UNIVERSIDADE
LUSÓFONA

Deteção de Arritmias através de modelos de Machine Learning

Trabalho Final de Curso

Entrega Final

Sebastião Coelho, 22202310, LIG

Orientador: Prof. Iolanda Velho

Coorientadores: Prof. Lúcio Studer, Dr. Luís Rosário

Entidades Externas: Hospital de Santa Maria, FMUL, IST, ISCTE

Departamento de Engenharia Informática e Sistemas de Informação

Universidade Lusófona, Centro Universitário de Lisboa

13/07/2025

Direitos de cópia

Resumo

As doenças cardiovasculares são a principal causa de morte no mundo. Em 2016 este tipo de doença representou 31% dos óbitos a nível global. “Mais de três quartos das mortes por doenças cardiovasculares ocorrem em países de baixa e média renda.”, destacando a disparidade no acesso a cuidados de saúde e a necessidade de soluções acessíveis para melhorar a prevenção e o tratamento destas condições[1].

Hoje, graças ao progresso tecnológico, os smartphones estão ao alcance de uma vasta parcela da população mundial. Segundo um estudo de 2019 do Pew Research Center, a posse de smartphones tem crescido rapidamente em diversas partes do mundo, incluindo um aumento notável em mercados emergentes como a Índia, o Quénia e o Vietname [2]. Esta ubiquidade permite que estes dispositivos, presentes no quotidiano de milhões, sejam explorados como ferramentas de saúde.

O objetivo do trabalho foi contribuir para a identificação precoce de arritmias através de um modelo de Machine Learning e mais tarde, já fora do âmbito deste projeto, integrar o mesmo numa aplicação móvel previamente desenvolvida capaz de fazer a leitura da frequência cardíaca através de uma técnica chamada Fotopletismografia, com o objetivo final de prever eventos cardiovasculares e, assim, facilitar a intervenção precoce e a prevenção eficaz.

Para validar essa abordagem, foi desenvolvido um modelo de Machine Learning treinado com métricas extraídas de sinais de ECG. Após o tratamento dos dados, os algoritmos testados permitiram avaliar a capacidade do sistema em distinguir entre registos com e sem arritmias irregulares. O modelo final alcançou resultados muito positivos, com um F1-score de 0.9000 e uma AUC de 0.9361, confirmando o potencial da análise automática de sinais cardíacos como ferramenta complementar no apoio ao diagnóstico precoce de condições cardíacas.

Este projeto foi desenvolvido no âmbito do Trabalho Final de Curso (TFC) da Licenciatura de Informática de Gestão (LIG), sob a orientação da Professora Iolanda Velho e coorientação do Professor Lúcio Studer, docentes do Departamento de Engenharia Informática e Sistemas de Informação (DEISI) da Universidade Lusófona. Em parceria com o Professor Dr. Luís Rosário, médico cardiologista no Hospital de Santa Maria e docente na Faculdade de Medicina da Universidade de Lisboa (FMUL) e no Instituto Superior Técnico (IST). Foi o Dr. Luís Rosário e a sua equipa de investigadores do ISCTE que desenvolveram a aplicação móvel no âmbito do projeto AIMHealth [4], atualmente em fase de testes e ainda não disponível ao público. Este trabalho foi proposto como um contributo complementar para a evolução da aplicação, com vista à futura integração de capacidades de deteção automática de arritmias.

Palavras-chave: Machine Learning; Arritmias; Monitorização Cardíaca; Eletrocardiograma; Fotopletismografia; Dispositivos Móveis; Saúde Digital

Abstract

Cardiovascular diseases are the leading cause of death worldwide. In 2016, these diseases accounted for 31% of global deaths. "More than three-quarters of deaths from cardiovascular diseases occur in low and middle class countries"[1], highlighting disparities in access to healthcare and the need for accessible solutions to improve prevention and treatment.

With the widespread adoption of mobile technology, smartphones are now accessible across diverse socioeconomic groups [2]. This ubiquity makes them promising tools for health monitoring. The objective of this project is to support the early detection of arrhythmias by developing a machine learning model, which will be later integrated into a previously developed mobile application capable of measuring heart rate using photoplethysmography. The ultimate goal is to predict adverse cardiovascular events and thus enable timely intervention and effective prevention.

The objective of this project was to contribute to the early identification of arrhythmias through a Machine Learning model. This model is intended to be integrated, outside the scope of this specific project, into a previously developed mobile application. This application utilizes Photoplethysmography to measure heart rate, with the ultimate goal of predicting adverse cardiovascular events, thereby facilitating timely intervention and effective prevention.

To validate this approach, a Machine Learning model was developed and trained using metrics extracted from ECG signals. After data preprocessing, the tested algorithms demonstrated the system's capability to distinguish between records with and without irregular arrhythmias. The final model achieved highly positive results, with an F1-score of 0.9000 and an AUC of 0.9361, confirming the potential of automated cardiac signal analysis as a complementary tool to support the early diagnosis of cardiac conditions.

This project was developed as part of the Final Course Project (TFC) for the Bachelor's degree in Information Technology Management at Universidade Lusófona, under the supervision of Professor Iolanda Velho and co-supervision of Professor Lúcio Studer, both faculty members of the Department of Computer Engineering and Information Systems (DEISI). This work was done in collaboration with Professor Dr. Luís Rosário, a cardiologist at Hospital de Santa Maria and faculty member at the Faculdade de Medicina da Universidade de Lisboa (FMUL) and Instituto Superior Técnico (IST). Dr. Luís Rosário and his research team from ISCTE developed the mobile application within the scope of the AIMHealth project [4], which is currently in testing phases and not yet publicly available. This project serves as a complementary contribution to the evolution of the application, aiming for the future integration of automated arrhythmia detection capabilities.

Keywords: Machine Learning; Arrhythmias; Cardiac Monitoring; Electrocardiogram; Photoplethysmography; Mobile Devices; Digital Health

Índice

Lista de Figuras

Lista de Tabelas

Lista de Siglas

OMS	Organização Mundial da Saúde
TFC	Trabalho Final de Curso
LIG	Licenciatura de Informática de Gestão
DEISI	Departamento de Engenharia Informática e Sistemas de Informação
ULHT	Universidade Lusófona de Humanidades e Tecnologia
FMUL	Faculdade de Medicina da Universidade de Lisboa
IST	Instituto Superior Técnico
PPG	*Photoplethysmography* Inglês – Fotopletismografia em Português
AVC	Acidente Vascular Cerebral
ODS	Objetivos de Desenvolvimento Sustentável
FA	Fibrilhação Auricular
ECG	Eletrocardiograma
SVM	*Support Vector Machines*
BPM	Batimentos por Minuto
HRV	*Heart Rate Variability*

1 Introdução

As doenças cardiovasculares, como as arritmias, são uma das principais causas de morte a nível mundial [1], com um impacto significativo na qualidade de vida e na sobrecarga dos sistemas de saúde. Apesar da relevância do diagnóstico precoce, tradicionalmente, este diagnóstico dependia da utilização de equipamento especializado e infraestrutura hospitalar avançada. Contudo, avanços recentes têm permitido a utilização de dispositivos portáteis, aliados a modelos de Machine Learning, para a deteção deste tipo de condições [3]. Facilitando o acesso a diagnósticos em contextos de classes sociais baixa e média. Esta evolução evidencia a necessidade crescente de soluções acessíveis, capazes de proporcionar monitorização contínua e deteção precoce de condições críticas como as arritmias.

Este trabalho surge no contexto do projeto "AIM Health" [4], uma aplicação móvel previamente desenvolvida pelo Dr. Luís Rosário e seus alunos do IST [5] e [6]. Este software utiliza uma técnica chamada fotopletismografia (PPG) para medir a frequência cardíaca através da câmara e da lanterna de dispositivos móveis, a aplicação já foi autorizada pelo Infarmed, instituição reguladora de dispositivos de saúde, com esta autorização será considerada um dispositivo médico quando for disponibilizada ao público.

A aplicação já demonstrou eficácia na medição de parâmetros fisiológicos básicos, como a frequência cardíaca e a frequência respiratória quando os resultados são comparados com medições provenientes de eletrocardiogramas hospitalares [5] e [6]. No entanto, ainda não possui funcionalidades para identificar padrões associados a arritmias. A proposta deste Trabalho Final de Curso (TFC) é desenvolver um modelo de Machine Learning capaz de detetar ritmos cardíacos anómalos, com o objetivo de, futuramente, poder ser integrado na aplicação, ampliando o seu contributo para o rastreio e monitorização de doenças cardiovasculares.

Complementando o trabalho realizado, espera-se tornar a aplicação móvel numa ferramenta acessível, capaz de realizar monitorização cardíaca em tempo real, com impacto direto na prevenção de complicações graves, como Acidente Vascular Cerebral (AVC) e insuficiência cardíaca. Além disso, a solução alinha-se aos Objetivos de Desenvolvimento Sustentável (ODS) das Nações Unidas [7], promovendo saúde e bem-estar pelo meio de inovação tecnológica.

Este trabalho destaca-se não apenas pela sua relevância clínica, mas também pelo seu suporte científico. A aplicação base foi fundamentada em estudos publicados pelo Dr. Luís Rosário e equipa [5] e [6], e a integração do modelo de Machine Learning será desenvolvida com base em dados clínicos fornecidos pelo Hospital de Santa Maria, além de bases de dados públicas reconhecidas.

Esta parceria surgiu a partir de um convite da Professora Iolanda ao Dr. Luís, especialista em cardiologia e com vasta experiência na colaboração com estudantes, com o objetivo de unir os campos da tecnologia e da saúde. A colaboração entre diferentes instituições de ensino superior, nomeadamente a Universidade Lusófona, o Instituto Superior Técnico e o ISCTE, demonstra o valor do trabalho conjunto entre universidades, promovendo a inovação e permite dar continuidade a projetos já desenvolvidos, acrescentando-lhes valor.

1.1 Enquadramento

As arritmias cardíacas são distúrbios que afetam a frequência ou o ritmo dos batimentos cardíacos [8]. A frequência cardíaca refere-se ao número de batimentos por minuto, enquanto o ritmo cardíaco descreve o padrão e a regularidade com que os batimentos ocorrem. Distúrbios na frequência podem resultar em batimentos demasiado rápidos (taquicardia) ou demasiado lentos (bradicardia), enquanto alterações no ritmo podem causar irregularidades nos intervalos entre os batimentos, como na fibrilhação auricular (FA). Estas condições podem comprometer a eficiência do coração em bombear sangue, afetando a oxigenação dos tecidos e órgãos vitais. A identificação precisa do tipo de arritmia é crucial para a implementação de tratamentos adequados e para a prevenção de complicações graves, como AVC e insuficiências cardíacas [8].

A variabilidade da frequência cardíaca (Heart Rate Variability - HRV) representa as variações nos intervalos entre batimentos cardíacos consecutivos [9]. Embora possa parecer desejável que o coração bata sempre num ritmo constante, uma ligeira variação é, na verdade, sinal de um sistema cardiovascular saudável. Isto acontece porque o organismo está constantemente a adaptar-se a estímulos internos (como emoções, respiração ou digestão) e externos (como mudanças de temperatura ou movimento físico), e essa adaptação reflete-se nas flutuações da frequência cardíaca.

No entanto, na ausência de qualquer estímulo, espera-se que o ritmo cardíaco se mantenha relativamente estável. Uma HRV excessivamente irregular em repouso pode indicar problemas no controlo autonómico do coração ou a presença de arritmias. A fibrilhação auricular (FA), por exemplo, é uma das condições mais comuns associadas a uma perda clara dessa variabilidade normal, refletindo uma atividade elétrica inconstante nas cavidades superiores do coração, conhecidas como aurículas.

A deteção precoce de casos de ritmo cardíaco irregular é crucial para a implementação de tratamentos que previnam complicações graves. Os procedimentos mais comuns após diagnóstico são:

- Prevenção de eventos tromboembólicos, utilizando anticoagulantes orais para reduzir o risco de AVC.
- Controlo da frequência cardíaca, através de medicamentos que regulam a resposta dos ventrículos, mantendo a frequência cardíaca dentro dos valores normais.
- Controlo do ritmo cardíaco, com estratégias para restaurar e manter o ritmo normal do coração, incluindo cardioversão elétrica ou farmacológica e procedimentos de ablação, conforme o caso.

A monitorização contínua e precisa dos ritmos cardíacos é, portanto, fundamental para a deteção precoce e gestão eficaz de eventos cardiovasculares, melhorando significativamente os prognósticos dos pacientes [10].

1.2 Motivação e Identificação do Problema

A escolha deste trabalho foi impulsionada pelo meu interesse pessoal em atuar na área da saúde, aliando competências tecnológicas para resolver problemas reais e relevantes.

A saúde é uma área de impacto direto na qualidade de vida das pessoas, e contribuir para a criação de soluções que possam melhorar diagnósticos e tratamentos sempre foi um dos principais pilares da minha trajetória académica. Este projeto oferece uma oportunidade única de trabalhar num tema que combina tecnologia inovadora com necessidades críticas do setor de saúde, proporcionando um desafio tanto técnico quanto social.

Além disso, o projeto destaca-se pelo seu caráter colaborativo e multidisciplinar. Trabalhar num contexto que envolve várias entidades e especialistas de diferentes áreas, como análise de dados, tecnologia e cardiologia, representa um cenário enriquecedor e motivador. A parceria com o Hospital de Santa Maria, o Instituto Superior Técnico e ISCTE não só eleva o nível técnico do projeto, como também reforça a sua relevância prática, algo que eu dei valor na escolha do TFC.

O meu TFC foi originalmente aceite com um outro tema também na área da saúde, análise de dados e tecnologia. No entanto, durante conversas com o Dr. Luís Rosário, que aceitou coorientar o trabalho a convite da Professora Iolanda Velho, surgiu a oportunidade de colaborar diretamente num projeto já em andamento. Este projeto destacou-se não apenas pela sua relevância clínica, mas também pelo seu caráter multidisciplinar, envolvendo profissionais de cardiologia, engenharia e análise de dados. A proposta de integrar um modelo de machine learning numa aplicação móvel para monitorização cardíaca apresentou-se como um desafio técnico e um possível contributo para a saúde pública.

A deteção precoce de arritmias cardíacas, continua a ser um desafio significativo na prática clínica. Apesar dos avanços tecnológicos, o diagnóstico de arritmias ainda depende, na maioria dos casos, de exames realizados em ambiente hospitalar [10]. Esta realidade limita o acesso de grande parte da população mundial, especialmente em regiões com infraestruturas de saúde limitadas ou entre classes sociais com menores recursos financeiros, contribuindo para o subdiagnóstico e atraso na intervenção. Existe, assim, uma oportunidade clara para o desenvolvimento de soluções baseadas em tecnologia móvel e modelos de Machine Learning capazes de identificar sinais precoces de arritmias fora do contexto hospitalar. Estas abordagens visam democratizar o acesso ao diagnóstico e acelerar a resposta às necessidades de saúde pública.

1.3 Objetivos

O TFC teve como principal objetivo o desenvolvimento de um modelo de Machine Learning capaz de identificar sinais de ECG com uma variabilidade da frequência cardíaca (HRV) suficientemente irregular para serem considerados indicativos de um ritmo cardíaco anormal ou de uma condição cardiovascular. Pretende-se que este modelo alcance um elevado nível de precisão, garantindo a sua aplicabilidade em contextos clínicos. Numa fase posterior, será considerada a integração do modelo numa aplicação móvel já existente, potenciando a sua utilização em ambientes reais de monitorização e triagem.

O trabalho foi, por isso, organizado em duas fases complementares. A Fase 1, concluída neste TFC, centrou-se na pesquisa teórica, tratamento dos dados clínicos e construção do modelo preditivo. Já a Fase 2, que inclui a integração do modelo na aplicação móvel e a realização de testes em ambiente real, encontra-se fora do âmbito deste trabalho

e está prevista como uma possível continuidade, a ser desenvolvida por futuros alunos em próximos TFCs. Abaixo encontra-se a descrição detalhada das duas fases propostas.

Fase 1: Criação do Modelo de Machine Learning

- Pesquisa sobre Arritmias e Cardiologia
  - Compreender as suas causas, características e sinais associados às arritmias
  - Estudar os métodos tradicionais de diagnóstico
  - Consultar literatura científica, para fundamentar o desenvolvimento do modelo
- Tratamento dos Dados:
  - Identificar repositórios de dados médicos online
  - Limpar, normalizar e estruturar os dados recolhidos
  - Identificar os picos R dos sinais ECG
  - Criar métricas representativas da variabilidade da frequência cardíaca (HRV) dos sinais ECG
- Desenvolvimento do Modelo:
  - Testar e comparar o desempenho de diferentes modelos de Machine Learning
  - Selecionar o modelo mais eficaz para prever a presença de arritmias

Fase 2: Implementação e Integração na Aplicação Móvel

- Integração na Aplicação:
  - Incorporar o modelo na aplicação já existente
  - Implementar a funcionalidade de deteção de FA em tempo real
- Testes de Usabilidade e Validação Clínica:
  - Realizar testes com utilizadores finais, como profissionais de saúde e pacientes
  - Validar os resultados da aplicação em cenários reais, comparando-os com diagnósticos feitos por profissionais de saúde

1.4 Estrutura do Documento

A estrutura do documento é a seguinte:

- **Secção 1:** Introdução ao tema do trabalho
- **Secção 2:** Discutem-se a pertinência e a viabilidade do trabalho, incluindo a relevância no contexto clínico e técnico.
- **Secção 3:** Apresentam-se os conceitos teóricos fundamentais e os algoritmos relevantes para o desenvolvimento da solução.
- **Secção 4:** Explora-se o estado da arte, analisando soluções existentes e destacando a proposta de inovação e as mais-valias do projeto.
- **Secção 5:** Detalha-se a solução proposta, incluindo as metodologias, a recolha e o pré-processamento de dados, e a análise exploratória dos mesmos.
- **Secção 6:** Descreve-se o método e o planeamento seguido no desenvolvimento do projeto, incluindo o cronograma previsto (Gantt) e a análise crítica da execução face ao planeado.
- **Secção 7:** Apresentam-se os resultados obtidos com os modelos desenvolvidos, discutindo o desempenho dos diferentes algoritmos, as limitações identificadas e a relevância dos resultados.
- **Secção 8:** Conclui-se o trabalho, sintetizando os principais contributos, e são propostas direções futuras para evolução da solução.

