Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6402

Artigo de Pesquisa/Research Paper

Investigação de Enviesamento de Modelos de Aprendizado de Máquina para Diagnóstico de Doenças Neurodegenerativas via Registros de Marcha e Voz

Investigation of Bias in Machine Learning Models for the Diagnosis of Neurodegenerative Diseases Using Gait and Voice Records

Ana Luísa de Bastos Chagas [Universidade Federal de Goiás | *analuisa23@discente.ufg.br*]
Giordana de Farias F. B. Bucci [Universidade Federal de Goiás | *giordanabucci@discente.ufg.br*]
Juliana Paula Félix [Pontifícia Universidade Católica de Goiás e Universidade Federal de Goiás | *julianafelix@ufg.br*]
Rogerio Lopes Salvini [Universidade Federal de Goiás | *rogeriosalvini@ufg.br*]
Hugo A. D. do Nascimento [Universidade Federal de Goiás | *hadn@ufg.br*]
Fabrizzio Soares [Universidade Federal de Goiás | *fabrizzio@ufg.br*]

Instituto de Informática (INF) - UFG, Alameda Palmeiras, Quadra D, Câmpus Samambaia, 74690-900, Goiânia, Goiás, Brasil.

Resumo. Este trabalho sintetiza os principais achados de uma pesquisa sobre vieses em modelos de aprendizado de máquina para o diagnóstico de doenças neurodegenerativas por análise da marcha e da voz. Avaliamos como técnicas de sobreamostragem, como o janelamento da marcha e o uso indiscriminado de múltiplas amostras de voz por pessoa, inflam métricas de desempenho ao tratar amostras do mesmo sujeito como independentes. Comparando protocolos que ignoram ou preservam essas dependências em dois conjuntos de dados, observamos que desconsiderá-las inflaciona as métricas, enquanto sua preservação fornece avaliações mais fiáveis. Os resultados destacam a importância da segregação adequada das amostras para obter modelos diagnósticos confiáveis.

Abstract. This paper synthesizes the main findings of a study on biases in machine learning models for diagnosing neurodegenerative diseases through gait and voice analysis. We evaluated how oversampling techniques, such as gait windowing and the indiscriminate use of multiple voice samples per person, inflate performance metrics when treating samples from the same subject as independent. By comparing protocols that ignore or preserve these dependencies in two datasets, we observed that disregarding them inflates the metrics, while preserving them provides more reliable assessments. The results emphasize the importance of proper segregation of samples to obtain diagnostic models.

Palavras-chave: Doenças neurodegenerativas, diagnóstico, enviesamento, aprendizado de máquina

Keywords: Neurodegenerative diseases, diagnosis, biases, machine learning

Recebido/Received: 22 August 2025 • **Aceito/Accepted:** 17 December 2025 • **Publicado/Published:** 29 December 2025

1 Introdução

Doenças neurodegenerativas (DNDs), como Doença de Parkinson (DP), Doença de Huntington (DH) e Esclerose Lateral Amiotrófica (ELA), são de caráter progressivo e incurável, causando deterioração neuronal e trazendo risco de morte ao paciente [Heemels, 2016]. Embora cada uma apresente características próprias, pacientes costumam manifestar sintomas como perda de memória, movimentos involuntários, dificuldades motoras, problemas na fala e instabilidade na marcha [Berman and Bayati, 2018]. A maioria das DNDs também conta com a ausência de exames diagnósticos definitivos, fazendo com que a identificação da doença dependa da observação clínica dos sintomas e sua progressão, o que frequentemente leva a um diagnóstico tardio e impreciso [Mayeux, 2003].

Considerando os impactos debilitantes dessas enfermidades, a detecção precoce pode influenciar significativamente a evolução dos pacientes, permitindo intervenções oportunas, melhorando sua qualidade de vida e otimizando o uso dos recursos de saúde. Nesse contexto, estratégias alternativas para diagnóstico têm sido amplamente investigadas, e o aprendizado de máquina (ML, ou *machine learning*) vem se destacando como uma ferramenta promissora para a análise de dados clínicos, auxiliando na identificação de padrões relevantes de diversas maneiras, como no uso de dados de *smartwatch* [Varghese *et al*., 2024], exames de imagens como ressonância magnética e PET-CT [Vyas *et al*., 2022; Noella and Priyadarshini, 2023], e principalmente em dados acústicos [Yasar *et al*., 2019] e de marcha [Erdaş *et al*., 2021].

Apesar do bom desempenho de modelos de ML na distinção entre pessoas saudáveis de controle (CO) e pacientes com DNDs, limitações metodológicas podem comprometer a validade dos achados. A escassez de dados, comum em doenças raras, frequentemente exige técnicas que aumentem a quantidade de dados, como janelamento na marcha e o uso direto de múltiplas amostras de voz coletadas de uma mesma pessoa. No entanto, a falta de controle na separação dessas amostras fruto de técnicas de aumento pode introduzir vieses. Pesquisas anteriores [Felix *et al*., 2022] indicam que esse problema pode gerar estimativas de desempenho excessivamente otimistas, não refletindo a real capacidade do modelo no diagnóstico.

