CLASSIFICAÇÃO DE DOENÇAS NEURODEGENERATIVAS A PARTIR DA MARCHA: UMA ABORDAGEM COM APRENDIZADO DE MÁQUINA E VALIDAÇÃO SEM VAZAMENTO ENTRE SUJEITOS

Reginaldo José da Silva[1], Mara Lúcia Martins Lopes[2], Edvaldo J. R. Cardoso[3], Angela Leite Moreno[3]

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho", Ilha Solteira, Brasil (reginaldo.silva@unesp.br)
[2]Universidade Estadual Paulista "Júlio de Mesquita Filho", Ilha Solteira, Brasil
[3]Universidade Federal de Alfenas, Alfenas, Brasil

Resumo: Este trabalho compara *Random Forest* e *HistGradientBoosting* na classificação de Parkinson, Huntington, ELA e controles a partir de características de marcha, usando validação *leave-one-subject-out* para evitar vazamento de dados entre janelas do mesmo paciente. O *HistGradientBoosting* obteve acurácia de 73,44% e AUC-ROC de 0,892 por sujeito. Parkinson foi a classe mais difícil de classificar, e testes de sensibilidade confirmaram a configuração adotada.

Palavras-chave: Parkinson; Huntington; Esclerose lateral amiotrófica; *Random Forest*; *Gradient Boosting*

INTRODUÇÃO

As doenças neurodegenerativas figuram entre os maiores desafios da neurologia contemporânea, não apenas pela ausência de tratamentos curativos, mas também pela dificuldade em identificá-las em estágios iniciais, quando intervenções têm maior chance de preservar a qualidade de vida do paciente (GBD 2021 *Nervous System Disorders Collaborators*, 2024). Parkinson, Huntington e esclerose lateral amiotrófica (ELA) compartilham uma característica clínica relevante: comprometem, de formas distintas, o controle motor e, por extensão, a marcha. Essa manifestação comum abre espaço para uma linha de investigação que vem ganhando força nos últimos anos, a de que padrões sutis na forma de caminhar podem funcionar como marcadores objetivos e não invasivos dessas condições, complementando o diagnóstico clínico tradicional.

A doença de Parkinson é um distúrbio neurodegenerativo progressivo associado à perda de neurônios dopaminérgicos na substância negra, manifestando-se clinicamente por tremor de repouso, rigidez, bradicinesia e instabilidade postural. Acomete predominantemente pessoas acima dos 60 anos, embora formas de início precoce, antes dos 50, representem uma parcela não desprezível dos casos. O diagnóstico permanece de natureza clínica, apoiado em critérios motores e, quando disponível, em exames de imagem complementares; não existe cura, e o tratamento concentra-se no controle sintomático, sendo a levodopa considerada o padrão-ouro terapêutico, ainda que seu uso prolongado esteja associado a flutuações motoras e discinesias (Armstrong; Okun, 2020).

A doença de Huntington, por sua vez, tem origem genética bem definida: decorre da expansão de repetições CAG no gene HTT, localizado no braço curto do cromossomo 4, e é transmitida de forma autossômica dominante. Quanto maior o número de repetições, mais cedo tende a surgir a doença, cujo quadro clássico envolve coreia, ou seja, movimentos involuntários e abruptos, além de declínio cognitivo e distúrbios psiquiátricos. O início dos sintomas costuma ocorrer entre os 30 e os 50 anos de idade. O diagnóstico combina histórico familiar, quadro clínico e confirmação por teste genético, e, assim como no Parkinson, não há tratamento capaz de deter a progressão da doença, restando apenas o manejo dos sintomas motores e comportamentais (Roos, 2010).

Já a esclerose lateral amiotrófica é uma doença do neurônio motor caracterizada pela degeneração simultânea de neurônios motores superiores e inferiores, o que provoca fraqueza muscular progressiva, espasticidade e, eventualmente, insuficiência respiratória. Costuma se manifestar entre os 50 e os 70 anos e apresenta prognóstico reservado, com sobrevida mediana de dois a quatro anos após o diagnóstico. O diagnóstico é fundamentalmente clínico e eletrofisiológico, muitas vezes obtido por exclusão de outras condições, e o riluzol permanece como uma das poucas opções farmacológicas capazes de oferecer algum ganho de sobrevida, ainda que modesto (Molligoda, 2025).

Nas três condições, o atraso diagnóstico tem custo real. Sintomas motores mais evidentes só aparecem depois de uma perda substancial de neurônios ou de progressão significativa da doença, o que reduz a janela de tempo em que terapias sintomáticas, cuidados multidisciplinares e inclusão em ensaios clínicos podem ser mais eficazes. O diagnóstico das três condições permanece, em grande parte, dependente da avaliação clínica subjetiva, marcada por dificuldades de diagnóstico precoce e diferencial já documentadas na literatura (Pratihar; Sankar, 2024). Diante desse cenário, cresce o interesse por ferramentas computacionais capazes de auxiliar, não substituir, o julgamento clínico.

O aprendizado de máquina tem sido aplicado a diferentes tipos de dado com esse propósito. Um exemplo consolidado é o uso de medidas de disfonia extraídas da voz para rastrear Parkinson: Little et al. (2009) treinaram um classificador de vetores de suporte sobre gravações de vogais sustentadas de 31 participantes, alcançando acurácia média de 91,4%, resultado que impulsionou uma linha de pesquisa inteira sobre telemonitoramento por voz. Outros trabalhos exploraram imagens de ressonância magnética, sinais de eletroencefalografia e dados de acelerômetros vestíveis com finalidade semelhante (Pratihar; Sankar, 2024). Cada tipo de dado, porém, tem suas próprias limitações de custo, disponibilidade de equipamento e necessidade de ambiente controlado para coleta.

