

# Avaliando a Sobreamostragem de Dados Temporais de Marcha no Diagnóstico Automático de Doenças Neurodegenerativas

Ana Luísa de Bastos Chagas<sup>1</sup>, Giordana de Farias F. B. Bucci<sup>1</sup>, Juliana Paula Félix<sup>1</sup>, Afonso Ueslei da Fonseca<sup>1</sup>, Hugo A. D. do Nascimento<sup>1</sup>, Fabrizzio Soares<sup>1</sup>

<sup>1</sup>Instituto de Informática, Universidade Federal de Goiás

{analuisa23, giordanabucci}@discente.ufg.br,  
{julianafelix, afonsoueslei, hadn, fabrizzio}@ufg.br

**Abstract.** *Neurodegenerative diseases (NDDs) cause, among other symptoms, impairment of gait. Numerous studies analyze gait to, with the aid of artificial intelligence, assist in the diagnosis of NDDs. Due to the difficulty of collecting new data, the technique of oversampling through data windowing is often used. However, a previous study pointed to a possible bias in the training phase of classification algorithms using windowing techniques. This work investigates this bias, evaluating, in addition to the traditionally used cross-validation, a second approach, in which we address the pointed bias. The results indicate the need for extra care when dealing with oversampling of gait temporal data.*

**Resumo.** *Doenças neurodegenerativas (DNDs) causam, dentre outros sintomas, o comprometimento da marcha. Diversos estudos analisam a marcha para, com o auxílio da inteligência artificial, auxiliar no diagnóstico de DNDs. Devido à dificuldade de coleta de novos dados, a técnica de sobreamostragem através do janelamento de dados é frequentemente utilizada. No entanto, um estudo anterior apontou para um possível enviesamento na fase de treinamento de algoritmos de classificação utilizando técnicas de janelamento. Este trabalho investiga esse viés, avaliando, além da validação cruzada tradicionalmente utilizada, uma segunda abordagem, em que tratamos o enviesamento apontado. Os resultados indicam a necessidade de cuidados extras quando se lida com sobreamostragem de dados temporais de marcha.*

## 1. Introdução

A integração da computação na área da saúde tem revolucionado a maneira como lidamos com diagnósticos, tratamentos e monitoramento de pacientes [Bohr and Memarzadeh 2020]. Com o avanço da Inteligência Artificial (IA) e a capacidade de processar grandes conjuntos de dados, de maneira rápida e eficaz, diversos sistemas inteligentes para auxiliar no diagnóstico de doenças tem surgido. Dentre as aplicações de IA para este fim, pode-se citar aquelas focadas no diagnóstico de doenças pulmonares, como COVID-19, tuberculose e pneumonia [Jaeger et al. 2013, Li et al. 2020, Abideen et al. 2020, Çallı et al. 2021], doenças cardiovasculares [Ghumbre and Ghatol 2012], e também doenças neurodegenerativas [Dutta et al. 2009, Ning et al. 2018, Erdaş et al. 2021, Fraiwan and Hassanin 2021].

As Doenças Neurodegenerativas (DNDs), como as Doenças de Parkinson (DP), Doença de Huntington (DH), e Esclerose Lateral Amiotrófica (ELA), são doenças que afetam o cérebro e causam a degeneração progressiva dos neurônios. As DNDs não tem cura, e o seu diagnóstico é puramente clínico, através da análise e acompanhamento

da progressão dos sintomas, sendo geralmente diagnosticadas em fases já avançadas da doença [Mayeux 2003]. A falta de testes únicos para diagnosticar essas doenças, somadas à semelhança dos sintomas entre as diversas DNDs, justificam o interesse de se realizar o diagnóstico de maneiras alternativas.

Soluções alternativas para auxiliar no diagnóstico de DNDs vem sendo exploradas através da análise da marcha, visto que essa é uma das atividades prejudicadas com o comprometimento do sistema nervoso. Grande parte dos estudos que utilizam IA para este fim foram conduzidos utilizando-se a base de dados *Gait in Neurodegenerative Diseases (GaitNDD)*, como discutido em [Cicirelli et al. 2021]. A base GaitNDD está disponível publicamente no sítio da *Physionet* [Hausdorff 2000], e provê dados de até cinco minutos de caminhada de ritmo auto-determinado e sem apoio de um total de 64 indivíduos, sendo 15 diagnosticados com as doenças de Parkinson (DP), 20 com a doença de Huntington (DH), 13 com Esclerose Lateral Amiotrófica (ELA), e 16 de indivíduos de controle (CO).

Apesar da popularidade da base de dados *GaitNDD*, a aparente pequena quantidade de participantes do estudo (64 participantes) pode ser motivo de questionamento quando se trata de aplicações que envolvem inteligência artificial. De modo geral, quanto maior a quantidade de dados disponíveis para alimentar as fases de treinamento dos algoritmos de IA, maior a chance de generalização e confiabilidade do resultados reportados [Bujang and Adnan 2016, Rajput et al. 2023]. No entanto, coletar novos dados de marcha de pessoas com DNDs para novos estudos computacionais não é uma tarefa fácil, seja pela natureza rara das doenças, pela dificuldade de seu diagnóstico precoce, ou simplesmente pela dificuldade que pessoas com DNDs tem de manter uma marcha estável, sem apoio, por longos períodos de tempo.

