UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

Tiago Almeida Zanetti

Previsão de Pressão Arterial Utilizando Métodos de Aprendizado de Máquina

Campinas
20/11/2025

Tiago Almeida Zanetti

Previsão de Pressão Arterial Utilizando Métodos de Aprendizado de Máquina*

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Petra Maria Bartmeyer.

*Este trabalho foi financiado pelo PIBIC/UNICAMP, cota 2025-2026.

Resumo

Este projeto propõe o desenvolvimento de um modelo de aprendizado de máquina para previsão contínua e não invasiva da pressão arterial em indivíduos com lesão medular, utilizando sinais multivariados obtidos por sensores vestíveis. Para isso, serão consolidadas bases de dados de macacos com lesão medular, aplicando-se técnicas de pré-processamento (filtragem de intervalos comuns, remoção de *outliers* e *time warping*) e de *feature engineering*. Inicialmente, foi estabelecido um modelo baseline com regressão linear; em seguida, exploraram-se modelos com termos penalizadores para melhorar o desempenho preditivo e realizar seleção automática de variáveis. A otimização de hiperparâmetros foi feita via validação cruzada, e o desempenho avaliado por métricas de regressão (RMSE, MRE e R^2). Espera-se oferecer uma alternativa acessível e precisa aos métodos convencionais de aferição, com potencial de integração em *smartwatches* e aplicativos de saúde, beneficiando não apenas pacientes com lesão medular, mas ampliando as opções de monitoramento contínuo no cuidado à saúde.

Abstract

This project proposes the development of a machine learning model for continuous and non-invasive blood pressure prediction in individuals with spinal cord injury, using multivariate signals obtained from wearable sensors. To achieve this, databases from monkeys with spinal cord injury will be consolidated, applying preprocessing techniques (common interval filtering, outlier removal, and time warping) and feature engineering. Initially, a baseline model was established using linear regression; subsequently, models with penalty terms were explored to improve predictive performance and perform automatic variable selection. Hyperparameter optimization was performed via cross-validation, and performance was evaluated using regression metrics (RMSE, MRE, and R^2). It is expected to offer an accessible and accurate alternative to conventional measurement methods, with the potential for integration into smartwatches and health applications, benefiting not only patients with spinal cord injury but also expanding the options for continuous monitoring in healthcare.

Conteúdo

1 Introdução

O presente projeto propõe o desenvolvimento de um modelo de aprendizado de máquina capaz de prever a pressão arterial de indivíduos com lesão medular a partir de um conjunto diversificado de *features*. Essa abordagem tem o potencial de contornar as limitações impostas pelos métodos convencionais, os quais, ao serem aplicados, podem desencadear alterações na pressão do paciente. A ideia é, por meio de técnicas de *machine learning*, analisar padrões e tendências derivadas de dados que representem com precisão o estado fisiológico dos pacientes, sem a interferência dos métodos tradicionais de medição.

A relevância deste estudo é duplamente evidenciada: primeiramente, pela possibilidade de oferecer uma alternativa mais acessível e precisa para a aferição da pressão arterial em pessoas com lesão medular; e, em segundo lugar, pelo potencial de aplicação ampla na área da saúde, beneficiando inclusive indivíduos sem as referidas limitações, ampliando as opções de monitoramento contínuo e não invasivo. Para a implementação, foram utilizados modelos de regressão linear com termos de regularização.

Além disso, o desenvolvimento deste modelo abrirá caminho para a integração dessa ferramenta em aplicativos de saúde executados em *smartwatches*, promovendo uma medição mais simples e eficiente no dia a dia dos usuários. Essa integração não só facilitará o monitoramento, mas também ampliará o acesso a informações essenciais para a gestão da saúde, contribuindo para a melhoria da qualidade de vida dos pacientes.

2 Descrição do problema

2.1 Monitoramento não invasivo e seu potencial para pacientes com lesão medular

O monitoramento de parâmetros fisiológicos pode ser realizado de forma invasiva ou não-invasiva. Métodos invasivos, como a inserção de cateteres arteriais para monitoramento hemodinâmico contínuo, embora ofereçam medições precisas e em tempo real, apresentam riscos significativos de infecção, trombose, desconforto e são inadequados para monitoramento rotineiro ambulatorial. Por outro lado, o monitoramento não-invasivo refere-se à obtenção de parâmetros fisiológicos sem a necessidade de inserção de

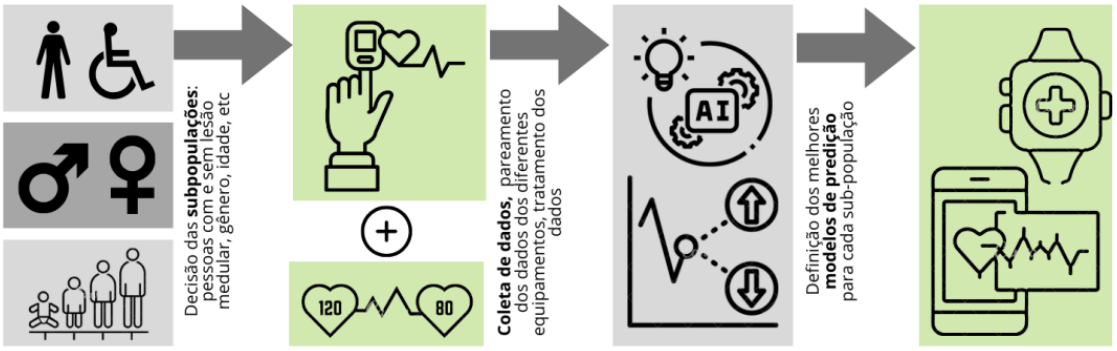


Figura 1: Fluxo do projeto de previsão de pressão arterial via IA, desde a coleta de dados até a aplicação em *smartwatches*.

dispositivos no corpo do paciente. Para pessoas com lesão medular, métodos não-invasivos são particularmente promissores.

Diante dessas limitações dos métodos convencionais, alternativas baseadas em sensores vestíveis (*wearables*), como *smartwatches*, surgem como uma solução promissora. Esses dispositivos oferecem uma oportunidade de obter sinais fisiológicos de maneira contínua, confortável e com menos riscos de saúde [Vijayalakshmi et al., 2021]. Isso contribui para um acompanhamento mais preciso da saúde cardiovascular e pode ser uma importante ferramenta de empoderamento desses pacientes no controle da própria saúde [Adeghe et al., 2024].

