Redução do Viés Intra-Sujeito na Classificação da Doença de Parkinson pela Voz com Otimização por Colônia de Formigas

Pedro Lemes Sixel Lobo[1], **Juliana Paula Felix**[1,2], **Rogerio Salvini**[1], **Clarimar Coelho**[2], **Fabrizzio Soares**[1]

[1]Instituto de Informática, Universidade Federal de Goiás
[2]Escola Politécnica e de Artes, Pontifícia Universidade Católica de Goiás

pedro_lemes@discente.ufg.br,
{julianafelix, rogeriosalvini, fabrizzio}@ufg.br,
clarimarc@gmail.com

Abstract. *Parkinson's Disease (PD) is a neurodegenerative condition that affects voice production. Previous studies have employed machine learning techniques for PD diagnosis based on voice data. However, the vast majority of existing approaches overlook intra-subject variation in their methodologies. This study proposes a methodology that accounts for such variation while also employing the bio-inspired Ant Colony Optimization algorithm for feature selection, thereby reducing overfitting. An accuracy of 84.80% was achieved in distinguishing individuals with PD from healthy controls, providing a fairer and reliable approach and highlighting the potential of bio-inspired algorithms in PD diagnosis.*

Resumo. *A Doença de Parkinson (DP) é uma condição neurodegenerativa que afeta a voz. Estudos anteriores utilizaram aprendizado de máquina para diagnóstico de Parkinson a partir de dados obtidos pela voz. Entretanto, a grande maioria das abordagens existentes desconsideram a variação intra-sujeito nas abordagens propostas. Este estudo propõe um método que aborda essa variação, além de empregar o algoritmo bioinspirado Otimização por Colônia de Formigas para seleção de características, reduzindo o sobreajuste. Foi obtida uma acurácia de 84,80% na distinção entre indivíduos com DP e controles saudáveis em uma abordagem mais justa e confiável, destacando o potencial de algoritmos bioinspirados no diagnóstico da DP.*

1. Introdução

A Doença de Parkinson (DP) é uma enfermidade neurodegenerativa progressiva que afeta predominantemente o sistema motor [Hayes 2019]. A condição atinge entre 2% e 3% da população com mais de 65 anos [Poewe et al. 2017], e se caracterizada pela perda progressiva e anormal de neurônios na região cerebral, resultando em uma queda da produção de dopamina, neurotransmissor essencial para o controle dos movimentos [Braak and Braak 2000]. Isso leva a sintomas motores clássicos, como tremores, rigidez muscular, bradicinesia (lentidão dos movimentos), instabilidade postural e distúrbios de fala e voz [Ho et al. 1998]. A disfonia parkinsoniana envolve hipofonia, monotonia, rouquidão e dificuldades na articulação [Hayes 2019]. Essas mudanças podem prejudicar a comunicação do paciente e levar a dificuldades sociais e emocionais, impactando sua interação com o meio e sua qualidade de vida.

Um dos desafios no tratamento da DP está na dificuldade de diagnóstico precoce. Muitos pacientes enfrentam um diagnóstico tardio, devido a sintomas iniciais sutis e à ausência de biomarcadores confiáveis. Nesse contexto, diversas abordagens vêm sendo propostas na literatura para auxiliar no diagnóstico de Parkinson de forma automatizada, utilizando técnicas de aprendizado de máquina. Dentre essas alternativas, destacam-se métodos que analisam dados relacionados à marcha [Daliri 2012], atividades motoras [Varghese et al. 2024] e voz [Rana et al. 2022, Tsanas et al. 2012, Ouhmida et al. 2021, Govindu and Palwe 2023].

Dentre as abordagens existentes para classificação de sinais de voz para auxiliar no diagnóstico de DP, o uso de bases de dados com amostras de voz replicadas é recorrente. A tática de se captar e utilizar mais de uma amostra de voz de um mesmo indivíduo pode ser justificada tanto pela dificuldade em se obter um conjunto representativo da população com a Doença de Parkinson, quanto pela variabilidade intra-indivíduo, o que justificaria o uso desses dados como amostras independentes. Entretanto, embora essa tática permita aumentar o conjunto de dados, estudos recentes tem apontado que a forma como os dados de um mesmo indivíduo são utilizados na criação de modelos de machine learning, em especial na separação dos conjuntos de treinamento e teste, pode causar viéses e resultados superestimados [Naranjo et al. 2016, da Silva et al. 2024, Felix et al. 2025]. Isso porque, quando amostras distintas de um mesmo indivíduo figuram tanto no conjunto de treino como no de teste, é possível que os modelos aprendam a reconhecer características pessoais ao invés dos padrões das doenças investigadas.

