Centro de Ciências Exatas e Tecnologias - CCET
Programa de Pós-graduação em Engenharia Elétrica

Classificação de Subclasses de Fibrilação Atrial utilizando Estatística de Alta Ordem e Aprendizado de Máquina

Luis Fillype da Silva Lago Cutrim Barros

São Luis - MA, 2022

Luis Fillype da Silva Lago Cutrim Barros

Classificação de Subclasses de Fibrilação Atrial utilizando Estatísticas de Alta Ordem e Aprendizado de Máquina

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica do Centro de Ciências Exatas e Tecnologias, da Universidade Federal do Maranhão, como requisito para a obtenção do grau de Mestre em Engenharia Elétrica.

Orientador: Allan Kardec Duailibe Barros Filho

São Luis, 18 Fevereiro de 2022

CCET - Centro de Ciências Exatas e Tecnologias
Universidade Federal do Maranhão

Dissertação apresentada ao Programa de Pós-graduacão em Engenharia Elétrica titulada ***Classificação de Subclasses de Fibrilação Atrial utilizando Estatística de Alta Ordem e Aprendizado de Máquina*** de autoria de Luís Fillype da Silva Lago Cutrim Barros, aprovada pela banca examinadora constitu´ıda pelos seguintes professores:

Prof. Dr. Allan Kardec Duailibe Barros Filho -
Orientador
Programa de Pós-graduação em Engenharia Elétrica
- UFMA

Prof. Dr Ewaldo Eder Carvalho Santana
Membro da banca
Programa de Pós-graduação em Engenharia Elétrica
- UFMA

Prof. Dr. Carlos Alberto Bezerra Tomaz
Membro da banca
Coordenação Pesquisa em Neurociências - CEUMA

São Luis, 18 de Fevereiro de 2022

”Quero mudar o mundo, caminhar sem olhar para trás; Com você eu encontrei a paz, nas asas de um sonho não vou me perder jamais”

Inuyasha

DEDICATÓRIA

Dedico essa dissertação à minha mãe e ao meu irmão Daniel que não estão mais entre nós, mas sei que olham por mim em todo o meu processo de crescimento e amadurecimento. Dedico também aos meus pais, Amazonina e Carlindo, por todo o apoio até aqui, foi essencial, obrigado. À minha namorada, que me ajuda e apoia sempre. Amo todos vocês.

AGRADECIMENTOS

Queria agradecer fortemente aos meus pais por todo apoio. À minha namorada pela ajuda e entendimento. Aos meus amigos de UEMA, os famosos deuses emissários, que sempre me ajudam.

Aos meus amigos de laboratório do PIB, Davi, Marta, Caio, Juliana, Claudyane, Elias. Em especial ao meu grande amigo Jonathan Queiroz por toda a ajuda, todas as horas que me tirou dúvida e incentivo nos estudos. Ao meu orientador prof Allan que me deu a oportunidade de ingressar no laboratório e ser seu orientando, acreditando no meu potencial.

RESUMO

O eletrocardiograma (ECG) é um procedimento simples e rotineiro de grande importância para o diagnóstico de patologias cardíacas. Esse exame nos fornece uma representação gráfica da atividade elétrica do coração, que resulta em sua interpretação, pois apresenta ondas, segmentos e possíveis intervalos para mensuração e identificação das alterações presentes no órgão cardíaco. Esta dissertação tem como objetivo desenvolver um modelo de classificação baseado nos batimentos de quatro grupos de indivíduos: com fibrilação atrial paroxística, fibrilação atrial intracardíaca, fibrilação atrial e ritmo sinusal normal. A metodologia de extração de características baseada na amplificação das características, a fim de classificar indivíduos com Fibrilação Atrial, seus subtipos e saudáveis, com e sem o uso da técnica de Análise de Componentes Independentes (ICA). As classificações foram realizadas com base nas características das estatísticas das quatro bases de dados, avaliando as métricas dos algoritmos K-vizinhos mais próximos (KNN), Máquina de vetores de suporte (SVM), Rede Neural Artificial, (RNA), obtendo acurácia de 93,4% a 99,85%.

Palavras-chave: Fibrilação Atrial. Eletrocardiograma. Aprendizado de máquina. Estatística.

ABSTRACT

The electrocardiogram (ECG) is a simple and routine procedure of great importance for the diagnosis of cardiac pathologies. This exam gives us a graphic representation of the electrical activity of the heart, which results in its interpretation, as waves, segments and possible intervals for measuring and identifying the changes it presents in the cardiac organ. This dissertation aims to develop a classification model based on the beats of four groups of desired: with paroxysmal atrial fibrillation, intracardiac atrial fibrillation, atrial fibrillation and normal sinus rhythm. The methodology of extraction of characteristics based and adapted to classify with Atrial Fibrillation, its subtypes and healthy, with and without the use of the Independent Component Analysis (ICA) technique. As evaluated, they were evaluated based on the characteristics of the statistics of the four databases, evaluating as metrics the K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) algorithms, obtaining accuracy of 93.4% to 99.85%.

Key-words: Atrial Fibrillation. Eletrocardiogram. Machine Learning. Statistics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ECG	Eletrocardiograma
K-NN	K-Nearest Neighbors
SVM	Support Vector Machine
FA	Fibrilação Atrial
EOS	Estatísticas de Ordem Superior
OMS	Organização Mundial da Saúde
MLP	Multi-layer Perceptron
PCA	Principal Component Analysis
ICA	Independent Component Analysis

Conteúdo

1 INTRODUÇÃO

As doenças cardiovasculares matam mais e mais a cada ano que passa. Números divulgados pela OMS [1] mostram que cerca de 17,3 milhões de pessoas em todo o mundo são vítimas de doenças cardíacas a cada ano. A Fibrilação Atrial (FA), o tipo mais comum de arritmia cardíaca, é uma das principais causas de morbidade e mortalidade em todo o mundo. O diagnóstico oportuno de FA é uma tarefa igualmente importante e desafiadora devido à sua natureza assintomática e episódica [2] [3]. A OMS [1] reitera, ainda, que as principais causas de morte no mundo atualmente estão fortemente relacionadas a problemas decorrentes do coração, principalmente a FA. Diante do exposto, é necessário que o diagnóstico das cardiopatias seja mais eficaz.

Essa necessidade tem impulsionado o desenvolvimento de métodos autônomos que auxiliem na detecção e previsão dessas doenças cardíacas. Um exame que quantifica a atividade elétrica do coração, possibilitando detectar a frequência cardíaca e o número de batimentos por minuto, é a análise do eletrocardiograma (ECG). Desta maneira, o ECG é essencial para predizer, detectar e diagnosticar diversos problemas cardíacos, como a Fibrilação Atrial, pois é uma das técnicas não invasivas mais utilizadas para auxiliar neste diagnóstico (Silva et al., 2020, p. 1) [2] .

Trabalhos relacionados a esse tema são encontrados na literatura, Queiroz et al., [4] (2017), onde os autores investigam a variação da tensão que ocorre em um intervalo t de batimento cardíaco por meio de curtose. Kachue et al [3] propõem um método baseado em redes neurais convolucionais profundas para a classificação dos batimentos cardíacos, capaz de classificar com precisão cinco diferentes arritmias. Silva et al [2][4] realizam uma classificação da fibrilação atrial utilizando a mesma metodologia desta dissertação.

Este trabalho propõe extrair todo o batimento cardíaco de um ECG de quatro bancos de dados, a fim de agrupá-los por meio de Estatísticas de Alta Ordem. Ao longo da dissertação, são mostrados os resultados desde o começo da pesquisa, aonde classificou-se dicotomicamente as classes, até o problema de multiclassificação, juntando os 4 grupos. Utilizou-se Análise dos Componentes Principais e Análise dos Componentes Independentes para descorrelacionar os dados, a fim de na etapa de classificação obter-se melhores resultados com os quatro algoritmos de Aprendizado de Máquina.

1.1 Trabalhos relacionados

Diversos autores desenvolveram métodos para apoiar o diagnóstico de FA utilizam o intervalo R-R. No entanto, a análise dos intervalos R-R não é capaz de medir as distorções morfológicas na onda P, ou mesmo, para ciclo cardíaco. Já o método proposto Queiroz et al [4] usa variabilidade da tensão em cada batimento cardíaco, ao contrário do intervalo R-R, no qual cada ciclo cardía-co está associado a um único número real, o método proposto associa cada ciclo cardíaco a um conjunto de pontos, ou seja, a um vetor. Este método utiliza a variação de

tensão em cada ciclo cardíacos. Portanto, nos basearemos neste método para classificação de subclasses de Fibrilação Atrial utilizando Estatísticas de Alta Ordem e Aprendizado de Máquina.

Outro trabalho que utiliza os batimentos cardíacos ao invés do intervalo R-R é o de Silva et al (2020) [5][4], onde é realizada uma classificação dicotômica de Fibrilação Atrial e individuos saudáveis. A metodologia utilizada neste trabalho incorpora parte da utilizada neste artigo. Como contribuições e trabalhos relacionados a essa dissertação, o autor publicou quatro artigos relacionados a classificação de doenças cardíacas, em dicotomia. Nos resultados, foram mostrados valores de acurácia de 90%.

Já na perspectiva multi classe, trabalhos como o de Kachuee et. al. (2018) [3][6] propõe um método baseado em convoluções profundas redes neurais para a classificação de batimentos cardíacos, capaz de classificar com precisão cinco diferentes arritmias. Ullah et. al. (2020) [7] realiza um classificação de 8 tipos de arritmia também usando redes neurais convolucionais. Em seu trabalho, Ma et al. (2020) [8] utilizou o intervalo RR para a classificação, obtendo uma precisão de 98,3%.

Em outro estudo, os mesmos autores também utilizaram o intervalo RR, classificando com CNN-LSTM, obtendo um precisão de 97,21%. Alhusseini et al. (2020) [9] desenvolveu um CNN aplicado a imagens de 35 pacientes, que tomou decisões semelhantes às de especialistas, com 95% de precisão. Khriji et al. (2020) [10] usado ANN para classifique três tipos diferentes de doenças cardíacas também, obtendo 93,1% de precisão.

Em demais contribuições podemos destacar Silva et al, o qual se baseou a classificação multi classe igual a três, e quatro, já utilizando Analise dos Componentes Principais na metodologia.

Como supracitado, a comunidade acadêmica desprende uma variedade de esforços para identificação de FA baseadas e análise de padrões cardíacos, ocasionando em possíveis opções para acompanhamento no tratamento de pacientes que sofrem cardiopatias, bem como o auxílio ao diagnóstico de pessoas com suspeita.

Ressalta-se que a maioria das metodologias propostas para classificação baseia-se apenas no intervalo R-R. Em outras palavras, os classificadores ficam limitados a esse cálculo e não muito generalista. Desta maneira, neste trabalho a metodologia utilizada é a de Estatística de Alta Ordem, aliada a aplicação de Análise dos Componentes Independentes em situações de multi classes.

1.2 Objetivo Geral

Desenvolver um modelo, baseado em aprendizado de máquina, para a classificação generalizada dos batimentos cardíacos de indivíduos saudáveis e com Fibrilação Atrial

1.3 Objetivos Específicos

- Avaliar resultados de classificação de vetores de características;
- Investigar quais estatíticas separam com eficiência os pacientes saudáveis e com FA;
- Aplicar metodologia dicotomicamente e no problema de multiclasse;
- Realizar comparativo entre resultados deste trabalho e resultados anteriormente propostos.

2 CORAÇÃO E ELETROCARDIOGRAMA

A FA é a arritmia cardíaca sustentada mais frequente e é responsável por 33% de todas as internações por arritmia. Ela ocorre entre 1% e 2% na população geral, aumentando significativamente com o envelhecimento e com a presença de doenças cardíacas, conforme mostrado na Figura 1 na Introdução deste trabalho.

Ressaltando a importância desta pesquisa, esse tópico de Fundamentação Teórica irá guiar as pessoas que lerão esta dissertação nos conceitos principais para a elaboração deste trabalhos, tais como o coração, o ECG, as ondas provevientes do exame, e as cardiopatias aborda na problemática trazida aqui.

2.1 Coração

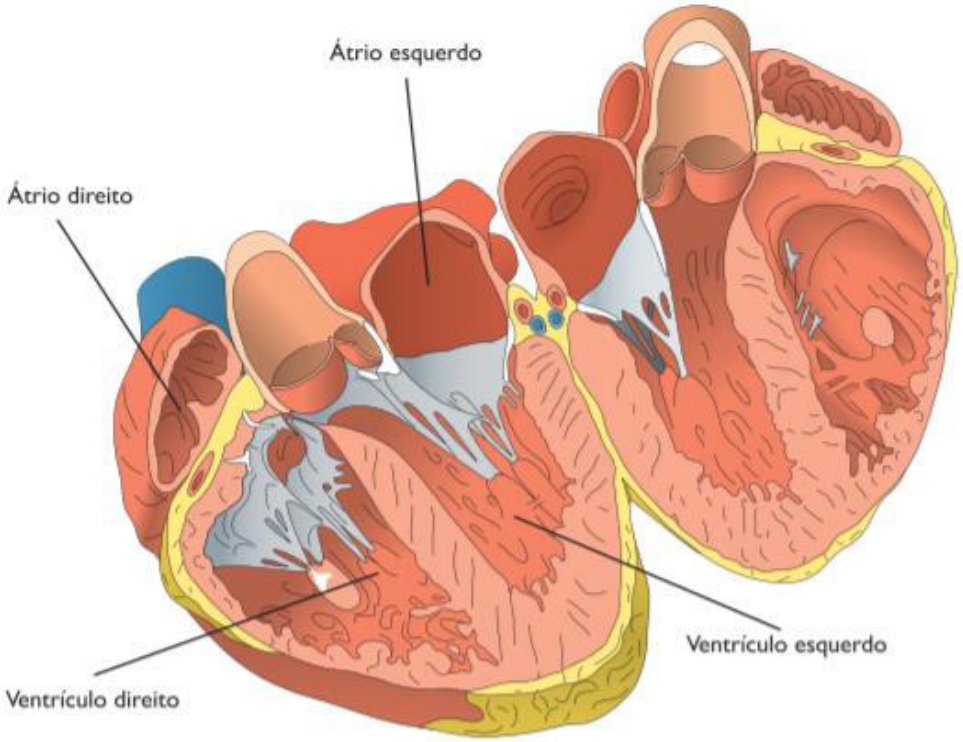
O coração é um órgão muscular, oco, tem forma de cone e funciona de modo similar a duas bombas, contrátil e propulsora. O órgão realiza dois movimentos básicos: sístole (contração) e diástole (relaxamento), de acordo com a despolarização e repolarização de suas cargas elétricas intra e extracelulares, estimuladas por íons como: sódio, potássio, magnésio, cálcio. São conduzidas por um sistema nervoso próprio, capaz de produzir automaticamente seus estímulos elétricos, iniciados por células especializadas que formam o nódulo sinoatrial, localizado na parede posterior do átrio direito. [11]

Sua divisão é conhecida como ápice, base e mais Enfermagem em Cardiologia Intervencionista 4 três faces: esternocostal, diafragmática e pulmonar. A base é formada pelos átrios direito e esquerdo. As veias cavas superior e inferior e as veias pulmonares penetram no coração pela base. É também a porção posterior do coração em posição anatômica. O ápice é contralateral a base e tem formato arredondado, formada pela parte inferolateral do ventrículo esquerdo e onde ocorre o batimento apical. [11]

Quanto às cavidades do coração, são subdivididas em quatro câmaras: átrios e ventrículos localizados à direita e à esquerda. O átrio direito se comunica com o ventrículo direito por meio do óstio atrioventricular direito, no qual existe uma estrutura direcionadora do fluxo, a valva atrioventricular direita (tricúspide). O mesmo ocorre à esquerda, por meio do óstio atrioventricular esquerdo, cuja comunicação de fluxo é por meio da valva atrioventricular esquerda (mitral). [12]

As cavidades direitas são separadas das esquerdas pelos septos interatrial e interventricular. A câmara esquerda (ventrículo) proporciona a força necessária para o sangue circular por todos os tecidos do corpo. Sua função é vital porque, para sobreviver, os tecidos necessitam receber continuamente oxigênio.

Figura 1: Morfologia coração.



Fonte: Dutra, 2018.

2.2 Eletrocardiograma e Ondas

O eletrocardiograma (ECG) é o registro dos fenômenos elétricos que se originam durante a atividade cardíaca por meio de um aparelho denominado eletrocardiógrafo, que é um galvanômetro, o qual mede pequenas intensidades de corrente que recolhe a partir de dois eletrodos dispostos em determinados pontos do corpo humano. Ele serve como um auxiliar valioso no diagnóstico de grande número de cardiopatias. [13]

Strapazzon et al., (2016) [14] definem o eletrocardiograma (ECG) como um procedimento simples e rotineiro de grande importância para diagnósticos de patologias cardíacas. Este procedimento corresponde a uma representação gráfica da atividade elétrica do coração, que resulta na sua interpretação por apresentar ondas, segmentos e intervalos possíveis de medir e identificar alterações no coração.

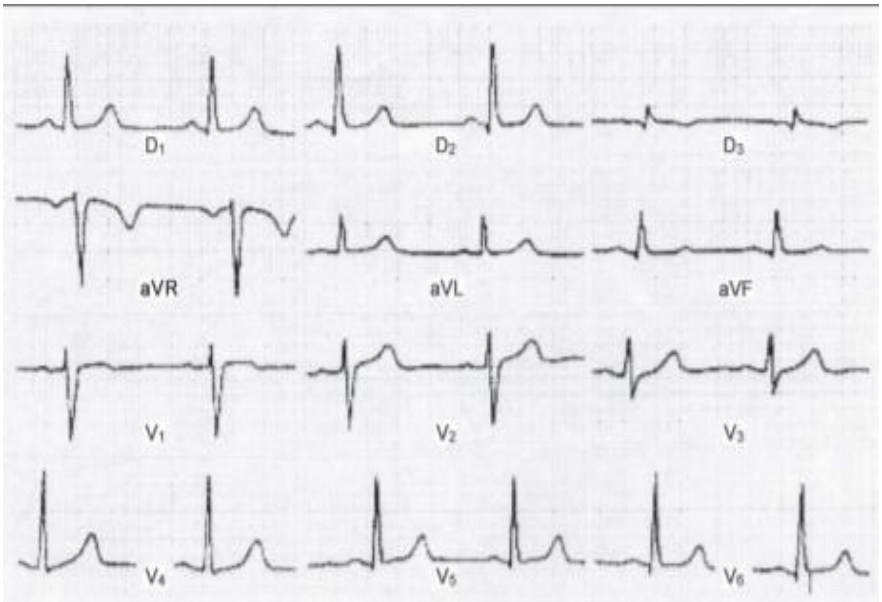
O ECG estabeleceu-se como um dos exames complementares de maior capacidade informativa, utilizado no diagnóstico, na avaliação de gravidade e no planejamento terapêutico de praticamente todas as doenças cardiovasculares. O fato de ser um método não invasivo, o baixo custo, a facilidade de transporte e de manuseio (que permite a realização de exames à beira do leito, no centro cirúrgico ou ambulatorialmente, por exemplo) muito contribuíram para que o ECG constituísse um método de rotina nas clínicas

e consultórios. Juntamente com o exame clínico é extremamente útil para detectar problemas cardíacos. É método soberano tanto nas arritmias, quanto nos distúrbios de formação, como aqueles de condução do estímulo. Na maioria dos casos de infarto do miocárdio, ainda que a clínica e os exames de laboratório sejam suficientes para suspeitar ou fazer um diagnóstico, este é confirmado pelo ECG que, além disso, fornece valiosas informações sobre a localização e evolução do processo, esclarecendo as dúvidas nos casos menos típicos, levando a um sinalizador do acidente coronário. [14] [15]

Percebe-se assim a suma importância do conhecimento aprofundado sobre o ECG normal, incluindo as informações pertinentes a este, como duração, amplitude e forma das ondas e segmentos, anatomia do coração e os locais de ativação relacionados às ondas eletrocardiográficas, eletrofisiologia cardíaca, com enfoque aos canais iônicos; além das 12 derivações cardíacas (periféricas e precordiais). Este conhecimento é fundamental para que a análise clínica seja eficaz, obtendo uma interpretação satisfatória e um diagnóstico preciso, identificando prováveis patologias ou a ausência destas. [16]

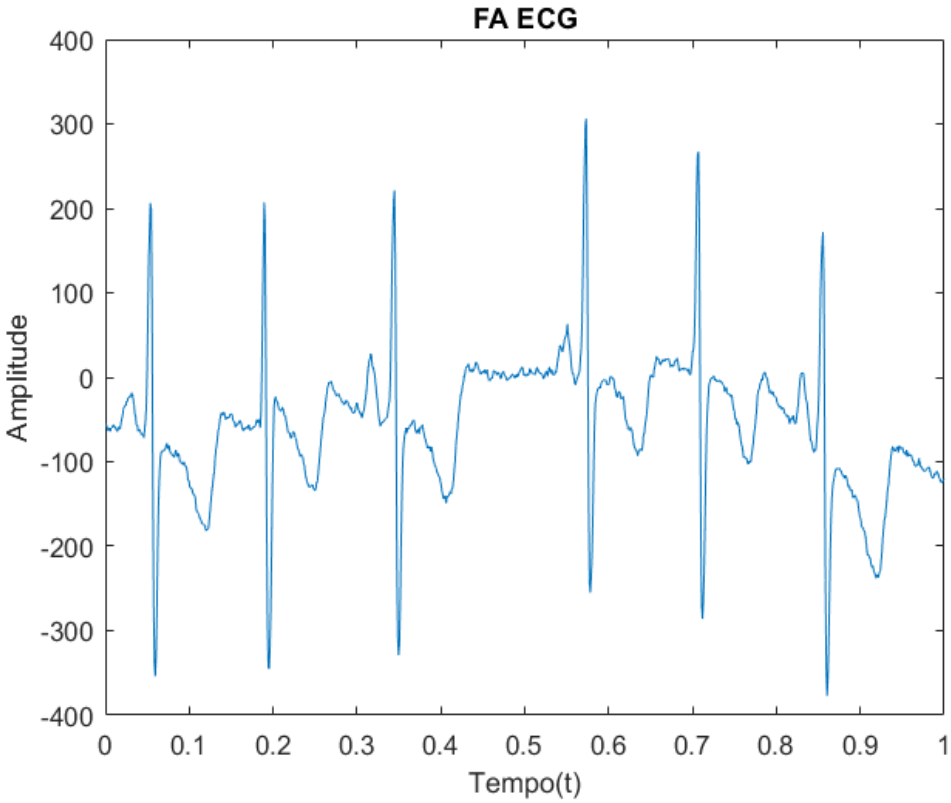
Deve-se destacar ainda a importância do processamento computacional que associa os sinais do ECG com patologias e morfologias cardíacas, permitindo uma avaliação mais precisa do quadro clinico do paciente. Esta idéia fundamenta os sistemas atuais de telemedicina, inclusive o cálculo automático computacional de escore. Além disso, o ECG é essencial para dois tipos principais de informações, intervalos de tempo do ECG para determinar quanto tempo o onda elétrica passa pelo coração sistema de condução elétrica. Esta informação encontra para descobrir se a atividade elétrica é regular ou irregular, rápido ou lento. Em segundo lugar, medindo o força da atividade elétrica, o cardiologista é capaz para descobrir se as partes do coração são muito grandes ou sobrecarregado (Ebrahimi et al., 2020) [17]

Figura 2: Eletrocardiograma



Fonte: Autor, 2021.