Uma alternativa à coleta de novos dados consiste em se utilizar estratégias de sobreamostragem, ou *oversampling*. Com esta técnica, novos conjuntos de dados podem ser gerados ‘artificialmente’ a partir de um conjunto inicial, aumentando a quantidade de amostras que alimentam os algoritmos de inteligência artificial, e contribuindo com uma melhor generalização do sistema inteligente final. Uma das formas de se realizar a sobreamostragem de dados temporais, como é o caso dos dados de marcha disponíveis na GaitNDD, consiste na realização de um ‘janelamento’ de dados, fazendo com que uma única amostra de um participante selecionado sejam divididas em diversas amostras temporais de menor duração.

Essa estratégia é utilizada em [Ning et al. 2018], em que a partir dos sinais brutos dos pés esquerdo e direito, janelas de 640 pontos de dados foram construídas a partir dos 5 min de sinais coletados dos pés esquerdo e direito, com 50% de intersecção, resultando em um total de 16.612 amostras, das quais 75% foram utilizadas na fase de treinamento da rede neural convolucional (CNN) proposta, enquanto os outros 25% de amostras foram reservadas para a fase de teste. Os autores realizam a multiclassificação entre as quatro classes estudadas (DP, DH, ELA, e CO), e reportam uma acurácia de 99,50%. Em [Fraiwan and Hassanin 2021], uma abordagem similar de sobreamostragem foi utilizada, em que janelas de 30s de dados, sem intersecção, foram construídas a partir dos 5 minutos de dados disponíveis na base. Características estatísticas foram extraídas das janelas obtidas a partir das séries de intervalo de passada, apoio e balanço dos pés esquerdo e direito. Uma rede Adaboost foi avaliada utilizando-se *10-fold cross-validation*, e uma acurácia média 99,17% foi reportada para a tarefa de multiclassificação.

No entanto, os resultados reportados por [Ning et al. 2018] e [Fraiwan and Hassanin 2021] correspondem à classificação das janelas de dados

resultantes do processo de sobreamostragem, e não à classificação direta dos indivíduos analisados. Como anteriormente apontado por [Felix et al. 2022], essa observação leva a questionar se os classificadores construídos e os resultados reportados não estariam enviesados, uma vez que, sem o devido cuidado, janelas distintas, porém originárias de um mesmo indivíduo, poderiam figurar ao mesmo tempo nos conjuntos de treinamento e de teste. Esse trabalho desenvolveu uma hipótese que parte do princípio de que, mesmo num grupo de pessoas com a mesma condição patológica, cada indivíduo tem suas particularidades que se fazem presentes na marcha. Assim, acredita-se que janelas de dados originadas de um mesmo indivíduo carregam informações semelhantes e, por isto, não deveriam ser tratadas de modo independente na fase de avaliação.

Neste sentido, este trabalho tem como objetivo investigar a existência desse viés na classificação automática de doenças neurodegenerativas por meio da marcha. Para isso, classificadores clássicos da literatura, alimentados por características estatísticas, são comparados em 4 tarefas binárias de classificações: Parkinson vs. Controle (DP vs. CO), Huntington vs. Controle (DH vs. CO), Esclerose Lateral Amiotrófica vs. Controle (ELA vs. CO), e também Doença Neurodegenerativa vs. Controle (DND vs. CO), sendo este o caso em que dados de indivíduos com as 3 doenças investigadas são agrupadas em uma única classe e diferenciadas de indivíduos saudáveis. Para avaliar a hipótese da existência de viés nas classificações de janelas de dados, são apresentadas e analisadas duas formas de avaliação dos classificadores, tratando o viés em uma delas. Por fim, postula-se que os resultados apresentados com a avaliação sem viés são inferiores que os da abordagem com viés.

O restante desse trabalho é disposto da seguinte maneira: a Seção 2 detalha os materiais e métodos, passando por um detalhamento a respeito da base de dados usada, a operação de sobreamostragem dos dados, fase de extração de características, cenários e algoritmos de classificação implementados, e as duas abordagens de avaliação dos resultados utilizadas. Os resultados e sua discussão são apresentados na Seção 3. As conclusões seguem na Seção 4.

## 2. Materiais e Métodos

A Figura 1 apresenta de modo geral o método discutido neste trabalho. As seções seguintes detalham cada uma das etapas analisadas.



Figura 1. Fluxograma do método avaliado.

### 2.1. Base de Dados

Este estudo foi conduzido utilizando a base de dados *Gait in Neurodegenerative Diseases Database* (GaitNDD) [Hausdorff 2000], disponível publicamente no sítio da *Physionet* [Goldberger et al. e 13] e fornecida por [Hausdorff et al. 1997, Hausdorff et al. 2000]. A base provê dados de até cinco minutos de caminhada de 15 indivíduos com Doença de Parkinson (10 homens e 5 mulheres), 20 indivíduos com Doença de Huntington (6 homens e 14 mulheres), 13 indivíduos com Esclerose Lateral Amiotrófica (10 homens e 3 mulheres) e 16 indivíduos saudáveis de controle (2 homens e 14 mulheres). Informações clínicas de cada um dos 64 participantes, como idade, altura (m), peso (kg), e velocidade da marcha (m/s) também estão disponíveis na base. A Tabela 1 sumariza essas informações,

sendo  $\bar{x}$  o valor médio reportado, e SE o erro médio padrão (dado pela fórmula  $\sigma/\sqrt{N}$ , em que  $\sigma$  é o desvio-padrão da amostra, e  $N$  é o tamanho dessa amostra).