2.2 Previsão de pressão arterial como parte da medicina de precisão

O problema de prever a pressão arterial a partir de outros sinais vitais está fortemente alinhado com os princípios da Medicina de Precisão, abordagem médica que busca adaptar diagnósticos, tratamentos e estratégias de monitoramento ao perfil individual de cada paciente. Como destacado por Aziz et al. [2024], o uso de *wearables IoT*, aliados a técnicas de aprendizado de máquina e inteligência artificial, permite capturar sinais fisiológicos em tempo real, analisar padrões e tendências e, a partir disso, gerar intervenções personalizadas. No contexto da saúde de pacientes com lesão medular, isso representa um avanço significativo, já que possibilita um cuidado contínuo, remoto e individualizado. Além disso, tais tecnologias contribuem para a detecção precoce de

alterações, possibilitando respostas clínicas mais ágeis e eficazes.

3 Desenvolvimento

3.1 Fontes e características dos dados

A etapa de coleta de dados será realizada pelo Centro de Primatas de Davis (Califórnia). Os dados estão sendo coletados sob os protocolos IACUC. Os protocolos contam com etapas de anonimização e dessensibilização dos dados desta pesquisa, dispensando este projeto de pesquisa da aprovação junto ao Comitê de Ética em Pesquisa conforme Resolução nº 466, de 12 de dezembro de 2012, do Conselho Nacional de Saúde. Os dados utilizados neste projeto são fornecidos por pesquisadores da instituição mencionada acima e estão organizados em formato CSV. Entre as principais variáveis registradas estão a taxa de oxigenação do sangue, índice de perfusão, pulsação (bpm), pressão sistólica (mmHg), pressão diastólica (mmHg), pressão arterial média (mmHg), respirações por minuto (bpm). Essas variáveis permitem a análise detalhada do perfil fisiológico dos indivíduos e serão a base para a previsão da pressão arterial.

Dois tipos de aparelhos serão usados para compor o banco de dados. Oxímetros do tipo Radical-7 (R7)[†] serão responsáveis pela coleta de dados de batimentos cardíacos, oxigenação e índice de perfusão. Já os dados de pressão arterial e batimento cardíaco serão de responsabilidade de um aparelho de pressão arterial do tipo *beat-to-beat* (B2B) da empresa Caretaker (CT). A Figura 2 apresenta os aparelhos utilizados na pesquisa.

Figura 2: Aparelhos utilizados para a coleta de dados utilizada no projeto. Na esquerda, o aparelho R7, responsável pelos dados de oxigenação, perfusão e pulsação. Na direita, aparelho CT-B2B responsável pela pressão arterial e pulsação.

[†]https://www.masimo.com/products/bedside-solutions/radical-7/

Este aparelho fornece medidas de pressão arterial não-invasivas a cada batimento cardíaco e apresenta resultados semelhantes aos métodos invasivos de medição [Kwon et al., 2022]. Assim, os dados do aparelho CT-B2B e serão utilizados como padrão-ouro para este trabalho. No que, mesmo o aparelho CT-B2B fornecendo dados de forma não invasiva, a aquisição de dados via diferença de pressão não permite que o monitoramento ocorra de maneira efetiva fora do ambiente hospitalar.

3.2 Pré-processamento e integração dos dados

Devido aos dados serem provenientes de diferentes dispositivos com taxas de amostragem distintas, o pré-processamento é uma etapa crucial. As principais ações previstas incluem:

- **Filtragem de intervalos comuns:** Garantir que os dados sejam recortados para um intervalo de tempo comum entre as diferentes fontes, assegurando a sincronização temporal das medições realizadas pelos aparelhos R7 e CT-B2B.

- **Remoção de outliers:** Identificação e tratamento de valores atípicos que possam comprometer a qualidade do modelo.

- **Time Warping:** Aplicação de técnicas de Dynamic Time Warping (DTW) para ajustar e alinhar séries temporais vindas de aparelhos de coleta diferentes e com estruturas temporais distintas. O DTW é um algoritmo de programação dinâmica que permite comparar e alinhar sequências que podem variar em velocidade ou fase temporal.

  A Figura abaixo ilustra a diferença conceitual entre o alinhamento Euclidiano convencional e o alinhamento por DTW. No alinhamento Euclidiano (painel superior), os pontos são comparados de forma rígida baseando-se apenas em suas posições temporais, o que pode resultar em correspondências inadequadas quando as séries apresentam distorções temporais. Já o DTW (painel inferior) permite um alinhamento flexível, onde cada ponto de uma série pode ser mapeado para um ou mais pontos da outra série, compensando diferenças de velocidade e fase entre os sinais.

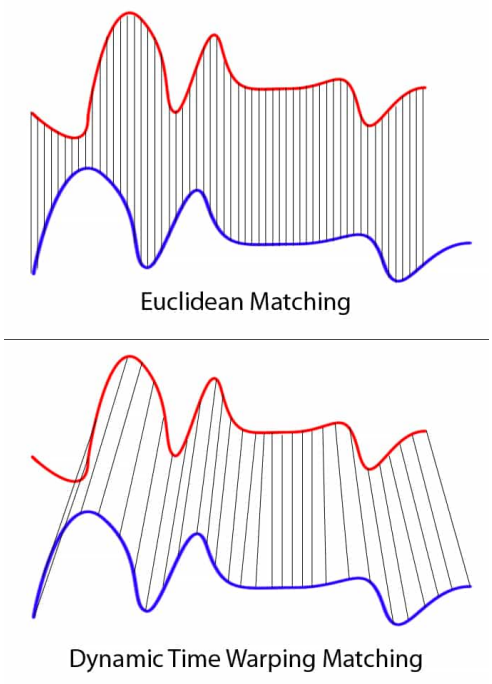


Figura 3: Comparação entre alinhamento Euclidiano (superior) e Dynamic Time Warping (inferior). Fonte: Databricks [2019].

O algoritmo constrói uma matriz de custos acumulados $D(i, j)$, onde cada elemento representa a distância mínima para alinhar as subsequências até os pontos i e j das séries X e Y, respectivamente. A relação de recorrência utilizada é:

$$D(i, j) = d(x_i, y_j) + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\} \quad (1)$$

onde $d(x_i, y_j)$ representa a distância Euclidiana entre os pontos correspondentes. Este alinhamento é essencial para garantir a sincronização adequada entre os sinais coletados por diferentes dispositivos, possibilitando a construção de um banco de dados integrado e coerente para o treinamento dos modelos de aprendizado de máquina.

