CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Sobre a análise de sinais de voz para o diagnóstico da doença de Parkinson

On the voice signal analysis for the diagnosis of Parkinson's disease

Sobre el análisis de la señal de voz para el diagnóstico de la enfermedad de Parkinson

Matheus Isac da Silva[1], Juliana Paula Felix[2], Thiago de Stecca Prado[1], Ana Luísa de Bastos Chagas[1], Giordana de Farias Franco Bueno Bucci[1], Afonso Ueslei da Fonseca[2], Fabrizzio Soares[2]

1 Graduando(a), Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil
2 Prof(a). Dr(a), Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil

Autor correspondente: Profa. Dra. Juliana Paula Felix
E-mail: julianafelix@ufg.br

Resumo

Objetivo: Este estudo investiga se o possível viés na sobreamostragem via janelamento de dados de marcha em indivíduos com Doença de Parkinson (DP) também ocorre em sinais vocais. Um estudo anterior levantou a hipótese de que amostras distintas de um mesmo indivíduo não devem ser tratadas independentemente, dado o risco de enviesamento dos modelos. Método: Usamos sinais de voz de 24 indivíduos com DP e 8 saudáveis, e os algoritmos K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Random Forest (RF). A validação cruzada foi feita com Leave-one-out (LOOCV), adaptada para cenários com e sem viés nos dados de treinamento. Resultados: Modelos avaliados sem considerar o viés apresentaram performances inflacionadas, enquanto a abordagem rigorosa mostrou resultados mais modestos. Conclusão: Amostras do mesmo indivíduo em treinamento e teste podem inflar a performance dos modelos. A correta aplicação da sobreamostragem é crucial para desenvolver modelos confiáveis para o diagnóstico de DP.

Descritores: Doença de Parkinson; Aprendizado de Máquina; Diagnóstico.

Abstract

Objective: This study investigates whether the potential bias in oversampling through windowing gait data in individuals with Parkinson's Disease (PD) also occurs in vocal signals. A previous study hypothesized that distinct samples from the same individual should not be treated independently, due to the risk of biasing the models. Method: We used voice signals from 24 individuals with PD and 8 healthy subjects, and the algorithms K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Cross-validation was performed with Leave-one-out (LOOCV), adapted for scenarios with and without bias in the training data. Results: Models evaluated without considering bias showed inflated performances, while the rigorous approach yielded more modest results. Conclusion: Samples from the same individual in training and testing may inflate the models' performance. Proper application of oversampling is crucial for developing reliable models for PD diagnosis.

Keywords: Parkinson Disease; Machine Learning; Diagnosis

Resumen

Objetivo: Este estudio investiga si el posible sesgo en el sobremuestreo a través del enventanado de datos de marcha en individuos con enfermedad de Parkinson (EP) también ocurre en señales vocales. Un estudio previo planteó la hipótesis de que muestras distintas de un mismo individuo no deben tratarse independientemente, debido al riesgo de sesgar los modelos. Método: Utilizamos señales de voz de 24 individuos con EP y 8 sanos, y los algoritmos K-Nearest Neighbors (KNN), Support Vector Machine (SVM) y Random Forest (RF). La validación cruzada se realizó con Leave-one-out (LOOCV), adaptada para escenarios con y sin sesgo en los datos de entrenamiento. Resultados: Los modelos evaluados sin considerar el sesgo presentaron desempeños inflados, mientras que el enfoque riguroso mostró resultados más modestos. Conclusión: Las muestras del mismo individuo en entrenamiento y prueba pueden inflar el desempeño de los modelos. La correcta aplicación del sobremuestreo es crucial para desarrollar modelos confiables para el diagnóstico de EP.

Descriptores: Enfermedad de Parkinson; Aprendizaje de Máquina; Diagnóstico

Introdução

A Doença de Parkinson (DP) é uma doença neurodegenerativa progressiva que afeta a mobilidade, a fala e a postura, causando tremores, rigidez muscular e bradicinesia[1]. A doença é causada pela morte de neurônios, resultando na diminuição dos níveis de dopamina no cérebro e, por sucessão, dificulta a comunicação entre sinapses, que causa deterioramento das funções motoras[2]. A DP tem uma prevalência de aproximadamente 0,5 a 1 por cento entre aqueles com 65 a 69 anos de idade, aumentando para 1 a 3 por cento entre pessoas com 80 anos de idade ou mais[3], sendo a segunda doença neurodegenerativa mais comum depois do Alzheimer[4]. Espera-se que tanto a prevalência como a incidência da DP aumentem em mais de 30% até 2030, com o envelhecimento da população[3].

