UPF EDITORA
UNIVERSIDADE DE PASSO FUNDO

RBCA
ISSN 2176-6649

Revista Brasileira de Computação Aplicada, Julho, 2023

DOI: 10.5335/rbca.v15i2.13556
Vol. 15, Nº 2, pp. 88–104
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Classificação de sinais de voz para auxílio no diagnóstico da doença de Parkinson

Classification of voice signals to aid in the diagnosis of Parkinson's disease

Mateus Melo[1], and Thiago Gouveia[1]

[1]Instituto Federal da Paraíba (IFPB)

*mateus.kopelman@academico.ifpb.edu.br; thiago.gouveia@ifpb.edu.br

Recebido: 23/05/2022. **Revisado:** 09/04/2023. **Aceito:** 02/06/2023.

Resumo

A doença de Parkinson é uma patologia neurodegenarativa que afeta a capacidade motora e de fala, além de provocar alterações comportamentais, de humor e de raciocínio. Ela atinge, mais usualmente, a população idosa e seu diagnóstico é feito por meio de um exame clínico, através da observação dos sintomas apresentados por um paciente. Uma vez que os sintomas mais notórios costumam aparecer em estágios mais avançados da doença, o que dificulta o tratamento, e que o mundo vem passando por um processo de inversão da pirâmide etária, a tendência é que o Parkinson venha a se tornar um problema de saúde pública mundial. Dentro deste contexto, propostas para a utilização de sinais de voz como forma de diagnóstico precoce do Parkinson veem obtendo resultados. Este trabalho propõe a construção de uma ferramenta de auxílio no dignóstico da doença de Parkinson utilizando sinais de voz associados a técnicas de Aprendizado de Máquina. Fazendo uso de um conjunto de dados com atributos diversos extraídos da fala de portadores e não portadores da patologia, obteve-se uma acurácia 93.8 % utilizando um algoritmo de Random Forest e efetuando uma validação cruzada com a técnica k-fold.

Palavras-Chave: Aprendizado de Máquina; doença de Parkinson; random forest; validação cruzada k-fold

Abstract

The Parkinson's disease is a neurodegenerative illness which impairs motor and speech skills, in addition to provoke behavior, mood and thinking changes. It hits, more usually, the elderly population and its diagnosis is done by a clinical exam, by the observation of a patient's symptoms. Since the most notorious symptoms appear in advanced stages of the disease, which makes the treatment more difficult, and that the world is passing through a process of age pyramid inversion, the tendency is that the Parkinson will become a global public health issue. Within this context, proposals for the use of voice signals as a form of early diagnosis of Parkinson's have been achieving results. This work proposes the construction of a tool to aid in the diagnosis of Parkinson's disease using voice signals associated with Machine Learning techniques. Using a set of data with different attributes extracted from the speech of carriers and non-carriers of the pathology, an accuracy of 93.8 % was obtained using a Random Forest algorithm and performing a cross-validation with the k-fold technique.

Keywords: K-fold cross validation; machine Learning; parkinson's disease; random Forest

1 Introdução

A doença de Parkinson é uma patologia neurodegenerativa que prejudica e diminui a capacidade motora e de fala, bem como provoca alterações comportamentais, de humor e de raciocínio aos acometidos por ela. O Parkinson não possui predileções por etnia, gênero ou classe social e atinge mais frequentemente a idosos.

O Parkinson é diagnosticado de forma clínica por um neurologista através da exclusão de outras doenças e, atualmente, não há um teste ou biomarcador que possa indicar a presença da doença (Steidl et al., 2007). Isto faz com que o Parkinson seja frequentemente diagnosticado tardiamente, uma vez que os principais sintomas aparecem apenas em estágios mais avançados da doença.

Estima-se que a doença de Parkinson atinja até 1% da população mundial acima de 65 anos (Massano and Cabreira, 2019) e com o processo de aumento de expectativa de vida, a tendência é que o Parkinson se torne um problema de de saúde pública mundial. No Brasil há o agravante que até o ano de 2050 terá cerca de 19% de sua população com idade superior a 65 anos (Nasri, 2008).

Neste cenário, diversas pesquisas que utilizam a análise acústica vocal associada a técnicas de Aprendizado de Máquina se apresentam como alternativa que possibilitaria um diagnóstico precoce da doença de Parkinson, uma vez que analisando-se sinais de voz, é possível observar alterações em seus atributos que seriam impossíveis de identificar apenas pela audição.

Além disso, extraindo-se os principais atributos de um sinal de voz e efetuando-se uma análise estatística é possível determinar quais destes atributos possuem relação significativa com a doença, possibilitando um entendimento geral maior a seu respeito, o que pode impulsionar pesquisas no campo.

Desta forma, este trabalho apresenta o processo de construção de uma ferramenta para auxílio no processo de diagnóstico da doença de Parkinson que foi desenvolvida utilizando-se um algoritmo de Random Forest treinado e testado por meio de um conjunto de dados com atributos extraídos da fala de portadores e não portadores da doença de Parkinson.

Este trabalho se encontra dividido em 5 seções, sendo esta introdução a primeira delas. Na Seção 2, são apresentados os principais conceitos relacionados à doença de Parkinson, uso de sinais de voz na medicina, análise estatística e os fundamentos do Aprendizado de Máquina, que são fundamentais à compreensão deste trabalho.

Na Seção 3 são apresentados trabalhos de autores que utilizaram o mesmo conjunto de dados para construção de modelos semelhantes, fazendo uso de algoritmos de aprendizado de máquina como rede neural, máquina de vetores de suporte, regressão logística e árvore de decisão. Os resultados obtidos por cada modelo também são apresentados nessa Seção.

A Seção 4 apresenta o processo de seleção dos atributos utlizados no treinamento do modelo proposto, que foi feita através de uma análise estatística constituída de duas etapas: análise de correlação entre os atributos e análise de significância estatística dos atributos utilizando-se os valores p. Neste capítulo também é apresentado como os atributos selecionados foram utilizados no treinamento de um modelo que fez uso do algoritmo de Aprendizado de Máquina Random Forest.

Por fim, a Seção 5 traz as considerações finais sobre este trabalho, apresentando sua relevância, a validade dos resultados obtidos, bem como caminhos que possam levar a futuras melhorias.

2 Fundamentação Teórica

Esta Seção descreve os principais conceitos que motivaram e possibilitaram o desenvolvimento do trabalho aqui exposto. A Seção 2.1 apresenta informações gerais sobre a Doença de ParKinson (DP). A Seção 2.2 trata da aquisição e processamento dos sinais de voz, bem como sua relação com os mais variados distúrbios médicos. A Seção 2.3 e a Seção 2.4 apresentam respectivamente as principais ferramentas e técnicas estatísticas e de Aprendizado de Máquina relacionadas a construção de modelos preditivos nos quais este trabalho se baseia.

2.1 A Doença de Parkinson

A DP foi inicialmente denominada como paralisia agitante e era descrita como uma doença que provocava movimentos trêmulos involuntários, diminuição da força muscular, propensão para dobrar o tronco para frente e alteração no ritmo da marcha mesmo que o portador não apresentasse lesões nos sentidos e intelecto (Parkinson, 1817).

Anos mais tarde, o médico e pesquisador Jean-Martin Charcot utilizou o termo "Doença de Parkinson" em homenagem a James Parkinson que foi quem primeiro dissertou a respeito da condição clínica em 1817. O termo rapidamente se popularizou, tornando-se o nome mais frequentemente utilizado na designação da doença.

Atualmente, a DP é descrita como uma doença degenerativa e crônica que atua no sistema nervoso central envolvendo os gânglios da base, sendo causada pela deficiência da dopamina – neurotransmissor – na via nigroestriatal e cortical, interferindo principalmente no sistema motor (Steidl et al., 2007).

Os sintomas da DP dividem-se em duas categorias: motores, que também são chamados de parkinsonismo, e não motores. Segundo Massano (2011), o parkinsonismo se caracteriza principalmente pela acnésia ou bradicinésia, rigidez, tremos de repouso e alterações posturais e da marcha.

A acnésia se caracteriza não apenas pela diminuição progressiva da velocidade, mas também pela diminuição da amplitude do movimento. Outras manifestações possíveis são a hipominia (face inexpressiva ou imóvel), hipofonese (menor volume na fala) e a micrografia (caligrafia menor), sendo este último costumeiramente mais imperceptível.

A rigidez se caracteriza pela sensação de resistência ou oposição ao movimento, flexão ou extensão de um membro. A velocidade da movimentação não influencia negativamente ou positivamente a sensação de rigidez, mas ela aumenta com a ativação simultânea de um outro membro.

O tremor de repouso é o sintoma mais típico da DP (Massano and Cabreira, 2019). Trata-se do tremor que ocorre em membros que estejam relaxados e apoiados numa superfície de forma que não haja ação da gravidade que jus-

tifique a movimentação. O "tremor a contar moedas"é o mais usual e é caracterizado pala adução-abdução do polegar.

As alterações posturais e da marcha do parkinsionismo se caracterizam pelo andar lento e com passos curtos e pela postura fletida ou curvada de forma muito acentuada sem que esse comportamente se observe com o corpo em decúbito, o que a distingue das deformidades ósteo-articulares.

É importante destacar que o parkinsonismo pode ter causas para além da DP. O que distingue o parkinsonismo da DP é a apresentação de características próprias como o fato dos sintomas aparecerem e progredirem em apenas um dos lados do corpo nos anos iniciais da doença.

O parkinsonismo pode ser dividido em três categorias: primário ou idiopático, secundário e *plus*. O tipo mais comum é o idiopático, que abrange cerca de 75% dos casos e sendo ele a própria DP. O dois últimos tipos não possuem relação com a DP e podem ter origem por meio de infecções, medicamentos, acidentes, entre outras causas (Steidl et al., 2007).

Os sintomas não motores (SNM) podem ser peça chave no diagnóstico antecipado da DP, uma vez que eles podem surgir anos antes dos sintomas motores. Segundo Massano (2011), alguns dos principais SNM são a depressão, apatia, disfunção sexual, alterações no sono, ansiedade, fadiga, deteriorização cognitiva, alterações psicóticas, entre outros.

A DP é considerada cosmopolita, uma vez que ela não faz distinção entre os indivíduos, acometendo pessoas das mais diferentes classes sociais. Ocorre em homens e mulheres, principalmente, na faixa etária entre 55 e 65 anos, embora ocorra mais frequentemente em pessoas do sexo masculino (Steidl et al., 2007).

O Parkinson é a segunda doença neurodegenerativa mais frequente em todo o mundo, sendo superada apenas pela Doença de Alzheimer. Como a epidemologia exata da DP é de difícil determinação, não há consenso sobre a quantidade de casos. Estima-se que a incidência média mundial seja entre 15 e 20 casos por 100 mil habitantes por ano (Massano and Cabreira, 2019).