Este trabalho tem como objetivo auxiliar no diagnóstico da DP a partir da voz, utilizando técnicas de aprendizado de máquina. Para isso, uma base de dados pública de voz, obtida de 252 participantes através de três repetições da vogal sustentada /a/, é utilizada. De cada amostra, mais de 700 características foram extraídas e estão disponíveis para uso. A fim de evitar o mal da dimensionalidade, exploramos o desempenho do algoritmo de colônia de formigas para seleção de características, o qual é avaliado sob a ótica de vários classificadores. Diferentemente de outras abordagens existentes na literatura, o trabalho aqui reportado garante uma classificação não enviesada, assegurando que amostras de voz originárias de um mesmo participante sejam devidamente tratadas para evitar vazamentos de dados nas fases de treinamento e teste. Caso isso não seja feito, o sistema pode aprender mais sobre o paciente do que sobre a doença. Por isso, reduzir esse viés é essencial para que os modelos funcionem bem em cenários reais e realmente sejam generalizáveis para novos pacientes.

O artigo está estruturado da seguinte maneira. A Seção 2 revisa os principais trabalhos relacionados ao diagnóstico da doença de Parkinson a partir da voz. A Seção 3 apresenta a fundamentação teórica do algoritmo utilizado para a seleção de características. A Seção 4 descreve os materiais e métodos aplicados no estudo. Por fim, as Seções 5 e 6 apresentam, respectivamente, os resultados obtidos e as conclusões do trabalho.

2. Trabalhos Relacionados

A seguir, são discutidos trabalhos relevantes que abordam o diagnóstico de Parkinson com diferentes bases de dados de voz.

Um artigo com a base de dados *"Oxford Parkinson's Disease Detection Dataset"* investigou o viés causado pelo uso de amostras replicadas de voz como amostras

independentes no processo de treinamento e avaliação de modelos de *machine learning* para diagnóstico da Doença de Parkinson (DP) [da Silva et al. 2024]. A análise foi realizada com sinais de 24 indivíduos com DP e 8 saudáveis, aplicando algoritmos como *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Random Forest* (RF) com validação cruzada *Leave-one-out* (LOOCV). Quando todas as amostras de voz foram tratadas como independentes, ignorando a similaridade intra-sujeito, os modelos apresentaram acurácias elevadas, com destaque para o *Random Forest*, que atingiu 92,31%. No entanto, como cada participante possuía três amostras de voz, um novo experimento foi conduzido, agrupando as amostras de um mesmo indivíduo para que fossem usadas exclusivamente no treino ou no teste dos modelos. Nesse cenário mais realista, o desempenho dos modelos reduziu significativamente, sendo o SVM com *kernel* linear o mais eficaz, com 83,33% de acurácia, 94% de sensibilidade e 52% de especificidade. O estudo enfatiza a importância de um processamento adequado dos dados para evitar a superestimação do desempenho dos modelos e garantir sua real eficácia na detecção da DP.

Diversos outros estudos utilizam essa mesma base de dados com métodos que desconsideram a variação intra-sujeito nas abordagens de aprendizado de máquina propostas. Em especial, destaca-se o trabalho dos criadores da base de dados, em que características acústicas, como *jitter*, *shimmer*, *Recurrence Period Density Entropy* (RPDE), *Detrended Fluctuation Analysis* (DFA) e *Pitch Period Entropy* (PPE) [Little et al. 2009]. Foi utilizado o classificador *Support Vector Machine* (SVM) com *kernel* gaussiano, com validação por reamostragem *bootstrap* com 50 replicações. Os autores obtiveram uma acurácia média de 91,4%, semelhante ao obtido por da Silva et al. (2024) quando as amostras de voz foram todas consideradas amostras independentes.

Naranjo et al. (2016) apresentaram um estudo com uma nova base de dados contendo com gravações de voz de 80 pessoas, sendo 40 diagnosticadas com Doença de Parkinson (DP) e 40 saudáveis, das quais 44 atributos acústicos foram extraídos. Entre eles, incluíram-se medidas de variação de frequência e amplitude, relação harmônico-ruído e coeficientes cepstrais de frequência Mel (MFCCs). A classificação foi conduzida por um modelo Bayesiano baseado em variáveis latentes, levando em conta a correlação entre gravações do mesmo participante. A avaliação, utilizando validação cruzada estratificada e considerando a dependência entre medições replicadas para evitar viés na classificação, atingiu uma precisão de 75,2%. Além disso, identificou-se uma diferença de desempenho entre os gêneros, com maior acurácia para mulheres em comparação aos homens.