2 Pertinência e Viabilidade

2.1 Pertinência

Este trabalho é relevante devido ao elevado número de pessoas que podem beneficiar de um diagnóstico precoce, reduzindo significativamente o risco de complicações graves como casos de AVC e de insuficiência cardíaca. O projeto procura responder à necessidade de ferramentas acessíveis, portáveis e eficientes para a deteção precoce destas condições, tornando possível utilizar tecnologias móveis não apenas para a medição da frequência cardíaca mas também para a identificação de padrões patológicos associados à variabilidade cardíaca. Além disso, a possibilidade de um paciente conseguir obter, a qualquer momento, um registo PPG através do seu telemóvel permite-lhe guardar dados de momentos específicos em que se sentiu mal, facilitando a posterior análise clínica e contribuindo para diagnósticos mais precisos e informados.

A relevância clínica do projeto é reforçada pela colaboração com o Dr. Luís Rosário, Médico cardiologista e também Professor, envolvido em projetos de investigação na área.

O trabalho já desenvolvido deu origem a dois artigos científicos publicados em revistas internacionais de alto impacto:

- "Benchmarking of Sensor Configurations and Measurement Sites for Out-of-the-Lab Photoplethysmography" (2024) [5]
- "Validation of an mHealth System for Monitoring Fundamental Physiological Parameters in the Clinical Setting" (2023) [6]

2.2 Viabilidade

Alinhamento com os Objetivos de Desenvolvimento Sustentável

Os Objetivos de Desenvolvimento Sustentável (ODS) foram estabelecidos pela Organização das Nações Unidas como parte da Agenda 2030. Estes objetivos visam erradicar a pobreza, proteger o planeta e garantir condições de paz e prosperidade para todos. Cada objetivo define áreas estratégicas prioritárias, incluindo a promoção da saúde, da inovação e do desenvolvimento de infraestruturas.

O projeto está alinhado com os seguintes ODS [7]:

- ODS 3 - Saúde e Bem-Estar: Contribui para melhorar o diagnóstico precoce de arritmias, prevenindo complicações graves e promovendo o bem-estar da população.
- ODS 9 - Indústria, Inovação e Infraestrutura: Promove a integração de tecnologias inovadoras, como Machine Learning e dispositivos móveis, no campo da saúde.
- ODS 10 – Resolução das Desigualdades: O projeto contribui diretamente para a redução das desigualdades ao democratizar o acesso à monitorização de saúde cardíaca, oferecendo uma solução amplamente acessível via smartphone, o que pode beneficiar populações em regiões com menor acesso a cuidados médicos especializados.

Viabilidade Técnica

A implementação técnica do projeto é viável devido à escolha de ferramentas consolidadas e amplamente utilizadas, como Python e bibliotecas de Machine Learning (por exemplo: pandas, sklearn, numpy). Estas ferramentas permitem um desenvolvimento eficiente e flexível, adequando-se ao processamento dos dados e ao desenvolvimento do modelo de Machine Learning.

Para o desenvolvimento e validação do modelo, foi utilizada exclusivamente uma base de dados pública de elevada qualidade disponibilizada pelo site PhysioNet disponível em: https://physionet.org/ , nomeadamente o dataset *A Large Scale 12-lead Electrocardiogram Database for Arrhythmia Study* [11].

Embora tenha sido inicialmente considerada a possibilidade de incluir dados clínicos provenientes do Hospital de Santa Maria, essa colaboração permanece como uma perspetiva para trabalhos futuros.

Adicionalmente, existem outros bancos de dados relevantes e fiáveis para investigação médica, como:

- CDC WONDER (wonder.cdc.gov)

- UK Biobank (ukbiobank.ac.uk)
- MIMIC Database (mimic.mit.edu)
- PhysioNet CharisDB (physionet.org)

Alguns destes repositórios exigem pedidos formais de acesso específicos para investigação, os quais passam por processos de aprovação ética e legal que devido à complexidade do processo, pode ser demorado. Dado o tempo limitado disponível para o desenvolvimento deste projeto, optou-se por utilizar dados de acesso imediato e público, assegurando o cumprimento dos prazos sem comprometer a qualidade científica do trabalho.

A aplicação móvel, embora ainda não disponível ao público, já obteve aprovações da comissão de ética do Infarmed. Estas aprovações conferem-lhe o estatuto de dispositivo médico, tornando-a uma fonte de informação fidedigna para uso em ambiente hospitalar. Este estatuto oferece uma base sólida para a futura integração do modelo proposto, oportunidade que se encontra fora do âmbito deste TFC, mas que representa um caminho promissor para trabalhos futuros, eliminando a necessidade de desenvolver uma infraestrutura nova.

Viabilidade Económica

O desenvolvimento do projeto é viável de um ponto de vista económico, uma vez que:

- Não existem custos associados a licenças de software ou plataformas, dado que todas as tecnologias utilizadas são de livre acesso (open-source) ou estão disponíveis gratuitamente através de parcerias e acordos estabelecidos com a universidade.
- O dataset utilizado é público e acessível para fins de investigação, não implicando custos adicionais.

Em comparação com os métodos tradicionais de diagnóstico de arritmias, como o eletrocardiograma que exige um Eletrocardiógrafos (dispositivo médico utilizado para realizar um ECG), infraestruturas hospitalares e a supervisão de profissionais de saúde, a solução proposta não tem custos adicionais para o utilizador final. No setor público, o custo de um ECG pode ser de apenas 1,40€, contudo, existe a incerteza quanto aos tempos de espera para a marcação do exame e o seguimento subsequente. Em contraste, no setor privado, os custos de um ECG variam significativamente: 23€ na CUF, 44,60€ no Hospital da Luz, e entre 35€ e 40€ no Hospital da Cruz Vermelha. A deteção de arritmias em casa via smartphone é essencialmente gratuita para o utilizador, aproveitando um dispositivo que já possui, e elimina por completo os tempos de espera associados aos sistemas de saúde tradicionais [12] [13] [14] [15].

Assim, o projeto demonstra ser financeiramente sustentável e poderá ser continuado sem a necessidade de financiamento externo. Adicionalmente, esta abordagem tem o potencial de gerar benefícios económicos indiretos significativos. A identificação precoce de arritmias e a consequente intervenção atempada podem levar a uma redução drástica de internamentos hospitalares, visitas a urgências e diagnósticos tardios de condições cardíacas graves, resultando em poupanças substanciais para os sistemas de saúde e melhoria da qualidade de vida dos pacientes.

Viabilidade Social

A viabilidade social deste projeto está no seu impacto positivo para pacientes com problemas cardíacos. A solução proposta é uma ferramenta acessível e prática, terá em conta a sua usabilidade, sendo desenvolvida a pensar num vasto leque de utilizadores, incluindo idosos e indivíduos de diversos extratos socioeconómicos.

Esta abordagem democratiza a monitorização cardíaca, tornando-a inclusiva e contínua, e é particularmente benéfica para aqueles com dificuldades de acesso a dispositivos médicos tradicionais e a um diagnóstico precoce. O seu impacto é especialmente relevante em zonas remotas ou rurais, onde a escassez de cardiologistas e infraestruturas de saúde especializadas limita drasticamente o acesso a cuidados. Ao permitir que a monitorização ocorra de forma remota, a aplicação preenche uma lacuna crucial nestes contextos.

Além da deteção, a aplicação tem um grande potencial para a educação em saúde. Poderá fornecer alertas personalizados, explicações claras sobre os dados recolhidos e recomendações baseadas em evidências, capacitando os utilizadores a gerir melhor a sua condição e a adotar hábitos mais saudáveis, promovendo assim um maior controlo sobre a sua própria saúde.

3 Conceitos Fundamentais

3.1 Conceitos Teóricos

3.1.1 Coração

O coração é o órgão central do sistema cardiovascular, responsável por bombear o sangue por todo o corpo, garantindo que os tecidos e órgãos recebam o oxigénio e os nutrientes de que necessitam para funcionar corretamente, ao mesmo tempo que remove dióxido de carbono e outros resíduos para serem eliminados por outros órgãos. Além disso, o coração controla o ritmo e a velocidade dos batimentos cardíacos e mantém a pressão arterial.

O funcionamento do coração, exemplificado na Figura 1, assenta num ciclo contínuo de circulação do sangue através das suas quatro câmaras, dois átrios (superiores) e dois ventrículos (inferiores), e num sistema de válvulas que assegura a direção correta do fluxo sanguíneo [16].

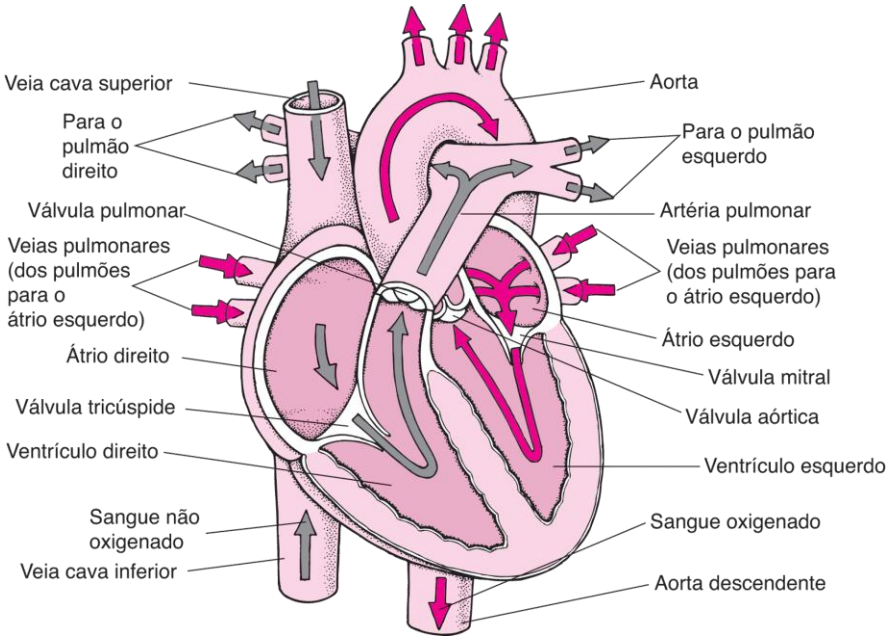

Figura 1 - Estrutura interna do coração e circulação do sangue (fonte: [17])

O lado direito do coração recebe sangue pobre em oxigénio proveniente do corpo, através das veias cava superior e inferior, conduzindo-o ao átrio direito. Este sangue é então transportado para o ventrículo direito ao passar pela válvula tricúspide. Durante a contração ventricular, o sangue é impulsionado pela válvula pulmonar para a artéria pulmonar, que o conduz aos pulmões para oxigenação. Na Figura 1, este percurso está representado pelas setas a cinzento, que indicam o trajeto do sangue não oxigenado.

Após a troca gasosa nos pulmões, o sangue oxigenado retorna ao coração pelas veias pulmonares, entrando no átrio esquerdo. De seguida, atravessa a válvula mitral em direção ao ventrículo esquerdo, a câmara com maior força de contração, que impulsiona o sangue pela válvula aórtica para a aorta, permitindo a sua distribuição por todo o organismo. Este trajeto do sangue oxigenado está representado na Figura 1 pelas setas cor-de-rosa, que mostram o fluxo do sangue rico em oxigénio.

Este mecanismo de bombeamento unidirecional é garantido pelo funcionamento sincronizado das válvulas cardíacas (tricúspide, pulmonar, mitral e aórtica), que evitam o refluxo e promovem a eficiência do ciclo cardíaco.

Este ciclo é coordenado por um sistema elétrico interno, responsável pela geração e propagação de impulsos elétricos que controlam a contração do coração, assegurando um ritmo estável e eficaz.

3.1.2 Eletrocardiograma

O eletrocardiograma (ECG) é um exame que regista o ritmo e a atividade elétrica do coração [18]. O traçado do ECG é composto por ondas que refletem diferentes fases do ciclo cardíaco, como demonstrado na Figura 2 [19]:

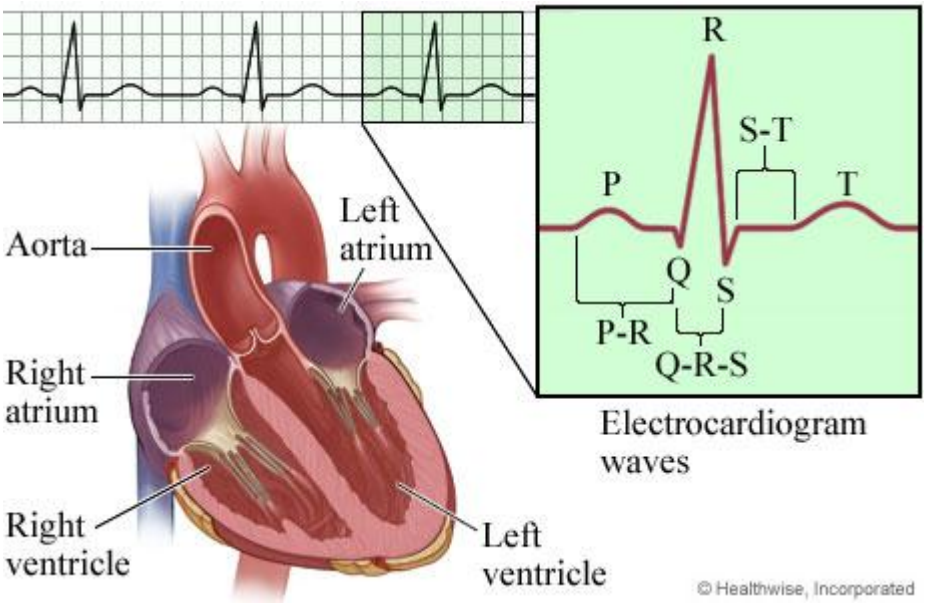

Figura 2 - Funcionamento ECG (fonte: [19])

- **Onda P:** Registo da atividade elétrica nas câmaras superiores do coração (aurículas).
- **Complexo QRS:** Registo do movimento dos impulsos elétricos através das câmaras inferiores do coração (ventrículos).
- **Segmento ST:**
  - Representa o momento em que o ventrículo está a contrair-se, mas sem fluxo de eletricidade.
  - Geralmente aparece como uma linha reta e nivelada entre o complexo QRS e a onda T.
- **Onda T:** Mostra quando as câmaras inferiores do coração (ventrículos) estão a reajustar-se eletricamente e a preparar-se para a próxima contração muscular.

Dentro do complexo QRS destaca-se o pico R, um ponto máximo positivo que corresponde ao momento em que o impulso elétrico atinge o seu auge nos ventrículos.

A deteção precisa do pico R é essencial para a identificação dos batimentos cardíacos. É possível concluir através da Figura 3 que a partir da posição temporal dos picos R, é possível calcular o intervalo RR, definido como o tempo entre dois batimentos consecutivos.

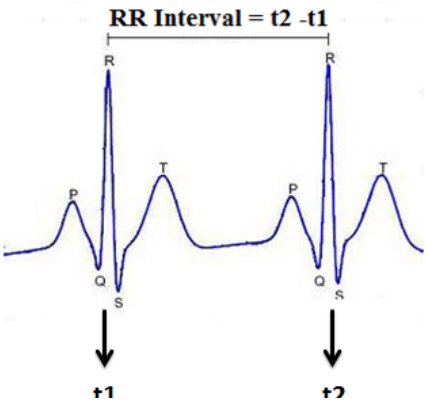

Figura 3 - Exemplo de Intervalo RR (fonte: [14])

A análise dos intervalos RR fornece métricas fundamentais para a avaliação da variabilidade da frequência cardíaca (HRV), esta variabilidade é um dos indicadores principais de arritmias [15].

Desta forma, a deteção dos picos R e a análise subsequente dos intervalos RR constituem ferramentas centrais na identificação precoce de disfunções cardíacas.

3.1.3 Fotopletismografia

A Fotoplestimografia (PPG) é uma técnica não invasiva utilizada para medir variações no volume sanguíneo na superfície da pele. Este método baseia-se na emissão de luz por um sensor e na deteção da quantidade de luz refletida através dos tecidos [22].

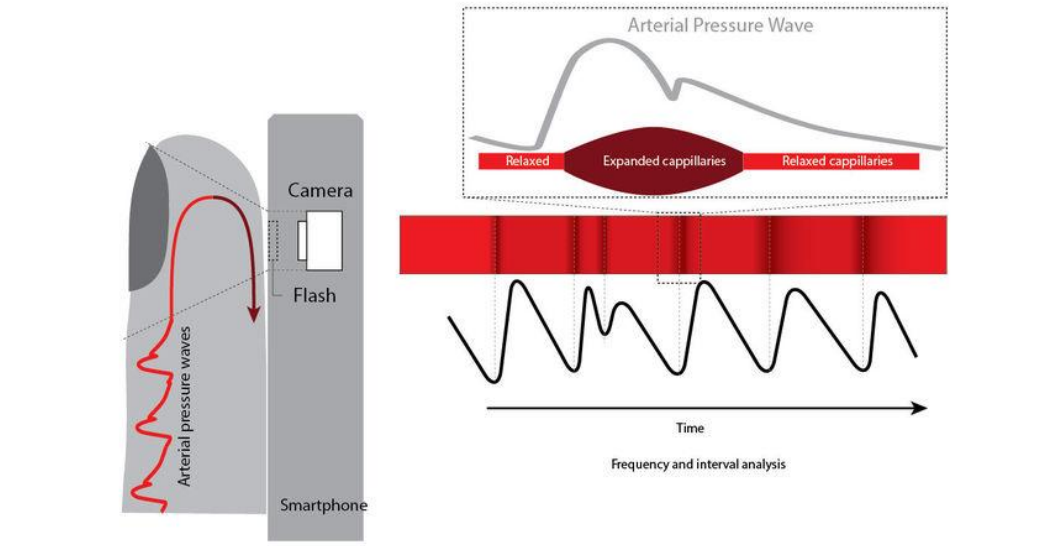

Figura 4 - Funcionamento PPG (fonte: [23])

O sinal PPG apresenta um sinal que reflete o ciclo de contração e relaxamento do coração, permitindo a extração de informações como a frequência cardíaca, a HRV e, em alguns casos, a saturação de oxigénio [24].