Este estudo tem como objetivo avaliar o viés presente

nos algoritmos de aprendizado de máquina na classificação de doenças neurodegenerativas (DNDs), com foco na análise de dados de marcha e dados de voz, que são comumente utilizados para auxiliar no diagnóstico automático dessas doenças. A avaliação foi conduzida em diferentes domínios, avaliando as consequências desse viés na real eficácia dos modelos de diagnóstico. Para isso, o estudo realiza três tarefas de classificação, começando com a análise do impacto desse viés na diferenciação entre pacientes com DP, DH e ELA no domínio da marcha. Em seguida, a análise avança para a distinção entre DP e indivíduos controle (CO) a partir de dados acústicos, destacando a persistência do viés em decorrência da má organização dos conjuntos de treinamento dos dados.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta a motivação e uma contextualização teórica referente ao trabalho. A Seção 3 descreve o método proposto. Os resultados são apresentados na Seção 4, e sua consequente discussão está na Seção 5. Nossas conclusões são apresentadas na Seção 6.

2 Motivação

Nos últimos anos, diversos estudos têm aplicado aprendizado de máquina à análise de padrões de marcha e voz para o diagnóstico de doenças neurodegenerativas (DNDs), atingindo altos níveis de acurácia na diferenciação entre indivíduos saudáveis e pacientes [Naranjo *et al*., 2016; Bielby *et al*., 2020; Dutta *et al*., 2009; Ning *et al*., 2018; Fraiwan and Hassanin, 2021]. Em muitos desses trabalhos, técnicas como a segmentação dos sinais em múltiplas janelas são utilizadas para aumentar a quantidade de amostras disponíveis, favorecendo o desempenho dos modelos de inteligência artificial. No entanto, a forma como essas amostras são distribuídas entre os conjuntos de treino e teste pode impactar significativamente a avaliação do modelo, o que será discutido posteriormente nesta seção.

A sobreamostragem (*oversampling*) tem sido amplamente empregada para contornar a escassez de dados disponíveis para o estudo de DNDs. Para dados temporais, essa técnica pode ser realizada por meio de 'janelamento', que segmenta uma única amostra de um participante em várias amostras menores, isto é, com menos pontos de dados. Um conjunto de notáveis trabalhos desse campo de pesquisa faz uso dessa abordagem.

Ning *et al*. [2018] aplicaram janelas de 640 pontos de dados com 50% de sobreposição, obtendo 16.612 amostras totais de sinais de marcha. Treinando uma CNN com 75% dos dados, alcançaram 99,50% de acurácia para as mesmas classes. Fraiwan and Hassanin [2021] extraíram janelas de 30s, sem sobreposição, a partir de 5 minutos de marcha, usando características estatísticas extraídas das fases da passada para treinar uma rede Adaboost, alcançando 99,17% de acurácia na classificação entre DP, DH, ELA e CO com 10-*fold cross-validation*. Nos dois trabalhos citados, o indivíduo de origem referente a cada janela não é levado em consideração na fase de separação dos folds.

A análise de sinais de voz para o diagnóstico de DP surge inspirada pelo fato de a maioria dos diagnosticados com a doença adquirir volume vocal reduzido, voz monótona e soprosa/rouca, e articulação imprecisa de palavras – um conjunto de sintomas vocais comumente nomeado de disartria hipocinética [Atarachi and Uchida, 1959; Ho *et al*., 1998]. Um dos trabalhos pioneiros é o de Little *et al*. [2009], que usou disfonia para diferenciar indivíduos saudáveis de pessoas com DP. Com 31 participantes (23 com DP) e 195 amostras vocais, os autores extraíram características do sinal de voz e aplicaram SVM com kernel gaussiano, atingindo 91,4% de acurácia. Ouhmida *et al*. [2021] utilizaram CNNs e ANNs, alcançando 93,10% de acurácia na mesma base usada por Little *et al*. [2009] e 88,89% em outro conjunto público. Em ambos os estudos, as múltiplas amostras de um mesmo indivíduo foram tratadas como amostras independentes na divisão de treino e teste.

A relevância das similaridades em amostras de um mesmo indivíduo ao analisar sinais de voz foi destacada por Naranjo *et al*. [2016]. O estudo avaliou dados de voz replicados, indicando que, embora não idênticas, as amostras de um mesmo participante tendem a ser mais semelhantes entre si do que em relação às de outros indivíduos. Essa característica sugere a necessidade de um tratamento diferenciado para as características vocais do mesmo sujeito ao conduzir experimentos. Quando essas amostras são distribuídas entre treino e teste sem o devido controle, o modelo pode simplesmente aprender padrões individuais do sujeito, e não características gerais da patologia. Isso configura vazamento de dados, pois a avaliação superestima a capacidade de generalização do modelo para novos pacientes.