A maioria dos indivíduos diagnosticados com DP desenvolve distúrbios de voz e fala durante o curso da doença[5]. Volume vocal reduzido, voz monótona e soprosa ou rouca, e articulação imprecisa são as principais características da fala parkinsoniana[6]. Esses distúrbios de voz e fala, denominados coletivamente de disartria hipocinética, podem estar entre os primeiros sinais da DP[7]. Não há cura para a DP, de forma que os pacientes dependem de detecção precoce e tratamentos personalizados para retardar o progresso da doença[7] e assegurar uma melhor qualidade de vida. Neste sentido, dados acústicos têm sido utilizados para descrever as características vocais de indivíduos com DP[7], e são diversos os trabalhos que propõem o uso de aprendizado de máquina para auxiliar no diagnóstico da DP a partir da classificação de sinais de voz[7,8,9,10,11].

Devido à raridade da doença, as bases de dados de voz de pessoas com DP disponíveis publicamente geralmente têm um número limitado de participantes (poucas dezenas). Como solução, muitos estudos coletam múltiplas amostras de um mesmo indivíduo, aumentando assim a representatividade do conjunto total de amostras e permitindo uma análise mais abrangente da população em estudo.

Entretanto, um estudo recente levantou a hipótese de que, ao realizar experimentos de aprendizado de máquina para classificação de doenças neurodegenerativas, como a Doença de Parkinson, amostras obtidas a partir de um mesmo indivíduo não deveriam ser tratadas de forma independente na modelagem e avaliação dos algoritmos de classificação[12], como é frequentemente observado em

estudos encontrados na literatura. Essa hipótese é avaliada pelos autores do estudo utilizando-se sinais de marcha de pessoas com a Doença de Parkinson, Doença de Huntington e Esclerose Lateral Amiotrófica, todas doenças neurodegenerativas que possuem como sintoma comum alterações na marcha. Os autores do estudo avaliam a performance de diferentes classificadores considerando dois cenários de avaliação distintos: Cenário 1, em que as amostras são todas tratadas de forma independente; e Cenário 2, em que a avaliação do classificador considera que amostras distintas de um mesmo indivíduo devem figurar exclusivamente no conjunto de treinamento ou no conjunto de teste, nunca em ambos os conjuntos no mesmo ciclo de avaliação.

Neste trabalho, investigamos a hipótese descrita por Chagas et al.[12] em um novo domínio, agora considerando a classificação de sinais de voz de pessoas com a Doença de Parkinson. Tendo em vista que distúrbios de voz e fala podem estar entre os primeiros sinais de DP[4], acreditamos que o comportamento observado anteriormente para sinais de marcha de pessoas com doenças neurodegenerativas se repetirá na análise automática de sinais de voz. No restante deste artigo, apresentamos uma breve revisão da literatura. Na sequência, os materiais e métodos são descritos, incluindo a base de dados utilizada e posterior esclarecimento sobre a metodologia utilizada. Finalmente, os resultados são apresentados e discutidos, e as conclusões do estudo são apresentadas.

Referencial Teórico

Diversos são os trabalhos que propõem a análise de características vocais por meio da inteligência artificial para auxiliar no diagnóstico da DP. Um dos estudos pioneiros neste campo foi realizado por Little et al. (2009)[7], que avaliaram medidas de disfonia para discriminar pessoas saudáveis de pessoas com DP. O estudo contou com a participação de 31 pessoas, sendo 23 com DP, das quais foram coletadas um total de 195 amostras vocais, com média de 6 amostras por participante. Diversas características foram geradas a partir de cada amostra de sinal de voz, e métodos de seleção de características foram aplicados. Os autores reportam uma acurácia média de 91,4% obtida com uso da Máquina de Vetores de Suporte (SVM) com kernel de base radial gaussiana. A base de dados coletada por

Little et al.[7] foi disponibilizada publicamente no repositório da University of California Irvine (UCI), sob o nome "*Oxford Parkinson's Disease Detection Dataset*", e vem servindo de base para novos estudos que envolvem aprendizado de máquina.

