Classification of Individuals with Type 1 Diabetes by Neural Network from Cardiac Variability Data.

João Lucas Pereira dos Santos de Paula
Mestrando Eng. Elétrica - IEB-UFSC
Universidade Federal de Santa Catarina
Florianópolis, SC, Brasil
joaolucassantospaula@gmail.com

Débora Cristina Soares Bandeira
Mestranda Eng. Elétrica - IEB-UFSC
Universidade Federal de Santa Catarina
Florianópolis, SC, Brasil
debora.csbandeira@hotmail.com

***Abstract*—Type 1 diabetes is a chronic disease that affects millions of people around the world. Early detection and accurate classification of this condition are essential for adequate and effective treatment. In this context, the use of machine learning techniques, such as neural networks, has shown promise for improving accuracy in classifying individuals with type 1 diabetes. This article presents a study that uses data from people with cardiomyopathy due to type I diabetes mellitus based on ECG signals. Cardiac variability refers to variations in the time interval between heartbeats and is influenced by the autonomic nervous system. The study analyzed heart rate variability data from a group of subjects with type 1 diabetes and a group of healthy subjects. These data were pre-processed and used as input to train and test the neural network and thereby classify the patients.**

***Index Terms*—Type 1 Diabetes, Classification, Cardiac Variability, Machine Learning.**

I. Introdução

Um eletrocardiograma (ECG) é simplesmente uma gravação da atividade elétrica gerada pelo coração [1]. A análise do ECG é importante para o diagnóstico de doenças cardiovasculares, que são a principal causa de morte no mundo, de acordo com a Organização Mundial da Saúde. As arritmias cardíacas são as doenças cardiovasculares mais comuns e, por isso, sua classificação precisa é objeto de estudos biomédicos [2].

Diabetes mellitus é um distúrbio metabólico que causa hiperglicemia devido à incapacidade do corpo de metabolizar a glicose adequadamente, exigindo insulina para a absorção de glicose pelas células. Na diabetes do tipo 1, diagnosticado na infância, onde o sistema imunológico do corpo destrói as próprias células beta, resultando em deficiência de insulina; o tipo 2, comum em adultos acima de 40 anos, onde as células tornam-se insensíveis à insulina produzida ou não a usam adequadamente, conhecido como resistência à insulina. O diabetes gestacional, que é a intolerância à glicose desenvolvida durante a gravidez. O diabetes tipo 2 é o mais ocorrente [3].

O diabetes pode causar danos aos nervos, aumentar o risco de doenças cardíacas e derrames, levar à amputação de membros, causar retinopatia diabética e insuficiência renal, além de afetar outros órgãos vitais como coração, vasos sanguíneos, rins, nervos, pés e olhos [3].

A análise da VFC é realizada utilizando os intervalos R-R, que representam os espaços de tempo entre os picos R consecutivos no eletrocardiograma (ECG). Esses intervalos R-R compõem uma série temporal de eventos, em contraste com um sinal contínuo amostrado de forma uniforme [7].

A neuropatia autonômica cardiovascular (CAN) induzida pelo diabetes pode causar alterações no ECG, como mudanças no segmento ST-T, taquicardia sinusal, alterações na variabilidade da frequência cardíaca e prolongamento do QTc(QT corrigido). Foi confirmado que as dispersões do QT, QTc e ST são preditores de morte em pacientes diabéticos [4], [5].

Mesmo quando o organismo parece estar em equilíbrio fisiológico, a frequência cardíaca não permanece constante, apresentando variações de aceleração e desaceleração durante a inspiração e expiração, respectivamente [6].

O objetivo deste trabalho é utilizar de variáveis e características do ECG, para treinar uma rede neural afim de classificar pessoas com diabetes e sem diabetes do tipo 1.

II. Metodologia

A. Métodos e Features no Domínio do Tempo

As medidas de domínio de tempo envolvem o cálculo estatístico da média e variância do intervalo R-R dos dados de Variabilidade da frequência cárdica (VFC ou do inglês HRV). Os parâmetros importantes incluem a média da frequência cardíaca, RMSSD e SDNN. O RMSSD reflete a atividade parassimpática, indicando alterações de alta frequência que afetam a frequência cardíaca. No entanto, as medidas de domínio de tempo são propensas a outliers (valores discrepantes) e artefatos, o que exige a eliminação desses dados durante a análise [3].