Alguns artigos que utilizam a mesma base de dados avaliada neste trabalho podem ser encontrados na literatura. Em um primeiro estudo, foi aplicada a Transformada Wavelet de Fator Q Ajustável (TQWT) na extração de características dos sinais vocais de pacientes com Doença de Parkinson (DP), comparando-se com a Transformada Wavelet Discreta (DWT) tradicional [Sakar et al. 2019]. Esse estudo extraiu características avaliadas por meio de validação cruzada *Leave-One-Group-Out*, incluindo abordagens de *ensemble* para aprimorar a acurácia dos modelos. Os resultados indicaram que a TQWT (*Tunable Q-factor Wavelet Transform*) apresentou desempenho comparável ou superior a técnicas tradicionais de extração de características na classificação da DP. A melhor acurácia obtida foi de 86%, utilizando o classificador *Support Vector Machine* (SVM) com *kernel* RBF e um subconjunto de 50 características selecionadas pela técnica MRMR (*Minimum Redundancy Maximum Relevance*).

Em um segundo estudo com a mesma base de dados, foi explorada a seleção de características por meio do algoritmo *Wrappers*, obtendo-se um conjunto de 8 a 20 características para a classificação [Solana-Lavalle et al. 2020]. Quatro classificadores foram aplicados na detecção da DP com base nas características vocais: *k-Nearest Neighbors* (kNN), *Perceptron* Multicamadas (MLP), *Support Vector Machine* (SVM) e *Random Forest* (RF). O melhor desempenho foi alcançado com o classificador SVM, que obteve uma acurácia de 94,7%, sensibilidade de 98,4%, especificidade de 92,68% e precisão de 97,22%. A seleção das características foi feita utilizando validação cruzada com 10 *folds*.

No trabalho de [Parlar 2021], o autor avalia a eficácia de diferentes métodos de seleção de características, como *Information Gain* e *ReliefF*, combinados com o algoritmo bioinspirado *Wolf Search Algorithm* (WSA). Os resultados indicaram que o método *ReliefF* obteve os melhores resultados de classificação, superando os outros métodos quando combinado com a Rede Neural Artificial (ANN), que atingiu um F-score de 92,5% utilizando um conjunto de 100 características. Contudo, embora a ANN tenha proporcionado o melhor desempenho, ela exigiu de 20 a 700 vezes mais tempo de computação em comparação aos outros métodos. Os classificadores empregados foram *Logistic Regression* (LR), *Support Vector Machines* (SVM), *Random Forest* (RF) e ANN, com os resultados avaliados pela média micro do F-score, utilizando validação cruzada com 5 *folds*.

No entanto, estes dois últimos estudos não mencionaram o uso de abordagens para garantir que os dados de um mesmo sujeito não sejam utilizados tanto no treinamento quanto no teste, para evitar o vazamento de dados. O presente trabalho realiza essa separação e avalia diferentes modelos em um cenário que leva em consideração o relacionamento da amostra com cada sujeito participante. Além disso, o algoritmo *Ant Colony Optimization* foi utilizado para seleção de características, o que não foi explorado em trabalhos anteriores.

3. Fundamentação Teórica

Algoritmos bioinspirados são métodos computacionais que bulem soluções para problemas complexos com base em comportamentos e processos observados na natureza. Esses algoritmos se baseiam em princípios de otimização presentes em sistemas biológicos, tais como a forma como organismos interagem entre si, ou como resolvem desafios do seu ambiente de forma eficiente [Darwish 2018]. Este é o caso do algoritmo da colônia de formigas, utilizado neste trabalho.

Introduzido por Dorigo et al. na década de 1990, a Otimização por Colônia de Formigas (*Ant Colony Optimization* – ACO) foi criada como um método computacional para a resolução de problemas de otimização combinatória. O ACO constitui uma meta-heurística inspirada no comportamento coletivo das formigas durante a busca por alimento. O desenvolvimento desse algoritmo fundamenta-se na observação de que certas espécies de formigas depositam feromônios no solo para marcar trajetórias mais curtas e eficientes entre o ninho e uma fonte de alimento [Dorigo et al. 2006]. Esse mecanismo permite que a colônia encontre rotas otimizadas.

Seu funcionamento é baseado na construção iterativa de soluções por agentes artificiais, as formigas. Esses agentes percorrem um espaço de busca definido pelo problema e tomam decisões influenciadas pela quantidade de feromônio já depositado em determi-

nadas soluções e uma heurística baseada nas características do problema. Para evitar que o algoritmo fique preso em soluções subótimas, os feromônios passam por um processo de atualização, no qual trilhas mais antigas perdem influência gradualmente, permitindo uma exploração mais ampla e aumentando as chances de encontrar soluções otimizadas. A atualização dos feromônios também reforça os caminhos associados a boas soluções e enfraquece aqueles vinculados a soluções de menor qualidade [Dorigo et al. 2006].

O desempenho do algoritmo é diretamente influenciado pelos seus parâmetros, os quais impactam a eficiência e a qualidade das soluções encontradas [Dorigo et al. 1996]. O peso do feromônio (α) determina a influência da memória acumulada na tomada de decisão das formigas, favorecendo soluções previamente bem-sucedidas. A taxa de evaporação do feromônio (ρ) evita a estagnação prematura em soluções subótimas, promovendo a exploração de novas regiões do espaço de busca. Além disso, a quantidade de feromônio depositada (Q) regula o reforço atribuído às soluções de maior qualidade, garantindo que caminhos mais promissores tenham maior probabilidade de serem selecionados em iterações subsequentes.