- **Integração:** Como os dados estão em formato CSV, a integração é facilitada através do uso de bibliotecas como pandas e numpy, permitindo a fusão e padronização dos dados provenientes das diferentes fontes de coleta.

3.3 Modelagem

3.3.1 Definição da variável alvo

A variável alvo escolhida para este projeto é a Pressão Arterial Média (MAP), medida em mmHg. A MAP representa a pressão média nas artérias durante um ciclo cardíaco completo e é calculada a partir da pressão sistólica (SBP) e pressão diastólica (DBP), uma aproximação para o cálcula da MAP pode ser feita usando a seguinte relação:

$$MAP = \frac{SBP + 2 \cdot DBP}{3} \tag{2}$$

A escolha da MAP como variável alvo se justifica por diversos motivos. Primeiramente, a MAP é um indicador clinicamente relevante da perfusão tecidual e da função cardiovascular, apresentando menor variabilidade intra-individual em curtos períodos de tempo quando comparada às pressões sistólica ou diastólica isoladamente [Kwon et al., 2022]. Em segundo lugar, prever a pressão sistólica e diastólica simultaneamente exigiria maior poder computacional e aumentaria a complexidade do modelo, uma vez que seria necessário treinar múltiplos outputs ou modelos independentes. Além disso, caso a pressão de pulso (PP, do inglês *Pulse Pressure*), definida como:

$$PP = SBP - DBP \tag{3}$$

seja aproximadamente constante para um determinado indivíduo, é possível estimar essas pressões a partir da MAP prevista, utilizando as relações inversas:

$$DBP = MAP - \frac{1}{3}PP \tag{4}$$

$$SBP = MAP + \frac{2}{3}PP \tag{5}$$

Esta estratégia simplifica o problema de regressão para uma única variável alvo, reduzindo o custo computacional e facilitando a interpretação e avaliação do modelo.

Os valores de MAP obtidos pelo aparelho CT-B2B serão utilizados como padrão-ouro para o treinamento e validação dos modelos de aprendizado de máquina, tendo como entrada as demais variáveis fisiológicas coletadas pelos sensores R7 e CT-

B2B.

3.3.2 Arquiteturas de modelagem

Para a previsão da MAP, foram exploradas diferentes abordagens de regressão linear, permitindo comparar o desempenho de modelos simples com modelos regularizados que incorporam termos penalizadores.

Regressão linear simples Como modelo baseline inicial, foi empregada a regressão linear clássica [Géron, 2019], que busca minimizar o Erro Quadrático Médio (MSE). O modelo faz predições através de uma combinação linear das *features* de entrada, buscando ajustar os parâmetros que minimizem a função de custo:

$$J(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \tag{6}$$

onde y_i é o valor observado da MAP, $\boldsymbol{x}_i$ é o vetor de features e $\boldsymbol{\beta}$ é o vetor de coeficientes a ser estimado.

Modelos com termos penalizadores Para lidar com problemas de multicolinearidade, *overfitting* e realizar seleção automática de variáveis, foram explorados modelos de regressão linear com termos de regularização. Conforme descrito por Géron [2019], a regularização é alcançada adicionando um termo de penalização à função de custo MSE, restringindo os pesos do modelo e reduzindo sua capacidade de sobreajustar os dados:

- **Ridge Regression (L2)**: Adiciona penalização quadrática aos coeficientes (ℓ_2), mantendo todos os coeficientes no modelo mas reduzindo sua magnitude [Géron, 2019]. A função de custo é:

$$J(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{7}$$

- **Lasso Regression (L1)**: Utiliza penalização baseada no valor absoluto dos coeficientes (ℓ_1), com a característica distintiva de eliminar completamente os pesos das features menos importantes, definindo-os como zero e realizando seleção automática

de features [Géron, 2019]. A função de custo é:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{8}$$

- **Elastic Net**: Representa um meio-termo entre Ridge e Lasso, combinando ambas as penalizações L1 e L2 [Géron, 2019]:

$$J(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\left[\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right] \tag{9}$$

  onde $\alpha \in [0, 1]$ controla o balanço entre L1 e L2.

O termo de penalização L1 ($\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$) promove esparsidade, levando alguns coeficientes a zero e, consequentemente, realizando seleção automática de variáveis. Já o termo L2 ($\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p}\beta_j^2$) reduz a magnitude dos coeficientes, lidando com multicolinearidade entre features correlacionadas. A combinação de ambas as penalizações oferece maior estabilidade quando há alta correlação entre preditores, situação comum em dados fisiológicos multivariados.

4 Resultados e discussão

4.1 Ambiente computacional

Todos os experimentos foram executados em um notebook pessoal equipado com processador Intel Core i5-1135G7 (2.4 GHz), 24 GB de memória RAM e GPU Intel Iris Xe Graphics. O ambiente de desenvolvimento foi configurado com Python 3.11, utilizando as bibliotecas pandas 2.0.3 para manipulação de dados, numpy 1.26.4 para operações numéricas, fastdtw para o alinhamento temporal, e scikit-learn 1.4.2 para os modelos de machine learning.

4.2 Pré-processamento dos Dados

O pré-processamento dos dados foi implementado através de uma função customizada que integra múltiplas fontes de dados provenientes de diferentes dispositivos de

medição. A Figura 4 ilustra o fluxo completo do pipeline de pré-processamento, desde a conversão inicial dos timestamps até a consolidação final do dataset unificado.

Figura 4: Pipeline de pré-processamento dos dados mostrando as etapas principais.

4.2.1 Filtragem de intervalos comuns

O primeiro passo do pré-processamento consistiu na identificação e extração do intervalo temporal comum entre os diferentes dispositivos. Dada a natureza assíncrona da coleta de dados, cada aparelho (R7 e CT-B2B) possui seus próprios timestamps de início e término de gravação. Para garantir a sincronização adequada, foi implementado o seguinte procedimento:

1. Conversão das colunas temporais para o formato datetime de todos os DataFrames;
2. Identificação dos limites temporais de cada dispositivo:

$$t_{min}^{(i)} = \min(\text{Time}_i) \tag{10}$$

$$t_{max}^{(i)} = \max(\text{Time}_i) \tag{11}$$

onde i representa cada dispositivo;

3. Determinação do intervalo comum:

$$t_{start}^{comum} = \max(t_{min}^{(1)}, t_{min}^{(2)}, t_{min}^{(3)}, t_{min}^{(4)}) \tag{12}$$

$$t_{end}^{comum} = \min(t_{max}^{(1)}, t_{max}^{(2)}, t_{max}^{(3)}, t_{max}^{(4)}) \tag{13}$$

4. Filtragem de todos os DataFrames para manter apenas os dados dentro do intervalo $[t_{start}^{comum}, t_{end}^{comum}]$.