**Tabela 1. Informações dos participantes da base de dados GaitNDD.**

Grupo	Idade (anos)		Altura (m)		Peso (kg)		Velocidade (m/s)	
	$\bar{x}$	SE	$\bar{x}$	SE	$\bar{x}$	SE	$\bar{x}$	SE
DP	66,8	2,8	1,87	0,04	75,1	4,4	0,99	0,05
DH	46,6	2,8	1,83	0,02	72,1	3,9	1,15	0,08
ELA	55,6	3,6	1,74	0,03	77,1	5,9	1,05	0,06
CO	39,3	4,6	1,83	0,02	66,8	2,8	1,35	0,04

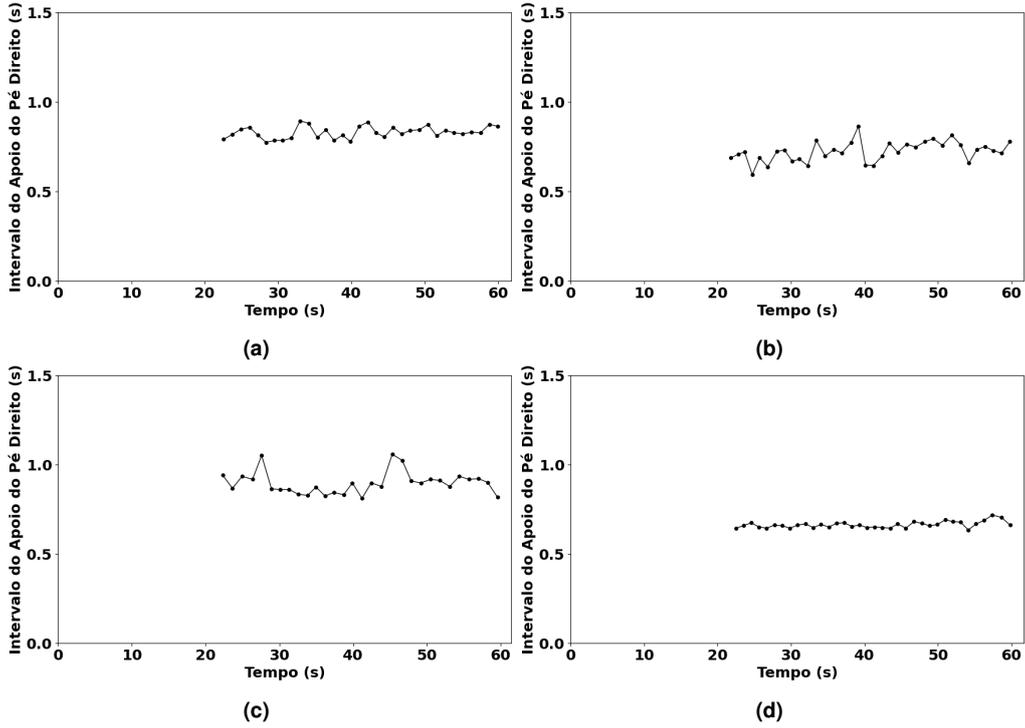
Como reportado por [Hausdorff et al. 1997, Hausdorff et al. 2000], sensores de força foram acoplados nas solas dos sapatos de cada participante. Em seguida, foi pedido que os participantes caminhassem por 5 minutos, num ritmo livre e sem utilização de qualquer tipo de apoio, por um corredor de 77 metros. Devido à limitação do corredor, era necessário que os participantes, eventualmente, dessem meia-volta para continuar a caminhada. Alguns efeitos de inicialização, relacionados ao momento em que os participantes iniciavam a caminhada, foram reportados pelos criadores da base.

A partir dos dados brutos de cada pé dos participantes (esquerdo e direito), referente à força aplicada ao solo durante a caminhada e amostrados a 300 Hz, algumas séries temporais de fases da marcha foram derivadas e disponibilizadas pelos criadores da base. São elas: o intervalo total da passada, ou *stride* (tempo decorrido entre o primeiro contato de dois passos consecutivos de um mesmo pé com o solo), o balanço, ou *swing* (tempo no qual o pé se mantém suspenso durante a passada), e o apoio, ou *stance* (período em que o pé permanece em contato com o solo). Em decorrência dos efeitos de inicialização, os primeiros 20 s de coleta de dados foram descartados pelos criadores da base para a geração desses intervalos.