A marcha desponta como uma alternativa atraente justamente por evitar boa parte dessas limitações. Sua análise pode ser feita com sensores relativamente simples e baratos, dispensa exames de imagem, e capta diretamente o produto final de circuitos neurais responsáveis pelo controle motor, entre eles os núcleos da base e o próprio neurônio motor. O trabalho pioneiro de Hausdorff et al. (2000) demonstrou que a variabilidade do intervalo entre passadas, medida por sensores de força sob o pé, distingue estatisticamente pacientes com Parkinson, Huntington e ELA de indivíduos saudáveis, lançando as bases do banco de dados de marcha em doenças neurodegenerativas hoje amplamente utilizado em pesquisa.

Desde então, diversos estudos aplicaram aprendizado de máquina sobre esse tipo de série temporal para automatizar a classificação dessas condições. Erdaş, Sümer e Kibaroğlu (2023) combinaram transformada wavelet com classificadores estatísticos e reportaram acurácia acima de 97% na diferenciação entre as quatro classes. Setiawan, Liu e Lin (2022) converteram o sinal de marcha em espectrogramas de coerência wavelet e aplicaram redes convolucionais pré-treinadas, chegando a acurácias próximas de 96%. Esses números são expressivos, e por isso merecem uma leitura cuidadosa. Parte considerável dessa literatura reporta validação do tipo *leave-one-out* sem deixar claro se a exclusão ocorre no nível do sujeito ou no nível da amostra individual, o que é relevante porque muitos desses estudos, incluindo o de Hausdorff et al. (2000), fragmentam o sinal de cada paciente em múltiplos segmentos antes do treinamento. Quando segmentos de um mesmo paciente aparecem tanto no treino quanto no teste, ainda que em uma validação cruzada nominalmente rigorosa, o modelo tende a aprender a assinatura individual daquele paciente em vez de um padrão da doença que generalize para casos novos, inflando artificialmente a acurácia relatada. Essa ambiguidade metodológica não invalida os achados, mas impõe cautela na comparação direta entre estudos e reforça a necessidade de protocolos de validação que separem sujeitos, não apenas amostras, entre treino e teste.

O presente trabalho parte dessa lacuna. Utilizando o banco de dados de marcha em doenças neurodegenerativas originalmente descrito por Hausdorff et al. (2000), composto por séries temporais de intervalos de passada de pacientes com Parkinson, Huntington, ELA e controles saudáveis, o objetivo consiste em desenvolver e avaliar um modelo de classificação multiclasse baseado em características estatísticas e não lineares extraídas da marcha, comparando *Random Forest* e *Gradient Boosting* como classificadores. A validação adota o esquema *leave-one-subject-out*, em que cada paciente é utilizado como teste exatamente uma vez e nenhuma amostra sua participa do treinamento correspondente, de modo a produzir uma estimativa de desempenho mais fiel à capacidade de generalização do modelo para pacientes ainda não vistos.

MODELOS DE CLASSIFICAÇÃO

A tarefa de classificação foi conduzida com dois algoritmos baseados em conjuntos de árvores de decisão, *Random Forest* (RF) e *Histogram-based Gradient Boosting* (HGB), escolhidos por lidarem bem com um número moderado de amostras e por não exigirem normalização prévia das variáveis.

O *Random Forest* constrói um conjunto de Bárvores de decisão independentes, cada uma treinada sobre uma amostra *bootstrap* dos dados originais e considerando, em cada divisão de nó, apenas um subconjunto aleatório das variáveis disponíveis (Breiman, 2001). Essa dupla aleatorização (nas amostras e nas variáveis) reduz a correlação entre as árvores e, consequentemente, a variância do conjunto, mantendo baixo o viés do modelo. A predição final para uma nova observação x é obtida pelo voto majoritário das B árvores:

$$\hat{y}(x) = \text{moda}\{T_1(x),\ T_2(x),\ \cdots, T_B(x)\} \qquad (1)$$

em que $T_b(x)$ é a classe prevista pela b-ésima árvore. Cada árvore é construída de forma independente e, em

cada nó, seleciona a divisão que proporciona a maior redução da impureza. Quando utilizado o índice de Gini, a impureza de um nó contendo K classes, com proporções p_k, é definida por:

$$G = 1 - \sum_{k=1}^{K} p_k \quad (2)$$

O *Histogram-based Gradient Boosting* segue uma lógica distinta: em vez de treinar árvores independentes, constrói o modelo de forma aditiva e sequencial, em que cada nova árvore corrige os erros deixados pelo conjunto já treinado (Friedman, 2001). O modelo após M etapas é dado por:

$$\mathrm{F}_M(x) = \mathrm{F}_0(x) + \sum_{m=1}^{\mathrm{M}} \alpha\, \mathrm{h}_m(x) \quad (3)$$

em que $F_0(x)$ representa a predição inicial do modelo, $h_m(x)$é a árvore ajustada na m-ésima etapa e α é a taxa de aprendizado (*learning rate*), que controla a contribuição de cada nova árvore para o modelo final. Cada h_m é treinada para aproximar o gradiente negativo da função de perda em relação às previsões correntes:

$$r_i = -\frac{\partial L(\mathrm{y}_i, F(\mathrm{x}_i)}{\partial F(x_i)}\Big|_{F=F_{m-1}} \quad (4)$$

em que L é a função de perda, y_i é o valor observado e $F_{m-1}(x_i)$ corresponde à previsão do modelo antes da inclusão da m-ésima árvore. No caso da perda logística utilizada em problemas de classificação binária, esse gradiente corresponde à diferença entre a classe observada e a probabilidade prevista pelo modelo, de modo que cada nova árvore é ajustada para reduzir esse erro. A variante *histogram-based*, inspirada no LightGBM (Ke et al., 2017), discretiza previamente as variáveis contínuas em um número fixo de compartimentos (*bins*) antes da busca pelos pontos de divisão. Essa estratégia reduz o custo computacional do treinamento e o consumo de memória, preservando a lógica geral do algoritmo de *Gradient Boosting*.