Na Europa, a incidência é entre 75 e 300 casos por 100 mil habitantes por ano, mas alguns estudos apontam que este número pode chegar a 12500 (Steidl et al., 2007). Já na América do Norte, mais de um milhão de pessoas têm a sua vida afetada pela DP (Little * et al., 2009). Estima-se 200 mil portadores da DP no Brasil (ROCHE, 2018).

A DP afeta principalmente os idosos. Estima-se que a doença acometa cerca de 1% da população mundial com mais de 65 anos (Massano and Cabreira, 2019). Com o aumento da expectativa de vida, cresce também o número de idosos de forma que é possível teorizar que a DP vá provocar grande impacto nas estruturas econômicas, sociais e de saúde em todo o mundo.

Esse impacto deve ser ainda mais grave no Brasil, uma vez que o país vem passando por um acelerado processo de envelhecimento populacional. Estima-se que até o a no de 2050, cerca de 19% da população brasileira terá idade superior aos 65 anos (Nasri, 2008). As mudanças na pirâmide populacional podem ser observadas nas Figs. 1 e 2.

O diagnóstico da DP é clínico. Ou seja, não há atualmente um teste ou biomarcador que garanta a presença da doença (Massano and Cabreira, 2019). A ausência de um exame definitivo é que leva às discrepância no número de doentes que encontramos na Europa, cuja incidência

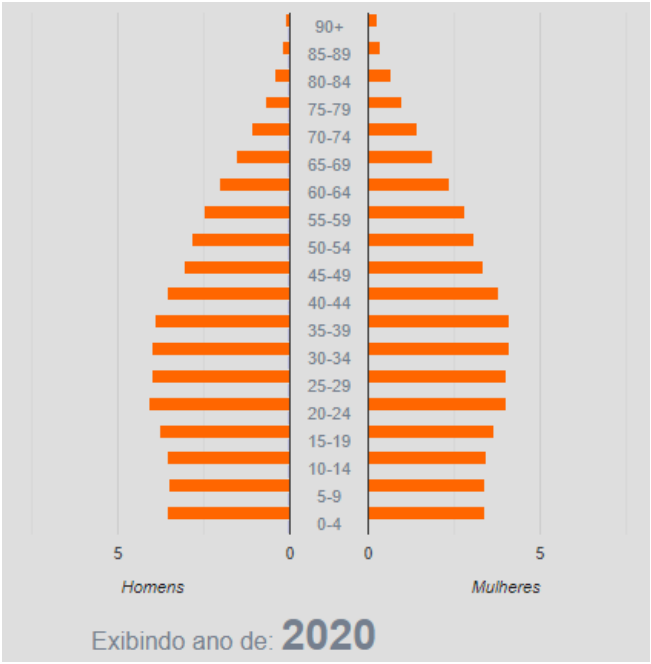


Figura 1: Projeção da pirâmide populacional no ano de 2020
Fonte: IBGE (2008)

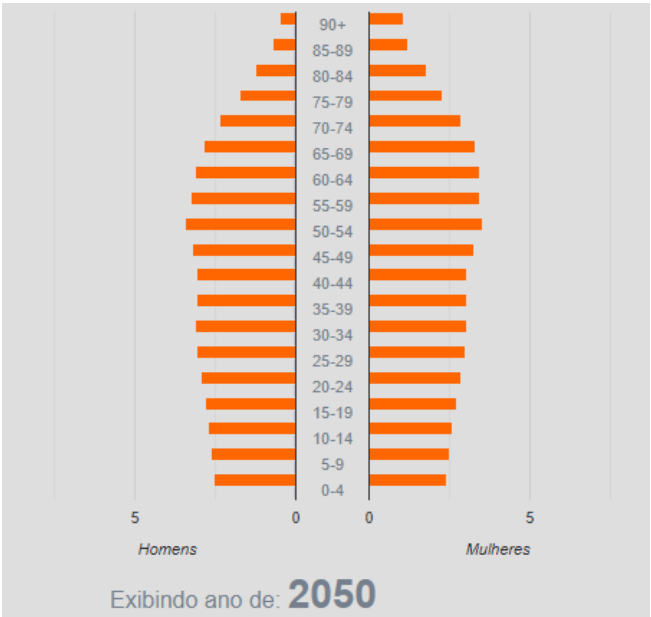


Figura 2: Projeção da pirâmide populacional no ano de 2050
Fonte: IBGE (2008)

Tabela 1: Classificação de Hoehn e Yahr (Steidl et al., 2007)

Estágio	Descrição
0	Nenhum sinal da doença.
1	Doença unilateral.
1.5	Envolvimento unilateral e axial.
2	Doença bilateral sem déficit de equilíbrio.
2.5	Doença bilateral leve, com recuperação no "teste do empurrão"
3	Doença bilateral moderada; alguma instabilidade postural; capacidade de viver independente.
4	Incapacidade grave, ainda capaz de caminhar ou permanecer de pé sem ajuda.
5	Confinado à cama ou cadeira de rodas a não ser que receba ajuda

média anual varia de 75 até 12500 pessoas por 100 mil habitantes.

A DP é diagnosticada por um neurologista através da exclusão de outras doenças. Após a descrição dos sintomas pelo paciente, o médico solicita uma série de exames (eletroencefalograma, tomografia computadorizada, ressonância magnética e análise do líquido espinhal) a fim de atestar a ausência de outras doenças no cérebro (Steidl et al., 2007).

Após a conclusão do diagnóstico, pode-se iniciar o tratamento. Atualmente, o tratamentos proposto levam em consideração o estágio da doença em que o paciente se encontra. Os estágios da DP são definidos de acordo com a escala Hoehn e Yahr criada em 1967 e estão apresentados na Tabela 1.

Dado que a DP é uma doença incurável e degenerativa, os tratamentos procuram atenuar os sintomas e retardar a sua progressão. Os tratamentos são inúmeros e variam de acordo com o estágio em que o paciente se encontram, bem como o sintoma que se deseja tratar.

Os medicamentos mais utilizados são os colinérgicos. Além disso, faz-se necessário o acompanhamento do paciente por parte de diversos profissionais (fisioterapeutas, psicólogos, fonoaudiólogos, nutricionistas e neurologistas) formando um tratamento multidisciplinar a fim de melhorar a qualidade de vida do portador da DP (Steidl et al., 2007).

Há ainda a possibilidade de um tratamento cirúrgico, conforma apresentado na Tabela 2. Independentemente das medidas terapêuticas clínicas ou cirúrgicas adotadas, a DP evolui progressivamente. Desta forma, o tratamento cirúrgico apenas proporciona uma melhora na qualidade de vida do pacientes com DP.

Tabela 2: Procedimentos Cirúrgicos e Disfunções (Steidl et al., 2007)

Procedimento	Condição
Talamotomia	Tremor
Campotomia de Forel	Tremor e Rigidez
Polidotomia	Acinésia e alterações axiais
Estimulação elétrica do Vim	Tremor contralateral
Estimulação palidal	Acinedia contralateral
Estimulação do NST	Alterações axiais, rigidez
Implante tecidual	Doentes Jovens

2.2 Análise de Sinais de Voz na Medicina

A análise de sinais de voz ou análise acústica vocal tem como principal objetivo quantificar e caracterizar um sinal sonoro proveniente da fala humana. Seu uso no Brasil e, mais especificamente, na clínica fonoaudiológica vem se intensificando desde o início do século XXI (de Carvalho Teles and Rosinha, 2008).

Inúmeras técnicas são utilizadas no campo médico a fim de atestar a qualidade vocal do paciente, sendo a forma mais usual a análise perceptivo-auditiva. Esta técnica pode levar a diferentes resultados, a depender da experiência do profissional envolvido. Este cenário impulsionou o desenvolvimento da análise acústica (Teixeira et al., 2013).

Características como rouquidão, soprosidade e rugosidade podem indicar a presença de patologias como nódulos vocais, laringite, cistos, endema de Reinke, câncer de laringe, entre outras. Essas características podem ser facilmente detectadas através da análise acústica vocal (Teixeira et al., 2013).

A análise acústica na medicina não se restringe apenas a doenças diretamente relacionadas à voz. Suas aplicações são diversas, uma vez que mudanças na fala de um indivíduo podem apresentar indícios de variados tipos de patologia, desde doenças respiratórias, como a pneumonia, até doenças psicológicas, como a depressão (Lenain et al., 2020).

Doenças neurodegenerativas também podem afetar a voz dos indivíduos. Em pesquisa realizada com 200 portadores da DP, verificou-se que 89% apresentavam algum problema vocal (Ho et al., 1999). Outras pesquisas mostram que até 90% dos portadores da DP apresentam alguma deficiência vocal (Little * et al., 2009).

Através do processamento digital de um sinal de voz, diversos atributos podem ser extraídos e analisados. Esses atributos podem ser utilizados no processo de diagnóstico de diversas doenças. Os atributos mais comuns são a frequência fundamental (F0), jitter, shimmer e a razão harmônica-ruído (HNR, do inglês harmonic-noise ratio).

A F0 de um sinal de voz é medida em Hertz (HZ) e é definida como a quantidade de vezes que a onda sonora produzida pelas cordas vocais se repete em um determinado período. Ela também representa o número de ciclos de abertura/fecho da glótis (abertura entra as cordas vocais responsável pela sonorização de vogais) (Teixeira et al., 2013).

Medidas de distúrbio na F0, jitter e shimmer, se mostram úteis na descrição de características vocais. Jitter é definido como a variação de frequência de ciclo para ciclo.

Tabela 3: Atributos de jitter (Teixeira et al., 2013)

Atributo	Definição
Jitta (μs)	Representa a diferença média absoluta entre 2 períodos consecutivos.
Jitt (%)	Mesmo que jitta dividida pelo período médio.
RAP (%)	Distúrbio médio entre 3 períodos consecutivos divido pelo período médio.
PPQ5 (%)	Distúrbio médio entre 5 períodos consecutivos divido pelo período médio.

Tabela 4: Atributos de shimmer (Teixeira et al., 2013)

Atributo	Definição
Shim (%)	Diferença média de amplitude entre 2 períodos consecutivos dividida pela amplitude média.
ShdB (dB)	Mesmo que shim em dB.
APQ3 (%)	Distúrbio médio de amplitude entre 3 períodos consecutivos divido pela amplitude média.
APQ5 (%)	Distúrbio médio de amplitude entre 5 períodos consecutivos divido pela amplitude média.

Shimmer está relacionado com a variação na amplitude da onda sonora (Teixeira et al., 2013). As Tabelas 3 e 4 apresentam os principais atributos de jitter e shimmer.