Seguindo a linha de outros estudos [Felix et al. 2020, Felix et al. 2022], e visando contribuir com uma maior facilidade e conforto para o paciente, neste trabalho utiliza-se apenas o primeiro minuto da coleta de dados, e não os cinco minutos totais disponíveis na base de dados. Após análise cuidadosa dos dados disponíveis, observou-se que um dos indivíduos com a doença de Huntington (indivíduo #20) não possui dados coletados no primeiro minuto do experimento. Por este motivo, este indivíduo foi descartado da análise. Assim, as análises aqui apresentadas são feitas com base nas séries temporais do intervalo de apoio do pé direito extraídos de 63 indivíduos (15 com DP, 19 com DH, 13 com ELA e 16 indivíduos de controle). A Figura 2 ilustra essas séries temporais para um indivíduo de cada classe investigada. Note que os 20 primeiros segundos de amostras não estão disponíveis.

## 2.2. Sobreamostragem dos Dados

De posse do primeiro minuto das séries referentes ao intervalo da fase de apoio extraídas do pé direito de cada participante, inicia-se a fase de sobreamostragem dos dados. Para realizar este procedimento, optou-se pela técnica de sobreamostragem a partir da criação de janelas temporais de intervalos mais curtos de modo a maximizar a quantidade de dados provenientes de um mesmo participante. O tamanho de 10 s foi definido para cada uma das janelas criadas, e a separação foi feita sem que houvesse intersecção entre as



**Figura 2.** Intervalo de apoio do pé direito de um indivíduo com a Doença de Parkinson (a), Doença de Huntington (b), Esclerose Lateral Amiotrófica (c) e de Controle (d), tomados ao longo dos 60 segundos iniciais da coleta de dados.

janelas de dados. Deste modo, os dados de um único indivíduo passam a ser representados por 4 novas séries temporais, extraídas dos momentos de coleta referente aos segundos 20–30, 30–40, 40–50, e 50–60, visto que os primeiros 20s não estão disponíveis na base. O resultado é um conjunto de 252 janelas disjuntas de dados (63 indivíduos  $\times$  4 janelas).

### 2.3. Extração de características

Para cada uma das janelas de 10 segundos, foram extraídas quatro características significativas, sendo elas a média ( $\mu$ ), calculada conforme a Equação (Eq.) 1, o desvio-padrão ( $\sigma$ ) (Eq. 2), a potência média referente ao sinal da marcha (Eq. 3), sendo definida como a soma dos quadrados absolutos de suas amostras no domínio do tempo dividida pelo comprimento do sinal, e a entropia (Eq. 4), também chamada de entropia de informação. A entropia  $H(X)$  de uma variável aleatória  $X$  pode ser entendida como a taxa média na qual um resultado de  $X$  é produzido [Shannon 1948]. A escolha das características foi feita com base em um trabalho anterior [Felix et al. 2020], e o cálculo destas é descrito pelas Equações 1–4,

$$\mu(X) = \frac{1}{N} \sum_{i=1}^N x_i, \quad (1)$$

$$\sigma(X) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu(X))^2}, \quad (2)$$

$$Potência(X) = \frac{\sum_{i=1}^N x_i^2}{N}, \quad (3)$$

$$H(X) = - \sum_{i=1}^N p(x_i) \cdot \log(p(x_i)), \quad (4)$$

em que  $X = \{x_1, x_2, x_3, \dots, x_N\}$  é a série temporal de marcha com um número  $N$  de observações, sendo  $x_i$  a  $i$ -ésima observação de tal série, e  $p(x_i)$  a probabilidade do ponto  $x_i$ , calculada a partir de sua frequência de ocorrência na série temporal  $X$  dividida por  $N$ . Como foram utilizadas as séries temporais derivadas representativas dos períodos de apoio,  $N$  será o número de apoios ocorridos dentro dos 10 segundos da janela.

O grupo dessas quatro características a serem extraídas de cada janela é denominado conjunto de características, ou ainda *feature set*. A Equação 5 representa o formato geral dos *feature sets* obtidos. Tais conjuntos servirão para alimentar os algoritmos classificadores de aprendizado de máquina discutidos a seguir.

$$FS(X) = [\mu(X), \sigma(X), P(X), H(X)]. \quad (5)$$

## 2.4. Classificação

Após extraídas as características e definido o *feature set* de cada amostra, o próximo passo consiste em definir um modelo classificativo. Os cenários considerados foram: Doença de Parkinson versus Controle (DP vs. CO), Doença de Huntington versus Controle (DH vs. CO), Esclerose Lateral Amiotrófica versus Controle (ELA vs. CO), e um cenário combinado de doenças neurodegenerativas, que considera as três doenças em uma mesma classe, também versus Controle (DND vs. CO).

Para realizar as classificações, cinco classificadores clássicos da literatura são utilizados e comparados: *Support Vector Machine* (SVM, com kernel linear), *K-Nearest Neighbors* (KNN, com  $k = 5$ ), *Naive Bayes* (NB), *Linear Discriminant Analysis* (LDA) e *Decision Tree* (DT). A escolha dos algoritmos de aprendizado de máquina utilizados para classificação também se baseia em trabalhos anteriores [Felix et al. 2019, Felix et al. 2021], que já investigaram estes mesmos cenários de classificação para diagnóstico de doenças neurodegenerativas.