MATERIAL E MÉTODOS

Conjunto de dados

O estudo utiliza o *Gait in Neurodegenerative Disease Database*, descrito originalmente por Hausdorff et al. (2000) e disponibilizado publicamente no PhysioNet. O banco reúne séries temporais de marcha de 64 indivíduos, divididos em quatro grupos: 16 controles saudáveis (Grupo Controle - GC), 15 pacientes com doença de Parkinson (DP), 20 com doença de Huntington (DH) e 13 com esclerose lateral amiotrófica (ELA). Cada sujeito caminhou em ritmo confortável por alguns minutos enquanto sensores de força posicionados sob cada pé registravam o instante de contato e de saída do solo, dado do qual derivam sete canais por passo: os intervalos de passada, balanço e apoio de cada perna, além do intervalo de duplo apoio. O número de passos varia entre sujeitos, já que a duração da caminhada não foi padronizada.

Pré-processamento e extração de características

Como a série de cada paciente tem comprimento próprio e o objetivo é capturar variações locais da marcha, e não apenas um resumo estático de toda a caminhada, cada série foi segmentada em janelas deslizantes de 50 passos, com sobreposição de 50% entre janelas consecutivas. Esse tamanho de janela não foi arbitrário: comparou-se o desempenho de janelas de diferentes comprimentos (30, 50, 60, entre outros) por meio de uma validação cruzada agrupada por sujeito, e o valor de 50 passos foi o que produziu o melhor resultado quando confirmado na validação final descrita adiante.

Cada janela deu origem a um vetor de 68 características: para cada um dos sete canais, calcularam-se média, desvio-padrão, coeficiente de variação, energia, valores mínimo e máximo, o expoente de escala obtido por análise de flutuação destendenciada, ou DFA (Peng et al., 1994), a entropia amostral (Richman; Moorman, 2000), e uma medida de desvio em relação à normalidade, definida como a soma dos afastamentos além de uma faixa de referência (média ± dois desvios-padrão) calculada a partir do grupo controle. Somaram-se ainda medidas de assimetria entre os lados esquerdo e direito e a razão entre os intervalos de balanço e apoio. Quando a entropia amostral não convergia, o que ocorre ocasionalmente em janelas curtas ou muito ruidosas, o valor foi imputado pela mediana da respectiva coluna.

Desenho experimental

Cada janela foi tratada como uma unidade de treinamento independente, rotulada com a classe do paciente a que pertence, mas sempre respeitando o vínculo com o sujeito de origem nas etapas de validação, de modo a impedir que janelas do mesmo paciente aparecessem simultaneamente no treino e no teste. Essa é a distinção central em relação a parte da literatura discutida na introdução. A decisão final, contudo, é reportada por paciente: as previsões de todas as janelas de um mesmo sujeito são agregadas por voto majoritário para gerar a classe prevista, e a média das probabilidades preditas por janela é usada para calcular métricas baseadas em limiar de decisão, como a área sob a curva *Receiver Operating Characteristic* (AUC-ROC). Por isso, os resultados são reportados em dois níveis, o de janela, que reflete o desempenho bruto do classificador sobre cada segmento, e o de sujeito, que reflete o desempenho após a agregação e corresponde à decisão clinicamente relevante.

Validação cruzada e otimização de hiperparâmetros

A validação seguiu um esquema em duas etapas, adotado para equilibrar rigor metodológico e custo computacional. Na primeira, os hiperparâmetros de cada modelo foram ajustados por otimização bayesiana, técnica em que um modelo probabilístico da função-objetivo orienta a escolha de novas combinações de hiperparâmetros a partir dos resultados já observados, em vez de testá-las de forma exaustiva ou puramente aleatória. Adotou-se o estimador de Parzen estruturado em árvore, ou TPE (Bergstra et al., 2011), com poda por mediana para interromper combinações pouco promissoras antes de concluídas.

A função-objetivo foi o F1-macro médio obtido em uma validação cruzada estratificada de cinco dobras, agrupada por sujeito, de modo que, mesmo nessa etapa exploratória, nenhuma dobra continha janelas de um paciente presente em outra dobra. Foram avaliadas 60 combinações de hiperparâmetros por modelo, dentro dos intervalos listados na Tabela 1. Optou-se por não aninhar essa busca dentro de cada dobra da validação final, o que exigiria repetir a otimização para cada um dos 64 sujeitos excluídos, tornando o custo computacional proibitivo sem ganho de generalização proporcional em um conjunto de dados desse tamanho.