A função objetivo exerce um papel central no ACO, pois permite avaliar a qualidade das soluções geradas em cada iteração do algoritmo. Neste estudo, a função foi utilizada para avaliar diferentes subconjuntos de características, guiando o processo de seleção para otimizar o desempenho do modelo. Dessa forma, subconjuntos que resultam em melhores desempenhos recebem um reforço maior na matriz de feromônios, aumentando a probabilidade de serem selecionados em iterações subsequentes. Esse processo possibilita a identificação de subconjuntos de características que maximizam a eficiência do modelo preditivo, reduzindo redundâncias e melhorando a capacidade geral de generalização.

4. Materiais e Métodos

A Figura 1 apresenta o método aplicado neste trabalho. As seções seguintes detalham essas etapas.

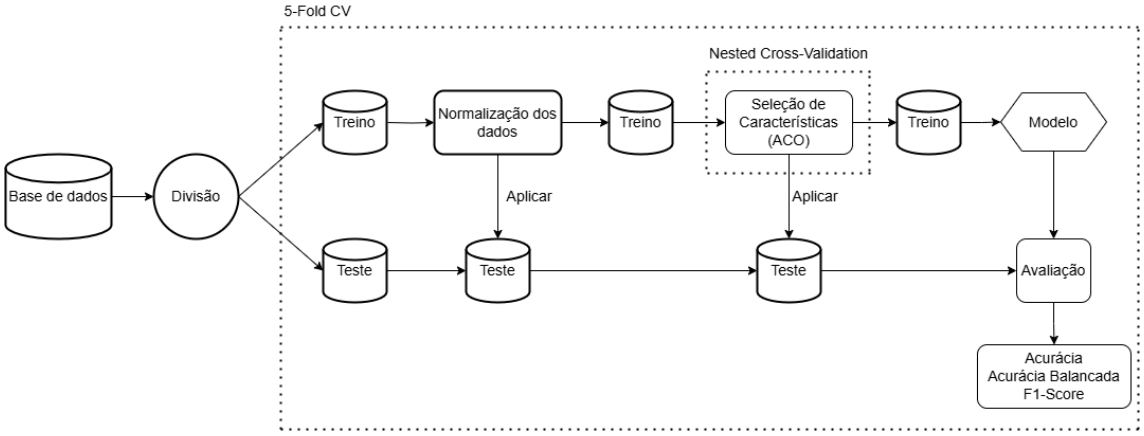

Figura 1. Método utilizado neste trabalho, com validação cruzada com 5 folds (5-fold CV) e *nested cross- validation* na etapa de seleção de características.

4.1. Base de dados

Neste estudo, a base *Parkinson's Disease Classification dataset*, foi utilizada. A base foi construída a partir de dados coletados pelo Departamento de Neurologia da Faculdade

de Medicina Cerrahpaşa, Universidade de Istambul, e está disponível publicamente no University of California Irvine (UCI) Machine Learning Repository[1].

Os dados foram coletados a partir da gravação da vogal sustenida /a/, com 3 coletas por indivíduo. Como haviam 252 participantes, o conjunto de dados resultante contém 756 instâncias. No total, 754 características acústicas foram extraídas de cada instância e disponibilizadas na base de dados. Essas características foram divididas em categorias distintas, as quais estão detalhadas na Tabela 1.

Tabela 1. Categorias das características disponíveis na base de dados.

Grupo	Descrição
Baseline	Medidas estatísticas de *jitter*, *shimmer*, frequência fundamental, relação sinal-ruído, entropia do *pitch*, DFA (*Detrended Fluctuation Analysis*) e RPDE (*Recurrence Period Density Entropy*).
Frequência Temporal	Intensidade do sinal, frequências dos formantes e largura de banda.
MFCCs	MFCC (*Mel-Frequency Cepstral Coefficients*), coeficientes cepstrais na escala Mel para modelagem da resposta auditiva.
Wavelet Transform	Características extraídas por DWT (*Discrete Wavelet Transform*) e TQWT (*Tunable Q-factor Wavelet Transform*), analisando o contorno da frequência fundamental.
Vocal Fold	Medidas da vibração das pregas vocais: GQ (*Glottis Quotient*), GNE (*Glottal to Noise Excitation*), VFER (*Vocal Fold Excitation Ratio*) e EMD (*Empirical Mode Decomposition*).

Os participantes foram diagnosticados por especialistas e divididos em dois grupos: 188 pacientes com Doença de Parkinson, representando 564 instâncias do conjunto de dados, e 64 indivíduos saudáveis, totalizando 192 instâncias.