Existem outras medidas baseadas nas diferenças de intervalo R-R, como a contagem de pares de intervalos RR adjacentes que diferem em mais de 50 ms (NN50) e a porcentagem da contagem de NN50 em relação à contagem total de intervalos R-R (pNN50). Essas medidas são principalmente influenciadas por variações de alta frequência (HF) na frequência cardíaca e são praticamente independentes das tendências de longo prazo. Além disso, essas medidas estão altamente correlacionadas entre si e, portanto, podem ser consideradas substitutas umas das outras [7].

B. Plataforma para Programação e sua Linguagem

O Colaboratory (Colab) é um produto do Google Research que permite a escrita e execução de código Python diretamente do navegador. É ideal para aprendizado de máquina, análise de dados e educação. E também um serviço de notebook Jupyter hospedado que não requer configuração e oferece acesso gratuito a recursos de computação, incluindo GPUs [8].

C. Dataset Utilizado e Limpeza dos Dados

Neste estudo de caso foi utilizado duas bases de dados, nelas contidas as pastas NoCan e EstCan, continham dados de dois tipos, a pasta NoCan pertencente a classe de dados de indivíduos saudáveis, e a pasta EstCan com classe de indivíduos com diabetes tipo 1. Sendo que existiam 8 pacientes no NoCan e 8 pacientes EstCan.

Deve-se levar em consideração que antes de retirar as *features* foi realizado uma analise nos sinas para ver se eram ECGs convencionais ou possuíam alguma diferença, o que fez ser retirado 3 pacientes da pasta NoCan.

1) Filtragem do Sinal: Neste trabalho utilizou-se a metodologia proposta por Neto e Seisdedos, a qual considerou o efeito de filtro passa alta do tipo Butterworth na obtenção do final da onda T de sinais de ECG [9], onde o trabalho [10], também utilizou a metodologia o obteve bons resultados. O qual é um item importante para achar o período QT em versões futuras deste estudo sobre a detecção da diabetes por meio do ECG.

A seguir temos as características de filtragem do sinal:

- Filtro passa baixa do tipo Butterworth de 4ª ordem, com frequência de corte de 40 Hz;
- Filtro passa alta do tipo Butterworth de 3ª ordem, com frequência de corte de 0,5 Hz.

Nesta etapa foi utilizada a biblioteca scipy.signal, para que fosse possível utilizar o filtro do tipo Butterworth.

2) Valores dos Picos R e Tempo dos Picos R: Para a detecção dos picos foi utilizado o algoritmo de Thomas S. Hamilton, que está citados em [11], [12]. Em cima dos conceitos abordados por Thomas S. Hamilton existe a função em Python deste algoritmo no pacote biosppy.signals, que auxilia na detecção dos picos R no exame de ECG, assim abrindo a possibilidade de calcular *features* de VFC no domínio do tempo.

3) Remoção de Outliers: A ideia inicial foi fazer um teste de treinamento com os dados discrepantes e os dados sem os outliers, utilizando a seguinte razão:

- Limite inferior = $\bar{x} - 2s$;
- Limite superior = $\bar{x} + 2s$;
- Logo $\bar{x} \pm 2s$.

Onde o $\bar{x}$ é a média da amostra e s o desvio padrão. Lembrando que foi definido uma função dentro do código que pode ser alterada para retirar mais ou menos janelas de outliers.

Para calcular a *features*, utilizou-se a definição para modelar a partir das funções que o Python traz, ficando da seguinte maneira:

```
mean_rr_ms = np.mean(rr_intervals)
sdnn_ms = np.std(rr_intervals) * 1000
rmssd_ms = np.sqrt(np.mean(np.diff(rr_intervals)**2)) * 1000
bpm = 60 / mean_rr_ms

diff_rr = np.diff(rr_intervals) * 1000
pnn50 = sum(abs(diff_rr) > 50) / len(diff_rr) * 100
pnn30 = sum(abs(diff_rr) > 30) / len(diff_rr) * 100
pnn20 = sum(abs(diff_rr) > 20) / len(diff_rr) * 100
```

Os valores que são multiplicados por 1000, são para trazelos para milisegundos pois os intervalos R-R foram captados em segundos.

D. Treinamentos

A biblioteca scikit-learn oferece diversos algoritmos de aprendizado, além de funções para pré-processar dados e avaliar modelos [19].

Para o treinamento foram retirados 4 pacientes, sendo eles 2 diabéticos(EstCan) e 2 não diabéticos(NoCan), para que fosse feita a validação dos treinos, escolhidos aleatoriamente pelos comandos do ”sklearn”.

1) MLP Classifier: É um classificador que treina iterativamente, atualizando os parâmetros a cada passo de tempo. Ele calcula as derivadas parciais da função de perda para atualizar os parâmetros e pode usar regularização para evitar o overfitting. Essa implementação funciona com matrizes numpy densas ou esparsas de valores de ponto flutuante [13].