Figura 3: Eletrocardiograma de paciente com Fibrilação Atrial (FA)



Fonte: Autor, 2021.

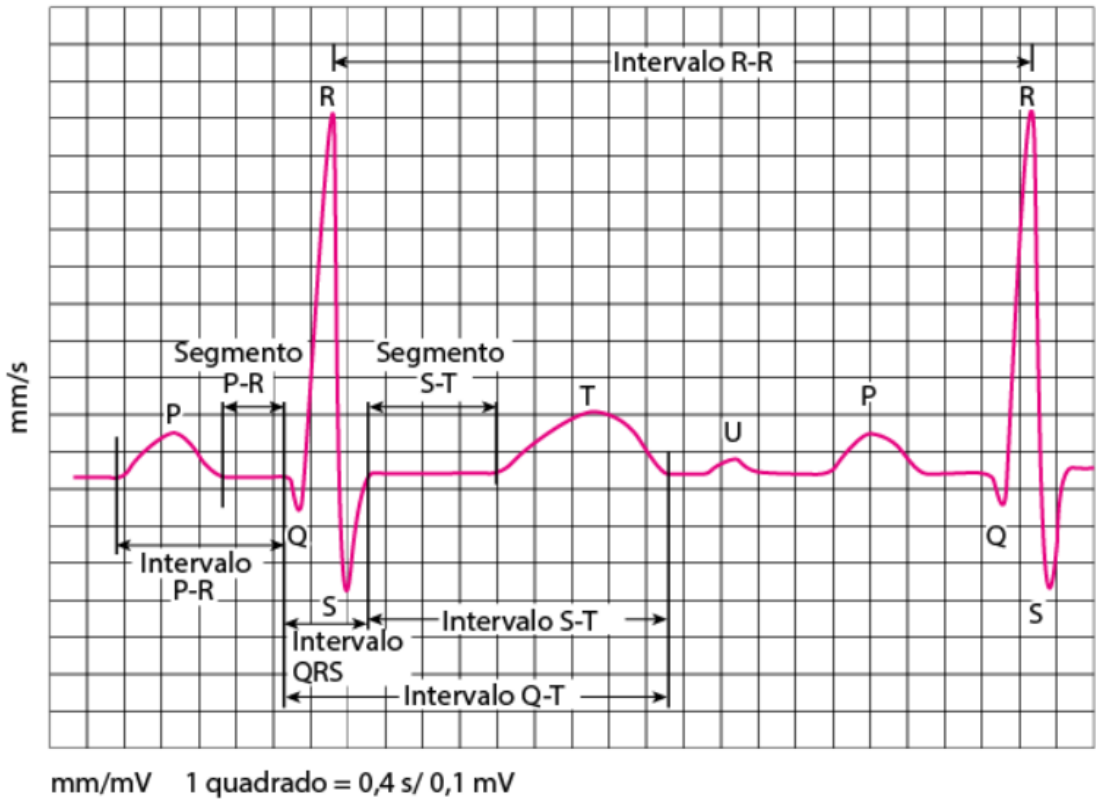
O ritmo secundário à atividade elétrica organizada forma um macrocircuito reentrante, que se propaga ao longo das paredes do átrio direito. O circuito pode apresentar duas direções de ativação: a)Sentido anti-horário: é a forma mais comum (90% casos), em que a frente de onda desce pela parede anterior e lateral e sobe pela parede posterior e septal do AD, com freqüência entre 240 e 340 bpm. O ECG apresenta um padrão característico de ondas “F”, com aspecto de dentes de serrote, negativas. [18] [19]

Como pode-se perceber, o ECG é composto por ondas e segmentos. Tais segmentos são compostos pelo intervalo RR, onda P, onda T, complexo QRS. A onda P é denominada pela despolarização dos átrios, onde deve preceder um QRS no ECG normal, com características arredondada e monofásica. Além disso, é positiva na maioria das derivações, com exceção de aVR. Pode ser bifásica na derivações II e V1; o componente inicial representa a atividade atrial direita e o 2º componente representa a atividade atrial esquerda. Normalmente, o eixo da onda P situa-se entre 0° e 75°. Já o intervalo RR, bastante importante em trabalhos de extração de características, é o intervalo de tempo entre dois complexos QRS. [18]

O complexo QRS representa a despolarização ventricular. A onda Q é a deflexão negativa inicial e tem duração menor que 0,05 segundos em todas as derivações, com exceção de V1–3, em que qualquer onda Q é considerada anormal, indicando infarto antigo ou atual. A onda R é a primeira deflexão positiva e os critérios normais de amplitude e

duração não são absolutos, mas ondas R mais amplas podem ser causadas por sobrecarga ventricular. Uma 2ª deflexão positiva do complexo QRS é designada R'. A onda S é a 2ª deflexão negativa; se houver onda Q ou se não houver onda Q é a primeira deflexão negativa. O complexo QRS pode ter onda R isolada, QS (sem R), QR (sem S), RS (sem Q) ou RSR', dependendo da derivação eletrocardiográfica, vetor e existência de cardiopatias. [19]

Figura 3: Ondas



Fonte: Sanarmed, 2020.

2.3 Ritmo Sinusal Normal

O ritmo sinusal é o ritmo considerado normal do coração. Assim, relembrando a eletrofisiologia cardíaca, o estímulo inicial deve ser realizado no nó sinusal. Então podemos definir, em poucas palavras, que o ritmo sinusal é aquele em que os estímulos elétricos estão sendo corretamente gerados pelo nó sinusal [44**]**. Nesse sentido, um desequilíbrio nessa região acarreta um comprometimento do ritmo normal do coração.

É o ritmo fisiológico do coração, que se origina no átrio direito alto, e, por isso, é visualizado no ECG de superfície pela presença e ondas P positivas nas derivações inferiores, com orientação vetorial média de 60 graus, sendo monofásico em DII, com duração inferior a 110 ms e amplitude máxima de 3 mm. Podem ocorrer modificações em sua morfologia em função da freqüência cardíaca [44.

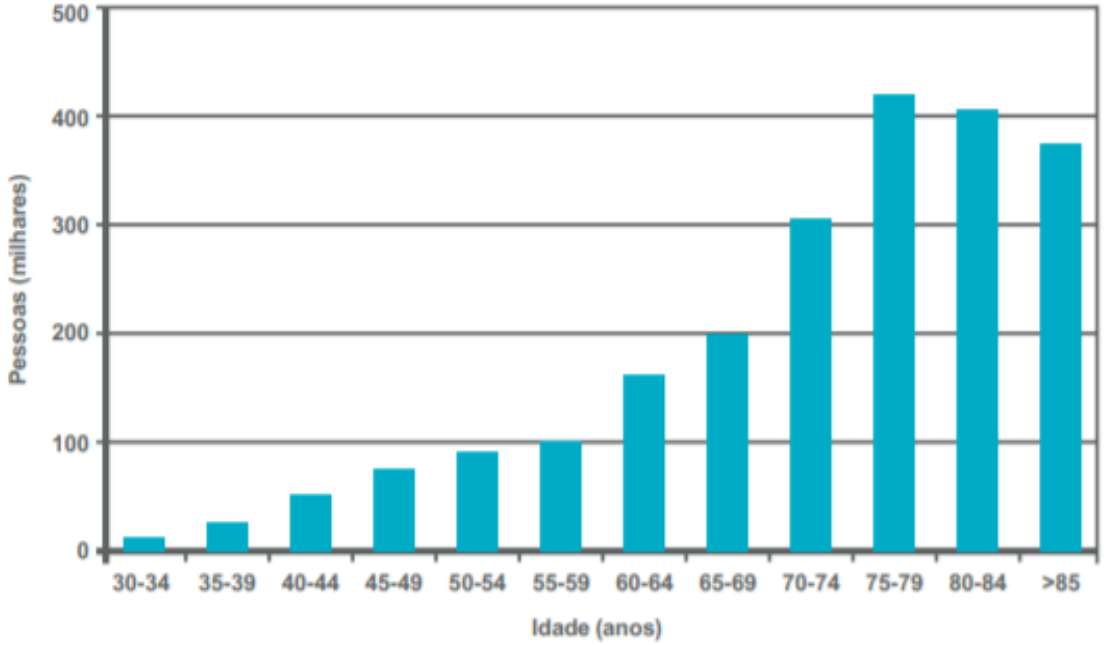
2.4 Fibrilação Atrial

Segundo o Hospital Israelita Albert Einstein (2020), [20] a fibrilação atrial (FA) é a arritmia cardíaca sustentada mais frequente e é responsável por 33% de todas as internações por arritmia. Ocorre entre 1% e 2% na população geral, aumentando significativamente com o envelhecimento e com a presença de doenças cardíacas. O hospital também define que a FA é uma arritmia cardíaca caracterizada pela desorganização completa da atividade elétrica dos átrios (câmaras superiores do coração) e consequente perda da contração atrial. Neto et al. (2018) [21] também definem a FA como sendo uma arritmia supraventricular caracterizada por atividade elétrica atrial desorganizada, secundária a múltiplos focos de despolarização atrial. Dentre as manifestações clínicas da FA, destacam-se a gravidade do tromboembolismo e a instabilidade hemodinâmica.

Fibrilação atrial (FA) como sendo um supraventricular arritmia caracterizada por atrial desorganizado atividade elétrica, secundária a múltiplos focos de atrial despolarização [21]. Apesar dos avanços recentes no tratamento da FA, pacientes com esta doença cardíaca ainda têm alta mortalidade. Isso ocorre porque existem outras maneiras que AF. Tem sido amplamente utilizado devido à natureza de seu observações, como mostrado: o padrão elétrico complexo observada durante a FA pode ser explicada com vários ondas que se propagam ao longo de várias rotas ao longo do átrios; os dados disponíveis também suportam um foco mecanismo, de acordo com o qual condutores, localizados principalmente nas veias pulmonares, gatilho e suporte a propagação da atividade elétrica nos átrios (Richter et al., 2010) [22]

A Fibrilação Atrial (FA) já é um problema de saúde pública, e representa tanto em instituições públicas como privadas, importante causa de internação. Nos Estados Unidos, sua prevalência será de 15,9 milhões em 2050, sendo metade desses pacientes com idade superior a 80 anos. É a arritmia cardíaca mais frequente, com prevalência na população geral estimada em 1%. Dados de estudos europeus mais recentes sugerem uma prevalência de até 2,9% [23], esses números ainda estão subestimados, uma vez que muitos casos (10 a 25%) não provocam sintomas. A figura 2.4 demonstra a relação da idade com a FA. Nos pacientes com menos de 60 anos quando prevalência é inferior a 0,1%, e nos idosos com mais de 80 anos que passa a ser de 8% [23]. Além do envelhecimento populacional, a prevalência cada vez maior de doenças cardiovasculares (DCV), a ampliação dos métodos diagnósticos, e maior atenção da comunidade médico-científica dedicados à FA podem explicar tal crescimento [23].

Figura 5: Progressão de casos de FA por idade no mundo.



Fonte: Autor, 2021.

2.5 Fibrilação Atrial intracardíaca

Para Richter [22], o registro dos potenciais cardíacos dos eletrodos em contato direto com o coração é denominado eletrograma intracardíaco (EGM). Os EGMs intracardíacos, portanto, registram a atividade elétrica local do coração, ou seja, o tecido cardíaco ao redor do eletrodo em contato. Isso pode auxiliar no propósito de guiar o cateter de ablação até os locais atriais de origem da arritmia ou que representam substratos da arritmia.

Richter [22] apontam que a abordagem intracardíaca tem sido amplamente utilizada, devido às seguintes observações: o complexo padrão elétrico observado durante a FA é explicado por várias ondas que se propagam por várias rotas ao longo dos átrios, além do os dados disponíveis também suportam um mecanismo focal, segundo o qual condutores, localizados principalmente nas veias pulmonares, desencadeiam e sustentam a propagação da atividade elétrica nos átrios.

2.6 Fibrilação Atrial Paroxística

Lip e Hee [24] argumentam que a FA também pode ocorrer de forma intermitente, e a importância da FA paroxística (PAF) ganhou destaque recentemente. Um erro comum no manejo clínico é tratar a FA e a FA paroxística de maneira semelhante, apesar de algumas diferenças nos objetivos do manejo.

O PAF pode estar associado a riscos de acidente vascular cerebral e tromboembolismo semelhantes aos da FA, e muitos pacientes sofrem morbidade significativa. Avanços recentes nas áreas de eletrofisiologia e fisiopatologia da FA também reacenderam muito interesse no

PAF. Define-se "fibrilação atrial paroxística" aquela que é revertida espontaneamente ou com intervenção médica em até 7 dias de seu início [24].

Episódios com duração superior a 7 dias têm o nome de "fibrilação atrial persistente". Alguns estudos utilizam a terminologia de "fibrilação atrial persistente de longa duração" para designar os casos com duração superior a 1 ano. Finalmente, o termo "fibrilação atrial permanente" é utilizado nos casos em que as tentativas de reversão ao ritmo sinusal não serão mais instituídas [24].

2.7 Aprendizado de Máquina

Para Russel [45]., aprendizado de máquina é um campo que está atraindo muito atualmente, sendo responsável por muitos novos avanços, o termo foi cunhado por Arthur Samuel em 1959 a quem foi atribuída a criação do primeiro programa mundial de autoaprendizagem. O programa que ele desenvolveu jogou damas e usou uma árvore de busca do tabuleiro para determinar os possíveis movimentos baseados no estado do tabuleiro. Deng [46]. define o aprendizado de máquina como o campo da inteligência artificial cujo interesse é a construção de programas de computadores que se aperfeiçoam automaticamente com a experiência. O aprendizado de máquina tem sido usado com sucesso em quase todas as áreas do conhecimento que utilizam computadores, como classificação e reconhecimento de padrões, controle, jogos, entre outros. desenvolveu jogou damas e usou uma árvore de busca do tabuleiro para determinar os possíveis movimentos baseados no estado do tabuleiro.

De acordo com Bianchi [47]., o aprendizado de máquina pode ser classificado pela maneira na qual o agente interage com o ambiente em que atua para construir seu conhecimento em três classes: supervisionado, não supervisionado e por reforço. Segundo Haykin, no aprendizado supervisionado tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: entrada, saída desejada. Segundo Silva [48]., o algoritmo de aprendizado de máquina extrai a representação do conhecimento a partir desses exemplos.

O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente. No aprendizado não-supervisionado não há a presença de um professor, ou seja, não existem exemplos rotulados. O algoritmo de aprendizado de máquina aprende a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade[49] [50]. Essas técnicas são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados. Este trabalho tem seu foco no aprendizado supervisionado, visando a classificação entre indivíduos saudáveis e no ritmo sinusal normal.

3 MOMENTOS ESTATÍSTICOS E ESTATÍSTICAS DE ALTA ORDEM

Houve um aumento do interesse pelo EOS e suas aplicações. Conforme aponta Boreli (2018) [25], verificou-se a aplicação de cumulantes em diversas áreas do conhecimento, como sonar, biomedicina, processamento de dados, reconstrução de imagens etc. Essas estatísticas fornecem mais informações do que as simplesmente fornecidas pela média e variância de um processo. Assim, pode-se dizer que permitem uma melhor forma de discriminar os processos. Assim, para melhor compreender e iniciar uma abordagem para além da variância e média dos conjuntos, neste artigo utiliza-se a curtose, a assimetria, que são definidas na secção 3 .

3.1 Variância

A variância de uma variável aleatória X é definida como o momento central de segunda ordem, de modo que [26]:

$$\sigma^2 = \mu_2. \tag{1}$$

Partindo disso é possível estabelecer a distância de cada valor do conjunto em relação ao valor médio [25] [26], verificando a dispersão do conjunto de dados.

3.2 Assimetria

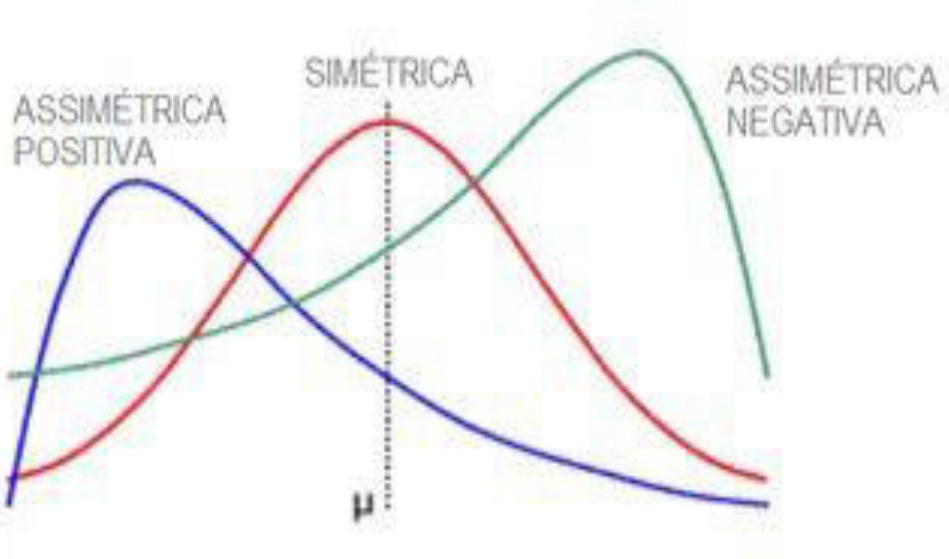
A assimetria, mais conhecida como skewness ou coeficiente de assimetria, define o grau de assimetria de uma distribuição de probabilidades. Ou seja, trata-se do afastamento de uma distribuição da unidade de simetria [27] .

O coeficiente de assimetria pode ainda ser classificado levando em consideração os valores de média (μ), mediana($\tilde{x}$) e moda(m_0), dadas da seguinte maneira para $\mu = \tilde{x} = m_0$ tem-se a classificação de simétrica; para para $m_0 < \tilde{x} < \mu$, tem-se como assimétrica positiva e assimétrica negativa para $\mu < \tilde{x} < m_0$.

A assimetria é a medida em que os dados não são simétricos. Valores de assimetria iguais a zero, positivos ou negativos revelam informações sobre a forma dos dados. Por exemplo, conforme os dados tornam-se simétricos, seu valor de assimetria aproxima-se de zero. Uma distribuição normal, por definição, tem assimetria relativamente pequena. Ao traçar uma linha abaixo do meio deste histograma de dados normais pode-se constatar que os dois lados refletem um ao outro. Mas a falta de assimetria simplesmente não significa normalidade.

Concomitante a isto, vale a pena ressaltar as seguintes observações: Uma distribuição simétrica não necessariamente será gaussiana ou assumirá um formato semelhante ao de um sino; Além disso, dados reais muitas das vezes apresentam valores extremos em uma das causas, facilitando assim a olhada de outliers; ainda em relação a dados reais, pode-se inferir que os mesmos podem ter distribuições bimodais e multimodais. Na Figura 6 é possível ver a forma das distribuições em cada situação.