Os estudos que buscam realizar a classificação de doenças neurodegenerativas, sejam por marcha ou por voz, mostram a tendência de tratar amostras isoladamente, sem considerar sua origem. No aumento de dados, a distribuição inadequada de múltiplas amostras do mesmo indivíduo entre treino e teste pode induzir viés, pois essas amostras compartilham características individuais que não refletem a variabilidade da doença. Isso pode levar os modelos a aprenderem peculiaridades dos participantes, e não padrões gerais da patologia, resultando em vazamento de dados (*data leakage*) e inflando artificialmente a acurácia, comprometendo a capacidade de generalização e confiabilidade do modelo.

3 Materiais e Métodos

A Figura 1 apresenta a metodologia adotada neste trabalho. Duas bases de dados são exploradas para a classificação de doenças neurodegenerativas, uma contendo dados de marcha, e outra contendo dados acústicos. A proposta de avaliação dos vieses na construção dos modelos são apresentados na sequência.

3.1 Dados de Marcha

Para os experimentos de marcha, foi utilizada a base de dados *Gait in Neurodegenerative Diseases Database* (GaitNDD) [Hausdorff, 2000], que inclui até cinco minutos de dados de caminhada de 15 pacientes com Doença de Parkinson (10 homens e 5 mulheres), 20 indivíduos com Doença de Huntington (6 homens e 14 mulheres), 13 pessoas com Esclerose Lateral Amiotrófica (10 homens e 3 mulheres) e 16 controles saudáveis (2 homens e 14 mulheres). A base de dados está disponível publicamente no site Physionet[1].

[1] `https://physionet.org/content/gaitndd/1.0.0/`

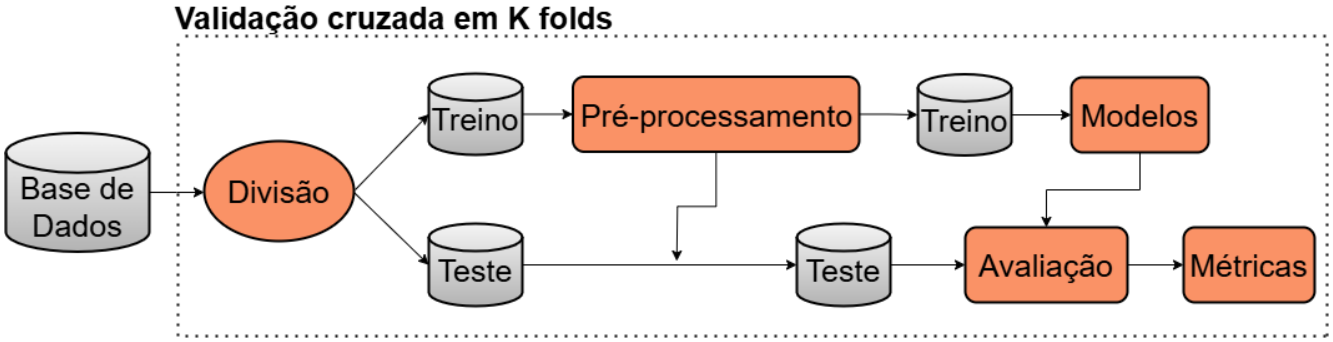

Figura 1. Fluxograma da metodologia adotada nesse trabalho.

Os dados de marcha foram coletados com sensores de força, colocados nos pés esquerdo e direito. A fase de apoio da marcha, que corresponde ao momento em que o pé está em contato com o solo e sustentando o peso do corpo, já é fornecida como uma série temporal pré-processada na base de dados. Da série temporal do intervalo de apoio do pé direito, apenas o primeiro minuto foi utilizado, seguindo a abordagem de estudos anteriores [Felix *et al*., 2022] que visam contribuir com uma maior facilidade e conforto para o paciente. Os 20 primeiros segundos de coleta, entretanto, já foram descartados pelos criadores da base de dados devido a efeitos de inicialização (citação). Portanto, apenas 40s úteis de dados de marcha são utilizados para os experimentos descritos aqui.

Esses 40 segundos restantes de cada série são divididos em quatro janelas de 10 segundos não sobrepostas, conforme ilustrado na Figura 2. De cada janela, são extraídas quatro características: média, desvio padrão, entropia e potência média do sinal. Ao final do pré-processamento, cada um dos indivíduos passa a ser representado por 4 janelas de dados, que, por sua vez, são representadas por vetores de 4 características. Esses vetores servem como entrada para os classificadores na fase futura. No total, 252 janelas de dados são analisadas, fornecendo a base para os experimentos de classificação.

