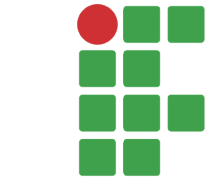


Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

MATEUS MELO

APOIO AO DIAGNÓSTICO DE PARKINSON POR INTELIGÊNCIA ARTIFICIAL E SINAIS DE VOZ

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2024

Mateus Melo

Apoio ao Diagnóstico de Parkinson por Inteligência Artificial e Sinais de Voz

Dissertação de Mestrado apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dr. Thiago Gouveia

João Pessoa

2024

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *Campus* João Pessoa, PB.

K81a Kolpeman, Mateus de Lima Melo.

Apoio ao diagnóstico de *Parkinson* por inteligência artificial e sinais de voz / Mateus de Lima Melo Kolpeman. - 2024.
101 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2024.
Orientação : Prof. Dr. Thiago Gouveia da Silva.

1.Doença de *Parkinson.* 2. Aprendizado de máquina. 3. *Random Forest*.4. Espectrograma. 5. Redes neurais convolucionais. I.Título.

CDU 616.8:004.8(043)

Bibliotecária responsável Lucrecia Camilo de Lima – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

MATEUS DE LIMA MELO KOLPEMAN

Apoio ao Diagnóstico de Parkinson por Inteligência Artificial e Sinais de Voz

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 20 de dezembro de 2024

Membros da Banca Examinadora:

Dr. Thiago Gouveia da Silva

IFPB - PPGTI

Dr. Diego Ernesto Rosa Pessoa

IFPB - PPGTI

Dr. Gilberto Farias de Sousa Filho

UFPB

João Pessoa/2024

Documento assinado eletronicamente por:

- **Thiago Gouveia da Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 30/12/2024 12:09:41.
- **Diego Ernesto Rosa Pessoa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 30/12/2024 12:36:57.
- **Gilberto Farias de Sousa Filho**, **PROFESSOR DE ENSINO SUPERIOR NA ÁREA DE ORIENTAÇÃO EDUCACIONAL**, em 23/01/2025 09:06:37.

Este documento foi emitido pelo SUAP em 06/12/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código 642857
Verificador: 12ca434ecb
Código de Autenticação:

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 pelo 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 desse contexto, propostas para a utilização de sinais de voz como forma de diagnóstico precoce do Parkinson vêm obtendo resultados. Este trabalho propõe a utilização de técnicas de Aprendizado de Máquina no problema de classificação de sinais de voz para diagnóstico da doença de Parkinson. Fazendo uso de um conjunto de dados com áudios proveninentes da fala de portadores e não portadores da patologia, obteve-se uma acurácia superior a 91% utilizando um comitê de classificadores que mescla características de modelos de Random Forest e Redes Neurais Convolucionais.

Palavras-chave: Doença de Parkinson. Aprendizado de Máquina. Random Forest. Espectrogramas. Redes Neurais Convolucionais. Comitê de Classificadores.

ABSTRACT

Parkinson's disease is a neurodegenerative pathology that affects motor and speech ability, in addition to causing behavioral, mood and reasoning changes. It most commonly affects the elderly population and its diagnosis is made through a clinical examination, through observation of the symptoms presented by a patient. Since the most notorious symptoms tend to appear in more advanced stages of the disease, which makes treatment difficult, and since the world is going through a process of inversion of the age pyramid, the tendency is for Parkinson's to become a problem of global public health. Within this context, proposals for the use of voice signals as a means of early diagnosis of Parkinson's are getting results. This work proposes the use of Machine Learning techniques in the problem of classifying voice signals for diagnosing Parkinson's disease. Using a dataset with audio from the speech of people with and without the disease, an accuracy of over 91% was obtained using a ensemble learning model that mixes characteristics of Random Forest and Convolutional Neural Networks models.

Keywords: Parkinson's disease. Machine Learning. Random Forest. Spectrograms. Convolutional Neural Networks. Ensemble Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

De acordo com Massano e Cabreira (2019), 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, atingindo mais frequentemente a idosos. Seu diagnóstico é feito de forma clínica por um neurologista através da exclusão de outras doenças, não havendo atualmente um teste ou biomarcador que possa indicar a sua presença (MASSANO; CABREIRA, 2019). Isto faz com que o Parkinson seja frequentemente diagnosticado tardiamente, uma vez que os principais sintomas aparecem apenas em estágios mais avançados da doença.

Nesse cenário, a análise acústica vocal, método que consiste na quantificação e caracterização de um sinal sonoro proveniente da fala humana, se apresenta como alternativa que possibilitaria um diagnóstico precoce da doença de Parkinson, levando-se em consideração que, como observado por Ho et al. (1999) e Little et al. (2009), cerca de 90% dos portadores do Parkinson apresentam alguma deficiência vocal.

Associando-se o Aprendizado de Máquina com a análise acústica vocal, é possível identificar alterações nos sinais de voz cuja percepção seria impossível utilizando-se apenas a audição humana. Diversas pesquisas vêm obtendo ótimos resultados na classificação de vozes de pessoas saudáveis das vozes de pessoas portadoras do Parkinson, como evidenciado por Little et al. (2009), Bhattacharya e Bhatia (2010), Das (2010) e Sakar e Kursun (2009), Govindu e Palwe (2023) e Melo e Gouveia (2023). Esses procedimentos podem ainda ser empregados no estudo de outras doenças neurodegenerativas, como foi observado por Riad et al. (2020) ao analisar a doença de Huntington.

Karaman et al. (2021) constatou que o uso de representações visuais bidimensionais se apresenta como alternativa promissora na classificação de sinais de voz. Essas representações são chamadas de espectrogramas e são costumeiramente obtidas por meio da transformada de Fourier de curta duração, podendo ser associados com métodos de Aprendizado de Máquina avançados, como as redes neurais convolucionais.

Além disso, informações relevantes podem ser extraídas dos modelos de Aprendizado de Máquina que, quando associadas a modelos de análise estatística, podem apresentar relações e correlações entre a doença de Parkinson e os sinais de voz, indicando novas possibilidades aos estudos que buscam tanto o entendimento do surgimento da patologia, como uma possível cura.

Assim, esta pesquisa terá como foco o desenvolvimento de um modelo de Aprendizado de Máquina para classificação de sinais de voz de portadores da doença de Parkinson. Além disso, também serão apresentadas técnicas de extração de atributos e espectrogramas do sinais

de voz provenientes da fala humana que possam ser relevantes no entendimento da doença como um todo.

De forma mais específica, este trabalho propõe-se a investigar, analisar, implementar e avaliar métodos de Aprendizado de Máquina na classificação de pacientes portadores da doença de Parkinson utilizando sinais de voz. Pretende-se atestar a viabilidade desses métodos como alternativa para o diagnóstico dessa patologia.

1.1 Definição do problema

O objeto dessa pesquisa é o estudo das principais técnicas de análise de dados, inferência estatística e processamento de sinais de voz em conjunto com técnicas de Aprendizado de Máquina que possam ser utilizadas na construção de um modelo que auxilie profissionais da saúde no diagnóstico da doença de Parkinson, utilizando sinais de voz provenientes de pacientes portadores e não portadores da doença. Isso será feito por meio de um classificador que possa indicar com elevada precisão a presença ou ausência da patologia, e também pelo fornecimento de informações sobre os diferentes atributos da voz humana, bem como suas relações e correlações com a presença do Parkinson que possam ser úteis tanto no entendimento do surgimento da doença, como apresentando alternativas para o seu tratamento.

A pesquisa se concentrará nos atributos de sinais de voz mais tradicionalmente utilizados na literatura como os de frequência fundamental, variação da frequência fundamental e variação de amplitude do sinal, como apresentados por Little et al. (2007a). Os atributos não lineares, que foram introduzidos por Little et al. (2009) e Tai et al. (2021), também serão analisados. Técnicas de inferência estatística serão utilizadas a fim de se verificar a relação desses atributos com a doença de Parkinson. Espectrogramas também serão investigados uma vez que Karaman et al. (2021) obteve resultados promissores com seu uso.

Os atributos e espectrogramas observados como mais estatisticamente relevantes na detecção da doença de Parkinson serão utilizados no processo de treinamento de modelos de aprendizado de máquina baseados nos diversos algoritmos que vêm sendo usados na literatura, destacando-se as redes neurais convolucionais e random forest.

1.2 Motivação

O Parkinson é a segunda doença neurodegenerativa mais frequente no mundo, sendo superada apenas pela doença de Alzheimer. Ele possui uma incidência média mundial na faixa de 15 a 20 casos para cada 100 mil habitantes por ano (MASSANO; CABREIRA, 2019). Na Europa, o valor é ainda mais elevado, chegando a uma incidência média na faixa de 75 a 300 casos para cada 100 mil habitantes por ano (STEIDL; ZIEGLER; FERREIRA, 2007). Na América do Norte, estima-se que mais de 1 milhão de pessoas tenham as suas vidas afetadas pela doença de Parkinson (LITTLE et al., 2009).

A doença de Parkinson afeta principalmente os idosos. Estima-se que a doença acometa 1% da população mundial com idade acima de 65 anos (MASSANO; CABREIRA, 2019). Como o processo de inversão da pirâmide etária está se acentuando, o Parkinson deverá representar cada vez mais um problema grave de saúde pública, especialmente no Brasil onde estima-se que até o ano de 2050, a quantidade de idosos representará 19% de toda a sua população (NASRI, 2008).

Dentro desse contexto e uma vez que não há ainda hoje um teste ou biomarcador que ateste a presença da doença de Parkinson (MASSANO; CABREIRA, 2019), faz-se extremamente relevante a investigação de métodos alternativos que possam auxiliar no diagnóstico da doença. Isso traria vantagens não apenas aos profissionais da área da saúde, que contariam com um método mais objetivo para detecção da patologia, diminuindo o ruído no diagnóstico, como também para os pacientes, uma vez que a identificação precoce de doenças neurodegenerativas influencia diretamente na qualidade de vida dos portadores.

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral desta pesquisa é o desenvolvimento de um modelo de aprendizado de máquina que faça uso de sinais de voz para classificar portadores da doença de Parkinson.

1.3.2 Objetivos específicos

- Propor e desenvolver métodos computacionais de extração de atributos de sinais de voz que possam ter relação com a doença de Parkinson.
- Propor e desenvolver métodos computacionais de obtenção de espectrogramas de sinais de voz.
- Construir e analisar estatisticamente uma base de dados proveniente de atributos de áudio extraídos de sinais de voz de portadores e não portadores da doença de Parkinson.
- Desenvolver um classificador por meio de métodos de Aprendizado de Máquina, comparando os resultados obtidos com os da literatura.
- Avaliar a viabilidade do classificador como alternativa para diagnóstico da doença de Parkinson

1.4 Estrutura do documento

Este documento está dividido em 6 capítulos, sendo esta introdução o primeiro deles. O Capítulo 2 apresenta a fundamentação teórica necessária à compreensão do trabalho, explicando

conceitos envolvendo a doença de Parkinson, a análise acústica vocal e o aprendizado de máquina. O Capítulo 3 trata dos principais trabalhos relacionados ao diagnóstico de Parkinson com sinais de voz, além de tratar das oportunidades identificadas no tema. Já o Capítulo 4 descreve a metodologia do trabalho e os modelos propostos, incluindo os experimentos computacionais realizados. O Capítulo 5 apresenta os principais resultados obtidos, além de abordar as discussões acerca da aplicabilidade da proposta. Por fim, o Capítulo 6 apresenta as considerações finais do trabalho e apresenta possibilidades para a obtenção de melhores resultados em pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para o melhor entendendimento deste trabalho. A Seção 2.1 apresenta os principais conceitos e dados relacionados a doença de Parkinson. A Seção 2.2 apresenta a análise acústica vocal, como ela vem sendo utilizada na medicina e quais os principais atributos que podem ser extraídos de um sinal de voz. Por fim, a Seção 2.3 apresenta os principais conceitos, métricas e algoritmos de aprendizado de máquina necessários para o entendimento dos métodos propostos.

2.1 A doença de Parkinson

Esta seção apresenta os principais conceitos e definições relacionados à doença de Parkinson (DP). Nela serão apresentadas informações gerais e descritivas da patologia, seus principais sintomas, o contexto socioeconômico em que ela está inserida, a forma de diagnóstico e as formas de tratamento mais tradicionais.

2.1.1 Definição

A DP, originalmente conhecida como paralisia agitante, manifesta-se como uma condição que resultava em movimentos involuntários tremulantes, fraqueza muscular, inclinação do tronco para frente e perturbação no padrão da marcha, mesmo quando o indivíduo não exibe lesões nos sentidos ou no intelecto (PARKINSON, 1817).

O médico e pesquisador Jean-Martin Charcot cunhou o termo “Doença de Parkinson” como uma homenagem a James Parkinson, o primeiro a abordar essa condição clínica em 1817. Esse termo logo se difundiu e se tornou a designação mais comum para a doença.

Na atualidade, a DP é caracterizada como uma doença crônica e degenerativa que afeta o sistema nervoso central, envolvendo os gânglios da base. Sua origem está associada à deficiência do neurotransmissor dopamina na via nigroestriatal e cortical, o que causa principalmente interferências no sistema motor (STEIDL; ZIEGLER; FERREIRA, 2007).

2.1.2 Principais sintomas

Os sintomas da DP podem ser classificados em duas categorias: os motores, também conhecidos como parkinsonismo, e os não motores. De acordo com Massano e Cabreira (2019), o parkinsonismo é marcado principalmente por acinesia ou bradicinesia, rigidez, tremores em repouso e alterações na postura e marcha.

A acinesia se caracteriza não apenas pela redução progressiva da velocidade dos movimentos, mas também pela diminuição da amplitude dos mesmos. Outras manifestações possíveis

incluem a hipomimia, que se refere a uma expressão facial reduzida ou imóvel, a hipofonia, que é a diminuição do volume na fala, e a micrografia, que se traduz em uma caligrafia menor, sendo este último geralmente mais difícil de ser percebido.

A rigidez é marcada pela sensação de resistência ou oposição ao movimento, flexão ou extensão de um membro. A velocidade da movimentação não exerce influência negativa ou positiva na sensação de rigidez, porém essa sensação aumenta quando há ativação simultânea de outro membro.

O tremor de repouso é o sintoma mais característico da DP. Ele se refere ao tremor que ocorre nos membros quando estão relaxados e apoiados em uma superfície, sem ação da gravidade que justifique o movimento. O chamado "tremor a contar moedas" é o tipo mais comum e é caracterizado pela movimentação de adução-abdução do polegar (MASSANO; CABREIRA, 2019).

É fundamental ressaltar que o parkinsonismo pode ser causado por condições além da DP. O que diferencia o parkinsonismo da DP é a manifestação de características distintas, como o fato dos sintomas surgirem e progredirem inicialmente em apenas um lado do corpo durante os estágios iniciais da doença.

O parkinsonismo pode ser classificado em três categorias: primário ou idiopático, secundário e *plus*. O tipo mais comum é o idiopático, que representa aproximadamente 75% dos casos, sendo ele a própria DP. Os dois últimos tipos não estão relacionados à DP e podem surgir devido a infecções, efeitos de medicamentos, acidentes ou outras causas diversas (STEIDL; ZIEGLER; FERREIRA, 2007).