Na segunda etapa, com os hiperparâmetros já fixados, a avaliação final foi conduzida por validação cruzada do tipo *leave-one-subject-out* (LOSO): a cada rodada, um sujeito é inteiramente excluído do treinamento, o modelo é treinado com os demais 63, e a previsão é gerada apenas para as janelas do sujeito excluído. O processo se repete até que todos os 64 sujeitos tenham servido como teste exatamente uma vez, produzindo previsões fora da amostra para o conjunto de dados inteiro sem que nenhuma informação de um paciente jamais influencie sua própria previsão.

A Tabela 1 apresenta o espaço de busca de hiperparâmetros explorado pela otimização bayesiana para cada modelo; os valores finais selecionados para o classificador de melhor desempenho são reportados na seção de resultados.

Tabela 1. Espaço de busca de hiperparâmetros explorado pela otimização bayesiana.

Modelo	Hiperparâmetro	Espaço de busca
RF	número de árvores	100 a 600 (passo 50)
	profundidade máxima	3 a 20
	amostras mínimas para divisão	2 a 20
	amostras mínimas por folha	1 a 10
	número de variáveis por divisão	raiz quadrada, log2 ou total
HGB	número de iterações	100 a 500 (passo 50)
	profundidade máxima	2 a 12
	taxa de aprendizado	0,01 a 0,3 (escala log)
	regularização L2	1e-4 a 10 (escala log)
	número máximo de folhas	7 a 63

Métricas de avaliação

O desempenho dos modelos foi avaliado com um conjunto amplo de métricas, calculadas tanto de forma agregada quanto por classe individual, e tanto no nível de janela quanto no nível de sujeito. Para uma classe k tratada como positiva frente às demais, sejam VP, FP, VN e FN o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. A acurácia mede a proporção geral de acertos:

$$Acurácia = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}} \quad (5)$$

A precisão indica, entre as previsões de uma classe, quantas estavam corretas, enquanto a sensibilidade, também chamada de *recall*, indica quantos casos reais daquela classe foram corretamente identificados:

$$Precisão = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \quad (6)$$

$$\mathrm{Sensibilidade} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \quad (7)$$

A especificidade mede a capacidade de reconhecer corretamente os casos que não pertencem à classe:

$$\mathrm{Especificidade} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}} \quad (8)$$

O F1-*score* resume precisão e sensibilidade em um único valor, por meio de sua média harmônica:

$$\mathrm{F1} - score = \frac{\mathrm{VP}}{\mathrm{VP} + \frac{1}{2}(FP + FN)} \quad (9)$$

Como o problema envolve quatro classes, cada métrica foi calculada primeiro por classe, tratando-a como positiva frente às demais, e depois combinada de duas formas: a média macro, que calcula a métrica separadamente para cada classe k e tira a média simples entre as K classes,

$$Métrica_{macro} = \frac{1}{K}\sum Métrica_k \quad (10)$$

e a média micro, que soma os verdadeiros e falsos positivos e negativos de todas as classes antes de calcular a métrica uma única vez, dando peso proporcional ao número de amostras de cada classe. A acurácia balanceada, por sua vez, corresponde à média macro da sensibilidade, sendo menos sensível ao desbalanceamento entre classes do que a acurácia simples.

Por fim, a área sob a curva característica de operação do receptor (*Area Under the Receiver Operating Characteristic Curve* - AUC-ROC) e a área sob a curva precisão-recall (*Area Under the Precision-Recall Curve* - AUC-PR) avaliam a qualidade das probabilidades previstas, e não apenas da classe de maior probabilidade, ao considerar todos os limiares de decisão possíveis. A AUC-ROC é obtida pela soma das áreas entre pontos consecutivos da curva:

$$\text{AUC} - \text{ROC} = \sum_{i=1}^{n-1} \text{TPR}\,(t_i)\,\Delta\text{FPR}(t_i), \qquad (11)$$

em que t_i representa o i-ésimo limiar de decisão, $\text{TPR}(t_i)$ é a sensibilidade e $\text{FPR}(t_i) = \frac{FP(t_i)}{FP(t_i)+VN(t_i)}$ é a taxa de falsos positivos. De forma análoga, a AUC-PR é calculada por:

$$\text{AUC-PR} = \sum_{i=1}^{n-1} \text{Precisão}\,(t_i)\,\Delta\text{Sensibilidade}(t_i). \qquad (12)$$

Ambas foram calculadas nas variantes macro e micro, bem como por classe individual, considerando cada classe isoladamente frente às demais.

RESULTADOS E DISCUSSÃO

Hiperparâmetros selecionados

A otimização bayesiana convergiu para configurações distintas para cada modelo, refletindo diferenças na forma como cada algoritmo lida com o número relativamente pequeno de sujeitos. O HGB selecionou uma taxa de aprendizado baixa (0,0153) combinada a uma regularização L2 alta (9,71), um padrão que sugere um ajuste conservador, coerente com o risco de sobreajuste em um conjunto de dados deste porte. A Tabela 2 apresenta a combinação final de cada modelo.

Tabela 2. Hiperparâmetros selecionados pela otimização bayesiana para cada modelo.

Modelo	Hiperparâmetro	Valor
RF	número de árvores	100
	profundidade máxima	18
	amostras mínimas para divisão	5
	amostras mínimas por folha	9
	número de variáveis por divisão	total
HGB	número de iterações	250
	profundidade máxima	10
	taxa de aprendizado	0,0153
	regularização L2	9,715
	número máximo de folhas	53

Nível de janela

Como descrito na metodologia, cada janela de 50 passos foi tratada como uma unidade de treinamento e teste independente durante a validação cruzada *leave-one-subject-out*, ainda que respeitando o vínculo com o sujeito de origem. As métricas neste nível refletem o desempenho bruto do classificador sobre cada segmento de marcha, antes de qualquer agregação por paciente.