Esta tecnologia é utilizada em dispositivos de monitorização contínua, como smartwatches, pulseiras de fitness e oxímetros de pulso, oferecendo uma solução acessível e cómoda para a recolha de dados fisiológicos fora do ambiente hospitalar.

No âmbito deste trabalho esta tecnologia seria utilizada pela aplicação para recolher as métricas essenciais para identificar os casos com ritmos irregulares.

3.1.4 Arritmias

As arritmias cardíacas são perturbações do ritmo normal do coração, que resultam em anomalias na geração ou na condução dos impulsos elétricos responsáveis pela contração cardíaca. Estas alterações, podem manifestar-se de três formas principais, representadas na Figura 5 [25]:

- Batimentos demasiado rápidos (taquicardia), caracterizados por uma frequência cardíaca superior a 100 batimentos por minuto (BPM);
- Batimentos demasiado lentos (bradicardia), com frequência cardíaca inferior a 60 BPM;
- Batimentos irregulares, com variação inconsistente no intervalo entre os batimentos.

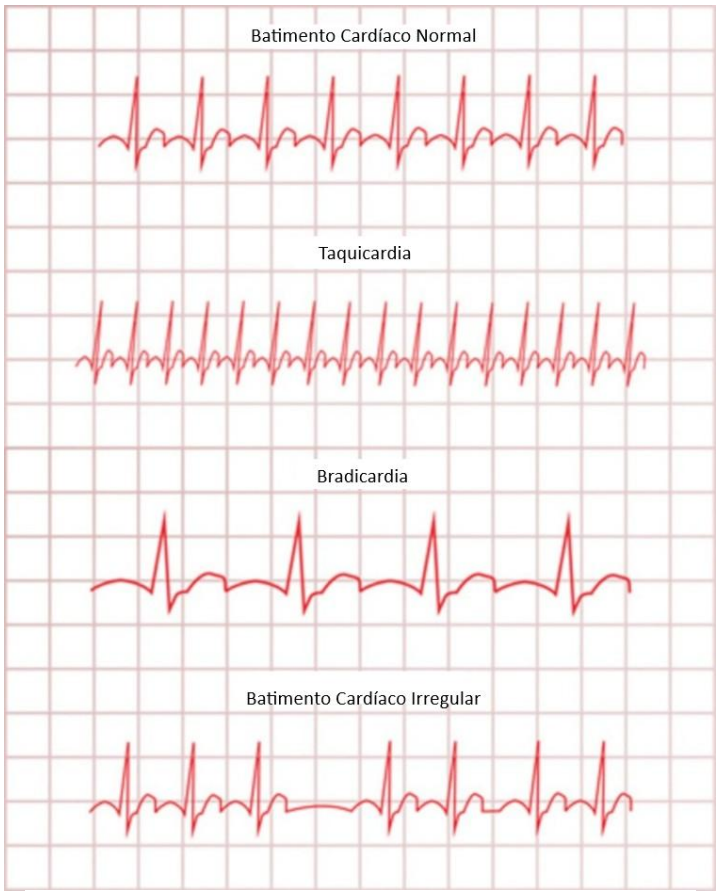

Figura 5 - Comparação de ritmos (adaptada, fonte: [26])

As arritmias podem comprometer a eficiência da contração cardíaca, prejudicando o fornecimento de sangue aos tecidos e aumentando o risco de eventos clínicos graves, como acidentes vasculares cerebrais, insuficiência cardíaca ou morte súbita.

3.1.5 Arritmias com Batimentos Irregulares

Este tipo de arritmia pode surgir em diferentes zonas do coração e inclui tanto alterações benignas como condições clínicas mais sérias.

De seguida, são apresentadas as principais arritmias com este padrão identificadas no âmbito do projeto.

Fibrilhação Auricular (FA) - *Atrial Fibrillation (AFIB)*:

A Fibrilhação Auricular é o tipo mais comum de arritmia cardíaca. Esta condição pode comprometer significativamente o funcionamento cardíaco, aumentando o risco de acidente vascular cerebral, insuficiência cardíaca e outras complicações graves [27].

Nas arritmias com FA, as ondas P consistentes são substituídas por ondas de fibrilhação, que variam em amplitude, forma e "timing" [28]. Estas alterações podem ser identificadas no traçado de ECGs e de PPGs através dos intervalos RR como demonstrado na Figura 6.

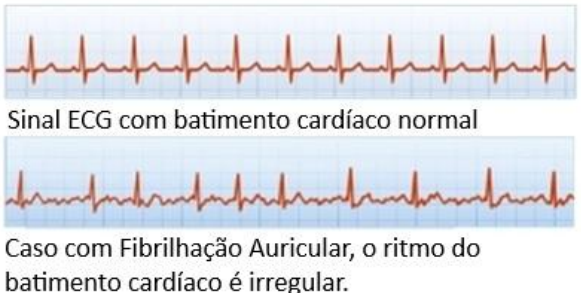

Figura 6 - ECG normal e com FA (adaptada, fonte: [28])

Flutter Atrial - *Atrial Flutter (AF)*:

Atrial Flutter é uma arritmia cardíaca caracterizada por contrações rápidas e regulares das aurículas, geralmente com uma frequência atrial entre próxima dos 300 BPM. Este ritmo origina-se normalmente por um circuito elétrico reentrante no átrio direito, criando um padrão em "dente de serra" visível no ECG [29].

A Figura 7 mostra um exemplo da irregularidade dos intervalos RR, marcada pela ausência da onda P e pela variação dos intervalos entre os batimentos ventriculares.

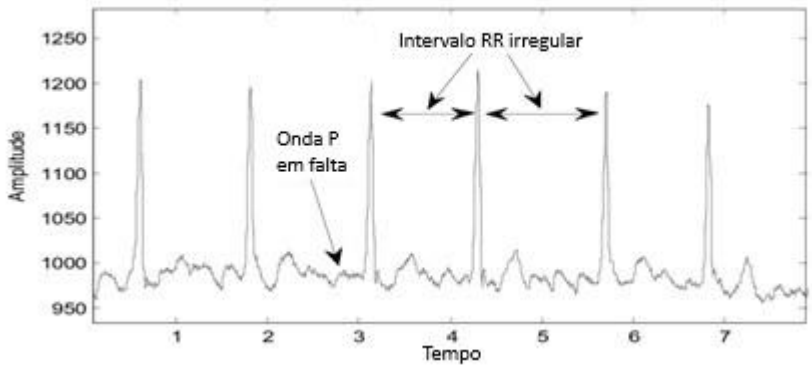

Figura 7 - ECG com AF (adaptada, fonte: [30])

Arritmia Sinusal - *Sinus Irregularity (SA)*:

A SA é um tipo de arritmia benigna caracterizada por variações no intervalo entre os batimentos cardíacos, normalmente relacionadas com o ciclo respiratório. Durante a inspiração, os batimentos tendem a acelerar, e durante a expiração, abrandam, o que resulta em intervalos RR irregulares [31].

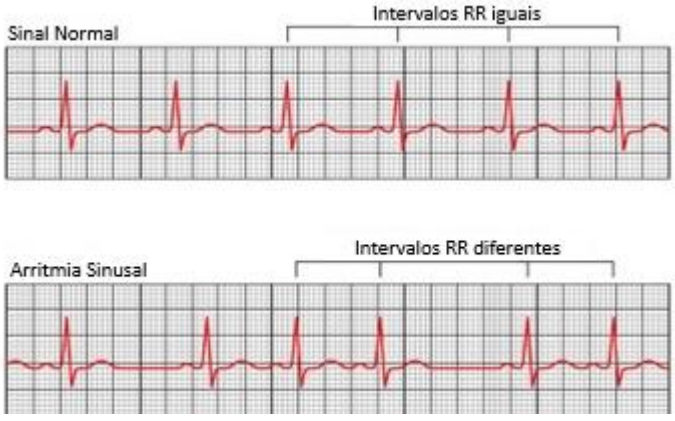

Figura 8 - Normal Sinus Rhythm e SA (adaptada, fonte: [31])

Como mostra a Figura 8, o ritmo sinusal normal apresenta intervalos regulares entre os batimentos, enquanto que na SA esses intervalos variam. Apesar de ser considerada uma arritmia, não representa perigo e é comum em pessoas jovens e saudáveis, sendo um reflexo normal da interação entre o sistema nervoso autónomo e o coração.

Ritmo Atrial Migratório do Nodo Sinusal – *Sinus Atrium to Atrial Wandering Rhythm (SAAWR)*:
O Sinus Atrium to Atrial Wandering Rhythm, também conhecido como Wandering Atrial Pacemaker, é uma arritmia benigna caracterizada por uma variação no local de origem dos impulsos elétricos nos átrios, resultando em ondas P com morfologias diferentes num mesmo traçado ECG [32].
Na Figura 9, é possível observar essa diferença: enquanto as ondas P de origem sinusal mantêm um aspeto consistente, as ectópicas (fora do nó sinusal) variam na forma e direção.

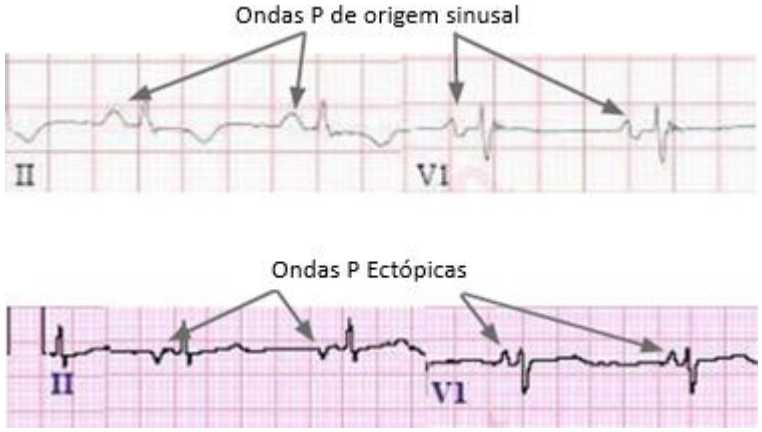

Figura 9 - Comparação ondas P sinusal e Ectópicas (adaptada, fonte: [32])

Contração Atrial prematura - *Atrial Premature Beat (APB)*:
O Atrial Premature Beat, é identificado quando um batimento acontece mais cedo do que o normal e vem de uma parte diferente dos átrios, fora do nó sinusal. Este batimento extra pode alterar o ritmo normal do coração, causando uma pausa ligeiramente maior logo a seguir. Apesar de muitas vezes ser inofensivo e sem sintomas, pode provocar sensação de batimentos "falhados" ou palpitações, e está associado a fatores como cansaço, stress ou consumo de cafeína [33].
Na Figura 10, estão identificados vários batimentos (marcados com setas) que surgem mais cedo do que os outros, e tem uma forma diferente, indicando que tiveram origem noutros pontos dos átrios. Depois destes batimentos prematuros, o coração faz pausas ligeiramente mais longas antes de voltar ao ritmo normal.

Figura 10 - Exemplo de APB (adaptada, fonte: [33])

Contração Ventricular Prematura - *Ventricular Premature Beat (VPB)*:

O *Ventricular Premature Beat* (VPB), também conhecido como *Premature Ventricular Contraction* (PVC), é um batimento cardíaco extra que se origina nos ventrículos antes do tempo esperado. Estes batimentos não seguem o ritmo normal do coração e não passam pelo nó sinusal, o que os torna visivelmente diferentes num ECG. Apesar de geralmente serem benignos em pessoas saudáveis, podem estar associados a doenças cardíacas se forem frequentes ou sintomáticos [34].

Na Figura 11, existem dois casos de VPB, que surgem mais cedo do que o ritmo regular. Estes batimentos apresentam morfologia alargada e diferente, indicando a sua origem ventricular. Após cada VPB, ocorre uma "pausa compensatória", um intervalo mais longo até o próximo batimento normal, que restabelece o ritmo do coração.

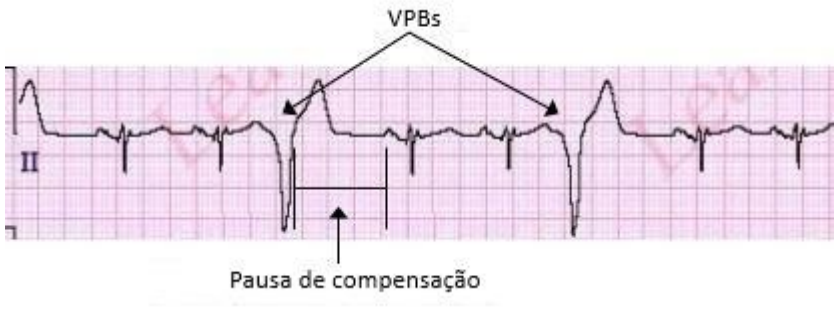

Figura 11 - Exemplo de VPB (adaptada, fonte: [34])

Bigeminismo Atrial - *Atrial Bigeminy (ABI)*:

O Bigeminismo Atrial é um tipo de arritmia em que cada batimento normal é seguido por um batimento atrial prematuro. Embora possa ocorrer em indivíduos saudáveis, também pode estar associado a stress, uso de estimulantes ou doenças cardíacas subjacentes [35].

Na Figura 12, é possível observar um traçado típico de ABI. Cada batimento normal é seguido por um batimento ectópico atrial, criando um padrão facilmente identificável.

Figura 12 - Exemplo de ABI (fonte: [35])

Bigeminismo Ventricular - *Ventricular Bigeminy (VB)*:

O Bigeminismo Ventricular é uma arritmia em que cada batimento normal do coração é seguido por um batimento ectópico de origem ventricular, conhecido como contração ventricular prematura (VPB). Este padrão alternado entre batimento normal e VPB pode indicar irritabilidade ventricular e pode ser causado por fatores como isquemia, desequilíbrios eletrolíticos ou efeito de medicamentos [36].

Na figura 13, é possível observar um traçado com padrão típico de VB. Os batimentos normais alternam-se com batimentos ventriculares prematuros e ausência de onda P. Este padrão cria uma irregularidade nos intervalos RR.

Figura 13 - Exemplo de VB (fonte: [36])

Junctional Escape Beat (JEB):

O *Junctional Escape Beat* é um batimento de substituição, feito a partir do nó atrioventricular, geralmente quando o nódulo sinusal falha ou atrasa. Atua como um mecanismo de segurança do coração para evitar longas pausas no ritmo [37].

Na figura 14, observa-se um ritmo de escape juncional, é evidente o papel de segurança desempenhado pelo nó atrioventricular quando a atividade sinusal falha.

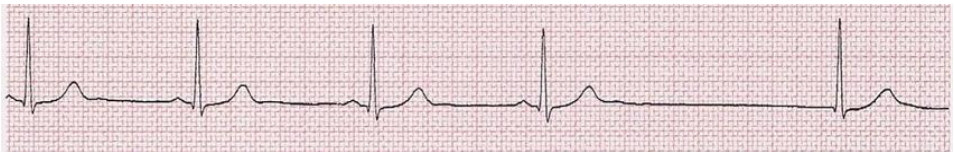

Figura 14 - Exemplo de JEB (fonte: [37])

Junctional Premature Beat (JPT):

O Junctional Premature Beat é um batimento que tem uma origem precoce na junção atrioventricular, antes do batimento sinusal esperado. Frequentemente, está associado a estimulação aumentada do nó atrioventricular, podendo ocorrer em pessoas saudáveis, mas também em contextos de irritação miocárdica, como em casos de isquemia ou uso de certos medicamentos [38].

No sinal apresentado na Figura 15, o quinto batimento é o JPT, identificável pela sua ocorrência precoce em relação ao ritmo regular.

Figura 15 - Exemplo de JPT (fonte: [38])

***Ventricular Escape Beat (VEB)*:**

O Ventricular Escape Beat é um batimento de "recurso" gerado nos ventrículos quando o ritmo cardíaco normal falha em produzir um impulso atempadamente. É uma resposta protetora do coração e pode ocorrer em situações de bloqueio AV completo, bradicardia severa ou paragem sinusal [39].

Na Figura 16, o sexto batimento é um VEB, aconteceu depois de uma pausa prolongada, destacando-se do ritmo regular até ao momento.

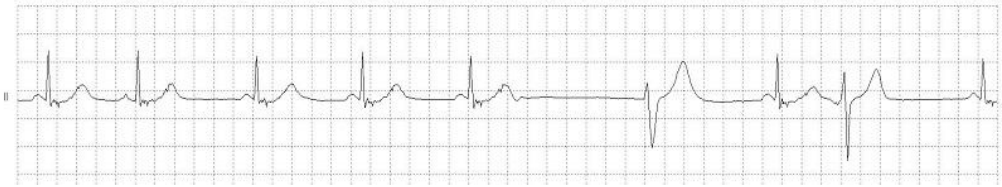

Figura 16 - Exemplo de VEB (fonte: [39])

Trigeminismo de Escape Ventricular - *Ventricular Escape Trigeminy (VET)*:

O *Ventricular Escape Trigeminy* é um padrão rítmico em que um batimento ventricular ectópico (geralmente um PVC) ocorre a cada três batimentos cardíacos. Este fenómeno representa uma resposta do ventrículo, frequentemente em situações em que os impulsos normais estão comprometidos ou há uma supressão do ritmo sinusal. O padrão é cíclico e pode indicar instabilidade elétrica do coração [40].