Por fim, os sintomas não motores (SNM) podem desempenhar um papel fundamental no diagnóstico precoce da DP, uma vez que eles podem surgir anos antes dos sintomas motores. De acordo com Massano e Cabreira (2019), alguns dos principais SNM incluem depressão, apatia, disfunção sexual, alterações no sono, ansiedade, fadiga, deterioração cognitiva e alterações psicóticas Identificar e avaliar esses sintomas não motores pode ser essencial para um diagnóstico mais antecipado e preciso da doença.

2.1.3 Contexto socioeconômico

A DP é considerada uma doença cosmopolita, pois não faz distinção entre os indivíduos, afetando pessoas de diferentes classes sociais. Ela ocorre tanto em homens quanto em mulheres, principalmente na faixa etária entre 55 e 65 anos, embora seja mais frequente em pessoas do sexo masculino. Isso ressalta a importância da conscientização e do diagnóstico precoce, independentemente do gênero ou status social (STEIDL; ZIEGLER; FERREIRA, 2007).

O Parkinson é a segunda doença neurodegenerativa mais comum em todo o mundo, ficando atrás apenas da Doença de Alzheimer. Devido à complexidade na determinação exata da epidemiologia da DP, não há consenso sobre o número exato de casos. Estima-se que a

incidência média mundial esteja entre 15 e 20 casos por 100 mil habitantes por ano (MASSANO; CABREIRA, 2019). A incidência da DP varia em diferentes regiões do mundo. Na Europa, estima-se que a taxa fique entre 75 e 300 casos por 100 mil habitantes por ano, mas alguns estudos sugerem que esse número pode chegar a 12.500 casos (STEIDL; ZIEGLER; FERREIRA, 2007). Na América do Norte, mais de um milhão de pessoas são afetadas pela DP (LITTLE et al., 2009). Estima-se 200 mil portadores da DP no Brasil (ROCHE, 2018).

A DP tem uma forte incidência entre os idosos, afetando principalmente essa faixa etária. Estima-se que cerca de 1% da população mundial com mais de 65 anos seja acometida pela doença (MASSANO; CABREIRA, 2019). Com o aumento da expectativa de vida em muitas regiões, o número de idosos está crescendo, o que indica que a DP pode causar um impacto significativo 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). Esse processo já vem sendo observado na última decada, conforme pode ser visto na Figura 1.

2.1.4 Diagnóstico

O diagnóstico da DP é fundamentalmente clínico e atualmente não existe um teste específico ou biomarcador que possa confirmar definitivamente a presença da doença (MASSANO; CABREIRA, 2019). A falta de um exame conclusivo é uma das razões que levam às discrepâncias nos números de casos relatados em diferentes regiões, como na Europa, onde a incidência média anual varia significativamente, indo de 75 até 12.500 pessoas por 100 mil habitantes. A ausência de um teste definitivo faz com que o diagnóstico dependa da avaliação clínica dos sintomas motores e não motores apresentados pelo paciente, o que pode levar a variações nos registros epidemiológicos em diferentes populações e sistemas de saúde.

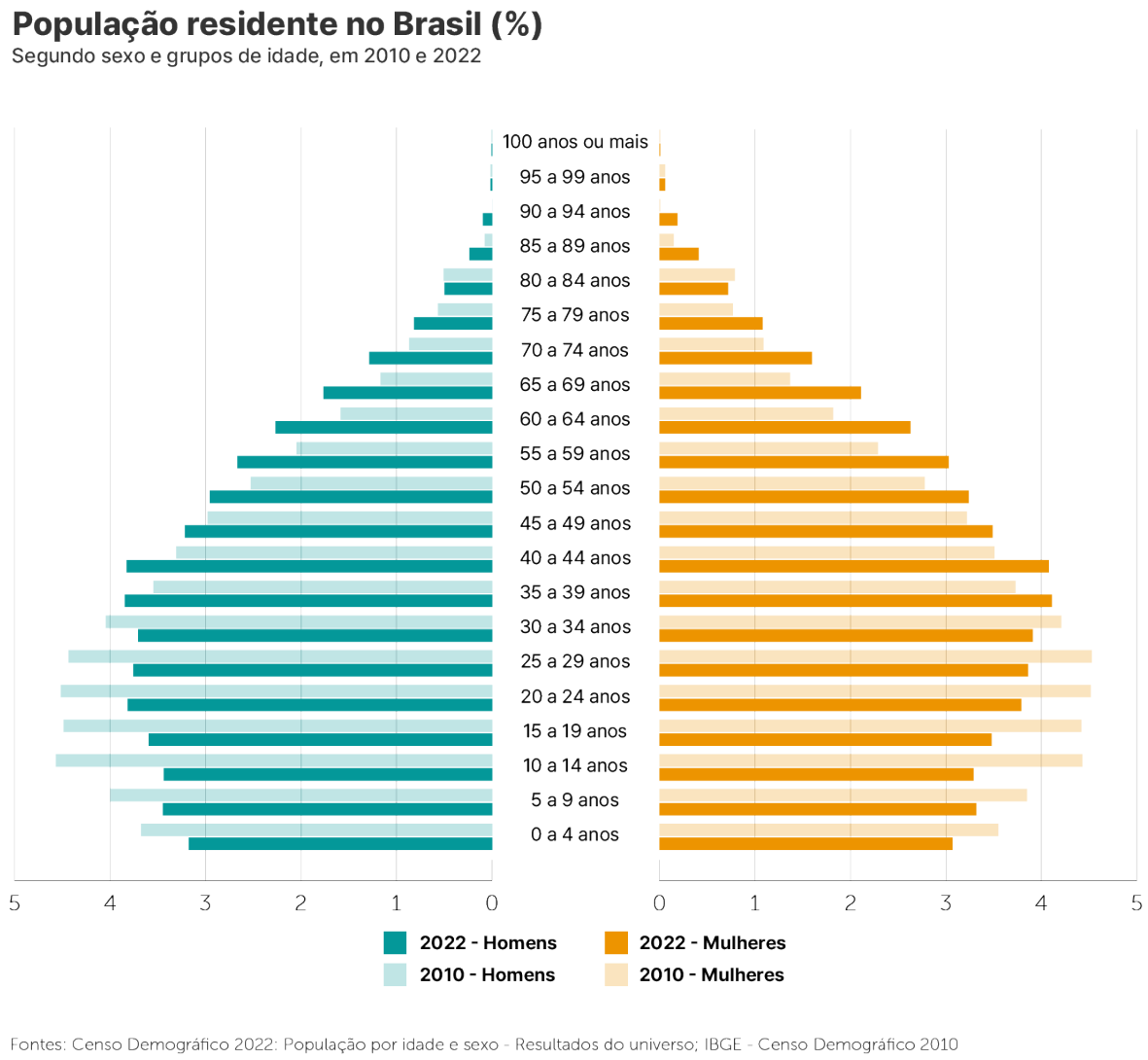
A DP é diagnosticada por um neurologista através da exclusão de outras doenças. Após a descrição dos sintomas pelo paciente, o médico solicita uma série de exames (eletroencefalograma, tomografia computadorizada, ressonância magnética e análise do líquido espinhal) a fim de atestar a ausência de outras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

2.1.5 Tratamento

Atualmente, os tratamentos propostos 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.

Uma vez que a DP é uma doença degenerativa e sem cura, os tratamentos visam amenizar os sintomas e desacelerar sua evolução. Existem diversas abordagens terapêuticas, adaptadas ao estágio em que o paciente se encontra e ao sintoma específico a ser tratado, sendo os colinérgicos

Figura 1 – Inversão da pirâmide populacional



Fonte: IBGE (2022)

os medicamentos mais comumente empregados no tratamento da DP.

Além disso, é fundamental contar com a supervisão de uma equipe diversificada de profissionais, incluindo fisioterapeutas, psicólogos, fonoaudiólogos, nutricionistas e neurologistas, trabalhando em conjunto para fornecer um tratamento multidisciplinar visando melhorar a qualidade de vida do indivíduo com DP (STEIDL; ZIEGLER; FERREIRA, 2007). Essa abordagem abrangente e integrada é essencial para enfrentar os desafios associados à doença e proporcionar um suporte adequado em todas as áreas afetadas pela condição.

Existe também a opção de tratamento cirúrgico, como apresentado na Tabela 2. No entanto, independentemente das medidas terapêuticas clínicas ou cirúrgicas adotadas, a DP tem uma evolução progressiva. Nesse sentido, o tratamento cirúrgico busca melhorar a qualidade de

Tabela 1 – Classificação de Hoehn e Yahr

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

Fonte: Steidl, Ziegler e Ferreira (2007)

vida dos pacientes com DP, proporcionando alívio dos sintomas e uma melhora no bem-estar geral.

Tabela 2 – Procedimentos cirúrgicos e disfunções

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

Fonte: Steidl, Ziegler e Ferreira (2007)

2.2 Análise de sinais de voz na medicina

Esta seção discute a utilizção de sinais de voz na medicina para detecção de doenças. São apresentados os principais conceitos relacionados à análise acústica vocal, os principais atributos extraídos, bem como atributos alternativos que podem ser extraídos de sinais sonoros e serem aplicados em problemas de classificação. Além disso, também será apresentado como os sinais sonoros podem ser representados de forma visual por meio de espectrogramas.

2.2.1 Definição e motivação

A análise de sinais de voz, ou análise acústica vocal, tem como objetivo principal quantificar e caracterizar um sinal sonoro proveniente da fala humana. No Brasil, e mais especificamente na clínica fonoaudiológica, o uso dessa análise tem se intensificado desde o início do século XXI (TELES; ROSINHA, 2008).

No campo médico, diversas técnicas são empregadas para avaliar a qualidade vocal do paciente, sendo a análise perceptivo-auditiva a mais comum. No entanto, essa técnica pode levar a resultados variados, dependendo da experiência do profissional envolvido. Diante desse cenário, houve um impulso para o desenvolvimento da análise acústica como uma abordagem complementar. A análise acústica se baseia em medições objetivas das características do som vocal, oferecendo uma avaliação mais precisa e padronizada, o que pode aprimorar a compreensão e o acompanhamento dos distúrbios vocais e suas intervenções terapêuticas (TEIXEIRA; OLIVEIRA; LOPES, 2013).

Características vocais como rouquidão, soprosidade e rugosidade podem indicar a presença de várias patologias, como nódulos vocais, laringite, cistos, endema de Reinke, câncer de laringe, entre outras condições. Através da análise acústica vocal, essas características podem ser facilmente detectadas, o que auxilia os profissionais da saúde, especialmente os fonoaudiólogos, a identificarem e diagnosticarem essas patologias de forma mais precisa e objetiva (TEIXEIRA; OLIVEIRA; LOPES, 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).

As doenças neurodegenerativas têm impacto na voz dos indivíduos. Em um estudo envolvendo 200 portadores de DP, constatou-se que 89% deles apresentavam algum problema vocal (HO et al., 1999). Segundo Little et al. (2009), até 90% dos portadores de DP possuem alguma deficiência vocal.

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 que possam afetar a qualidade vocal de um indivíduo.

2.2.2 Atributos tradicionais

Os atributos mais usualmente extraídos e investigados são a frequência fundamental (F0), jitter, shimmer, a razão harmônica-ruído (HNR, do inglês harmonic-noise ratio) e os coeficientes cepstrais da frequência de Mel (MFCCs, do inglês Mel-frequency cepstral coefficients).

A F0 de um sinal de voz é medida em Hertz (Hz) e representa a quantidade de vezes que a onda sonora produzida pelas cordas vocais se repete em um determinado período. Além disso, a F0 também representa o número de ciclos de abertura/fechamento da glote (espaço entre as cordas vocais responsável pela sonorização das vogais) (TEIXEIRA; OLIVEIRA; LOPES, 2013).

As medidas de distúrbios na frequência fundamental (F0), como jitter e shimmer, são ferramentas úteis na descrição de características vocais. Jitter é definido como a variação na

frequência de ciclo para ciclo, ou seja, a irregularidade nas repetições da onda sonora produzida pelas cordas vocais. O shimmer está relacionado com a variação na amplitude da onda sonora, indicando a instabilidade na intensidade da voz ao longo do tempo (TEIXEIRA; OLIVEIRA; LOPES, 2013). As Tabelas 3 e 4 apresentam os principais atributos de jitter e shimmer.

Tabela 3 – Atributos de jitter

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

Fonte: Teixeira, Oliveira e Lopes (2013)

Tabela 4 – Atributos de shimmer

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

Fonte: Teixeira, Oliveira e Lopes (2013)

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; OLIVEIRA; LOPES, 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; OLIVEIRA; LOPES, 2013). Os valores de limiar para vozes patológicas são apresentado na Tabela 5.

O MFCC é um dos atributos mais utilizados no processamento de sinais de voz, sendo utilizado em inúmeras aplicações, em especial no reconhecimento de locutor, reconhecimento de voz e identificação de gênero (ABDUL; AL-TALABANI, 2022). O MFCC tem como objetivo mimetizar o processamento sonoro efetuado pelos ouvidos humanos, agrupando as frequências que compõem um sinal em faixas representativas (ZHOU et al., 2011).

Tabela 5 – Limiares para vozes patológicas

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

Fonte: Teixeira, Oliveira e Lopes (2013)

2.2.3 Atributos não lineares

Atributos não lineares vêm sendo utilizados na análise acústica vocal. Esses atributos eram mais usualmente usados no processamento de sinais de eletroencefalograma, que, de forma similar a sinais de voz, também é difícil de ser caracterizado. Vários desses atributos estão relacionados com a teoria do caos, um ramo que tem como objetivo o entendimento do comportamento de sistemas dinâmicos sensíveis às condições iniciais. O processo de produção da voz pode ser interpretado desta forma (WANG et al., 2020).

A análise de flutuação sem tendência (DFA, do inglês Detrended Fluctuation Analysis) caracteriza a auto-similaridade de um sinal. Quando tratamos de ondas sonoras, informações valiosas sobre vibrações vocais de baixa frequência, associadas aos ruídos respiratórios, tornam-se acessíveis através da DFA, permitindo análises em busca de patologias (LITTLE et al., 2007b)

A entropia da densidade de probabilidade do período de recorrência (RDPE, do inglês Recurrence Density Period Entropy) calcula a incerteza média de um determinado valor em uma densidade de probabilidade discreta. Considerando que um sinal de voz processado digitalmente pode ser visto como uma densidade de probabilidade, a RDPE se mostra uma ferramenta relevante na avaliação de distúrbios vocais. Ela permite representar a incerteza média do período do sinal, auxiliando na detecção de possíveis irregularidades (LITTLE et al., 2007b).

Assim como a RDPE, a entropia do período do tom (PPE, do inglês Pitch Period Entropy) também simboliza uma entropia - ou seja, a medida da incerteza média em uma distribuição de probabilidade. Enquanto na RDPE a entropia reflete a incerteza da distribuição de probabilidade do período de recorrência, na PPE é calculada a entropia da distribuição de probabilidade do período do tom (ou seja, a percepção sensorial da F0) (OZKAN, 2016).

Os expoentes de Lyapunov quantificam a divergência entre 2 sistemas como parâmetros iniciais semelhantes. O maior expoente caracteriza o caos num sistema, enquanto que o seu inverso é denominado tempo de Lyapunov e é utilizado para definir por quanto tempo o comportamento de um sistema pode ser previsto (WANG et al., 2020).

A dimensão fractal também é usada na análise de sistemas dinâmicos. Ela representa a razão da variação logarítmica em detalhe para a variação logarítmica em escala de um sinal (WANG et al., 2020). A dimensão fractal está relacionada a complexidade de um sinal e foi demonstrado que a análise da quantificação de recorrência determinística de balanço (que é similar a dimensão fractal) pode diferenciar portadores da DP de grupos de controle (SCHMIT et al., 2006).