Figura 6: Classificação de função de distribuição a partir da assimetria.



Fonte: Adaptada pelo autor

3.3 Curtose

A curtose, também chamada de coeficiente de curtose, trata-se da medida de in- tensidade dos picos de uma distribuição de probabilidades. A mesma é definida por:

$$\kappa_2 = \frac{\mu_4}{\mu^2} - 3 \qquad (2)$$

Em outras palavras a partir do coeficiente de curtose é possível indicar a concentração dos valores da distribuição em torno do centro desta. Tento obtido esse valor a função de distribuição de probabilidades pode ser classificador em:

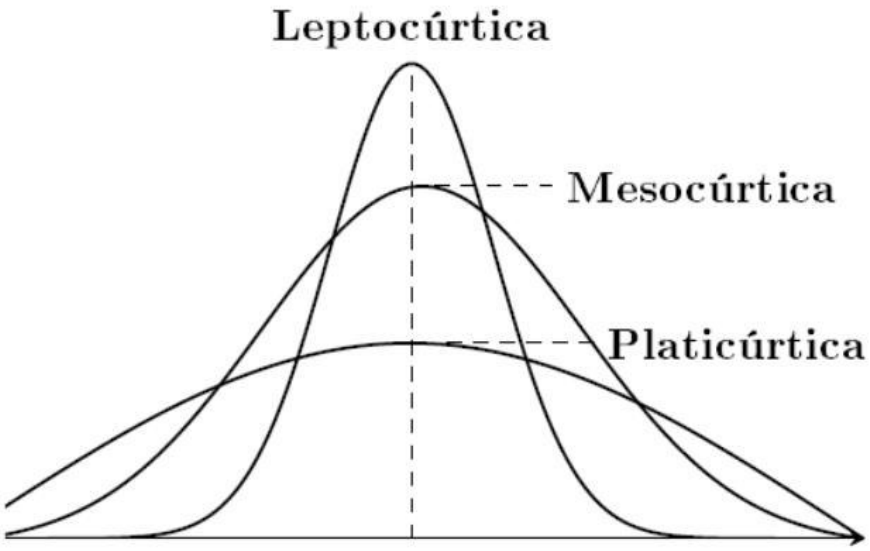
- *Mesocúrtica*: para $\alpha_4 = 0$, a função de distribuição tem o mesmo achatamento da distribuição normal.
- *Leptocúrtica*: para $\alpha_4 > 0$, a função de distribuição possui a curva da função de distribuição mais afunilada com um pico mais alto do que a distribuição normal.
- *Platicúrtica*: para $\alpha_4 < 0$, a função de distribuição é mais achatada do que a distribuição normal.

Os dados que seguem uma distribuição normal perfeitamente têm um valor de 0. Normalmente, os dados distribuídos estabelecem a linha de base para curtose. A curtose da amostra que se desvia significativamente de 0 pode indicar que os dados não estão normalmente distribuídos. Uma distribuição com um valor de curtose positiva indica que a distribuição tem caudas mais pesadas do que a distribuição normal. Por exemplo, os dados que se seguem a distribuição T tem um valor de curtose positiva. Uma distribuição com um valor de curtose negativa indica que a distribuição tem caudas mais leves do que a distribuição normal. Por exemplo, os

dados que seguem uma distribuição beta com primeiro e segundo parâmetros de forma igual a 2 têm um valor de curtose negativo. A linha contínua mostra a distribuição normal e a linha pontilhada mostra uma distribuição com um valor de curtose negativa.

A Figura 7 ilustra a classificação da função de distribuição com base nos valores de curtose.

Figura 7: Classificação de função de distribuição a partir da curtose.



Fonte: Medium: curtose

Esse momento estatístico será de extrema importância para a realização deste trabalho, pois a curtose é bastante aplicável em sinais esparsos como o ECG.

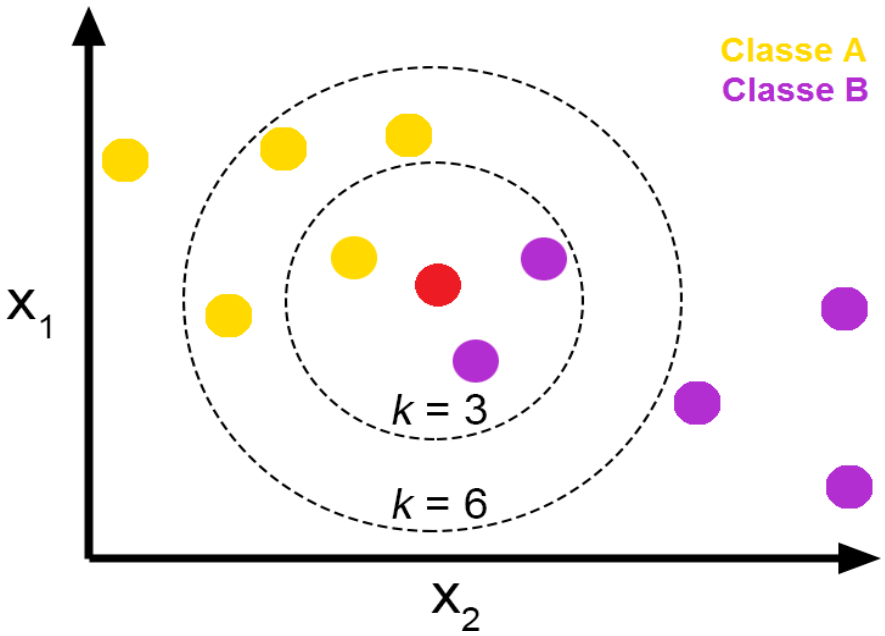
4 CLASSIFICADORES E TÉCNICAS UTILIZADAS

4.1 k-Vizinhos Próximos - k-NN

O k-vizinho mais próximo (k-NN, do inglês k-Nearest Neighbor) está entre os mais populares e mais simples algoritmos de aprendizado de máquina. Em síntese, ocorre a memorização do conjunto de treinamento e ao ser inserido um novo dado, a este será atribuído o rótulo de seus vizinhos mais próximos [28]. Inicialmente a proximidade de um novo dado em relação aos seus vizinhos é definida pela distância euclidiana de seus vetores de atributos [28]

KNN é um dos algoritmos de classificação estatística prospectiva usados para classificar objetos com base em exemplos de treinamento mais próximos. Segundo os autores, trata-se de um algoritmo lento, devido ao modelo ou aprendizado real não estar sendo realizado durante a fase de treinamento. Neste caso, este conjunto é usado apenas para preencher uma amostra do espaço com instâncias cuja classe é conhecida. Nesta fase, os rótulos de vetor e classe das amostras de treinamento constantes definidas pelo usuário, uma consulta ou ponto de teste (vetor não rotulado) e os dados são classificados atribuindo um rótulo, que é o mais recorrente entre os K exemplos de cursos de treinamento mais próximos àquele ponto consultado (Noi e Kappas, 2018) [29]. Neste artigo, KNN com configurações de K iguais a 10 foi usado para a classificação de doenças cardíacas

Figura 8: Ilustração de classificação com o k-NN.



Fonte: Noi e Kappas, 2018.

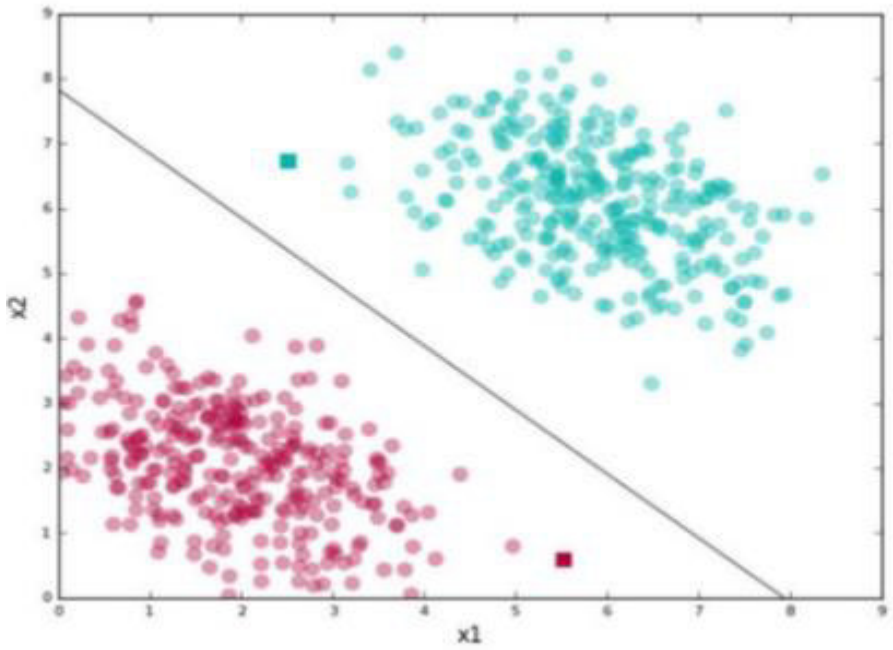
Tendo em vista o cálculo da distância euclidiana, ao novo dado será atribuído o rótulo da Classe B se k = 3, porém tendo k = 6, o novo dado pertencerá a Classe A. De acordo como foi

supracitado, o k-NN trata-se de um algoritmo de fácil implementação, podendo ser aplicável em problemas complexos de classificação e naturalmente incre- mentável, visto que basta apenas inserir novos exemplos de treinamento na memória [28]. Um aspecto negativo desse classificador é que seu método de aprendizado é preguiçoso, tendo um custo computacional grande. Além disso, não é aplicada uma discriminação de dados mas há apenas a memorização dos dados de treino.

4.2 Máquina De Vetores De Suporte - SVM

O objetivo de classificadores binários e lineares é estabelecer um limite satisfatório entre duas classes. Porém tais limites podem não ser ótimos. O ideal é ser tais como a Figura 9

Figura 9: Limites de classificação otimizados.



Fonte: Autor

O SVM como um método poderoso para construir um classificador. Esse método visa criar um limite de decisão entre duas classes que possibilitam prever os rótulos de um ou mais vetores de recursos. Para os autores, este fronteira de decisão, conhecida como hiperplano, é orientada de modo que esteja o mais longe possível dos dados mais próximos pontos para cada uma das aulas presentes. Tão perto pontos são chamados de vetores de suporte, dando origem ao nome do método (Huang et al., 2017) [30]. Desta forma, o hiperplano ideal pode ser definido como aquele que separa os dados e maximiza a margem, respeitando o seguinte equações (Huang et al., 2017).Temos **X** como um conjunto de dados com *n* objetos $\mathbf{x}_i$, seus rótulos i_i e *X* como espaço de entrada e $Y = -1, +1$. Partindo do pressuposto que **X** é linearmente separável sendo possível separar as classes -1 e $+1$ por um hiperplano.

A equação do hiperplano é definida por

$$h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b, \tag{3}$$

dividir o espaço *X* em regiões $\mathbf{w} \cdot \mathbf{x} + b > 0$ e $\mathbf{w} \cdot \mathbf{x} + b < 0$.

isso é definida uma função sinal $g(\mathbf{x})$ que será a saída da função *sigmoid*, tendo como parâmetro $h(\mathbf{x})$, tal como a equação a seguir

$$g(\mathbf{x}) = sgn(h(\mathbf{x})) \quad \begin{cases} +1 & \text{se } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ -1, & \text{se } \mathbf{w} \cdot \mathbf{x} + b < 0 \end{cases} \tag{4}$$

Define-se então, o hiperplano canônico em relação ao conjunto X, de forma que exemplos próximos ao hiperplano $\mathbf{w} \cdot \mathbf{x} + b = 0$ satisfaçam a equação

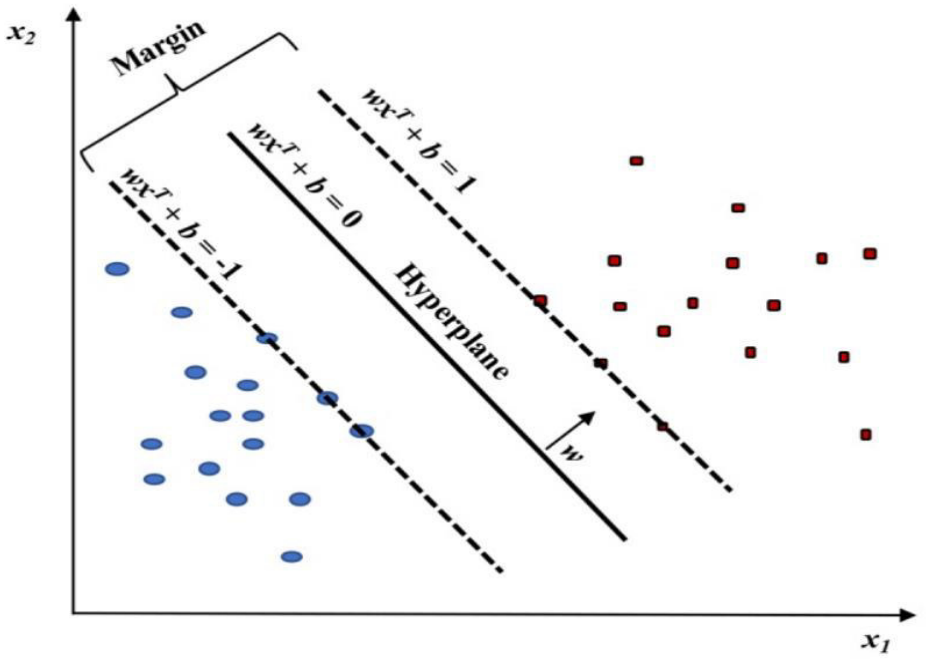
$$|\mathbf{x} \cdot \mathbf{x}_i + b| = 1. \tag{5}$$

As condições são resumidas pela seguinte expressão:

$$y_i(\mathbf{x} \cdot \mathbf{x}_i + b) - 1 \geq 0, \; \forall(\mathbf{x}_i, y_i) \in X \tag{6}$$

De forma que é possível a separação de classes por um hiperplano otimizado, estabelecendo a maior distância entre eles, tala como a Figura 10 [2].

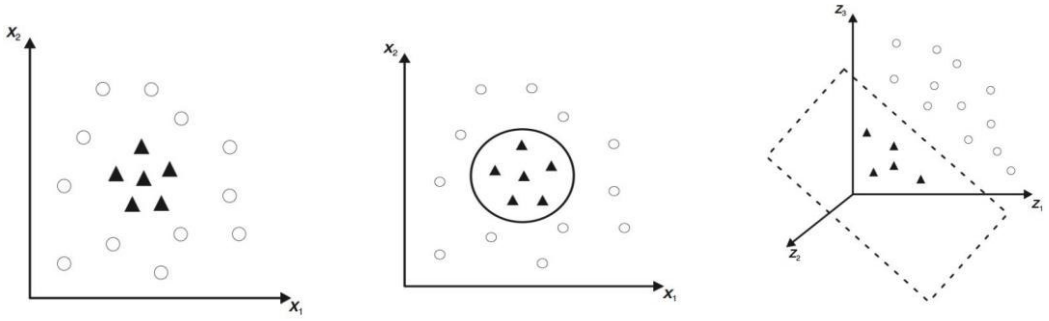
Figura 10: Ilustração de hiperplano de classificação do SVM.



Fonte: Huang, 2017.

Há casos nos quais não é possível dividir os dados satisfatoriamente por um hi- perplano de treino. Essa situação é exemplificada pela Figura 11, em que o uso de uma curva é mais adequado para fazer a separação de classes.

Figura 11: Transformação realizada em conjunto de dados não linear.



Fonte: Huang, 2017

É necessário então realizar o mapeamento do conjunto de treinamento em seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão, este denominado de espaço da características [30]. Esse procedimento é chamado de teorema de Cover [31], definido por:

$$\Phi : X \longrightarrow S \tag{7}$$

No qual X corresponde ao espaço de entradas e S ao espaço de características. por uma escolha adequado de Φ, o teorema afirma que o espaço X pode ser transfor- mado em um espaço de características S no qual há alta probabilidade dos objetos serem linearmente separáveis. De modo que, precisão serem satisfeitas duas duas condições: a transformação precisa ser não linear; a dimensão do espaço de características seja suficientemente alta.

Sabendo que S pode ter alta dimensão, ou mesmo infinita, o cálculo de Φ pode ser extremamente custoso. Para facilitar esse processo são utilizadas as funções kernel K, as quais recebem dois ponto $\mathbf{x}_i$ e $\mathbf{x}_j$ no espaço de entradas e obtém-se o produto escalar desse objetos no espaço de características [31], tendo então:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j). \tag{8}$$

4.3 Redes Neurais Artificiais - RNA

Haykin (2001) [32] define RNAs, mais conhecidas como redes neurais, como estruturas complexas interconectadas por elementos de processamento simples (neurônios), que têm a capacidade de realizar operações, como cálculos paralelos, para processamento de dados e representação de conhecimento (HAYKIN, 2001, p. .27).

O autor também enfatiza que as propriedades e capacidades que tornam os RNAs potencialmente úteis são: não linearidade: um neurônio artificial pode usar funções lineares ou não lineares; Mapeamento de entrada-saída: com base em exemplos de entrada e saída, a ANN é capaz de se adaptar para minimizar o erro de mapeamento. Dentre as estruturas conhecidas desses modelos, temos o MLP (Multilayer Perceptron), que, em geral, possui uma camada de entrada (sem função computacional), uma ou mais camadas ocultas e uma camada de saída.

O Perceptron trata-se da arquitetura mais simplista de redes neurais. Tendo sua simplicidade observada pela sua constituição contendo apenas uma camada neural, havendo somente um único neurônio artificial na camada [32]. A Figura 4.6 ilustra a arqui- tetura de uma rede neural Perceptron.

Para o seu treino usa-se uma algoritmo supervisionado de correção de erro e uma função de ativação. Na qual apar um objeto $\mathbf{x}_i$ os pesos são ajustados de acordo com:

$$w_j(t+1) = w_j(t) + \eta x^j(y_j - \hat{f}(\mathbf{x}_i)) \tag{9}$$

Onde $w_j(t)$ trata-se do peso da j-ésima conexão de entrada no instante de tempo e t, η corresponde a taxa de aprendizagem, x e o valor do j-ésimo atributo do vetor de entrada $\mathbf{x}_i$, $\mathbf{x}_i$) refere-se a saída da rede neural no instante t e por fim, y_i é a saída desejada da rede (o rótulo de $\mathbf{x}_i$) [32].

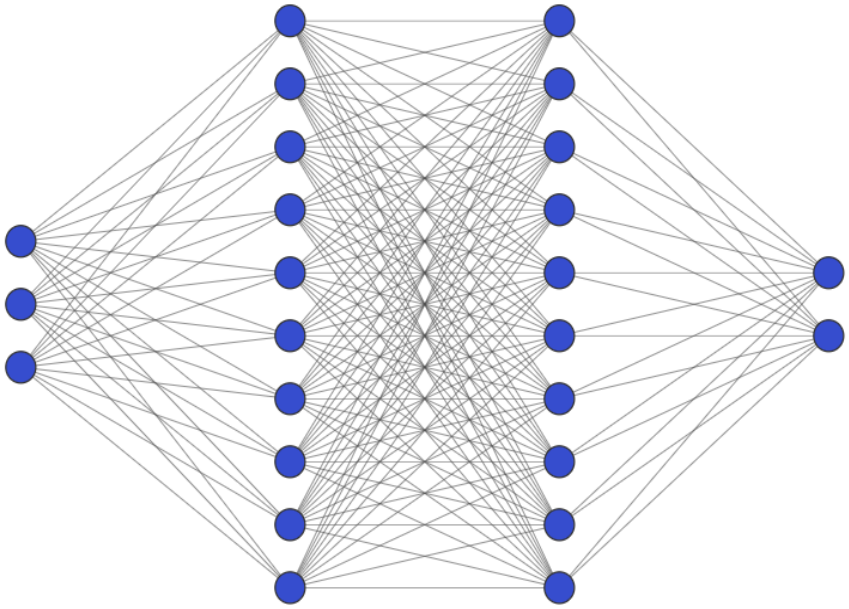
Com relação a função de ativação podemos destacar as de uso mais comum:

- Identidade: $f(z) = z$;
- Log´ıstica: $f(z) = \frac{1}{1+e^{-z}}$;
- Tangente Hiperbólica: $f(z) = \frac{e^z - e^{-z}}{z \quad -\tilde{e}\epsilon}$;
- ReLu: $f(z) = max(0, z)$;

A partir de 9 e sabendo apenas os pesos para neurônios da camada de saída, é necessária a estimação de pesos de camadas intermediárias. Para isso usa-se o algoritmo *back-propagation*, que estima o erro dos neurônios de determinada camada observando o

erro dos neurônios de uma camada posterior [32] [33]. De forma que o erro da camada em que se encontra o neurônio é calculado a seguir:

Figura 12: Arquitetura de Rede Neural MLP



Fonte: Autor

4.4 Análise Dos Componentes Independentes - ICA

A maioria das quantidades medidas são, na verdade, misturas de outras quantidades. Exemplos típicos são (a) o som em uma sala em que várias pessoas estão falando simultaneamente, (b) um sinal de eletrocardiograma (ECG), que contém contribuições de muitas regiões diferentes do coração, e (c) a altura de uma pessoa, que é determinada por contribuições de muitos genes diferentes e fatores ambientais. A ciência está, em grande medida, preocupada em estabelecer o natureza precisa dos processos de componentes responsáveis por um determinado conjunto de quantidades medidas, quer envolvam sinais de ECG, altura humana ou mesmo QI. Sob certas condições, os sinais subjacentes às quantidades medidas podem ser recuperados fazendo uso de análise de componentes (ICA), que é membro de uma classe de separação de fonte cega (BSS) métodos [34].