Na Figura 17, é possível observar a ocorrência de um complexo ventricular prematuro (etiquetado como "PVC") a cada terceiro batimento, formando o padrão típico de trigeminy.

Figura 17 - Exemplo de VET (fonte: [40])

3.2 Modelos e Algoritmos Relevantes

Para o desenvolvimento deste trabalho, serão considerados e comparados diferentes modelos e algoritmos, selecionados com base na sua adequação à natureza dos dados e ao problema a ser resolvido.

Os algoritmos propostos para este trabalho incluem a Regressão Logística, *SVM*, *Random Forest* e *Gradient Boosting*, cada um com características distintas e aplicações específicas na classificação e análise de dados.

3.2.1 Regressão Logística

A Regressão Logística é um modelo linear utilizado para resolver problemas de classificação binária, como distinguir entre ritmos cardíacos normais e anormais. [41].
O modelo estima a probabilidade de uma amostra pertencer à classe positiva (y = 1) ao aplicar uma função logística (sigmoide) sobre uma combinação linear das variáveis de entrada, resultando numa curva em forma de "S", como ilustrado na Figura 18.

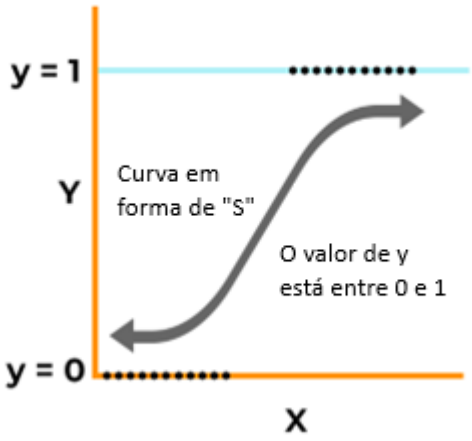

Figura 18 - Curva Regressão Logística (adaptada, Fonte: [41])

Apesar da sua simplicidade, a regressão logística serve como uma linha de base confiável para comparação com algoritmos mais complexos.

3.2.2 Support Vector Machines (SVM)

O SVM ou *Support Vector Classifier* (SVC) é um modelo eficaz na identificação de padrões complexos, mesmo em contextos com dados de alta dimensionalidade [42].
Como ilustrado na Figura 19, o SVM procura encontrar o hiperplano ótimo (linha a tracejado) que separa duas classes, neste caso, representadas por círculos e cruzes, maximizando a distância entre os pontos mais próximos de cada classe (os chamados vetores de suporte, destacados com contornos). Esta maximização da margem ajuda a melhorar a generalização do modelo, reduzindo o risco de *overfitting* (casos em que o modelo se ajusta demasiado bem aos dados de treino, que acaba por perder desempenho ao lidar com dados novos).

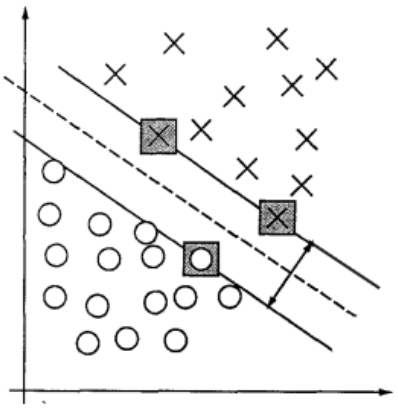

Figura 19 - Exemplo de SVM (fonte: [42])

3.2.3 Random Forest

O *Random Forest* baseia-se na construção de múltiplas árvores de decisão, combinando os seus resultados para melhorar a precisão da classificação [43].

Como ilustrado na Figura 20, cada árvore é treinada com subconjuntos diferentes dos dados, e quando um novo exemplo (ponto roxo) precisa de ser classificado, é passado por todas as árvores. Cada árvore dá a sua previsão (neste caso seria classe verde ou cinzenta), e a decisão final é feita através de uma votação maioritária ou média, resultando numa previsão mais robusta do que se fosse apenas classificado por uma árvore.

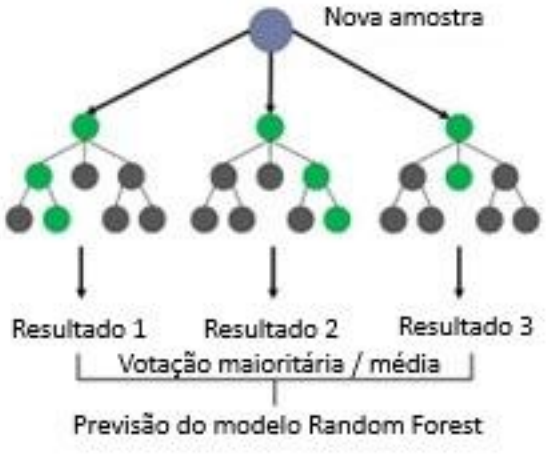

Figura 20 - Funcionamento *Random Forest* (adaptada, Fonte: [44])

3.2.4 Gradient Boosting

O *Gradient Boosting* é uma abordagem de aprendizagem por reforço sequencial, onde cada novo classificador é treinado para corrigir os erros cometidos pelos modelos anteriores [45].

Na Figura 21, é possível ver como o processo começa com um conjunto de dados original e, a cada iteração, os pesos dos dados são ajustados para dar mais importância aos exemplos que foram mal classificados.

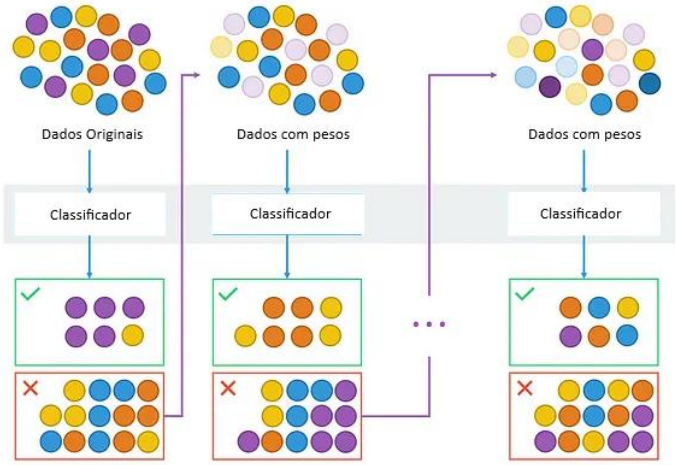

Figura 21 - Funcionamento *Gradient Boosting* (adaptada, Fonte: [46])

Neste projeto, foi utilizado o *XGBoost (Extreme Gradient Boosting)*, uma implementação otimizada do *Gradient Boosting*, conhecida pelo seu elevado desempenho e eficiência computacional. Ao contrário de métodos como o *Random Forest*, que constroem árvores em paralelo, o *XGBoost* constrói-as de forma encadeada, tornando-o especialmente eficaz em conjuntos de dados desafiantes com padrões complexos e desequilíbrios entre as classes.

3.3 Métricas de Avaliação de Modelos de Machine Learning

Para avaliar o desempenho dos modelos desenvolvidos, foram utilizadas as métricas:

- **Precisão - *Precision***: mede a proporção de casos verdadeiramente positivos entre todas as previsões positivas feitas pelo modelo. Esta métrica é útil para perceber quantos dos diagnósticos positivos identificados pelo o modelo estão realmente corretos. A precisão é calculada pela seguinte fórmula:

$$Precision = \frac{TP}{(TP + FP)}$$

  Onde TP são os verdadeiros positivos e FP os falsos positivos. Um valor alto de precisão significa que o modelo é confiável nas suas previsões positivas [47].
- **Sensibilidade - *Recall***: indica a capacidade do modelo em identificar corretamente todos os casos positivos reais. No contexto da deteção de arritmias, o Recall mostra quantos dos pacientes com arritmia foram corretamente detetados. A sensibilidade é calculada pela fórmula:

$$Recall = \frac{TP}{(TP + FN)}$$

  Sendo FN os falsos negativos. Uma sensibilidade alta garante que poucos pacientes doentes são ignorados, sendo, portanto, prioritário em sistemas de apoio à decisão clínica onde a sensibilidade do rastreio é essencial. [47].
- **Especificidade - *Specificty:*** mede a proporção de casos verdadeiramente negativos entre todos os casos negativos reais. Ou seja, quantos dos pacientes que não têm arritmia foram corretamente identificados como negativos pelo modelo. Esta métrica é particularmente importante em contextos clínicos onde é necessário evitar alarmes falsos, assegurando que os pacientes saudáveis não são classificados incorretamente como doentes. A especificidade é calculada como:

$$Specificity = \frac{TN}{(TN + FP)}$$

  Onde TN representa os verdadeiros negativos e FP os falsos positivos [48].
- **Acurácia - *Accuracy*:** indica a proporção total de previsões corretas (tanto positivas como negativas) entre todas as previsões realizadas pelo modelo. Em outras palavras, mede a capacidade geral do modelo em classificar corretamente os casos. No entanto, em conjuntos de dados desequilibrados, onde uma classe é muito mais frequente do que a outra, a acurácia pode ser enganadora. A fórmula é:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

  Sendo TP os verdadeiros positivos, TN os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos [49].
- **F1-Score:** representa o equilíbrio entre a precisão e a sensibilidade, sendo a média "harmónica" destas duas métricas, penalizando fortemente valores muito baixos em qualquer uma das duas métricas. Esta métrica é especialmente relevante quando é necessário encontrar um compromisso entre evitar falsos positivos e não falhar casos verdadeiros, como ocorre na deteção de condições clínicas sensíveis. A sua fórmula é:

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

  Um valor alto de F1-Score indica que o modelo tem bom desempenho tanto em identificar corretamente os casos positivos como em evitar classificações incorretas [47].
- **AUC (Área Sob a Curva ROC):** uma medida que quantifica a capacidade de um modelo de classificação binária em distinguir corretamente entre as classes positiva e negativa. A curva ROC (*Receiver Operating Characteristic*) representa diferentes combinações de *True Positive Rate* (sensibilidade) e *False Positive Rate* (1 - especificidade) ao variar o limiar de decisão. A AUC corresponde à área total sob essa curva, e seu valor varia entre 0,5 (equivalente a classificar ao acaso) e 1,0 (modelo perfeito). Um valor mais próximo de 1 indica melhor desempenho do modelo em separar as classes [50].
- **Matriz de Confusão:** A matriz de confusão é uma ferramenta fundamental na avaliação do desempenho de modelos de classificação, especialmente em problemas de classificação binária como a deteção de arritmias. Esta matriz organiza os resultados das previsões do modelo em quatro categorias principais, permitindo analisar detalhadamente onde o modelo acerta e onde falha [51]. Na Figura 22 é demonstrado

um exemplo da Matriz de confusão no diagnóstico médico binário (doente e saudável).

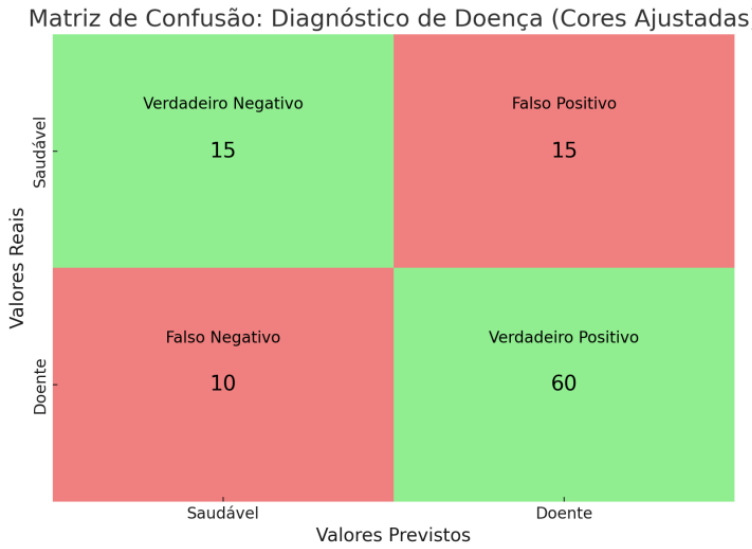

Figura 22 - Matriz de confusão (fonte: [51])

A matriz está dividida em quatro quadrantes:

- Verdadeiros Positivos - 60 casos: representam os pacientes que efetivamente têm a doença e foram corretamente identificados como doentes pelo modelo.
- Falsos Positivos - 15 casos: são os pacientes saudáveis que foram incorretamente classificados como doentes.
- Falsos Negativos - 10 casos: referem-se aos pacientes com a doença que o modelo classificou incorretamente como saudáveis, o que pode representar uma falha crítica em contextos clínicos.
- Verdadeiros Negativos - 15 casos: correspondem aos pacientes saudáveis que foram corretamente identificados como tal.

Assim, a matriz de confusão é uma ferramenta indispensável na interpretação e validação de modelos de ML aplicados à saúde, permitindo não só quantificar a performance geral, mas também identificar pontos críticos de melhoria do modelo.

3.4 Tecnologias e Ferramentas Utilizadas

Python (Linguagem de Programação - Desenvolvimento do Modelo de ML)

Python foi escolhido pela sua simplicidade e eficiência, sendo amplamente utilizado para projetos de machine learning graças à grande quantidade de bibliotecas e recursos disponíveis.

Todas as bibliotecas foram obtidas através do repositório oficial PyPi [52].

Bibliotecas principais Python:

- **Numpy:** Cálculos matemáticos e manipulação de arrays.
- **Pandas:** Análise e manipulação de dados estruturados.
- **Matplotlib:** Criação de gráficos e visualizações dos dados.
- **Scikit-learn (Sklearn):** Biblioteca essencial para a criação de modelos de Machine Learning, divisão de dados (como train_test_split), métricas de avaliação (ex: F1-score e classification_report) e para a visualização da matriz de confusão com confusion_matrix e ConfusionMatrixDisplay.
- **XGBoost:** Biblioteca utilizada na criação de modelos de ML de alto desempenho.
- **WFDB:** Leitura e processamento de sinais de ECG no formato dos datasets provenientes do repositório PhysioNet.

- **SciPy:** Biblioteca usada tanto para processamento de sinais (como deteção de picos com scipy.signal) como para cálculo de estatísticas (como skewness e kurtosis com scipy.stats) aplicadas aos intervalos RR.
- **Systole-Detectors**: Algoritmos clássicos de deteção de picos, em especial o algoritmo de Pan-Tompkins.
- **NeuroKit2:** Extração automática dos picos R dos sinais de ECG.
- **Joblib**: Guardar e carregar modelos de machine learning de forma eficiente e persistente.
- **Warnings:** Biblioteca que permite suprimir ou controlar mensagens de aviso durante execução de código.

PyCharm (IDE - Integrated Development Environment)

O PyCharm foi utilizado como ambiente de desenvolvimento, devido à sua facilidade de uso e suporte avançado para programação em Python, incluindo ferramentas de debugging e gestão de bibliotecas.

Flutter (Futura Integração do modelo na Aplicação Móvel)

A aplicação móvel desenvolvida no âmbito do projeto AIMhealth foi construída utilizando a tecnologia Flutter. Esta escolha estratégica permite o desenvolvimento multiplataforma, possibilitando a criação de aplicações nativas para Android e iOS a partir de uma única base de código, reduzindo o esforço de manutenção e desenvolvimento.

Tendo em conta esta base tecnológica, a futura integração do modelo de Machine Learning deverá ser realizada utilizando o ecossistema do Flutter. Esta abordagem garante compatibilidade com a arquitetura existente da aplicação e aproveita as vantagens da performance nativa e da crescente popularidade da framework no mercado de desenvolvimento móvel.

4 Estado da Arte

4.1 Estado da Arte

Deteção de Arritmias com Fotopletismografia e Machine Learning

Voisin et al. (2018) [53] desenvolveram um algoritmo capaz de identificar episódios de FA utilizando sinais de PPG obtidos por "Wearables" em condições ambulatórias. O modelo, baseado numa rede neural convolucional de 50 camadas, alcançou uma área sob a curva (AUC) de 95%, demonstrando robustez face a artefactos de movimento inerentes aos sinais PPG.

Whiting et al. (2018) [54] propuseram um método automático para reconhecer anomalias cardíacas em sinais PPG utilizando uma rede neural recorrente do tipo "Long Short-Term Memory". Treinado com 400.000 amostras de PPG, o modelo identificou com sucesso anomalias que correspondem a casos de FA, sem necessidade de eletrocardiograma.

Bulut et al. (2025) [55]: Propuseram um modelo de "Deep Convolutional Neural Network" para a deteção de distúrbios do ritmo cardíaco usando sinais de PPG de dispositivos wearable. Os resultados mostraram um F1-score de 0.94, precisão de 0.93, sensibilidade de 0.95 e acurácia de 0.94 para a classificação de arritmias como FA, contrações auriculares prematuras, e ritmo sinusal normal.

Além dos avanços com sinais de PPG, Qin et al. (2017) [56] demonstraram a possibilidade de identificar picos R diretamente a partir de gráficos de sinais ECG. O estudo apresentou métodos automáticos para extrair as localizações dos complexos QRS, permitindo posteriormente calcular métricas relevantes como a variabilidade dos intervalos RR, fundamentais para a deteção de arritmias.