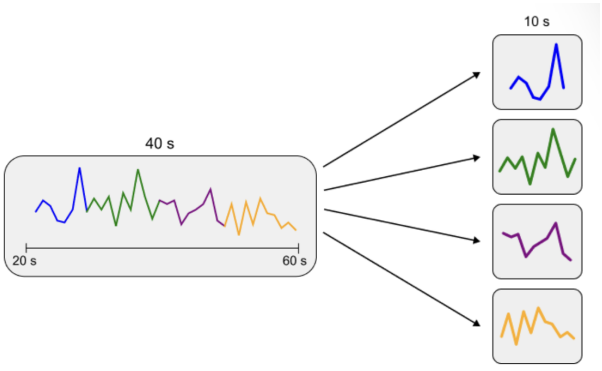

Figura 2. Representação gráfica do janelamento de dados praticado, resultando em 4 janelas disjuntas de 10 segundos.

3.2 Dados de Voz

No domínio da voz, foi utilizado o conjunto de dados *Parkinson Dataset With Replicated Acoustic Features* [Prez, 2016] para classificar a Doença de Parkinson em comparação com controles saudáveis com base em dados vocais. Este conjunto contém 80 participantes, dos quais 40 têm diagnóstico de Doença de Parkinson (22 homens e 18 mulheres) e 40 são indivíduos saudáveis (27 homens e 13 mulheres). Cada participante forneceu três amostras de voz, totalizando 210 gravações. A base está disponível publicamente no site da *UC Irvine*[2].

Cada amostra acústica inclui 27 características diferentes, diretamente acessíveis na base de dados, como variação de pitch, perturbações de amplitude, HNR (do inglês *'harmonic-to-noise ratio'*), entre outras características do sinal. Visto que o objetivo dessa investigação consiste em avaliar os vieses da classificação de dados de voz quando amostras replicadas estão disponíveis, nenhuma etapa de seleção específica de características foi realizada. Portanto, todas as características oferecidas pela base são usadas como entrada para os modelos de aprendizado de máquina.

3.3 Métodos de Classificação e Cenários de Avaliação

Para ambos os conjuntos de dados (marcha e voz), foram aplicados algoritmos clássicos de aprendizado de máquina para avaliar o desempenho desses algoritmos em diferentes cenários de avaliação e aplicar experimentos para comparação. Os modelos empregados foram: *Support Vector Machine* (SVM) com kernel linear, *k-Nearest Neighbors* (KNN) com k=5, *Naïve Bayes* (NB), *Linear Discriminant Analysis* (LDA) e *Decision Tree* (DT).

Para os dados de marcha, os modelos foram avaliados nas seguintes tarefas de classificação: PD vs. Controle (CO), HD vs. CO, ALS vs. CO e Doença Neurodegenerativa (NDD) vs. Controle (CO), onde NDD representa uma classe combinada incluindo todos os três grupos de doenças neurodegenerativas. Para os dados de voz, o foco foi na classificação PD vs. CO.

Considerando que a distribuição inadequada de amostras do mesmo indivíduo pode induzir viés e comprometer a capacidade de generalização dos modelos, foram avaliados dois cenários distintos para analisar como essa divisão dos dados influencia os resultados, conforme ilustrado nas Figuras 3 e 4.

Cenário 1: Validação cruzada realizada com separação baseada no total de amostras disponíveis em cada base, permitindo que informações de um mesmo indivíduo estejam simultaneamente nos conjuntos de treino e teste.

- Para os dados de marcha, foi utilizada validação cruzada Leave-One-Out (LOOCV) sobre as janelas de 10 segundos. Este método foi escolhido por otimizar o uso dos dados e fornecer uma estimativa precisa do desempenho

[2] `https://archive.ics.uci.edu/dataset/489/parkinson+dataset+with+replicated+acoustic+features`

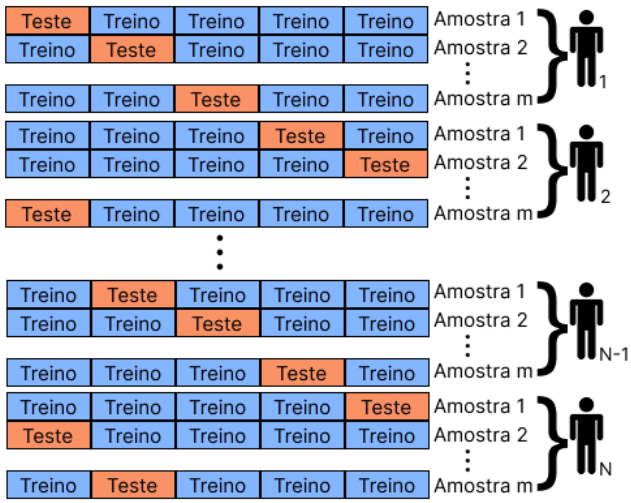

Figura 3. Cenário 1, representando a validação por amostras.