Tabela 3. Métricas globais no nível de janela.

Métrica	RF	HGB
Acurácia	0,6296	**0,6589**
Acurácia balanceada	0,6316	**0,6581**
Precisão macro	0,6138	**0,6485**
Precisão micro	0,6296	**0,6589**
Sensibilidade macro	0,6316	**0,6581**
Sensibilidade micro	0,6296	**0,6589**
Especificidade macro	0,8769	**0,8853**
F1 macro	0,6168	**0,6521**
F1 micro	0,6296	**0,6589**
AUC-ROC macro	0,8368	**0,8511**
AUC-ROC micro	0,8518	**0,8645**

Métrica	RF	HGB
AUC-PR macro	0,6418	**0,6675**
AUC-PR micro	0,6629	**0,6955**

O HGB superou o RF em todas as métricas neste nível, embora por margens modestas: a acurácia balanceada passou de 0,6316 para 0,6581, e a área sob a curva ROC macro, de 0,8368 para 0,8511. A especificidade macro permaneceu a métrica mais alta para ambos os modelos, o que é esperado em um problema de quatro classes, já que confundir uma classe com qualquer uma das outras três conta como acerto de especificidade.

Tabela 4. Métricas por classe no nível de janela.

Modelo	Classe	Sensib.	Especif.	F1	ROC
RF	GC	0,821	0,874	0,762	0,898
	ELA	0,687	0,881	0,597	0,910
	DH	0,598	0,897	0,658	0,830
	DP	0,421	0,855	0,451	0,709
HGB	GC	0,729	0,903	0,734	0,926
	ELA	0,699	0,909	0,644	0,913
	DH	0,689	0,865	0,698	0,844
	DP	0,516	0,863	0,533	0,722

A quebra por classe revela um padrão que se repete em ambos os modelos: a DP é consistentemente a classe mais difícil de identificar, com sensibilidade de apenas 0,421 no RF e 0,516 no HGB, e a menor área sob a curva ROC entre as quatro classes em ambos os casos (0,709 e 0,722, respectivamente). Em contraste, a classe controle mantém a maior sensibilidade nos dois modelos.

Tabela 5. Matrizes de confusão no nível de janela.

RF				
Verdadeiro \ Predito	GC	ELA	DH	DP
GC	115	6	6	13
ELA	14	57	0	12
DH	17	18	98	31
DP	16	27	30	53
HGB				
GC	102	7	13	18
ELA	11	58	5	9
DH	16	9	113	26
DP	9	23	29	65

As matrizes de confusão no nível de janela apresentadas na Tabela 5 confirmam esse padrão de forma mais granular. No RF, dos erros da classe Parkinson, 30 janelas foram confundidas com Huntington e 27 com ELA, uma dispersão praticamente simétrica entre as demais classes. O HGB exibe uma tendência semelhante, embora com maior concentração dos erros de Parkinson em direção à Huntington (29 janelas) do que à ELA (23 janelas). Essa dispersão, em vez de uma confusão concentrada em uma única classe, sugere que a dificuldade em separar o Parkinson não decorre de uma sobreposição pontual com outra condição específica, mas de uma limitação mais geral das características utilizadas para capturar sua assinatura de marcha.

Nível de sujeito

A decisão final por paciente, obtida por voto majoritário das janelas de cada sujeito, corresponde à unidade clinicamente relevante e é a que melhor estima a capacidade do modelo de generalizar para um paciente ainda não visto. Em ambos os modelos, a agregação melhorou o desempenho em relação ao nível de janela, um indício de que o voto majoritário absorve parte do ruído presente em segmentos individuais.

Tabela 6. Métricas globais no nível de sujeito.

Métrica	RF	HGB
Acurácia	0,6875	**0,7344**
Acurácia balanceada	0,6944	**0,7329**

Métrica	RF	HGB
Precisão macro	0,6884	**0,7310**
Precisão micro	0,6875	**0,7344**
Sensibilidade macro	0,6944	**0,7329**
Sensibilidade micro	0,6875	**0,7344**
Especificidade macro	0,8974	**0,9117**
F1 macro	0,6813	**0,7290**
F1 micro	0,6875	**0,7344**
AUC-ROC macro	0,8665	**0,8922**
AUC-ROC micro	0,8781	**0,8999**
AUC-PR macro	0,7509	**0,7879**
AUC-PR micro	0,7389	**0,7953**

A vantagem do HGB sobre o RF se mantém e se acentua neste nível: a acurácia sobe de 0,6875 para 0,7344, e a área sob a curva ROC macro, de 0,8665 para 0,8922. A área sob a curva precisão-*recall* macro, mais sensível ao desbalanceamento entre classes do que a ROC, também favorece o HGB (0,7879 contra 0,7509).

Tabela 7. Métricas por classe no nível de sujeito.

Modelo	Classe	Sensib.	Especif.	F1	ROC
RF	GC	0,875	0,875	0,778	0,909
	ELA	0,769	0,882	0,690	0,919
	DH	0,600	0,955	0,706	0,894
	DP	0,533	0,878	0,552	0,744
HGB	GC	0,813	0,875	0,743	0,941
	ELA	0,769	0,922	0,741	0,928
	DH	0,750	0,932	0,789	0,922
	DP	0,600	0,918	0,643	0,778

No nível de sujeito, o Parkinson permanece a classe mais difícil para os dois modelos, com sensibilidade de 0,533 no RF e 0,600 no HGB, e novamente a menor área sob a curva ROC (0,744 e 0,778, respectivamente), destacada em relação às demais classes, que superam 0,89 em ambos os modelos.