Estes estudos reforçam a viabilidade de aplicar métodos automáticos, baseados em Machine Learning, tanto em sinais de PPG como em ECG, para a deteção de irregularidades cardíacas utilizando dispositivos móveis ou sistemas de monitorização remota.

Aplicações Móveis para Monitorização Cardíaca

Atualmente, existem várias aplicações disponíveis para monitorizar a saúde cardíaca. Durante o estudo, foram testadas as seguintes opções:

- Cardiio - Cardiio: Heart Rate Monitor on the App Store
- Heartify - Heartify: Heart Health Monitor on the App Store
- Instant Heart Rate: HR Monitor - Instant Heart Rate: HR Monitor on the App Store

Embora sejam ferramentas úteis, todas as aplicações apresentaram barreiras de acesso, exigindo pagamento para desbloquear funcionalidades completas, o que pode limitar a sua utilização. Além disso, antes de realizar a primeira medição, todas exibiram um aviso semelhante ao presente na Figura 23:

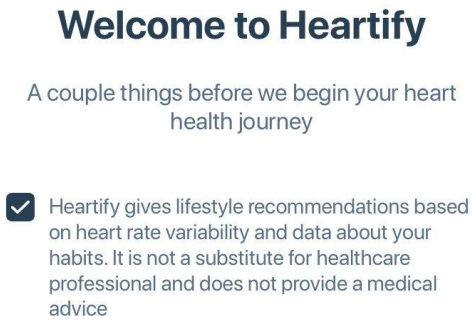

Figura 23 - Aviso app "Heartify"

Muitas aplicações no âmbito da saúde incluem avisos indicativos que não são dispositivos médicos e que as informações fornecidas têm fins exclusivamente educativos ou de orientação para hábitos de vida saudáveis. Esses avisos refletem a ausência de certificação regulatória necessária para dispositivos médicos e limitam a aplicação clínica direta dessas ferramentas. No entanto, o presente projeto diferencia-se ao já possuir autorização do Infarmed

para a utilização da aplicação como um dispositivo médico validado. Esta certificação garante que o projeto cumpre os requisitos técnicos e regulamentares necessários para oferecer um serviço confiável na medição da frequência cardíaca e na sua utilização como ferramenta complementar no contexto clínico. Assim, a solução proposta transcende as limitações das aplicações convencionais, promovendo um impacto direto na prática médica e no diagnóstico preventivo.

4.2 Proposta de inovação e mais-valias

A solução proposta destaca-se pela sua capacidade de detetar diferentes tipos de arritmias cardíacas com base na análise da variabilidade dos intervalos RR, utilizando modelos de Machine Learning.

Adicionalmente, este trabalho contribui diretamente para a evolução do projeto AIMHealth. A solução desenvolvida pode ser adaptada para que mais tarde fosse integrada na aplicação, permitindo incorporar a funcionalidade de detetar de forma automática arritmias, com potencial para funcionar com dados recolhidos diretamente pela câmara do telemóvel.

O facto de ser uma solução pensada para funcionar em smartphones, sem necessidade de dispositivos adicionais, torna-a acessível a uma grande parte da população. Combinando acessibilidade, fiabilidade e possibilidade de uso clínico, esta abordagem representa um passo significativo para a monitorização cardíaca remota.

5 Solução Proposta

5.1 Introdução

A solução desenvolvida está no seguinte repositório Git:

https://github.com/DEISI-ULHT-TFC-2024-25/TFC-Aluno2114-DetecaoDeArritmiaML

A Figura 24 detalha a organização dos ficheiros:

```
TFC/
├── data/
│   ├── WFDBRecords/     # Registos brutos descarregados da PhysioNet
│   ├── processed/       # Dados transformados (pickles, .csv, etc.)
├── df_ml/               # DataFrames finais prontos para treino/teste
├── models/              # Modelos treinados (.joblib)
│   └── afib_only/       # Modelos específicos para Fibrilação Auricular (AFIB)
├── notebooks/           # Jupyter Notebooks de exploração e análise
├── src/
│   ├── data/            # Carregamento de dados brutos
│   ├── features/        # Extração de métricas RR/HRV
│   ├── models/          # Treino e seleção de modelos
│   ├── pipelines/       # Scripts principais de execução
│   └── legacy/          # Scripts antigos não utilizados no pipeline atual
├── .gitignore
├── README.md
```

Figura 24 - Estrutura ficheiros (fonte: repositório Git)

Para complementar a descrição do processo de desenvolvimento e treino do modelo, foi criado um vídeo demonstrativo. Este vídeo ilustra as etapas chave do código, desde o pré-processamento dos dados até à criação e avaliação do modelo de Machine Learning.

O vídeo pode ser acedido através do seguinte link: https://www.youtube.com/watch?v=NDYwvJRMCfM

5.2 Metodologia

A metodologia do projeto pode ser dividida em duas fases do trabalho, a segunda fase está dependente da parceria com o projeto AIMHealth, e seria desenvolvida em TFCs de anos futuros. Estas fases podem ser divididas em vários objetivos e estão representadas na Figura 25:

- Fase 1: Criação do Modelo de Machine Learning
  - Pesquisa conceitos médicos
    - Estudo sobre, funcionamento do coração, arritmias relevantes e interpretação de ECGs;
  - Pesquisa sobre o Estado da Arte
    - Análise de abordagens já existentes para deteção de arritmias com Machine Learning.
  - Procura e Seleção de Dataset
    - Identificação de datasets públicos;
    - Avaliação da qualidade dos dados (número de amostras, duração do sinal, frequência de amostragem e anotações sobre o diagnóstico de cada caso).
  - Análise Exploratória dos Dados
    - Visualização dos sinais;
    - Analisar distribuição dos diagnósticos
    - Verificação de missing values e formatação incorreta
  - Tratamento e Preparação dos Dados
    - Limpeza dos dados;
    - Identificação dos picos R
    - Extração de características (intervalos RR e métricas sobre a variabilidade deles)

- Criação dos Modelos
  - Escolha e parametrização dos algoritmos
  - Treino dos modelos
  - Comparar Métricas de desempenho
- Fase 2: Implementação e Integração do modelo na Aplicação Móvel

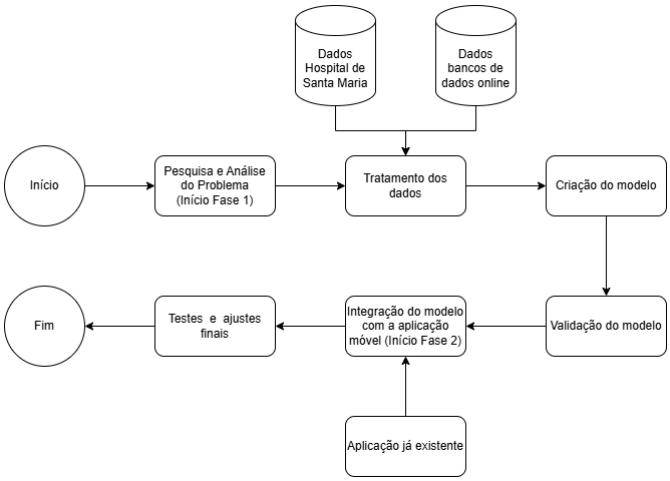

Figura 25 - Fluxo das fases do TFC

5.3 Recolha de Dados

A recolha de dados foi uma etapa fundamental no desenvolvimento do modelo. Foram avaliados vários conjuntos de dados públicos, mas muitos acabaram por ser excluídos por diferentes motivos: .

- **Restrições de acesso:** sendo dados médicos, era comum que alguns datasets exigissem um pedido formal por parte de um investigador responsável. Dado o tempo limitado para a execução deste projeto, este processo burocrático tornava-os inviáveis.
- **Falta de datasets com PPG**: inicialmente, o objetivo era usar sinais PPG. No entanto, por ser uma tecnologia mais recente, ainda existem poucos datasets públicos com dados PPG, ao contrário dos sinais ECG que já são amplamente utilizados e documentados.
- **Número insuficiente de registos**: foram encontrados alguns datasets com potencial, mas que continham poucos registos, o que dificultaria o treino eficaz dos modelos de Machine Learning.
- **Amostras demasiado curtas:** nalguns casos, a duração dos sinais era demasiado reduzida, não permitindo identificar padrões relevantes na linha do ECG nem extrair métricas fiáveis de variabilidade do ritmo cardíaco.
- **Qualidade do sinal**: alguns conjuntos de dados apresentavam sinais com muito ruído ou obtidos com baixa resolução (frequência de amostragem reduzida), o que prejudicava a deteção precisa dos picos R e comprometia a fiabilidade do modelo.
- **Ausência de diagnóstico clínico**: para treinar modelos de forma correta, era essencial que os sinais estivessem associados a diagnósticos validados por profissionais de saúde. Sem esta informação, os dados não podiam ser utilizados de forma adequada.

Para o desenvolvimento do projeto, foram utilizados dados disponibilizados na plataforma PhysioNet, em particular o conjunto de dados *A Large Scale 12-Lead Electrocardiogram Database for Arrhythmia Study* [11]. Disponível em: https://physionet.org/content/ecg-arrhythmia/1.0.0/ . Esta base de dados fornece sinais de ECG recolhidos de pacientes reais, com registos clínicos anonimizados todos eles com diagnósticos feitos por vários profissionais de saúde.

Como já foi mencionado, inicialmente, previa-se a utilização de sinais PPGs, dada a sua relevância na aplicação prática em dispositivos móveis. Contudo, durante a fase de pesquisa, constatou-se uma elevada dificuldade em encontrar bases de dados públicas que disponibilizassem sinais PPG com qualidade suficiente para a análise pretendida. Esta limitação levou à necessidade de reorientar o projeto para trabalhar com sinais de ECGs.

Em paralelo, estava prevista a utilização de dados clínicos do Hospital de Santa Maria e de sinais recolhidos pela aplicação móvel desenvolvida no âmbito do projeto AIMHealth. Contudo, não foi possível cruzar estes dados com os do dataset público. Esta situação foi acompanhada em diversas reuniões com os responsáveis pelo projeto AIMHealth e pelo Hospital de Santa Maria, com o objetivo de viabilizar a integração futura destes dados no trabalho.

5.4 Descrição dos dados

O dataset [11], contém 45.150 registos de sinais de eletrocardiograma provenientes de pacientes reais, do Shaoxing People's Hospital na cidade Shaoxing na China, são registos anonimizados, e acompanhados de informações demográficas e clínicas relevantes.

Cada registo é constituído pelas seguintes variáveis principais:

- **Sinal de ECG**: Série temporal contínua que representa o traçado do sinal elétrico do coração. Cada registo tem a duração de 10 segundos e foi registado com uma frequência de amostragem de 500 Hz, o que significa que foram recolhidos 500 valores por segundo. Esta alta resolução temporal permite capturar com precisão as rápidas variações elétricas do ciclo cardíaco, essenciais para a identificação de eventos como os picos R, garantindo a qualidade dos dados utilizados para treinar os modelos de ML.
- **Idade**: Valor numérico, indica a idade do paciente no momento da recolha do sinal.
- **Sexo**: Informação categórica (masculino ou feminino).
- **Diagnóstico**: Lista de condições clínicas associadas a cada paciente, codificadas segundo a terminologia SNOMED-CT [57] e mapeadas para acrónimos mais interpretáveis, como por exemplo "AFIB" ("Atrial fibrillation") para fibrilhação auricular.

Das 63 condições as 12 apresentadas na tabela 1 são as que fazem com que o intervalo RR do sinal do ECG seja irregular.

Tabela 1 - Condições relevantes para o trabalho

Acrónimo do diagnóstico	Nome diagnóstico	Código Snomed_CT
ABI	Bigeminismo Atrial	251173003
APB	Contração Atrial prematura	284470004
JEB	*Junctional escape beat*	426995002
JPT	*Junctional premature beat*	251164006
VB	Bigeminismo Ventricular	11157007
VEB	*Ventricular escape beat*	75532003
VPB	Contração Ventricular Prematura	17338001
VET	Trigeminismo de Escape Ventricular	251180001
AFIB	Fibrilhação Auricular	164889003
AF	Flutter Atrial	164890007
SA	Arritmia Sinusal	427393009
SAAWR	Ritmo Atrial Migratório do Nodo Sinusal	195101003

5.5 Pré-processamento dos dados

Dada a natureza sensível dos dados na área da saúde, é fundamental adotar medidas cautelosas durante a limpeza e remoção de valores em falta, assegurando que o significado clínico dos dados seja preservado.

5.5.1 Recolha dos sinais ECG

Foi realizada a leitura dos sinais ECG presentes no dataset, os quais incluem as 12 derivações convencionais de um eletrocardiograma, correspondentes aos valores registados por elétrodos colocados em diferentes pontos do corpo. Cada registo encontrava-se originalmente armazenado num ficheiro individual, o que exigiu a agregação manual de todos os sinais num único DataFrame (uma estrutura de dados bidimensional, semelhante a uma tabela, composta por linhas e colunas), de forma a facilitar o seu processamento e análise.

A segunda derivação do sinal de ECG evidencia mudanças rápidas no sinal, como as associadas aos picos R, tendo isto em conta apenas utilizei os valores associados à 2º derivação dos ECGs [58]. Estes sinais foram armazenados, como uma lista de números decimais (floats) numa coluna de um DataFrame. Cada linha representa um registo diferente que é identificado de forma única pela coluna record_id. A Figura 26 mostra o sinal ECG do registo com o ID JS00004.

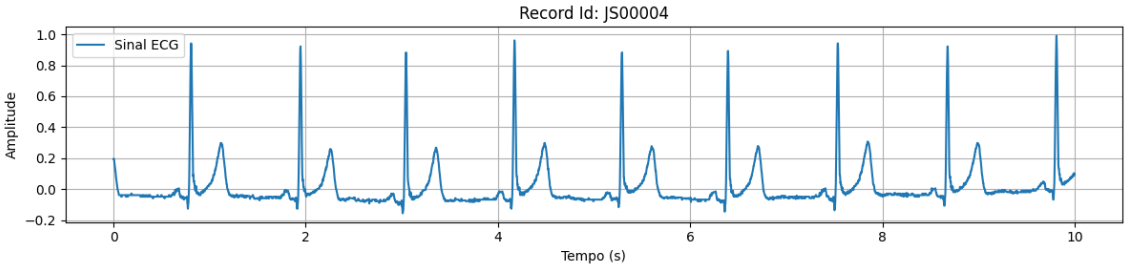

Figura 26 - Exemplo de sinal

5.5.2 Adição de informação clínica

Cada registo no dataset era constituído por dois ficheiros distintos: um que continha o sinal de ECG e outro com informações clínicas adicionais. A este segundo ficheiro foram extraídos dados como a idade, o sexo do paciente e o diagnóstico clínico atribuído por especialistas, originalmente codificado segundo o sistema SNOMED-CT.

Para facilitar a interpretação dos diagnósticos, esses códigos SNOMED-CT foram convertidos para siglas padronizadas correspondentes a cada condição clínica, utilizando uma tabela de mapeamento externa. Esta transformação teve como objetivo simplificar a leitura e análise dos dados durante as fases de processamento e modelação. Na Figura 27 encontra-se o mesmo registo que na figura anterior, mas agora com as respetivas informações clínicas.

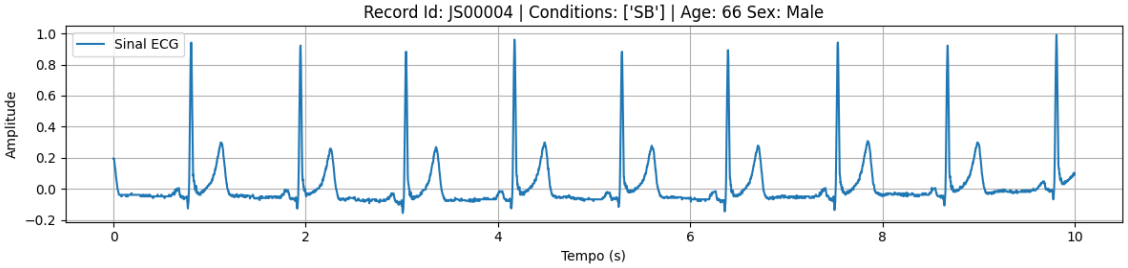

Figura 27 - Exemplo de sinal com informação clínica

5.5.3 Deteção dos batimentos cardíacos (picos R)

Foram testados três métodos diferentes para identificar os picos R dos sinais ECG: find_peaks da biblioteca Scipy, o algoritmo pan_tompkins e ecg_peaks da biblioteca NeuroKit2. Após uma comparação dos resultados, o método da NeuroKit2 foi o escolhido, por ter sido o que detetou o maior número de registos com frequências cardíacas dentro do intervalo entre 40 e 120 batimentos por minuto (BPM), intervalo considerado neste trabalho como a margem aceitável de BPM em repouso para um adulto saudável.

A função ecg_peaks da NeuroKit2 deteta os picos R com base num algoritmo próprio da biblioteca, que identifica os complexos QRS através da análise da variação rápida da amplitude do sinal ao longo do tempo. Essa variação, conhecida como gradiante absoluto, permite detetar regiões onde o sinal muda de forma mais brusca, características típicas dos complexos QRS. Após localizar essas regiões, o algoritmo identifica os picos R como os valores máximos locais dentro desses segmentos [59].

Com base nos valores de BPM extraídos por este método, foram filtrados todos os registos cuja frequência cardíaca média se encontrava fora do intervalo, 40 a 120 BPM, por se considerarem fora do padrão fisiológico de repouso e potenciais sinais de deteção incorreta. A Figura 28 apresenta o mesmo registo da figura anterior, agora com os picos R assinalados ao longo do sinal ECG e com a respetiva frequência cardíaca (BPM) calculada com base nesses picos.