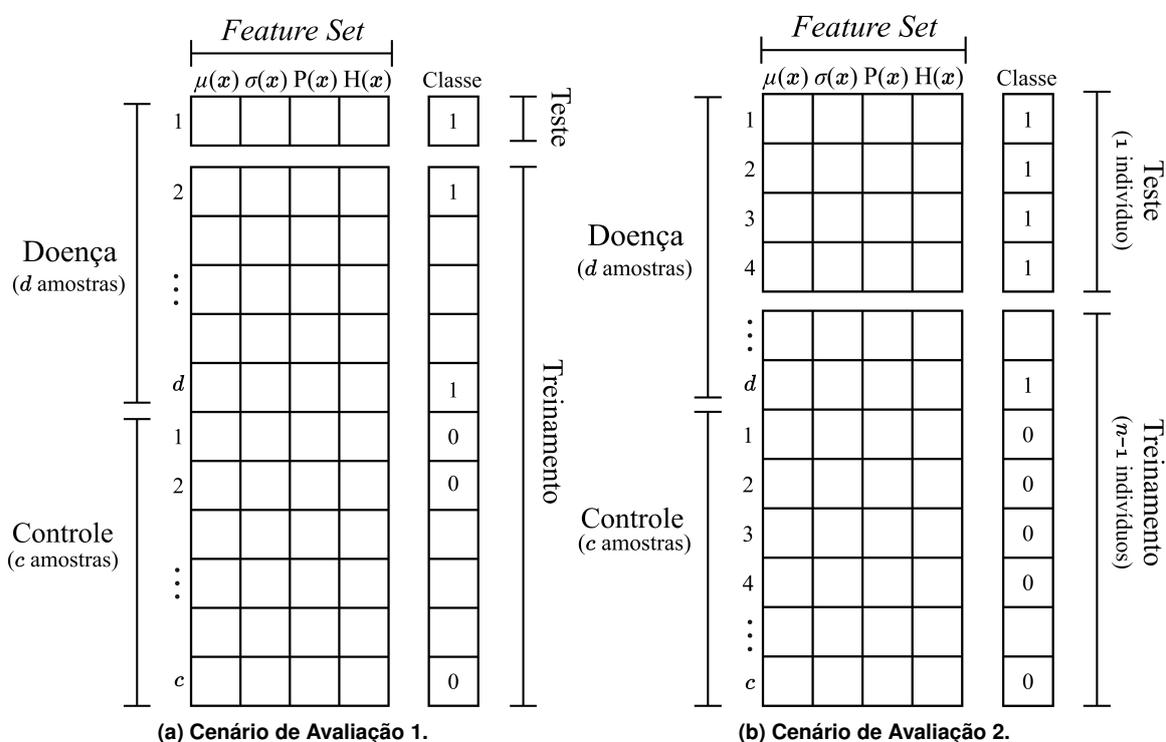
## 2.5. Avaliação

Uma das formas mais comuns de se avaliar a capacidade de generalização dos modelos classificativos é através de técnicas como a validação-cruzada, ou *cross-validation*. Em especial, quando se lida com uma quantidade pequena de dados em cada classe investigada, como é o caso da base de dados GaitNDD, um método especial da validação-cruzada, conhecida como *leave-one-out cross-validation* (LOOCV), é recomendada [Duda et al. 2000]. No LOOCV, considerando-se um total de  $n$  amostras,  $n - 1$  são utilizadas para a fase de treino, e a amostra que foi deixada de lado é utilizada para teste. O procedimento é, então, repetido  $n$  vezes, de modo que cada amostra seja usada exatamente uma vez como o conjunto de teste. Finalizadas as  $n$  repetições, o resultado reportado corresponde à média dos resultados obtidos em cada caso de teste avaliado.

Entretanto, com a realização de sobreamostragem dos dados, questionou-se a necessidade de adaptação da técnica de validação para este cenário. A hipótese parte do pressuposto de que janelas de dados originadas de um mesmo indivíduo carregam informações semelhantes e, por isso, não deveriam ser tratadas de modo independente

na fase de avaliação. Em outras palavras, ao realizar técnicas como a validação-cruzada, as janelas de dados originárias de um mesmo indivíduo não deveriam figurar ao mesmo tempo nos conjuntos de treino e de teste, de modo a não comprometer o resultado.

Desta forma, este trabalho propõe a avaliação dos resultados de classificação em dois cenários distintos, ilustrados na Figura 3. No primeiro cenário (Figura 3a), os dados obtidos de cada amostra (janela de 10 s) são considerados individualmente. Assim, em cada tarefa de classificação analisada (DP vs. CO, DH vs. CO, ELA vs. CO ou DND vs. CO), o LOOCV opera sobre  $n$  amostras, sendo  $n$  a quantidade total de janelas geradas pelas classes de dados envolvidas em cada tarefa. No segundo cenário (Figura 3b), a separação de amostras para treino e teste do LOOCV é feita em torno da quantidade total de participantes envolvidos na tarefa de classificação analisada. Com isso, cada participante selecionado para a fase de treinamento/teste contribuem com 4 amostras.