Tabela 8. Matrizes de confusão no nível de sujeito.

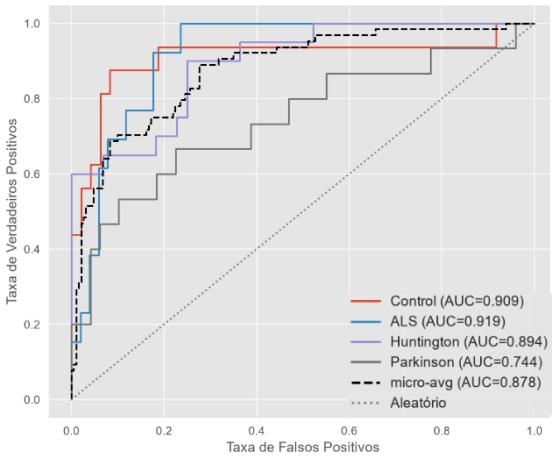
RF

Verdadeiro \ Predito	GC	ELA	DH	DP
GC	14	1	0	1
ELA	2	10	0	1
DH	2	2	12	4
DP	2	3	2	8

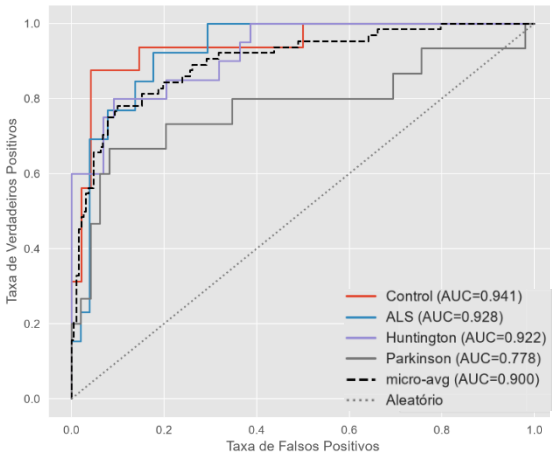
HGB

GC	13	1	1	1
ELA	2	10	0	1
DH	3	0	15	2
DP	1	3	2	9

As matrizes de confusão por sujeito tornam esse padrão concreto. O HGB classifica corretamente 13 dos 16 controles, 10 dos 13 pacientes com ELA, 15 dos 20 com Huntington e 9 dos 15 com Parkinson. O RF apresenta um perfil parecido, com 14, 10, 12 e 8 acertos, respectivamente, para as mesmas classes. Em ambos os modelos, os erros de Parkinson se distribuem entre as três classes restantes sem uma concentração clara, enquanto os erros de Huntington se dirigem majoritariamente para a classe controle, principalmente no *HistGradientBoosting* (3 dos 5 erros).



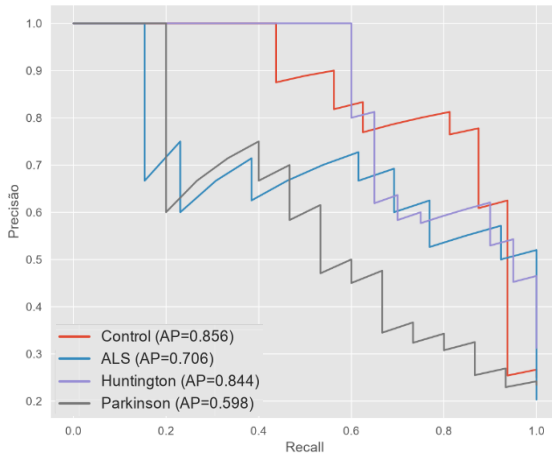
a) *Random Forest*



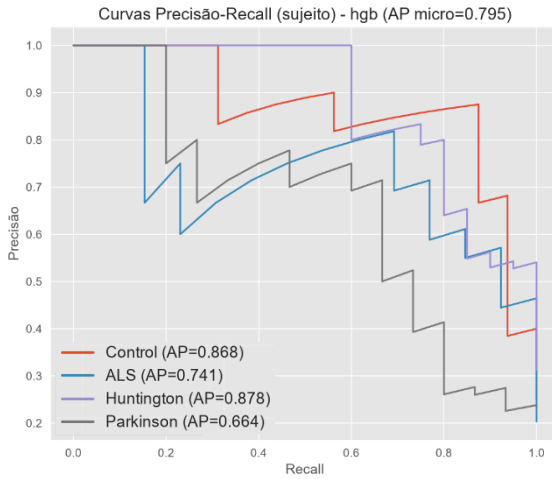
b) *HistGradientBoosting*

Figura 1. Curvas ROC por classe no nível de sujeito.

As curvas ROC por sujeito permitem avaliar a separabilidade de cada classe independentemente do limiar de decisão utilizado. A classe Parkinson exibe a curva mais próxima da diagonal de referência em ambos os modelos, confirmando, com uma métrica que não depende da escolha de limiar, que a dificuldade nessa classe é estrutural e não um artefato da regra de decisão padrão. As demais classes ultrapassam 0,89 de área sob a curva nos dois modelos, patamar que indica boa separabilidade.



a) *Random Forest*



b) *HistGradientBoosting*

Figura 2. Curvas precisão-*recall* por classe no nível de sujeito.