2.2.4 Representações visuais de sinais de voz

Os sinais sonoros podem ser representados como uma combinação de senóides. Dessa forma, o sinal mais simples que pode ser representado é exatamente uma senóide. Uma forma de representar visualmente um sinal é através de um gráfico de amplitude pelo tempo. Uma forma alternativa de representá-los é utilizando a transformada de Fourier, que extrai as frequências que compõem um sinal. Essas frequências nada mais são do que o inverso do período das senóides que compõem o sinal. A Figura 2 apresenta um sinal no domínio do tempo e o mesmo sinal no domínio da frequência. Como o sinal é uma senóide simples com frequência de 5Hz e amplitude unitária, sua representação no domínio da frequência é composto por um único ponto com frequência também igual a 5Hz e amplitude unitária. Na teoria, o sinal no domínio da frequência deveria ser representado por uma única reta no ponto de 5Hz com altura unitária. Na prática, tem-se esse formato triângular devido a imperfeições na representação da senóide. Uma senóide é composta por infinitos pontos, o que é impossível de se representar computacionalmente. A senóide apresentada foi construída com 1000 pontos para cada segundo.

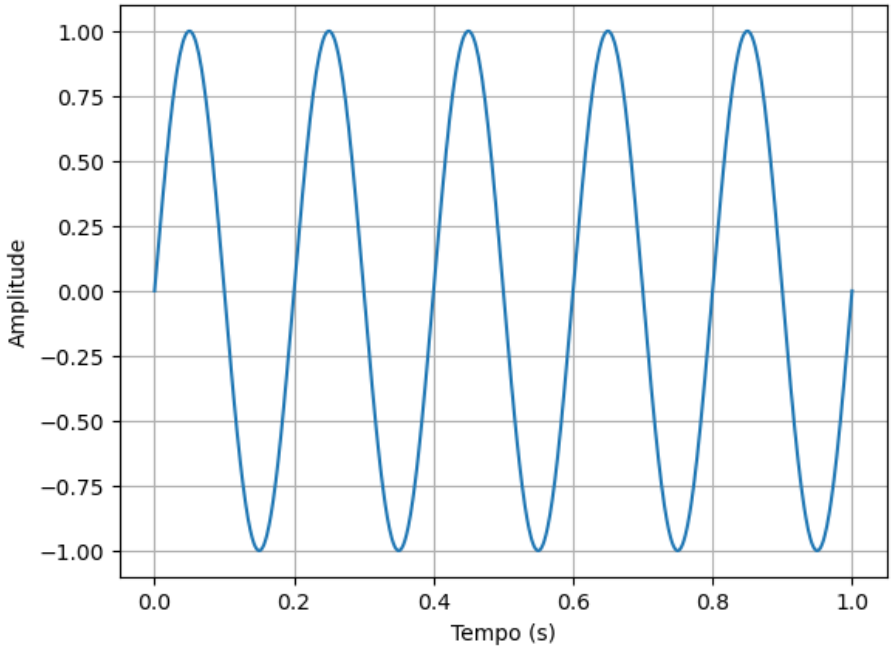
A medida que o número de senóides que compõem um sinal aumenta, ele se torna mais complexo no domínio do tempo e a sua análise no domínio da frequência se torna mais vantajosa. A Figura 3 apresenta um sinal composto pela soma de 3 senóides com frequências de 5Hz, 37Hz e 17Hz. Suas amplitudes são respectivamente 1, 0.5 e 2. Para cada nova senóide, um novo ponto é adicionado no domínio da frequência. A Figura 4 apresenta as 3 senóides que compõem o sinal.

Uma maneira alternativa de caracterizar o sinal de voz e apresentar a informação associada aos sons é através da representação espectral. O espectrograma de um sinal de voz é uma representação em três dimensões da intensidade da voz, em diferentes bandas de frequência, sobre o tempo. A forma mais usual de obtenção do espectrograma é através da Transformada de Fourier de Tempo Curto (STFT, do inglês Short-Time Fourier Transform) (LOPES, 2013).

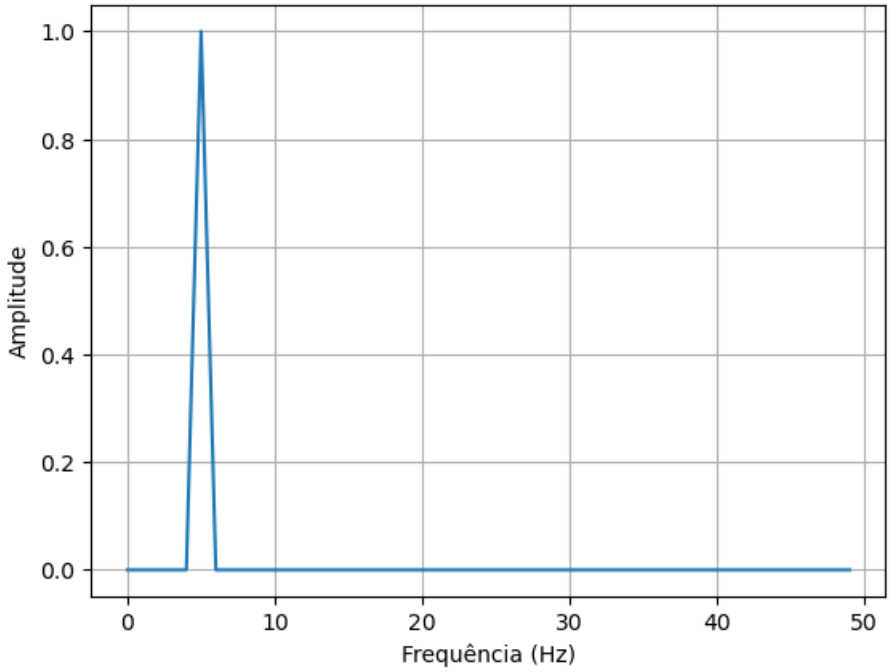
A SFTF funciona por meio do janelamento de um sinal. A Figura 5 apresenta um espectrograma extraído para o mesmo sinal apresentado na Figura 3. Para gerar o espectrograma, uma janela de duração de um segundo foi definida, iniciando-se no instante inicial (0 segundos). Essa janela forma um “sub-sinal” para o qual é aplicado a transformada de Fourier. Como o sinal é composto por apenas 3 senóides, apenas 3 frequências são extraídas: 5Hz, 37Hz e 17Hz. Em seguida, a janela foi deslocada 1 segundo para a direita. Ou seja, o instante inicial da janela

Figura 2 – Sinal composto por uma única senóide

(a) Domínio do tempo



(b) Domínio da frequência

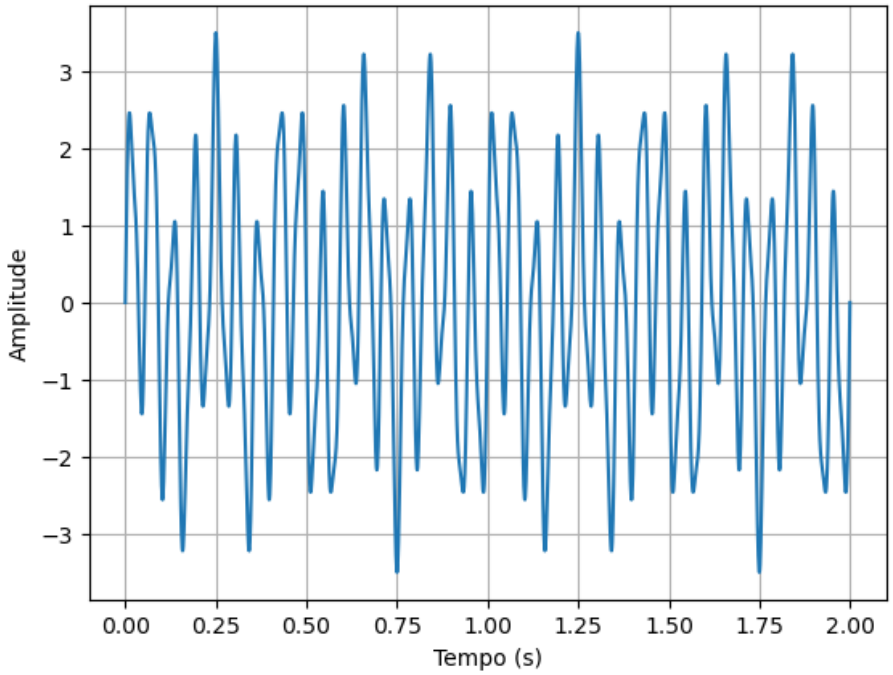


Fonte: Autoria Própria

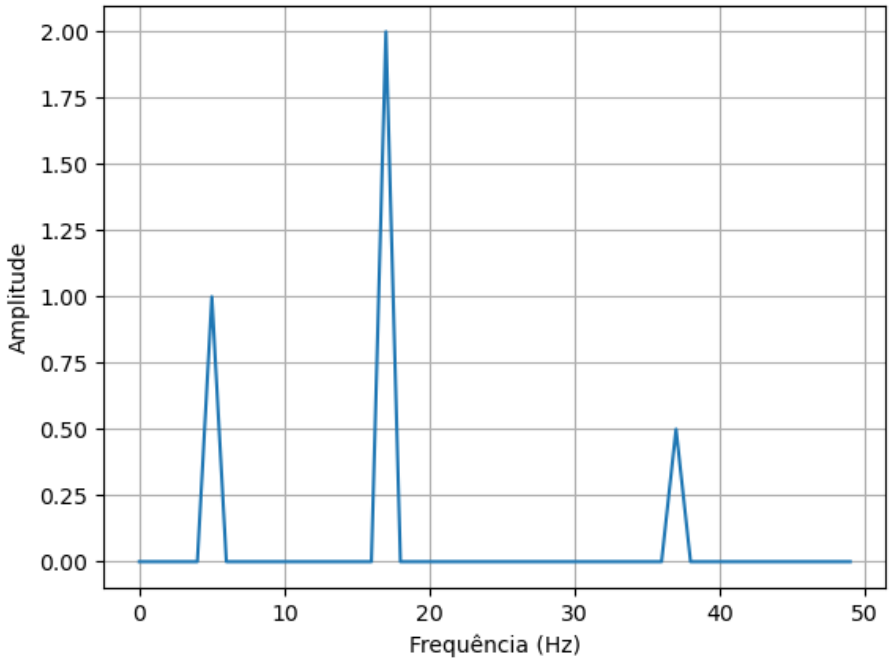
seria 1 segundo e o instante final 2 segundos. Aplica-se novamente a transformada e obtém-se os mesmos resultados. O eixo X do gráfico representa o instante inicial no qual a transformada foi aplicada, o eixo Y representa as frequências e as cores representam a amplitude de uma

Figura 3 – Sinal composto por três senóides

(a) Domínio do tempo



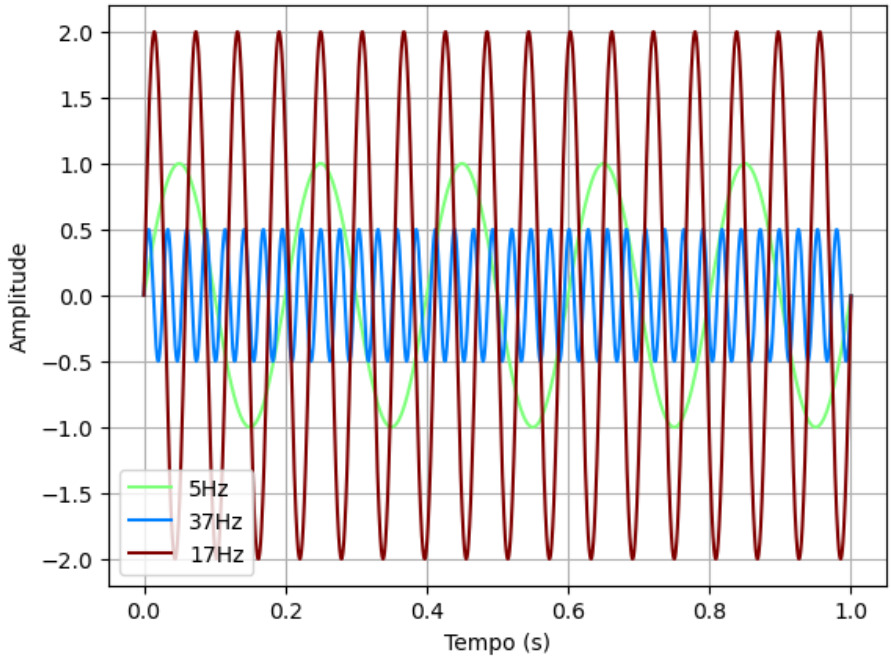
(b) Domínio da frequência



Fonte: Autoria Própria

determinada frequência em um determinado instante de tempo. Como o sinal possui apenas 3 componentes (senóides), observa-se 3 faixas no espectrograma na altura de suas frequências correspondentes. A faixa na altura de 17Hz é mais próxima de um vermelho escuro por ser a

Figura 4 – Decomposição de um sinal composto por três senóides



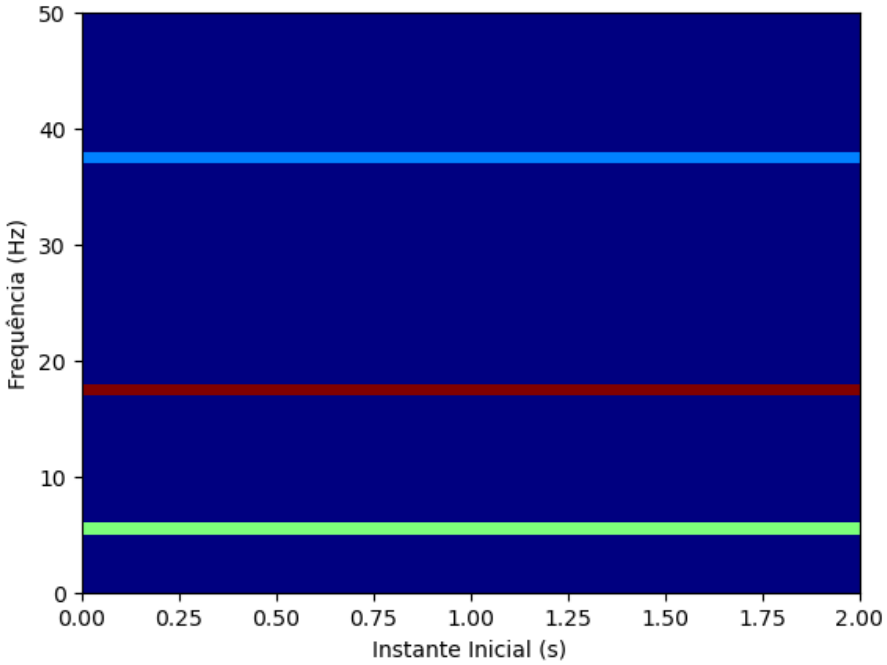
Fonte: Autoria própria

de maior amplitude. A faixa na altura de 37 Hz é mais próxima de um azul escuro por ser a de menor amplitude. O fundo em azul escuro pode ser interpretado como um "silêncio" das outras frequências.