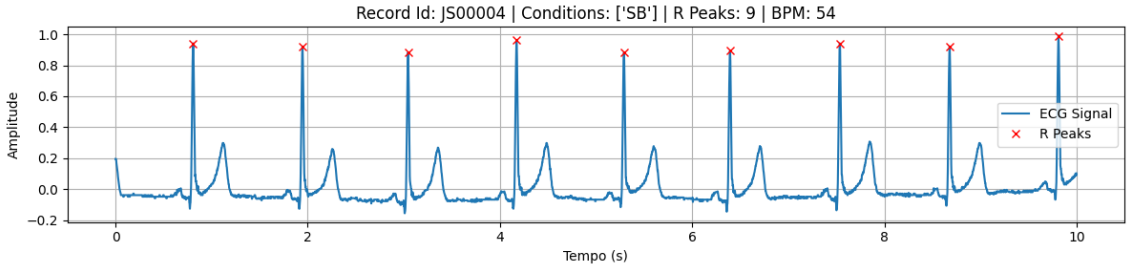

Figura 28 - Exemplo de sinal com picos R detetados

5.5.4 Cálculo de métricas do ritmo cardíaco

Através dos picos R depois foram calculados os intervalos RR, permitindo extrair métricas de HRV, como:

- Desvio padrão dos intervalos RR: É uma métrica clássica de variabilidade da frequência cardíaca [60].
- Coeficiente de variação dos intervalos RR: Divisão do desvio padrão dos intervalos pela média dos mesmos [60].
- Percentagem de intervalos RR consecutivos que diferem mais de 50 milissegundos (pNN50) e *Root mean square of successive differences* (RMSSD): Métricas utilizadas frequentemente em análise de HRV [60].
- Assimetria dos intervalos RR (*Skewness* e *kurtosis*): Medidas estatísticas que avaliam o grau de concentração dos valores em torno da média. Ambas relevantes para análise de variabilidade [61].

A comparação entre os sinais representados nas Figuras 29 e 30 evidencia diferenças claras nas métricas extraídas, refletindo a regularidade ou irregularidade dos intervalos RR. O sinal da Figura 28, caracterizado por intervalos regulares, apresenta valores muito baixos nas métricas de variabilidade, como o STD (0.02), CV (0.01), PNN50 (0.00), RMSSD (0.02), Skewness (−0.60) e Kurtosis (−1.17), indicando uma frequência cardíaca estável e ritmada. Já o sinal da Figura 29 revela grande irregularidade nos intervalos RR, com um aumento nas mesmas métricas: STD (0.24), CV (0.32), PNN50 (0.82), RMSSD (0.35), Skewness (0.93) e Kurtosis (0.05), sugerindo variabilidade significativa entre batimentos e um padrão cardíaco menos regular.

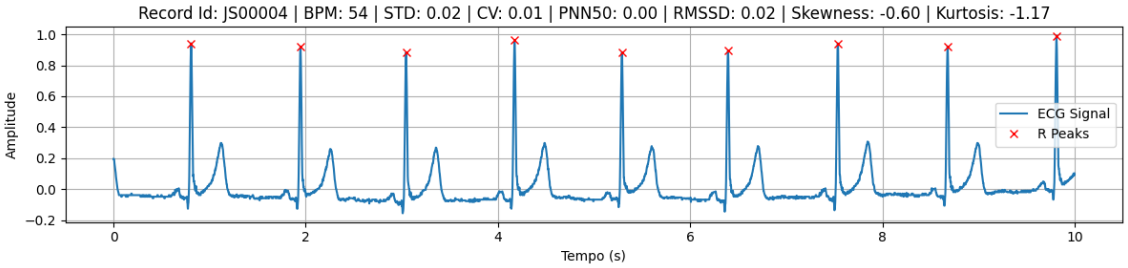

Figura 30 – Exemplo de sinal com intervalos RR regulares

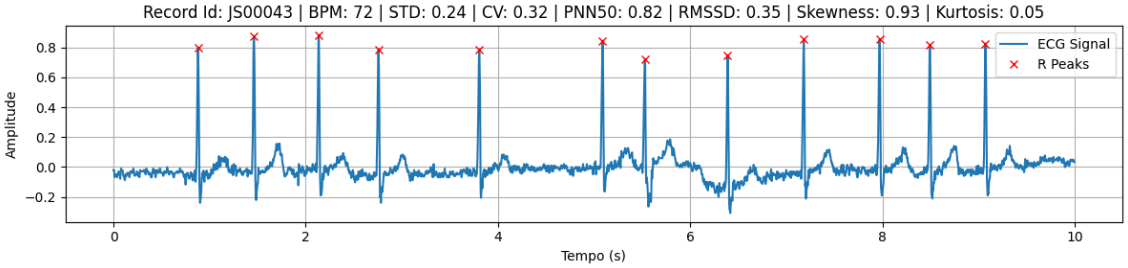

Figura 29 – Exemplo de sinal com intervalos RR irregulares

5.5.5 Criação de coluna binária para presença de arritmias

Foi adicionada uma nova coluna ao DataFrame chamada has_diagnosis, que indica se um determinado registo apresenta alguma das arritmias de interesse. Esta coluna assume o valor 1 (presença) ou 0 (ausência), com base na lista de diagnósticos associados a cada sinal.

Foram considerados os seguintes tipos de arritmia presentes na tabela da secção 5.4 do relatório, as 13 condições indicam que os intervalos RR não são regulares. Esta classificação permitiu separar os dados entre casos positivos e negativos, o que foi essencial para o treino dos modelos de ML.

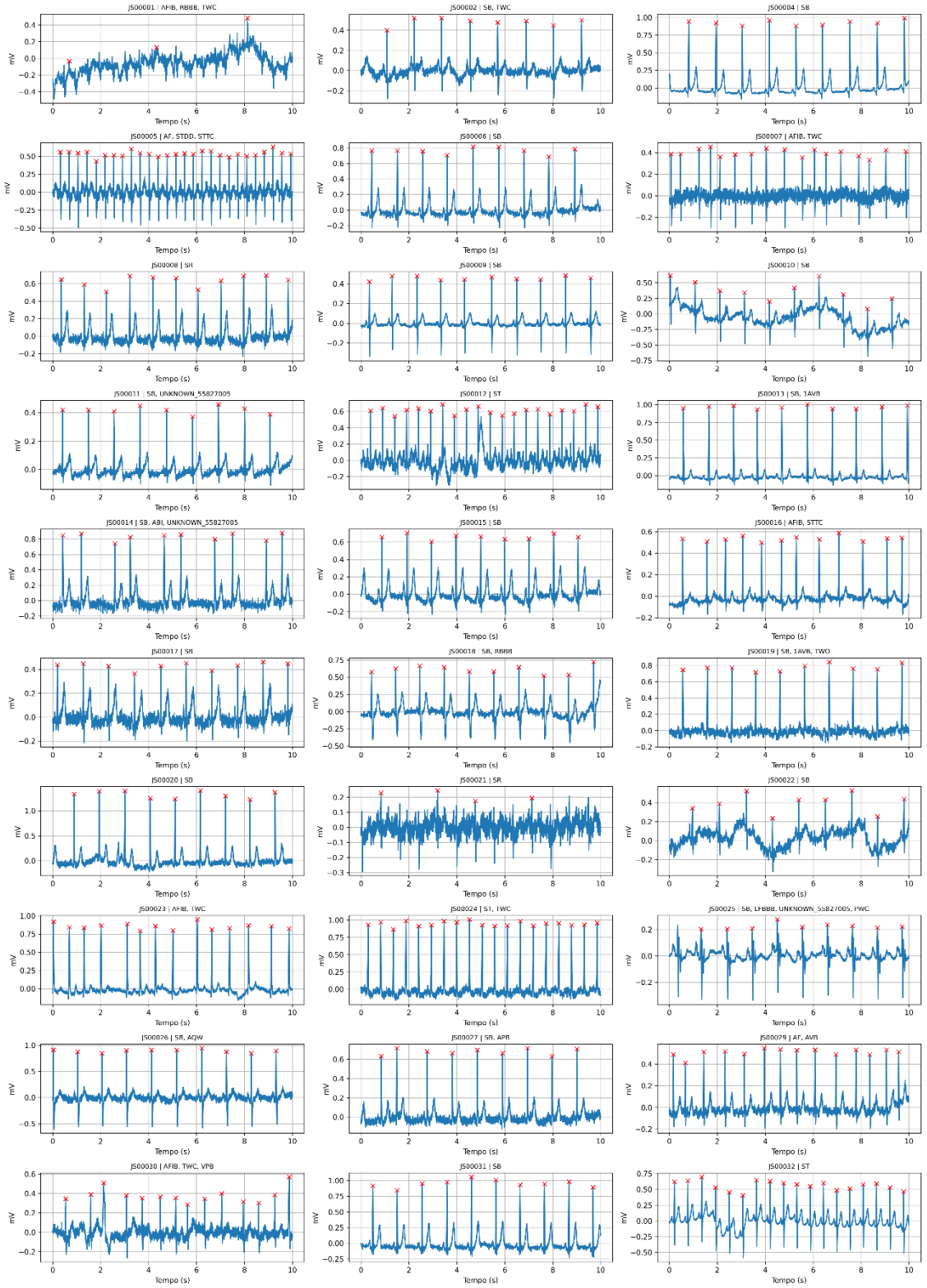

Figura 31 - Exemplos de registos e os picos R detetados

5.6 Análise Exploratória dos Dados

Após o pré-processamento foi realizada uma análise exploratória dos dados com o objetivo de identificar padrões, relações e características relevantes para o desenvolvimento do modelo de Machine Learning.

Inicialmente, o conjunto de dados continha 45.150 registos. Após a filtragem de sinais com um número de picos R fora dos limites considerados fisiologicamente plausíveis, foram mantidos 39.681 registos válidos, o que corresponde a 87,89% do total. A média de picos R por registo foi de 12,06, o que equivale a uma frequência cardíaca média de 72,36 batimentos por minuto (BPM), considerando que cada registo tem 10 segundos de duração.

$$\frac{12.06\ Picos\ R\ em\ 10\ Segundos\ X\ 60\ Segundos}{10\ Segundos} = 72.36\ Picos\ R\ em\ 60\ segundos\ (BPM)$$

Em relação aos diagnósticos, 11.356 registos apresentam pelo menos uma das arritmias consideradas, enquanto 28.325 não apresentam nenhuma, resultando numa proporção de 28,62% de casos positivos.

Importa referir que o objetivo inicial do projeto era detetar exclusivamente casos de fibrilhação auricular (FA), que representam apenas 3,94% dos diagnósticos no dataset. Este desequilíbrio é comum em contextos clínicos, mas representa um desafio significativo na construção de modelos de previsão, pois pode levar os algoritmos a favorecer a classe maioritária (ausência de FA).

Para enfrentar este problema, foram consideradas duas abordagens:

- Utilizar algoritmos com ajuste automático de pesos: Para compensar o desequilíbrio das classes, é possível recorrer a algoritmos de machine learning que permitem ajustar automaticamente os pesos atribuídos a cada classe durante o treino do modelo. Com a opção class_weight=balanced, o modelo penaliza mais fortemente os erros cometidos na classe minoritária (FA), forçando o algoritmo a prestar mais atenção aos exemplos menos representados.
- Aplicar técnicas de undersampling: Consiste em reduzir a quantidade de exemplos da classe maioritária (casos sem AFIB) até igualar aproximadamente o número de exemplos da classe minoritária (casos com AFIB). Esta técnica de undersampling permite criar conjuntos de treino mais equilibrados, minimizando o viés dos modelos para a classe dominante.

5.7 Modelos e Algoritmos Escolhidos

Conforme descrito na secção 3.2, foram aplicados os algoritmos Random Forest, XGBoost, Logistic Regression e Support Vector Classifier (SVC).

Antes do treino dos modelos, os dados foram preparados recorrendo a uma função de separação (split_features_target) que permite extrair as colunas relevantes do dataset. As variáveis selecionadas como features foram: rr_std, rr_cv, pnn50, rmssd, skewness e kurtosis, todas elas calculadas a partir dos intervalos RR dos sinais ECG. A coluna has_diagnosis foi utilizada como variável-alvo (target), indicando a presença ou ausência de um diagnóstico de arritmia.

A separação entre treino e teste foi feita com uma proporção de 80/20, utilizando a função train_test_split da biblioteca Scikit-learn, com o parâmetro stratify=y para garantir que a proporção entre classes se mantinha semelhante em ambos os conjuntos.

5.8 Abrangência

A solução proposta integra conhecimentos adquiridos em diversas unidades curriculares do curso.

Abaixo estão descritas as principais disciplinas e como os conceitos aprendidos serão utilizados:

- **Data Mining:** Aplicação de técnicas de tratamento de dados, análise exploratória e desenvolvimento do modelo de machine learning.
- **Probabilidade e Estatística:** Aplicação de conceitos como estatística descritiva, regressão linear e probabilidade para a análise dos dados clínicos e avaliação da variabilidade da frequência cardíaca.
- **Sistemas Móveis Empresariais:** Integração do modelo de machine learning com a aplicação móvel existente, garantindo as funcionalidade e uma boa usabilidade nos dispositivos móveis.
- **Base de Dados:** Estruturação dos dados clínicos para facilitar a análise e utilização eficiente pelo modelo.

- **Algoritmia e Estrutura de Dados:** Análise da complexidade computacional das tarefas envolvidas e implementação de soluções eficientes para processamento de dados e execução do modelo.

6 Método e Planeamento

6.1 Planeamento Inicial

O desenvolvimento do TFC foi organizado em fases, planeadas com base na metodologia de gestão de projeto e acompanhadas através de um cronograma Gantt.

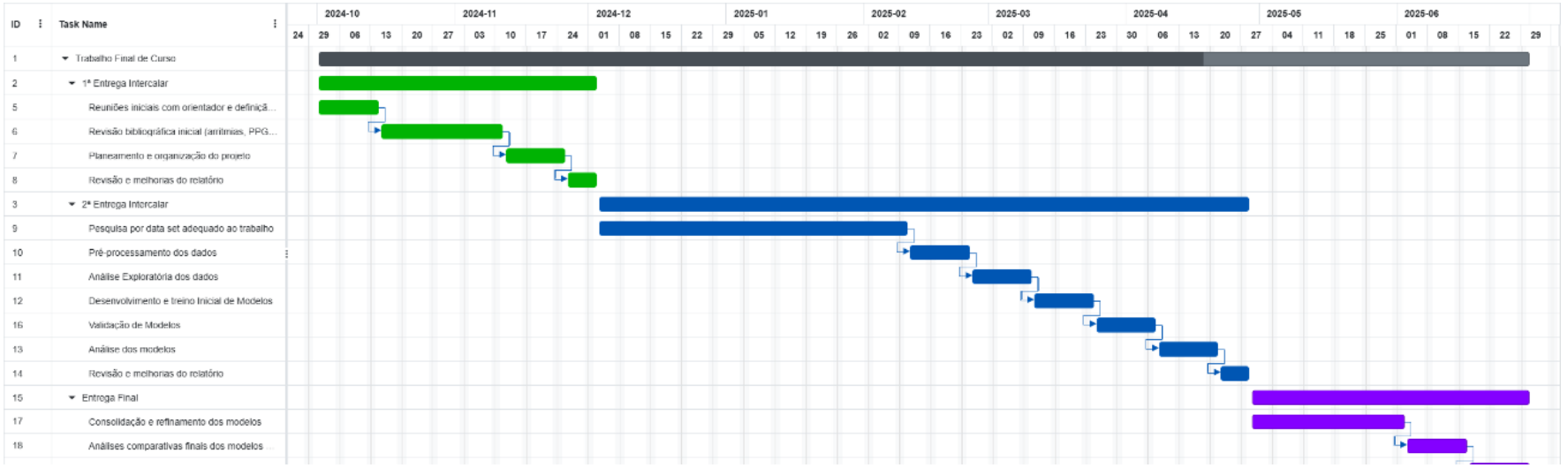

Figura 32 - Diagrama de Gantt

6.2 Análise Crítica ao Planeamento

Inicialmente, foi feita uma proposta de um tema diferente para o TFC, foi submetida por iniciativa própria, sendo posteriormente aceite pela Professora Iolanda Velho. No entanto, surgiu a oportunidade de integrar o projeto AIMHealth [4], projeto coordenado pelo Dr. Luís Rosário, em colaboração com o Hospital de Santa Maria, o Instituto Superior Técnico e o ISCTE, o que representou uma mais-valia significativa para o enriquecimento do projeto, aumentando também a sua complexidade e exigência.

Apesar de o objetivo central do projeto se manter, desenvolver um modelo de Machine Learning para auxiliar na deteção de arritmias cardíacas, o plano inicial era mais ambicioso: criar um modelo capaz de detetar apenas os casos de fibrilhação auricular (FA) com base em sinais de fotopletismografia (PPG) recolhidos por uma aplicação móvel previamente desenvolvida, e integrá-lo diretamente nessa aplicação. No entanto, devido à dificuldade em aceder a bases de dados públicas com sinais PPG rotulados para FA, não foi possível avançar com essa abordagem.

Como alternativa viável, foi decidido focar o trabalho no desenvolvimento de um modelo de Machine Learning baseado em sinais de ECG de curta duração, com foco na variabilidade da frequência cardíaca (HRV) e não apenas nos casos de FA devido ao desequilíbrio dos dados.

Apesar da alteração do tipo de dados utilizados, de sinais PPG para ECG, os princípios metodológicos aplicados neste projeto poderão futuramente ser adaptados para desenvolver um modelo semelhante baseado em PPG. Este modelo poderá então ser integrado na aplicação móvel do projeto AIMHealth, tirando partido da infraestrutura já existente e aproximando-se do objetivo inicial de detetar arritmias de forma acessível e remota.