4.2. Seleção de Características

Devido à alta dimensionalidade do conjunto de dados, foi aplicada a seleção de características a fim de otimizar a eficácia dos modelos empregados para o diagnóstico da Doença de Parkinson em um cenário sem sobreajuste causado pelo *mal da dimensionalidade*. O método adotado foi o algoritmo de Colônia de Formigas, operando de maneira iterativa, explorando diferentes subconjuntos de características com o objetivo de encontrar a combinação ideal que maximize o desempenho do modelo de classificação.

A cada iteração, as formigas selecionam subconjuntos de características, sendo guiadas pela quantidade de feromônio presente em cada uma delas. Características com maior concentração de feromônio têm maior probabilidade de serem escolhidas, direcionando o algoritmo para soluções mais promissoras, ao mesmo tempo em que mantém a exploração de diferentes combinações. Formalmente, a probabilidade de seleção de uma característica é calculada dividindo a quantidade de feromônio dessa característica, elevada ao parâmetro alpha, pelo somatório dos níveis de feromônio de todas as características, também elevados a esse parâmetro.

[1]Repositório Parkinson's Disease Classification dataset: `https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification`

Após a seleção, a qualidade de cada subconjunto de características é avaliada com base na acurácia de um modelo de *Support Vector Machine* (SVM) com *kernel* linear. O modelo é treinado com os subconjuntos escolhidos pelas formigas e validado nos dados de teste. Quando uma solução apresenta uma acurácia superior à anterior, o feromônio das características selecionadas é reforçado, aumentando a probabilidade de elas serem escolhidas nas iterações seguintes. Em contrapartida, as características associadas a soluções de menor qualidade têm seu feromônio reduzido, permitindo que o algoritmo se concentre nas combinações mais eficazes. Para evitar que o algoritmo se limite a soluções subótimas, aplica-se uma evaporação do feromônio a cada iteração. Esse processo diminui progressivamente a influência das soluções mais antigas, incentivando a exploração de novas combinações de características. Esse mecanismo assegura uma busca diversificada, prevenindo a estagnação e aumentando as chances de encontrar o subconjunto ideal de características. Os parâmetros utilizados no algoritmo são:

- Número de formigas (n_{ants}): 20
- Número de iterações ($n_{\text{iterations}}$): 100
- Parâmetro α (influência do feromônio): 1
- Taxa de evaporação do feromônio (ρ): 0.5
- Fator de reforço do feromônio (Q): 1

Para avaliar o impacto da quantidade de características selecionadas no desempenho do modelo, foram realizados testes com diferentes números de características finais – 100, 200, 300, 400 e 500, como feito por Parlar (2021). Os parâmetros $\alpha = 1$ e $\rho = 0.5$ foram definidos com base nas configurações padrão utilizadas por Dorigo et al. (1996), que realizaram testes experimentais com esses valores para avaliar seu impacto no desempenho do algoritmo de colônia de formigas. O valor de 20 formigas foi assim definido para assegurar uma exploração eficiente do espaço de soluções, enquanto as 100 iterações foram estabelecidas para garantir a convergência adequada do algoritmo.

4.3. Classificação

O processo de classificação entre indivíduos saudáveis e com Doença de Parkinson foi conduzido com diferentes classificadores e foram avaliados em dois cenários distintos.

- **Cenário 1** – Utilizando todas as características disponíveis na base de dados.
- **Cenário 2** – Utilizando características selecionadas pelo algoritmo da colônia de formigas.

Os dois cenários foram analisados por meio de uma validação cruzada estratificada com 5 *folds*, conforme apresentado anteriormente na Figura 1. O *k-fold* estratificado foi utilizado para garantir que a distribuição das classes seja preservada em cada divisão, o que é crucial em conjuntos de dados com classes desbalanceadas. Essa abordagem proporciona uma avaliação mais confiável do desempenho do modelo em todas as classes, reduzindo os vieses que poderiam surgir devido à representação desigual das classes nos conjuntos de validação [Prusty et al. 2022]. Nos dados destinados ao treino em cada fold, foi realizada a normalização com *StandardScaler*, a qual foi seguida pelo processo de seleção de características (apenas no cenário 2). Posteriormente, as etapas de transformação dos dados e seleção de características foram aplicadas ao conjunto separado para teste. No cenário 2, para mitigar o problema do mal da dimensionalidade existente no cenário 1, empregou-se o método de *Nested Cross-Validation*, apresentado

na Figura 2, para a seleção de características e a avaliação dos modelos. Esse método envolve dois níveis de validação cruzada, um externo, responsável pela avaliação do desempenho dos modelos, e um interno, dedicado à seleção de características.