As curvas de precisão-*recall* reforçam essa leitura e a tornam mais sensível ao desbalanceamento entre classes: a precisão média de Parkinson é a mais baixa em ambos os modelos (0,598 no RF e 0,664 no HGB), enquanto Huntington e Controle mantêm precisão média acima de 0,84 nos dois casos.

Importância de características

A Tabela 9 lista as características mais relevantes para cada modelo, considerando a importância nativa por impureza no RF e a importância por permutação no HGB, já que este último não expõe uma medida de importância nativa. Apesar da diferença de método, os dois modelos convergem para a mesma característica como a mais relevante: a assimetria do intervalo de passada entre os lados esquerdo e direito, seguida pelo valor mínimo do intervalo de apoio da perna direita.

Tabela 9. Cinco características mais importantes por modelo (importância nativa por impureza no RF; importância por permutação no HGB).

Ranking	RF	HGB
1º	Assimetria do intervalo de passada esq./dir. (0,249)	Assimetria do intervalo de passada esq./dir. (0,198)
2º	Valor mínimo do intervalo de apoio, perna dir. (0,232)	Valor mínimo do intervalo de apoio, perna dir. (0,075)
3º	Valor mínimo do intervalo de passada, perna dir. (0,051)	Média do intervalo de balanço, perna esq. (0,057)
4º	Média do intervalo de balanço, perna dir. (0,039)	Valor mínimo do intervalo de duplo apoio (0,033)
5º	Média da razão balanço/apoio (0,039)	Valor mínimo do intervalo de passada, perna esq. (0,020)

Essa convergência entre dois algoritmos com mecanismos internos distintos é um indício de que a assimetria de marcha carrega, de fato, sinal discriminativo relevante para essas condições, e não é apenas um artefato de ajuste de um modelo específico.

Gravidade clínica e padrão de erro

Uma análise exploratória cruzou os acertos e erros do HGB, por classe, com a gravidade clínica informada no banco de dados. Para a doença de Huntington, cuja gravidade é medida pela capacidade funcional total, sujeitos classificados corretamente apresentaram mediana de 4,5 pontos, contra 12,0 pontos entre os classificados incorretamente, diferença estatisticamente significativa pelo teste de Mann-Whitney ($p = 0{,}010$, $n = 14$ acertos e 5 erros). Como valores mais baixos de capacidade funcional total indicam doença mais avançada, o resultado sugere que o modelo identifica com mais facilidade os casos de Huntington em estágio avançado, presumivelmente por sua assinatura de marcha mais alterada, e tende a errar nos casos mais leves. Para a doença de Parkinson, cuja gravidade é medida pela escala de Hoehn-Yahr, não houve diferença significativa entre acertos e erros ($p = 0{,}952$), indicando que, ao contrário do observado em Huntington, a dificuldade do modelo com essa classe não parece decorrer do estágio da doença. Por se tratar de uma análise exploratória, conduzida sobre um único modelo e com poucos casos por grupo, esse achado deve ser interpretado como indício, não como confirmação.

Análise de sensibilidade

Três aspectos do desenho experimental foram testados de forma adicional para verificar se a configuração adotada era, de fato, a mais adequada, e não apenas uma escolha arbitrária inicial.

O primeiro teste comparou tamanhos de janela de 30, 50 e 60 passos, cada um avaliado com o mesmo protocolo completo de validação (busca de hiperparâmetros seguida de *leave-one-subject-out*). A janela de 50 passos, adotada neste trabalho, produziu o melhor resultado nos três casos, com acurácia por sujeito de 0,734 contra 0,688 para 30 passos e 0,672 para 60 passos.

Tabela 10. Comparação de tamanhos de janela (HGB).

Tamanho de janela	Acurácia	F1-macro	AUC-ROC macro
30 passos	0,6875	0,6811	0,8759
50 passos (adotado)	**0,7344**	**0,7290**	**0,8922**
60 passos	0,6719	0,6619	0,8735

O segundo teste avaliou se um subconjunto menor de características, selecionado pelo ranking de importância, superaria o uso do conjunto completo. Subconjuntos de 15 e 25 características produziram área sob a curva ROC ligeiramente superior à do conjunto completo (0,907 e 0,907, contra 0,892), mas acurácia e F1-macro inferiores, o que levou à manutenção das 68 características na configuração final.

Tabela 11. Comparação de números de características selecionadas (HGB).

Quantidade de *Features*	Acurácia	F1-macro	AUC-ROC macro
15	0,7031	0,6932	0,9067
25	0,7031	0,6939	0,9066
68 (todas, adotado)	**0,7344**	**0,7290**	**0,8922**

O terceiro teste avaliou se ajustar o ponto de corte de decisão por classe, usando o índice de Youden,

recalculado a cada rodada do *leave-one-subject-out* exclusivamente a partir dos sujeitos de treino, para evitar que o sujeito excluído influenciasse seu próprio limiar de decisão, melhoraria a classificação final em relação à regra padrão de maior probabilidade. O ajuste produziu uma pequena melhora no nível de janela (acurácia de 0,6706 contra 0,6589), mas piora no nível de sujeito (0,7188 contra 0,7344), o que levou à manutenção da regra de decisão padrão como resultado principal.

Tabela 12. Efeito do ajuste de ponto de corte por índice de Youden (HGB).