**Figura 3. Ilustração dos dois cenários de validação cruzada avaliados neste trabalho, sendo (a) o cenário de avaliação por janelas de dados e (b) o cenário de avaliação por indivíduo.**

Para exemplificar, no caso da tarefa de classificação DP vs. CO, que contém dados de 15 pessoas com DP e 16 CO, o Cenário de Avaliação 1 opera sobre um total de 124 amostras de 10s ( $15 \cdot 4 + 16 \cdot 4$ ), sendo que cada execução do LOOCV reserva uma amostra distinta para teste, enquanto as outras 123 amostras são utilizadas para treinamento do classificador. Neste cenário, o procedimento é repetido  $n = 124$  vezes. No Cenário de Avaliação 2, cada execução do LOOCV reserva as quatro amostras de um mesmo indivíduo para teste, enquanto as outras 120 amostras são utilizadas para treinamento do classificador. Neste cenário, o procedimento é repetido por  $n = 31$  vezes.

Após a finalização de todos os casos de teste considerados em cada cenário, os resultados médios de acurácia (taxa de acertos entre o número total de casos examinados), sensibilidade (taxa de amostras de indivíduos com uma das doenças foram corretamente

identificados), e especificidade (taxa de amostras de sujeitos saudáveis corretamente classificados) foram calculadas conforme segue,

$$\text{Acurácia (Acc.): } \frac{VP + VN}{VP + VN + FP + FN}, \quad (6)$$

$$\text{Sensibilidade (Sens.): } \frac{VP}{VP + FN}, \quad (7)$$

$$\text{Especificidade (Espec.): } \frac{VN}{VN + FP}, \quad (8)$$

onde VP são os verdadeiros positivos, VN são os verdadeiros negativos, FP são os falsos positivos e FN são os falsos negativos. O código para realizar a classificação e avaliação das métricas aqui reportadas foi desenvolvido utilizando Python 3.10.12, com a biblioteca Scikit-Learn.

### 3. Resultados e Discussão

Os resultados obtidos pelos 5 classificadores avaliados neste trabalho (SVM, KNN, NB, LDA e DT) para o diagnóstico das doenças de Parkinson, Huntington, Esclerose Lateral Amiotrófica e também para o diagnóstico de doenças neurodegenerativas (DNDs) utilizando-se a estratégia de sobreamostragem de dados e extração de características apresentadas na Seção 2 são apresentados nas Tabelas 2-5. Para cada tarefa avaliada, os resultados são apresentados considerando-se os dois cenários de avaliação discutidos na Seção 2.5. A acurácia média, sensibilidade e especificidade de cada modelo também são apresentadas.

Considerando-se, inicialmente, apenas o Cenário de Avaliação 1, em que as amostras resultantes do processo de *oversampling* são tratadas de forma independente, observa-se que os 5 classificadores analisados retornam acurácias médias que variam de 43,57% (SVM na tarefa DH vs. CO – Tabela 3) a 88,79% (DT na tarefa ELA vs. CO – Tabela 4).

Analisar os resultados apenas em termos de acurácia, no entanto, pode gerar uma falsa impressão de acuidade. Ao analisar os valores de sensibilidade e especificidade presentes nas tabelas, é possível compreender melhor o comportamento dos algoritmos construídos. Na Tabela 5, por exemplo, podemos perceber que, apesar de 74,60% de acurácia ser obtido pelo SVM na tarefa DND vs. CO, a especificidade é de 0%, ou seja, o modelo construído classificou erroneamente todas as amostras de controle como sendo

**Tabela 2. Resultados obtidos para DP vs. CO.**

Modelo	Cenário 1 ( <i>Leave-One-Out</i> por janelas)			Cenário 2 ( <i>Leave-One-Out</i> por pessoas)		
	Acc. Média(%)	Sens.(%)	Espec.(%)	Acc. Média(%)	Sens.(%)	Espec.(%)
SVM	66,94	53,33	79,60	62,90	45,00	79,69
KNN	74,19	65,00	82,81	71,77	61,67	81,25
NB	72,58	53,33	90,62	72,58	53,33	90,62
LDA	79,03	73,33	84,38	79,03	75,00	82,81
DT	81,45	80,00	82,81	70,16	61,67	78,12

**Tabela 3. Resultados obtidos para DH vs. CO.**

Modelo	Cenário 1 ( <i>Leave-One-Out</i> por janelas)			Cenário 2 ( <i>Leave-One-Out</i> por pessoas)		
	Acc. Média(%)	Sens.(%)	Espec.(%)	Acc. Média(%)	Sens.(%)	Espec.(%)
SVM	43,57	43,42	43,75	22,14	40,79	0,00
KNN	67,86	59,21	78,12	61,43	51,32	73,44
NB	77,86	67,11	90,62	77,86	67,11	90,62
LDA	79,29	67,11	93,75	75,00	63,16	89,06
DT	80,71	82,89	78,12	73,57	76,32	70,31

**Tabela 4. Resultados obtidos para ELA vs. CO.**

Modelo	Cenário 1 ( <i>Leave-One-Out</i> por janelas)			Cenário 2 ( <i>Leave-One-Out</i> por pessoas)		
	Acc. Média(%)	Sens.(%)	Espec.(%)	Acc. Média(%)	Sens.(%)	Espec.(%)
SVM	77,59	57,69	93,75	77,59	57,69	93,75
KNN	84,48	76,92	90,62	81,90	71,15	90,62
NB	77,59	59,62	92,19	78,45	59,62	93,75
LDA	81,03	65,38	93,75	80,17	63,46	93,75
DT	88,79	86,54	90,62	85,34	84,62	85,94

de indivíduos com alguma DND. Um comportamento similar também pode ser observado pelo LDA para essa mesma tarefa (74,60% acc, 93,62% sens. e 18,75% de espec.).