Este procedimento garantiu que apenas os dados coletados simultaneamente por todos os dispositivos fossem utilizados nas análises subsequentes, eliminando períodos onde algum aparelho não estava em operação.

4.2.2 Tratamento de valores ausentes e outliers

Após a filtragem temporal, foi realizado o tratamento de valores ausentes e outliers presentes nos dados. Definiu-se como outlier qualquer observação que apresentasse valor não-nulo na coluna `Events`, essa coluna continha anotações sobre a coleta, registrando principalmente erros ou intercorrências durante o processo de aquisição dos dados. Para garantir a qualidade do dataset, foram mantidas apenas as observações correspondentes a valores NaN nesta coluna (indicando ausência de erros registrados), e posteriormente a coluna foi removida, uma vez que não contribui diretamente para a previsão da pressão arterial.

Para as variáveis de saturação de oxigênio (SpO2) e índice de perfusão (Pi), valores representados como strings ('-') foram convertidos para zero. Posteriormente, todos os valores ausentes remanescentes foram preenchidos com zero.

4.2.3 Agregação multi-dispositivo e Dynamic Time Warping

Uma das etapas mais críticas do pré-processamento foi a integração dos dados provenientes de múltiplos dispositivos R7 (aparelhos 2, 3 e 4) com os dados do aparelho CT-B2B (aparelho 1, padrão-ouro). Este processo envolveu três sub-etapas principais:

Agregação dos dispositivos R7 Os três dispositivos R7 forneceram medições paralelas das mesmas variáveis fisiológicas (frequência cardíaca, SpO2, Pi). Para aumentar a

robustez das medições e reduzir o impacto de ruídos individuais de cada sensor, foi calculada a mediana dos valores de frequência cardíaca (PR bpm) entre os três dispositivos:

$$\mathrm{PR}_{mediana}(t) = \mathrm{mediana}(\mathrm{PR}_2(t), \mathrm{PR}_3(t), \mathrm{PR}_4(t)) \tag{14}$$

A escolha da mediana em vez da média se justifica por sua robustez a outliers, característica importante em dados fisiológicos que podem apresentar artefatos de movimento ou falhas momentâneas de leitura.

Alinhamento Temporal via Dynamic Time Warping Apesar da sincronização inicial baseada em timestamps, persistem diferenças sutis nas taxas de amostragem e latências de resposta entre os dispositivos R7 e CT-B2B. Para compensar essas distorções temporais e garantir um alinhamento preciso entre as séries, foi aplicado o algoritmo de Dynamic Time Warping entre a frequência cardíaca medida pelo CT-B2B (`HeartRate (bpm)`) e a mediana das frequências cardíacas dos dispositivos R7 (`Mediana_PR_bpm`).

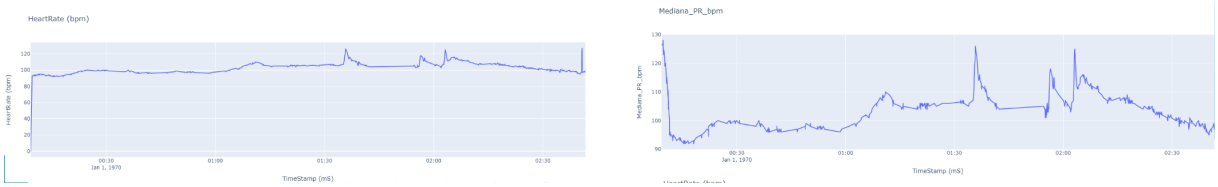
O DTW encontrou o caminho ótimo de alinhamento $\mathcal{P} = \{(i_1, j_1), (i_2, j_2), ..., (i_k, j_k)\}$ que minimiza a distância acumulada entre as séries. Este caminho estabelece uma correspondência entre os índices temporais dos dois conjuntos de dados:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \sum_{(i,j)\in\mathcal{P}} d(\mathrm{HR}_{CT-B2B}[i], \mathrm{PR}_{mediana}[j]) \tag{15}$$

onde d representa a distância Euclidiana entre pontos correspondentes.

Para avaliar quantitativamente a eficácia do alinhamento DTW, foi realizada uma comparação com o método de interpolação linear simples. Inicialmente, os dados dos dispositivos R7 foram interpolados para corresponder aos timestamps do CT-B2B, e o erro quadrático médio (MSE) entre as séries foi calculado. A Figura 5 apresenta as séries temporais antes do alinhamento DTW, onde é possível observar defasagens e descasamentos temporais entre os sinais.

A Figura 6 sobrepõe ambas as séries no mesmo gráfico, evidenciando as discrepâncias temporais. Utilizando interpolação linear simples, o MSE entre as séries foi de 32,8 $(\mathrm{bpm})^2$ indicando um desalinhamento significativo que poderia comprometer a qualidade do dataset integrado.



(a) HeartRate (CT-B2B) (b) Mediana PR (R7)

Figura 5: Séries temporais de frequência cardíaca antes do alinhamento DTW.

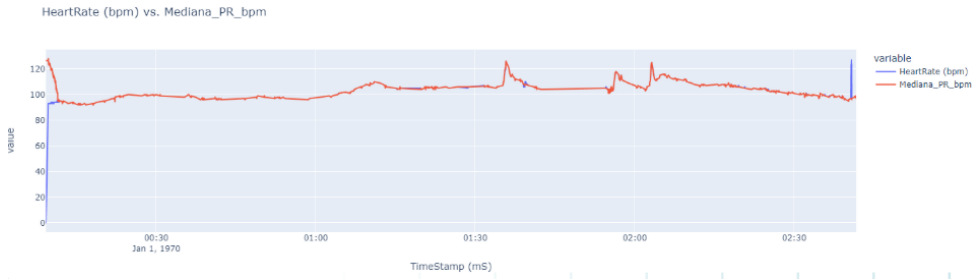


Figura 6: Comparação direta entre HeartRate (CT-B2B) e Mediana PR (R7) antes do alinhamento DTW.