O HNR é a razão entre componentes periódicas e não periódicas de um segmento de sinal de voz. A primeira componente representa a vibração das cordas vocais e a segunda representa o ruído glotal (Teixeira et al., 2013). Ela é expressa em dB e seu oposto, a relação ruído-harmônica (NHR) também é uma medida amplamente utilizada.

A relação entre os atributos de jitter, shimmer e HNR com a presença/ausência de determinadas patologias é amplamente estudada. A forma mais usual de extração desses atributos é através da sustentação vogal (Teixeira et al., 2013). Os valores de limiar para vozes patológicas são apresentado na Tabela 5.

Tabela 5: Limiares para vozes patológicas (Teixeira et al., 2013)

Atributos	Valor Limiar
Jitt (%)	1.04
Jitta (μs	83.2
RAP (%)	0.68
Shim (%)	3.81
ShdB (dB)	0.35
HNR	7

Pesquisas recentes apontam que a produção vocal pode ser tratado como um sistema não-linear dinâmico. Desta forma, além das medidas tradicionais mais utilizadas, novas medidas como a entropia da densidade de probabilidade do período de recorrência (RPDE) e a análise de flutuação sem tendência (DFA) apresentam resultados relevantes (Little * et al., 2009).

A DFA pode ser utilizada para se caracterizar a autosimilaridade de um sinal. No caso de ondas sonoras, a DFA pode obter informações a respeito de vibrações vocais de baixa frequência que estão relacionadas a ruídos respiratórios que podem ser analisados em busca de patologias (Little et al., 2007).

A RDPE calcula a incerteza média de um dado valor de uma densidade de probabilidade discreta. Uma vez que um sinal de voz processado digitalmente pode ser tratado como uma densidade de probabilidade, a RDPE pode ser utilizada na avaliação de distúrbios vocais, representando a incerteza média do período do sinal (Little et al., 2007).

Outra medida que vem sendo utilizada na detecção de patologias, especialmente doenças neurodegenerativas, é a entropia do período do tom (PPE). Um sintoma comum de portadores da DP é a dificuldade em controlar o tom de uma voz estacionária (como na sustentação vogal) (Little * et al., 2009).

Assim como a RDPE, a PPE também representa uma entropia, isto é, a quantidade de incerteza média de uma densidade probabilidade. Na RDPE, se obtem a entropia da densidade de probabilidade do período de recorrência. Já na PPE, a entropia calculada é a da densidade de probabilidade de período do tom (percepção sensorial da F0) (Ozkan, 2016).

2.3 Inferência Estatística

A área da Saúde necessita de métodos matemáticos e estatísticos que permitam afirmar ou negar, a partir de evidências e dados clínicos, a influência ou não de determinado sintoma no desenvolvimento ou aparecimento de uma doença (da Silva et al., 2006). Nesse contexto, uma das principais ferramentas utilizadas é a inferência estatística.

Inferência estatística pode ser definida como o processo responsável pela geração de conclusões a respeito de uma população partindo de uma amostra aleatória dela. Sem a inferência estatística, dados não teriam qualquer utilidade, com ela é possível gerar novos conhecimentos a respeito de determinado fenômeno (Caffo, 2016).

A inferência estatística serve como descrição parsimoniosa (o modelo razoável mais simples que explica um determinado fenômeno) do mundo. Ela faz uso de diversos conceitos estatísticos, sendo os principais deles os de valor esperado, variância e desvio padrão, intervalos de confiança e teste de hipóteses.

O valor esperado ou média da população de uma variável aleatória é o centro de sua distribuição. Para uma variável discreto X com função massa de probabilidade (PMF, do inglês probability mass function) $p(x)$ o valor esperado é definido como:

$$E[X] = \sum_x xp(x) \tag{1}$$

onde o somatório é realizado por todos os valores pos-

síveis de x. O que não ocorre com a média amostral, que é calculada em cima dos valores amostrados. Desta forma, a média amostral é uma estimativa da média populacional e pode ser definida como:

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i) \tag{2}$$

Enquanto que o valor esperado de uma distribuição pode ser entendido de forma análoga ao centro de massa de um objeto, a variância representa o quão "espalhada"essa distribuição se encontra (Caffo, 2016). A variância populacional é definida como:

$$Var(X) = E[X^2] - E[X]^2 \tag{3}$$

Assim como a média da amostra representa uma estimativa da média populacional, a variância da amostra representa uma estimativa da variância populacional. A variância amostral é definida como:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{4}$$

A distribuição normal é a mais utilizada em toda a estatística. Ela pode ser utilizada na descrição de diversos fenômenos e segue o princípio da modelagem parcimoniosa, uma vez que para ser definida, faz-se necessário o conhecimento de apenas duas informações sobre uma população: a média e a variância.

Seja X uma variável aleatória cuja distribuição seja normal com $E[X] = \mu$ e $Var(X) = \sigma^2$, a notação utilizada para descrevê-la é $X \sim N(\mu, \sigma^2)$ e sua função de densidade associada é dada por:

$$(2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{x-\mu}{2\sigma^2}} \tag{5}$$

A distribuição normal é utilizada na matemática e outras ciências exatas, ciência humanas, sociais e biológicas. Algumas de suas aplicações incluem a balística, o quociente de inteligência, a mensuração da altura humana e do tamanho de bicos de aves, preços de determinadas commodities, entre outras.

O Teorema do Limite Central (CLT, do inglês central limit theorem) é um dos mais importantes na estatística. Segundo ele, a distribuição de médias de variáveis se torna a distribuição normal à medida que a quantidade de amostras cresce (Caffo, 2016). Ou seja, dada uma distribuição de n médias com $E[X] = \mu$ e $Var(X) = \sigma^2$, temos que $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

O CLT permite a criação de intervalos de confiança para estimativas. Os intervalos de confiança são métodos para quantificar a incerteza de uma estimativa. O intervalo de confiança mais usualmente adotado é o de 95% e pode ser obtido da seguinte forma:

$$\mu \pm \frac{2\sigma}{\sqrt{n}} \tag{6}$$

Caso o número de amostrar seja reduzido, pode-se utilizar a distribuição t na construção do intervalo de confiança. A expressão abaixo apresenta a construção deste intervalo, onde t_{n-1} representa o quantil relevante da distribuição t.

$$\bar{X} \pm \frac{t_{n-1}S}{\sqrt{n}} \tag{7}$$

O teste de hipótese está relacionado com a tomada de decisão utilizando dados. Em geral, o teste é realizado sobre duas possíveis decisões. A primeira é chamada de hipótese nula (H_0) e representa o status quo. A segunda é costumeiramente chamada de hipótese alternativa (H_a).

O teste de hipótese está intimamente associado com os intervalos de confiança. Inicialmente, a H_0 é assumida como verdadeira. Em seguida, escolhe-se a distribuição adequada, para a qual H_0 representará o centro, e verifica-se se a estatística encontrada ou não dentro do intervalo de confiança para aquela distribuição.

Caso a estatística encontrada se encontre dentro do intervalo, a H_a é rejeita. Caso contrário, entende-se que ou um fenômeno muito raro ocorreu, ou a H_0 é falsa. Neste caso, a H_0 é rejeitada e consequentemente a H_a é aceita. O tamanho do intervalo de confiança escolhido varia de acordo com o fenômeno estudado.

Os valores-p são a medida de significância estatística mais comumente utilizados. A sua ideia central é assumir que a H_0 é verdadeira e calcular o quão incomum ou extremo seria a maneira que os dados observados se apresentam (na forma de um teste estatístico), rejeitando ou aceitando H_a.

Os valores-p são interpretados da seguinte forma: Se o valor-p for pequeno, ou um evento extremamente raro foi observado ou H_0 é falsa. O que pode ser considerado um valor-p pequeno vai variar de acordo com o tamanho do intervalo de confiança escolhido. Para um intervalo de 95%, valores-p menos que 0.05 rejeitam H_0.

2.4 Fundamentos de Aprendizado de Máquina

Aprendizado de Máquina (ML, do inglês machine learning) pode ser definida como um conjunto de métodos computacionais que fazem uso da experiência para melhorar sua performance e/ou realizar previsões mais precisas. Entenda-se experiência como a capacidade de utilizar informações antepassadas para seu aprendizado (Mohri et al., 2018).

Usualmente, ML tem como objetivo prever uma determinada medida de resultado, frequentemente quantitativa ou categórica, baseada num conjunto de dados conhecidos, comumente chamados de atributos, que tenham alguma relação com esta medida (Hastie et al., 2008).

O aprendizado é denominado como supervisionado quando a medida de resultado é conhecida para um conjunto de instâncias com seus atributos associados. Caso contrário, o aprendizado é dito não supervisionado. A

Tabela 6: Categoria de problemas de ML (Mohri et al., 2018)

Categoria	Definição
Classificação	tipo de problema onde deseja-se atribuir uma categoria para cada instância.
Regressão	Tipo de problema onde deseja-se prever um valor real para cada instância.
Ranking	Tipo de problema onde deseja-se hierarquizar as instâncias de seguindo determinado critério.
Agrupamento	Tipo de problema onde deseja-se particionar um conjunto de itens em subconjuntos homogêneos.
Redução de dimensão	Tipo de problema onde partindo-se de uma representação inicial das instâncias, deseja-se obter uma representação com um número de dimensões menor que ainda preserve algumas propriedades da representação original.

Tabela 6 apresenta outras divisões para os problemas de Aprendizado de Máquina.

Partindo da hipótese de que os erros ϵ sejam aditivos e que o modelo $Y = f(X) + \epsilon$ seja razoável, aprendizado supervisionado tem como objetivo aprender a f através de um "professor". a performance de um modelo será tão melhor quanto forem o número de professores (Hastie et al., 2008).

O processo de treino, validação e teste de um modelo de ML envolve a divisão da base de dados em três subconjuntos distintos e sem sobreposição de elementos entre si. Na amostra de treino é feito o processo de aprendizagem do modelo através do exemplo.

O desempenho do modelo pode então ser avaliado preliminarmente na amostra de validação, onde podem ser definidas alterações no processo de treino de forma a melhorar sua performance. A performance final é avaliada com a amostra de teste e o modelo não pode sofrer novas alterações (Mohri et al., 2018).

A acurácia é a medida mais frequentemente utilizada na avaliação de modelos de ML. Dado que todas as instâncias do conjunto de dados tenham o mesmo peso, a acurácia de um algoritmo de classificação é definida como o número de previsões realizadas corretamente dividido pelo número total de instâncias (Wong, 2015).

Outras medidas que podem ser utilizadas na avaliação desses modelos, em especial nos de diagnóstico de patologias, são a sensibilidade, probabilidade do dignóstico positivo dado que o paciente possui a doença, e a especificidade, probabilidade do dignóstico negativo dado que o paciente não possui a doença (Caffo, 2016).