Aich et al. (2018)[8] apresentam uma abordagem de aprendizado de máquina supervisionado para distinguir dados de voz de pessoas com DP de indivíduos saudáveis. Uma abordagem de seleção de características utilizando Análise de Componentes Principais (PCA) e Algoritmo Genético (GA) foi utilizada. Os resultados são avaliados utilizando-se uma abordagem de separação de 70% para treino e 30% para teste. Os autores reportam uma acurácia média de 97,57% no conjunto de teste utilizando Máquinas de Vetores de Suporte (SVM) alimentadas por características selecionadas pelo algoritmo genético.

Ouhmida et al. (2021)[9] apresentaram um método para classificação de sinais acústicos de pessoas com a Doença de Parkinson e pessoas saudáveis. A abordagem reportada utiliza redes neurais convolucionais (CNN) e redes neurais artificiais (ANN), atingindo resultados de 93,10% de acurácia para a base de dados disponibilizada por Little et. al (2009)[7], e 88,89% em um segundo conjunto de dados disponível publicamente. Rana et al.(2022)[10] apresenta uma comparação de quatro algoritmos de aprendizado de máquina (SVM, Naive Bayes, K-Nearest Neighbors e ANN) para auxiliar no diagnóstico da doença de Parkinson a partir de gravações de voz. A abordagem investigada por Rana et al.[10] reporta uma acurácia média de 96,7% utilizando ANN, seguida por 87,17% atingidos por ambos SVM e KNN, e 74,11% com a Naive Bayes.

Govindu e Palwe (2023)[11] reportam a utilização de quatro modelos de aprendizado de máquina – regressão logística, SVM, Random Forest e KNN - para classificação de sinais de voz da base de dados de Little et al. (2009)[7]. Três abordagens diferentes foram comparadas utilizando os classificadores apontados: usando o conjunto de dados completo, usando atributos identificados através da análise de componentes principais, e utilizando uma estratégia de balanceamento dos dados. Os resultados reportados sugerem que o KNN, quando utilizado com dados balanceados, atinge uma acurácia de 91,83%.

Diversos outros trabalhos que realizam a classificação de sinais de voz de pessoas saudáveis e pessoas com Parkinson podem ser encontrados na literatura,

como apontados na revisão de literatura por Ngo et al.[13]. Entretanto, apesar da relevância dos estudos prévios, e altos valores de acurácia reportados, percebe-se que, predominantemente, os resultados discutidos nos estudos anteriores referem-se à taxa de acerto da classificação das amostras de voz, que são tratadas como amostras independente pelos algoritmos de aprendizado de máquina investigados, mesmo quando provenientes de um mesmo indivíduo. Neste sentido, este estudo propõe a investigação desse viés reportado por Chagas et al.[12] sob a ótica de dados de voz de pessoas com Parkinson, propondo a verificação dos resultados reportados por modelos de machine learning quando há o cuidado de não comprometer o conjunto de treinamento com amostras de um mesmo indivíduo que pode figurar no conjunto de teste. A seção seguinte descreve os materiais e o método proposto.

Materiais e Métodos

Neste trabalho foi utilizada a base de dados "*Oxford Parkinson's Disease Detection Dataset",* desenvolvida por Little et al. (2009)[7] e disponibilizada publicamente no *UC Irvine Machine Learning Repository*. Tal dataset contém um conjunto de dados de 195 fonações de vogais sustentadas, colhidas de 31 pessoas com faixa etária entre 46 e 85 anos (média de 65,8, desvio padrão de 9,8), sendo 24 indivíduos com diagnóstico de Doença de Parkinson, e 8 indivíduos saudáveis (grupo de controle). De cada participante obteve-se, em média, 6 áudios sustentados, que variam de 1 a 36 segundos de duração[7].