Um sinal proveniente da voz humana é muito mais complexo, sendo composto por infinitas senóides com diferentes frequências. Na prática, não se observa os "silêncios" como os apresentados na Figura 5. A Figura 6 apresenta um sinal de voz proveniente de uma fonação sustentada da vogal "a" e seu espectrograma. O sinal de voz tem duração de 0.1 segundos e o espectrograma foi extraído como uma janela de 0.01 segundos. Observa-se que para as frequências mais baixas, o sinal é mais forte, ou seja, a amplitude das senóides de frequência mais baixa que compõem o sinal é mais elevada que a amplitude das senóides de alta frequência.

Uma vez que também é possível representar um sinal através da soma de funções de cosseno oscilando em diferentes frequências, uma outra forma de se obter um espectrograma é através da Transformada Discreta de Cossenos (DCT, do inglês Discrete Cosine Transformation). A principal diferença entra a DCT e a Tranformada Discreta de Fourier (DFT, do inglês Discrete Fourier Transform), que é utilizada no cálculo da TSFT, é que a DCT contém apenas as componentes reais do sinal. Um tipo diferente de espectrograma pode ser obtido através da aplicação de DCTs em curtos períodos de tempo do sinal, técnica análoga ao cálculo da STFT, seguida da obtenção das frequências de Mel para cada um desses períodos (KARAMAN et al., 2021). A Figura 7 apresenta esse tipo de espectrograma para o mesmo sinal de voz apresentado na Figura

Figura 5 – Espectrograma de um sinal composto por três senóides



Fonte: Autoria própria

6, utilizando o mesmo janelamento. Foram extraídas 320 frequências de Mel.

2.3 Aprendizado de máquina

Esta seção aborda, de forma geral, a área de Aprendizado de Máquina, sendo fundamental para a compreensão do trabalho. São apresentados os principais conceitos e métricas de avaliação de performance e alguns dos principais modelos, técnicas e arquiteturas de Aprendizado de Máquina.

2.3.1 Principais conceitos

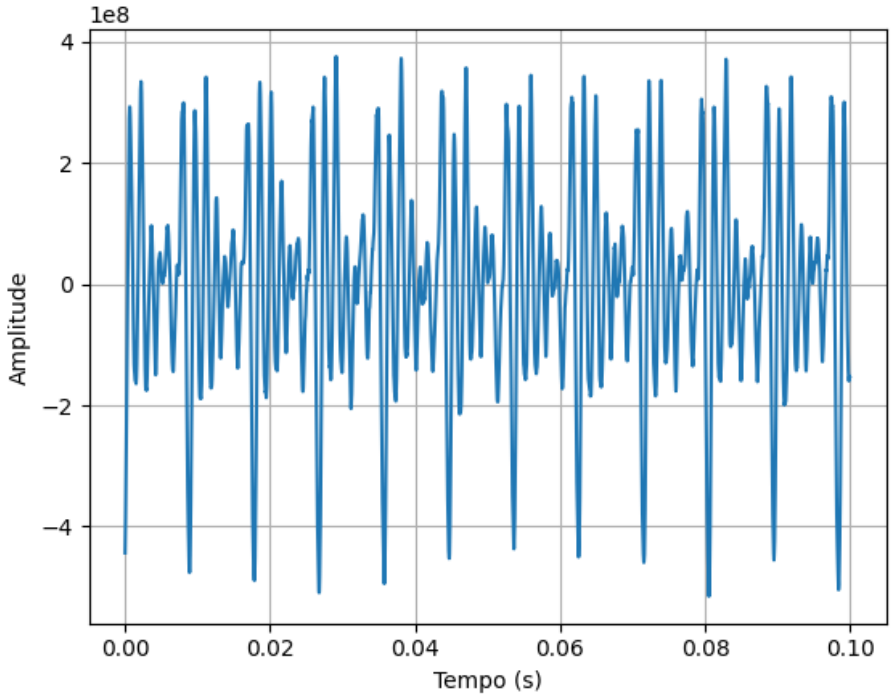
Aprendizado de Máquina (ML, do inglês *Machine Learning*), refere-se a um conjunto de abordagens computacionais que utilizam experiências passadas para aprimorar seu desempenho e/ou fazer previsões mais precisas. Nesse contexto, experiência se traduz na habilidade de utilizar informações prévias para o aprendizado (MOHRI; RASTAMIZADEH; TALWALKAR, 2018).

De maneira comum, ML tem como propósito fazer previsões de uma medida específica de resultado, frequentemente quantitativa ou categórica. Essas previsões são fundamentadas em um conjunto de dados conhecidos, chamados de atributos, os quais estão relacionados a essa medida em questão (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

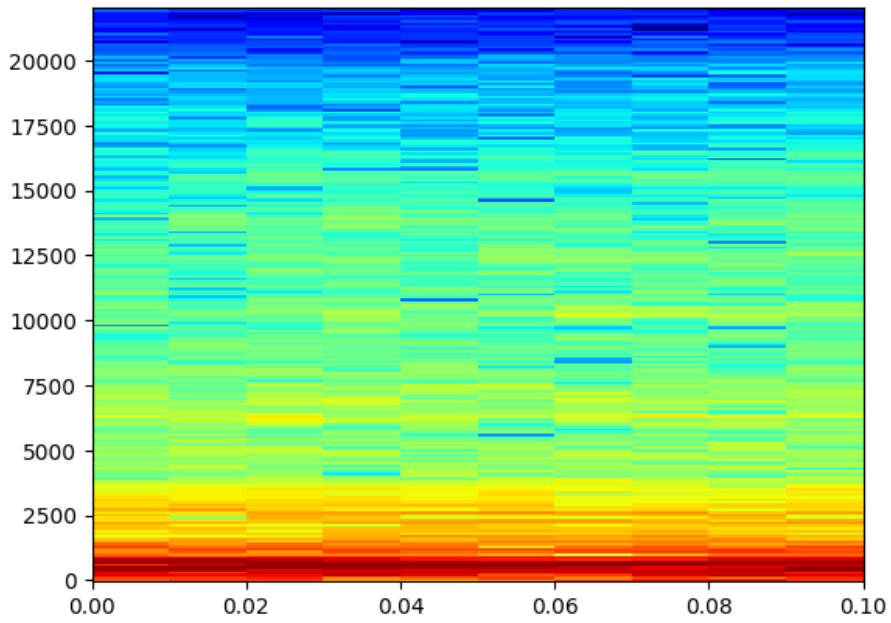
O aprendizado é classificado como supervisionado quando a medida de resultado é

Figura 6 – Sinal de voz

(a) Domínio do tempo



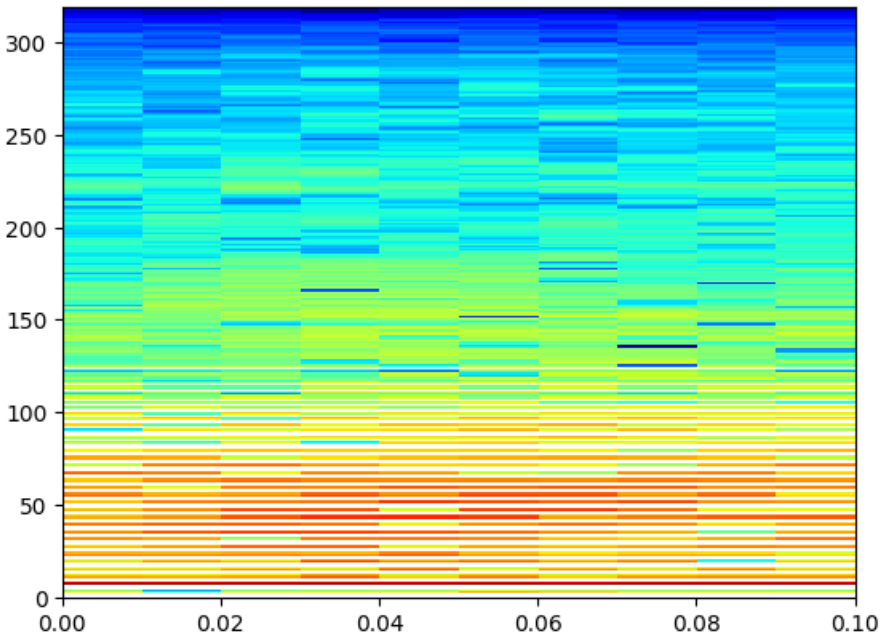
(b) Espectrograma



Fonte: Autoria Própria

conhecida para um conjunto de instâncias, juntamente com seus atributos associados. Por outro lado, se a medida de resultado não é conhecida, o aprendizado é considerado não supervisionado. Além dessas categorias, a Tabela 6 apresenta outras divisões para os problemas de ML.

Figura 7 – Espectrograma - frequências de mel



Fonte: Autoria própria

Tabela 6 – Categoria de problemas de ML

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.

Fonte: Mohri, Rastamizadeh e Talwalkar (2018)

Considerando X um conjunto de atributos que podem ser utilizados para prever um resultado Y e partindo-se da hipótese de que os erros ε sejam aditivos e que o modelo $Y = f(X) + \varepsilon$ seja razoável, tem-se que f representa a informação sistemática que X provém para Y. O aprendizado supervisionado é um conjunto de abordagens para se estimar f (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

Para verificar a capacidade de generalização de um modelo, utiliza-se a validação cruzada. O método mais tadicional é o *hold out*, que consiste em dividir a base de dados em três

subconjuntos distintos e sem sobreposição de elementos entre eles. O primeiro subconjunto contém as amostras de treinamento, que é utilizado no processo de aprendizagem do modelo. Esse processo visa ajustar os parâmetros do modelo com base nos dados fornecidos, para que ele possa fazer previsões ou classificações mais precisas em novos dados não vistos antes. Os outros dois subconjuntos são as amostras de validação e as amostras de teste, que são utilizadas para avaliar o desempenho do modelo e verificar sua capacidade de generalização em dados desconhecidos. As amostras de validação são usadas para ajustar hiperparâmetros do modelo, enquanto que as amostras de teste são reservadas para medir a eficácia final do modelo em dados nunca antes vistos, fornecendo uma estimativa mais realista de seu desempenho em aplicações do mundo real. Essa divisão em três subconjuntos proporciona uma avaliação imparcial e adequada do modelo de ML (MOHRI; RASTAMIZADEH; TALWALKAR, 2018).

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 sem sobreposição de elementos e de tamanho aproximado. O treino é efetuado em k-1 subconjuntos e a validação é feita 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; TIBSHIRANI; FRIEDMAN, 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 métricas colhidas na etapa de validação (WONG, 2015).

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 métrica colhida pode não ser realista (HASTIE; TIBSHIRANI; FRIEDMAN, 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; TIBSHIRANI; FRIEDMAN, 2008).

2.3.2 Principais métricas

A matriz de confusão, como seu nome sugere, é uma ferramenta que mostra quais classes foram confundidas durante a classificação. Ela provém informação em um nível bem mais detalhado do que outras métricas, como a acurácia ou o erro (SUSMAGA, 2004). A Tabela 7 apresenta uma matriz de confusão de um problema com duas classes. O rótulo Verdadeiro Positivo (TP, do inglês *True Positive*) representa a quantidade de instâncias que foram classificadas corretamente como positivas, enquanto que Falso Positivo (FP, do inglês *False Positive*)

representa a quantidade de instâncias que foram classificadas erroneamente como positivas. Os rótulos Falso Negativo (FN, do inglês *False Negative*) e Verdadeiro Negativo (TN, do inglês *True Negative*) apresentam um comportamento análogo.

Tabela 7 – Matriz de Confusão

		Classe Prevista	
		Positiva	Negativa
Classe Real	Positiva	Verdadeiro Positivo	Falso Negativo
	Negativa	Falso Positivo	Verdadeiro Negativo

A acurácia é a medida mais comumente utilizada na avaliação de modelos de ML. Essa métrica é calculada considerando que todas as instâncias do conjunto de dados têm o mesmo peso. Para algoritmos de classificação, a acurácia é definida como o número de previsões corretas realizadas dividido pelo número total de instâncias no conjunto de dados (WONG, 2015). Em outras palavras, é a proporção de predições corretas em relação ao tamanho total do conjunto de dados, e representa a capacidade geral do modelo em fazer previsões corretas. Ela pode ser obtida por:

$$\text{Acurácia} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Além da acurácia, outras medidas importantes podem ser utilizadas na avaliação de modelos de Aprendizado de Máquina, especialmente em diagnóstico de patologias. Três dessas medidas são a sensibilidade (também conhecida como *recall*), que é a probabilidade do diagnóstico ser positivo quando o paciente realmente possui a doença, a especificidade, que é a probabilidade do diagnóstico ser negativo quando o paciente não possui a doença e precisão, que é a proporção de verdadeiros positivos em relação ao total de previsões positivas (CAFFO, 2016). Elas podem ser obtidas por:

$$\text{Sensibilidade} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{Especificidade} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{Precisão} = \frac{TP}{TP+FP} \tag{4}$$

Sensibilidade e especificidade estão intimamente associadas com a 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:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D)+[1-P(-|D^C)][1-P(D)]} \tag{5}$$

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 problemas de classificação, é frequente encontrarmos conjuntos de dados com classes desbalanceadas. Nesses casos, a acurácia pode ser uma medida de avaliação enganosa. Por exemplo, um modelo que sempre prediz o rótulo A em uma base em que 90% das instâncias possuem esse rótulo teria uma acurácia de 90%.

Por sua vez, a estatística kappa (κ) é uma medida de avaliação que pode ser útil nos casos de desbalanceamento de classes. A fórmula a seguir apresenta seu cálculo para um problema de classificação binária utilizando as métricas da matriz de confusão.

$$\kappa = \frac{2X(TPXTN - FNXFP)}{(TP+FP)X(FP+TN)+(TP+FN)X(FN+TP)} \tag{6}$$

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 8.

Tabela 8 – Valor de Kappa e nível de concordância.

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

Fonte: (LANDIS; KOCH, 1977)

Outra métrica útil em casos de desbalanceamento é o F-score, também conhecido como F1-score, que combina as noções de precisão e sensibilidade em uma única pontuação, fornecendo uma medida geral do desempenho do modelo (POLAT; GüNEş, 2009). O F-score é calculado como a média harmônica entre a precisão P e a sensibilicade S, conforme a fórmula a seguir.

$$\kappa = \frac{2PS}{P+S} \tag{7}$$

2.3.3 Árvores de decisão e random forest

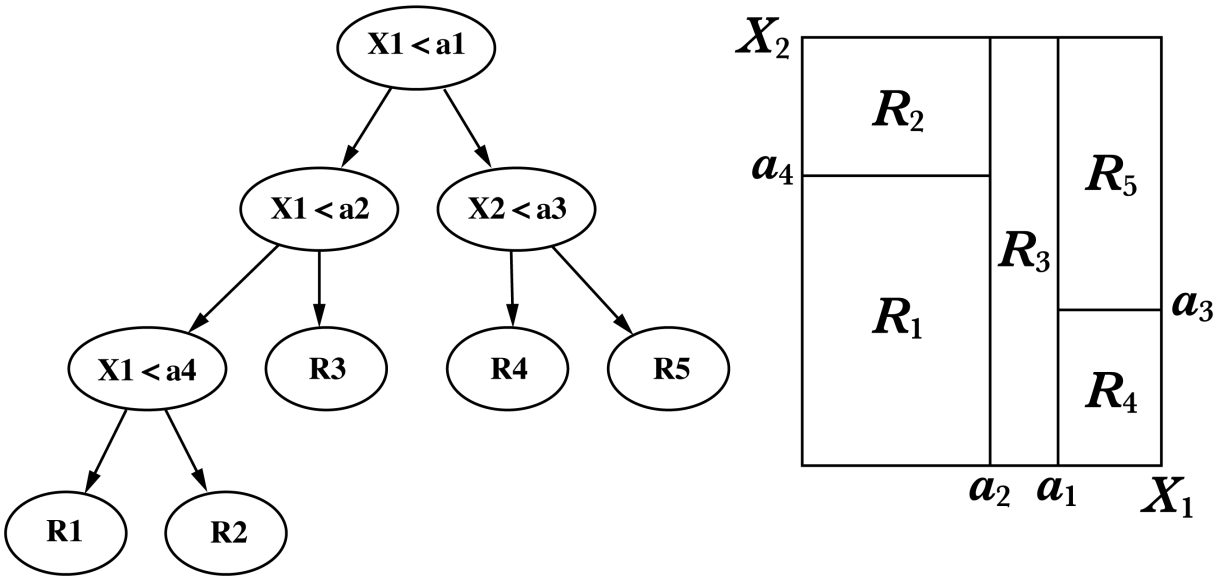
As árvores de decisão (DT, do inglês *Decision Trees*) são modelos de aprendizado rápidos que são frequentemente utilizados em problemas de classificação, apesar de também poderem ser

aplicados em tarefas de regressão e clusterização. Uma das principais vantagens das árvores de decisão é a sua interpretabilidade. Ou seja, esses modelos podem ser facilmente compreendidos e visualizados, permitindo que os resultados e decisões sejam explicados de forma intuitiva.