7 Resultados e Discussão

7.1 Resultados das Análises e comparação dos modelos

Para avaliar o desempenho dos modelos desenvolvidos, foram realizados três tipos distintos de treino: com os dados originais, com os dados equilibrados utilizando técnicas de ponderação de classes, e com under sampling para equilibrar as classes. Em cada cenário, foram aplicados os algoritmos Random Forest, Logistic Regression, Support Vector Classifier (SVC) e XGBoost, amplamente utilizados em problemas de classificação binária. Abaixo apresentam-se os resultados obtidos para cada configuração, incluindo as principais métricas de avaliação: F1 Score, Precisão (Precision), Sensibilidade (Recall), Especificidade (Specificity), Acurácia (Accuracy) e AUC (Área sob a Curva ROC).

Tabela 2 - Resultados dos modelos treinados com dados originais

Treino com Dados Originais	F1 Score	Precisão	Sensibilidade	Especificidade	Acurácia	AUC
Logistic Regression	0.7838	0.8034	0.7651	0.925	0.8793	0.8451
Random Forest	0.8999	0.909	0.8909	0.9643	0.9433	0.9276
SVC	0.8202	0.8691	0.7766	0.9532	0.9027	0.8649
XGBoost	0.8946	0.9029	0.8865	0.9618	0.9403	0.9242

Tabela 3 - Resultados dos modelos treinados com dados equilibrados

Treino com Dados Equilibrados	*F1 Score*	Precisão	Sensibilidade	Especificidade	Acurácia	*AUC*
Logistic Regression	*0.7838*	*0.7127*	*0.8706*	*0.8595*	*0.8627*	*0.8651*
Random Forest	*0.8977*	*0.9042*	*0.8914*	*0.9622*	*0.9419*	*0.9268*
SVC	*0.8295*	*0.7797*	*0.8861*	*0.8898*	*0.8959*	*0.8929*
XGBoost	*0.9*	*0.8772*	*0.9241*	*0.9482*	*0.9413*	*0.9361*

Tabela 4 - Resultados dos modelos treinados com dados *Under Sampling*

Treino com Dados *Under Sampling*	F1 Score	Precisão	Sensibilidade	Especificidade	Acurácia	AUC
Logistic Regression	0.7833	0.7119	0.8706	0.8589	0.8623	0.8648
Random Forest	0.8935	0.8536	0.9373	0.9357	0.9361	0.9365
SVC	0.822	0.7691	0.8826	0.8939	0.8907	0.8882
XGBoost	0.8906	0.8535	0.9311	0.936	0.9346	0.9336

Analisando os resultados obtidos, observa-se um claro destaque para os modelos *Random Forest* e *XGBoost*, que, independentemente do tipo de tratamento aplicado aos dados, apresentaram sempre valores de F1 Score muito próximos de 0.9. Além disso, mantiveram valores elevados de AUC, entre 0.92 e 0.94, o que demonstra uma excelente capacidade de distinguir corretamente entre as classes. Esta consistência evidencia a robustez e eficácia destes modelos para o problema em estudo.

Por outro lado, os modelos *Logistic Regression* e SVC apresentaram desempenhos consistentemente inferiores em todas as métricas. Tanto em F1 Score (variando de 0.7833 a 0.8295) quanto em AUC (variando de 0.8451 a 0.8929), estes modelos ficaram aquém dos algoritmos baseados em árvores. Embora a Regressão Logística e o SVC sejam algoritmos mais simples, a sua menor capacidade para capturar padrões complexos presentes neste tipo de dados, especialmente em cenários de dados desequilibrados, é evidente nos resultados.

Apesar da aplicação de diferentes abordagens de treino (utilizando dados originais, dados equilibrados e *under sampling*), os resultados obtidos revelaram poucas variações significativas entre si. Esta estabilidade pode ser justificada pelo facto de os dados originais já apresentarem uma distribuição relativamente equilibrada, com 28,62%

dos registos a corresponderem a casos positivos (diagnóstico presente). Assim, o impacto dos vários métodos de treino foi atenuado, permitindo que os modelos treinados com os dados originais já atingissem métricas de desempenho elevadas.

O modelo que apresentou o melhor desempenho global foi o *XGBoost*, treinado com o parâmetro *class_weight='balanced'* (correspondente aos Dados Equilibrados). Esta configuração alcançou um F1-score de 0.9000, o mais elevado entre todas as configurações testadas, demonstrando um equilíbrio excecional entre precisão e sensibilidade na identificação das arritmias. O seu AUC de 0.9361 também foi notavelmente alto, apenas ligeiramente inferior ao obtido pelo modelo *Random Forest* com *under sampling*, que atingiu um AUC de 0.9365. No entanto, esse modelo registou um F1-score de 0.8935, o que representa uma diferença mais expressiva em comparação com o *XGBoost*, destacando este último como a solução com melhor equilíbrio geral entre precisão e sensibilidade.

Analisando as demais métricas, o *XGBoost* com *class_weight='balanced'* manteve consistentemente valores elevados: Precisão de 0.8772, Sensibilidade de 0.9241, Especificidade de 0.9482 e Acurácia de 0.9413. Contudo, é importante notar que o Random Forest com dados originais obteve picos ligeiramente superiores em algumas métricas: Precisão de 0.909, Especificidade de 0.9643 e Acurácia de 0.9433. Similarmente, o Random Forest com under sampling apresentou a Sensibilidade mais alta, com 0.9373. Apesar destas exceções pontuais, a performance do XGBoost com dados equilibrados demonstrou a performance mais robusta e completa em todo o conjunto de métricas, alcançando o F1-score mais alto e um AUC muito competitivo, validando a sua escolha como o modelo mais adequado e com melhor equilíbrio geral para o problema em estudo.

A matriz de confusão apresentada na Figura 33 permite visualizar o desempenho do modelo XGBoost na classificação de dois grupos: pacientes sem diagnóstico (classe 0) e com diagnóstico (classe 1).

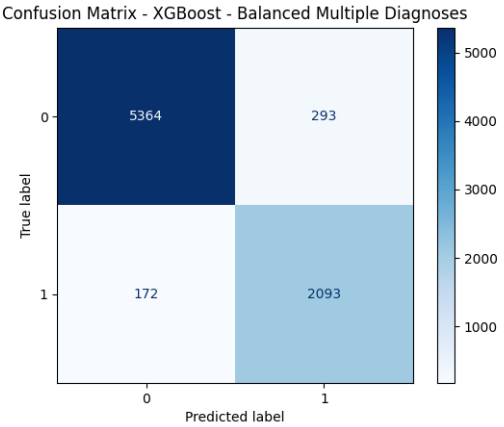

Figura 33 - Matriz de Confusão do modelo

- **Verdadeiros Negativos (5364**): Corresponde ao número de casos em que o modelo previu corretamente que o paciente não apresentava diagnóstico de arritmia.
- **Falsos Positivos (293):** Refere-se aos casos em que o modelo previu incorretamente que o paciente tinha um diagnóstico, quando na realidade não tinha.
- **Falsos Negativos (172):** São os casos em que o modelo não detetou a presença de arritmia, mesmo estando presente.
- **Verdadeiros Positivos (2093):** Casos em que o modelo identificou corretamente a presença de arritmia, o que confirma a eficácia do modelo na deteção de diagnósticos positivos.

Ao todo o modelo foi testado com 7922 casos, dos quais 7457 foram corretamente classificados e apenas 463 incorretamente classificados, estes valores correspondem a uma taxa de Accuracy de 94,13%.

A tabela 5, resume os principais indicadores de desempenho do modelo XGBoost para cada uma das classes:

Tabela 5 - Indicadores de desempenho modelo final XGBoost

	Precisão	Sensibilidade	f1-score
0 Sem Diagnóstico	0.97	0.95	0.96
1 Com Diagnóstico	0.88	0.92	0.90

Classe 0 (Sem diagnóstico):

- **Precisão (0.97):** Das previsões feitas como "sem diagnóstico", 97% estavam corretas, o que indica uma taxa muito reduzida de falsos positivos.
- **Sensibilidade (0.95):** O modelo foi capaz de identificar corretamente 95% dos casos realmente "sem diagnóstico", demonstrando elevada sensibilidade nesta classe.
- **F1-score (0.96):** Reflete um equilíbrio sólido entre precisão e recall, confirmando que o modelo é altamente eficaz na classificação dos casos negativos.

Classe 1 (Com diagnóstico):

- **Precisão (0.88):** Das previsões atribuídas à classe "com diagnóstico", 88% correspondiam efetivamente a pacientes com arritmia, indicando uma boa taxa de acerto.
- **Sensibilidade (0.92):** O modelo identificou corretamente 92% dos casos reais com diagnóstico, um valor elevado e clinicamente relevante.
- **F1-score (0.90):** Este resultado reforça a robustez do modelo mesmo na classe mais desafiante, onde a variabilidade é maior e a representação nos dados é menor.

Os resultados apresentados demonstram um excelente desempenho do modelo *XGBoost* na tarefa de classificação binária (com ou sem diagnóstico). Ao utilizar as funções *f1_score()* e *roc_auc_score()* da biblioteca sklearn.metrics, o modelo obteve um F1-score geral de 0.900 e um AUC de 0.9361, o que indica um ótimo equilíbrio entre precisão e sebsibilidade.

Em comparação com o modelo desenvolvido na 2ª entrega intermédia, que tinha como objetivo apenas a identificação de casos de Fibrilhação Auricular (FA), foi também utilizado o algoritmo *XGBoost* com o parâmetro *class_weight='balanced'*. Importa referir que, nessa fase, foram utilizados os mesmos dados de base, extraídos do mesmo dataset, com as mesmas métricas calculadas a partir dos intervalos RR (como o desvio padrão, o RMSSD, o PNN50, entre outras), e aplicando exatamente os mesmos métodos de deteção de picos e extração de características, bem como o mesmo pipeline de treino do modelo de machine learning.

Tabela 6 - Indicadores de desempenho modelo 2ª entrega

	Precisão	Sensibilidade	f1-score
0 Sem Diagnóstico	0.98	0.87	0.92
1 Com Diagnóstico	0.17	0.66	0.27

Apesar de ter existido consistência metodológica, os resultados demonstraram uma diferença substancial no desempenho, especialmente na classe "com diagnóstico", onde o F1-score evoluiu de 0.27 para 0.90. Esta diferença justifica-se, em grande parte, pela representatividade limitada da classe "Com Diagnóstico" (neste caso com FA) na 2ª entrega: o número de amostras com esse diagnóstico era bastante reduzido, o que limitava a capacidade do modelo de aprender padrões relevantes para identificar esses casos. Além disso, a elevada precisão de 98% na identificação dos casos "sem diagnóstico" revela que o modelo da 2ª entrega estava fortemente direcionado para essa classe, falhando na deteção de casos positivos. Isso é evidenciado pela baixa precisão de apenas 17% nos casos com FA, indicando que, quando o modelo previa a presença de diagnóstico, raramente estava correto, o que reforça a conclusão de que o modelo não estava devidamente treinado para identificar casos com diagnóstico.

Por outro lado, o modelo final desenvolvido neste trabalho incluiu todos os diagnósticos relacionados com os intervalos RR irregulares, abrangendo não apenas a FA, mas também outras condições clínicas relevantes, o que resultou numa base de dados mais rica, equilibrada e informativa, permitindo um treino mais robusto e eficaz.

Consequentemente, a capacidade preditiva do modelo aumentou de forma significativa, sobretudo nos casos positivos.

7.2 Interpretação dos resultados

Os resultados obtidos com o modelo XGBoost, utilizando class_weight=balanced, revelaram um desempenho bastante robusto, com um F1-score global de 0.90 e uma AUC (Área Sob a Curva ROC) de 0.9361. Este valor de AUC indica que o modelo tem uma excelente capacidade discriminativa para distinguir entre sinais com e sem arritmia, o que reforça a fiabilidade da abordagem baseada na análise dos intervalos RR. Estes resultados demonstram um equilíbrio sólido entre a precisão e o recall, essenciais para tarefas clínicas sensíveis onde é fundamental minimizar tanto os falsos negativos como os falsos positivos.

Este desempenho representa uma evolução notável face ao modelo anterior desenvolvido na 2ª entrega, que se focava apenas na deteção de casos de fibrilhação auricular (FA) e obteve um F1-score de apenas 0.27 para a classe positiva. A expansão do objetivo para identificar um conjunto mais abrangente de arritmias revelou-se decisiva para aumentar a aplicabilidade prática e a robustez do modelo.

Quando comparado com os estudos existentes na literatura, como os de Voisin et al. (2018) [53], que obtiveram AUCs na ordem dos 0.95 utilizando redes neuronais com sinais PPG, os resultados deste trabalho situam-se dentro de uma margem muito competitiva.

Por fim, os resultados suportam a relevância da hipótese inicial: que é possível detetar padrões patológicos fiáveis utilizando apenas os intervalos entre batimentos cardíacos (RR). Com o acesso a sinais PPG de qualidade ou com a futura integração com a aplicação AIMHealth, será possível evoluir para uma ferramenta completa de rastreio remoto, acessível e validada clinicamente, contribuindo para a prevenção ativa de doenças cardiovasculares com recurso à tecnologia.

7.3 Limitações da Análise

Durante o desenvolvimento deste trabalho, foi identificada uma limitação significativa no que diz respeito à disponibilidade de dados clínicos reais, elemento essencial numa investigação com aplicação médica. A utilização de dados reais é crucial para garantir a fiabilidade dos resultados e a sua eventual aplicabilidade em contextos clínicos.

Inicialmente, a intenção era trabalhar com sinais PPG, dado o seu carácter não invasivo e o potencial de recolha contínua através de dispositivos wearables. No entanto, verificou-se que os datasets públicos de PPG são extremamente escassos, sobretudo quando se exige uma anotação clínica fidedigna, como o diagnóstico de fibrilhação auricular (FA) ou de qualquer outro tipo de arritmia.

Mesmo ao expandir a pesquisa para sinais de ECG, dados mais comuns em datasets púbicos, foi necessário aplicar critérios rigorosos de seleção, pois não seria viável utilizar qualquer dataset disponível. Muitos dos sinais encontrados apresentavam uma baixa frequência de amostragem, o que compromete a qualidade da representação gráfica do sinal e dificulta a deteção precisa dos picos R, essenciais para o cálculo das métricas RR utilizadas no modelo.

Outro critério fundamental foi a exigência de que os sinais tivessem um diagnóstico médico associado, garantindo que o modelo fosse treinado com base em dados rotulados de forma correta e confiável. Para além disso, era necessário que cada registo tivesse uma duração mínima de sinal, de forma a permitir a extração de padrões significativos. Sinais muito curtos geram menos picos R, tornando as métricas menos robustas e o padrão de ritmo cardíaco mais difícil de identificar.

Estas limitações levaram à reformulação parcial do objetivo inicial do projeto. Embora o propósito inicial fosse a deteção de fibrilhação auricular através de sinais PPG, a escassez de dados adequados obrigou à utilização de sinais ECG e à expansão do diagnóstico para outros tipos de arritmias cardíacas. Ainda assim, o trabalho manteve alinhamento com a tese inicial, explorando o potencial de modelos de Machine Learning na deteção automática de padrões anómalos no ritmo cardíaco, mesmo que com outro tipo de dados.

8 Conclusão

8.1 Conclusão

O presente Trabalho Final de Curso abordou o desenvolvimento de um modelo de Machine Learning com o objetivo de detetar arritmias cardíacas com base em sinais de eletrocardiograma (ECG), com o objetivo de, futuramente, integrar um modelo semelhante numa aplicação móvel de monitorização da saúde cardíaca. Este TFC responde à necessidade clínica, realização de diagnósticos de condições cardíacas potencialmente graves, precoces e acessíveis.

O projeto teve como ponto de partida a definição do problema clínico e científico, seguindo-se um levantamento e estudo dos conceitos fundamentais da eletrofisiologia cardíaca e das principais métricas de variabilidade da frequência cardíaca (HRV). Um passo determinante foi a procura de um dataset com dados reais e adequado, que permitisse extrair intervalos RR com qualidade suficiente para análise. Após a escolha do dataset, foi realizada uma fase de pré-processamento e tratamento dos sinais, incluindo a deteção dos picos R e o cálculo de métricas estatísticas sobre a variabilidade dos intervalos RR, estas métricas serviram de base para a criação dos modelos, permitindo transformar os sinais fisiológicos em dados estruturados compreensíveis pelos algoritmos, o que foi essencial para que os modelos fossem capazes de reconhecer padrões e tomar decisões com base nesses dados.

Diversos algoritmos de classificação binária foram implementados e avaliados, incluindo Random Forest, Logistic Regression, SVC e XGBoost. Os resultados obtidos demonstraram que os modelos XGBoost e Random Forest foram consistentemente os mais eficazes, alcançando, no melhor cenário, um F1-score de 0.9000 e uma AUC de 0.9361, valores que indicam uma excelente capacidade de distinguir entre pacientes com e sem diagnóstico de arritmia. Estes dados confirmam a hipótese inicial de que é possível detetar padrões patológicos a partir dos intervalos RR, utilizando apenas métricas estatísticas derivadas desses sinais.

Apesar de não ser ainda viável para aplicação clínica direta, o modelo mostrou-se eficaz e promissor, abrindo portas à sua futura integração na aplicação AIMHealth. Esta integração permitiria a monitorização remota e em tempo real da saúde cardíaca, com recurso a sinais fisiológicos recolhidos de forma não invasiva, pelo próprio utilizador da aplicação.

O projeto permitiu, ainda, a consolidação prática de conhecimentos adquiridos ao longo do curso, cruzando áreas como probabilidade e estatística, data mining e inteligência artificial, e demonstrando como estas áreas podem ser integradas na resolução de problemas reais com impacto na saúde pública. Este projeto representa um ponto de partida sólido para investigações futuras, com elevado potencial de inovação e aplicabilidade no setor da saúde.