Cada fonação foi gravada em uma cabine acústica da *Industrial Acoustics Company (IAC)*, utilizando um microfone head-mounted (AKG C450) posicionado a 8 cm dos lábios. Foi colocado um medidor de nível sonoro classe 1 (B&K 2238) a 30cm do alto-falante. As gravações foram feitas pelo hardware *Computerized Speech Laboratory (CSL)* 4300B *(Kay Elemetrics)*, amostradas em 44,1 kHz com resolução de 16 bits. Os dados acústicos disponíveis nessa base são dados já calculados, tratados e preparados de acordo com o apresentado na Tabela 1. No total, 22 características (*features*) estão disponíveis.

As 22 características vocais foram usadas para alimentar os algoritmos selecionados: K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e

Random Forest Classifier (RF). Uma breve explicação sobre esses algoritmos e detalhes das configurações avaliadas são apresentadas na seção seguinte. A avaliação dos métodos foi realizada com a validação cruzada via Leave-one-out (LOOCV), e adaptada para os dois cenários discutidos neste trabalho: com e sem viés no conjunto de treinamento. Métricas de avaliação foram calculadas e serão apresentadas e discutidas. Os experimentos foram desenvolvidos utilizando a linguagem *Python*, versão 3.10.9, e a biblioteca *Scikit-learn.*

Tabela 1 – Características disponíveis na base de dados.

Features	Significado
F0 (Hz)	Média das Frequências da voz em relação ao tempo.
Fhi (Hz)	Pico da Frequência de voz.
Flo (Hz)	Menor valor da Frequência de voz.
Jitter (%)	Diferença média absoluta entre períodos consecutivos, dividida pelo período médio.
Jitter (µs)	Diferença média absoluta entre períodos consecutivos, em segundos.
RAP (%)	Perturbação Média Relativa, a diferença média absoluta entre um período e a média dele e dos seus dois vizinhos, dividida pelo período médio.
PPQ (%)	Quociente de Perturbação do Período de cinco pontos, a diferença média absoluta entre um período e a média dele e dos seus quatro vizinhos mais próximos, dividida pelo período médio.
DDP	Diferença média absoluta entre diferenças consecutivas entre períodos consecutivos, dividida pelo período médio.
Shimmer (%)	Diferença média absoluta entre as amplitudes de períodos consecutivos, dividida pela amplitude média.
Shimmer (dB)	Logaritmo médio absoluto de base 10 da diferença entre as amplitudes de períodos consecutivos, multiplicado por 20.
APQ11	Quociente de Perturbação de Amplitude de onze pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes de seus vizinhos, dividida pela amplitude média.
APQ3	Quociente de Perturbação de Amplitude de três pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes de seus vizinhos, dividida pela amplitude média.
APQ5	Quociente de Perturbação de Amplitude de cinco pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes dele e de seus quatro vizinhos mais próximos, dividida pela amplitude média.
DDA	Diferença média absoluta entre as amplitudes de períodos consecutivos.
NHR	Relação ruído-harmônicos.
HNR	Razão Harmônica-Ruído
RPDE	Medida de entropia de densidade do período de recorrência
DFA	Análise de flutuação de tendência
Spread 1	Duas medidas não lineares da frequência fundamental
Spread 2	Variação de frequência
PPE	Entropia de período de pitch

K-Nearest Neighbours (KNN)

O algoritmo de classificação *K-Nearest Neighbours*, ou KNN, é um dos algoritmos de aprendizado de máquina mais simples. Ele consiste na disposição dos

dados em um espaço dimensional n, sendo n o número de atributos em questão sendo analisados. Para inferir a classe de um novo dado amostrado, necessitamos dispô-lo neste espaço e contabilizar as classes dos dos "k" (coeficiente a determinar) elementos mais próximos, baseados em uma métrica de distância a ser definida. A classe majoritária determinará a classe do novo dado analisado[14].