A árvore de decisão binária (BDT, do inglês *Binary Decision Tree*) é um tipo específico de árvore de decisão que se caracteriza por ter apenas dois ramos em cada nó interno. Esses ramos representam as duas opções de resposta possíveis para o teste aplicado no nó. Em outras palavras, cada nó interno da árvore de decisão binária apresenta uma pergunta sobre uma determinada característica dos dados, e as respostas possíveis são geralmente "sim"ou "não", "verdadeiro"ou "falso", ou outras opções binárias relevantes para o problema em questão (MOHRI; RASTAMIZADEH; TALWALKAR, 2018).

A Figura 8 apresenta um exemplo de BDT aplicada a um espaço bidimensional com 2 atributos numéricos, X1 e X2. Cada nó corresponde a uma pergunta númerica relacionada a esses atributos com a exceção dos nós folha que trazem valores categóricos indicativos dos rótulos correpondentes.

Figura 8 – Exemplo prático de uma BDT



Fonte: Mohri, Rastamizadeh e Talwalkar (2018)

O processo de previsão das BDTs é simples. Dada uma instância X, começa-se pela raiz e em cada nó, os atributos de x são utilizados para responder a pergunta correspondente. Em caso positivo, parte-se para o nó da direita. Caso contrário, nó da esquerda. No fim, chega-se a uma folha com o rótulo indicativo.

O processo de aprendizado de uma BDT é complexo. Trata-se de um problema NP-difícil, ou seja, até o momento, não há solução em tempo polinomial para ele e suspeita-se de que não haja uma, embora ainda não tenha sido provado. Para contornar este problema, é possível

a utilização de técnicas de algoritmos gulosos. Um algoritmo é dito guloso se eles agirem de acordo com o princípio de que tomando-se uma decisão local ótima a cada estágio do problema, chega-se a uma decisão global ótima (AREFIN, 2006). Os algoritmos gulosos são construídos em cima de 5 pilares que estão apresentados na Tabela 9.

Tabela 9 – Pilares dos algoritmos gulosos

Pilar	Descrição
Conjunto de candidatos	Conjunto em cima da qual a solução é criada.
Função de seleção	Função que seleciona melhor candidato a ser adicionado na solução.
Função de viabilidade	Função que determina se o candidato pode ser usado na solução.
Função objetiva	Função que associa um valor para a solução final ou parcial.
Função de solução	Função que indica que uma solução foi encontrada.

Fonte: Arefin (2006)

Importante ressaltar que um algoritmo guloso não necessariamente leva a uma solução ótima, uma vez que ele não opera exaustivamente sobre os dados, analisando todas as possibilidades. Isto é compensado pela sua velocidade. Se é provado que um algoritmo guloso leva a solução ótima, ele se torna, tipicamente, o método escolhido (AREFIN, 2006).

O método guloso de aprendizado da BDT consiste em iniciar com um único nó que recebe o rótulo majoritário nas instâncias da amostra. Em seguida, em cada estágio, um nó n_t é dividido de acordo com uma pergunta q_t. O par (n_t,q_t) é escolhido de forma a minimizar a impureza nodal medida por uma função F (MOHRI; RASTAMIZADEH; TALWALKAR, 2018).

A impureza de um nó n é dada por $F(n)$. Seja $n_+(n,q)$ o nó filho da direita e $n_-(n,q)$ o filho da esquerda da divisão de n e $\eta(n,q)$ a fração de pontos na região definida por n movida para $n_-(n,q)$, temos que o decaimento de impureza é dado por:

$$\tilde{F}(n,q) = F(n) - [\eta(n,q)F(n_-(n,q)) + (1-\eta(n,q))F(n_+(n,q)] \tag{8}$$

Para todo nó n e rótulo $l \in [1,k]$, seja $p_l(n)$ a fração do número de pontos na região de n que pertencem à classe l, as 3 funções mais utilizadas como medida de impureza nodal são erro de classificação (EC), entropia (ENT) e índice de gini (IG) (MOHRI; RASTAMIZADEH; TALWALKAR, 2018), que podem ser calculadas da seguinte forma:

$$F(n) = \begin{cases} 1 - max_{l\in[1,k]} p_l(n) & EC \\ -\sum_{l=1}^{k} p_l(n) log_2 p_l(n) & ENT \\ \sum_{l=1}^{k} p_l(n)(1-p_l(n)) & IG. \end{cases} \tag{9}$$

sendo a entropia e o índice de Gini as mais utilizadas.

Como dito anteriormente, as BDTs são considerados classificadores fracos. Isto significa que seu desempenho não é muito superior ao da escolha aleatória de rótulos para cada instância. A fim de obter uma performance superior com classificadores fracos, pode-se utilizar o método de *boosting*.

Boosting é uma técnica de ML que busca melhorar a performance de um modelo combinando várias versões fracas desse modelo em um modelo mais forte e preciso. O objetivo do *boosting* é reduzir a variância do modelo, aumentando a sua capacidade de generalização (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). O processo de funcionamento do *boosting* é dividido em 4 etapas: Inicialização, ponderação dos dados, treinamento em etapas e agregação.

Na inicialização, o algoritmo começa a partir de um modelo fraco treinado em uma parte dos dados de treinamento. Na ponderação dos dados, O algoritmo atribui pesos iniciais a cada instância do conjunto de treinamento, dando mais peso às instâncias que foram classificadas incorretamente pelo modelo fraco anterior. No treinamento em etapas, o modelo fraco é treinado em várias iterações (etapas). A cada iteração, o modelo fraco é treinado para se concentrar mais nas instâncias classificadas incorretamente nas etapas anteriores, ajustando-se assim a esses erros. Por fim, na agregação, o modelo fraco é adicionado à combinação final em cada iteração, e sua contribuição para a previsão final é ponderada de acordo com sua performance. Esse processo é repetido até que um número pré-definido de iterações seja atingido ou até que a precisão do modelo seja considerada aceitável.

Bagging, por sua vez, é uma técnica que pode ser usada por diferentes métodos de classificação e regressão para reduzir a variância associada à previsão e, assim, melhorar o processo de previsão. Ele faz uso de uma técnica de bootstrap, que consiste em reamostragem de um conjunto de dados com reposição de elementos (HESTERBERG, 2011). Bagging é uma ideia relativamente simples: muitas amostras de bootstrap são extraídas dos dados disponíveis, algum método de previsão é aplicado a cada amostra de bootstrap e, em seguida, os resultados são combinados, calculando a média para regressão e votação simples para classificação, para obter a previsão geral. Dessa forma, a variância é reduzida e os modelos tendem a ter o *overfitting* também reduzido (SUTTON, 2005).

Por fim, *Random Forest* ou floresta aleatória, como o nome sugere, consiste na combinação de BDTs que operam em conjunto formando um classificador mais forte do que elas seriam individualmente (YIU, 2019). Ou seja, a floresta aleatória pode ser obtida através da aplicação de uma técnica de bagging em uma BDT.

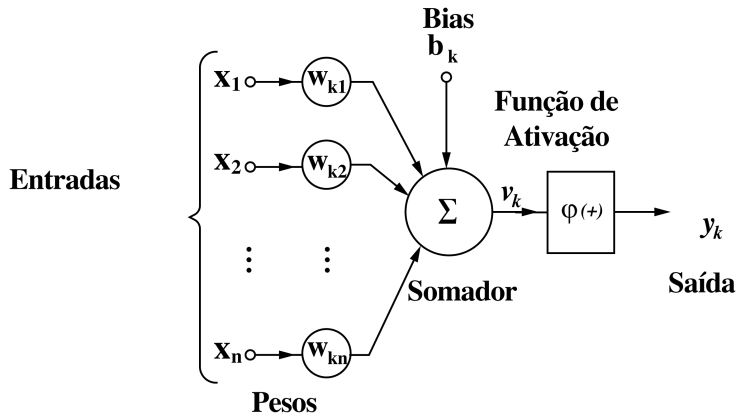
2.3.4 Redes neurais

A Rede Neural (NN, do inglês Neural Network) é um dos modelos de ML mais utilizados. Ela apresenta uma arquitetura profundamente inspirada nos processos biológicos, manifestando uma habilidade intrínseca de ser treinada e de internalizar representações que permanecem inalteradas diante de mudanças de escala, translação, rotação e metamorfoses afins. A estrutura

da NN é construída a partir de elementos computacionais simples, denominados neurônios, interligados de maneira altamente entrelaçada, criando uma configuração distribuída em paralelo capaz de armazenar, bem como processar, informações com o propósito de cumprir uma tarefa específica (SOUZA et al., 2020).

A Figura 9 apresenta modelo base de um neurônio de uma NN. Ele tem como entrada um vetor *m* dimensional *x*. Suas componenente são combinadas em uma soma ponderada definida por pesos *wk*. Para possibilitar ajustes nos valores de saída do neurônio utiliza-se o limiar de ativação inerente (Bias), introduzido como um grau de liberdade adicional para manipular a saída do neurônio. O resultado da soma entre o limiar de ativação e as entradas multiplicadas pelos pesos produz o potencial de ativação *vk*. A saída é indicada pelo vetor *yk*, que corresponde ao potencial de ativação transformado pela função de ativação $\phi(.)$. Essa função é normalmente não-linear, como a ReLU, que é dada por $f(x) = max(x, 0)$, ou uma função do tipo softmax, que mapeia a saída em um intervalo entre 0 e 1 (SILVA, 2023).

Figura 9 – Modelo base de um neurônio

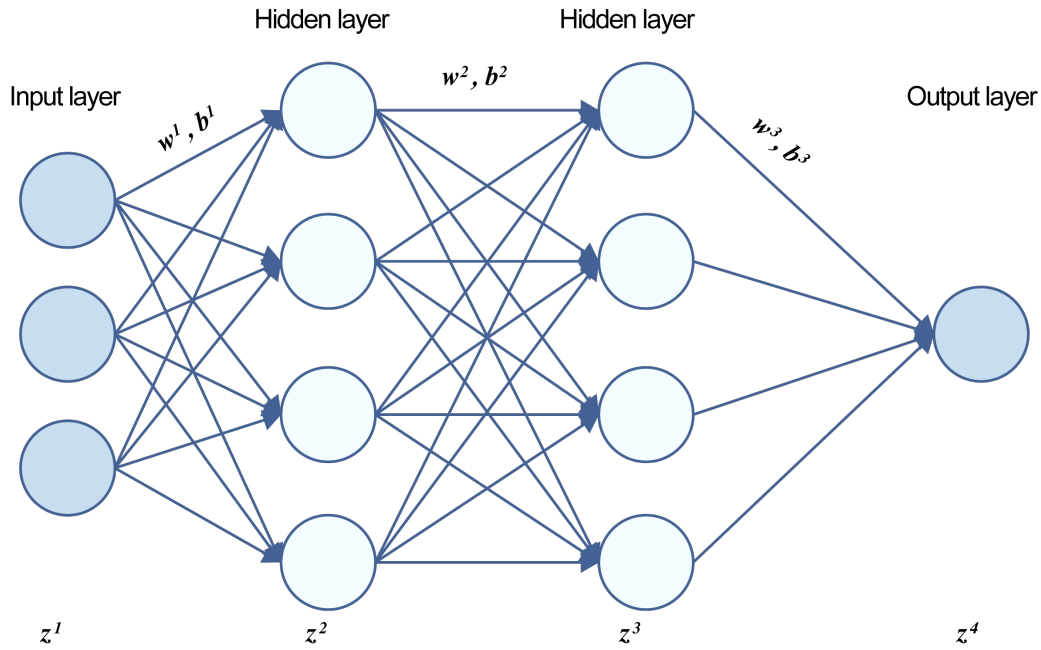


Fonte: Silva (2023)

Uma NN é composta por camadas que estão encadeadas e mapeiam os dados de entrada para fazer predições de acordo com as classes definidas. Cada camada de entrada da rede contém um conjunto de neurônios. Como as camadas da NN estão encadeadas, os valores de saída das camadas iniciais tornam-se os valores de entrada das camadas posteriores. Dessa forma, o aprendizado do modelo depende da interação entre as camadas e ocorre com o ajuste dos pesos entre os neurônios e dos limiares de ativação ao longo da rede. Esses parâmetros são alterados durante a fase de treinamento e os pesos são utilizados para calcular a taxa de crescimento da função que o algoritmo tenta modelar. Os limiares de ativação também são necessários para deslocar a saída da função (RUAN et al., 2019).

A Figura 10 apresenta o modelo de uma NN, no qual cada círculo representa um neurônio artificial da rede, o conjunto inicial é a camada de entrada, os conjuntos ao centro representam as camadas intermediárias ou ocultas e a esfera final representa a camada de saída.

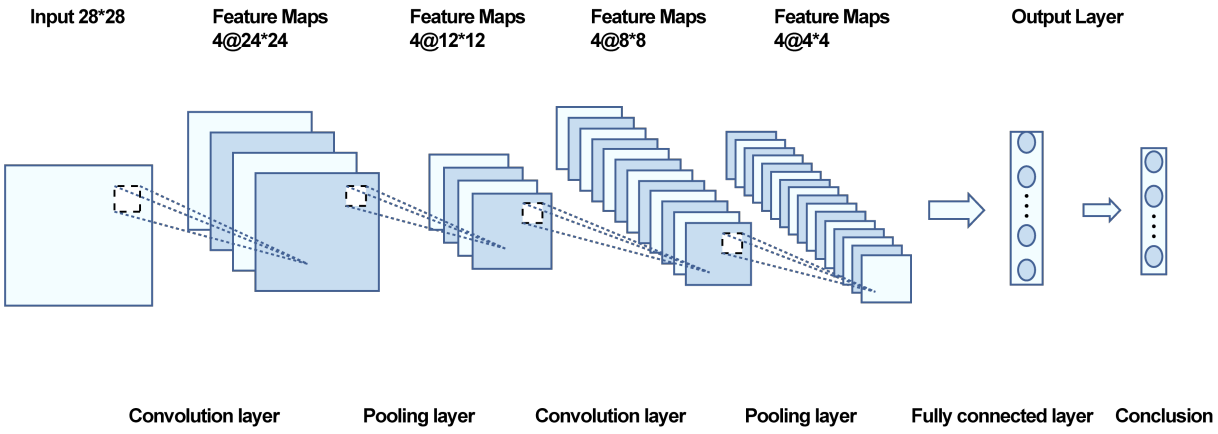
Figura 10 – Modelo básico de uma NN



Fonte: Silva (2023)

A Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*) é um tipo especial de NN. Sua estrutura apresentada na Figura 11, onde é possível notar a arquitetura base que consiste nas seguintes camadas: Camada Convolucional (*Convolucional Layer*), Camada de Agrupamento (*Pooling Layer*), Camada Totalmente Conectada ou Densa (*Fully Connected Layer*) e as camadas de entrada e saída de dados. Ela é amplamente utilizada no contexto de classificação de imagens.

