(0,7)	152,5	(0,7)
C	194,8	(2,7)	7,7	(0,0)	202,5	(2,7)	11,7	(0,1)	152,9	(0,3)	164,6	(0,4)
D	165,9	(1,0)	**6,0**	(0,0)	171,9	(0,9)	**8,7**	(0,1)	105,4	(1,4)	114,1	(1,5)
E	536,1	(58,9)	9,1	(0,8)	545,2	(58,1)	15,5	(0,1)	213,5	(2,2)	229,0	(2,3)
F	553,3	(26,2)	17,0	(6,8)	570,3	(32,9)	41,6	(0,3)	638,5	(5,1)	680,1	(5,3)

Fonte: Autor

Os valores de tempo de treino para os classificadores BC, OPF-L1 e OPF-SCS são próximos aos tempos obtidos com 5 classes, mostrando que esses classificadores são robustos no que diz respeito à quantidade de classes envolvidas na classificação. O classificador SVM-RBF apresentou menores tempos com a redução do número de classes. Em relação à fase de teste, o OPF-L1 e OPF-SCS também apresentaram tempos compatíveis com os tempos para 5 classes,

sem alterações significativas, enquanto, para o SVM-RBF e o BC, houve redução de custo para a maioria dos conjuntos.

O menor custo computacional para treino é apresentado pelo classificador BC, seguido do OPF-L1, para todos os conjuntos, assim como ocorre considerando 5 classes. Com exceção dos conjuntos C e F, onde o OPF levou mais tempo para treinar, o SVM-RBF é o classificador com maior custo computacional de treino, devido à busca em grade para obter-se os parâmetros C e γ.

Na etapa de teste, o SVM-RBF obteve os melhores tempos, sendo muito superior aos outros classificadores. Classificando o conjunto D, levou apenas 6,0 segundos, enquanto o OPF-L1 levou 53,3 segundos, sendo aproximadamente 9 vezes mais lento. O BC e OPF-SCS também obtiveram os melhores tempos classificando o conjunto D com 105,4 e 134,1 segundos, respectivamente.

Quando considera-se a soma dos dois tempos, de treino e de teste, o classificador mais eficiente é o OPF-L1, com 92,4 segundos de tempo total, seguido do BC com 112,38 segundos. O SVM-RBF obteve tempo total de 113,60 segundos para enquanto o OPF-SCS levou 268,52 segundos para completar as tarefas de treino e teste. Os melhores tempos totais foram todos obtidos através do conjunto D, mostrando que, assim como para 5 classes, a técnica de extração proposta em Yu e Chen (2007) resulta em eficiência dos classificadores. Isso pode ser explicado pela baixa dimensionalidade dos vetores de atributos para esse conjunto.

4.1.4 Discussão dos resultados

Entre as distâncias testadas com o OPF, destacam-se as distâncias *Manhattan* e *Squared Chi-Squared*, sendo a primeira com maior taxas de acerto, classificando o conjunto D e a segunda com menor generalização, obtendo maiores valores de M_H, classificando o conjunto C.

O menor custo computacional entre as distâncias testadas foi obtido com a distância *Bray-Curtis*, porém com grande generalização da classe N, o que explicaria o baixo tempo para processar as amostras.

Comparando com os outros classificadores, as maiores taxas de acerto são obtidas pelo classificador SVM, classificando também o conjunto (YU; CHEN, 2007).

Analisando o parâmetro M_H, considerando 5 classes, constata-se que o SVM-RBF foi o classificador com melhores resultados, apenas 2% acima do OPF utilizando a métrica *Squared Chi-Squared*, que obteve o segundo melhor resultado. O SVM, porém, apresentou péssima sensibilidade para a classe $SVEB$, de grande importância para a classificação. Considerando 3 classes, o OPF obteve o melhor valor M_H, sendo 1% maior que o SVM, em especial na classe $SVEB$, onde a sensibilidade foi 15% maior.

Entre as técnicas de extração,o método de Yu e Chen (2007) mostrou maiores taxas de acerto, e em especial para o OPF, os atributos propostos por Song et al. (2005) apresentou maiores valores para H_m, indicando menor generalização das classes.

5 CONCLUSÃO

Foi realizado um estudo detalhado do desempenho e custo computacional de algoritmos de classificação, em especial o classificador Floresta de Caminhos Ótimos, aplicado a sinais de ECG para a detecção de arritmias, em que conclui-se que:

- Seis distâncias foram avaliadas com o uso do OPF, dentre as quais as melhores taxas de acerto foram obtidas pela métrica *Manhattan*, enquanto uma melhor generalização é obtida pela distânica *Square Chi-Square*;
- Seis técnicas de extração de atributos foram avaliadas, dentre as quais as melhores taxas de acerto e melhor generalização dependem da técnica utilizada para cada classificador.
- Comparando com os algoritmos de classificação de Máquinas de Vetores de Suporte (SVM) e classificador Bayesiano (BC), o OPF foi menos generalista, enquanto o SVM obteve maiores taxas de acerto.