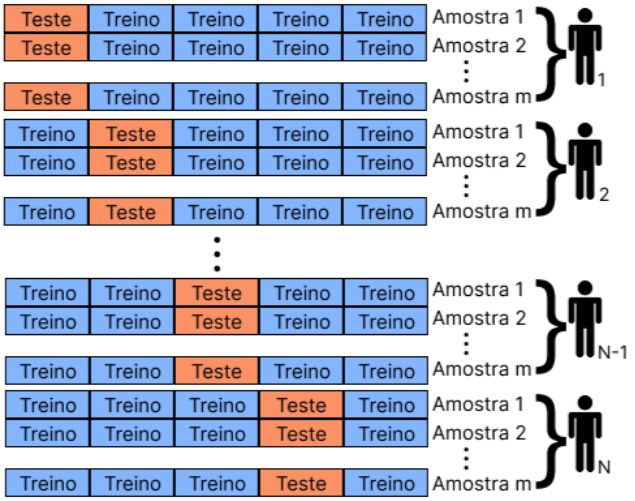

Figura 4. Cenário 2, representando a validação por pessoas.

do modelo, sendo particularmente adequado devido ao tamanho reduzido de cada uma das quatro classes (PD, 15 amostras; HD, 20 amostras, ELA. 13 amostras; CO, 16 amostras).

- Para os dados de voz, foi empregada validação cruzada de 5 folds, método preferido neste caso devido ao maior número de pacientes por classe (PD, 40 amostras; CO, 40 amostras), o que elimina a necessidade de lidar com o alto custo computacional da LOOCV ao mesmo tempo que resultados confiáveis já são oferecidos.

Cenário 2: Validação cruzada realizada considerando-se o total de indivíduos dos quais os dados foram gerados, garantindo que todas as amostras de um mesmo indivíduo estejam exclusivamente em um dos conjuntos (treino ou teste).

- Na marcha, isso significa que todas as quatro janelas de um participante são alocadas exclusivamente ao conjunto de treino ou ao conjunto de teste.
- Na voz, todas as três amostras de um indivíduo são mantidas no mesmo grupo.

Esta abordagem rigorosa tem como objetivo prevenir o vazamento de dados entre os conjuntos de treino e teste, uma vez que a presença de múltiplas amostras de um mesmo indivíduo pode levar à superestimação do desempenho do modelo. Assim, a avaliação torna-se mais representativa do desempenho dos modelos em indivíduos não previamente observados durante o treinamento.

Todo o trabalho foi desenvolvido em Python 3.10.12, com a biblioteca *Scikit Learn* para tarefas de classificação, fazendo uso de suas configurações padrões para os métodos.

4 Resultados

No domínio de voz, os resultados para DP vs. CO em ambos os cenários de avaliação são apresentados na Tabela 1. No Cenário 1, onde os dados de voz repetidos foram tratados como amostras independentes, a acurácia variou entre 71,67% (DT) e 83,33% (NB), com o modelo KNN atingindo 82,50%. No Cenário 2, em que a influência do indivíduo de origem foi mantida durante as fases de treinamento e teste, o intervalo de acurácia foi menor, variando entre 66,67% (DT e LDA) e 82,92% (NB). O KNN, que foi um dos modelos de melhor desempenho no primeiro cenário, apresentou uma queda notável de 10,42 pontos percentuais, reduzindo sua acurácia para 72,08%. Essa diferença nos resultados sugere que considerar dados do mesmo participante como independentes pode inflar artificialmente os resultados, potencialmente comprometendo a validade e a confiabilidade dos modelos. Ademais, observou-se que o Naive Bayes apresentou a menor queda de desempenho no Cenário 2.

Tabela 1. Resultados obtidos para DP *vs*. CO (dados de voz).

Cenário 1 (*5 fold* por amostras)			
Algoritmo	Acc.(%)	Sens.(%)	Espec.(%)
SVM	74.58	75.82	73.18
KNN	82.50	81.69	83.39
NB	83.33	81.65	84.96
LDA	76.25	75.71	76.66
DT	71.67	69.91	73.35
XGB	76.67	71.68	81.55
MLP	77.50	75.92	79.09

Cenário 2 (*5 fold* por pessoa)			
Algoritmo	Acc.(%)	Sens.(%)	Espec.(%)
SVM	70.00	69.60	71.48
KNN	72.08	70.62	74.05
NB	82.92	81.42	84.44
LDA	66.67	65.91	68.24
DT	66.67	66.22	67.33
XGB	72.92	69.39	76.83
MLP	69.17	62.50	75.83

As Tabelas 2 e 3 exibem os resultados das quatro tarefas de classificação e dos dois cenários de avaliação no experimento relacionado à marcha. No Cenário 1, em que as amostras de dados foram consideradas como independentes para o treinamento do modelo, o XGB superou todos os modelos, alcançando até 90,52% de precisão (ELA vs. CO), enquanto o SVM apresentou o pior desempenho, chegando a classificar incorretamente todas as amostras de controle em

Tabela 2. Resultados do Cenário 1 com dados de marcha (LOOCV por amostra).