Figura 11 – Modelo básico de uma CNN

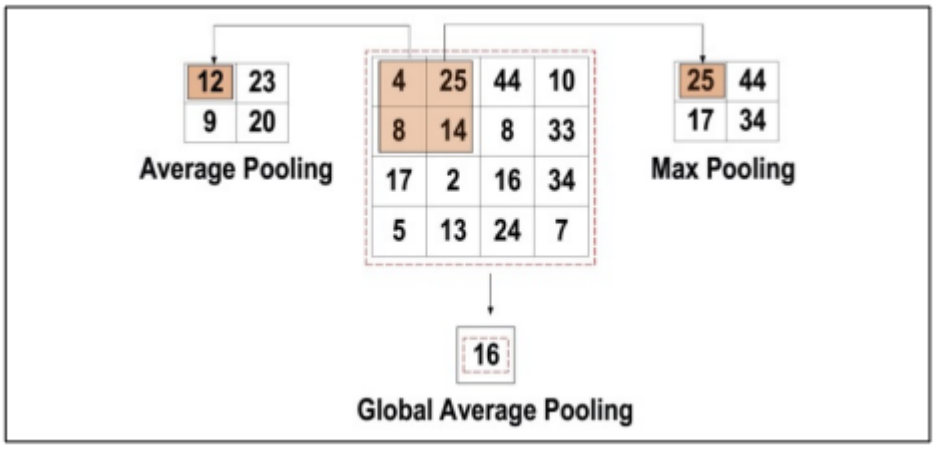


Fonte: Silva (2023)

Dentro da estrutura de uma CNN, o elemento de maior destaque é a camada convolucional. Essa camada é composta por um conjunto de filtros convolucionais, também referidos como *kernels*. A imagem de entrada, representada como métricas de N dimensões, passa por um processo de convolução com esses filtros. Cada filtro é, essencialmente, uma matriz de pesos. A convolução realizada em uma CNN tem como principal objetivo extrair características da imagem original, resultando na criação de um mapa de características de saída (ALZUBAIDI et al., 2021).

Outro elemento de destaque é a camada de agrupamento. Ela desempenha um papel central na redução da dimensionalidade dos mapas de características que são gerados a partir das camadas convolucionais. Em essência, essa etapa realiza uma amostragem sub-representativa dos mapas de características mais abrangentes, resultando na formação de mapas de características mais compactos, mas ainda retendo as informações de maior relevância. Uma variedade de métodos são empregados na execução dessa operação de redução de dimensionalidade, dentre os principais, temos o agrupamento por valor médio (*Average Pooling*), por valor máximo (*Max Pooling*) e média global (*Global Average Pooling*), conforme ilustrados na Figura 12.

Figura 12 – Principais operações de *pooling*



Fonte: Alzubaidi et al. (2021)

Por último, a camada que conclui a estrutura fundamental de uma CNN é a camada completamente conectada. Geralmente posicionada como a camada final, cada neurônio nessa camada está conectado a todos os neurônios da camada precedente. Essa camada é empregada como o componente classificador da CNN (ALZUBAIDI et al., 2021).

2.3.5 Análise de componentes principais

A análise de componentes principais (PCA, do inglês *Principal Component Analysis*) é a técnica de redução de dimensão mais empregada na exploração e análise de dados. Ela é particularmente utilizada em conjuntos de dados que possuem múltiplos atributos altamente

correlacionados, tendo como objetivo a representação dos dados originais num subespaço com menos dimensões sem que a perda de informação seja elevada. Possui como vantagens uma menor necessidade de espaço de armazenamento, remoção de colinearidade e redução de ruído, sendo esses 2 últimos especialmente úteis no contexto da maioria dos algoritmos de ML (KHERIF; LATYPOVA, 2020).

Dada uma matriz X que representa um conjunto de dados com n instâncias de p atributos, a obtenção dos componentes principais é realizada por meio da diagolaziação de matrizes simétricas positivas semi-definidas. Isso é feito calculando-se a matriz de covariância $\sum$. Com ela, é possível obter os pares de autovalores e autovetores $(\lambda_1, e_1), (\lambda_1, e_1), ..., (\lambda_p, e_p)$. Esses pares podem ser utilizados para calcular as componentes (HONGYU; SANDANIELO; JUNIOR, 2016). A i-ésima componente Z é dada por:

$$Z_i = \sum_{n=1}^{p} e_{in} X_n \tag{10}$$

Os dados são projetados nas componentes principais obtidas, resultando em uma representação de dimensionalidade reduzida dos dados. Quanto mais componentes forem utilizadas, mais próximos se estará do conjunto original. Dessa forma, o número máximo de componentes será sempre limitado pela quantidade de atributos do conjunto original.

2.3.6 Comitê de classificadores

Ensemble learning é um termo genérico utilizado para descrever métodos que combinam diferentes algoritmos de ML para construir um modelo mais robusto e capaz de aprendizado em ambientes mais complexos, combinando as melhores características de cada modelo (SAGI; ROKACH, 2018).

Modelos de *ensemble learning* exploram múltiplos métodos de ML que produzem resultados preditivos fracos, mesclando seus resultados com mecanismos de "eleição" para obter uma melhor performance. De forma geral, os métodos de ML utilizados trabalham com os atributos sob diferentes perspectivas ou projeções (DONG et al., 2020).

No caso específico do problema de classificação de sinais de voz de portadores da DP, os espectrogramas e os atributos de áudio podem ser tratados como projeções diferentes de um mesmo conjunto de dados. Assim, um modelo que fizesse uso de um método que trabalhasse com os atributos de áudio e de um outro método que trabalhasse com os espectrogramas poderia ser definido como um modelo de *ensemble learning*.

O teorema do jury de Condorcet é considerado o precursor dos métodos de *ensemble learning*. Ele foi apresentado por Marie Jean Antoine Nicolas de Caritat, matemático e marquês de Condorcet, em 1785. O teorema se refere ao processo de votação de um juri tratando de uma assunto com resultado binário. Considerando que cada membro do juri tenha uma probabilidade

superior a 50% de dar um veredito correto, então adicionar mais membros votantes ao juri tende a aumentar a probabilidade de se chegar a decisão correta, assumindo-se que os votantes são independentes (SAGI; ROKACH, 2018).

O termo comitê de classificadores vem do inglês *classifier ensemble*, sendo um subgênero do *ensemble learning* que pode ser entendido como um sistema de classificação composto por classificadores individuais treinados de forma isolada que são posteriormente combinados por meio de um regra (OLIVEIRA, 2021). Um comitê de classificadores pode ser tratado como a aplicação de *boosting* em modelos que utilizam algoritmos distintos.

Considerando-se que um conjunto de classificadores com desempenhos de treinamento semelhantes pode ter desempenhos de generalização diferentes, principalmente nos casos em que o conjunto de dados de teste utilizados para determinar o desempenho de generalização não for suficientemente representativo dos dados, combinar as saídas de vários classificadores pode ajudar a reduzir o risco de uma seleção incorreta de um classificador de baixo desempenho (POLIKAR, 2006).

Outra vantagem do uso de um comitê de classificadores está em problemas muito complexos para um determinado classificador resolver, como por exemplo um classificador linear que não é capaz de classificar dados não lineares complexos, entretanto, uma combinação adequada de vários classificadores lineares pode ser utilizada para modelar esse limite não linear (POLIKAR, 2006).

3 TRABALHOS RELACIONADOS

Este capítulo aborda os principais trabalhos relacionados com o uso de sinais de voz para o diagnóstico da DP. A Seção 3.1 apresenta revisões sistemáticas realizadas que demonstram a relevância do tema. Já a Seção 3.2 aborda os trabalhos realizados com uma base composta por atributos de áudio previamente extraídos desenvolvida por Little et al. (2009), enquanto que a Seção 3.3 apresenta os trabalhos realizados na base desenvolvida por Bot et al. (2016), que contém gravações de voz de portadores e não portadores da DP. Por fim, a Seção 3.4 apresenta as oportunidades para o desenvolvimento desta pesquisa.

3.1 Mapeamentos sistemáticos

A utilização de sinais e informações oriundos da fala como potenciais indicadores da presença da DP tem experimentado um aumento notável ao longo dos últimos 15 anos. De acordo com a pesquisa conduzida por Ngo et al. (2022), no período de 2010 a 2021, foi identificado um total de aproximadamente 838 estudos relacionados a esse tema. Desse conjunto, 189 estudos foram criteriosamente selecionados, dos quais 147 foram considerados apropriados para inclusão na revisão sistemática. Os resultados demonstram que tanto a fala quanto a voz emergem como biomarcadores de relevância no contexto das investigações sobre a DP.

Já Amato et al. (2023), focando de maneira mais específica na aplicação de métodos de ML e técnicas de inferência estatística no contexto das características vocais, observou por meio de uma análise abrangente de 102 estudos realizados entre os anos de 2017 e 2022 que determinados atributos como jitter, shimmer, HNR, F0 e DFA são utilizados de forma mais predominante.

3.2 A base de Little

Little et al. (2009) desenvolveram 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 10. As medidas mais tradicionais (F0, shimmer, jitter, entre outras) foram computadas utilizando-se o software

Praat[1]. Já as medidas não tradicionais (RDPE, DFA, PPE, entre outras) foram extraídas a partir de algoritmos implementados pelos autores .

Tabela 10 – Descrição dos atributos do conjunto de dados de Parkinson

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)					

Fonte: Sakar e Kursun (2009)

O modelo concebido por Little et al. (2009) utilizou um classificador fundamentado em uma Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) e implementou uma abordagem inicial de filtragem de atributos por meio de uma análise de correlação. Dentro desse contexto, para cada par de atributos, aqueles que exibiam um coeficiente de correlação superior a 0.95 teriam um dos atributos removido do conjunto de dados.

Um outro atributo descartado foi a F0, com o argumento de que, embora existisse evidência de uma relação estatística entre os valores absolutos de F0 e a presença da DP, optou-se por não incorporar essa medida, dada a sua forte influência vinculada ao gênero dos pacientes.

Após a etapa de filtragem, apenas 10 atributos permaneceram no conjunto de dados, conforme indicado na Tabela 11. A totalidade dos 1023 subconjuntos viáveis foi empregada tanto no treinamento quanto no teste de um modelo SVM. Os melhores resultados, acompanhados de um intervalo de confiança de 95%, são apresentados na Tabela 12.

[1] Acesso em: praat.softonic.com.br

Tabela 11 – Atributos mantidos após filtragem

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

Fonte: Little et al. (2009)

Tabela 12 – Resultados da performance de classificação do SVM

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

Fonte: Little et al. (2009)

Com base no mesmo conjunto de dados, Bhattacharya e Bhatia (2010) elaborou um modelo alternativo empregando igualmente uma SVM. O desenvolvimento desse modelo foi conduzido utilizando o software Weka[2], que desempenhou funções cruciais na filtragem de atributos, bem como no processo de treinamento e avaliação do modelo proposto .

Os atributos que passaram pelo processo de filtragem resultaram no mesmo subconjunto de características identificado no modelo anterior. Para a validação do modelo, a abordagem adotada foi o método k-fold, no qual k foi definido como 3. O melhor resultado de acurácia alcançado foi de 95.3% para o conjunto de treinamento e 60.9% para o conjunto de teste.

Utilizando a plataforma SAS base[3] (uma ferramenta que engloba recursos de filtragem e mineração de dados, assim como métodos de aprendizado de máquina), Das (2010) explorou 4 diferentes abordagens de ML: Rede Neural, DMNeural, Regressão Logística e Árvore de Decisão .

[2] Acesso em: www.weka.io
[3] Acesso em: www.sas.com

A validação dos modelos foi realizada por meio da divisão do conjunto de dados em dois conjuntos distintos, um para treinamento e outro para teste, garantindo que não houvesse sobreposição entre as instâncias. A Tabela 13 reúne os melhores resultados de acurácia alcançados nos conjuntos de treinamento e teste para cada um dos modelos empregados.

Tabela 13 – Resultados da performance de classificação de múltiplos modelos

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%

Fonte: Das (2010)

Sakar e Kursun (2009) também conceberam um modelo fundamentado em SVM, o qual adotou um procedimento de discretização para os atributos que passaram por um prévio processo de filtragem usando a abordagem Máxima Relevância - Mínima Redundância (mRMR). O método de validação utilizado foi o *leave-one-individual-out*.

A discretização dos atributos ocorreu em 9 níveis. Para essa finalidade, a média μ e o desvio padrão σ foram utilizados, transformando valores entre $\mu - \sigma$ e $\mu + \sigma$ em 0. Os quatro intervalos de tamanho σ à direita de 0 foram convertidos em níveis discretos de 1 a 4, e de modo correspondente, aqueles à esquerda receberam níveis discretos de -1 a -4.

O método mRMR se baseia na premissa de que atributos individualmente informativos não necessariamente resultam em um melhor desempenho de classificação. A metodologia consiste em selecionar atributos altamente relevantes, evitando redundância e maximizando a dependência conjunta.

No que diz respeito ao método de validação *leave-one-individual-out*, sua particularidade reside em manter todas as instâncias relacionadas separadas. No contexto do conjunto de dados da Doença de Parkinson, todas as 6 ou 7 instâncias provenientes de um sinal de voz de um dos 32 pacientes eram mantidas isoladas durante o processo de treinamento e teste.

Este método difere do método tradicional *leave-one-individual-out*, no qual apenas uma única instância é excluída no processo de treinamento e teste. A justificativa é que dessa maneira as amostras de treinamento e teste são verdadeiramente independentes, resultando em estimativas de acurácia e outras estatísticas mais confiáveis.

O melhor resultado obtido foi uma acurácia de 92.75%, com um intervalo de confiança de 1.21%. Esse desempenho foi alcançado utilizando um subconjunto de treinamento composto pelos atributos spread1, Fo(Hz), Shimmer:APQ3 e D2.

Govindu e Palwe (2023) adotou três diferentes estratégias, cada uma delas explorando quatro modelos de aprendizado de máquina distintos: SVM, regressão logística, *Random Forest*

e k-vizinhos mais próximos (KNN, do inglês *K-nearest neighbors*). Na primeira abordagem, todos os 22 atributos do conjunto de dados foram empregados no treinamento dos modelos. Na segunda abordagem, uma PCA foi aplicada para reduzir o número de atributos utilizados no treinamento para 5. A terceira abordagem focou na questão de balanceamento dos dados, já que a base continha 109 amostras de pacientes com DP e apenas 40 de indivíduos saudáveis. Para remediar isso, a base foi balanceada por meio da duplicação das amostras de não portadores da DP até que o número de registros fosse igualado. As amostras duplicadas eram escolhidas aleatoriamente. Os modelos foram então treinados com todos os atributos disponíveis. Em todas as abordagens, a normalização dos dados foi realizada usando a técnica do desvio padrão.