O OPF, por ser menos generalista no que se refere às classes VEB e $SVEB$, de maior interesse clínico, em relação à classe N, demonstra ser a ferramenta mais adequada para a classificação de arritmias em sinais de ECG.

Apesar do OPF ser menos generalista considerando todas as classes, caso seja de interesse o estudo de um tipo específico de arritmia, pode-se optar por uma combinação de classificador, distância e método de extração que aumente a eficiência na classificação da classe. Se o interesse for de classificar especificamente a classe N, o SVM mostra ser mais eficiente.

5.1 Trabalhos Futuros

Como tema de trabalhos futuros, é sugerida a aplicação de novas métricas de distância, não implementadas na biblioteca do OPF, outros algorítimos de seleção de protótipos, a aplicação em sinais de ECG em tempo real, e avaliação dos classificadores em outros sinais fisiológicos.

REFERÊNCIAS

ABAWAJY, J.; KELAREV, A.; CHOWDHURY, M. Multistage approach for clustering and classification of ECG data. *Computer Methods and Programs in Biomedicine*, v. 112, p. 720–730, 2013.

ADDISON, P. S. Wavelet transforms and the ECG: a review. *Physiological Measurement*, v. 26, n. 5, p. 155–199, 2005.

AFONSO, V. X. et al. ECG beat detection using filter banks. *IEEE Transactions on Biomedical Engineering*, v. 46, n. 2, p. 192–202, 1999.

AHLSTROM, M. L.; TOMPKINS, W. J. Automated high-speed analysis of holter tapes with microcomputers. *Biomedical Engineering, IEEE Transactions on*, v. 30, p. 651–657, 1983.

ANSI/AAMI. *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*. 2008. Association for the Advancement of Medical Instrumentation -AAMI / American National Standards Institute, Inc.-ANSI. ANSI/AAMI/ISO EC57, 1998-(R)2008.

CAPPABIANCO, F. A. M. et al. Brain tissue mr-image segmentation via optimum-path forest clustering. *Computer Vision and Image Understanding*, p. 1047–1059, 2012.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, p. 27:1–27:27, 2011. Versão 3.18, disponível em <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, acessado em 03 de abril de 2014.

CHAZAL, P.; O'DWYER, M.; REILLY, R. B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, v. 51, n. 7, p. 1196–1206, 2004.

CHEN, S. Y. Y. Selection of effective features for ECG beat recognition based on nonlinear correlations. *Artificial Intelligence in Medicine*, v. 54, p. 43–52, 2012.

CLIFFORD, G.; AZUAJE, F.; MCSHARRY, P. *Advanced methods and tools for ECG data analysis*. [S.l.]: Artech House, 2006. (Artech House engineering in medicine & biology series).

CONOVER, M. B. *Understanding electrocardiography*. 8. ed. [S.l.]: Mosby, 2002. ISBN 0323019056.

CORTES, C.; VAPNIK, V. Support vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.

DAAMOUCHEA, A. et al. A wavelet optimization approach for ECG signal classification. *Biomedical Signal Processing and Control*, v. 7, n. 4, p. 342–349, 2012.

DUTTA, S.; CHATTERJEE, A.; MUNSHI, S. Correlation technique and least square support vector machine combine for frequency domain based ECG beat classification. *Medical Engineering & Physics*, v. 32, n. 10, p. 1161–1169, 2010.

FALCÃO, A. X.; STOLFI, J.; LOTUFO, R. A. The image foresting transform theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 26, n. 1, p. 19–29, 2004.

FREITAS, G. M. de et al. Estimativa de ocorrência de precipitaćão em áreas agrícolas utilizando floresta de caminhos ótimos. *Revista Brasileira de Meteorologia*, v. 25, n. 1, p. 13–23, 2010.

FRIESEN, G. M. et al. A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on Biomedical Engineering*, v. 37, n. 1, p. 85–98, 1990.

GAO, D. et al. Bayesian ann classifier for ECG arrhythmia diagnostic system: A comparison study. In: *Proceedings of International Joint Conference on Neural Networks*. Montreal, Canadá: IEEE, 2005. v. 4, p. 2383–2388.

GOLDBERGER, A. L. et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation Journal*, v. 101, n. 23, p. e215–e220, 2000.

GUILHERME, I. R. et al. Petroleum well drilling monitoring through cutting image analysis and artificial intelligence techniques. *Engeneering Applications of Artficial Inteligence*, v. 24, p. 201–207, 2011.