Após a aplicação do algoritmo DTW, o alinhamento temporal entre as séries foi substancialmente melhorado. O MSE foi reduzido para 6,6 $(\text{bpm})^2$ representando uma redução de aproximadamente 80% no erro de alinhamento. Esta melhoria significativa demonstra a eficácia do DTW em compensar as distorções temporais e estabelecer correspondências mais precisas entre os dispositivos, garantindo a qualidade e a coerência do dataset final para o treinamento dos modelos de aprendizado de máquina.

Reamostragem estratificada Devido às diferentes taxas de amostragem entre os dispositivos, o alinhamento DTW resultou em um número de amostras do CT-B2B superior ao dos dispositivos R7. Para equalizar o número de observações mantendo a representatividade temporal dos dados, foi implementada uma estratégia de reamostragem estratificada:

1. O conjunto de dados do CT-B2B mapeado pelo DTW foi dividido em 10 partições temporalmente sequenciais;

2. De cada partição, foi amostrado aleatoriamente um número proporcional de ob-

servações, de modo que:

$$n_i = \left\lfloor \frac{n_{total}}{10} \right\rfloor + \begin{cases} 1, & \text{se } i < (n_{total} \mod 10) \\ 0, & \text{caso contrário} \end{cases} \tag{16}$$

onde n_i é o número de amostras selecionadas da partição i e n_{total} é o número total de observações dos dispositivos R7;

3. As amostras selecionadas foram ordenadas temporalmente para preservar a sequencialidade dos dados.

Esta abordagem garantiu que a distribuição temporal fosse mantida, evitando viés de seleção que poderia surgir de uma amostragem aleatória simples ou de downsampling uniforme.

4.2.4 Consolidação do dataset final

Após o alinhamento e reamostragem, os dados dos dispositivos R7 e CT-B2B foram concatenados horizontalmente, resultando em um dataset unificado contendo:

- Variáveis dos três dispositivos R7: SpO2, Pi e PR bpm de cada sensor;
- Mediana das frequências cardíacas dos R7;
- Variáveis do CT-B2B: frequência cardíaca, pressão arterial (SBP, DBP, MAP), entre outras;
- Timestamp unificado para todas as observações.

Colunas redundantes relacionadas a datas, horários duplicados e identificadores de época foram removidas para simplificar o dataset. O resultado final foi um conjunto de dados alinhado temporalmente, livre de valores ausentes e pronto para a etapa de feature engineering e modelagem.

A Tabela 1 apresenta um resumo estatístico do dataset após o pré-processamento, incluindo o número de observações, features disponíveis e estatísticas descritivas das principais variáveis.

Tabela 1: Resumo estatístico do dataset após pré-processamento (n = 4.894 observações)

Variável	**Média**	**Desvio Padrão**	**Mín**	**Máx**
MAP (mmHg)	26,82	4,35	20,00	64,00
SBP (mmHg)	50,69	4,28	41,00	95,00
DBP (mmHg)	15,86	2,17	11,00	38,00
HeartRate (bpm)	102,04	6,06	0,00	127,00
Respiration (bpm)	19,97	3,76	0,00	24,00
Mediana SpO2 (%)	93,39	7,44	0,00	100,00
Pi Device2	2,47	1,12	0,00	5,90
Pi Device3	2,52	1,05	0,00	4,30
Pi Device4	4,52	2,27	0,00	11,00
Mediana PR (bpm)	102,71	6,20	92,00	128,00

A Figura 7 apresenta a evolução temporal das principais variáveis cardiovasculares para um dos indivíduos (Loris) durante o experimento. O gráfico ilustra a dinâmica simultânea da frequência cardíaca (bpm) , pressão arterial média (mmHg), pressão sistólica (mmHg) e pressão diastólica (mmHg) ao longo do tempo.

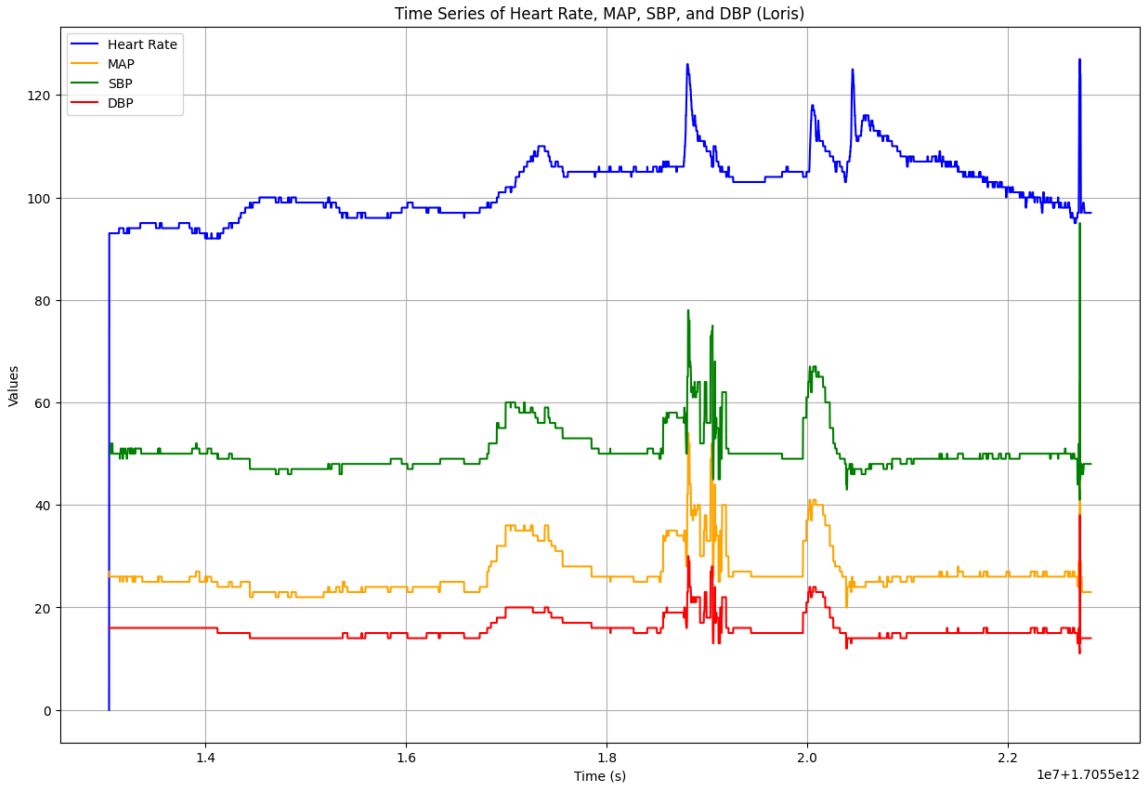


Figura 7: Série temporal das variáveis cardiovasculares para o indivíduo Loris.