A sensibilidade e especificidade estão intimamente associados com o Lei de Bayes, que é amplamente utilizada na avaliação de teste de diagnóstico. Seja $P(D)$ a prevalência de uma doença, $P(+|D)$ a sensibilidade e $P(-|D^c)$ a especificidade, a Lei de Bayes diz que:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + [1 - P(-|D^C)][1 - P(D)]} \quad (8)$$

Ou seja, com a sensibilidade e especificidade do teste e a prevalência de uma doença, é possível obter a probabilidade de que o paciente possua determinada patologia dado que seu teste indicou a presença desta patologia. Importante observar que $P(D|+)$ difere da acurácia, dado que a acurácia trata dos acertos do modelo para casos conhecidos.

Em problema de classificação, é comum que as classes do conjunto de dados não sejam balanceadas. Nestas situações, a acurácia pode ser uma medida de avaliação ilusória. Um modelo que sempre indicasse um rótulo A em uma base em que 90% das intâncias tivessem este mesmo rótulo, teria uma acurácia de 90%, por exemplo.

A estatística kappa (κ) é uma medida de avaliação que pode ser útil nesses casos de desbalanceamento de classes. A fórmula a seguir apresenta seu cálculo, onde ρ_o representa a concordância observada e ρ_e representa a concordância esperada.

$$\kappa = \frac{\rho_o - \rho_e}{1 - \rho_e} \quad (9)$$

O valor de κ é sempre inferior a 1. Um valor negativo indicaria que o classificador avaliado é inútil. Não há uma forma padronizada de se interpretar o valor de κ. Quanto mais próximo de 1, melhor. Landis e Koch propõem a interpretação dos valores de κ obtidos de acordo com a Tabela 7.

Tabela 7: Valor de Kappa e nível de concordância (Landis and Koch, 1977)

Valor de Kappa	Nível de concordância
< 0	Sem Condorância
0 - 0.20	Concordância Mínima
0.21 - 0.40	Concordância Razoável
0.41 - 0.60	Concordância Moderada
0.61 - 0.80	Concordância Substâncial
0.81 - 1.00	Concordância Quase Perfeita

Com um número reduzido de dados, dividir o conjunto de dados em 3 subconjuntos (treino, validação e teste) pode dificultar o aprendizado do modelo de ML, uma vez que o número de instâncias disponíveis para treino seria diminuído. Para contornar este problema, pode-se utilizar a validação cruzada k-fold.

A k-fold consiste na divisão do conjunto de dados em k subconjuntos. O treino é efetuado em k-1 subconjuntos e o teste é feito no subconjunto que sobrar. O processo é repetido k vezes até que todos os subconjuntos tenham sido usados como teste. Ao fim do processo, o modelo é treinado com todos os dados disponíveis (Hastie et al., 2008).

A vantagem deste método de validação cruzada está na utilização de toda a base de dados para treino. A avaliação da performance do modelo é feita calculando a média das

k acurácias, ou outra medida que se utilize, colhidas na etapa de teste (Wong, 2015). Importante ressaltar que os k subconjuntos possuem tamanho aproximado e não são sobrepostos.

A quantidade k de subconjuntos varia de acordo com o problema. Um k grande leva a um aumento na variância, e consequentemente, na incerteza da acurácia calculada. Um k pequeno diminui a variância, mas pode levar a resultados enviesados, ou seja, a acurácia colhida pode não ser realista (Hastie et al., 2008).

Os valores de k mais utilizados são entre 5 e 10. Quando k é idêntico a quantidade de instâncias do conjuntos de dados, temos um caso especial de validação cruzada k-fold que passa a ser chamada de validação cruzada leave-one-out (deixe uma de fora) (Hastie et al., 2008).

O intervalo de confiança da acurácia pode ser calculado partindo-se da hipótese de que ela siga uma distribuição normal. Para que esta hipótese seja válida, é necessário que o conjunto de dados de teste tenha pelo menos 30 instâncias e que a quantidade de erros e acertos seja pelo menos 5 (Wong, 2015).

3 Trabalhos Relacionados

Esta Seção descreve a construção do conjunto de dados utilizado no trabalho aqui exposto, bem como os principais trabalhos que o utilizaram no desenvolvimento de seus modelos. A Subseção 3.1 apresenta os resultados de duas revisões sistemáticas sobre o tema. A Subseção 3.2 apresenta os equipamentos e métodos usados por Little et al. (2009) na confecção do conjunto de dados. A Subseção 3.3 apresenta o modelo e os resultados obtidos por Little et al. (2009). As Subseções 3.4, 3.5, 3.6 e 3.7 apresentam respectivamente os modelos e resultados obtidos por Bhattacharya e bhatia (2010), Das (2010), Sakar e Kursun (2009) e Govindu e Palwe (2023). A Subseção 3.8 faz um apanhado geral sobre os resultados observados. Por fim, a Subseção 3.9 traz os resultados de trabalhos que fizeram uso de outras bases.

3.1 Estado da Arte

A quantidade de trabalhos que fazem uso de sinais e dados provenientes da fala como indicativo da presença da DP vêm crescendo de forma significativa nos últimos 15 anos. No estudo realizado por Ngo et al. (2022) entre os anos de 2010 e 2021, cerca de 838 artigos relacionados à temática foram encontrados. Destes, 189 foram selecionados e 147 foram considerados aptos à participarem da revisão sistemática. Foi constatado que a fala e a voz se apresentam como biomarcadores de relevância no estudo da DP.

Focando mais especificamente no uso de métodos de aprendizado de máquina e inferência estatística associados com características vocais, Amato et al. (2023) constatou, através de uma revisão que investigou 102 trabalhos entre os anos de 2017 e 2022, que os atributos mais predominanetemente utilizados são jitter, shimmer, HNR, F0 e DFA.

3.2 Construção do Conjunto de Dados

Little et al. (2009) desenvolveu um conjunto de dados através de 195 fonações de vogais sustentadas colhidas de 31 pacientes, homens e mulheres, dos quais 23 foram diagnosticados como portadores da DP, com faixa etária entre 46 e 85 anos (média 65.8, desvio padrão 9.8). Cada paciente gravou, em média, 6 áudios (alguns chegaram a 7 gravações).

As fonações tinham duração entre 1 e 36 segundos e foram gravadas em uma cabine acústica utilizando-se um microfone (AKG C420) posicionado a 8 centímetros dos lábios do paciente. Os sinais de voz foram gravados com um hardware CSL 4300B, amostrados a 44.1 kHz, com 16 bits de resolução, tendo sido normalizados em amplitude.

Os atributos extraídos dos sinais de voz são apresentados na Tabela 8. As medidas mais tradicionais (F0, shimmer, jitter, entre outras) foram computadas utilizando-se o software Praat. Já as medidas não tradicionais (RDPE, DFA, PPE, entre outras) foram extraídas a partir de algoritmos implementados pelos autores.

3.3 O Modelo de Little et. al. (2009)

delo desenvolvido por Little et al. (2009) utilizou um classificador com base em uma máquina de vetores de suporte (SVM) e foi iniciado com a filtragem de alguns atributos utilizando-se uma análise de correlação. Para cada par de atributos, aqueles que tivessem um coeficiente de correlação superior a 0.95, um dos atributos seria removido do conjunto de dados.

Outro atributo descartado foi a F0, argumentando-se que embora tenha sido demonstrado que há uma relação estatística entre os valores absolutos de F0 e a presença da DP, optou-se por não utilizar essa medida, uma vez que ela era fortemente influenciada pelo gênero do paciente.

Após o processo de filtragem, sobraram apenas 9 atributos do conjunto de dados, conforme apresentado na Tabela 9. Todos os 1023 subconjuntos viáveis foram utilizados no treinamento e teste de um modelo de SVM. Os melhores resultados obtidos, com um intervalo de confiança de 95% são apresentados na Tabela 10.

3.4 O Modelo de Battacharya e Bathia (2010)

Utilizando o mesmo conjunto de dados, Bhattacharya and Bhatia (2010) desenvolveu outro modelo também utilizando a SVM. O modelo foi construído fazendo-se uso do software Weka. Através dele, foi efetuada a filtragem dos atributos, o treinamento e o teste do modelo proposto.

Os atributos que passaram pela filtragem resultaram no mesmo subconjunto do modelo anterior. O processo de validação foi utilizando o método k-fold, com k=3. O melhor resultado de acurácia obtido foi de 95.3% para o conjunto de treino e 60.9% para o conjunto de teste.

3.5 O Modelo de Das (2010)

Utilizando o SAS base (software que integra ferramentas de filtragem e mineração de dados, bem como métodos de ML) Das (2010) construiu 4 modelos distintos, cada um

Tabela 8: Descrição dos atributos do conjunto de dados de Parkinson (Sakar and Kursun, 2009)

Descrição	Identificador	Mínimo	Máximo	Média	Desvio Padrão
F0 médio	Fo(Hz)	88.33	260.11	154.23	41.39
F0 máximo	Fhi(Hz)	102.15	592.03	197.11	91.50
F0 mínimo	Flo(Hz)	65.48	239.17	116.33	43.52
Medidas de Jitter	Jitter(%)	0.002	0.033	0.006	0.005
	Jitter(Abs)	7E-06	26E-05	4.4E-05	3.48E-05
	RAP	0.001	0.021	0.003	0.003
	PPQ	0.001	0.020	0.003	0.003
	Jitter:DDP	0.002	0.064	0.010	0.009
Medidas de Shimmer	Shimmer	0.01	0.119	0.03	0.019
	Shimmer(dB)	0.085	1.302	0.282	0.195
	Shimmer:APQ3	0.005	0.056	0.016	0.010
	Shimmer:APQ5	0.006	0.079	0.018	0.012
	Shimmer:APQ	0.007	0.138	0.024	0.017
	Shimmer:DDA	0.014	0.169	0.047	0.030
Razões harmônico/ruído	HNR	8.441	33.047	21.886	4.426
	NHR	0.001	0.315	0.024	0.017
Medidas de complexidade	RPDE	0.257	0.685	0.499	0.104
dinâmicas não lineares	D2	1.423	3.671	2.382	0.383
Medidas não lineares	Spread1	-7.965	-2.434	-5.684	1.090
de variação de F0	Spread2	0.006	0.450	0.227	0.083
	PPE	0.045	0.527	0.207	0.090
Análise de flutuação	DFA	0.574	0.825	0.718	0.055
sem tendência (expoente)					

Tabela 9: Atributos mantidos após filtragem (Little * et al., 2009)