2) Decision Tree Classifier: Uma árvore de decisão é uma ferramenta que utiliza um modelo em forma de árvore para auxiliar na tomada de decisões, considerando diferentes possíveis consequências, incluindo resultados aleatórios, custos de recursos e utilidade. As árvores de decisão são amplamente aplicadas em pesquisa operacional, especialmente na análise de decisões, para ajudar a identificar estratégias que tenham maior probabilidade de alcançar um objetivo específico. Além disso, as árvores de decisão também são uma técnica popular no campo de aprendizado de máquina [14].

3) Logistic Regression: A regressão logística é um modelo de aprendizado de máquina amplamente empregado na previsão de desfechos binários, inclusive no contexto do diagnóstico médico, e alguns casos para resoluções de problemas pode dar resultados bons quanto as redes neurais artificiais [15], [16].

4) Random Florest: É um método de aprendizado em conjunto que gera vários classificadores e combina os resultados deles. O RF cria múltiplas árvores de classificação e regressão (CART), cada uma treinada em uma amostra de inicialização dos dados de treinamento originais e busca em um subconjunto aleatoriamente selecionado de variáveis de entrada para determinar a divisão. As CARTs são árvores de decisão binárias que são construídas dividindo os dados em um nó em nós filhos repetidamente, começando com o nó raiz que contém a amostra completa de aprendizado [17].

5) KNN: é um algoritmo que usa casos anteriores para prever valores semelhantes em novos dados. Ele usa semelhança entre os recursos para prever valores para novos pontos. A regressão KNN tem duas abordagens: média dos K vizinhos mais próximos e média ponderada da distância inversa dos K vizinhos mais próximos. Usa funções de distância como

Euclidiana, Manhattan e Minkowski. Inicialmente, eliminamos valores nulos e substituímos valores ausentes antes de usar os dados para classificadores [18].

III. Resultados e Discussões

Em resumo, o código realizou pré-processamento de dados de ECG, aplicou filtros, extraiu informações de frequência e calculou medidas de variabilidade cardíaca a partir dos intervalos R-R dos complexos QRS. E assim conseguiu-se obter resultados de aprendizagem de maquina.

A. Resultados de Filtragem

Na Figura 1, pode-se ver o sinal original passando pelo filtro passa alta no segundo gráfico, e passa baixa no terceiro gráfico.

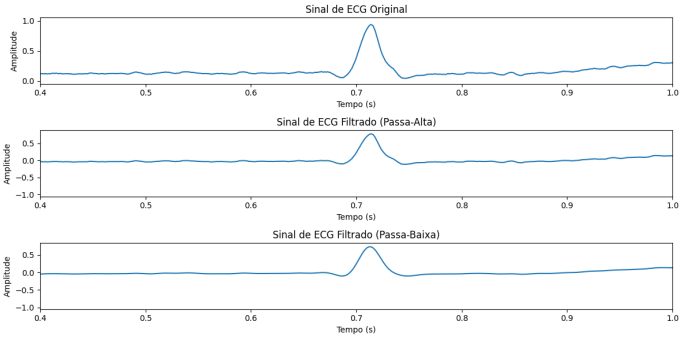

Fig. 1. Exemplo de Filtragem de Sinal

Pode-se ver que Utilização dos filtros, teve diferenças perceptíveis no sinal.

B. Resultados do Algoritmo para Adquirir os períodos R-R

Na Figura 2, é mostrado uma plotagem de um período de 10 segundos onde mostra a aquisição dos picos, para que abra a possibilidade de analises utilizando o R-R.

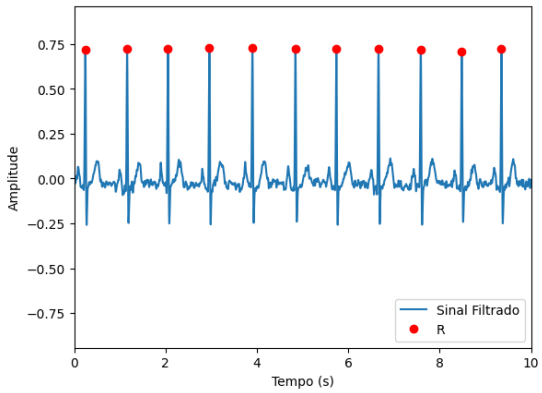

Fig. 2. Exemplo de Seleção de Picos

A resposta do algoritmo de Hamilton ao dataset foi um dos critérios utilizados para descartar os dados, pois o não funcionamento dele acarretaria em maus resultados de VFC.