	DP vs CO			DH vs CO			ELA vs CO			DND vs CO		
	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)
SVM	66,94	53,33	79,60	43,57	43,42	43,75	77,59	57,69	93,75	74,60	100,00	0,00
KNN	74,19	65,00	82,81	67,86	59,21	78,12	84,48	76,92	90,62	76,19	81,91	59,38
NB	72,58	53,33	90,62	77,86	67,11	90,62	77,59	59,62	92,19	66,67	58,51	90,62
LDA	79,03	73,33	84,38	79,29	67,11	93,75	81,03	65,38	93,75	74,60	93,62	18,75
DT	81,45	80,00	82,81	80,71	82,89	78,12	88,79	86,54	90,62	84,52	90,43	67,19
XGB	79,03	78,33	79,69	83,57	81,58	85,94	90,52	96,15	85,94	87,30	93,62	68,75
MLP	81,45	76,67	85,94	82,14	75,00	90,62	80,17	63,46	93,75	80,16	92,02	45,31

Tabela 3. Resultados do Cenário 2 com dados de marcha (LOOCV por pessoa).

	DP vs CO			DH vs CO			ELA vs CO			DND vs CO		
	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)
SVM	62,90	45,00	79,69	22,14	40,79	0,00	77,59	57,69	93,75	74,60	100,00	0,00
KNN	71,77	61,67	81,25	61,43	51,32	73,44	81,90	71,15	90,62	73,02	79,79	53,12
NB	72,58	53,33	90,62	77,86	67,11	90,62	78,45	59,62	93,75	66,67	58,51	90,62
LDA	79,03	75,00	82,81	75,00	63,16	89,06	80,17	63,46	93,75	72,62	91,49	17,19
DT	70,16	61,67	78,12	73,57	76,32	70,31	85,34	84,62	85,94	81,75	87,77	64,06
XGB	72,58	71,67	73,44	78,57	80,26	76,56	86,21	86,54	85,94	84,13	91,49	62,50
MLP	80,65	75,00	85,94	77,14	67,11	89,06	78,45	59,62	93,75	76,98	92,55	31,25

certos casos (DND vs. CO).

Entretanto, no Cenário 2, em que a avaliação considera a separação de apenas um sujeito por vez para o conjunto de teste, a acurácia, sensibilidade e especificidade foram, de modo geral, menores. Isso sugere que tratar amostras de marcha do mesmo indivíduo como unidades independentes leva a métricas de desempenho inflacionadas, distorcendo os resultados. Notavelmente, neste cenário em que evita-se o vazamento de dados, o MLP obteve o melhor desempenho para PD vs. CO (80,65% de acurácia), enquanto o XGB se destacou para HD vs. CO (78,57% de acurácia). Vale notar que, no caso dos dados de marcha, o desempenho do Naive Bayes no Cenário 2 se manteve relativamente estável, e até levemente superior em ELA vs. CO, destoando da tendência geral. Esses resultados enfatizam a relevância de uma separação rigorosa entre os conjuntos de treino e teste para assegurar uma classificação confiável no diagnóstico de NDDs.

5 Discussão

Os resultados obtidos sugerem que a escolha da estratégia de validação exerce um impacto significativo na avaliação do desempenho dos modelos de aprendizado de máquina voltados para o diagnóstico de doenças neurodegenerativas, independentemente do tipo de dado analisado, seja ele proveniente da marcha ou da voz. Em especial, a discrepância observada entre os dois cenários de validação testados neste estudo evidencia que abordagens comumente adotadas na literatura científica podem, mesmo que de forma não intencional, introduzir vazamento de dados. Esse fenômeno resulta na superestimação das taxas de acurácia, oferecendo uma impressão incorretamente otimista sobre a eficácia dos modelos desenvolvidos.

Dessa forma, embora diversos trabalhos prévios relatem desempenhos aparentemente elevados, como discutido na Seção 2, as variações nos protocolos de validação empregados tornam inviável uma comparação direta entre os resultados, uma vez que esses métodos não são equivalentes em termos de rigor metodológico. Os achados apresentados neste estudo sugerem que, caso metodologias mais rigorosas e realistas, capazes de levar em conta a variabilidade intra-sujeito, tivessem sido aplicadas nos estudos anteriores, as acurácias relatadas provavelmente seriam consideravelmente menores.