Neste trabalho, analisamos o desempenho de três variações do KNN, considerando os k = 5, 7 ou 10 vizinhos mais próximos. As três variações do algoritmo consideram pesos iguais ao inverso da distância entre os dois pontos, ou seja, os demais vizinhos terão menor influência do que os vizinhos mais próximos de um ponto de consulta, e contam com a mesma quantidade de memória a ser alocada para o armazenamento da árvore (*'leaf_size'*=100). A distância utilizada nos algoritmos é a distância euclidiana[14], tal que para dois pontos de coordenadas (x_1, y_1) e (x_2, y_2) no plano cartesiano, pode ser calculada conforme a Equação (1).

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

Support Vector Machine (SVM)

A Máquina de Vetores de Suporte, ou *Support Vector Machine* (SVM), é um modelo classificador bastante utilizado na literatura, sendo capaz de fornecer resultados precisos e altamente robustos. O objetivo desse modelo é classificar dados de treinamento separando as classes[11]. Uma das vantagens do SVM é sua versatilidade, possibilitando o uso de diferentes *kernels*, que podem ser especificados para a função de decisão. Os kernels utilizados neste trabalho foram:

- **Linear:** Define uma fronteira linear a partir de dados linearmente separáveis, separando tais dados com um hiperplano definido pela Equação (2),

$$f(x) = wx + b = 0, \quad (2)$$

  onde *w* é o vetor de pesos perpendicular ao hiperplano de separação, *b* é um escalar e x é um objeto do conjunto de treinamento[14].

- **Radial Basis Function (RBF):** Um dos kernels mais utilizados, o RBF é capaz de combinar vários kernels polinomiais, de diferentes graus, e múltiplas vezes. Usa a chamada função de base radial, que pode ser definida pela Equação (3)[14], na qual o valor de sigma (σ) utilizado foi 1.

$$K(x_i, x_j) = exp\left(-\sigma\left\|x_i - x_j\right\|^2\right). \quad (3)$$

Random Forest Classifier (RF)

O algoritmo *Random Forest Classifier*, ou RF, consiste na elaboração de árvores de decisão utilizando técnicas mais simples como regressão linear ou KNN. Todavia, o principal diferencial é o treinamento de várias árvores de decisão com subconjuntos menores dos dados originais, processo conhecido como *bootstrap*. Por fim, a média ou a moda dos resultados de cada árvore treinada será escolhida para definir a classificação de quaisquer dados novos apresentados.[14] Neste trabalho, definimos em 100 o número limite de árvores de decisão a serem treinadas para cada algoritmo de RF.

Leave-one-out Cross-Validation

Para determinar o desempenho de cada classificador, geralmente um modelo é treinado com os dados disponíveis. Em seguida, o desempenho da classificação é avaliado usando dados recém-coletados. Quando não há a disponibilidade de dados novos específicos para a fase de teste, uma parte do conjunto de dados original é separado para a fase de teste. Para superar limitações como o tamanho do banco de dados, o desequilíbrio de dados e a possibilidade de *overfitting*, ao invés de treinar um modelo fixo apenas uma vez, como em uma divisão de treinamento/teste, a abordagem de *cross-validation* (CV) é fortemente recomendada[15].

Na abordagem conhecida como k-fold cross-validation, a validação cruzada é realizada por k vezes, cada vez usando um particionamento diferente dos dados em conjuntos de treinamento e teste, sendo reportado a média dos resultados obtidos para as k dobras da CV. Neste trabalho, utilizamos o caso especial do k-fold *cross-validation* conhecido como *leave-one-out* (LOOCV), em que cada divisão do conjunto de dados separa n-1 amostras para a fase de treino, e exatamente uma amostra é 'deixada de fora', sendo então utilizada para a fase de teste. Essa divisão ocorre por k=n vezes, sendo n o número de amostras disponíveis no conjunto de dados, de modo que todas as amostras tenham sido utilizadas, em algum momento, na fase de teste. Para avaliar a hipótese do enviesamento dos modelos que realizam a classificação de dados de voz de pessoas com DP, conduzimos os experimentos considerando-se dois cenários do LOOCV, apresentados com detalhe na Figura 1:

- ➔ **Cenário 1**: As 195 amostras (147 DP, 48 CO) disponíveis na base de dados são consideradas como amostras independentes. Neste caso, para cada iteração do LOOCV, uma amostra por vez é 'deixada de fora' do treinamento e utilizada para teste, como mostrado na Figura 1a.
- ➔ **Cenário 2**: O LOOCV opera sobre a quantidade total de indivíduos (24 DP, 8 CO) dos quais as amostras foram coletadas. Em cada iteração do LOOCV, as aproximadamente 6 amostras coletadas de cada indivíduo são destinadas ao mesmo conjunto, de treinamento ou de teste, de forma exclusiva. Neste cenário, garantimos que o modelo a ser testado nunca tenha sido alimentado por amostras de um mesmo indivíduo separado para teste.