C. Remoção de Outliers

As Figuras 3 e 4, trazem exemplos de boxplot e histograma, onde são mostrado pontos discrepantes no gráfico da esquerda, e no gráfico da direita a remoção destes dados por meio dos limites superiores e inferiores pré definidos.

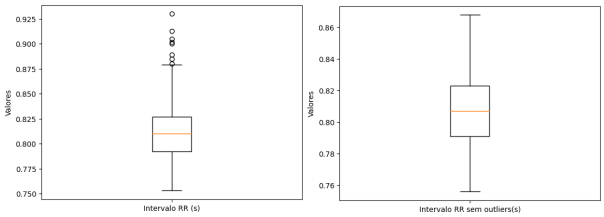

Fig. 3. Exemplo de Remoção de Outliers do BoxPlot

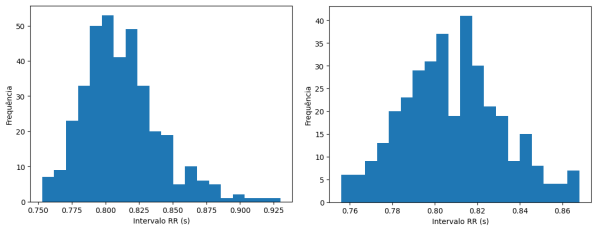

Fig. 4. Exemplo de Remoção de Outliers do Histograma

Foram feitos testes com e sem outliers, mas não houve diferença significativas nos dados e também em relação ao treinamento.

D. Features

Na Tabela I, podemos ver os valores de VFC(ou domínio do tempo) estão visualmente baixos para pessoas com diabetes, o que pode ser um indicativo ou caraterística crucial para os métodos classificarem esses dois grupos de indivíduos.

Classe	mean_rr	bpm	sdnn	rmssd	pnn50	pnn30	pnn20
0	0.6806	88.1569	18.6370	6.9578	0.0000	0.0000	0.7212
0	0.8083	74.2289	23.4728	14.2965	0.0000	2.5714	17.1429
0	1.0945	54.8215	26.8318	29.3131	4.1045	27.9851	49.2537
0	0.9343	64.2184	35.5977	36.7087	12.4579	35.3535	50.1684
0	0.8884	67.5378	30.4451	31.6952	10.5919	40.8100	60.4361
1	1.1395	52.6561	18.3586	21.9050	0.0000	16.3424	45.9144
1	0.6955	86.2632	6.4195	7.4078	0.0000	0.0000	0.0000
1	0.9086	66.0379	34.7694	22.8539	3.1348	10.9718	25.0784
1	0.7196	83.3832	9.9446	4.6925	0.0000	0.0000	0.2451
1	0.8882	67.5504	11.4654	8.5350	0.0000	0.0000	1.9169
1	0.7990	75.0921	13.3795	6.1432	0.0000	0.0000	0.2717
1	0.9517	63.0433	7.1995	8.1382	0.0000	0.0000	0.6689
1	0.6798	88.2653	6.2055	3.9503	0.0000	0.0000	0.0000

TABLE I
Classe 0 - não diabéticos ; Classe 1 - diabéticos

E. Mapa de Correlação da Features

Foi realizado a analise da relação entre a *features* para escolher a melhor, para que em estudos futuros minimize o processamentos dos algoritmos de aprendizagem de maquina pela escolhas das que possuem maior correlação (Figura 5), o

mapa de calor com correlação entres as *features*, neste mapa mostra um dos pontos mais importantes da pesquisas, que é a alta relação

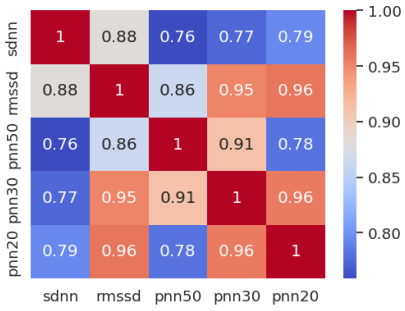

Fig. 5. Mapa de Calor e Correlação

F. Matriz de Confusão e Resultados de Acurácia dos Métodos Utilizados

Na Figura 6, temos um dos melhores resultados de treinamento que a utilização da Random Florest ou Floresta Aleatória, onde a acurácia foi de 100%, mas que não é um resultado factivel considerando a baixa quantidade de pacientes utilizados.