Atributo	Descrição
Jitter(Abs)	Jitter em percentual
Jitter:DDP	Diferença absoluta entre os ciclos dividido pela período médio
Shimmer:APQ	Quociente de perturbação de amplitude
Shimmer:DDA	Média das diferenças absolutas de diferenças de amplitudes conscutivas
NHR	Razão ruído/harmônico
HNR	Razão harmônico/ruído
RPDE	Densidade de entropia do período de recorrência
DFA	Análise de flutuação sem tendência
D2	Correlação dimensional
PPE	Entropia do período do tom

Tabela 10: Resultados da performance de classificação do SVM (Little * et al., 2009)

Atributos	Acurácia	Verdadeiro positivo	Verdadeiro negativo
HNR, RPDE, DFA, PPE (4)	91.4 ± 4.4	91.1 ± 4.9	92.3 ± 7.0
Todos (10)	90.6 ± 4.1	90.7 ± 4.3	89.1 ± 8.6
RPDE, DFA, PPE (10)	89.5 ± 3.9	89.6 ± 4.3	89.1 ± 8.6
DFA, PPE (2)	88.2 ± 3.8	88.2 ± 4.2	88.0 ± 8.1
PPE (1)	85.6 ± 5.4	85.9 ± 5.5	84.5 ± 10.8
Jitter(Abs) (1)	80.6 ± 9.9	80.7 ± 10.1	80.3 ± 10.9
RPDE, DFA (2)	79.2 ± 4.2	79.2 ± 4.5	79.0 ± 7.5
HNR (1)	77.4 ± 2.8	77.6 ± 3.1	76.9 ± 4.1
Shimmer:APQ (1)	76.7 ± 4.1	76.8 ± 4.3	76.2 ± 6.5

fazendo uso de um método de ML diferente: Rede neural, DMNeural, Regressão logística e árvore de decisão.

A validação dos modelos foi feita utilizando-se o método de dividir o conjunto de dados em um de treino e outro para teste, sem sobreposição de instâncias. A Tabela 11 apresenta os melhores resultados obtidos de acurácia nos conjuntos de treino e teste para cada modelo utilizado.

3.6 O Modelo de Sakar e Kursun (2009)

Sakar and Kursun (2009) também desenvolveram um modelo com base em uma SVM que fez uso de um processo de discretização dos atributos que foram filtrados utilizando a abordagem Máxima Relevância - Mínima Redundância (mRMR) e utilizou um processo de validação chamado pelos autores de leave-one-individual-out.

Os atributos foram discretizados em 9 níveis. Para isso, utilizou-se a respectiva média μ e desvio padrão σ onde

Tabela 11: Resultados da performance de classificação de múltiplos modelos (Das, 2010)

Método	Acurácia no conjunto de treino	Acurácia no conjunto de teste
Rede neural	100%	92.9%
DMNeural	89.6%	84.4%
Regressão logística	89%	88.6%
Árvore de decisão	93.6%	84.3%

os valores entre $\mu - \sigma$ e $\mu + \sigma$ são convertidos em 0. Os 4 intervalos de tamanho σ a direita de 0 foram convertidos nos níveis discretos de 1 a 4 e, de forma análoga, os da esquerda receberam níveis discretos de -1 a -4.

O método mRMR se baseia no fato de que atributos individualmente bons não necessariamente levam a uma melhor performance de classificação. O método consiste em selecionar os atributos mais relevantes, evitando aquelas que sejam redundantes, maximizando a dependência conjunta.

O método de validação leave-one-individual-out consiste em deixar todas as instâncias relacionadas separadas. No caso do conjunto de dados de Parkinson, todas as 6 ou 7 instâncias obtidas a do sinal de voz de um dos 32 pacientes seriam separadas no processo de treino e teste.

Ele difere assim do tradicional método leave-one-out que consiste em deixar uma única instância de fora no processo de treino e teste. Argumenta-se que, desta forma, as amostras de treino e teste seriam realmente independentes e a estimativa da acurácia e demais estatísticas seriam mais fidedignas.

O melhor resultado obtido para a acurácia foi de 92.75% com um intervalo de confiança de 1.21%. Este resultado foi obtido com um subconjunto de treino contendo os atributos spread1, Fo(Hz), Shimmer:APQ3 e D2 que têm seus significados apresentados na Tabela 9.

3.7 O Modelo de Govindu and Palwe (2023)

Govindu and Palwe (2023) utilizaram 3 diferentes abordagens, cada uma com 4 distintos modelos de aprendizado de máquina: SVM, regressão logística, random forest e k-vizinhos mais próximos (KNN, do inglês K-nearest neighbors). A primeira abordagem utilizava todos os 22 atributos do conjunto de dados no treimaneto dos modelos. A segunda abordagem utilizava uma análise de componentes principais (PCA, do inglês principal component analysis) para reduzir o número de atributos a serem usados no treinamento dos modelos para 5. A última abordagem tratou de balancear a base, que contém 109 amostras de pacientes portadores da PD e apenas 40 de pessoas saudáveis. O balanceamento foi feito atráves da reamostragem das amostras de não portadores da DP atém que a quantidade de registros fosse igualada. Os modelos foram em seguida treinados com todos os atributos disponíveis. Para todas as abordagens, foi feito o escalonamento dos dados usando a técnica do desvio padrão. A validação foi feita dividindo-se os dados em conjuntos de treinos e testes, mantendo sempre 75% dos dados nos conjuntos de treino. Para cada abordagem, foram colhidas as métricas de acurácia, precisão e sensibilidade. A Tabela 12 apresenta os resultados obtidos pela primeira abordagem. A Tabela 13 apresenta os resultados obtidos pela segunda abordagem e a Tabela 14 traz os resultados da última abordagem.

Tabela 12: Resultados da performance de classificação da primeira abordagem (Govindu and Palwe, 2023)

Método	Acurácia	Precisão	Sensibilidade
Regressão logística	83.67%	100%	83%
Random forest	91.83%	95%	86%
SVM	85.71%	100%	84%
KNN	85.71%	95%	86%

Tabela 13: Resultados da performance de classificação da segunda abordagem (Govindu and Palwe, 2023)

Método	Acurácia	Precisão	Sensibilidade
Regressão logística	83.67%	100%	83%
Random forest	83.67%	100%	90%
SVM	91.75%	100%	86%
KNN	83.67%	92%	90%

Tabela 14: Resultados da performance de classificação da terceira abordagem (Govindu and Palwe, 2023)

Método	Acurácia	Precisão	Sensibilidade
Regressão logística	85.71%	89%	92%
Random forest	85.71%	89%	92%
SVM	81.63%	82%	94%
KNN	91.83%	95%	95%

3.8 Resultados

A Tabela 15 apresenta os principais resultados obtidos pelas pesquisas supracitadas. O modelo de melhor resultado foi o de Das (2010) que, utilizando uma rede neural, alcançou uma acurária de 92.9. Contudo, o modelo de Sakar and Kursun (2009) alcançou um resultado bem próximo, chegando a uma acurácia de 92.8.

O modelo de Sakar and Kursun (2009) ainda tem como vantagem perando os outros resultados o fato de ter obtido o menor intervalo de confiança, cerca de 1.2, o que indica que a acurária real do modelo é bem próxima a estimada. O modelo de Das (2010) não teve o intervalo de confiança informado.

Além disso, considerando-se o intervalo de confiança informado, ambos os modelos possuem acurácia na mesma faixa, o que indica que os resultados obtidos entre eles é praticamente idêntico. O mesmo não pode ser dito

do modelo de Little * et al. (2009), cuja acurácia de 91.4 se encontra fora do intervalo de confiança, havendo assim, evidencia estatística de que o resultado obtido por Sakar and Kursun (2009) é de fato superior.

Tabela 15: Comparativo com outros modelos.

Modelo	Acurácia
Rede neural (Das, 2010)	92.9
SVM (Sakar; Kursun, 2009)	92.8 ± 1.2
Random forest (Govindu; Palwe, 2023)	91.83
SVM (Little et al., 2009)	91.4 ± 4.4
SVM (Bhattacharya; Bhatia, 2010)	60.9

Analisando-se os atributos utilizados nos treinamentos dos modelos que obtiveram os melhores resultados, constata-se que os atributos não lineares se mostram mais promissores que os atributos de jitter, shimmer e frequência fundamental na detecção da DP.

3.9 Outros Trabalhos

Almeida et al. (2019) construiu uma base com atributos provenientes de fonações sustentadas por 5 segundos da vogal "a". Cada fonação era repetida 3 vezes e os áudios foram captados por meio de um cardióide acústico (AKG Perception 220) e um smartphone com microfone interno (Samsung Galaxy Note 3) posicionado a cerca de 10 cm da boca dos pacientes. O formato de áudio utilizado foi mono PCM wav (16 bits e 44.1 kHz de taxa de amostragem). Os indivíduos participantes da pesquisa eram homens e mulheres, portadores e não portadores da DP, somando-se 99 participantes. A média de idade do grupo de controle era de 41.8 anos, enquanto que a média do grupo de portadores da DP era de 61.5 anos. Foram treinados modelos de KNN e SVM utilizando-se ferramentas como o Praat e OpenCV. A validação foi feita dividindo-se o conjunto de dados em um conjunto de treino e outro de teste. O melhor resultado obtido foi de 94.55% de acurácia com o modelo de KNN.

Karaman et al. (2021) e Tai et al. (2021) utilizaram a mesma base composta de mais de 65 mil áudios de fonações sustentadas da vogal "a"desenvolvida pelo projeto mPower no ano de 2015. Os atributos extraídos por Tai et al. (2021) podem estão listados na Tabela 16. Esses atributos foram filtrados através de uma análise de alta correlação e de uma PCA. Em seguida eles foram utilizados no treinamento de 4 modelos de aprendizado de máquinas: rede neural, random forest, SVM e regressão logística. A validação foi feita dividindo-se os dados em conjuntos de treino e teste e os resultados de acurácia são apresentados na Tabela 17.

Já Karaman et al. (2021) utilizou uma transformada de cosseno discreta para gerar espectrogramas que em seguida foram utilizados no treinamento de uma rede neural convolucional. O melhor resultado obtido foi de uma acurácia de 89.75%, superior a todos os resultados obtidos por Tai et al. (2021). Esses resultados não podem ser comparados com os resultados apresentados por Almeida et al. (2019) nem com os resultados apresentados pelos trabalhos apresentados nas Subseções 3.4, 3.5, 3.6 e 3.7, pois fazem uso de bases diferentes, com diferentes atributos.