A validação dos modelos envolveu a divisão dos dados em conjuntos de treinamento e teste, com 75% dos dados sendo atribuídos aos conjuntos de treinamento. Para cada uma das abordagens, métricas de acurácia, precisão e sensibilidade foram coletadas. Os resultados da primeira abordagem estão resumidos na Tabela 14, enquanto os resultados da segunda abordagem são apresentados na Tabela 15, e os resultados da terceira abordagem estão destacados na Tabela 16.

Tabela 14 – Resultados da performance de classificação da primeira abordagem

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%

Fonte: Govindu e Palwe (2023)

Tabela 15 – Resultados da performance de classificação da segunda abordagem

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%

Fonte: Govindu e Palwe (2023)

Melo e Gouveia (2023) utilizaram o *Random Forest* na implementação de seu modelo. A seleção dos atributos foi feita através de teste de hipótese utilizando a distribuição t e análise de correlação. Para cada atributo, foi obtido o seu valor-p correspondente ao teste de hipótese onde a hipótese nula H_0 representava a igualdade entra as médias de cada status (portador da DP ou paciente saudável). A Tabela 17 apresenta os valores-p obtidos para cada atributo.

Foi adotado um intervalo de confiança de 99%. Desta forma, o limiar utilizado como critério de rejeição da H_0 foi de 0.01. Ou seja, o atributo que obtivesse valor-p inferior a 0.01

Tabela 16 – Resultados da performance de classificação da terceira abordagem

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**%

Fonte: Govindu e Palwe (2023)

Tabela 17 – 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)		

Fonte: Melo e Gouveia (2023)

teria a diferença de médias atestada. Observa-se que o único atributo rejeitado nessa etapa foi a F0 máxima.

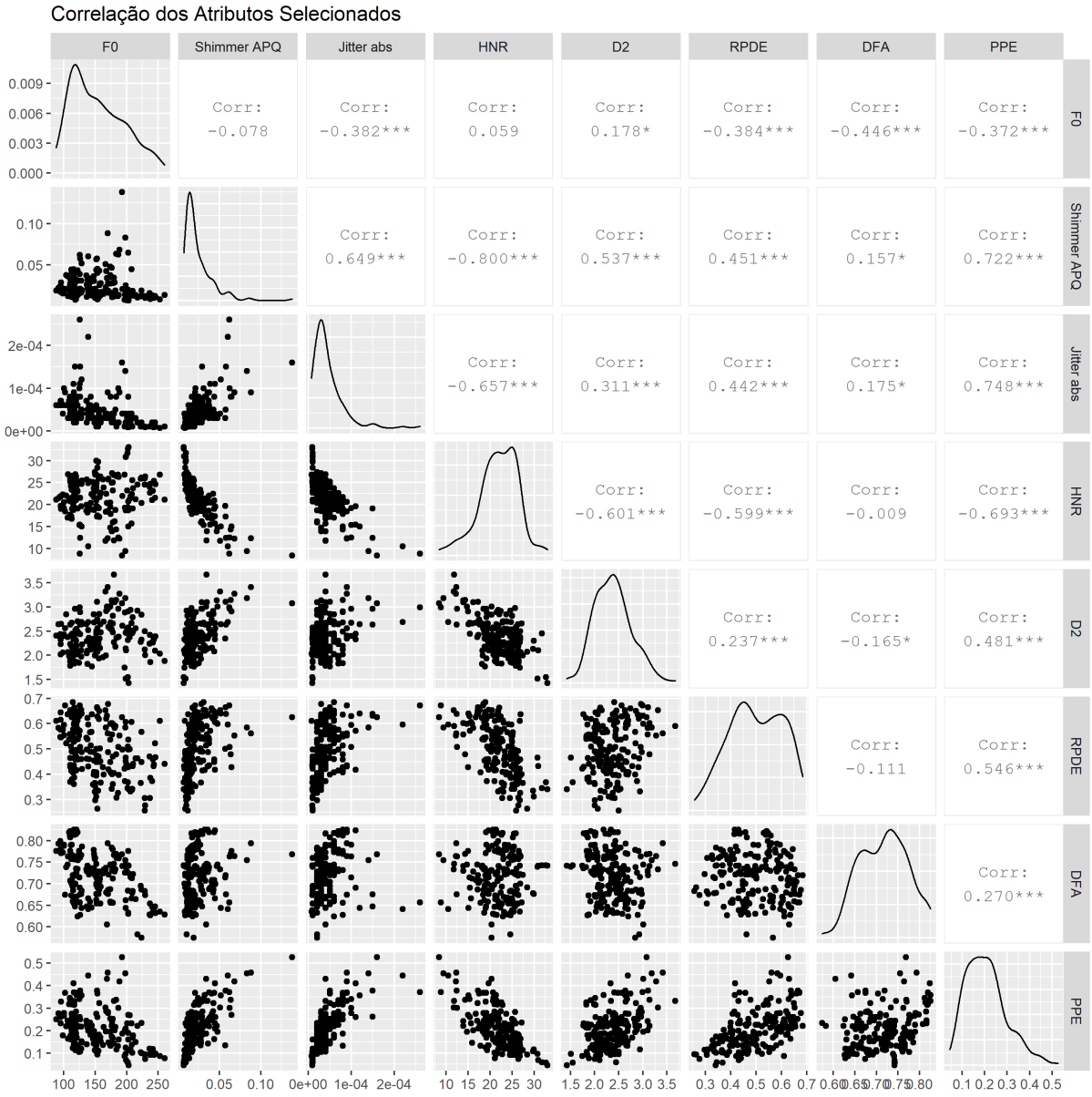
A segunda fase do procedimento de filtragem envolve uma análise de correlação entre os atributos que abordam uma mesma característica vocal. O propósito desse passo é evitar redundâncias no modelo. Dado que várias medidas abordam aspectos semelhantes ou até idênticos, é razoável esperar que haja uma correlação substancial entre elas.

Para cada característica do sinal vocal presente no conjunto de dados, calculou-se o

índice de correlação. Quando o valor absoluto da correlação entre um par de atributos excedia 0.75, o atributo com menor relevância estatística era identificado e removido. A determinação da relevância estatística foi conduzida através da avaliação do valor-p obtido no teste de hipótese.

Após a filtragem, restaram 8 atributos que podem ser visualizados na Figura 13, bem como suas correlações entre si. Esses atributos foram divididos em 6 subconjuntos distintos que são apresentados na Tabela 18. Cada subconjunto foi usado no treinamento de um modelo de *Random Forest* que foi validado com o método k-fold com um k igual a 5. Os resultados obtidos são apresentados na Tabela 19.

Figura 13 – Análise de correlação da base final



Fonte: Melo e Gouveia (2023)

Tabela 18 – 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

Fonte: Melo e Gouveia (2023)

Tabela 19 – 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

Fonte: Melo e Gouveia (2023)

Por fim, a Tabela 20 resume os principais resultados obtidos pelas pesquisas supracitadas. O modelo de melhor resultado foi o de Melo e Gouveia (2023) que, utilizando um modelo de *Random Forest*, alcançou uma acurária de 93.8%. Contudo, considerando-se o intervalo de confiança informado, todos os modelos, com a exceção do modelo de Bhattacharya e Bhatia (2010), possuem acurácia na mesma faixa, o que indica que os resultados obtidos entre eles são praticamente idênticos.

Tabela 20 – 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
Random Forest (Melo; Gouveia, 2023)	93.8 ± 4.4

Fonte: Autoria própria

3.3 A base do estudo mPower

Bot et al. (2016) desenvolvem o projeto mPower[4], que coleta dados de portadores da doença de Parkinson e um grupo de controle composto por pessoas não diagnosticadas com a

doença. Os participantes da pesquisa executam tarefas em 4 modalidades: locomoção, batida, memória e voz. Além disso, os participantes respondiam a um questionário demográfico que, entre outros pontos, abordava questões como idade, gênero, hábitos e histórico médico, status educacional e socioeconômico dos pacientes. A Tabela 21 apresenta quantos participantes e tarefas únicas foram desempanhadas por tipo de tarefa .

Tabela 21 – Tarefas do estudo mPower

Tarefa	Participantes	Atividades
Demográfico	6805	6805
Memória	968	8569
Batida	8003	78887
Voz	5826	65022
Locomoção	3101	35410

Fonte: Bot et al. (2016)

As atividades de voz consistiam em gravações de fonações sustentadas da vogal 'a' por 10 segundos. As gravações foram realizadas utilizando-se o microfone dos celulares dos participantes. Todos os participantes habitavam nos Estados Unidos durante a pesquisa e o celular dos participantes eram modelos de iPhone.

Fazendo uso da base mPower, Wang et al. (2020) construíram um modelo de SVM. Inicialmente, foram selecionadas 1600 amostras de forma aleatória. Cerca de 400 foram rejeitadas aplicando-se critérios como a presença de barulhos ambientais, vento e afins nas gravações. Em seguida, energias de curta duração foram extraídas das amostras restantes, sendo calculadas em intervalos de 0.1 segundos utilizando a biblioteca OpenSMILE, desenvolvida por Eyben, Wöllmer e Schuller (2010). Medidas simples como médias e variâncias foram calculadas e utilizadas para filtrar o resto dos dados. Após essa etapa, restaram apenas 4000 pacientes, dos quais 900 eram portadores da DP. Por fim, foram aplicados critérios de balanceamento de gênero e idade, resultando em 1022 pacientes restantes, sendo metade de portadores, e 1700 áudios. Foram extraídos atributos de disfonia, transformada, entropia, entre outros das gravações. Esses atributos foram utilizados no treinamento do modelo. O método de validação foi o k-fold, com k igual a 10. A acurácia obtida foi de cerca de 58%.

Por sua vez,Karaman et al. (2021) construíram um modelo de CNN. A filtragem dos dados foi feita selecionando-se apenas pacientes com um diagnóstico médico profissional e que não tenham realizado o procedimento cirúrgico de estimulação cerebral profunda. Além disso, foram utilizadas apenas gravações realizadas imediatamente antes do uso de medicação contra a DP (no caso dos pacientes portadores).

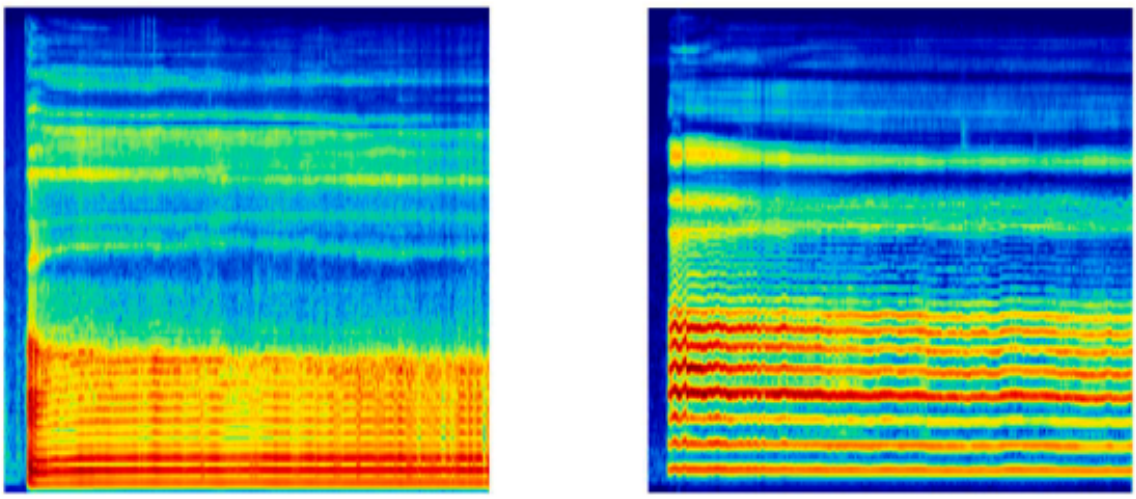
Ao fim, restaram 33877 gravações, sendo 10589 de pacientes portadores da DP. 18660 gravações de pacientes saudáveis e 8442 de portadores da DP foram utilizadas no processo

[4] Acesso em: www.synapse.org

de treinamento do modelo. O teste foi realizado com 4628 gravações de pacientes saudáveis e 2147 de portadores. De forma mais específica, foram utilizados 400 pacientes totalmente independentes no conjunto de teste, sendo 200 portadores. Isto significa que nenhuma gravação dos pacientes do conjunto de teste foram utilizadas no conjunto de treinamento.

O modelo foi treinado com espectrogramas que foram gerados a partir da DCT. A DCT era performada num frame de tamanho 2048 do áudio multiplicado pela janela de Hann. Em seguida, o frame era deslocado em 512 amostras e o processo era repetido para todo o comprimento do áudio. Por fim, era aplicada uma matriz de transformação de bancos de filtro para converter a DCT em 320 frequências de Mel. O resultado final do processo era uma matriz bidimensional com informações do domínio da frequência e do tempo. Os elementos dessa matriz eram então coloridos de forma logarítmica com o mapa de cor "jet" da biblioteca Matplotlib[5] do Python. Os Espectrogramas eram salvos no formato PNG com uma resolução de 448px por 448px. A Figura 14 traz dois exemplos desses espectrogramas, um de paciente portador, à direita, e outro de um paciente não portador da DP, à esquerda. .

Figura 14 – Exemplo de espectrogramas



Fonte: Karaman et al. (2021)

Todos os áudios passaram por esse processo e os espectrogramas foram utilizados no treinamento de modelos de CNN com as arquiteturas ResNet50, DenseNet161 e SequeezeNet1_1. Esses modelos haviam sido previamente treinados com os dados da base ImageNet. Com isso, espera-se melhorar e diminuir o tempo de treinamento no novo conjunto de dados. Essa técnica é chamada de transferência de aprendizado. O melhor resultado obtido foi uma acurácia de 89.75%.

[5] Acesso em: matplotlib.org

3.4 Oportunidades

De forma geral, os resultados obtidos pelos modelos que utilizaram a base de Little foram bem superiores aos do estudo mPower. A maior discrepância se verifica no modelo de Wang et al. (2020), cuja a acurácia ficou abaixo da faixa de 60%. Esse caso chama a atenção, pois tanto os atributos quanto o modelo utilizado foram semelhantes ao utilizado por Little et al. (2009), que obtiveram uma acurácia superior a 90%.

Assim, tem-se a oportunidade verificar se os atributos de áudio de fato não se mostram aptos para o diagnóstico da DP quando se utiliza uma base maior. Outra possibilidade é que as implementações para extração de atributos realizada por (WANG et al., 2020) sejam ineficazes e, dessa forma, ainda é possível atingir as acurácias verificadas na base de Little com uma abordagem diferente.

Além disso, a qualidade das gravações do estudo mPower podem explicar a discrepância de resultados encontrada. A base de Little foi desenvolvida em um ambiente extremamente controlado, enquando que a do estudo mPower foi feita pelos próprios pacientes por meio de um aplicativo de celular (BOT et al., 2016). Assim sendo, os atributos de áudio podem ser distorcidos por ruídos e interferências de outros sinais.