Matematicamente descrevendo, temos o vetor aleatório $[X_1, X_2, X_3, \dots X_n]^T$, cujos n elementos são gerados pela mistura de n componentes estatisticamente independentes entre si de um vetor aleatório $[S_1, S_2, S_3, \dots S_n]^T$, . O modelo ICA expressa cada X_i como uma combinação linear de componentes independentes, dada por

$$X_i = a_1 S_1 + a_2 S_2 + a_3 S_3 + \cdots + a_n S_n \quad (10)$$

para todo $i = 1, 2, 3 \dots n$.

Abordando matricialmente, pode-se escrever da seguinte maneira:

$$X = A.S \tag{11}$$

Em que A é a matriz dos coeficientes = $a_i$das combinações lineares, e Sendo $a_i$um coeficiente que pondera a mistura dos componentes independentes (sinais ou fontes originais), a matriz A é denominada matriz de mistura. Tanto os coeficientes = a_i como os componentes independentes = S_i são desconhecidos e devem ser estimados a partir da observação dos sinais misturados= X_i. Este é um modelo generativo, pois descreve como os dados observados são gerados a partir de um processo de mistura dos componentes S_i.

Alternativamente pode-se definir ICA como o problema de determinar uma transformação linear dada pela matriz W,

$$Y = W.X \tag{12}$$

em que *Y* é o vetor aleatório de componentes $[Y_1, Y_2, Y, \dots Y_n]$, que são estimativas dos componentes independentes e *W* é a matriz inversa de *A*, denominada Matriz de Separação, descrita em (11).

O sucesso do ICA depende de uma única suposição altamente plausível em relação à natureza do mundo físico: variáveis independentes ou sinais são gerados por diferentes processos físicos subjacentes. Se dois sinais são independentes, o valor de um sinal não pode ser usado para prever qualquer coisa sobre o outro sinal. Na prática, a maioria dos sinais medidos são derivados de muitos processos físicos independentes e, portanto, são misturas de sinais independentes. Dado esse conjunto de sinais medidos (ou seja, misturas), o ICA funciona encontrando uma transformação dessas misturas, que produz componentes de sinal independentes, na suposição de que cada um desses sinais de componentes independentes está associado a um processo físico diferente [34].

Na linguagem do ICA, os sinais medidos são conhecidos como misturas de sinais e os sinais independentes são conhecidos como sinais de origem. O ICA foi aplicado para separação de diferentes sinais de fala [45] 2, análise de dados de ECG [45], dados de imagem de ressonância magnética funcional (fMRI) [45], processamento de imagem [45] e como um modelo de processamento biológico de imagens [11]. Uma revisão dos avanços recentes em ICA pode ser encontrada em [6]. No entanto, deve-se notar que este exemplo poderia igualmente se aplicar a qualquer conjunto de sinais medidos fisicamente, e para qualquer número de sinais (por exemplo, imagens, dados biomédicos ou preços de ações).

Fundamentalmente, ICA em biomedicina envolve a extração e separação de fontes estatisticamente independentes subjacentes a múltiplas medições de sinais biomédicos. Avanços técnicos em desenvolvimentos algorítmicos implementando ICA são revisados junto com novas direções no campo. Esses avanços são resumidos especificamente com aplicações a sinais biomédicos em mente. As suposições básicas feitas ao aplicar o ICA são discutidas,

junto com suas implicações quando aplicadas particularmente a sinais biomédicos. [45]

Sinais biomédicos de várias fontes, incluindo corações, cérebros e sistemas endócrinos, representam um desafio para os pesquisadores que podem ter que separar os sinais fracos que chegam de fontes múltiplas contaminadas com artefatos e ruído. A análise desses sinais é importante tanto para pesquisa quanto para diagnóstico e tratamento médico. As aplicações da Independent Component Analysis (ICA) para sinais biomédicos é uma área de pesquisa em rápida expansão e muitos grupos estão agora ativamente engajados na exploração do potencial da separação cega de sinais e deconvolução de sinais para revelar novas informações sobre o cérebro e o corpo.

Várias questões importantes na aplicação do ICA aos dados biomédicos podem ser ilustradas pela análise dos sinais elétricos do coração. Os sinais registrados na superfície do tórax e abdômen, decorrentes do coração batendo, são usados pelos médicos para diagnosticar doenças cardíacas. Diferentes partes do coração, como os átrios e os ventrículos, produzem diferentes padrões espaciais e temporais de atividade elétrica na superfície do corpo. As gravações são normalmente feitas em vários locais, cada um refletindo uma mistura diferente de componentes do coração.

Os ECGs parecem satisfazer algumas das condições para ICA: 1) A corrente de diferentes fontes é misturada linearmente nos eletrodos de ECG; 2) Os atrasos na transmissão do sinal são insignificantes; 3) Parece haver menos fontes do que misturas; e 4) As fontes têm distribuições de tensão não gaussianas. A presença de ondas móveis de atividade elétrica através do coração também significa que a atividade de uma única câmara pode ser considerada como fontes múltiplas pela ICA.

4.5 Análise Dos Componentes Principais

Conforme descrito por Mishra et al., (2017) [35], a análise de componentes principais (PCA) é uma técnica multivariada que analisa dados em que as observações são descritas por várias variáveis dependentes quantitativas e correlacionadas. Seu objetivo é extrair informações importantes de dados estatísticos para representá-los como um conjunto de novas variáveis ortogonais chamadas de componentes principais, e mostrar o padrão de similaridade entre observações e variáveis como pontos em mapas de pontos.

A técnica de PCA é explicada por Castells et al., (2006) [36] como sendo uma técnica estatística que visa a condensação de informações de um grande conjunto de variáveis correlacionadas em algumas variáveis ("componentes principais"), mas não desperdiça a variabilidade presente em o conjunto de dados.

Os autores apontam que os componentes principais são derivados como uma combinação linear das variáveis do conjunto de dados, com pesos escolhidos de forma que esses componentes necessariamente se tornem não correlacionados, onde cada componente contém novas informações sobre o conjunto de dados e é ordenado de forma que o primeiro componentes são responsáveis pela maior parte da variabilidade.

Demonstrando a importância desta técnica, incluindo sinais de ECG, Castells et al., (2006) [36] relata que o PCA é usado para lidar com vários problemas na análise de ECG, como compressão de dados, detecção e classificação de batimento, redução de ruído, separação de sinal e extração de recursos.

A análise de componentes principais (PCA) se refere ao processo pelo qual os componentes principais são calculados e o uso subsequente desses componentes na compressão dos dados [7] [37]. O PCA é uma abordagem não supervisionada, pois envolve apenas um conjunto de recursos $x_1, x_2, \cdots, x_p$, e nenhuma resposta associada y.

Em uma abordagem simplista, o processo de aquisicão de componentes principais se inicia com o cálculo da covariância entre no mínimo duas dimensões (X e Y) [25] [38]:

$$cov(X, Y) = \frac{\sum_{i=1}^{n} [(X_i - \bar{X})(Y_i - \bar{Y})]}{n} \quad (13)$$

onde X e Y correspondem a listas de dados, em que X é a primeira e Y a segunda dimensão, $\bar{X}$ e $\bar{Y}$ suas médias, e por fim n o número de dados.

Havendo mais de duas dimensões é necessária a aplicação da covariância em cada par de dimensões, dai o surgindo a matriz de covariância. Onde, por exemplo, tendo três dimensões (x, y e z) . Em seguida, para M amostras de um determinado conjunto de dados, o vetor médio, ainda tendo a matriz de covariância de um certo conjunto de dados com M amostras, temos que o vetor médio.

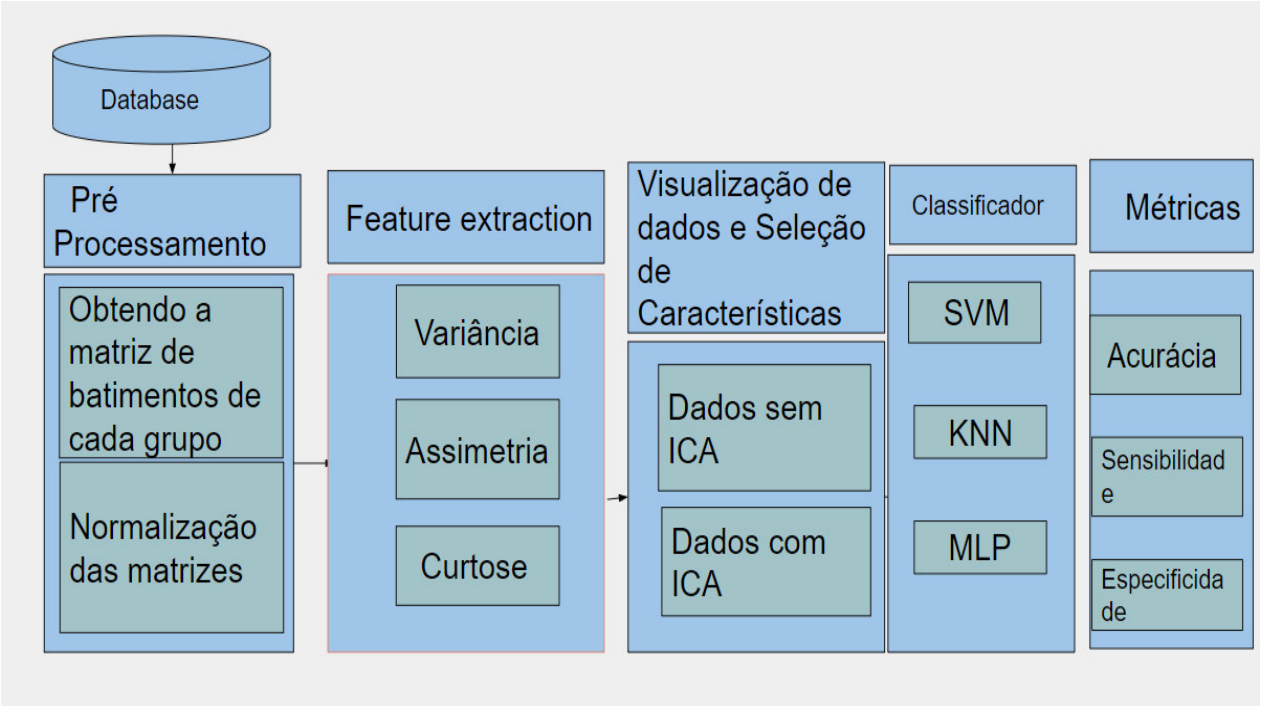
Além de produzir variáveis derivadas para uso em problemas de aprendizagem supervisionada e verificar qual eixo de componente principal tem maior variância do dados, o PCA também serve como ferramenta para visualização de dados, descorrelacionar dos dados, conforme previsto em Haikyn, além de servir de branqueamento para os dados na técnica de ICA, também utilizada nesta dissertação.

5 MATERIAIS E MÉTODOS

Na Figura 13, é ilustrada a metodologia utilizada neste artigo. Os bancos de dados a serem utilizados foram definidos, separando-os em quatro grupos: sinais de indivíduos com FA, indivíduos com FA intracardíaca, Fibrilação Atrial Paroxística e indivíduos com Ritmo Sinusal Normal. Os sinais do banco de dados foram pré-processados, organizando-os para a extração das características. Nesta etapa, são calculados os valores de variância, assimetria e curtose do conjunto de dados de cada base.

Resumindo o método, foi realizado o pré-processamento dos sinais do banco de dados e organização dos dados para a etapa de extração de recursos. Nessa etapa, são calculados os valores de variância, assimetria e curtose dos conjuntos de dados de cada base. Após essa etapa, foram selecionadas combinações de características, que são representadas pelas estatísticas citadas, e colocadas como entrada para os classificadores, dividindo os dados com a validação cruzada, em treino e teste. O ICA é utilizado aqui antes da validação cruzada, em cada fold, a fim de não corromper ou inviesar o processo. Ao final do processo, os valores da métrica de classificação são retornados para avaliação do algoritmo.

Figura 13: Metodologia



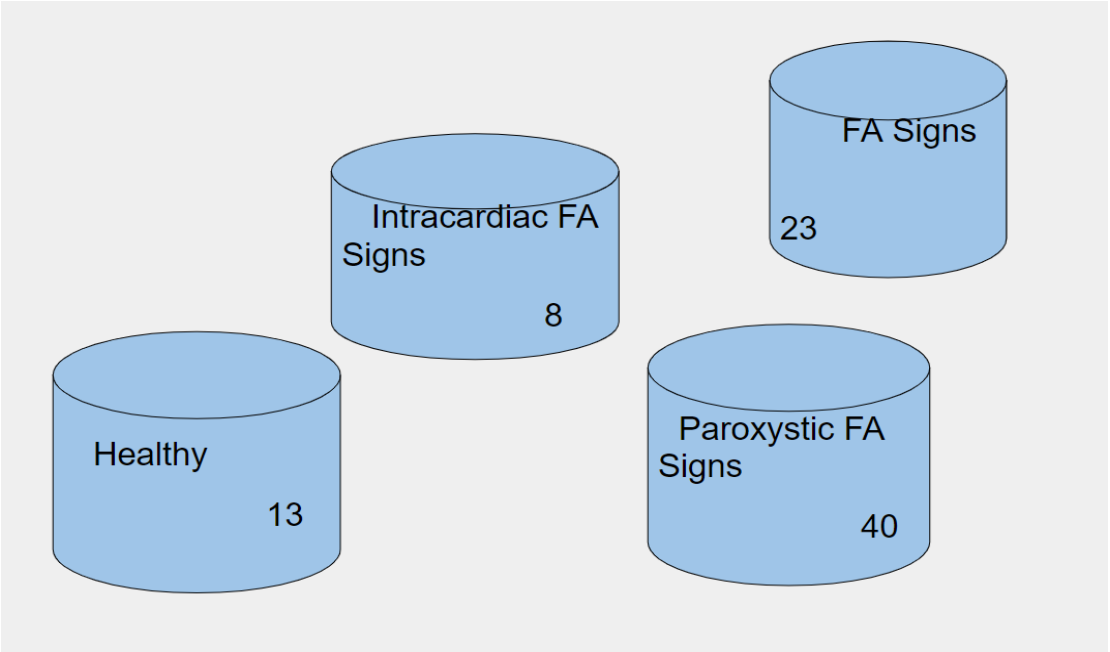
Fonte: Autor.

Em seguida, esses dados selecionados foram colocados como entrada para o classificador usado neste artigo. Ao final do processo, os valores da métrica de classificação são retornados para avaliação do algoritmo. Além disso, os algoritmos foram desenvolvidos na linguagem Python.

5.1 Base de Dados

Foram utilizados os conjuntos de dados Intracardiac Atrial Fibrillation Database, MIT-BIH Atrial Fibrillation Database, Intracardiac Atrial Fibrillation Database e MITBIH Rhythm Sinus Normal data sets, ambos disponíveis em Goldberger [39]. O banco de dados de sinais de pacientes com FA contém 23 prontuários, todos utilizados nesta análise. O banco de dados de sinais de pacientes com FA intracardíaca contém 8 pacientes, todos os quais são usados. O Teste de Fibrilação Atrial Paroxística contém 70 sinais, 40 dos quais são usados. O saudável contém 18, dos quais 13 são usados. A seguir, na Figura 4, as bases de dados são identificadas.

Figura 14: Bases de dados.



Fonte: Autor

5.2 Data Quality

Uma etapa de qualidade de dados dos dados de entrada foi realizada antes e depois do Aplicação ICA e foi notado que não há valores nulos ou outliers no conjunto de dados, além de todos os tipos serem de ponto flutuante, o que não compromete análise de dados.

5.3 Análise Exploratória de Dados

Nesta etapa, verificou-se a fundo os dados. Foi uma etapa fundamental para descobrir mais sobre a distribuição dos ECG. A partir disso, foi possível identificar que por não se tratarem dados normais, pode-se aplicar o ICA, tal como previsto nos pressupostos de as fonte não serem gaussianas.

Pode-se verificar também que alguns valores, tais como a curtose e assimetria, quando dispostos no plano cartesiano, iriam se sobrepor, não possibilitando uma boa separação dos

dados, também como uma classificação satisfatória, como evidenciado na Figura 15, do dataset construído.

5.4 Dataset

A construção do conjunto de dados foi realizada da seguinte forma. Uma coluna foi criada para cada estatística, representando a variância, assimetria e curtose de cada batimento. Também foi criado um rótulo, coluna 4 do conjunto de dados, que representa a classe pertencente ao respectivo batimento. Esta classe tem um valor de 0 para FA, 1 para FA paroxística, 2 para FA intracardíaca e 3 para saudável.

Figura 15: Dataset.

Variance	Skewness	Kurtosis	Class
5.4564e-05	-1.0606	4.0094	0
0.0023	-1.0679	12.1545	1
-4.88630	0.90579	-0.000843	2
1.5743e-14	-0.6562	6.6754	3

Fonte: Autor.

5.5 Pré-processamento

Foram adquiridos os sinais de ECG da derivação DII, a mais utilizada no mundo. Toda a duração do sinal, amostrada na frequência de 256 Hz, foi utilizada para extrair os batimentos de cada paciente para análise e posterior extração das características. Em seguida, cada sinal selecionado foi segmentado para obtenção do respectivo batimento, conforme proposto por Queiroz et al., [40] e Silva et al., [41].

Dessa forma, os batimentos de cada grupo foram agrupados, gerando uma matriz A pela concatenação dos batimentos do grupo com FA, e uma matriz B dos batimentos do grupo com FA intracardíaca, conforme descrito nas equações a seguir.

$$M = [Bn, a \;\; Bn, b \ldots Bn, z] \tag{14}$$

onde n representa o número de batidas e m representa o total de todas as colunas de todas as batidas.

Dessa forma, a média de seu conjunto foi subtraída do sinal, dividindo o resultado pela entropia de Shannon, dada pela Equação (15).

$$Z = M - \frac{\sum_1^N M\frac{1}{N}}{-\sum_1^N plog(\frac{1}{N})} \quad (15)$$

onde p representa a probabilidade associada a cada batida, e Z representa a nova matriz associada com a concatenação das batidas

5.6 Extração de Caracteństicas

A metodologia de extração foi adaptada utilizando estatísticas de alta ordem, propostas por Queiroz et al., [40]. e de Silva et al [41].. Foi obtido um vetor para cada uma das estatísticas associadas: variância, curtose e assimetria, que serão as entradas dos classificadores, representadas por σ_x^2, κ_x e λ_x respectivamente.

$$\sigma_X^2 = E(X^2) - ((E(X))^2 \quad (16)$$

$$\lambda_X = E\left[\left(\left(X - E(X)\right)\sigma^{-1}\right]^3 \quad (17)$$

$$\kappa_X = E\left[\left(\left(X - E(X)\right)\sigma^{-1}\right]^4 \quad (18)$$

5.7 Classificação

A etapa de classificação consiste em duas partes. A primeira utilizando os vetores de caracteŕística dos quatro grupos de indivíduos sem a utilização do ICA. E na segunda etapa ocorreu a classificação das componentes do ICA, oriundas a partir dos vetores de características.

Em ambas as etapas foram utilizados os classificadores já mencionados: k-Vizinhos Próximos (k-NN), Máquina de Vetores de Suporte (SVM), Rede Neural Perceptron de Multiplas Camadas (MLP) e o Radial Basis Functions Os hiper-parâmetros definidos em cada um dos classificadores foram definidos da seguinte forma: o RNA MLP foi usado com as configurações de 3 neurônios na camada de entrada, 2 camadas com 100 neurônios e 2 neurônios na camada de saída. O KNN com configurações de K igual a 10, para a classificação das doenças cardíacas. O SVM com as configurações de Kernel linear, para a classificação de doenças cardíacas.

5.8 Métricas de Avaliação

Neste artigo, foram utilizados os valores de acurácia, sensibilidade e especificidade, descritos pela Equação 16, Equação 17 e Equação 18 a seguir, para verificar o desempenho dos classificadores.

$$ac = \frac{VP + VN}{VP + VN + FN + FP} \times 100. \quad (19)$$

$$sens = \frac{VP}{VP + FN} \times 100. \quad (20)$$

$$esp = \frac{VN}{VN + FP} \times 100. \quad (21)$$

De modo que VP corresponde a quantidade de batimentos rotuladas com verdadeiros positivos para presença de FA, VN ao número de verdadeiros negativos para ausência de FA, FP para registros com FA classificados pelos algoritmos com ausência de FA e FN para classificações positivas para registros sem FA.