4.3 Análise exploratória de dados

A análise exploratória teve como objetivo compreender as relações entre as variáveis fisiológicas coletadas e a variável alvo MAP. Esta etapa foi fundamental para identificar padrões temporais, correlações e a relevância preditiva de cada *feature* no conjunto de treinamento.

4.3.1 Análise de correlação

Inicialmente, foi realizada uma análise de correlação de entre todas as variáveis do conjunto de treinamento e o MAP. A matriz de correlação foi calculada utilizando o método `corr()` do pandas, seguida pela extração e ordenação dos coeficientes de correlação absolutos com a variável alvo.

A Figura 8 apresenta o ranking das correlações observadas.

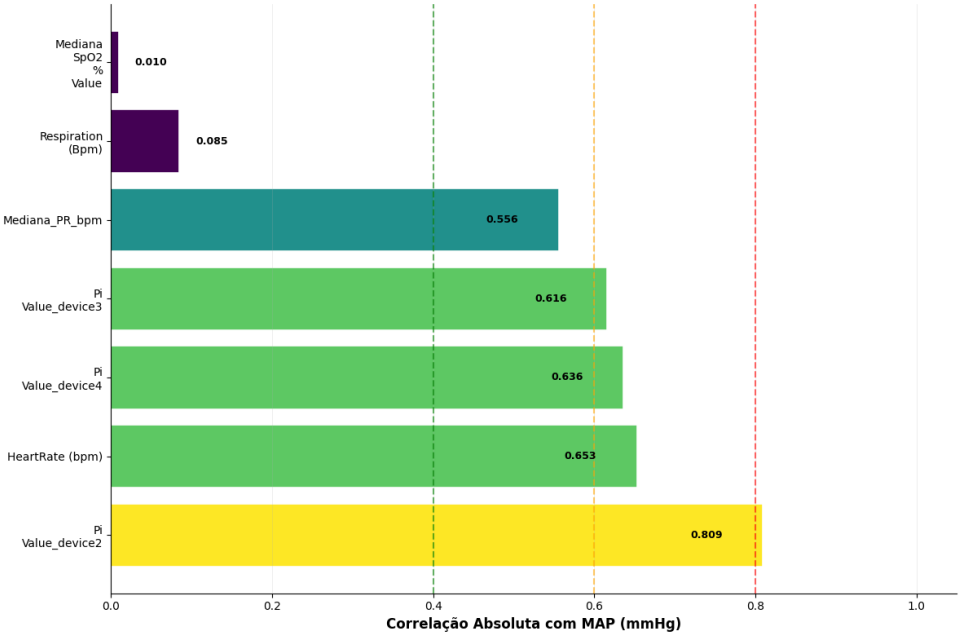


Figura 8: Correlação das features com a Pressão Arterial Média (MAP). As variáveis estão ordenadas por magnitude de correlação absoluta

Os resultados revelaram uma estratificação clara das *features* em diferentes níveis de correlação:

- **Maior correlação** : Pi Value_device2 ($r = 0.809$) apresentou a maior correlação com MAP entre as variáveis não-triviais. O *Perfusion Index Value* (Pi Value) é um

indicador da amplitude do sinal fotopletismográfico que reflete a perfusão periférica, justificando sua forte relação com a pressão arterial média.

- **Segunda e terceira maiores correlações**: HeartRate ($r = 0.653$) e Pi Value_device4 ($r = 0.636$) apresentaram, respectivamente, a segunda e terceira maiores correlações com MAP. A frequência cardíaca demonstrou relação substancial com MAP, consistente com a fisiologia cardiovascular, onde alterações na frequência cardíaca influenciam o débito cardíaco e, consequentemente, a pressão arterial. A presença de múltiplos dispositivos medindo Pi Value reforça a robustez desta variável como preditor.

- **Menores correlações**: Respiration ($r = 0.085$) e Mediana SpO2 ($r = 0.010$) apresentaram as menores correlações com MAP. A baixa correlação da SpO2 com a MAP é esperada neste contexto experimental, uma vez que ela deveria se manter relativamente constante durante todo o protocolo de coleta em condições controladas. Essa estabilidade da SpO2 reduz sua variabilidade e, consequentemente, sua capacidade preditiva para a pressão arterial média, mesmo sendo um importante indicador clínico em outras situações.

4.3.2 Análise temporal com Cross-Correlation

Para investigar relações temporais entre as variáveis, foi implementada uma análise de correlação cruzada (*cross-correlation*). Esta técnica permite identificar se existe defasagem temporal (*lag*) entre as variações de uma *feature* e as mudanças correspondentes no MAP.

A correlação cruzada foi calculada através da função `scipy.signal.correlate`, considerando lags de até 50 pontos temporais (ou 10% do conjunto de dados, o que fosse menor). Para cada *feature*, as séries temporais foram padronizadas antes do cálculo:

$$z_t = \frac{x_t - \mu_x}{\sigma_x} \tag{17}$$

onde μ_x e σ_x são a média e desvio padrão da série, respectivamente.

A correlação cruzada normalizada no *lag* τ foi então calculada como:

$$\rho_{xy}(\tau) = \frac{1}{n} \sum_{t} z_{x,t} \cdot z_{y,t+\tau} \tag{18}$$

Para cada *feature*, foram identificados:

- O *lag* ótimo (que maximiza $|\rho_{xy}(\tau)|$)
- A correlação no *lag* ótimo
- A correlação sem *lag* ($\tau = 0$)

A Figura 9 ilustra a análise de correlação cruzada para o índice de perfusão do Device2 (Pi Value_device2), a variável com maior correlação com MAP. O gráfico mostra a correlação em função do *lag* temporal, onde o ponto vermelho indica o *lag* ótimo encontrado ($lag = 5$), que maximiza a correlação em $r = 0.809$. A linha tracejada rosa representa a correlação sem *lag* ($lag = 0$), enquanto a linha tracejada verde indica o melhor lag identificado.