A análise conduzida neste trabalho reforça ainda a importância de uma prática metodológica cuidadosa: tratar como independentes os dados oriundos de diferentes amostras do mesmo indivíduo durante as fases de treinamento e teste pode induzir ao vazamento de dados, permitindo que os modelos memorizem padrões específicos de cada sujeito em vez de aprenderem de fato as características gerais associadas à doença. Esse viés metodológico compromete seriamente a capacidade de generalização dos classificadores, tornando-os menos eficazes na detecção de novos casos clínicos e reduzindo sua confiabilidade em contextos reais de aplicação, como em ambientes clínicos.

Além disso, é possível que as implicações metodológicas discutidas aqui também sejam relevantes para outros contextos clínicos, especialmente em cenários que envolvem dados fisiológicos ou biomédicos coletados de forma repetida. Assim, considerar cuidadosamente esses aspectos pode favorecer a transferência dos modelos para novos ambientes e populações, ainda que isso dependa de investigações adicionais.

Diante desses pontos, os achados aqui discutidos evidenciam a necessidade urgente de adoção de mecanismos de controle mais rigorosos durante o desenvolvimento e a validação de modelos de aprendizado de máquina aplicados à área da saúde. Isso é fundamental para garantir que o desempenho relatado nesses estudos seja compatível com cenários clínicos reais, promovendo maior segurança, confiabilidade e eficácia na aplicação dessas tecnologias em contextos diagnósticos.

6 Conclusão

Este estudo investigou possíveis vieses nas abordagens de aprendizado de máquina para o diagnóstico de doenças neurodegenerativas, com foco em como as técnicas de aumento de dados e o uso de amostras repetidas influenciam a confiabilidade dos modelos. Através de experimentos com sinais de marcha e dados acústicos provenientes de dois bancos de dados públicos, analisamos como o tratamento de amostras de marcha segmentadas e múltiplas gravações de voz do mesmo indivíduo como instâncias independentes afetam os resultados de classificação e a confiabilidade dos sistemas de suporte à decisão em saúde. Cinco algoritmos clássicos (SVM, KNN, Naive Bayes, LDA e Decision Trees) foram avaliados sob dois cenários de validação para investigar se tratar amostras individualmente poderia levar os modelos a aprender características dos indivíduos em vez de padrões relacionados às doenças.

Os achados deste estudo ressaltam a necessidade de protocolos de validação mais rigorosos, que minimizem o risco de vazamento de dados e assegurem que os modelos aprendam representações verdadeiramente discriminativas das patologias. A confiabilidade dos sistemas de suporte à decisão em saúde depende diretamente da qualidade das avaliações realizadas durante o treinamento e validação dos modelos. Assim, este trabalho contribui para o aprimoramento das práticas de desenvolvimento de sistemas de inteligência artificial para aplicações médicas, incentivando a adoção de metodologias que garantam uma generalização confiável para novos pacientes. Além disso, destaca-se que análises complementares, como testes estatísticos de significância entre os modelos e cenários avaliados, podem reforçar e validar de forma mais robusta os achados observados, constituindo uma direção importante para trabalhos futuros.

O presente artigo, apresentado no Concurso de Trabalhos de Iniciação Científica (CTIC) do XXV Simpósio Brasileiro de Computação Aplicada à Saúde [Chagas *et al*., 2025], é o compilado de um ano de resultados levantados durante a realização de uma iniciação científica. A partir dos resultados aqui mostrados, foram geradas publicações no formato de trabalhos completos [Chagas *et al*., 2024b; Felix *et al*., 2025], resumo expandido [Chagas *et al*., 2024c] em conferências, e um pôster em uma conferência local [Chagas *et al*., 2024a]. Além disso, também houve contribuições em artigos relacionados ao tema [da Silva *et al*., 2024; Bucci *et al*., 2025]. O projeto continua em andamento, com novas investigações previstas para explorar outras fontes de dados e aprimorar as metodologias, buscando mitigar possíveis vieses em modelos de aprendizado de máquina aplicados ao diagnóstico de doenças neurodegenerativas.

Declarações complementares

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, e Edital Nº 30/2022 – Programa Emergencial de Solidariedade Acadêmica (PDPG).

Contribuições dos autores

ALBC – metodologia, investigação, experimentos e escrita; GFFBB – experimentos e escrita; JPF – concepção, supervisão e validação; RLS – supervisão; HADN – supervisão; FS – supervisão, administração do projeto e obtenção de financiamento. Ana Luísa de Bastos Chagas é a principal contribuidora deste manuscrito. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

O conjunto de dados analisado durante o estudo atual está disponível publicamente em physionet.org e archive.ics.uci.edu.

Referências

Atarachi, J. and Uchida, E. (1959). A clinical study of parkinsonism. *Recent Advances in Research on the Nervous System*, 3:871–882.

Berman, T. and Bayati, A. (2018). What are neurodegenerative diseases and how do they affect the brain? *Frontiers for Young Minds*, 6. DOI: 10.3389/frym.2018.00070.