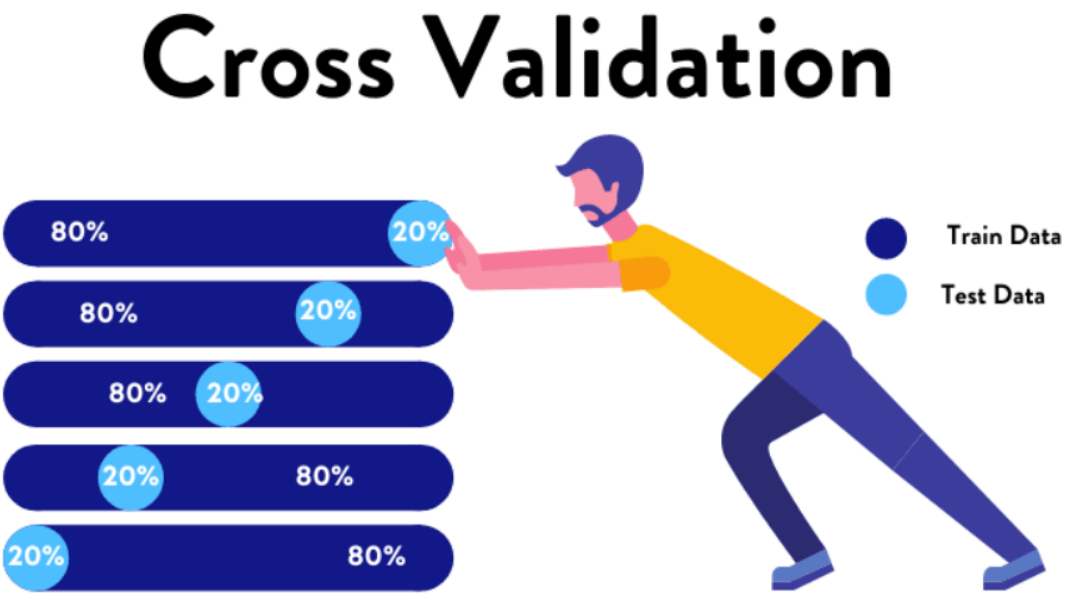
5.9 Validação Cruzada

A validação cruzada é uma técnica para avaliar a generalização de um modelo, com base em um conjunto de dados. No Aprendizado de Máquina, a validação cruzada é um método de re-amostragem usado para avaliação de modelo para evitar o teste de um modelo no mesmo conjunto de dados no qual ele foi treinado. Este é um erro comum, especialmente que um conjunto de dados de teste separado nem sempre está disponível.

No entanto, isso geralmente leva a medidas de desempenho imprecisas (já que o modelo terá uma pontuação quase perfeita, pois está sendo testado nos mesmos dados em que foi treinado). Para evitar esse tipo de erro, a validação cruzada é geralmente preferida. O conceito de validação cruzada é realmente simples: em vez de usar todo o conjunto de dados para treinar e, em seguida, testar nos mesmos dados, poderíamos dividir aleatoriamente nossos dados em conjuntos de dados de treinamento e teste.

Nesta dissertação, os dados são divididos usando o método de validação, que consiste em dividir os dados em 70 e 30 aleatoriamente. 70% dos pacientes foram usados para treinamento, 30% para teste e k-fold igual a 7.

Figura 16: Cross validation



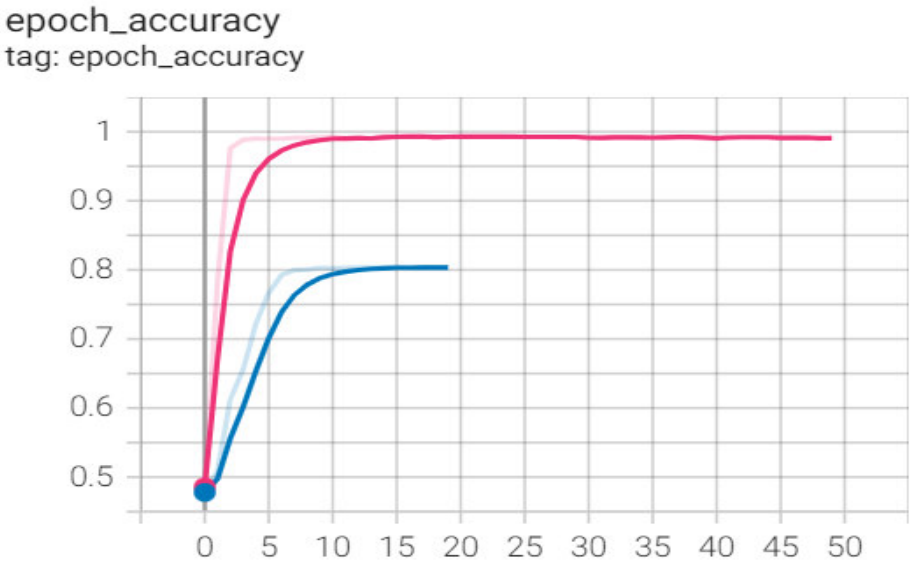
Fonte: Quora, 2021

6 RESULTADOS

6.1 Porcentuais dos avaliadores

Como mencionado anteriormente, foram feitas duas etapas de classificação: classificação dos vetores de característica de ECG, e classificação das componentes oriundas das do ICA a partir dos vetores de caracter´ıstica. Em vermelho podemos ver a média das classificações em termos de acurácia com a utilização do ICA. Em azul, vemos a média das classificações sem a utilização do ICA. Veja nas Figuras 16 e 17.

Figura 16: Iterações x Acurácia



Fonte: Autor.

Figura 17: Iterações x Função de perda

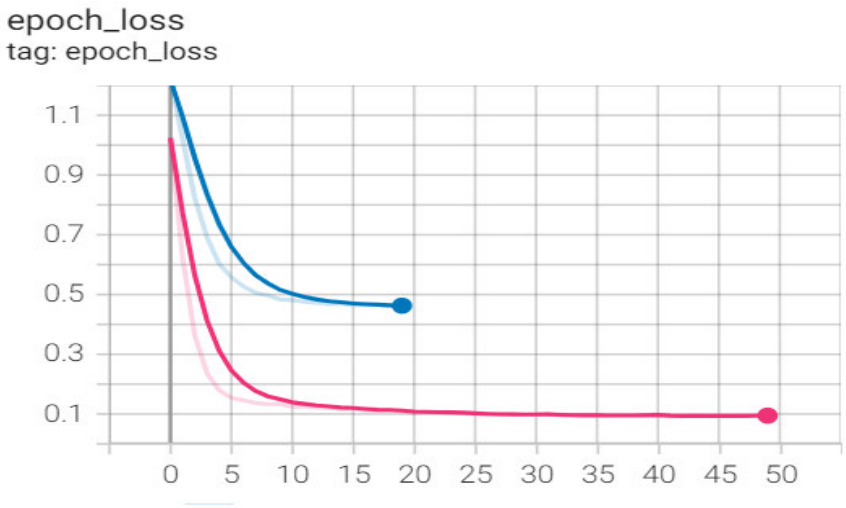


Tabela 2: Métricas de classificação para Ritmo Sinusal e FA.

Métricas	SVM-RBF	k-NN	MLP
Acurácia	100	**100**	100
Sensibilidade	100	99.89	100
Especificidade	100	96.23	100

Tabela 3: Métricas de classificação para Ritmo Sinusal e FA intracardíaca.

Métricas	SVM-RBF	k-NN	MLP
Acurácia	100	**100**	100
Sensibilidade	100	99.56	100
Especificidade	100	96.45	100

Tabela 4: Métricas de classificação para FA intracardíaca e FA sem PCA

Métricas	SVM-RBF	k-NN	MLP
Acurácia	78.56	**68.90**	80.43
Sensibilidade	75.4	64.5	80.2
Especificidade	70.53	63.4	78.54

Tabela 5: Métricas de classificação para FA intracardíaca e FA com PCA

Métricas	SVM-RBF	k-NN	MLP
Acurácia	96.45	92.45	95.3
Sensibilidade	93.76	90	94.2
Especificidade	94	90.4	92.54

Tabela 6: Métricas de classificação para FA paroxística e FA.

Métricas	SVM-RBF	k-NN	MLP
Acurácia	99.4	**96.6**	99.3
Sensibilidade	98.3	92.1	99.5
Especificidade	98.5	91.7	98.4

Tabela 7: Métricas de classificação para Ritmo Sinusal, FA e FA intracardíaca.

Métricas	SVM-RBF	k-NN	MLP
Acurácia	93.4	**93.6**	98.3

Sensibilidade	92.3	90.1	93.5
Especificidade	92.5	93.2	98.1

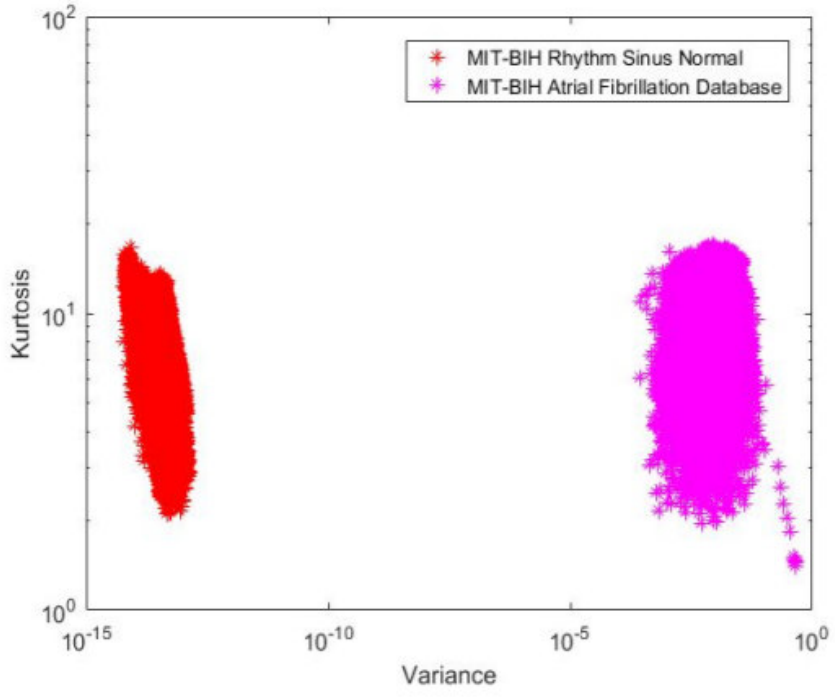
Tabela 8: Métricas de classificação para Ritmo Sinusal, FA, FA intracardíaca e FA paroxística.

Métricas	SVM-RBF	k-NN	MLP
Acurácia	98.4	**96.6**	100
Sensibilidade	98.3	94.0	100
Especificidade	96.5	95.2	99.1

6.2 Classificação dicotômica

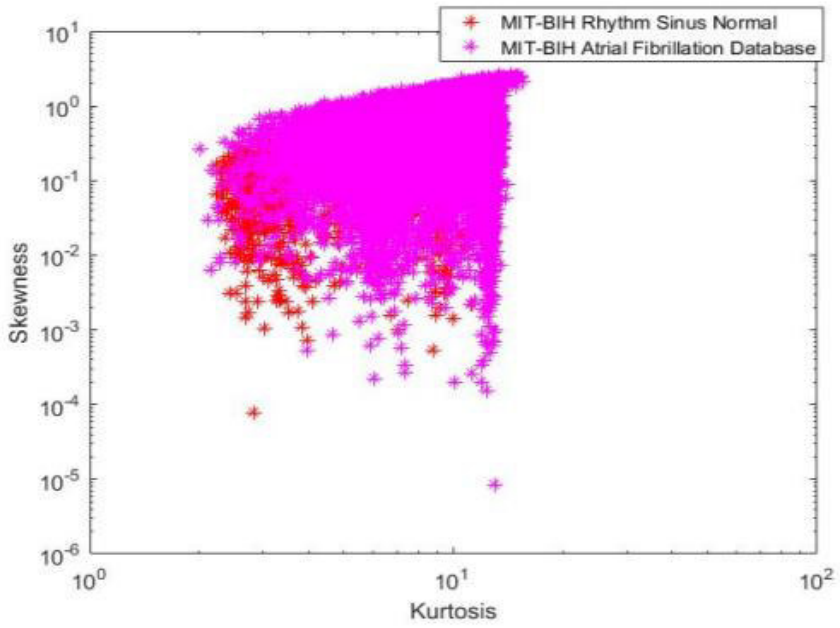
Os resultados aqui presentes são provenientes de classificações dicotômicas entre os grupos e explicam os resultados descritos nas Tabelas 2, 3, 4, 5 e 6. Primeiramente, utilizados os indivíduos saudáveis e com a FA. Foi verificado as características que mais apresentaram acurácia no método proposto, dentre a combinação de variância, curtose e assimetria. Nas Figuras a seguir, tem os resultados obtidos com essas combinações, conforme os resultados na Tabela 2.

Figura 18: Variância x Curtose



Fonte: Autor

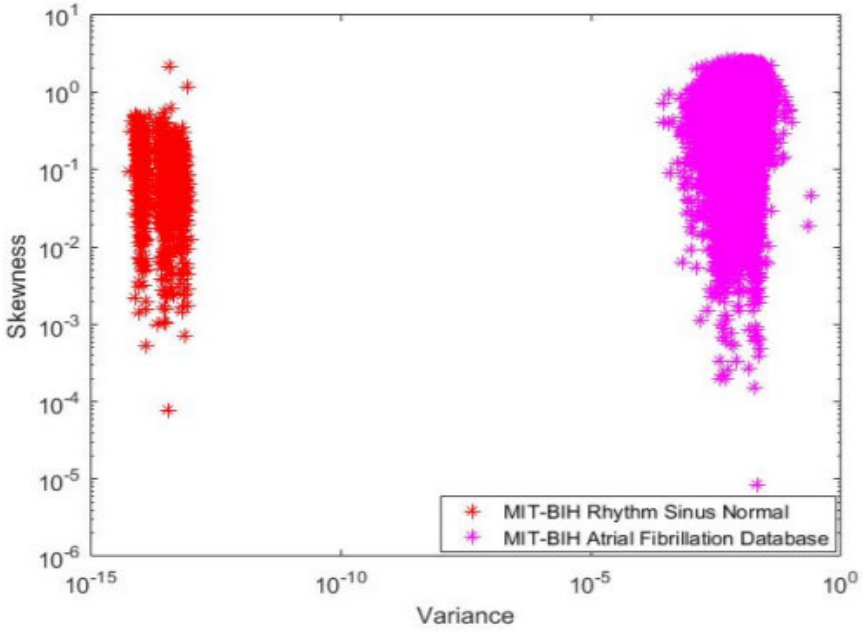
Figura 19: Curtose x Assimetria



Fonte: Autor.

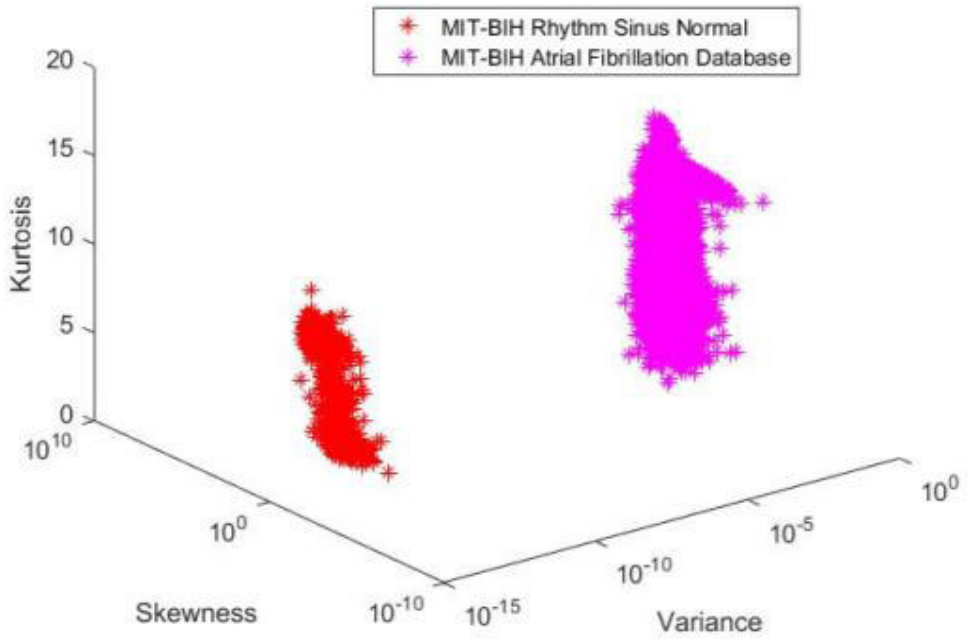
A representação da curtose e assimetria aglomerou bastante os dados, pois os vaores de assimetria são bem parecidos, e no plano cartesiano os dados ficam sobrepostos. Já para a Variância e a assimetria, a classificação foi aproximadamente 98%.

Figura 20: Variância x Assimetria



Fonte: Autor.

Figura 21: Variância x Assimetria x Curtose



Fonte: Autor

Pode-se observar que as combinações que utilizaram variância e curtose em suas representações de características obteve maior separação entre os dois grupos. Isso se deve ao fato de que os indivíduos com atrial fibrilação tem uma maior variância, enquanto indivíduos com ritmo sinusal normal têm menos variância nos dados.

Além disso, a análise em duas dimensões mostra uma separação dos dados usando variância e curtose, conforme mostrado nas Fig. 17, Fig.18 e Fig.19, Fig 20, Fig 21., reafirmando os resultados discutidos em Queiroz et. al. [1] e em Lucena et. al. [7], que apontam que a curtose pode ser uma abordagem apropriada para medir sinais esparsos, como ECG.

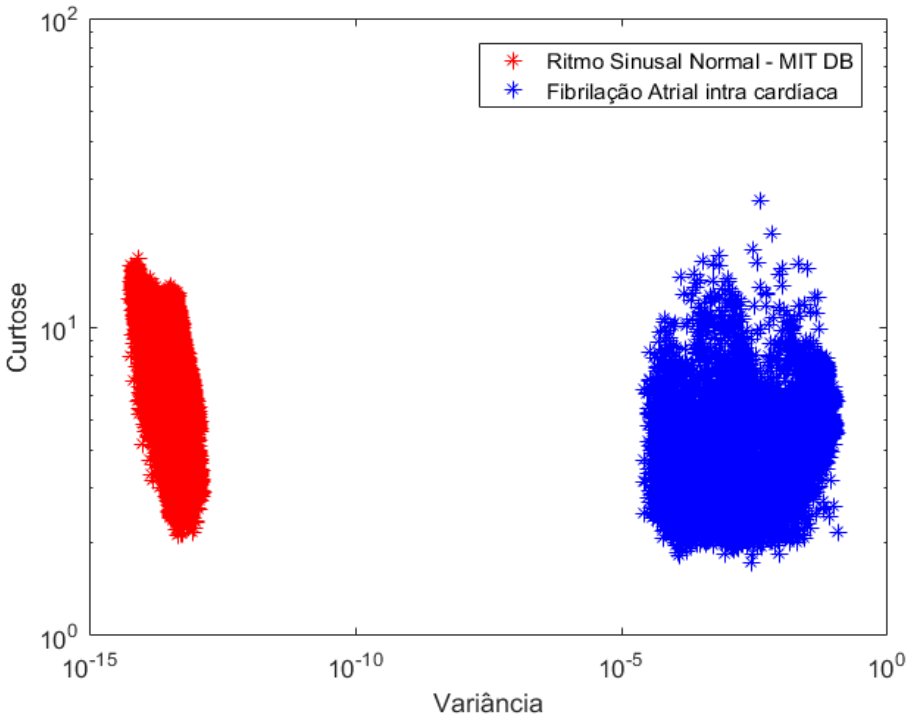
Alguns outros resultados provenientes deste estudo, incorporam a utilização ainda de mais classes, verificando se a extração se comporta da mesma maneira para as demais manifestações da FA.

Observou-se que as combinações que utilizam variância e curtose em suas representações das características obtiveram maior separação entre os dois grupos. Isso ocorre porque o ECG de indivíduos com fibrilação atrial apresenta a maior variância, enquanto indivíduos com ritmo sinusal normal apresentam menor variância nos dados.

Na Figura 20, é mostrada uma representação tridimensional, levando em consideração a variância, assimetria e curtose do conjunto de batimentos de cada grupo. Percebe-se que embora uma representação 3D proporciona uma melhor visualização dos dados, a acurácia de usar apenas duas características teve um resultado maior nos classificadores utilizados.