8.2 Trabalhos Futuros

Apesar dos resultados promissores obtidos até ao momento, este TFC pode representar apenas o início de um percurso mais amplo de investigação e desenvolvimento em parceria com o Hospital Santa Maria o IST e o ISCTE.

Neste sentido, identificam-se várias linhas de trabalho futuro que poderão dar continuidade e aprofundar este projeto, com vista à sua validação científica, aplicabilidade clínica e eventual implementação prática.

- **Acesso a dados PPG de qualidade clínica:**
  Uma das principais limitações deste trabalho foi a indisponibilidade de conjuntos de dados públicos com sinais PPG (Photoplethysmography) com qualidade clínica suficiente. Para que um modelo de Machine Learning seja eficaz na deteção de arritmias com base em PPG, é fundamental que os dados utilizados cumpram vários critérios: devem ter uma frequência de amostragem adequada, apresentar mínimos níveis de ruído, e estar anotados com diagnósticos médicos fiáveis, preferencialmente confirmados por profissionais de saúde.
  Embora existam alguns repositórios de investigação e bases de dados de instituições médicas internacionais com este tipo de informação, o seu acesso está muitas vezes condicionado por restrições éticas e legais, sendo necessário submeter um pedido formal de acesso, no qual se deve justificar a natureza do projeto, os objetivos científicos e os elementos envolvidos no estudo. Estes pedidos exigem tempo de aprovação e, por isso, não foram compatíveis com os prazos deste trabalho final de curso, que teve um calendário limitado ao ano letivo em vigor.
  Para trabalhos futuros, será fundamental iniciar este processo com antecedência, garantindo assim tempo suficiente para obter autorizações e integrar esses dados no desenvolvimento do modelo. O acesso a dados PPG de qualidade clínica permitirá, não só, retomar o objetivo original deste trabalho, a deteção de

fibrilhação auricular através de PPG, como também aumentar a relevância prática e científica da investigação, aproximando-a de uma possível aplicação real. Além disso, a utilização deste tipo de dados será essencial para validar a generalização do modelo e garantir a sua aplicabilidade em contextos reais, nomeadamente na futura integração com a aplicação móvel AIMHealth.

- **Integração com a aplicação móvel AIMHealth:**
  Uma das direções mais promissoras para a continuação deste trabalho é a integração do modelo preditivo na aplicação móvel AIMHealth, atualmente em desenvolvimento no âmbito de um projeto mais alargado. A ideia passa por utilizar os sinais PPG recolhidos diretamente a partir da câmara do smartphone, uma abordagem inovadora que dispensa sensores médicos dedicados e permite a monitorização de sinais fisiológicos de forma acessível e contínua.
  A integração do modelo de ML com a aplicação tem como objetivo a deteção automática de arritmias cardíacas com base em padrões irregulares nos intervalos RR extraídos dos sinais PPG. Esta funcionalidade poderá notificar o utilizador em tempo real sempre que forem detetadas irregularidades compatíveis com possíveis arritmias, funcionando como uma ferramenta preventiva e de triagem, com especial utilidade na saúde de pessoas que não têm fácil acesso a Hospitais.
  No entanto, este avanço requer um modelo altamente fiável, validado clinicamente e com métricas de desempenho rigorosas, nomeadamente em termos de precisão, sensibilidade e especificidade. Sendo uma aplicação com impacto direto na área da saúde, o modelo deverá cumprir com elevados padrões éticos e regulamentares. Para tal, será essencial garantir que a aquisição dos sinais PPG seja feita em condições controladas (ex: iluminação estável, posição fixa da câmara), e que o modelo seja treinado com dados PPG de qualidade e devidamente rotulados, o que ainda representa um desafio a ultrapassar.
  Esta integração poderá permitir transformar a aplicação AIMHealth numa ferramenta de apoio ao diagnóstico clínico remoto, com potencial para reduzir a pressão sobre os serviços de saúde e capacitar os utilizadores no acompanhamento proativo da sua saúde cardiovascular.

- **Validação clínica com dados reais:**
  Um dos passos mais importantes para reforçar a credibilidade e aplicabilidade prática do modelo desenvolvido será a validação clínica com dados reais de pacientes. A possibilidade de utilizar registos provenientes do Hospital de Santa Maria, nomeadamente de pacientes acompanhados pelo Dr. Luís Rosário, representa uma oportunidade valiosa para testar o modelo em cenários reais e clínicos.
  Este processo permitiria não só avaliar o desempenho do modelo com novos dados, distintos dos usados no treino, mas também realizar uma comparação direta entre os diagnósticos emitidos pelo sistema de Machine Learning e os diagnósticos clínicos feitos por profissionais de saúde. Tal comparação possibilitaria validar a eficácia do modelo em contexto real, identificar potenciais limitações e refinar a sua capacidade preditiva com base em feedback médico.
  Adicionalmente, estes dados clínicos poderiam ser utilizados, caso existisse volume e qualidade suficientes, para criar um novo modelo mais específico, adaptado ao perfil dos pacientes do hospital, o que contribuiria para um sistema mais personalizado e eficaz. Esta colaboração entre áreas da saúde e da tecnologia reforçaria o valor científico do projeto e permitiria dar um passo em direção à integração de soluções baseadas em ML no apoio ao diagnóstico médico, com base em dados reais, validados e eticamente tratados.

- **Ajuste do modelo:**
  Apesar dos bons resultados obtidos com modelos como o Random Forest e o XGBoost, uma linha de trabalho futuro será a exploração de outras arquiteturas de Machine Learning, nomeadamente redes neuronais artificiais e abordagens mais avançadas como redes neuronais recorrentes.
  Estas arquiteturas poderão permitir uma melhor captação de padrões complexos nos intervalos RR e um desempenho superior em contextos com dados ruidosos ou não lineares. Além disso, uma evolução natural do projeto será a transição de um modelo binário (com ou sem diagnóstico) para um modelo multiclasse, ou seja, um modelo classificador capaz de identificar especificamente o tipo de arritmia presente, como fibrilhação auricular, taquicardia ou bradicardia, a partir das métricas extraídas.
  Esta abordagem não só aumentaria a utilidade clínica do sistema, como também permitiria criar alertas mais específicos e personalizados para cada tipo de condição, contribuindo para uma triagem mais precisa e uma intervenção médica mais direcionada. O desenvolvimento destes classificadores exigirá, naturalmente, datasets mais amplos e detalhados, com diagnósticos precisos por tipo de arritmia.

Bibliografia

[1] World Health Organization: WHO. (2021, Junho 11). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Silver, L. (2019, February 5). Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/

[3] Yuan, P., Li, J., & Cheng, J. (2022). Automatic detection of atrial fibrillation using dual-path network based on two-dimensional ECG signal. *Computer Methods and Programs in Biomedicine, 215*, 106521. https://doi.org/10.1016/j.cmpb.2021.106521

[4] AIMHealth. (n.d.). AIMHealth – Artificial Intelligence in mHealth for Cardiovascular Diseases. ISCTE-IUL. Recuperado em abril de 2025, de https://istar.iscte-iul.pt/aimhealth/

[5] Supelnic, M. N., Ferreira, A. F., Bota, P. J., Brás-Rosário, L., & Da Silva, H. P. (2023). Benchmarking of sensor configurations and measurement sites for Out-of-the-Lab Photoplethysmography. Sensors, 24(1), 214. https://doi.org/10.3390/s24010214

[6] Martins, F., Fragoso, E., Plácido da Silva, H., Dias, M. S., & Rosário, L. B. (2024). Validation of an mHealth System for Monitoring Fundamental Physiological Parameters in the Clinical Setting. Sensors, 24(16), 5164. https://doi.org/10.3390/s24165164

[7] United Nations Regional Information Centre (UNRIC). (n.d.). *Objetivos de Desenvolvimento Sustentável (ODS)*. Recuperado em abril de 2025, de https://unric.org/pt/objetivos-de-desenvolvimento-sustentavel/

[8] What is an Arrhythmia? (2024, September 27). www.heart.org. https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia

[9] Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: a review. Medical & Biological Engineering & Computing, 44(12), 1031–1051. https://doi.org/10.1007/s11517-006-0119-0

[10] Sociedade Portuguesa de Cardiologia. (2020). *Recomendações de bolso de 2020 da ESC: Fibrilhação auricular*. https://spc.pt/profissional-de-saude/wp-content/uploads/2023/03/Pockets-Fibrilhacao-Auricular_compressed.pdf

[11] Zheng, J., Guo, H., & Chu, H. (2022). A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/wgex-er52.

[12] Margato, R. (2021, agosto 9). *Eletrocardiograma*. Saúde e Bem Estar. https://www.saudebemestar.pt/pt/medicina/cardiologia/eletrocardiograma/

[13] CUF. (2025, maio 26). *Tabela de preços de consultas, exames e tratamentos* [PDF]. https://www.cuf.pt/media/50900/download?inline

[14] Hospital da Luz. (2025, janeiro 7). *Preços e faturas*. https://www.hospitaldaluz.pt/lisboa/pt/para-clientes/precos-e-faturas

[15] Hospital Cruz Vermelha. (n.d.). *Tabela de preços*. https://hospitalcruzvermelha.pt/tabela-de-precos/

[16] Cleveland Clinic. (2024, janeiro 26). *Heart*. Cleveland Clinic. https://my.clevelandclinic.org/health/body/21704-heart

[17] Gupta, J. I., & Shea, M. J. (2025, abril). *Biologia do coração*. Em J. G. Howlett (Rev.), MSD Manuals. https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-do-cora%C3%A7%C3%A3o-e-dos-vasos-sangu%C3%ADneos/biologia-do-cora%C3%A7%C3%A3o-e-dos-vasos-sangu%C3%ADneos/biologia-do-cora%C3%A7%C3%A3o

[18] British Heart Foundation. (n.d.). *Electrocardiogram (ECG)*. https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/tests/electrocardiogram-ecg

[19] *Electrocardiogram (EKG) components and intervals*. (n.d.). https://myhealth.alberta.ca/Health/pages/conditions.aspx?hwid=zm2308

[20] Jezzini, A., Ayache, M., Ibrahim, Z. A. A., & Elkhansa, L. (2015). ECG classification for sleep apnea detection. In 2015 International Conference on Advances in Biomedical Engineering (ICABME) (pp. 101–104). IEEE. https://ieeexplore.ieee.org/document/7323312

[21] Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3), 230–236. https://pubmed.ncbi.nlm.nih.gov/8598068/

[22] Allen J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, *28*(3), R1–R39. https://doi.org/10.1088/0967-3334/28/3/R01

[23] Vandenberk, T., Stans, J., Van Schelvergem, G., Pelckmans, C., Smeets, C., Lanssens, D., De Cannière, H., Storms, V., Thijs, I., & Vandervoort, P. (2017). Clinical validation of heart rate apps: Mixed-methods evaluation study. *JMIR mHealth and uHealth*, 5(8), e129. https://mhealth.jmir.org/2017/8/e129/

[24] Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. *Current Cardiology Reviews*, 8(1), 14–25. https://doi.org/10.2174/157340312801215782

[25] What is an Arrhythmia? (2024, Septembro 27). www.heart.org. https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia

[26] KNIME. (2022, October 10). *ECG categorization to detect arrhythmia*. KNIME Blog. https://www.knime.com/blog/ecg-categorization-to-detect-arrhythmia

[27] Desai, D. S., & Hajouli, S. (2023, Junho 5). *Arrhythmias*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK558923/

[28] Atrial Fibrillation: Resources for Patients. (2020, August 27). *Understanding the EKG Signal - Atrial Fibrillation: Resources for patients*. https://a-fib.com/treatments-for-atrial-fibrillation/diagnostic-tests-2/the-ekg-signal/

[29] Lowry Lehnen, T. (2021, setembro 13). *Atrial fibrillation: An overview*. Medical Independent. https://www.medicalindependent.ie/clinical-news/atrial-fibrillation-an-overview/

[30] Maji, Uday & Mitra, Madhuchhanda & Pal, Saurabh. (2013). Automatic Detection of Atrial Fibrillation Using Empirical Mode Decomposition and Statistical Approach. Procedia Technology. 10. 45–52. 10.1016/j.protcy.2013.12.335.

[31] Cleveland Clinic. (2022, março 21). *Sinus arrhythmia*. https://my.clevelandclinic.org/health/diseases/21666-sinus-arrhythmia

[32] Lome, S. (n.d.). *Wandering atrial pacemaker review*. Healio. https://www.healio.com/cardiology/learn-the-heart/ecg-review/ecg-topic-reviews-and-criteria/wandering-atrial-pacemaker-review

[33] Burns, E., & Buttner, R. (2024, outubro 8). *Premature atrial complex (PAC)*. Life in the Fast Lane. https://litfl.com/premature-atrial-complex-pac/

[34] Lome, S. (n.d.). *Premature Ventricular Contractions (PVCs)*. Healio. 13 de junho de 2025, de https://www.healio.com/cardiology/learn-the-heart/cardiology-review/topic-reviews/premature-ventricular-contractions-pvcs

[35] Winter, J. L. (2016, 13 de novembro). *Atrial Bigeminy*. ECG Educator. https://ecg-educator.blogspot.com/2016/11/atrial-bigeminy.html

[36] Winter, J. L. (2016, 14 de novembro). *Ventricular Bigeminy*. ECG Educator. https://ecg-educator.blogspot.com/2016/11/ventricular-bigeminy.html

[37] Sampson, Michael. (2016). Understanding the ECG Part 4: Conduction blocks. British Journal of Cardiac Nursing. 11. 71-79. 10.12968/bjca.2016.11.2.71.

[38] Winter, J. L. (2016, 14 de novembro). *Premature Junctional Complex (PJC)*. ECG Educator. https://ecg-educator.blogspot.com/2016/11/premature-junctional-complex-pjc.html

[39] The W-Project. (2012). *Ventricular escape beat*. http://thew-project.org/Arrhythmia_LibSys/Definitions/Ventricular%20escape%20beat_DEF.htm

[40] Qaly. (n.d.). *What Ventricular Trigeminy Looks Like On Your Watch ECG*. 13 de junho de 2025, de https://www.qaly.co/post/what-ventricular-trigeminy-looks-like-on-your-watch-ecg

[41] Kanade, V. (2022, 8 de abril). *What is Logistic Regression?*. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[42] Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995). https://doi.org/10.1007/BF00994018

[43] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[44] Wei, D. (2024, 23 de fevereiro). *Demystifying Machine Learning Models: Random Forest*. Medium. https://medium.com/@weidagang/demystifying-machine-learning-models-random-forest-f992dc50b427

[45] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. arXiv preprint arXiv:1603.02754. https://arxiv.org/abs/1603.02754

[46] Torres, L. F. (2023, 28 de abril). *XGBoost: The king of Machine Learning Algorithms*. Medium. https://medium.com/latinxinai/xgboost-the-king-of-machine-learning-algorithms-6b5c0d4acd87

[47] Kashyap, P. (2024, 2 de dezembro). *Understanding Precision, Recall, and F1-Score Metrics*. Medium. https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

[48] GeeksforGeeks. (2025, 2 de junho). *What is Specificity*. https://www.geeksforgeeks.org/machine-learning/what-is-specificity/

[49] GeeksforGeeks. (2025, 2 de julho). *Evaluation Metrics in Machine Learning*. https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

[50] GeeksforGeeks. (2025, 12 de maio). *AUC ROC Curve in Machine Learning*. https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/

[51] Gomes, P. C. T. (2024, 19 de outubro). *Matriz de Confusão*. Data Geeks. https://www.datageeks.com.br/matriz-de-confusao/

[52] Python Package Index - PyPI. (n.d.). Python Software Foundation. Retrieved from https://pypi.org/

[53] Voisin, M., Shen, Y., Aliamiri, A., Avati, A., Hannun, A., & Ng, A. (2018, November 12). *Ambulatory Atrial Fibrillation Monitoring Using Wearable Photoplethysmography with Deep Learning*. arXiv.org. https://arxiv.org/abs/1811.07774#

[54] Whiting, S., Moreland, S., Costello, J., Colopy, G., & McCann, C. (2018, July 11). *Recognising Cardiac Abnormalities in Wearable Device Photoplethysmography (PPG) with Deep Learning*. arXiv.org. https://arxiv.org/abs/1807.04077

[55] Bulut, M. G., Unal, S., Hammad, M., & Pławiak, P. (2025). Deep CNN-based detection of cardiac rhythm disorders using PPG signals from wearable devices. *PLoS ONE*, *19*(1), e0314154. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0314154

[56] Qin, Q., Li, J., Yue, Y., & Liu, C. (2017). An Adaptive and Time-Efficient ECG R-Peak Detection Algorithm. *Journal of healthcare engineering*, *2017*, 5980541. https://doi.org/10.1155/2017/5980541

[57] SNOMED International. (n.d.). *SNOMED CT Starter Guide*. Retrieved from https://www.snomed.org/snomed-ct/what-is-snomed-ct

[58] Yuan, P., Li, J., & Cheng, J. (2022). Automatic detection of atrial fibrillation using dual-path network based on two-dimensional ECG signal. *Computer Methods and Programs in Biomedicine, 215*, 106521. https://doi.org/10.1016/j.cmpb.2021.106521

[59] Makowski, D. (2025). *ECG*. NeuroKit2 documentation. https://neuropsychology.github.io/NeuroKit/functions/ecg.html

[60] Shaffer, F., & Ginsberg, J. P. (2017). *An Overview of Heart Rate Variability Metrics and Norms*. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2017.00258

[61] Kamath, M. V., & Fallen, E. L. (1993). *Power spectral analysis of heart rate variability: a noninvasive signature of cardiac autonomic function*. Critical Reviews in Biomedical Engineering. https://pubmed.ncbi.nlm.nih.gov/8243093/