GÜLER, I.; ÜBEYLI, E. D. ECG beat classifier designed by combined neural network model. *Pattern Recognition*, v. 38, n. 2, p. 199–208, 2005.

HE, T.; GARI, C.; LIONEL, T. Application of independent component analysis in removing artefacts from the electrocardiogram. *Neural Computing and Applications*, v. 15, p. 105–116, 2006.

HEIMANN, T.; MEINZER, H.-P. P. Statistical shape models for 3D medical image segmentation: a review. *Medical image analysis*, v. 13, p. 543–563, 2009.

HOMAEINEZHAD, M. R. et al. Ecg arrhythmia recognition via a neuro-SVM-KNN hybrid classifier with virtual qrs image-based geometrical features. *Expert Systems with Applications*, v. 39, p. 2047–2058, 2012.

HU, Y. H.; PALREDDY, S.; TOMPKINS, W. J. A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering*, v. 44, n. 9, p. 891–900, 1997.

HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, v. 10, n. 3, p. 626–634, 1999.

ILIEV, A. I. et al. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*, v. 24, n. 3, p. 445–460, 2010.

KABIR, A.; SHAHNAZ, C. Denoising of ECG signals based on noise reduction algorithms in emd and wavelet domains. *Biomedical Signal Processing and Control*, v. 7, n. 5, p. 481–489, 2012.

KHAZAEE, A.; EBRAHIMZADEH, A. Classification of electrocardiogram signals with support vector machines and genetic algorithms using power spectral features. *Biomedical Signal Processing and Control*, v. 5, n. 4, p. 252–263, 2010.

KORÜEK, M.; DOĜAN, B. Ecg beat classificartion using particle swarm optimization and radial basis function neural network. *Expert Systems with Applications*, v. 37, p. 7563–7569, 2010.

KORÜREK, M.; DOGAN, B. ECG beat classification using particle swarm optimization and radial basis function neural network. *Expert Systems with Applications*, v. 37, n. 12, p. 7563–7569, 2010.

LLAMEDO, M.; MARTÍNEZ, J. P. Heartbeat classification using feature selection driven by database generalization criteria. *IEEE Transactions on Biomedical Engineering*, v. 58, n. 3, p. 616–625, 2011.

LUO, S.; JOHNSTON, P. A review of electrocardiogram filtering. *Journal of Electrocardiography*, v. 43, p. 486–496, 2010.

LUZ, E.; MENOTTI, D. How the choice of samples for building arrhythmia classifiers impact their performances. In: *Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE*. Boston, EUA: IEEE, 2011. p. 4988–4991.

MADEIRO, J. et al. An innovative approach of QRS segmentation based on first-derivative, hilbert and wavelet transforms. *Medical Engineering & Physics*, v. 34, p. 1236–1246, 2012.

MAR, T. et al. Optimization of ECG classification by means of feature selection. *IEEE Transactions on Biomedical Engineering*, v. 58, n. 8, p. 2168–2177, 2011.

MARK, R. G. et al. An annotated ECG database for evaluating arrhythmia detectors. *IEEE Transactions on Biomedical Engineering*, v. 29, n. 8, p. 600, 1982.

MARTIS, R. J.; ACHARYA, R.; ADELI, H. Current methods in electrocardiogram characterization. *Computers in Biology and Medicine*, v. 48, p. 133–149, 2014.

MARTIS, R. J. et al. Characterization of ECG beats from cardiac arrhythmia using discrete cosine transform in PCA framework,. *Kowledge-Based Systems*, v. 45, p. 76–82, 2013.

MARTIS, R. J. et al. Application of principal component analysis to ECG signals for automated diagnosis of cardiac health. *Expert Systems with Applications*, v. 39, p. 11792–11800, 2012.

MOAVENIAN, M.; KHORRAMI, H. A qualitative comparison of artificial neural networks and support vector machines in ECG arrhythmias classification. *Expert Systems with Applications*, v. 37, n. 4, p. 3088–3093, 2010.

MOODY, G. B.; MARK, R. G. QRS morphology representation and noise estimation using the karhunen-loeve transform. In: . [S.l.: s.n.], 1989. p. 269–272.

MOODY, G. B.; MARK, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, v. 20, n. 3, p. 45–50, 2001.

NEJADGHOLI, I.; MOHAMMAD, M. H.; ABDOLALI, F. Using phase space reconstruction for patient independent heartbeat classification in comparison with some benchmark methods. *Computers in Biology and Medicine*, v. 41, p. 411–419, 2011.

NOBLE, J. A.; BOUKERROUI, D. *IEEE Transactions on Medical Imaging*, v. 25, n. 8, p. 987–1010, 2006.