Tabela 16: Atributos do modelo de Tai et al. (2021)

Id	Atributo
1	HNR
2	apq11Shimmer
3	apq3Shimmer
4	apq5Shimmer
5	ddaShimmer
6	ddpJitter
7-19	desvMFCC
20-32	desvMFCCdelta
33	localJitter
34	localShimmer
35	localabsoluteJitter
36	localdbShimmer
37	max pitch
38	meanF0Hz
39-51	meanMFCC
52-64	meanMFCCdelta
65	min pitch
66	n periods
67	n pulses
68	ppq5Jitter
69	rapJitter
70	stdevF0Hz
71	Status

Tabela 17: Resultados do modelo de Tai et al. (2021)

Modelo	Filtro de alta correção	PCA
Rede neural	83%	86%
Random forest	75%	82%
SVM	88%	87%
Regressão logística	80%	71%

Pelo mesmo motivo, esses resultados também não podem ser comparados com os do modelo apresentado na Seção 4.

4 Modelo Proposto

Esta Seção descreve como o modelo proposto foi desenvolvido, bem como a aplicação responsável pela extração dos atributos de um sinal de voz. A Subseção 4.1 apresenta o processo de seleção de atributos. A Subseção 4.2 apresenta o treinamento e resultados do modelo.

4.1 Seleção dos Atributos

Fazendo uso do conjunto de dados elaborado por Little * et al. (2009), o desenvolvimento do modelo proposto foi iniciado com a seleção dos atributos a serem utilizados no processo de treino. A seleção pode ser dividida em 3 etapas. A primeira delas consiste numa análise exploratória, a fim de se observar visualmente a dispersão, média e presença de instâncias discrepantes por status (saudável ou portador da DP) em cada atributo.

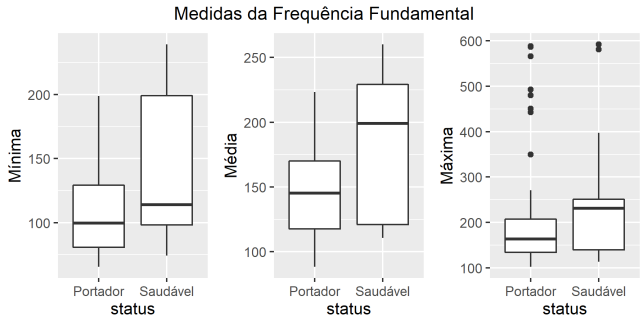
A segunda etapa consiste no uso de testes de hipóteses utilizando a distribuição t com um intervalo de confiança de 99%, a fim de filtrar os atributos que não sejam estatisticamente relevantes para a detecção do status do paciente.

Por fim, é feita uma análise de correlação dos atributos. Isto é feito para garantir que as bases de treino não possuam atributos com forte correlação, o que levaria a uma inflação no intervalo de confiaça estimado para a acurácia do modelo.

A forma de seleção dos atributos difere das observadas em pesquisas precedentes por adotar um critério de filtragem mais rígido, utilizando um intervalo de confiança de 99% ao invés do tradicional 95%. Além disso, as bases de treino foram separadas de forma a diminuir a correlação total entre os seus atributos.

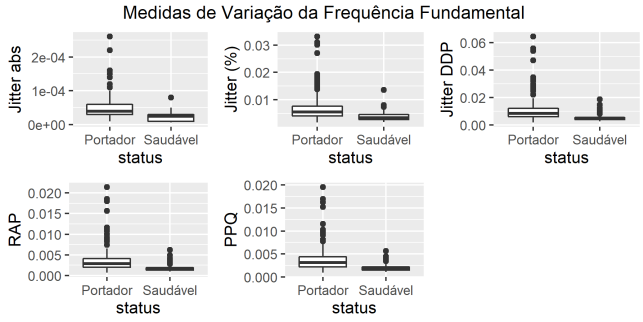
A Fig. 3 apresenta os gráficos de caixa para as medidas de F0. Observa-se que embora a haja uma distinção entre os valores médios para pacientes portadores e saudáveis, sendo a maior diferença observada na F0 média, há uma grande sobreposição no intervalo de valores. Observa-se também a presença de valores discrepantes na F0 máxima.

Figura 3: Gráfico de Caixa - F0



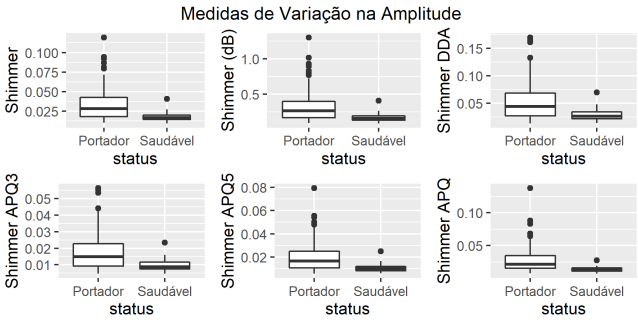
A Fig. 4 apresenta os gráficos de caixa para as medidas de Jitter. Observa-se um grande número de discrepância em todas elas. Além disso, como o intervalo de valores é muito grande e com muita sobreposição para ambos os status, é difícil determinar se há diferença significativa nas médias.

Figura 4: Gráfico de Caixa - Jitter



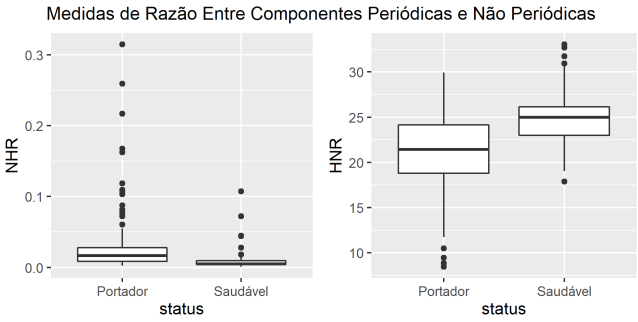
A Fig. 5 apresenta os gráficos de caixa para as medidas de Shimmer. Embora o número de discrepâncias também seja elevado em todas as medidas, a sobreposição de valores nos intervalos de cada status aparenta ser menor e a diferença de médias é mais observável (pessoas saudáveis possuem menor variação na amplitude da voz).

Figura 5: Gráfico de Caixa - Shimmer



A Fig. 6 apresenta os gráficos de caixa para os atributos HNR e NHR. Ambos possuem um número elevado de discrepâncias, mas este número é ainda maior para o NHR. Além disso, se observa uma moderada sobreposição de valores para cada status e a diferença de médias é bem aparente nos dois atributos.

Figura 6: Gráfico de Caixa - NHR e HNR



A Fig. 7 apresenta os gráficos de caixa para os atributos RPDE e D2. É possível observar que não há discrepâncias para o RPDE e um número reduzido para o D2. As médias para cada status são visivelmente distintas, embora a sobreposição dos intervalos seja muito grande.

O gráfico de caixa do atributos DFA é apresentado na Fig. 8. Não se observa a presença de discrepâncias. A diferença de médias é aparente e a sobreposição do intervalo de valores para cada status é a maior vista entre todos os atributos do conjunto de dados, sendo ambas muito próximas.

A Fig. 9 apresenta o gráfico de caixa das medidas não lineares de variação na F0. O número de discrepâncias é re-

Figura 7: Gráfico de Caixa - RPDE e D2

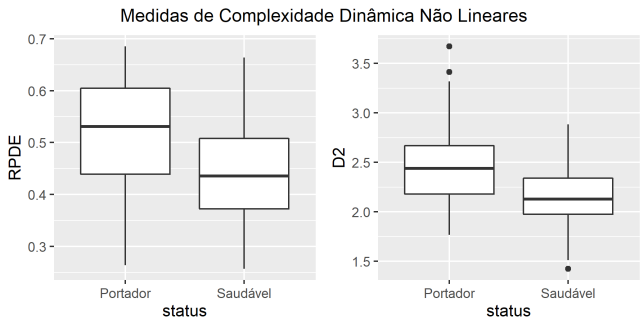
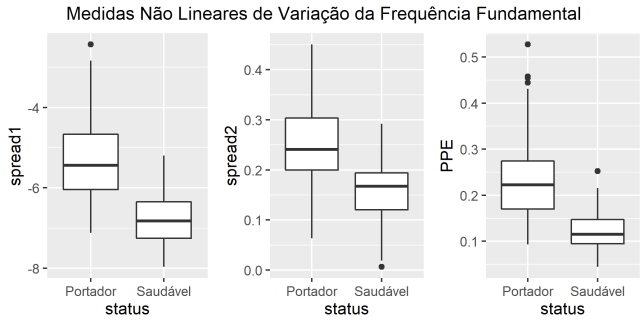


Figura 8: Gráfico de Caixa - DFA

duzido em spread1 e PPE e nulo em spread2. A diferença de médias para todas as medidas é grande e a sobreposição do intervalo dos valores é mínima, as menores encontradas entre todos os atributos do conjunto de dados.

Figura 9: Gráfico de Caixa - Medidas não lineares



Com os gráficos de caixa foi possível observar que nenhum atributos pode ser utilizado isoladamente como um classificador para a DP, uma vez que na maioria dos atributos houve grande sobreposição no intervalos de valores. Também observou-se que as medidas não lineares de variação de F0 são as que apresentam maior potencial para a classificação da doença.

Além disso, também foi observado que as medidas lineares mais tradicionais (F0, Jitter, Shimmer, HNR e NHR) possuem um grande volume de discrepâncias. Isto pode ocorrer em consequência desses atributos serem sensíveis a ruídos e interferências que possam ter ocorrido durante a captação dos sinais de voz dos pacientes.

O processo de filtragem dos atributos foi iniciado com um teste de hipótese utilizando a ditribuição t, onde a H_0 foi tida como a ausência de diferença entre as médias dos atributos para cada status. Desta forma, H_a representa que há diferença entre as médias para cada status.

O limiar utilizado com critério de rejeição da H_0 foi de 0.01, ou seja, o atributo que obtivesse valor-p inferior a 0.01 teria a diferença de médias atestada. Note-se que foi considerado um intervalo de confiança de 99%. A Tabela 18 apresenta o valor-p obtido com o teste para cada atributo. Observa-se que o único a ser rejeitado a F0 máxima.

Tabela 18: Valores-p obtidos para cada atributo.