Assim, em todas as tarefas de classificação analisadas, o SVM obteve os piores resultados em termos de acurácia, sensibilidade e especificidade, enquanto as árvores de decisão (DT) apresentaram os melhores resultados: 81,45% acc., 80% sens. e 82,81% espec. para DP vs. CO; 80,71% acc., 82,89% sens., 78,12% espec. para DH vs. CO; 88,79% acc., 86,54% sens., 90,62% espec. para ELA vs. CO; e 84,52% acc., 90,43% sens. e 67,19% espec. para DND vs. CO.

No Cenário 2, a validação cruzada *leave-one-out* foi realizada sobre a quantidade de pessoas envolvidas em cada tarefa de classificação, de modo a evitar o viés presente no conjunto de treinamento. Ao analisar os resultados apresentados nas Tabelas 2– 5, é possível perceber que os valores reportados para acurácia, sensibilidade e especificidade

**Tabela 5. Resultados obtidos para DND vs. CO.**

Modelo	Cenário 1 ( <i>Leave-One-Out</i> por janelas)			Cenário 2 ( <i>Leave-One-Out</i> por pessoas)		
	Acc. Média(%)	Sens.(%)	Espec.(%)	Acc. Média(%)	Sens.(%)	Espec.(%)
SVM	74,60	100,0	0,00	74,60	100,00	0,00
KNN	76,19	81,91	59,38	73,02	79,79	53,12
NB	66,67	58,51	90,62	66,67	58,51	90,62
LDA	74,60	93,62	18,75	72,62	91,49	17,19
DT	84,52	90,43	67,19	81,75	87,77	64,06

para todas as tarefas de classificação e todos os modelos analisados, são, em sua maioria, inferiores que os reportados para o Cenário 1. Essa observação sugere a confirmação da hipótese de que janelas de dados de marcha provenientes de um mesmo indivíduo podem enviar a fase de teste dos classificadores. Isso porque, apesar de o teste ser realizado com uma porção de dado nunca antes vista pelo classificador, a presença de outras janelas de dados de marcha do mesmo indivíduo no conjunto de treinamento podem contribuir para o acerto da amostra desse indivíduo separada para teste.

Além da queda nos valores reportados quando comparados com o mesmo modelo no Cenário 1, outra observação constatada foi de que houve alteração no modelo que melhor se comportou em cada tarefa de classificação. Enquanto no Cenário 1 as árvores de decisão (DT) obtiveram os melhores resultados em todas as tarefas, no Cenário 2 o LDA foi o modelo que apresentou os melhores resultados (79,03% acc., 75,00% sens. e 82,81% espec.) para DP vs. CO (Tabela 2), e o *Naive Bayes* se destacou para a tarefa DH vs. CO (77,86% acc., 67,11% sens. e 90,62% espec.), apresentada na Tabela 3.

Com essas observações, destaca-se a necessidade do cuidado da separação dos dados de treino e teste quando técnicas de sobreamostragem de dados temporais de marcha a partir de janelamento são utilizadas, de modo que o classificador construído seja melhor capaz de generalizar os dados para o objetivo específico, neste caso, de diagnosticar doenças neurodegenerativas. Apesar dos resultados reportados apontarem para a confirmação do viés presente no Cenário 1, é importante destacar que a quantidade de dados presente no conjunto de treinamento se difere em 3 amostras. Além disso, apesar de ser reportado o resultado médio obtido das  $n$  interações do LOOCV, a quantidade de execuções dos 2 cenários são distintas, comprometendo uma avaliação estatística justa.

#### 4. Conclusão

Neste trabalho, foram extraídas 4 características estatísticas (média, desvio padrão, potência e entropia) do sinal de um minuto de marcha de indivíduos com a Doença de Parkinson, Doença de Huntington, Esclerose Lateral Amiotrófica, e também indivíduos de controle. Os dados do primeiro minuto de marcha de cada indivíduo foram divididos em janelas de 10 segundos sem intersecção, tomando a fase de apoio do pé direito como referência. Foram usados 5 classificadores clássicos de aprendizado de máquina, alimentados por tais características, para conduzir os testes - SVM, KNN, NB, LDA e DT - e houve validação posterior por *Leave-One-Out*. Foram estabelecidos dois cenários: 1, em que as amostras de janelas são tratadas de modo independente, e 2, em que a validação é feita sobre a quantidade de pessoas envolvidas em cada tarefa de classificação.

Desse modo, foi observado que, o Cenário 2, sendo este o que evita o viés presente no conjunto de treinamento, apresenta, de modo geral, valores inferiores de acurácia, sensibilidade e especificidade. Os resultados obtidos sugerem que janelas de dados de marcha provenientes de um mesmo indivíduo podem enviar a fase de teste dos classificadores, confirmando a necessidade de cuidados extras, frequentemente ignorados, quando abordagens de classificação com técnicas de *oversampling* de dados são utilizadas para o diagnóstico de doenças neurodegenerativas por meio da marcha.