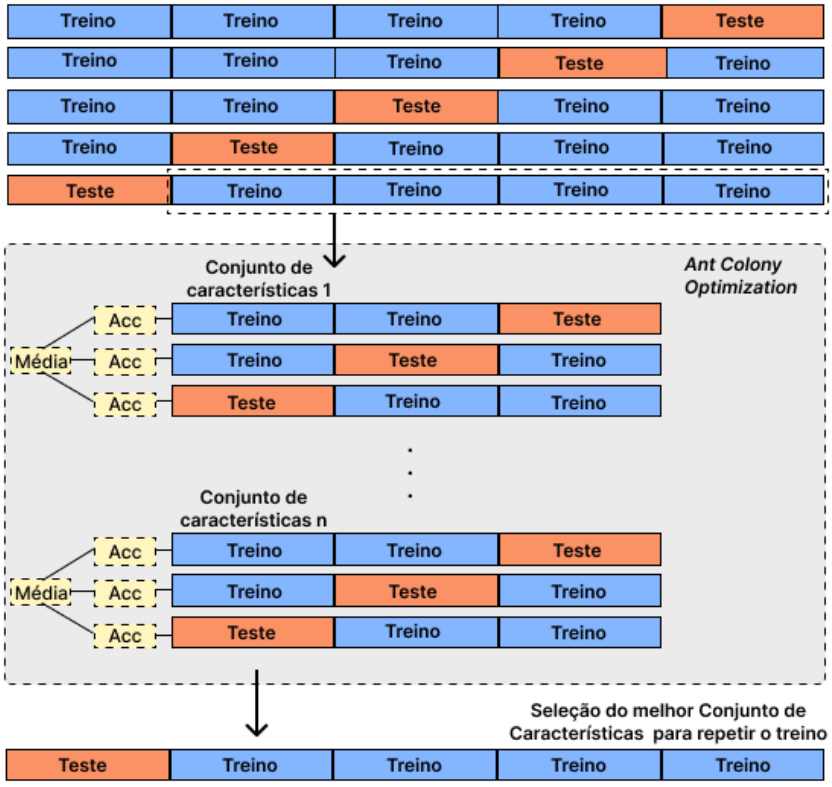

Figura 2. Ilustração do processo de *Nested Cross Validation*.

O laço externo foi dividido em *5-folds* estratificados. A cada iteração do processo de validação, o conjunto de treino desse laço era utilizado como entrada para o laço interno. No laço interno, o algoritmo ACO era empregado para a seleção de características, gerando diferentes subconjuntos de *features*, que eram avaliados com base na acurácia por meio de uma validação cruzada estratificada, interna, de *3-folds*. Após a seleção do melhor subconjunto de *features* dentro do laço interno, esse conjunto era utilizado para treinar os modelos no respectivo *fold* do laço externo. Em seguida, os modelos eram avaliados no conjunto de teste correspondente do laço externo. O desempenho final de cada modelo foi obtido a partir da média dos resultados dos *5-folds* do laço externo.

Esse processo garante que a avaliação de desempenho seja realizada em dados completamente independentes daqueles utilizados para a seleção de características, reduzindo o risco de sobreajuste e proporcionando uma estimativa mais confiável da performance dos modelos. Além disso, a redução da quantidade de *features* contribui para mitigar o *mal da dimensionalidade*, favorecendo a generalização dos modelos.

Como cada indivíduo possuía três instâncias de dados na base utilizada, a divisão dos *folds* foi realizada de maneira a preservar a integridade individual, garantindo que todas as instâncias de um mesmo participante permanecessem no mesmo subconjunto, como avaliado anteriormente por da Silva et al. (2024), porém para um conjunto de dados de voz distinto e utilizando outra estratégia de validação. Desta forma, assegurou-

se que um mesmo indivíduo não estivesse simultaneamente nos conjuntos de treino e teste. Essa abordagem é necessária para evitar vieses na avaliação dos modelos, uma vez que a presença de dados do mesmo indivíduo em ambos os conjuntos pode levar a um aumento inflacionado das métricas de desempenho, comprometendo a capacidade de generalização dos modelos para novos indivíduos.

As métricas de desempenho utilizadas para avaliar os modelos foram acurácia (proporção de previsões corretas), acurácia balanceada (média da sensibilidade para cada classe) e F1-Score ponderado (média harmônica entre precisão e sensibilidade, ajustada conforme o tamanho das classes).

Os classificadores utilizados durante este processo foram *Support Vector Machine* (SVM, com *kernel* 'rbf'), *K-Nearest Neighbors* (KNN, com K = 5), *Random forest* (RF, com seed = 42), *Extra trees* (ET, com seed = 42) e *Light Gradient-Boosting Machine* (LGBM, com seed = 42). O código foi feito utilizando Python, em sua versão 3.10.12.