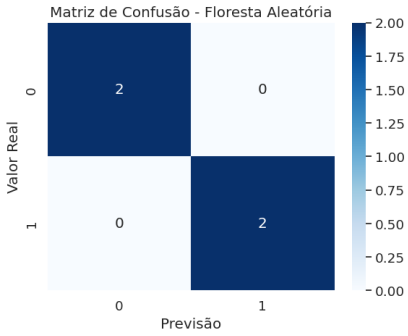

Fig. 6. Matriz de Confusão

Mais resultados são mostrado na Tabela II, que mostra os resultados dos outros métodos utilizados na pesquisa.

Método	**Acurácia (%)**
MLP Classifier	50
Decision Tree Classifier	100
Logistic Regression	75
Random Florest	100
KNN	75

TABLE II
COMPARAÇÃO DE MÉTODOS COM BASE NA ACURÁCIA

CONCLUSÃO

Nesta pesquisa conseguiu-se desenvolver um código que faz um pré-processamento dos dados afim de coloca-los em métodos de aprendizagem de maquina, mas para a obtenção de resultados é necessário a aquisição de uma base maior de dados para mais testes para melhores treinamentos. Em trabalhos futuros pode-se utilizar as ideias de "ultra-short HRV", que é utilizado para avaliar em tempos menores a variabilidade cardíaca, como mostrado em [20].

REFERENCES

[1] ALBERDI, Ane; AZTIRIA, Asier; BASARAB, Adrian. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. Journal of biomedical informatics, v. 59, p. 49-75, 2016.

[2] AFKHAMI, Rashid Ghorbani; AZARNIA, Ghanbar; TINATI, Mohammad Ali. Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals. Pattern Recognition Letters, v. 70, p. 45-51, 2016.

[3] SWAPNA, G.; SOMAN, K. P.; VINAYAKUMAR, R. Diabetes detection using ecg signals: An overview. Deep Learning Techniques for Biomedical and Health Informatics, p. 299-327, 2020.

[4] SAWICKI, P. T. et al. Prolonged QT interval as a predictor of mortality in diabetic nephropathy. Diabetologia, v. 39, p. 77-81, 1996.

[5] OKIN, Peter M. et al. Assessment of QT interval and QT dispersion for prediction of all-cause and cardiovascular mortality in American Indians: The Strong Heart Study. Circulation, v. 101, n. 1, p. 61-66, 2000.

[6] YASUMA, Fumihiko; HAYANO, Jun-ichiro. Respiratory sinus arrhythmia: why does the heartbeat synchronize with respiratory rhythm?. Chest, v. 125, n. 2, p. 683-690, 2004.

[7] KAMATH, Markad V.; WATANABE, Mari; UPTON, Adrian (Ed.). Heart rate variability (HRV) signal analysis: clinical applications. 2012.

[8] What is Colaboratory? Disponível em: https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory%2C%20or%20œColab%20for,learning%2C%20data%20analysis%20and%20education. Acesso 10 de maio de 2023.

[9] NETO, J. E.; SEISDEDOS, CR Vázquez. Influence of high pass filtering on the T-wave end estimation. In: VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014. Springer International Publishing, 2015. p. 627-630.

[10] RUFATO, Daiana Petry et al. Metodologia para a detecção de neuropatia autonômica subclínica em indivíduos com diabetes mellitus baseada na variabilidade de sinais fisiológicos. 2020.

[11] HAMILTON, Patrick S.; TOMPKINS, Willis J. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE transactions on biomedical engineering, n. 12, p. 1157-1165, 1986.

[12] HAMILTON, Patrick S. Open source ECG analysis. In: Computers in cardiology. IEEE, 2002. p. 101-104.

[13] MLPClassifier. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Acesso 10 de maio de 2023.

[14] ALBERT, Anitha Juliette; MURUGAN, R.; SRIPRIYA, T. Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. Research on Biomedical Engineering, v. 39, n. 1, p. 99-113, 2023.

[15] BOATENG, Ernest Yeboah; ABAYE, Daniel A. A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, v. 7, n. 4, p. 190-207, 2019.

[16] SONG, Jae H. et al. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. Academic radiology, v. 12, n. 4, p. 487-495, 2005.

[17] BREIMAN, Leo et al. Classification and regression trees. Wadsworth Int. Group, v. 37, n. 15, p. 237-251, 1984.

[18] PATNAIK, Srikanta; YANG, Xin-She; SETHI, Ishwar K. Advances in machine learning and computational intelligence. DOI: https://doi.org/10.1007/978-981-15-5243-4, 2021.

[19] RASHKA, S.; MIRDZHALILI, V. Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Birmingham, Mumbai. Packt, 2020.

[20] MELO, Hiago Murilo et al. Ultra-short heart rate variability reliability for cardiac autonomic tone assessment in mesial temporal lobe epilepsy. Epilepsy Research, v. 174, p. 106662, 2021.