Métricas de análise de desempenho

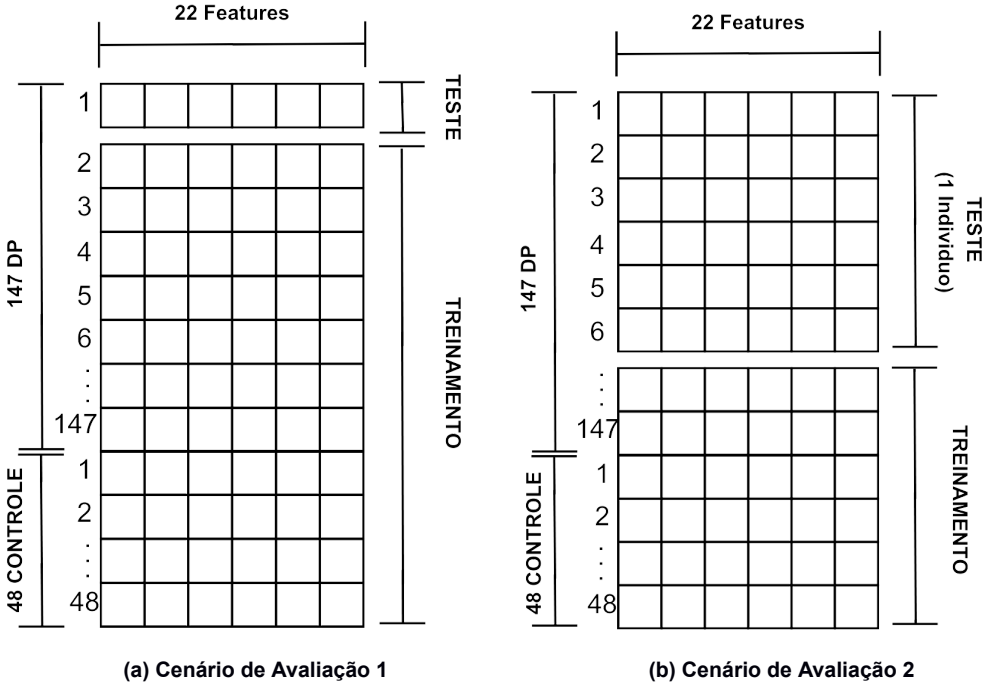
Ao calcular a performance de um modelo preditivo, torna-se essencial determinar uma ou mais métricas de avaliação. Na avaliação dos modelos comparados neste artigo, utilizamos as métricas de acurácia, sensibilidade e especificidade, recomendadas quando se trabalha com predição automática de diagnósticos[16]. Para calculá-las, faz-se necessário definir os seguintes valores:

- **Verdadeiro Positivo (VP):** Acerto do modelo em relação aos dados de pessoas com DP, predições corretas.
- **Verdadeiro Negativo (VN):** Acertos do modelo quanto aos dados de pessoas saudáveis, predições corretas.
- **Falso Positivo (FP):** Erro do modelo que, no contexto desse trabalho, falsamente acusa as amostras como sendo de indivíduos com DP.
- **Falso Negativo (FN):** Erro do modelo que, no contexto desse trabalho, falsamente prediz que a amostra analisada é de um indivíduo saudável.

A acurácia determina o quão próximo o valor real está da saída do classificador. Em outras palavras, é a divisão da soma dos acertos em relação à soma do total de amostras analisadas, dada pela Equação (4)[17].

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (4)$$

Figura 1 – Ilustração dos dois cenários de validação cruzada avaliados neste trabalho, sendo (a) o cenário de avaliação por dados independentes e (b) o cenário de avaliação por indivíduo.