5. Resultados e Discussão

O Cenário 1, avaliado com todas as características disponíveis na base de dados, apresentou em seu melhor resultado 85,86% de acurácia, 75,75% de acurácia balanceada e F1-Score de 84,69% com o algoritmo LGBM. O desempenho dos modelos nesta abordagem está contemplado na Tabela 2.

Tabela 2. Desempenho dos modelos utilizando todas as features disponíveis na base de dados – Cenário 1.

SVM			KNN			RF			ET			LGBM		
Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1
83,20	69,15	80,74	80,56	68,20	78,88	84,67	72,96	82,92	83,08	71,04	81,35	85,86	75,75	84,69

Os resultados do Cenário 2, apresentados na Tabela 3, detalham o desempenho dos classificadores SVM, KNN, RF, ET e LGBM explorando variações no número de *features* selecionadas pelo algoritmo *Ant Colony Optimization* (ACO). Dentre os modelos testados, o LGBM com 200 *features* selecionadas, destacado na tabela, apresentou o melhor desempenho em todas as métricas, com acurácia de 84,80%, acurácia balanceada de 75,52% e F1-score de 83,78%. A matriz de confusão para este modelo, apresentada na Tabela 4, demonstrou uma distribuição de resultados com 109 verdadeiros negativos (VN), 532 verdadeiros positivos (VP), 83 falsos positivos (FP) e 32 falsos negativos (FN).

Tabela 3. Desempenho dos modelos com diferentes quantidades de *features* selecionadas pelo Ant Colony Optimization (ACO) – Cenário 2.

	SVM			KNN			RF			ET			LGBM		
Features	Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1	Acc	Bal Acc	F1
100	81,88	68,48	79,54	78,56	67,14	76,92	82,43	69,78	80,41	**83,62**	**71,02**	**81,79**	82,81	72,83	81,75
200	82,82	68,38	80,23	81,09	68,60	79,22	82,70	70,47	80,93	82,95	70,49	81,02	**84,80**	**75,52**	**83,78**
300	82,54	69,00	80,34	81,23	68,81	79,47	82,42	69,73	80,44	82,43	70,50	80,82	**83,34**	**72,79**	**82,07**
400	82,95	68,64	80,44	80,15	67,11	78,24	82,96	70,62	81,06	83,50	72,03	81,94	**84,14**	**75,12**	**83,19**
500	82,81	68,57	80,29	80,96	68,72	79,32	83,89	72,24	82,22	82,82	70,42	81,07	**84,68**	**74,95**	**83,49**

A alta sensibilidade, de 94,4% (532 / (532 + 32)) indica que o modelo é eficaz na identificação das instâncias positivas, ou seja, ele consegue detectar a grande maioria dos casos positivos, o que é importante para a correta classificação de pacientes com a

Tabela 4. Tabela de confusão do modelo LGBM com 200 *features*.

	Classe Predita		
	DP	**CO**	
Classe Real	532 83	32 109	**DP** **CO**

doença de Parkinson. No entanto, a especificidade reduzida, de 56,9% (109 / (109 + 83)), sugere que algumas amostras do grupo de controle foram incorretamente classificadas como indivíduos com doença de Parkinson, o que pode indicar um viés do modelo em favor da classe majoritária. Foram utilizados os valores presentes na tabela 4 para calcular essas métricas. Os resultados apresentados na Tabela 3 também indicam que, embora o LGBM tenha se destacado, outros modelos apresentaram desempenho satisfatório. O Random Forest, por exemplo, obteve um desempenho semelhante ao LGBM, mas com métricas ligeiramente inferiores, o que sugere que sua capacidade de generalização não foi tão eficiente quanto a do LGBM.

A utilização do ACO para a seleção de *features* foi eficaz na redução da dimensionalidade dos dados sem prejudicar o desempenho dos modelos. A remoção de variáveis irrelevantes ou redundantes contribuiu para um aprendizado mais eficiente, evidenciado pelos bons resultados obtidos, especialmente pelo LGBM com 200 *features* selecionadas. Esses achados corroboram com estudos prévios que destacam o uso de técnicas de seleção de *features* como uma estratégia eficaz para treinar modelos de classificação em bases de dados com grande número de atributos, melhorando a performance e a capacidade de generalização dos modelos.

Apenas para fins de comparação, uma avaliação que desconsidera a influência da variação intra-indivíduo no treinamento dos modelos, tratando todas as amostras como independentes, foi realizada. Neste cenário, o melhor desempenho foi obtido com o LGBM (500 características), com 91,14% de acurácia, 84,45% de acurácia balanceada e 90,70% de F1-Score. Esse desempenho numericamente superior (6,34% de diferença), contudo, reforça a hipótese de que os resultados possam estar inflacionados quando analisados dessa maneira, incluindo aqueles de estudos anteriores presentes na literatura. Embora o desempenho obtido neste trabalho seja inferior ao reportado em alguns estudos, a separação por indivíduo representa uma contribuição relevante, proporcionando uma avaliação mais realista dos modelos, contribuindo para o estado da arte na área. Vale destacar, também, que o uso de técnicas de seleção de características já foi explorado anteriormente por [Parlar 2021], incluindo Information Gain, ReliefF e Wolf Search Algorithm. No entanto, uma comparação justa dos resultados não é viável, pois, ao contrário deste estudo, a variação intra-sujeito não foi considerada na avaliação dos modelos.