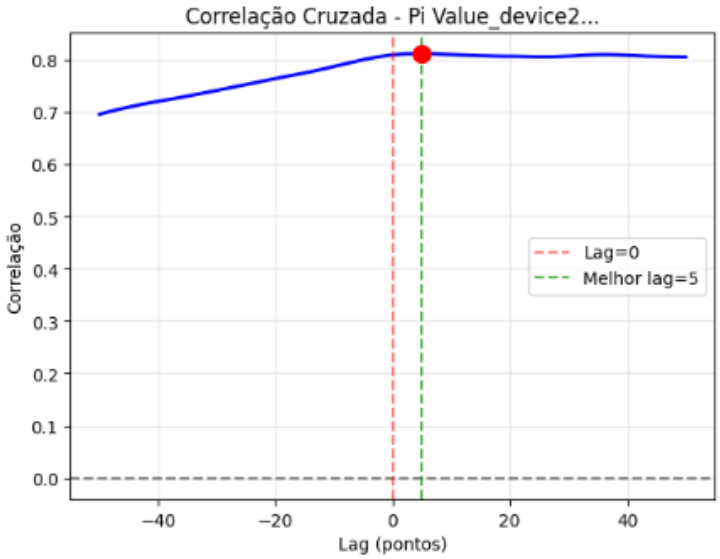


Figura 9: Correlação cruzada entre Pi Value_device2 e MAP. O pico de correlação ocorre com *lag* de 5 pontos temporais.

Os resultados revelaram que a maioria das correlações significativas ocorrem com *lag* pequeno (menor que 10), sugerindo relação temporal direta entre as variáveis fisiológicas e o MAP, sem atrasos substanciais entre causa e efeito.

4.3.3 Síntese e interpretação clínica

A análise exploratória multifacetada evidenciou que:

1. O Pi Value emerge como o indicador não-invasivo mais fortemente correlacionado com MAP ($r = 0.809$), sugerindo seu potencial como *feature* principal em modelos preditivos. A consistência desta correlação entre diferentes dispositivos (device2, device3, device4) reforça a robustez da medida.

2. A frequência cardíaca mantém relação significativa com MAP ($r = 0.653$), refletindo a interação cardiovascular entre débito cardíaco, frequência e resistência vascular periférica, descrita pela equação hemodinâmica fundamental.

3. A baixa correlação da SpO2 em relação ao MAP ($r = 0.010$) confirma que, no contexto experimental controlado deste estudo, a saturação de oxigênio permaneceu estável e relativamente constante, conforme esperado pelo protocolo de coleta.

4. A ausência de *lags* temporais significativos sugere que as variáveis respondem de forma quase simultânea às mudanças fisiológicas, facilitando a modelagem preditiva em tempo real sem necessidade de incorporar histórico temporal extenso.

4.4 Divisão dos dados e modelo base

Para a avaliação dos modelos, utilizou-se uma divisão temporal 80/20 dos dados, preservando a natureza temporal da série. Esta divisão foi implementada utilizando a biblioteca Scikit-learn, onde 80% dos registros temporais iniciais foram destinados ao treinamento e os 20% restantes para teste.

Inicialmente, aplicou-se uma Regressão Linear como modelo baseline, obtendo os seguintes resultados:

Tabela 2: Resultados da Regressão Linear

Métrica	**Treinamento**	**Teste**	**Diferença Absoluta**
MSE	6,2232	18,1443	11,9211
MAE	1,7561	3,4600	1,7039
R^2	0,7097	0,3564	0,3533

Os resultados indicam que o modelo sofre de overfitting, com desempenho significativamente superior no conjunto de treinamento em comparação com o teste, sugerindo a necessidade de técnicas de regularização.

4.5 Otimização com Grid Search e regularização

Os hiperparâmetros de regularização (λ e α) foram otimizados via validação cruzada, permitindo encontrar o equilíbrio ideal entre viés e variância de cada modelo.

- **Ridge**: 8 valores de λ testados: $\{0.001, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0\}$
- **Lasso**: 8 valores de λ testados: $\{0.001, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0\}$
- **Elastic Net**: 30 combinações de parâmetros, com $\lambda \in \{0.001, 0.01, 0.1, 0.5, 1.0, 5.0\}$ e $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

4.5.1 Resultados dos modelos regularizados

Tabela 3: Comparação dos modelos com regularização

Modelo	Parâm. escolhidos	R^2 Treino	R^2 Teste	MSE Teste	MAE Teste
Ridge	$\lambda = 5,0$	0,7097	0,3519	18,2702	3,4987
Lasso	$\lambda = 0,1$	0,6890	0,1477	24,0285	4,3721
Elastic Net	$\lambda = 0,1$, $\alpha = 0,1$	0,7097	0,3553	18,1751	3,4689

A Figura 10 apresenta o gráfico de dispersão comparando os valores reais de MAP com os valores previstos pelo modelo Elastic Net no conjunto de teste. A linha tracejada vermelha representa a predição perfeita (onde valores previstos seriam idênticos aos reais), permitindo avaliar visualmente a qualidade do ajuste e identificar padrões de erro sistemático.