NUNES, T. M. et al. Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals. *Expert Systems with Applications*, v. 40, n. 8, p. 3096–3105, 2013.

NUNES, T. M. et al. EEG signal classification for epilepsy diagnosis via optimum path forest - a systematic assessment. *Neurocomputing*, v. 136, p. 103–123, 2014.

PAGNIN, A. F.; ARTIOLI, S. S.; PAPA, J. P. Preliminary diagnosis of ophthalmological diseases through machine learning techniques. *Recent Patents on Signal Processing*, v. 1, p. 74–79, 2011.

PAPA, J.; FALCÃO, A.; SUZUKI, C. *LibOPF: A library for the design of optimum-path forest classifiers*. [S.l.], 2009. Versão 2.0, disponvél em <http://www.ic.unicamp.br/~afalcao/LibOPF>, acessado em 09 de janeiro de 2013.

PAPA, J. P. et al. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, v. 45, n. 1, p. 512–520, 2012.

PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, v. 19, n. 2, p. 120–131, 2009.

PAPA, J. P. et al. Computer techniques towards the automatic characterization of graphite particles in metallographic images of industrial materials. *Expert Systems with Applications*, v. 40, n. 2, p. 590–597, 2013.

PARK, M. S.; KIM, K.; OH, S. R. A fast classification system for decoding of human hand configurations using multi-channel sEMG signals. In: *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. [S.l.]: IEEE, 2011. p. 4483–4487.

PEREIRA, C. R. et al. An optimum-path forest framework for intrusion detection in computer networks. *Engineering Applications of Artificial Intelligence*, v. 25, p. 1226–1234, 2012.

PEREIRA, L. A. et al. Aquatic weed automatic classification using machine learning techniques. *Computer and Eletronics in Agriculture*, v. 87, p. 56–63, 2012.

PISANI, R. et al. Automatic landslide recognition through optimum-path forest. In: *IEEE international Geoscience and Remote Sensing Symposium 2012*. Munique, Alemanha: IEEE, 2012. p. 6228–6231.

PISANI, R. J. et al. Toward satellite-based land cover classification through optimum-path forest. *IEEE Transactions on Geoscience and Remote Sensing*, v. 52, p. 6075–6085, 2014.

POLI, R.; CAGNONI, S.; VALLI, G. Genetic design of optimum linear and nonlinear QRS detectors. *IEEE Transactions on Biomedical Engineering*, v. 42, n. 11, p. 1137–1141, 1995.

RAI, H. M.; TRIVEDI, A.; SHUKLA, S. Ecg signal processing for abnormalities detection using multi-resolution wavelet transform and artificial neural network classifier. *Measurement*, v. 46, p. 3238–3246, 2013.

RAMOS, C. C. O. et al. A novel algorithm for feature selection using harmony search and its application for non-technical losses detection. *Computers and Electrical Engineering*, v. 37, p. 886–894, 2011.

SCHÖLKOPF, B.; SMOLA, A. J. *Learning with Kernels*. Cambridge, EUA: MIT Press, 2002.

SONG, M. H. et al. Support vector machine based arrhythmia classification using reduced features. *International Journal of Control, Automation, and Systems*, v. 3, n. 4, p. 509–654, 2005.

SOUZA, A. N. de et al. Efficient fault location in underground distribution systems through optimum-path forest. *Applied Artificial Intelligence*, v. 26, n. 5, p. 503–515, 2012.

SUZUKI, C. T. N. et al. Automatic segmentation and classification of human intestinal parasites from microscopy images. *IEEE Transactiions on Biomedmedical Engineering*, v. 60, n. 3, p. 803–812, 2013.

TRAHANIAS, P. An approach to qrs complex detection using mathematical morphology. *IEEE Transactions on Biomedical Engineering*, v. 40, p. 201–205, 1993.

VAPNIK, V. N. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988–999, 1999.

WANG, J.-S. et al. ECG arrhythmia classification using a probabilistic neural network with a feature reduction method. *Neurocomputing*, v. 116, p. 38–45, 2013.

WHO. *World Health Organization - Global status report on noncommunicable diseases - 2010.* 2011.

YE, C.; COIMBRA, M. T.; KUMAR, B. V. K. V. Arrhythmia detection and classification using morphological and dynamic features of ECG signals. In: *IEEE International Conference on Engineering in Medicine and Biology Society*. Buenos Aires, Argentina: IEEE, 2010. p. 1918–1921.

YU, S.; CHEN, Y. Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recognition Letters*, v. 28, n. 10, p. 1142–1150, 2007.

YU, S.; CHOU, K. Integration of independent component analysis and neural networks for ECG beat classification. *Expert Systems with Applications*, v. 34, n. 4, p. 2841–2846, 2008.