6. Conclusões

Neste estudo, foi proposta uma abordagem para o diagnóstico da Doença de Parkinson a partir de dados de voz, utilizando o algoritmo de Otimização por Colônia de Formigas (ACO) na seleção de características. Experimentos com diferentes quantidades de features permitiram analisar o desempenho dos classificadores e o impacto da redução di-

mensional. O modelo LGBM obteve acurácia de 84,80% com apenas 200 características. Embora o desempenho tenha sido ligeiramente inferior ao obtido com todas as features, a principal vantagem está na mitigação do enviesamento causado pela alta dimensionalidade, comum quando há equilíbrio entre número de características e instâncias, além de considerar a variabilidade intra-sujeito nas amostras de voz. A estratégia mostra-se promissora para diagnósticos em saúde digital, como na telemedicina, ao possibilitar modelos mais robustos, acessíveis e não invasivos. Apesar dos resultados mais confiáveis, o ACO envolve maior custo computacional, o que deve ser considerado em cenários com restrições de tempo ou recursos.

Os resultados mostraram que o melhor modelo apresentou uma alta sensibilidade, que identifica corretamente a maioria dos casos positivos, mas apresenta uma taxa mais elevada de falsos positivos, resultando em uma especificidade relativamente baixa. Apesar de abordagens de divisão de dados estratificadas terem sido utilizadas, este comportamento ainda pode ser atribuído ao desbalanceamento das classes. Neste sentido, trabalhos futuros podem explorar técnicas para mitigar esse problema, como o uso de SMOTE, visando equilibrar o impacto das classes minoritária e majoritária durante o treinamento dos modelos e aplicar combinação de técnicas de seleção de características. Além disso, o uso de métodos alternativos de seleção de características pode melhorar consideravelmente a performance do modelo.

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 Nº 30/2022 – PDPG – Solidariedade Acadêmica.

Referências

Braak, H. and Braak, E. (2000). Pathoanatomy of parkinson's disease. *J Neurol*, 247 Suppl 2:II3–10.

da Silva, M. I., Felix, J. P., Prado, T. d. S., Chagas, A. L. d. B., 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).

Daliri, M. R. (2012). Automatic diagnosis of neuro-degenerative diseases using gait dynamics. *Measurement*, 45(7):1729–1734.

Darwish, A. (2018). Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. *Future Computing and Informatics Journal*, 3(2):231–246.

Dorigo, M., Birattari, M., and Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39.

Dorigo, M., Maniezzo, V., and Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1):29–41.

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 on Electrical and Computer Engineering (CCECE)*, pages 1–5, Vancouver, Canada. IEEE. To appear.

Govindu, A. and Palwe, S. (2023). Early detection of parkinson's disease using machine learning. *Procedia Computer Science*, 218:249–261. International Conference on Machine Learning and Data Engineering.

Hayes, M. T. (2019). Parkinson's disease and parkinsonism. *The American Journal of Medicine*, 132(7):802–807.

Ho, A., Iansek, R., Marigliani, C., Bradshaw, J., and Gates, S. (1998). Speech impairment in a large sample of patients with parkinson's disease. *Behavioural neurology*, 11:131–137.

Little, M., Mcsharry, P., Hunter, E., Spielman, J., and Ramig, L. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE transactions on bio-medical engineering*, 56:1015.

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.

Ouhmida, A., Terrada, O., Raihani, A., Cherradi, B., and Hamida, S. (2021). Voice-based deep learning medical diagnosis system for parkinson's disease prediction. In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pages 1–5.

Parlar, T. (2021). A heuristic approach with artificial neural network for parkinson's disease. *International Journal of Applied Mathematics Electronics and Computers*, 9:1–6.

Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A. E. (2017). Parkinson disease. *Nature Reviews Disease Primers*, 3(1):1–21.

Prusty, S., Patnaik, S., and Dash, S. K. (2022). Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4.

Rana, A., Dumka, A., Singh, R., Rashid, M., Ahmad, N., and Panda, M. (2022). An efficient machine learning approach for diagnosing parkinson's disease by utilizing voice features. *Electronics*, 11:3782.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., and Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. *Applied Soft Computing*, 74:255–263.

Solana-Lavalle, G., Galan-Hernandez, J., and Rosas-Romero, R. (2020). Automatic parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering*, 40.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Trans Biomed Eng*, 59(5):1264–1271.

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.