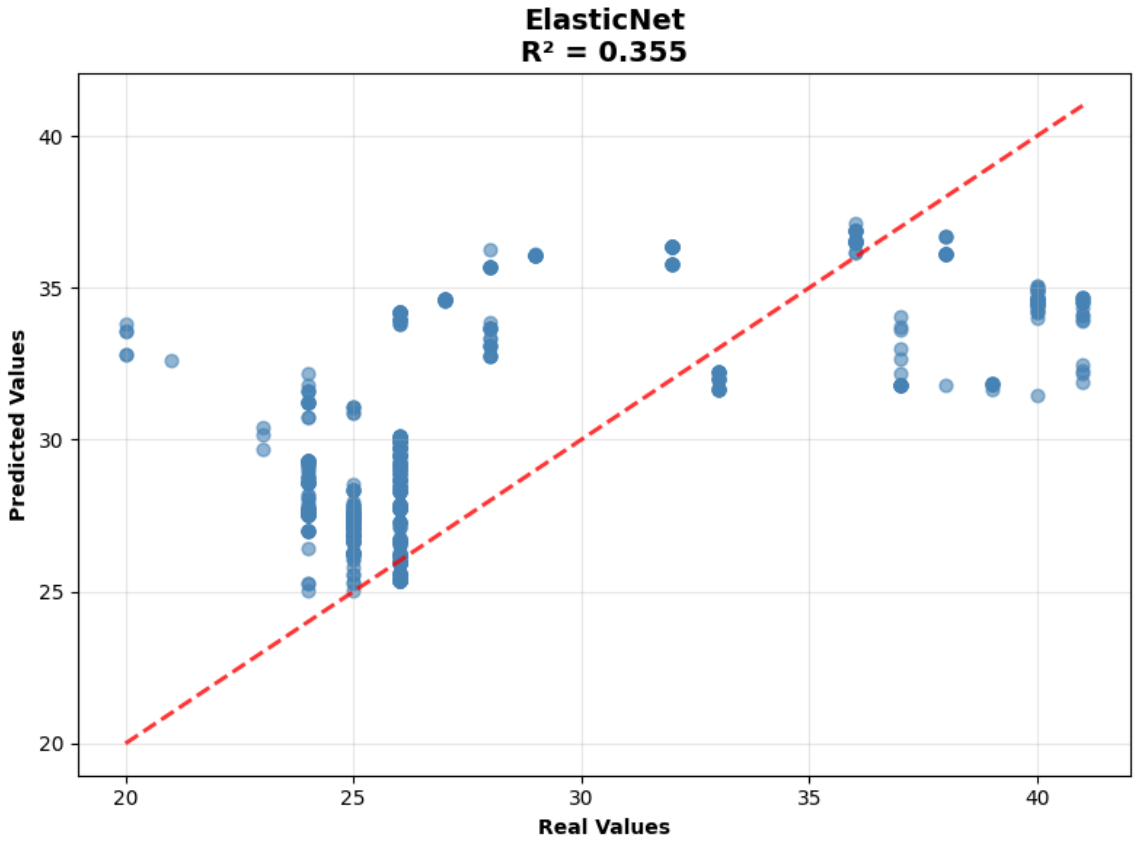


Figura 10: Comparação entre valores reais e previstos de MAP para o modelo Elastic Net.

4.6 Análise comparativa

Observa-se que o modelo Elastic Net obteve o melhor desempenho no conjunto de teste, com R^2 de 0,3553, MSE de 18,1751 e MAE de 3,4689. O modelo Ridge apresentou resultados similares, enquanto o Lasso demonstrou performance inferior, possivelmente devido à natureza dos dados onde muitas features podem ser relevantes.

A diferença persistente entre as métricas de treino e teste em todos os modelos sugere limitações intrínsecas nos dados ou a necessidade de abordagens de modelagem mais complexas para capturar adequadamente os padrões temporais subjacentes.

5 Conclusão

Este trabalho apresentou o desenvolvimento e avaliação de modelos de aprendizado de máquina para previsão não invasiva da MAP em pacientes com lesão medular,

utilizando dados multivariados de sensores vestíveis. A abordagem proposta demonstrou viabilidade técnica na integração de dados provenientes de múltiplos dispositivos de monitoramento, superando desafios significativos de pré-processamento através de técnicas como filtragem temporal, alinhamento via DTW e reamostragem estratificada.

A análise exploratória dos dados revelou que o Pi Value emerge como a variável preditora mais relevante para a estimativa da MAP. Esta descoberta está alinhada com os fundamentos fisiológicos, uma vez que o Pi Value reflete diretamente a perfusão tecidual periférica, intimamente relacionada com a dinâmica pressórica arterial. A modelagem através de técnicas de regressão linear com regularização permitiu explorar o equilíbrio entre capacidade preditiva e interpretabilidade do modelo, aspecto fundamental em aplicações clínicas.

Os resultados obtidos demonstram o potencial de sensores vestíveis como ferramentas de monitoramento cardiovascular contínuo para essa população específica, contribuindo para o desenvolvimento de soluções tecnológicas que promovam maior autonomia e qualidade de vida dos pacientes com lesão medular.

As limitações observadas no desempenho preditivo indicam a necessidade de abordagens mais sofisticadas para capturar completamente a complexidade dos padrões fisiológicos subjacentes. A persistente diferença entre métricas de treino e teste em todos os modelos sugere que fatores não capturados pelas variáveis disponíveis ou relações temporais mais complexas podem estar influenciando a dinâmica pressórica.

Como trabalhos futuros, recomenda-se a exploração de: (i) modelos baseados em séries temporais, como LSTM, capazes de capturar dependências temporais de longo prazo; (ii) engenharia de *features* mais sofisticada; (iii) validação em conjunto de dados mais amplo, com a inclusão de dados de outros macacos do estudo, permitindo avaliar a capacidade de generalização dos modelos desenvolvidos para diferentes indivíduos da mesma população experimental.

Apesar das limitações, este estudo representa um avanço significativo na direção de sistemas de monitoramento contínuo e não invasivo da pressão arterial, com potencial para integração em dispositivos vestíveis como *smartwatches*, oferecendo uma alternativa promissora aos métodos convencionais de aferição para pacientes com lesão medular e ampliando as opções de monitoramento cardiovascular na prática clínica.

Referências

Ehizogie Paul Adeghe, Chioma Anthonia Okolo, and Olumuyiwa Tolulope Ojeyinka. A review of wearable technology in healthcare: Monitoring patient health and enhancing outcomes. *OARJ of Multidisciplinary Studies*, 7(01):142–148, 2024.

Rabia Aziz, Firdaus Jawed, Sohrab Ahmad Khan, and Habiba Sundus. Wearable iot devices in rehabilitation: Enabling personalized precision medicine. In *Medical Robotics and AI-Assisted Diagnostics for a High-Tech Healthcare Industry*, pages 281–308. IGI Global, 2024.

Databricks. Understanding dynamic time warping, 2019. URL `https://www.databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html`. Acesso em: 05/11/25.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2 edition, 2019.

Younghoon Kwon, Patrick L Stafford, Kyle Enfield, Sula Mazimba, and Martin C Baruch. Continuous noninvasive blood pressure monitoring of beat-by-beat blood pressure and heart rate using caretaker compared with invasive arterial catheter in the intensive care unit. *Journal of cardiothoracic and vascular anesthesia*, 36(7):2012–2021, 2022.

A Vijayalakshmi, Deepa V Jose, and Sarwath Unnisa. Wearable sensors for pervasive and personalized health care. In *IoT in Healthcare and Ambient Assisted Living*, pages 123–143. Springer, 2021.