Descrição	Identificador	Valor-p
F0 médio	Fo(Hz)	2.65E-05
F0 máximo	Fhi(Hz)	0.02
F0 mínimo	Flo(Hz)	6.58E-05
Medidas de Jitter	Jitter(%)	1.23E-08
	Jitter(Abs)	7.03E-12
	RAP	1.30E-08
	PPQ	3.37E-16
	Jitter:DDP	1.30E-08
Medidas de Shimmer	Shimmer	1.06E-15
	Shimmer(dB)	1.88E-14
	Shimmer:APQ3	1.08E-13
	Shimmer:APQ5	8.48E-15
	Shimmer:APQ	3.36E-16
	Shimmer:DDA	1.08E-13
Razões harmônico/ruído	HNR	2.42E-08
	NHR	1.54E-04
Medidas de complexidade	RPDE	8.71E-06
dinâmicas não lineares	D2	2.68E-07
Medidas não lineares	Spread1	1.77E-21
de variação de F0	Spread2	4.33E-12
	PPE	6.76E-23
Análise de flutuação	DFA	9.63E-04
sem tendência (expoente)		

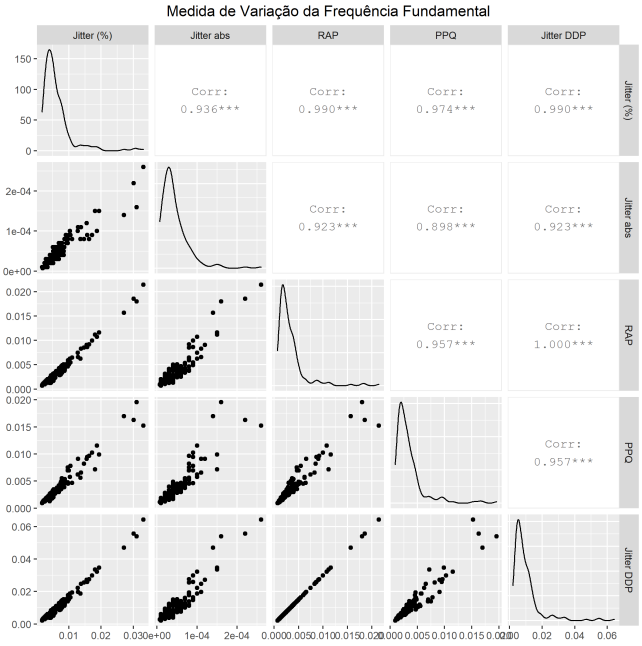
A segunda etapa do processo de filtragem consiste em uma análise de correlação entre os atributos que tratem de uma mesma característica da voz. Isto é feito a fim de evitar redundâncias no modelo. Como diversas medidas tratam de coisas semelhantes ou até mesmo idênticas, espera-se que haja entre elas uma elevada correlação.

Para cada característica de sinal de voz contida no conjunto de dados, observou-se a índice de correlação, para o par de atributos que tivessem uma correlação em módulo superior a 0.75, o atributo de menor relevância estatística, isto é, aquele com maior valor-p verificado no teste de hipótese era removido.

A Fig. 10 apresenta a análise de correlação entre os atributos de Jitter. Se observa que todos os pares de atributos

são fortemente correlacionados. Em módulo, o menor índice encontrado foi de 0.898. Desta forma, todos os atributos foram removidos do conjunto de dados com a exceção do Jitter(abs).

Figura 10: Análise de correlação - Jitter



A Fig. 11 apresenta a análise de correlação entre os atributos de Shimmer. Se observa que todos os pares de atributos são fortemente correlacionados. Em módulo, o menor índice encontrado foi de 0.897. Desta forma, todos os atributos foram removidos do conjunto de dados com a exceção do Shimmer:APQ.

Para as medidas não lineares de variação de F0, apenas a spread1 possuia uma correlação em módulo muito forte com PPE, acima de 0.9, conforme se observa na Fig. 12. Para as medidas de F0, não foi observada forte correlação entre elas, confome apresentado na Fig. 13. Logo a spread1 foi removida do conjunto de dados.

O conjunto de dados final utilizado é apresentado na Fig. 14 junto a análise de correlação dos atributos restantes. Observa-se que os atributos Jitter e Shimmer possuem forte correlação com PPE. Contudo, optou-se por mantê-los a fim de que o conjunto de dados final possua pelo menos um representante de característica.

Também se observa na Fig. 14 que os atributos de F0 mínima e spread2 foram removidos. Isto foi feito para manter reduzido o número de atributos do conjunto de dados final, uma vez que um número elevado de atributos leva a inflação da variância dos resultados obtidos o que consequentemente aumentaria a incerteza e imprecisão.

Figura 11: Análise de correlação - Shimmer

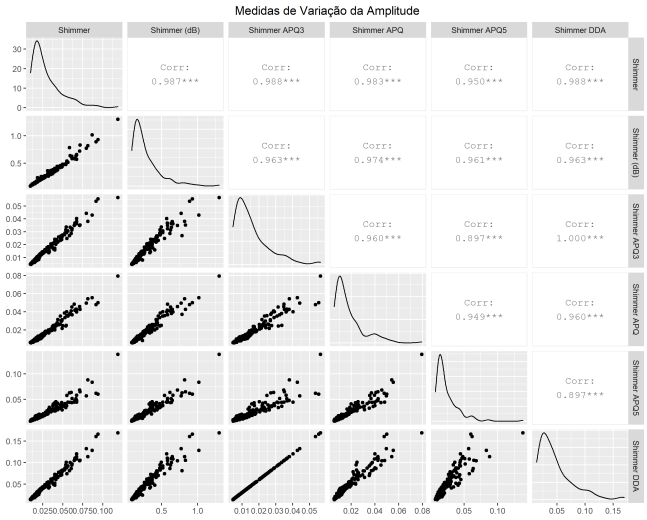
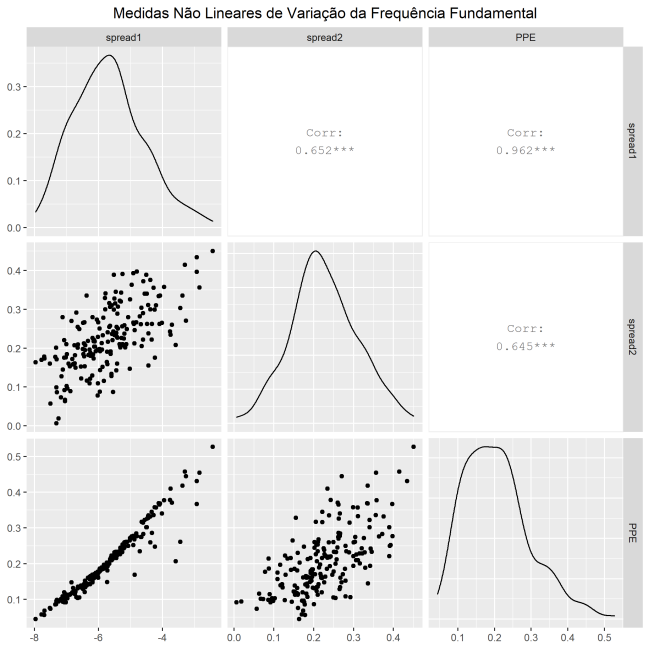


Figura 12: Análise de correlação - Medidas não lineares de variação de F0



4.2 Treinamento e Resultados

O conjunto de dados foi dividido em 6 bases de treino, a primeira delas contendo todos os atributos selecionados no processo de filtragem. As bases 2, 3 e 4 foram montadas seguindo o critério de baixa correlação entre os atributos. A base 5 é contém os atributos que obtiveram o melhor resultado no modelo de Little * et al. (2009). A base 6 é

Figura 13: Análise de correlação - F0

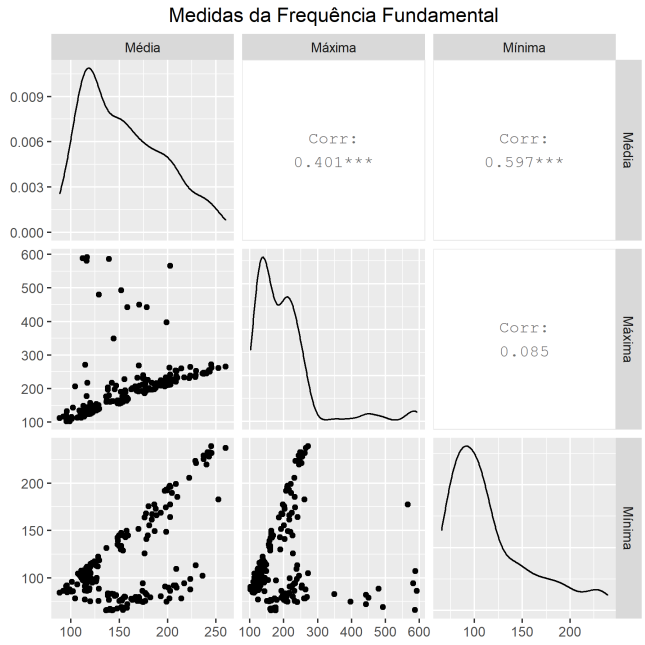
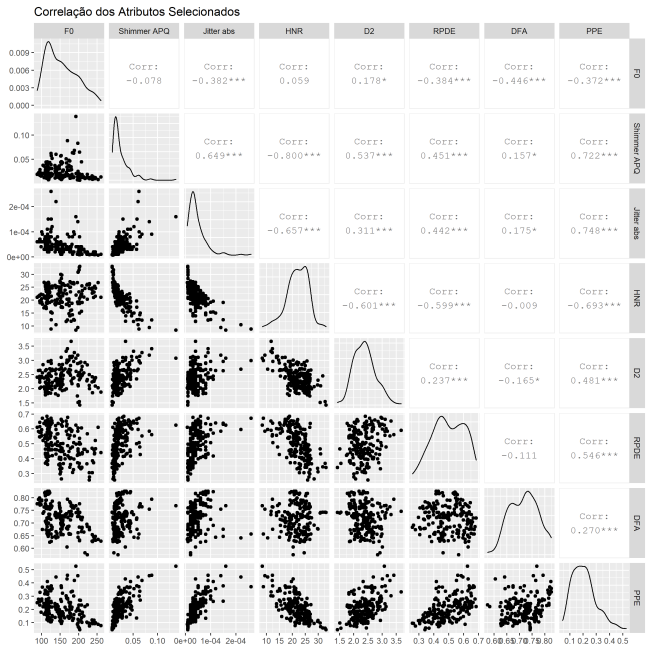


Figura 14: Análise de correlação - Base Final



idêntica a base 5 com a adição da F0, atributo que já foi observado como tendo relação com a DP Little * et al. (2009). As bases e suas descrições são apresentadas na Tabela 19.

Cada uma das bases foi utilizada num treinamento de um modelo de ML baseado na técnica de Random Forest. O treinamento foi realizado utilizando-se o pacote Caret da linguagem de programação R. A validação do modelo foi feita com a técnica K-fold com um k igual a 5, com cada um dos subconjuntos possuindo de 38 a 41 instâncias. Os resultados obtidos são apresentados na Tabela 20.