Como sugestão de trabalhos futuros, propomos a análise da hipótese descrita com técnicas de sobreamostragem que utilizam janelas de duração menor, bem como janelas construídas com intersecções de dados, dada a limitação do presente trabalho em usar apenas janelas de 10s. Deste modo, o processo de *oversampling* culminaria em uma quantidade ainda maior de amostras a serem analisadas e utilizadas para alimentar os algoritmos de aprendizado de máquina, contribuindo para uma melhor generalização dos

métodos propostos e permitindo uma análise estatística mais equiparada entre os cenários. Com o devido cuidado para evitar viés na classificação, a estratégia proposta pode contribuir com um resultado mais fidedigno para a classificação e diagnóstico de doenças neurodegenerativas a partir da marcha.

## Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e Edital N° 30/2022 - PDPG - Solidariedade Acadêmica. A autora Juliana Paula Félix é Bolsista CAPES/BRASIL.

## Referências

- Abideen, Z. U., Ghafoor, M., Munir, K., Saqib, M., Ullah, A., Zia, T., Tariq, S. A., Ahmed, G., and Zahra, A. (2020). Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks. *Ieee Access*, 8:22812–22825.
- Bohr, A. and Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, pages 25–60. Elsevier.
- Bujang, M. A. and Adnan, T. H. (2016). Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of Clinical and Diagnostic Research : JCDR*, 10:YE01–YE06.
- Çalli, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., and Murphy, K. (2021). Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125.
- Cicirelli, G., Impedovo, D., Dentamaro, V., Marani, R., Pirlo, G., and D’Orazio, T. (2021). Human gait analysis in neurodegenerative diseases: A review. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience.
- Dutta, S., Chatterjee, A., and Munshi, S. (2009). An automated hierarchical gait pattern identification tool employing cross-correlation-based feature extraction and recurrent neural network based classification. *Expert Systems*, 26(2):202–217.
- Erdaş, Ç. B., Sümer, E., and Kibaroglu, S. (2021). Neurodegenerative disease detection and severity prediction using deep learning approaches. *Biomedical Signal Processing and Control*, 70:103069.
- Felix, J., Fonseca, A. U., Nascimento, H., and Guimarães, N. (2022). Rede Neural Multicamadas para Classificação de Doenças Neurodegenerativas a partir de Sinais de Marcha. In *Anais do XXIV Congresso Brasileiro de Automática*, pages 1354–1361. SBA.
- Felix, J. P., do Nascimento, H. A. D., Guimarães, N. N., Pires, E. D. O., da Silva Vieira, G., and de Souza Alencar, W. (2020). An Effective and Automatic Method to Aid the Diagnosis of Amyotrophic Lateral Sclerosis Using One Minute of Gait Signal. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2745–2751. IEEE.
- Felix, J. P., Nascimento, H. A. D. d., Guimarães, N. N., Pires, E. D. O., Da Fonseca, A. U., and Vieira, G. D. S. (2021). Automatic Classification of Amyotrophic Lateral Sclerosis through Gait Dynamics. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1942–1947. IEEE.

- Felix, J. P., Vieira, F. H. T., da Silva Vieira, G., Franco, R. A. P., da Costa, R. M., and Salvini, R. L. (2019). An Automatic Method for Identifying Huntington's Disease using Gait Dynamics. In *31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2019, Portland, OR, USA, November 4-6, 2019*, pages 1659–1663. IEEE.
- Fraiwan, L. and Hassanin, O. (2021). Computer-aided identification of degenerative neuromuscular diseases based on gait dynamics and ensemble decision tree classifiers. *Plos one*, 16(6):e0252380.
- Ghumbre, S. U. and Ghatol, A. A. (2012). Heart disease diagnosis using machine learning algorithm. In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*, pages 217–225. Springer.
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000 (June 13)). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: <http://circ.ahajournals.org/content/101/23/e215.full> PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
- Hausdorff, J. (2000). Gait in neurodegenerative disease database. <https://physionet.org/content/gaitnidd/1.0.0/>. Acessado por último em 07 de março de 2024.
- Hausdorff, J. M., Lertratanakul, A., Cudkowicz, M. E., Peterson, A. L., Kaliton, D., and Goldberger, A. L. (2000). Dynamic Markers of Altered Gait Rhythm in Amyotrophic Lateral Sclerosis. *Journal of Applied Physiology*, 88(6):2045–2053.
- Hausdorff, J. M., Mitchell, S. L., Firtion, R., Peng, C.-K., Cudkowicz, M. E., Wei, J. Y., and Goldberger, A. L. (1997). Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease. *Journal of Applied Physiology*, 82(1):262–269.
- Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S., et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245.
- Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., et al. (2020). Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*.
- Mayeux, R. (2003). Epidemiology of neurodegeneration. *Annual Review of Neuroscience*, 26(1):81–104.
- Ning, Z., Li, L., and Jin, X. (2018). Classification of neurodegenerative diseases based on CNN and LSTM. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pages 82–85. IEEE.
- Rajput, D., Wang, W.-J., and Chen, C.-C. (2023). Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, 24.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.