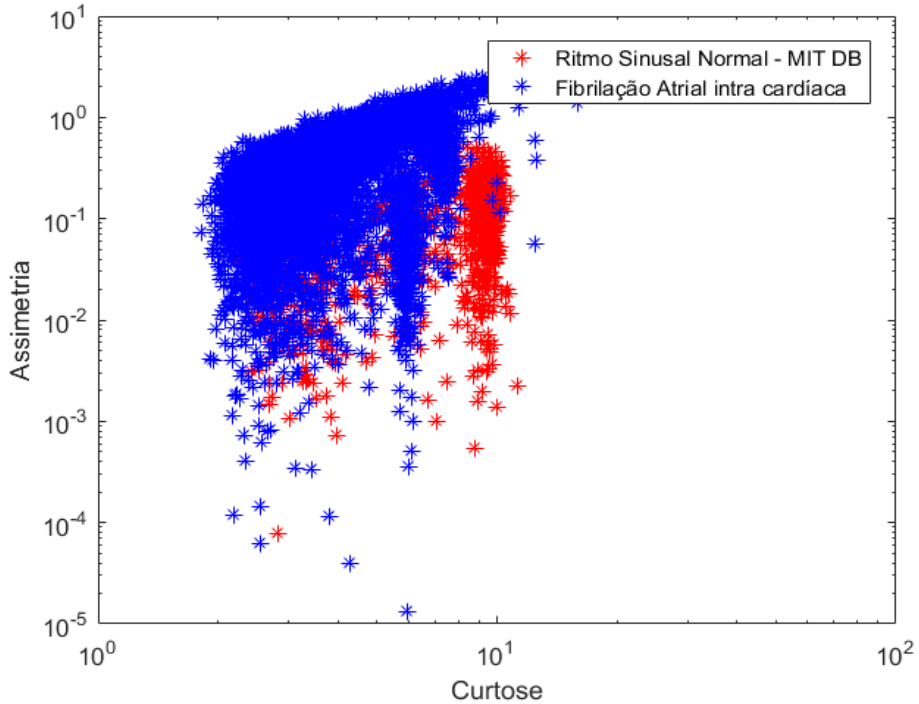
Os resultados de indivíduos saudáveis e com FA intra cardíaca são demonstrados a seguir, conforme descritos na Tabela 3.

Figura 22: Variância x Curtose



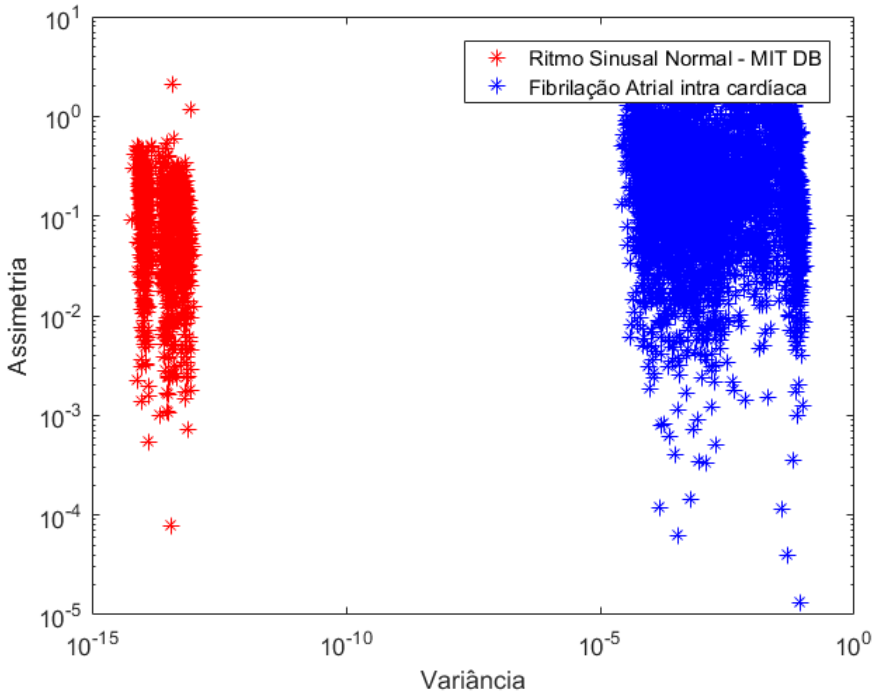
Fonte: Autor

Figura 23: Curtose x Assimetria



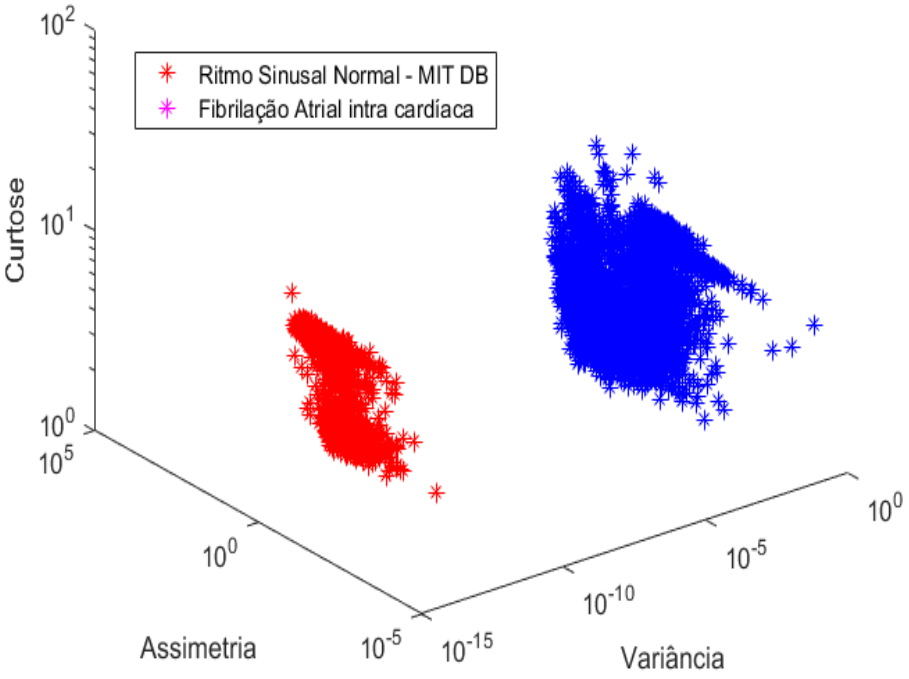
Fonte: Autor.

Figura 24: Variância x Assimetria



Fonte: Autor.

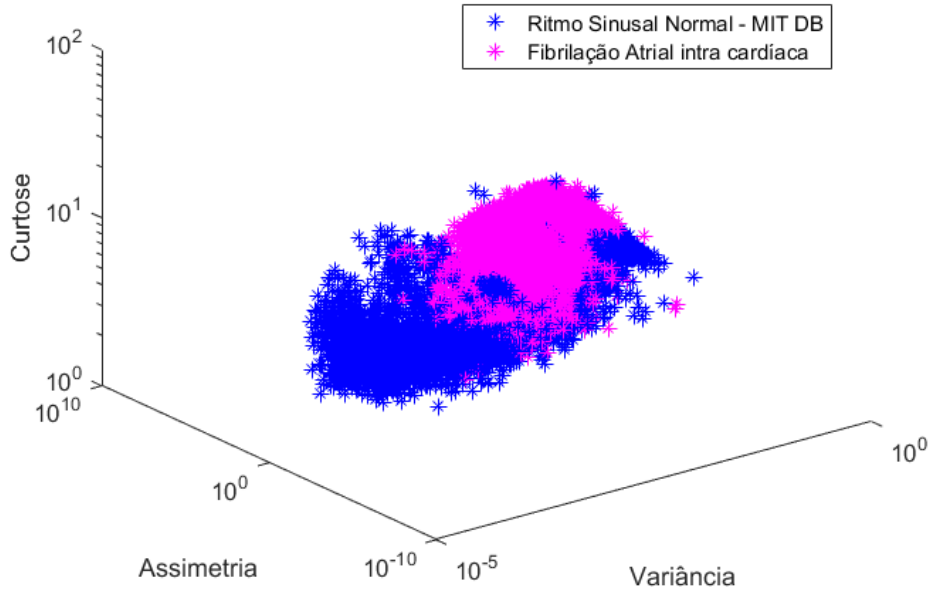
Figura 25: Variância x Assimetria x Curtose



Fonte: Autor.

Após a realização de classificações entre indivíduos saudáveis e FA e seus subtipos, agora, os resultados em classificações dicotômicas entre cardiopatias. Para os grupos de indivíduos com FA e FA intracardíaca em três dimensões, observa-se o seguinte.

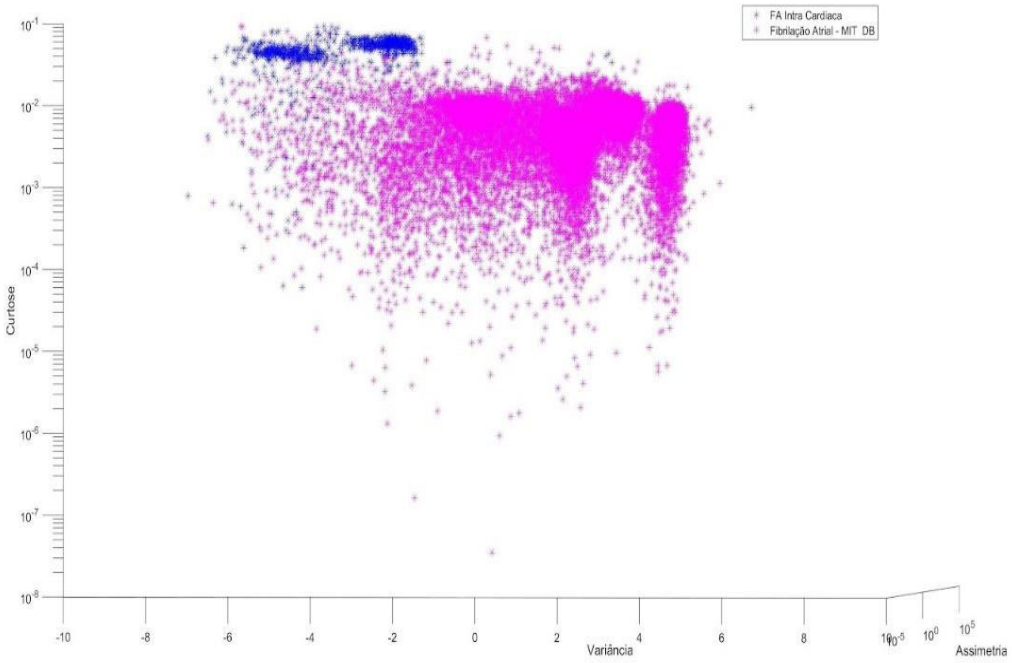
Figura 26: Variância x Assimetria x Curtose sem PCA



Fonte: Autor.

Pode-se perceber que mesmo aplicando a metodologia dos casos anteriores, não obteve-se sucesso na separação das classes, conforme descrito na Tabela 5. Isso se dá pelo seguinte motivo. A FA fisiologicamente é igual a FA intracardíaca. Entretanto, pode-se aplicar uma técnica para descorrelacionar esses dados e torná-los diferentes computacionalmente. Nessa etapa, utilizou-se o PCA nos conjuntos de dados para essa discriminação dos valores, conforme Haykyn explica em seu livro e se adequou bem nesses casos.

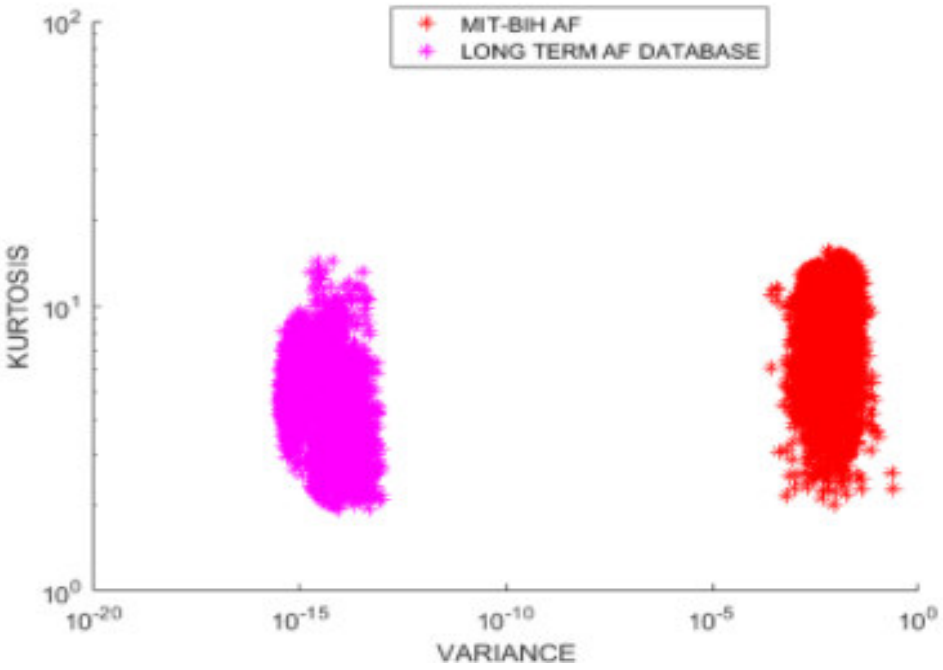
Figura 27: Variância x Assimetria x Curtose com PCA



Fonte: Autor.

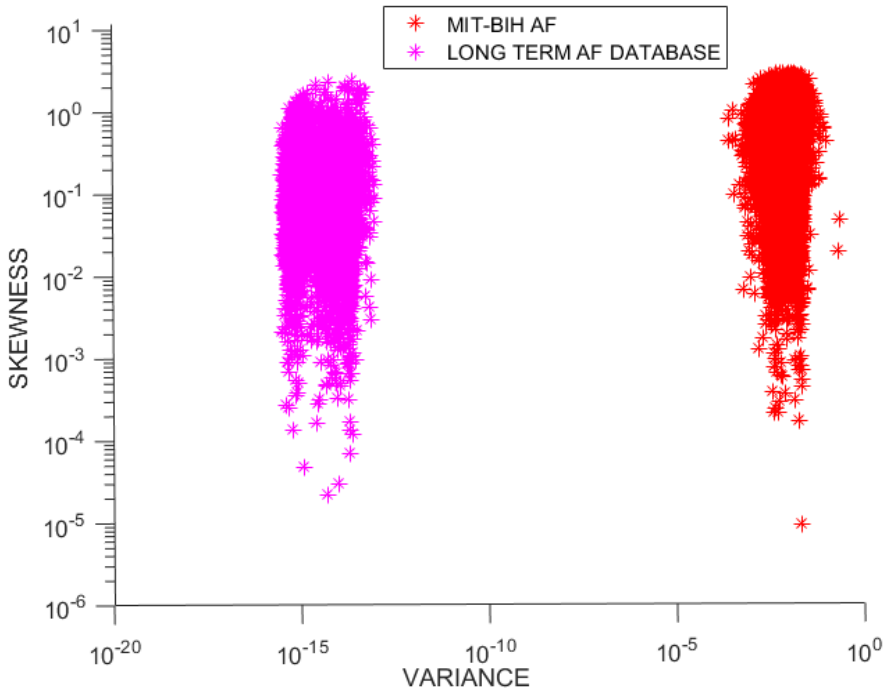
Outro resultado interessante é em relação a classificação de FA e FA paroxística. Conforme citado na Fundamentação Teórica, a FA paroxística se distingue da FA apenas pelo sua duração, uma hora o indivíduo está em ritmo sinusal normal, outra está em crise. Sendo assim, o sinal de FA paroxística é mais distintindo fisiologicamente do que os demais casos de FA. Os resultados da separação de classes estão dispostos a seguir.

Figura 28: Variância x Curtose



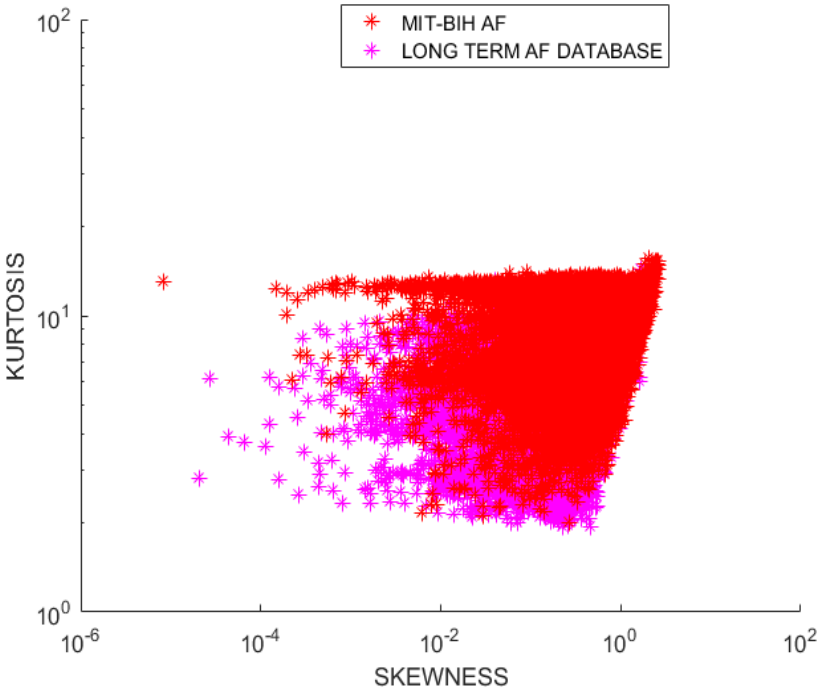
Fonte: Autor

Figura 29: Variância x Assimetria



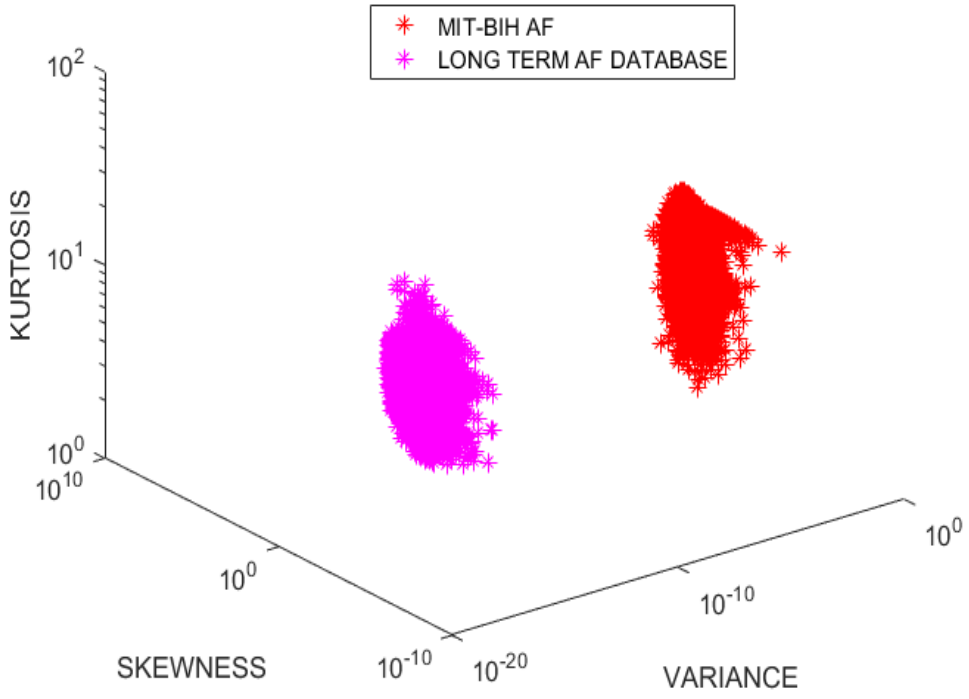
Fonte: Autor.

Figura 30: Assimetria x Curtose



Fonte: Autor.

Figura 31: Variância x Assimetria x Curtose



Fonte: Autor.

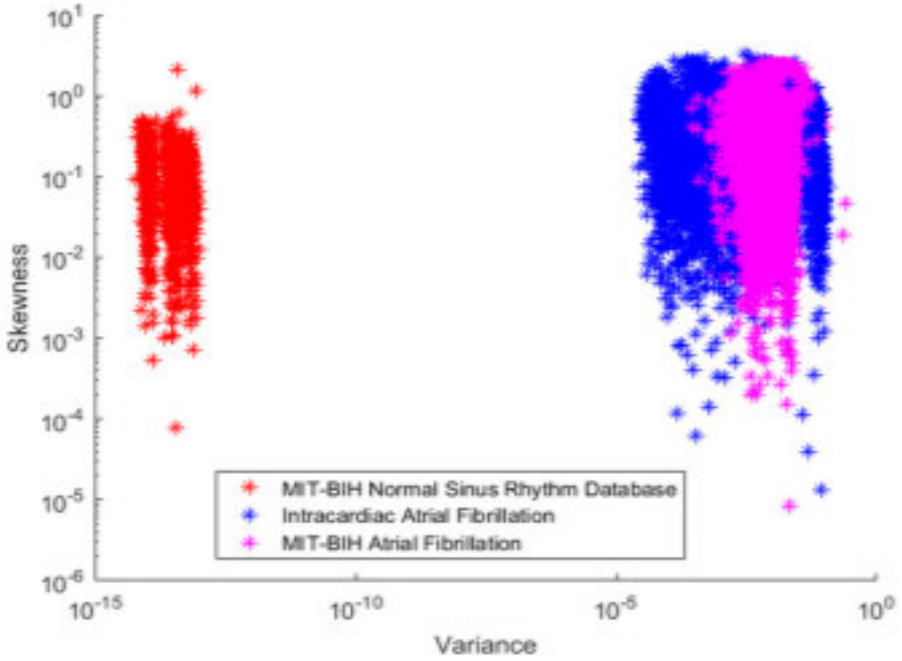
6.3 Classificação multiclasses

Mudando a perspectiva, agora os resultados serão proveninentes ao número de classes superior a 2. Os resultados aqui se refletem nas Tabelas 6, 7 e 8. Utilizou-se aqui indivíduos saudáveis e com duas ou mais cardiopatias simultaneamente, com a aplicação de PCA e sem a aplicação do mesmo.