Nível	Decisão	Acurácia	F1-macro
Janela	Padrão (argmax)	0,6589	0,6521
Janela	Ajustado (Youden)	0,6706	0,6607
Sujeito	Padrão (argmax)	0,7344	0,7290
Sujeito	Ajustado (Youden)	0,7188	0,7176

Em conjunto, os resultados indicam que características de marcha capturadas por sensores de força, combinadas a modelos baseados em árvore, distinguem as quatro condições com desempenho consistente e superior ao acaso, mas com uma diferença clara de dificuldade entre classes que uma métrica única de acurácia não deixaria evidente. Do ponto de vista prático, um sistema apoiado neste modelo dificilmente substituiria o diagnóstico clínico, sobretudo para Parkinson, cuja sensibilidade permanece a mais baixa entre as quatro classes, mas poderia funcionar como uma ferramenta de triagem complementar, sinalizando casos que mereçam avaliação especializada mais detalhada, especialmente para Huntington e ELA, em que a precisão se manteve elevada em ambos os modelos.

As principais limitações do estudo, o tamanho reduzido da amostra e a origem de um único protocolo de coleta, restringem a generalização dos resultados para outras populações e ambientes, e sugerem que a validação em bases de dados independentes e de maior porte seja o passo mais importante antes de qualquer aplicação prática.

CONCLUSÃO

Este trabalho propôs e avaliou um modelo de classificação multiclasse para distinguir doença de Parkinson, doença de Huntington, esclerose lateral amiotrófica (ELA) e controles saudáveis a partir de características extraídas da marcha, comparando *Random Forest* e *HistGradientBoosting* sob um protocolo de validação *leave-one-subject-out*. O *HistGradientBoosting* apresentou o melhor desempenho, com acurácia de 0,734 e área sob a curva ROC macro de 0,892 no nível de sujeito, superando o *Random Forest* em praticamente todas as métricas avaliadas.

A principal contribuição do trabalho está na forma como a validação foi conduzida. Ao segmentar as séries de marcha em janelas para gerar mais amostras de treinamento, mas garantir que janelas de um mesmo paciente nunca fossem divididas entre treino e teste, o protocolo adotado evita o vazamento de dados identificado como uma limitação comum na literatura que utiliza esse mesmo banco de dados. Essa escolha provavelmente explica por que os resultados obtidos aqui, embora sólidos, ficam abaixo dos valores acima de 95% relatados em parte dos trabalhos anteriores: a diferença não indica um modelo inferior, mas sim uma estimativa mais fiel da capacidade de generalização para pacientes ainda não vistos.

Os erros do modelo também se mostraram clinicamente interpretáveis, e não aleatórios. A doença de Parkinson se destacou como a classe mais difícil de classificar em ambos os modelos, um resultado consistente com a natureza mais sutil e heterogênea das alterações de marcha nessa condição, em comparação à coreia característica da doença de Huntington ou à fraqueza progressiva da esclerose lateral amiotrófica. Já para Huntington, os erros se concentraram nos pacientes em estágio mais leve, cuja marcha ainda não apresenta as alterações mais marcantes da doença avançada. Os três testes adicionais de sensibilidade, sobre tamanho de janela, número de características e ponto de corte de decisão, confirmaram que a configuração inicialmente adotada já correspondia à melhor opção validada, reforçando que buscas de configuração mais baratas computacionalmente nem sempre apontam na direção correta, e que a confirmação com o protocolo de validação completo é indispensável antes de qualquer conclusão sobre desempenho.

De forma mais ampla, este trabalho reforça que, em conjuntos de dados clínicos de porte reduzido como o utilizado aqui, protocolos de validação rigorosos no nível do sujeito devem preceder qualquer comparação de desempenho entre modelos ou entre estudos, mesmo quando isso resulta em números menos expressivos do que os reportados por trabalhos com validação menos criteriosa.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

Armstrong, M. J.; Okun, M. S. Diagnosis and Treatment of Parkinson Disease: A Review. JAMA, v. 323, n. 6, p. 548-560, 2020.

Bergstra, J. S. et al. Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems, v. 24, p. 2546-2554, 2011.

Breiman, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

Erdaş, Ç. B.; Sümer, E.; Kibaroğlu, S. Neurodegenerative diseases detection and grading using gait dynamics. Multimedia Tools and Applications, p. 22925-22942, 2023.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, v. 29, n. 5, p. 1189-1232, 2001.

GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet Neurology, v. 23, n. 4, p. 344-381, 2024.

Hausdorff, J. M. et al. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. Journal of Applied Physiology, v. 88, n. 6, p. 2045-2053, 2000.

Ke, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, v. 30, p. 3146-3154, 2017.

Little, M. A. et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, v. 56, n. 4, p. 1015-1022, 2009.

Molligoda A. S. P. Rethinking ALS: Current understanding and emerging therapeutic strategies. AIMS Neuroscience, v. 12, n. 3, p. 391-405, 2025.

Peng, C.-K. et al. Mosaic organization of DNA nucleotides. Physical Review E, v. 49, n. 2, p. 1685-1689, 1994.

Pratihar, R.; Sankar, R. Advancements in Parkinson's Disease Diagnosis: A Comprehensive Survey on Biomarker Integration and Machine Learning. Computers, v. 13, n. 11, p. 293, 2024.

Richman, J. S.; Moorman, J. R. Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology, v. 278, n. 6, p. H2039-H2049, 2000.

Roos, R. A. Huntington's disease: a clinical review. Orphanet Journal of Rare Diseases, v. 5, p. 40, 2010.

Setiawan, F.; Liu, A.-B.; Lin, C.-W. Development of neuro-degenerative diseases' gait classification algorithm using convolutional neural network and wavelet coherence spectrogram of gait synchronization. IEEE Access, v. 10, p. 38137-38153, 2022.