A sensibilidade, ou *recall,* é a porcentagem de acertos dos casos positivos, ou seja, é a porcentagem de amostras com a DP que foram corretamente classificadas, e pode ser calculada pela Equação (5)[17].

$$Sensibilidade = \frac{VP}{VP+FN} \quad (5)$$

A especificidade, também chamada de precisão, corresponde à porcentagem de amostras de pessoas saudáveis (verdadeiros negativos) corretamente classificadas, sendo calculada pela Equação (6)[17].

$$Especificidade = \frac{VN}{VN+FP} \quad (6)$$

Resultados e Discussão

Os resultados dos seis classificadores avaliados (KNN-5, KNN-7, KNN-10, SVM-Linear, SVM-RBF e RF) estão resumidos nas Tabelas 2 e 3 para os Cenários de Avaliação 1 e 2, respectivamente. Cada tabela apresenta as métricas de acurácia, sensibilidade e especificidade médias para os métodos de avaliação discutidos anteriormente.

Considerando-se, inicialmente, apenas o Cenário de Avaliação 1 (Tabela 2), observa-se que os 6 classificadores analisados retornam acurácias médias que variam de 75,38% (SVM-RBF) a 92,31% (RF). Ao analisar os valores de sensibilidade e especificidade, é possível compreender melhor o comportamento dos modelos construídos. Por exemplo, apesar do SVM-RBF obter uma acurácia média de 75,38%, a especificidade é de 0%, ou seja, o modelo construído classificou erroneamente todas as amostras de controle como sendo de indivíduos com DP.

No Cenário 2 (Tabela 3), o LOOCV foi realizado para os pacientes. Os resultados da Tabela 3 indicam que os valores de acurácia, sensibilidade e especificidade para todos os classificadores são menores em comparação com o Cenário 1. Isso sugere que, embora o Cenário 1 use dados de teste nunca vistos pelos modelos, a redução na acurácia sugere que a presença de amostras distintas do mesmo indivíduo no conjunto de treinamento pode influenciar na precisão dos modelos ao lidar com amostras desse mesmo indivíduo no conjunto de teste.

Tabela 2 – Resultados da abordagem LOOCV por amostras, Cenário 1.

Modelo	Acurácia	Sensibilidade	Especificidade
KNN5	86,15%	94,00%	62,00%
KNN7	85,64%	94,00%	56,00%
KNN10	84,62%	94,00%	56,00%
SVM-RBF	75,38%	100,00%	0,00%
SVM-Linear	86,15%	96,00%	56,00%
RF	92,31%	97,00%	75,00%

Tabela 3 – Resultados da abordagem LOOCV por pessoa, Cenário 2.

Modelo	Acurácia	Sensibilidade	Especificidade
KNN5	70,68%	85,00%	27,00%
KNN7	71,58%	86,00%	29,00%
KNN10	74,55%	90,00%	29,00%
SVM-RBF	75,00%	100,00%	0,00%
SVM-Linear	83,33%	94,00%	52,00%
RF	76,56%	92,00%	35,00%

Outra observação a ser constatada foi de que houve alteração no modelo que melhor se comportou em cada tarefa de classificação. Enquanto no Cenário 1, o *Random Forest Classifier* foi o modelo que apresentou melhor resultado (92,31% acc., 97% sens. e 75% espec.), este mesmo classificador no Cenário 2 apresentou apenas 76,56% de acurácia, com queda brusca de especificidade, que passou para 35%. No Cenário 2, o modelo de melhor resultado foi o *Support Vector Machine* com kernel linear (83,33% acc., 94% sens. e 52% espec.).