- **Ritmo Sinusal Normal, FA e FA intracardíaca**

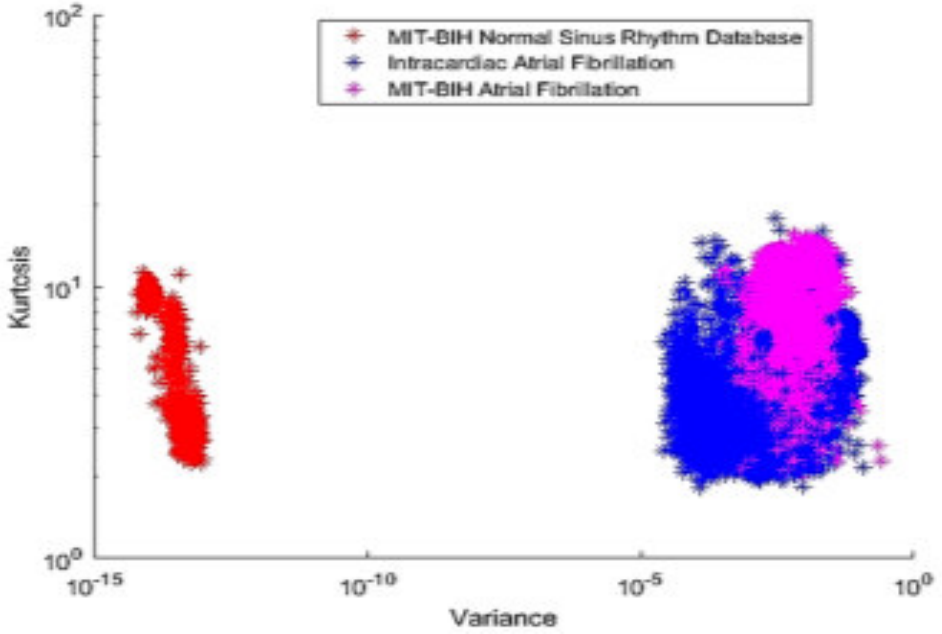
A análise sem PCA mostra que os dados AF são bastante agrupados, pois são a mesma doença cardíaca, mas com análises diferentes abordagens.

Figura 32: Variância x Assimetria sem PCA



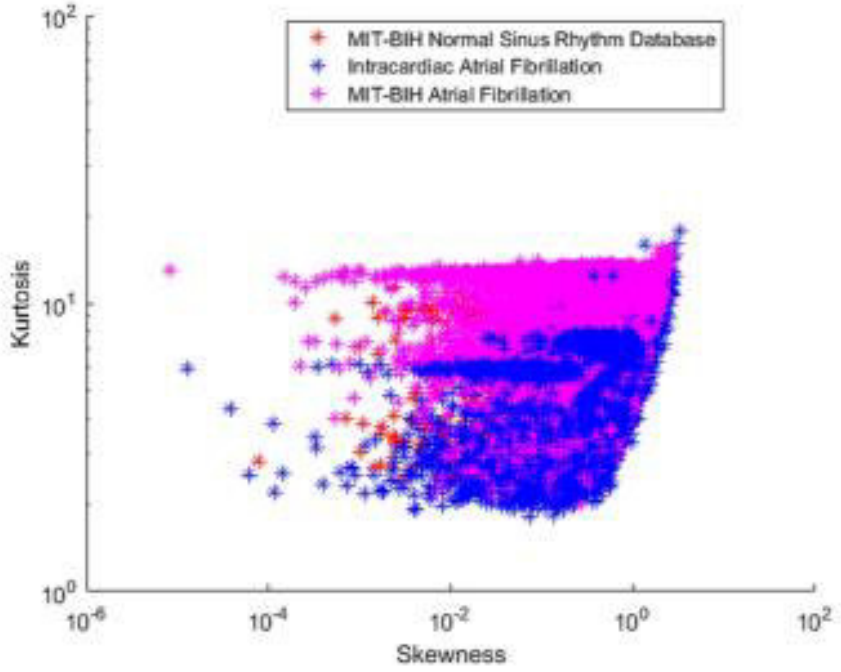
Fonte: Autor.

Figura 33: Variância x Curtose sem PCA



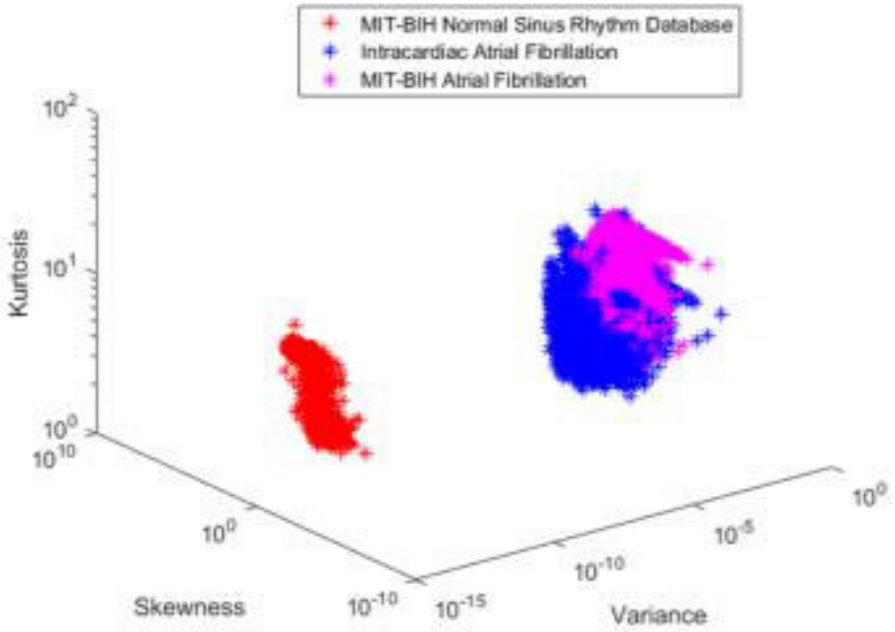
Fonte: Autor.

Figura 34: Assimetria x Curtose sem PCA



Fonte: Autor.

Figura 35: Variância x Assimetria x Curtose sem PCA



Fonte: Autor.

Pode-se perceber que os batimentos as características dos indivíduos com a cardiopatia encontram-se agrupados. A fim de separá-los, para descorrelacionar os dados, conforme evidenciado no livro do Haykin, aplicou-se o PCA e obteve-se os seguintes resultados.

Figura 36: Variância x Curtose com PCA

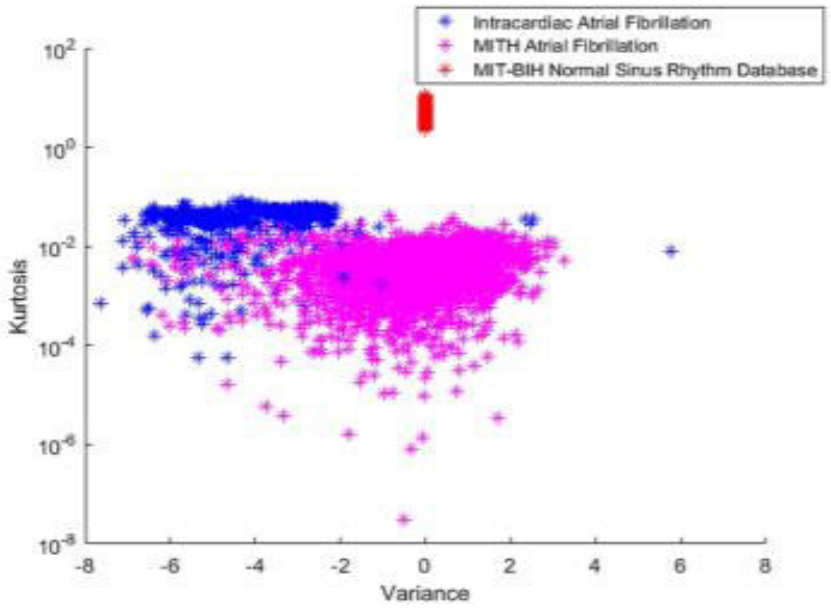
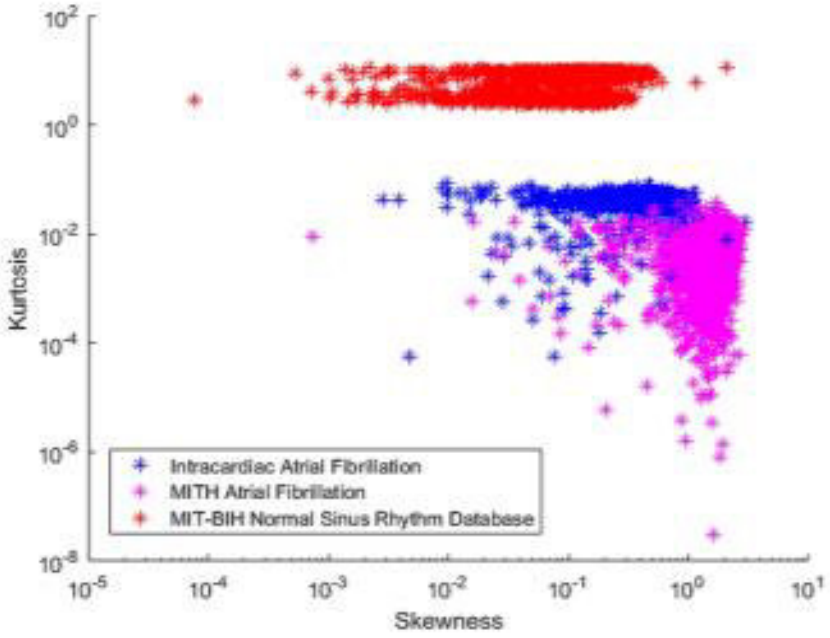


Figura 37: Assimetria x Curtose com PCA



Fonte: Autor.

Para os dados de assimetria e curtose, que anteriormente não eram satisfatórios, na classificação e separação os resultados foram maiores em 40%. Além disso, os dados foram descorrelacionados e aumentou a acurácia, que será explicitada no tópico 7.3 de Porcentuais de avaliadores utilizados.

Além disso, pode-se salientar que como o SVM não é comumente utilizado em classificações multiclasses, o resultado dele foi inferior ao da MLP. Isso reflete na forma como o SVM traça as retas ótimas, fazendo uma por uma em problemas de multiclassificação.

Diante do exposto, o uso de modificação de dados foi mostrado, mostrando uma diferença no desempenho do dados originais e os dados girados em sinais de ECG. Isto é também concluiu que embora sejam iguais patologia, FA computacionalmente e FA intracardíaca têm recursos diferentes. Também pode-se concluir que o uso de toda a batida em vez do intervalo RR pode ser uma boa metodologia para resolver este problema.

A MLP teve melhor desempenho devido ao seu parâmetro fácil definição. Para SVM, por outro lado, é necessário estimar e definir muito bem esses valores empiricamente para garantir convergência e generalização capacidade, além de adaptar a maneira de como é implementadas o traçado das retas ótimas que separam os planos. Ou seja, para alcançar o melhor resultado, é necessário testar várias arquiteturas diferentes, otimizando as retas e até mesmo o ângulo que elas fazem com o plano.

- **Ritmo Sinusal Normal, FA, FA intracardíaca e FA paroxística**

Em classificações dicotômicas, como discorrido por Silva, Queiroz, e Barros. (2020),

apenas as características de variância e curtose for am suficientes para a separação entre indivíduos saudáveis e doentes. Adicionando mais classes, como é mostrado nas Fig. 38, 39, os dados não separaram tão bem, devido a semelhança entre as classes, pois se tratam da mesma cardiopatia se manifestando de maneiras distintas**.** A classe tem um valor de 0 para FA, 1 para FA paroxística, 2 para FA intracardíaca e 3 para saudável.

Figura 38: Variância x Assimetria sem ICA

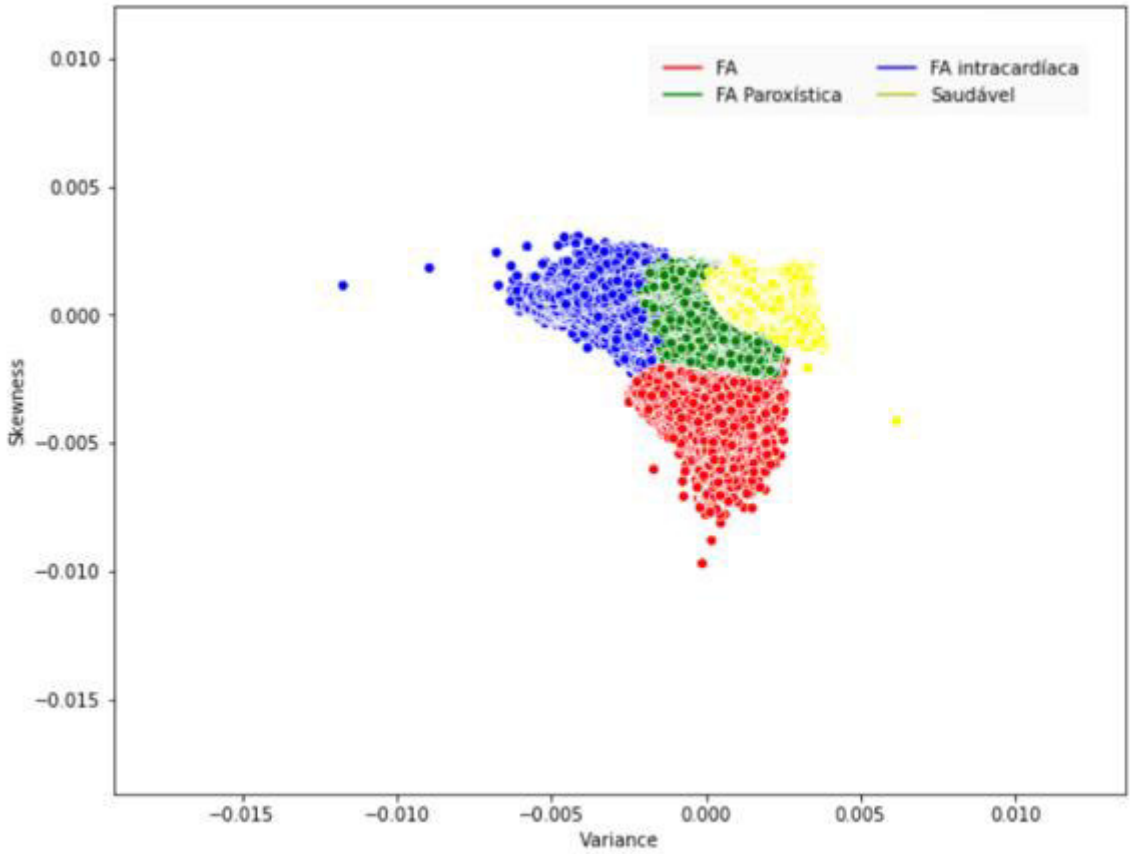
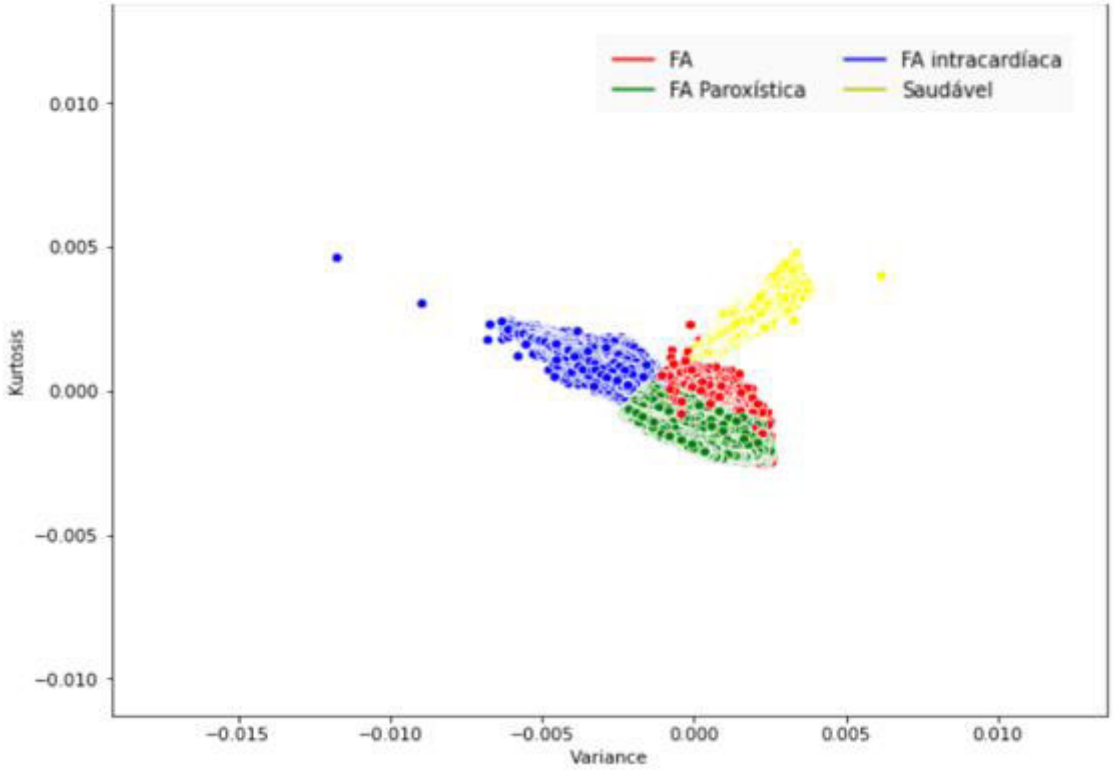


Figura 39: Variância x Curtose sem ICA

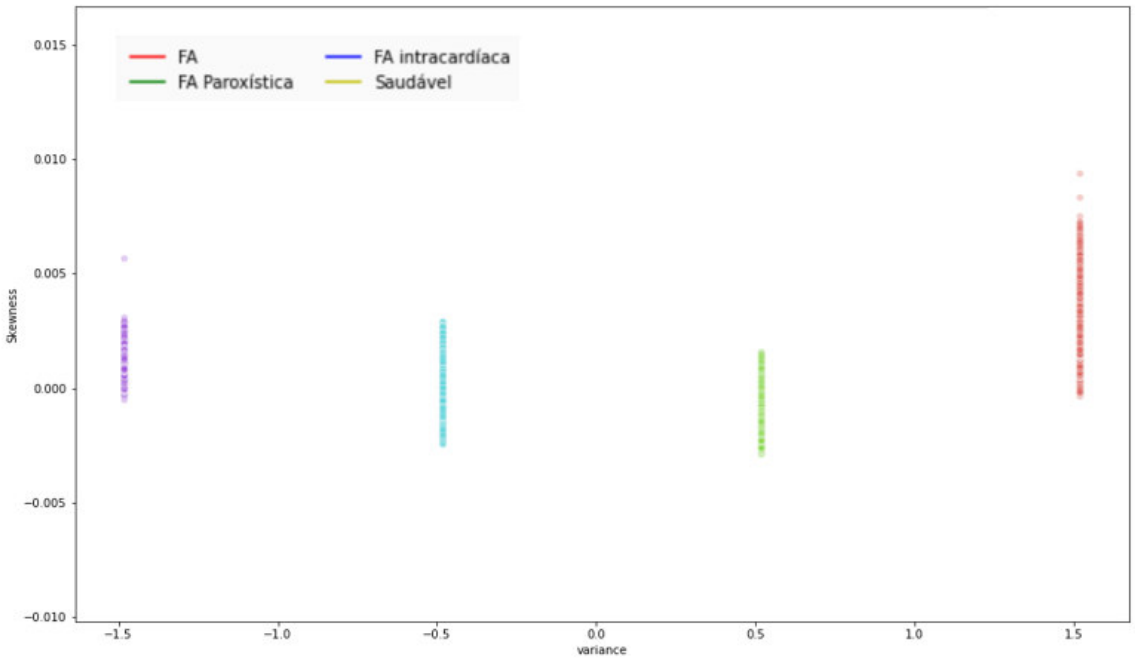


Desta forma, fez-se necessário um pre processamento a mais que pudesse descorrelacionar essas classes e provisionar a separação dos conjuntos de dados. Tal processamento é a análise com ICA. Para isto, realizou-se o branqueamento dos dados no ICA, com o PCA para descorrelacionar os dados, como proposto em [15]. A transformação ICA dividiu os sinais ECG em componentes independentes a partir da suposição que os sinais de ECG são uma combinação linear de componentes independentes. A transformação ICA separa os sinais complexos em componentes virtuais independentes.

No entanto, as suposições foram feitas sobre os sinais e como eles são combinados, admitindo que os sinais a serem reconstruídos são estatisticamente independentes um do outro. Então, com os resultados a seguir, verificou-se que as suposições estão corretas e se correspondem às condições fisiológicas reais.