Observa-se que para todas as bases o valor do índice Kappa é superior a 0.7, o que indica uma concordância pelo menos substância, chegando a ser uma concordância quase perfeita para a Base 6, que obteve acurácia de 93.8, sensibilidade de 93.3 e especificidade de 95.0, superando todos os resultados obtidos pelos modelos das referências.

Observa-se também que os valores Verdadeiro Positivo (VP) foram todos superiores a 95. Já os valores Verdadeiro Negativo (VN) foram bem abaixo disso, sendo inferiores a 80. Isto ocorre devido ao desbalanceamento do conjunto de dados que possui cerca de 75% do total de amostras provenientes de pacientes portadores da DP.

Outro aspecto que se destaca é que os melhores resultados obtidos foram através das bases que continham atributos não lineares, obtendo acurácias sempre superiores a 90. Já as bases que continham apenas os atributos tradicionais obtiveram resultados menores que 90. Isto mostra que os atributos não lineares são indicativos mais relevantes da presença da DP.

Paracadaacuráciafoicalculadoumintervalodeconfiança utilizando a distribuição t. Optou-se pelo seu uso em detrimento da normal pelo falo do critério de normalidade não ter sido atendido no processo de treino por todos os subconjuntos, pois alguns subconjuntos não chegaram a errar a classificação para pelo menos 5 instâncias.

A Tabela 21 apresenta os melhores resultados de acurácia obtidos pelo modelo proposto com os seus respectivos intervalos de confiança junto aos resultados obtidos pelos modelos da referência. Os intervalos de confiança que foram omitidos não se encontravam descritos nos seus respectivos trabalhos.

O modelo proposto treinado a partir da Base 6 apresentou o melhor resultado de acurácia entre os modelos analisados. Contudo, observando-se os intervalos de confiança, constata-se que todos os resultados expostos na Tabela 16 estão dentro da mesma faixa, o que indica que as performances são praticamente idênticas.

Partindo de uma prevalência de 0.02% (20 casos por 100 mil habitantes) e utilizando a Lei de Bayes, verifica-se que dado que o modelo proposto apontou a presença da DP, a probabalidade de que o paciente de fato seja portador é menor que 1%. Para se obter uma precisão maior que 50%, seria necessário uma sensibilidade e especificidade maior que 99.98.

Desta forma, se observa que modelos de ML dificilmente poderão ser utilizados isoladamente como testes de diagnóstico de doenças com baixa prevalência. O ideal é que eles sirvam como ferramenta de apoio a profissinais da saúde tanto no processo de diagnóstico como nas pesquisas que buscam entender o comportamento dessas patologias.

Tabela 19: Bases de treino.

Base	Atributos	Critério
Base 1	Todos (8)	Base completa
Base 2	D2, RPDE, DFA, PPE (4)	Apenas atributos não lineares
Base 3	F0, Shimmer:APQ, Jitter(Abs), HNR (4)	Apenas atributos tradicionais
Base 4	F0, D2, RPDE, DFA, PPE (5)	Apenas atributos não lineares e F0
Base 5	HNR, RPDE, DFA, PPE (4)	Base de Litte et al.
Base 6	F0,HNR, RPDE, DFA, PPE (4)	Base de Litte et al. e F0

Tabela 20: Resultados do modelo.

Base	Acurácia	Kappa	Sensibilidade	Especificidade	VP	VN
Base 1	90.8	0.72	90.1	94.1	98.6	66.7
Base 2	88.7	0.71	89.3	86.1	96.6	64.6
Base 3	89.7	0.71	91.5	83.3	95.2	72.9
Base 4	92.3	0.77	91.3	97.1	99.3	70.8
Base 5	91.3	0.75	91.1	91.9	98.0	70.1
Base 6	93.8	0.82	93.3	95.0	98.6	79.2

Tabela 21: Comparativo com outros modelos.

Modelo	Acurácia
Proposto (Base 6)	93.8 ± 4.4
Rede neural (Das, 2010)	92.9
SVM (Sakar; Kursun, 2009)	92.8 ± 1.2
Proposto (Base 4)	92.3 ± 4.7
Random forest (Govindu; Palwe, 2023)	91.83
SVM (Little et al., 2009)	91.4 ± 4.4

5 Considerações Finais

O objetivo deste trabalho foi propor a construção de uma ferramenta que auxiliasse o profissional da saúde no diagnóstico da DP. Isto foi realizado por meio de um modelo de ML fazendo uso de um algoritmo de Random Forest que obteve uma acurácia de 93.8% e um índice Kappa de 0.82, indicando uma concordância quase perfeita.

Foi demonstrado, utilizando-se a Lei de Bayes, que a probabilidade de que um determinado paciente seja portador da DP dado que o modelo apontou a presença do Parkinson é menor que 1% e que uma probabilidade maior que 50% necessitaria de sensibilidade e espeficificidade superiores a 99.98%, o que dificilmente é atingido por um modelo sem sobre ajuste.

Uma forma de melhorar este resultado seria através de uma base que contivesse apenas dados de pacientes com mais de 65 anos, uma vez que para estes casos, a prevalência da DP é de cerca de 1%. Contudo, ainda assim os valores de sensibilidade e especificidade teriam de ser superiores a 99% para se obter uma probabilidade de acerto no diagnóstico maior que 50%.

Outro resultado relevante foi a demonstração, através da inferência estatística, de que atributos da fala de fato possuem relação com a DP, especialmente os atributos não lineares, destacando-se a PPE. Esses atributos levantam novas possibilidades em pesquisas que buscam o entendimento do comportamento da patologia.

Referências

Almeida, J. S., Filho, P. P. R., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R. and de Albuquerque, V. H. C. (2019). Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques, *Pattern Recognition Letters* **125**: 55–62. https://www.sciencedirect.com/science/article/pii/S0167865519301163.

Amato, F., Saggio, G., Cesarini, V., Olmo, G. and Costantini, G. (2023). Machine learning- and statistical-based voice analysis of parkinson's disease patients: A survey, *Expert Systems with Applications* **219**. https://www.sciencedirect.com/science/article/pii/S0957417423001525.

Bhattacharya, I. and Bhatia, M. (2010). Svm classification to distinguish parkinson disease, *Proceedings of the A2CWiC'10* **Article 14**: 1–6. https://doi.org/10.1145/1858378.1858392.

Caffo, B. (2016). *Statistical Inference for Data Science*, Leanpub.

da Silva, P. P., Linares, K. S. C. and Patrício, C. M. M. M. (2006). Análise estatística no diagnóstico de doenças cardíacas, *RESI - Revista Eletrônica de Sistemas de Informação* **5**. https://doi.org/10.21529/RESI.2006.0503012.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease, *Expert Systems with Applications* **37**: 1568–1572. https://www.sciencedirect.com/science/article/pii/S0957417409006137.

de Carvalho Teles, V. and Rosinha, A. C. U. (2008). Análise acústica dos formantes e das medidas de perturbação do sinal sonoro em mulheres sem queixas vocais, não fumantes e não etilista, *Arquivos internacionais de otorrinolaringologia* **12**: 523–530. https://arquivosdeorl.org.br//conteudo/acervo_port.asp?id=567.

Govindu, A. and Palwe, S. (2023). Early detection of parkinson's disease using machine learning, *Procedia Com-*

puter Science **218**: 249–261. https://www.sciencedirect.com/science/article/pii/S1877050923000078.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning*, Springer, Stanford, CA.

Ho, A. K., Iansekb, R., Mariglianib, C., Bradshawa, J. L. and Gates, S. (1999). Speech impairment in a large sample of patients with parkinson's disease, *Behavioural Neurology* **11**: 131–137. https://doi.org/10.1155/1999/327643.

Karaman, O., Çakın, H., Alhudhaif, A. and Polat, K. (2021). Robust automated parkinson disease detection based on voice signals with transfer learning, *Expert Systems with Applications* **178**. https://www.sciencedirect.com/science/article/pii/S0957417421004541.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data, *Biometrics* **33**: 159–174. https://pubmed.ncbi.nlm.nih.gov/843571/.

Lenain, R., Weston, J., Shivkumar, A. and Fristed, E. (2020). Surfboard: Audio feature extraction for modern machine learning, *ArXiv* . https://doi.org/10.48550/arXiv.2005.08848.

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

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. and Moroza, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, *BioMedical Engineering OnLine* **6**. https://doi.org/10.1186%2F1475-925x-6-23.

Massano, J. and Cabreira, V. (2019). Doença de parkinsonn: Revisão clínica e atualização, *Acta Médica Portuguesa* **32**. https://doi.org/10.20344/amp.11978.

Mohri, M., Rastamizadeh, A. and Talwalkar, A. (2018). *Foudantions of Machine Learning*, MIT Press, London.

Nasri, F. (2008). O envelhecimento populacional no brasil, *Einstein* **24**. https://pesquisa.bvsalud.org/portal/resource/pt/lil-516986.

Ngo, Q. C., Motin, M. A., Pah, N. D., Drotár, P., Kempster, P. and Kumar, D. (2022). Computerized analysis of speech and voice for parkinson's disease: A systematic review, *Computer Methods and Programs in Biomedicine* **226**. https://www.sciencedirect.com/science/article/pii/S0169260722005144.

Ozkan, V. (2016). A comparison of classification methods for telediagnosis of parkinson's disease, *Entropy* **18**. https://doi.org/10.3390/e18040115.

Parkinson, J. (1817). An essay on the shaking palsy, *J Neuropsychiatry Clin Neurosci* **14**. https://doi.org/10.1176/jnp.14.2.223.

ROCHE (2018). *Dia Mundial do Parkinson*. Available at https://www.roche.com.br/pt/por-dentro-da-roche/dia-mundial-do-parkinson.html.

Sakar, C. O. and Kursun, O. (2009). Telediagnosis of parkinson's disease using measurements of dysphonia, *Journal of Medical Systems* **34**: 591–599. https://doi.org/10.1007/s10916-009-9272-y.

Steidl, E. M. d. S., Ziegler, J. R. and Ferreira, F. V. (2007). Doença de parkinson: Revisão bibliográfica, *Disciplinarum Scientia* **8**. https://periodicos.ufn.edu.br/index.php/disciplinarumS/article/view/921/865.

Tai, Y. C., Bryan, P. G., Loayza, F. and Peláez, E. (2021). A voice analysis approach for recognizing parkinson's disease patterns, *IFAC-PapersOnLine* **54**: 382–387. https://www.sciencedirect.com/science/article/pii/S2405896321016918.

Teixeira, J. P., Oliveira, C. and Lopes, C. (2013). Vocal acoustic analysis: Jitter, shimmer and hnr parameters, *Procedia Technology* **9**. https://www.sciencedirect.com/science/article/pii/S2212017313002788.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, *Pattern Recognition* **48**: 2839–2846. https://www.sciencedirect.com/science/article/pii/S0031320315000989.