Bielby, J., Kuhn, S., Colreavy-Donnelly, S., Caraffini, F., O'Connor, S., and Anastassi, Z. A. (2020). Identifying parkinson's disease through the classification of audio recording data. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–7. IEEE.

Bucci, G. d. F. F. B., Chagas, A. L. d. B., Félix, J. P., Nascimento, H. A. D. d., and Soares, F. (2025). Transformer: O poder da atenção no auxílio ao diagnóstico de múltiplas doenças neurodegenerativas. *Revista Eletrônica de Iniciação Científica em Computação*, 23(1):158–164. DOI: 10.5753/reic.2025.6048.

Chagas, A., Bucci, G., Félix, J., Salvini, R., Nascimento, H., and Soares, F. (2025). Exploring biases in machine learning models for neurodegenerative diseases diagnosis through gait and voice analysis. In *Anais Estendidos do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 1–6, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas_estendido.2025.7498.

Chagas, A., Lobo, P. S., Felix, J., do Nascimento, H., and Salvini, R. (2024a). Analyzing the impact of voice data replication on machine learning models for parkinson's disease diagnosis. In *Anais da XII Escola Regional de Informática de Goiás*, pages 263–264, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/erigo.2024.5092.

Chagas, A. L., Bucci, G., Felix, J., Fonseca, A., Nascimento, H., and Soares, F. (2024b). Avaliando a sobreamostragem de dados temporais de marcha no diagnóstico automático de doenças neurodegenerativas. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), Goiânia, GO, Brazil, June 25–28, 2024*, pages 1–12. SBC.

Chagas, A. L. B., Felix, J. P., and Nascimento, H. A. D. (2024c). Gait analysis for the diagnosis of neurodegenerative diseases. In *Anais Estendidos do XX Congresso Brasileiro de Informática em Saúde - CBIS'24*, pages 484–485, Belo Horizonte.

da Silva, M. I., Felix, J. P., de Stecca Prado, T., de Bastos Chagas, A. L., Bucci, G. d. F. F. B., da Fonseca, A. U., and Soares, F. (2024). Sobre a análise de sinais de voz para o diagnóstico da doença de parkinson. *Journal of Health Informatics*, 16(Especial).

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 Kibaroğlu, S. (2021). Neurodegenerative disease detection and severity prediction using deep learning approaches. *Biomedical Signal Processing and Control*, 70:103069.

Felix, J., da Silva, M. I., Chagas, A. L., Salvini, R., Nascimento, H., and Soares, F. (2025). Analyzing the effect of replicated voice samples in parkinson's disease classification. In *2025 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*. Aceito para publicação.

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.

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.

Hausdorff, J. (2000). Gait in neurodegenerative disease database. Disponível em: https://physionet.org/content/gaitndd/1.0.0/.

Heemels, M.-T. (2016). Neurodegenerative diseases.

Ho, A. K., Iansek, I. H., Marigliani, R. L., Bradshaw, J., and Gates, M. C. (1998). Speech impairment in a large sample of patients with parkinson's disease. *Behavioural Neurology*, 11(3):131–137.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022. DOI: 10.1109/TBME.2008.2005954.

Mayeux, R. (2003). Epidemiology of neurodegeneration. *Annual Review of Neuroscience*, 26(1):81–104.

Naranjo, L., Pérez, C. J., Campos-Roca, Y., and Martín, J. (2016). Addressing voice recording replications for parkinson's disease detection. *Expert Systems with Applications*, 46:286–292. DOI: 10.1016/j.eswa.2015.10.034.

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.

Noella, R. S. N. and Priyadarshini, J. (2023). Machine learning algorithms for the diagnosis of alzheimer and parkinson disease. *Journal of Medical Engineering & Technology*, 47(1):35–43. PMID: 36043506. DOI: 10.1080/03091902.2022.2097326.

Ouhmida, A., Fattah, J., Khaireddin, Y., and Maaroufi, M. (2021). Voice-based deep learning medical diagnosis system for parkinson's disease prediction. In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*. IEEE.

Prez, C. (2016). Parkinson Dataset with replicated acoustic features. UCI Machine Learning Repository. DOI: 10.24432/C5701F.

Varghese, J., Brenner, A., Fujarski, M., van Alen, C. M., Plagwitz, L., and Warnecke, T. (2024). Machine Learning in the Parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10(1):9.

Vyas, T., Yadav, R., Solanki, C., Darji, R., Desai, S., and Tanwar, S. (2022). Deep learning-based scheme to diagnose parkinson's disease. *Expert Systems*, 39(3):e12739. DOI: 10.1111/exsy.12739.

Yasar, A., Saritas, I., Sahman, M., and Cinar, A. (2019). Classification of parkinson disease data with artificial neural networks. In *IOP conference series: materials science and engineering*, volume 675, page 012031. IOP Publishing.