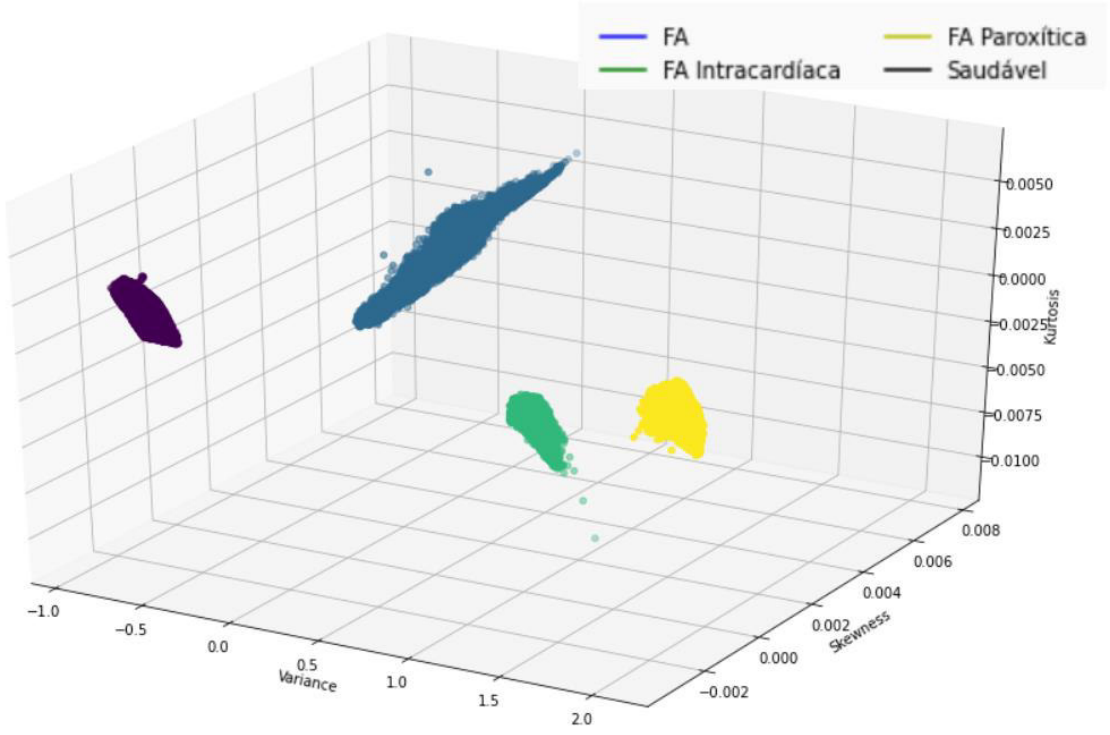
Além disso, foi possível identificar que por não se tratarem dados normais, pode-se aplicar o ICA, tal como previsto nos pressupostos de as fonte não serem gaussianas. Pode-se verificar também que alguns valores, tais como a curtose e assimetria, quando dispostos no plano cartesiano, iriam se sobrepor, não possibilitando uma boa separação dos dados, também como uma classificação satisfatória.

Figura 40: Variância x Assimetria com ICA



Fonte: Autor.

Figura 41: Variância x Assimetria x Curtose com ICA



Fonte: Autor.

7 DISCUSSÃO

De forma visual, é possível notar que após a extração de características ocorre um maior separação dos indivíduos em classificações dicotômicas. Nota-se, nesses casos, que no SVM, por exemplo, seria facilmente encontrada uma reta ótima que separaria os dados de maneira satisfatória. Quando o problema se torna de multiclassificação, e tendo a mesma doenças manifestando-se de maneiras distintas, o cenário muda e os indivíduos dos 3 grupos de FA se aglomeram em termos de características.

Para classificação dos indivíduos em ritmo sinusal normal e dois grupos cardiopatas obtemos melhores resultados na classificação após a aplicação do PCA, havendo uma mudança na disposicão dos dados, tal como nas Figuras 36, 37. Na classificação das componentes do PCA, os percentuais dos classificadores aumentaram, com destaque para os índices da MLP.

Para classificações com número de classes igual a 4, aumentou-se ainda mais a dificuldade. Além de ter mais características para processamento, as características da nova classe adicionada é também de uma subclasse de FA. Para isso, aplicou-se o PCA para branqueamento dos dados, como preparativo do método de ICA. Apenas dessa maneira foi obtida a separação de todas as classes e melhorando a classificação.

O SVM teve resultados excelentes nas classificações dicotômicas, pois performa melhor quando o número de classes é inferior a 3, por exemplo. Ele foi adaptado para o problema de multiclasse, e a acurácia e outras métricas refletem inferior devido ao traçado das retas ter de ser otimizado para mais de duas classes.

Tanto nas classificações dos vetores de característica quanto nas dos componentes do PCA e ICA, o k-NN teve melhores percentuais de acurácia, nos problemas de multiclassificação. Tendo em vista o uso da distância euclidiana para a classificação e seus índices de acurácia, podemos concluir que ambos os grupos foram bem separados. O que denota eficiência da metodologia de extração de características adotada.

Em seu trabalho, Ma [8 usou o intervalo RR para a classificação, obtendo uma precisão de 98,3%. Dentro outro estudo, os mesmos autores também utilizaram o RR intervalo, classificando com CNN-LSTM, obtendo um precisão de 97,21%. Alhusseini [9 desenvolveu uma CNN aplicada a imagens de 35 pacientes, que tomaram decisões semelhantes às dos especialistas, com 95% de precisão. Khriji usou ANN para classificar três tipos diferentes de doenças cardíacas também, obtendo 93,1% de acerto. Neste trabalho, o foi utilizada a abordagem de extração de características das batidas, com estatísticas de alta ordem e utilização do ICA, obtendo uma precisão de 99,95% para RNA.

Vale ressaltar que a contribuição deste trabalho é a caracterização da doença cardíaca por meio da extração de características do batimento cardíaco, e não do intervalo RR, como é

amplamente utilizado pelos autores já citados. Todavia, o método utilizado apresenta tendência para outliers, devido à utilização da média em todas as Estatísticas de Alta Ordem. A comparação é mostrada na Tabela 9.

Tabela 9: Comparativo de performance de metodologias anteriores com o a perfor- mance obtida neste trabalho.

Autor	Parâmetros	Acurácia
Ma et al. (2020)	Intervalo RR	98.3%
Ma et al. (2020)	Intervalo RR	97.21%
Khriji et al. (2020)	Intervalo RR	93.1%
Este Trabalho	Nosso método	99.95%

7.1 Trabalhos Futuros

Em trabalhos futuros, pode-se verificar outras ondas do coração como o ponto principal para a extração de características, como por exemplo a onda P, o complexo QRS, ao invés dos batimentos ou até mesmo o intervalo RR, comumente utilizado na literatura. Dessa maneira, poderia-se verificar e comparar quais das ondas oferecem uma melhor abordagem para classificação dos indivíduos em ritmo sinusal normal e cardiopatas.

Outro trabalho interessante é a mudança de parâmetros dos classificadores utilizados, e até utilizado outras técnicas de filtragem de sinais de ECG existentes na literatura, tais como filtros FIR, IRR, etc. Além disso, pode-se elaborar uma aplicação do ICA para filtragem de artefatos provenientes de aparelhos ou até mesmo de ruídos dos músculos, que acabam não deixando tão eficiente a obtenção dos ECG na frequência de 60 hz.

Diante do exposto, seria interessante também a confecção de uma aplicação que detecte o tipo de derivação dos ECG, como por exemplo utilizando os valores extraídos através da assimetria. Como o ECG tem uma distribuição não normal, poderia-se verificar em um conjunto de sinais, as suas respectivas derivações e confeccionar um banco de dados de assimetria extraídos desse banco e verificar os valores para cada sinal. Dessa maneira, poderia-se fazer um classificador de inversão de eletrodos, por exemplo, bastante útil para evitar diagnósticos errados na hora do exame de ECG.

8 CONCLUSÃO

Neste artigo, a eficácia do uso de estatísticas de alta ordem para extrair características e classificar as doenças cardíacas, como a fibrilação atrial, foi reforçada. Além disso, ao pré-processar os sinais utilizando ICA e PCA, foi evidenciado uma diferença no desempenho dos dados originais e os dados rotacionados nos sinais de ECG. Além disso, mostrou-se que a abordagem de utilizar o PCA para descorrelacionar os dados, como escrito no livro de Haykin, realmente funciona. Somado a isto, pode-se inferir e destacar que o PCA serve mais do que para verificar quais eixos tem maior variância dos dados. Ele serve também para branqueamento como pré-processo da técnica de ICA, também mostrada neste trabalho.

Conclui-se também que, embora sejam a mesma patologia, a FA e seus subtipos apresentam características diferentes computacionalmente. Além disso, pode-se inferior também que a utilização de todo o batimento ao invés do intervalo RR pode ser uma boa metodologia para solucionar esse problema. Entretanto, a metodologia restringe-se ao momento estatístico, permitindo margem para outliers.

No que diz respeito a classificação dos vetores de características a aplicacão do PCA e ICA proporcionou maior separação entre os grupos, obtendo melhores ındices de acurácia, com destaque para o classificador MLP com 99.6%. A partir da metodologia proposta também torna-se possível implementação de um modelo de classificação de FA, com uma menor complexidade matemática e uma implementação mais amigável em dispositivo embarcado, pois utiliza o Python.

A implementação da metodologia mostra-se relevante como alternativa para auxílio a triagem de pacientes com FA, pois são utilizadas características de fácil aquisição e foram necessários segmentos de ECG. Em trabalhos futuros, diferentes doenças cardiovasculares podem ser estudadas na metodologia e técnicas podem ser utilizadas para melhorar o pré-processamento, bem como aplicar outros classificadores para avaliar as métricas e testar hiperparâmetros dos algoritmos de classificação.

A extração de características se diferencia pois são usadas exclusivamente características obtidas no domínio do tempo, sem necessidade de técnicas mais complexas afim de obter informações no domínio da frequência. Aplicada a etapa de extração foram obtidos vetores estatisticamente distintos.

REFERÊNCIAS

[1] World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases. Acesso em: 18 de dez de 2020.

[2] L. Silva, J. Queiroz, and A. Barros,"Classificação de Fibrilação Atrial e Fibrilação Atrial Intracardíaca utilizando Estatística de Alta Ordem e Aprendizado de Máquina". 2020. from https://aprepro.org.br/conbrepro/2020/anais/arquivos/09272020_150900_5f70d62c23288.pdf.

[3] [M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation." 2020. ArXiv. Retrieved from: https://arxiv.org/abs/1805.00794.R. Brookmeyer, S. Gray, and C. Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88(9):1337–1342, 1998.

[4] J. Queiroz, A. Junior, F. Lucena, and A. Barros, "Diagnostic decision support systems for atrial fibrillation based on a novel electrocardiogram approach." Journal of electrocardiology, 51(2), 252-259, 2020. doi:10.1016/j.jelectrocard.2017.10.014.G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[5] L.Silva ., J.Queiroz. and A.Barros. Support method for the diagnosis of Atrial Fibrillation using Machine Learning. XLI CILAMCE, 16 – 19 November, 2020.

[6] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation." 2020. ArXiv. Retrieved from: https://arxiv.org/abs/1805.00794.R. Brookmeyer, S. Gray, and C. Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88(9):1337–1342, 1998..

[7] A.Ullah,, S.Anwar,. M, Bilal, and S.Mehmood. Classification of Arrhythmia by Using Deep Learning with 2-D ECG Spectral Image Representation. Arxiv. Retrieved from: https://arxiv.org/pdf/2005.06902.

[8] Ma, F, Zhang, J, Liang, W, and Xue, J. (2020). Automated Classification of Atrial Fibrillation Using Artificial Neural Network for Wearable Devices. Mathematical Problems in Engineering,J. S. Fonseca and G. A. Martins. *Curso de Estat´ıstica*. Atlas, 2010.

[9] Alhusseini, M. I, Abuzaid, F, Rogers, A. J, Zaman, J. A. B, Baykaner, T, Clopton, P, Bailis, P, Zaharia, M, Wang, P.J, Rappel, W.J, and Narayan, S. M. (2020). Machine Learning to Classify Intracardiac Electrical Patterns During Atrial Fibrillation. Circulation: Arrhythmia and Electrophysiology, 13(8). doi: 10.1161/CIRCEP.119. 008160..

[10] Khriji, L, Marwa, F, and Machhout, M. (2020). Deep Learning-Based Approach for Atrial

Fibrillation Detection. Development of Computer Vision (CV) Technology for Quality Assessment of Dates in Oman, LNCS 12157, 100–113. doi:10.1007/978-3-030-51517-1_9.

[11] Crespo X, Curell N, Curell J. Atlas de Anatomia e saúde. Paraná: Editora Bolsa Nacional do Livro, 2012.

[12] Crespo X, Curell N, Curell J. Atlas de Anatomia e saúde. Paraná: Editora Bolsa Nacional do Livro, 2012.

[13] .Sociedade Brasileira de Cardiologia. Diretriz de Interpretação de Eletrocardiograma de Repouso. Arq Bras Cardiol 2003;80

[14] Strapazzon et al.,. Interpretação básica de eletrocardiograma: o conhecimento dos enfermeiros. In: SEMINÁRIO DE INOVAÇÃO E TECNOLOGIA, 16., Porto Alegre. Anais... Porto Alegre: UNIJUI – Centro de Tecnologia, 2016. p. 2-4.

[15] M.Dubin,. Interpretação Rápida do ECG. Ed. de Publicações Científicas, 3ª ed. Rio de Janeiro, 1996

[16] G.D., Azuaje, F. and P.Msharry. Advanced Methods and Tools for ECG Data Analysis. Artech House Publishers, 1st edition. New York, 2006

[17] Ebrahimi, Z, Loni, M, Gharehbaghi, A, and Daneshtalab, M. (2020). A Review on Deep Learning Methods for ECG Arrhythmia Classification. Expert Systems with Applications X, V(7). doi: 10.1016/j.eswax.2020.100 033

[18] L.Oliveira, M.Artiaga. R.Penha. J.Batista. ANÁLISE DO ELETROCARDIOGRAMA (ECG) NORMAL – ASPECTOS ELÉTRICOS E FISIOLÓGICOS EM UMA ABORDAGEM INTERDISCIPLINAR. CEEL 2012

[19]] D. Julian., J.Cowan, Cardiologia. Ed. Santos, 6ª ed. São Paulo, 1996.

[20] Hospital Israelita Albert Einstein. Fibrilação Atrial. Disponível em: <einstein.br/especialidades/cardiologia/doencas-sintomas/fibrilacao-atrial> Acesso em: 17 set. 2020.

[21] Neto, J.F, Moreira, H. T and Miranda, C. H. (2018) Fibrilação Atrial – INÍCIO. Revista Qualidade. Retrived from: https://www.hcrp.usp.br/revista qualidade/edicaoselecionada.aspx?Edicao=6.

[22] Richter, U, Faes, L, Cristoforetti, A, Masè, M, Ravelli, F, Stridh, M, and Sörnmo, L., (2010). A Novel Approach to Propagation Pattern Analysis in Intracardiac Atrial Fibrillation Signals. Annals of Biomedical Engineering, 39(1),310-23. doi: 10.1007/s10439-010-0146-8.

[23] Associação Beneficente Síria - Hcor. Protocolo Assistencial de fibrilação Atrial. 2017

[24] Lip GY, Hee FL. Paroxysmal atrial fibrillation. QJM. 2001 Dec;94(12):665-78. doi: 10.1093/qjmed/94.12.665. PMID: 11744787.

[25] Borelli, A.F. (2018). Extração de Características em Sinais Biológicos Retrieved September 15, 2020 from: https://www.ppgee.ufmg.br/defesas/1479M.PDF

[26] T. Hastie. *Understanding Machine Learning: From Theory to Algorithms*. Springer, 2009.

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2009.

[28] I. José. Knn (k-nearest neighbors). https://towardsdatascience.com/ knn-k-nearest-neighbors-1-a4707b24bd1d. Acesso em: 25 de jun de 2019.

[29] Noi, P. T, Kappas, M. (2018). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors, 18(1),18. doi: 10.3390/s18010018.

[30] Huang, S, Cai, N, Pacheco, P. P, Narrandes, S, Wang, Y and Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer genomics & proteomics, 15(1), 41-51. doi: 10.21873/cgp.20063.

[31] R. Herbrich. *Learning kernel classifiers: theory and algorithms*. MIT press, 2001.

[32] S. Haykin. *Redes neurais: princípios e prática*. Bookman Editora, 2007.

[33] M. Kubat. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, 13(4):409–412, 1999

[34] E. Lucas. Extração do eletrocardiograma fetal utilizando métodos baseados na análise de componentes independentes e nas transformadas wavelet discreta e discreta redundante, (2019). Disponível em: http://repositorio.ufersa.edu.br/handle/prefix/6092

[35] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.

[36] Castells et al.,. Principal Component Analysis in ECG signal processing. 2006. Disponível em<https://www.researchgate.net/publication/26620236_Principal_Component_Analysis_in_ECG_Signal_Processing> Acesso em 13 set. 2020.

[37] R. Espirito. Utilização da Análise de Componentes Principais na compressão de imagens digitais (2019).

[38] RENCHER, A. C. Methods of Multivariate Analysis. 2.ed. Nova York: John Wiley & Sons, Inc, 2002

[39] J. Queiroz, A. Junior, F. Lucena, and A. Barros, “Diagnostic decision support systems for atrial fibrillation based on a novel electrocardiogram approach.” Journal of electrocardiology, 51(2), 252-259, 2020. doi:10.1016/j.jelectrocard.2017.10.014.G.

Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4):303–314, 1989.

[40] Fillype da Silva L., Queiroz J., Vanessa C., Barros A., Lopes G. and Cabral L. (2021). Separation Method of Atrial Fibrillation Classes with High Order Statistics and Classification using Machine Learning.In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: BIOSIGNALS, ISBN 978-989-758-490-9, pages 284-291. DOI: 10.5220/0010325402840291

[41] Goldberger, A.L, Amaral L.A.N, Glass, L, Hausdorff, J. M, Ivanov, P. C, Mark, R.G, Mietus, J.E, Moody, G.B, Peng, C.K, and Stanley, H. E. (2000). The MIT-BIH Atrial Fibrillation Database. Retrieved September 12, 2020, from: https://archive.physionet.org/physiobank/ database/afdb

[42] Goldberger A. L, Amaral L.A, Glass L, Hausdorff J. M, Ivanov P. C, Mark, R. G., Mietus, J. E, Moody G. B, Peng C. K., and Stanley H. E., (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23): e215-e220

[43] Hyv¨arinen, A., (2013). Independent component analysis: recent advances, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371(1984), 1471 – 2962.

[44] Sociedade Brasileira de Cardiologia. Diretriz de Interpretação de Eletrocardiograma de Repouso. Arq Bras Cardiol 2003;80(supl II).

[45] S, Russel; P.Norvig, Artificial Intelligence – a modern approach. Prentice-Hall, New Jersey, 1995. Disponível em: https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modernapproach.9780131038059.25368.pdf. Acesso em: 17 set. 2021, às 18h.

[46] L.Deng; , D.Yu. Deep learning: methods and aplications, journal of artificial intelligence research 4: 237-285. 1996. Disponível em: https://www.google.com/search?client=ubuntu&channel=fs&q=Deep+learning%3A+methods+and+aplications&ie=utf8&oe=utf-8. Acesso em: 1 out. 2021, às 15h.

[47] R, Bianchi. Uso de Heurísticas para a Aceleração do Aprendizado por Reforço. 2004. Tese (Doutorado em Engenharia) – Universidade de São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/3/3141/tde-28062005-191041/publico/tese- bianchi.pdf. Acesso em: 18 set. 2021, às 20h.

[48] A, Silva. Implementação e Aplicação de Aprendizado em um Sistema NeuroSimbólico. 2017. Dissertação (Mestrado em Engenharia de Computação e Elétrica) – Universidade Federal do Rio Grande do Sul. Disponível em: https://repositorio.ufrn.br/jspui/bitstream/123456789/22563/1/AndreQuintilianoBezerraSil va_DISSERT.pdf. Acesso em: 05 set. 2021, às 17h.

[49] D.Albright, D. 10 Examples of Artificial Intelligence. Disponível em:

https://www.techemergence.com/everyday-examples-of-ai/. Acesso em: 29 out. 2019, às 19hSILVA, A. Implementação e Aplicação de Aprendizado em um Sistema NeuroSimbólico. 2017. Dissertação (Mestrado em Engenharia de Computação e Elétrica) – Universidade Federal do Rio Grande do Sul. Disponível em: https://repositorio.ufrn.br/jspui/bitstream/123456789/22563/1/AndreQuintilianoBezerraSil va_DISSERT.pdf. Acesso em: 05 set. 2021, às 17h.

[50] E. Bezerra. Introdução à Aprendizagem Profunda. CEFET. RJ. Out. 2016. ResearchGate. Disponível em: https://www.researchgate.net/publication/309321510. Acesso em: 18 set. 2021, às 20.

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

Lago Cutrim Barros, Luis Fillype da Silva.
Classificação de SUbclasses de Fibrilação Atrial utilizando Estatística de Alta Ordem e Aprendizado de Máquina / Luis Fillype da Silva Lago Cutrim Barros. - 2022.
72 f.

Orientador(a): Allan Kardec Duailibe Barros Filho.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, Google Meet, 2022.

1. Aprendizado de Máquina. 2. Eletrocardiograma. 3. Fibrilação Atrial. I. Duailibe Barros Filho, Allan Kardec. II. Título.