Conclusão

Neste projeto, 22 características obtidas a partir de sinais de voz de pessoas com a Doença de Parkinson e indivíduos saudáveis alimentaram 6 diferentes classificadores de aprendizados de máquina (KNN5, KNN7, KNN10, SVM-RBF, SVM-Linear e RF), que forem avaliados pela técnica de validação cruzada *Leave-One-Out* em dois cenários: 1), em que as amostras são tratadas de modo independente, e 2), em que a validação é feita sobre a quantidade de pessoas envolvidas em cada iteração de classificação. Os resultados obtidos revelam que o Cenário 2 apresenta valores inferiores de acurácia, sensibilidade e especificidade

quando comparados ao Cenário 1, sugerindo a presença de um enviesamento no treinamento dos modelos avaliados conforme o primeiro cenário. Desta forma, conclui-se que amostras distintas de sinais de voz de um mesmo indivíduo, sejam elas de pessoas com Doença de Parkinson ou saudáveis, não deveriam ser tratadas de forma independente. A análise realizada neste trabalho corrobora com a hipótese levantada por Chagas et al.[12], agora avaliada para classificação automática de dados de voz de pessoas com a Doença de Parkinson. É importante ressaltar, no entanto, que o número reduzido de amostras disponíveis na base de dados utilizada neste trabalho (provenientes de 24 pessoas com DP e 8 saudáveis), pode comprometer a generalização dos resultados. Assim, novos experimentos com um número adequado de amostras a fim de assegurar a significância estatística dos resultados aqui apresentados é recomendada. Portanto, como trabalhos futuros, propomos a extensão desta avaliação para outros conjuntos de dados de voz de pessoas com a Doença de Parkinson, bem como a comparação de desempenho deste e de novos conjuntos de dados com outros classificadores.

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. A autora Juliana Paula Felix agradece a Bolsa CAPES/BRASIL.

Referências

1. Prabhavathi, K., and Shantanu Patil. Tremors and bradykinesia. Techniques for Assessment of Parkinsonism for Diagnosis and Rehabilitation (2022): 135-149.

2. Braak, Heiko, and Eva Braak. Pathoanatomy of Parkinson's disease. Journal of neurology 247 (2000): II3-II10.

3. Tanner, Caroline M. Epidemiology of Parkinson's disease. Neurologic clinics 10.2 (1992): 317-329.

4. Stewart A. Factor, William J. Weiner (2008) Parkinson Disease - Diagnosis and Clinical Management 2nd ed; 77-94

5. Ho, Aileen K., et al. Speech impairment in a large sample of patients with Parkinson's disease. Behavioural neurology 11.3 (1998): 131-137.

6. Atarachi, J., and E. Uchida. A clinical study of Parkinsonism. Recent Adv Res Nerv Syst 1959; 3: 871 882 (1959).

7. Little, Max, et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nature Precedings (2008): 1-1.

8. Aich, Satyabrata, et al. A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease. 2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019.

9. Ouhmida, Asmae, et al. Voice-Based Deep Learning Medical Diagnosis System for Parkinson's Disease Prediction. 2021 International Congress of Advanced Technology and Engineering (ICOTEN). IEEE, 2021.

10. Rana, Arti, et al. An efficient machine learning approach for diagnosing Parkinson's disease by utilizing voice features. Electronics 11.22 (2022): 3782.

11. Govindu, Aditi, and Sushila Palwe. Early detection of Parkinson's disease using machine learning. Procedia Computer Science 218 (2023): 249-261.

12. Chagas, A., Bucci, G., Félix, J., Fonseca, A., Nascimento, H., & Soares, F. (2024). Avaliando a Sobreamostragem de Dados Temporais de Marcha no Diagnóstico Automático de Doenças Neurodegenerativas. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, (pp. 567-578). Porto Alegre: SBC. doi:10.5753/sbcas.2024.2776

13. Quoc Cuong Ngo, Mohammod Abdul Motin, Nemuel Daniel Pah, Drotar P, Kempster P, Kumar D. Computerized analysis of speech and voice for Parkinson's disease: A systematic review. Computer Methods and Programs in Biomedicine. 2022 Nov 1;226:107133–3.

14. Faceli, K., et al. Inteligência Artificial: Uma abordagem de aprendizagem de máquina, LTC, Ed. Rio de Janeiro: Grupo Editorial Nacional (2011).

15. Duda, Richard O., and Peter E. Hart. Pattern classification. John Wiley & Sons, 2006.

16. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994 Jun 11;308(6943):1552. doi: 10.1136/bmj.308.6943.1552. PMID: 8019315; PMCID: PMC2540489.

17. Gunawardana, Asela, and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. Journal of Machine Learning Research 10.12 (2009).