Como Karaman et al. (2021) conseguiu uma acurácia mais próxima das verificadas nos modelos que utilizaram a base de Little, constata-se que a utilização de espectrogramas e CNN se mostra mais promissora do que o uso de atributos de áudio. Os espectrogramas podem ser menos sensíveis a "sujeiras" nos sinais de voz, sendo assim mais aptos para a detecção do Parkinson em ambientes pouco controlados.

A utilização de técnicas mais avançadas de ML, como o comitê de classificadores, e a utilização de critérios diferentes de seleção de amostras válidas para treinamento, validação e teste podem ser uma alternativa para alavancar os resultados obtidos por Karaman et al. (2021) e Wang et al. (2020).

Por fim, outra oportunidade que se apresenta é a avaliação da aplicabilidade dos modelos levando em consideração a incidência da DP. Tal informação é relevante para se atestar a utilidade de um método de diagnóstico e encontra-se ausente nas pesquisas supracitadas.

4 DESENVOLVIMENTO DO MODELO

Este capítulo apresenta os materiais e métodos utilizados no desenvolvimento da pesquisa. A Seção 4.1 trata da obtenção e preparação dos dados, enquanto que a Seção 4.2 busca dar um melhor entendimento aos dados, além de apresentar um conjunto de critérios para a seleção de amostras válidas para o treinamento dos modelos. As Seções 4.3 e 4.4 abordam respectivamente os modelos de *Random Forest* e CNN utilizados. Por fim, as Seções 4.5 e 4.6 tratam do processo de *data augmentation* e do modelo de comitê de classificadores utilizado.

4.1 Preparação dos dados

Esta seção apresenta todas as etapas relacionadas à preparação dos dados necessários para o treinamento dos modelos, partindo da obtenção até a extração de atributos de áudio e geração de espectrogramas.

4.1.1 Obtenção dos dados

Para o desenvolvimento da pesquisa, foram utilizados os dados relacionados às atividades de voz obtidos e disponibilizados pelo projeto mPower e que foram acessados por intermédio da plataforma synapse.org. Esses dados se subdividem em 2 grupos: dados tabulares simples no formato CSV e gravações de áudio no formato M4A.

Os dados tabulares são compostos por 2 arquivos e foram obtidos por meio de download simples na synapse.org. O primeiro deles apresenta dados gerais sobre os pacientes. Ele possui 6805 linhas e 33 colunas, sendo que cada linha corresponde a um paciente distinto. As suas colunas são apresentadas na Tabela 22, destacando-se a coluna“professional-diagnosis” que indica a presença ou ausência da DP no paciente.

O segundo arquivo apresenta informações gerais sobre as gravações. Ele possui 65022 linhas e 10 colunas, sendo que cada linha corresponde a uma gravação distinta de um paciente. As suas colunas são apresentadas na Tabela 23, destacando-se as colunas “recordId”,“healthCode” e “medTimePoint”, que apresentam, respectivamente, o identificador, o paciente e o momento da gravação em relação ao uso de medicamento relacionado à DP.

Foi utilizada uma biblioteca Python criada e disponibilizada pela synapse.org para obtenção das gravações de voz. Ela oferece, entre outras funcionalidades, um método para extração dos áudios e um arquivo JSON, que permite o mapeamento entre os nomes dos arquivos e os identificadores apresentados na tabela de gravações.

Foram extraídos 66427 áudios, mas observou-se que alguns deles estavam em formato inválido ou vazios (áudios com duração de 0 segundos). Além disso, foram removidas todas as

Tabela 22 – Colunas da tabela de pacientes

Coluna	Descrição
ROW_ ID	Identificador da linha
ROW_VERSION	Versão da linha
recordId	Identificador da entrevista
healthCode	Identificador do paciente
createdOn	Data de criação do registro
appVersion	Versão do aplicativo utilizada pelo paciente
phoneInfo	Tipo de celular utilizado pelo paciente
age	Idade do paciente
are-caretaker	Indicativo de se o paciente era cuidador portador da DP
deep-brain-stimulation	Indicativo de se o paciente passou por procedimento cirúrgico relacionado à DP
diagnosis-year	Idade em que o paciente foi diagnósticado com a DP
education	Nível de escolaridade do paciente
employment	Status empregatício do paciente
gender	Gênero do paciente
health-history	Histórico médico do paciente
healthcare-provider	Indicativo de se o paciente é profissional da área de Saúde
home-usage	Nível de uso de Internet
last-smoked	Ano da última vez que o paciente fumou
maritalStatus	Estado civil
medical-usage	Indicativo de se o paciente usa alguma medicação
medical-usage-yesterday	Indicativo de se o paciente usou a medicação no dia anterior à entrevista
medication-start-year	Ano em que o paciente começou a usar medicações
packs-per-day	Quantidade de medicações utilizadas por dia
past-participation	Indicativo de se o paciente já participou de outros estudos relacionados à DP
phone-usage	Nível de utilização de celular
professional-diagnosis	Indicativo se o paciente era portador da DP
race	Raça do paciente
smoked	Indicativo de se o paciente já foi ou é fumante
surgery	Descrição de procedimento cirúrgico realizado pelo paciente
video-usage	Nível de utilização de vídeo
years-smoking	Quantidade de anos como fumante

Fonte: Bot et al. (2016)

gravações que não tivessem um identificador com correspondência válida na tabela de gravações. Ao final do processo, restaram 65022 áudios com duração aproximada de 10 segundos.

Por fim, as tabelas de pacientes e gravações foram mescladas. Nessa etapa, observou-se que alguns dos pacientes não têm gravações associadas a eles e que algumas gravações não pertencem a nenhum dos pacientes. Também foi observado que nem todos os pacientes possuem diagnóstico conhecido. Dessa forma, foram descartados todos os pacientes e gravações avulsos,

Tabela 23 – Colunas da tabela de gravações.

Coluna	Descrição
ROW_ ID	Identificador da linha
ROW_VERSION	Versão da linha
recordId	Identificador da gravação
healthCode	Identificador do paciente
createdOn	Data de criação do registro
appVersion	Versão do aplicativo utilizada pelo paciente
phoneInfo	Tipo de celular utilizado pelo paciente
audio_audio.m4a	Nome do arquivo com a gravação
audio_countdown.m4a	Não especificado
medTimePoint	Momento em que a medicação específica para DP foi usada em relação à gravação

Fonte: Bot et al. (2016)

bem como os pacientes com diagnóstico incerto. Assim, restaram 4962 pacientes e 62848 gravações.

4.1.2 Extração de atributos de áudio

Foram extraídos diversos atributos de áudio para cada gravação. Os mais tradicionais, como F0, jitter e shimmer, foram obtidos utilizando-se a biblioteca OpenSmile (EYBEN; WöLLMER; SCHULLER, 2010). A Tabela 24 apresenta a lista de atributos, bem como o conjunto de métricas desses atributos que foram utilizadas. No total, foram extraídas 88 características dos áudios com a OpenSmile.

Outros atributos de áudio extraídos das gravações foram o expoente de Hurst e a DFA, utilizando a biblioteca Python Noldz[1] , o RPDE, utilizando a biblioteca Pyrpde[2] e o PPE que foi calculado com um algoritmo de implementação própria, de acordo com o apresentado por Little et al. (2009). Assim, foram extraídos um total de 92 atributos para cada áudio.

Nessa etapa, foram encontradas uma série de inconsistências nos valores obtidos para os atributos extraídos. Observou-se que essas inconsistências ocorrem por problemas nas gravações, como a ausência de som, ruídos com volume elevado, atraso no início ou término precoce da fonação sustentada, entre outros. Para contornar esse problema, optou-se por extrair os atributos apenas do segundo central de cada gravação. Ou seja, os atributos foram extraídos considerando o início de cada gravação no instante 4.5 segundos e seu término no instante 5.5 segundos. Por fim, os atributos de áudio foram persistidos em formatos tabulares para melhor adequação com os modelos que serão utilizados.

[1] Acesso em: pypi.org/project/nolds/

[2] pypi.org/project/pyrpde/

Tabela 24 – Atributos da OpenSmile.

Atributo	Métrica
F0semitoneFrom27.5Hz	Média, desvio padrão, mediana, 20º percentil, 80º percentil, média e desvio padrão da curva ascendente, média e desvio padrão da curva descendente
Loudness	Média, desvio padrão, mediana, 20º percentil, 80º percentil, média e desvio padrão da curva ascendente, média e desvio padrão da curva descendente
SpectralFlux	Média e desvio padrão
MFCC1	Média e desvio padrão
MFCC2	Média e desvio padrão
MFCC3	Média e desvio padrão
MFCC4	Média e desvio padrão
Shimmer	Média e desvio padrão
Jitter	Média e desvio padrão
HNR	Média e desvio padrão
LogRelF0-H1-H2	Média e desvio padrão
LogRelF0-H1-A3	Média e desvio padrão
F1frequency	Média e desvio padrão
F1bandwidth	Média e desvio padrão
F1amplitudeLogRelF0	Média e desvio padrão
F2frequency	Média e desvio padrão
F2bandwidth	Média e desvio padrão
F2amplitudeLogRelF0	Média e desvio padrão
F3frequency	Média e desvio padrão
F3bandwidth	Média e desvio padrão
F3amplitudeLogRelF0	Média e desvio padrão
alphaRatioV	Média e desvio padrão
hammarbergIndexV	Média e desvio padrão
slopeV0-500	Média e desvio padrão
slopeV500	Média e desvio padrão
spectralFluxV	Média e desvio padrão
MFCC1V	Média e desvio padrão
MFCC2V	Média e desvio padrão
MFCC3V	Média e desvio padrão
MFCC4V	Média e desvio padrão
alphaRatioUV	Média e desvio padrão
hammarbergIndexUV	Média e desvio padrão
slopeUV0-500	Média e desvio padrão
slopeUV500	Média e desvio padrão
loudness	Média
VoicedSegmentLengthSec	Média e desvio padrão
UnvoicedSegmentLength	Média e desvio padrão
equivalentSoundLevel_dBp	Média

Fonte: Autoria Própria

4.1.3 Geração de espectrogramas

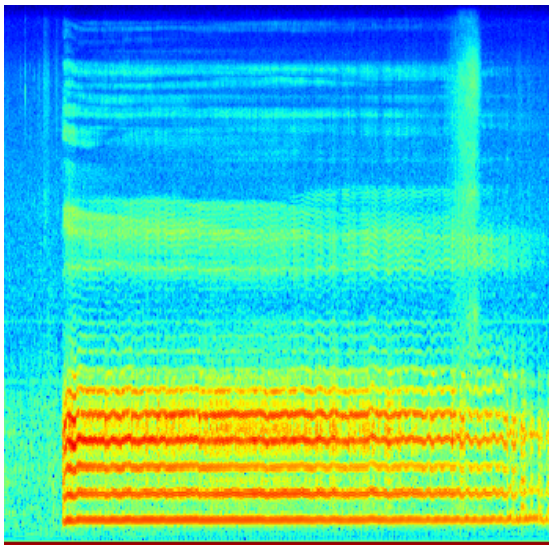
Os espectrogramas foram gerados por meio de implementação própria. Os áudios foram normalizados e, em seguida, passaram por uma SDCT, de acordo com o processo descrito por Karaman et al. (2021). A matriz gerada era então multiplicada por um banco de filtros de Mel. Esse banco foi gerado usando a implementação da biblioteca Python Librosa[3], considerando 4094 componentes da FFT e 320 frequências de Mel .

Por fim, a matriz resultante teve seus números transformados para uma escala logarítmica e os espectrogramas foram gerados utilizando a função colormesh da biblioteca Matplotlib. As imagens produzidas são quadradas, com 344 pixels de altura e largura e foram coloridas usando o mapeamento jet.

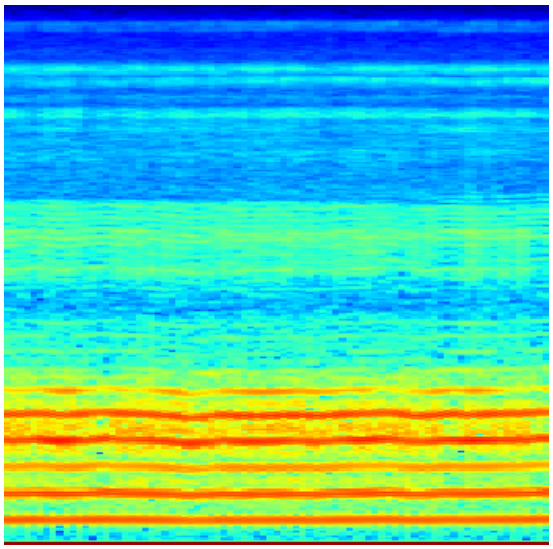
Os espectrogramas também foram gerados utilizando o segundo central de cada gravação, conforme o motivo apontado na Subseção 4.1.2. O principal impacto dessa escolha é eliminação dos vazios sonoros característicos no início e fim dos áudios. Além disso, analisando apenas o segundo central, tem-se um efeito semelhante ao de uma lupa, sendo possível observar aquele instante com um nível de detalhe maior. Esses efeitos podem ser observados na Figura 15, que apresenta um espectrograma considerando toda a duração da gravação e um espectrograma considerando apenas o segundo central.

Figura 15 – Espectrogramas com diferentes intervalos

(a) Gravação completa.



(b) Apenas o segundo central.



Fonte: Autoria Própria

Por fim, os espectrogramas foram transformados em arrays numpy[4] e em tensores com

[3] Acesso em: librosa.org/doc/latest/index.html

[4] Acesso em: numpy.org

o auxilio da biblioteca Tensorflow[5] para melhor adequação aos modelos de CNN que serão utilizados.

4.2 Seleção das Amostras

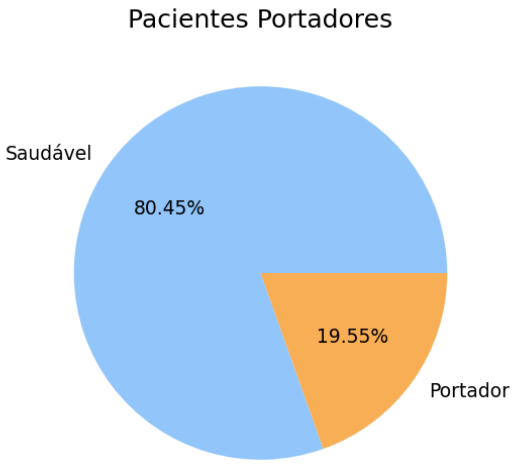
Esta seção apresenta os critérios utilizados na seleção das gravações válidas para uso no processo de treinamento dos modelos. Ela traz inicialmente uma análise exploratória dos dados demográficos, que servirá de suporte para a posterior filtragem dos dados.

4.2.1 Análise exploratória

Para a realização da análise exploratória dos dados, foram utilizadas as tabelas com os dados demográficos dos 4962 participantes do projeto mPower que passaram pelos critérios preliminares de seleção, conforme apresentado na Subseção 4.1.1. De forma mais específica, buscou-se entender o perfil dos pacientes, visando selecionar aqueles que melhor poderiam contribuir com a construção dos modelos.

A Figura 16 apresenta a proporção de pacientes portadores e não portadores da DP. Observa-se um claro desbalanceamento, com menos de 20% das amostras pertencentes a portadores. Isso afeta a escolha das métricas de avaliação dos modelos de ML.

Figura 16 – Proporção entre pacientes portadores e não portadores da DP



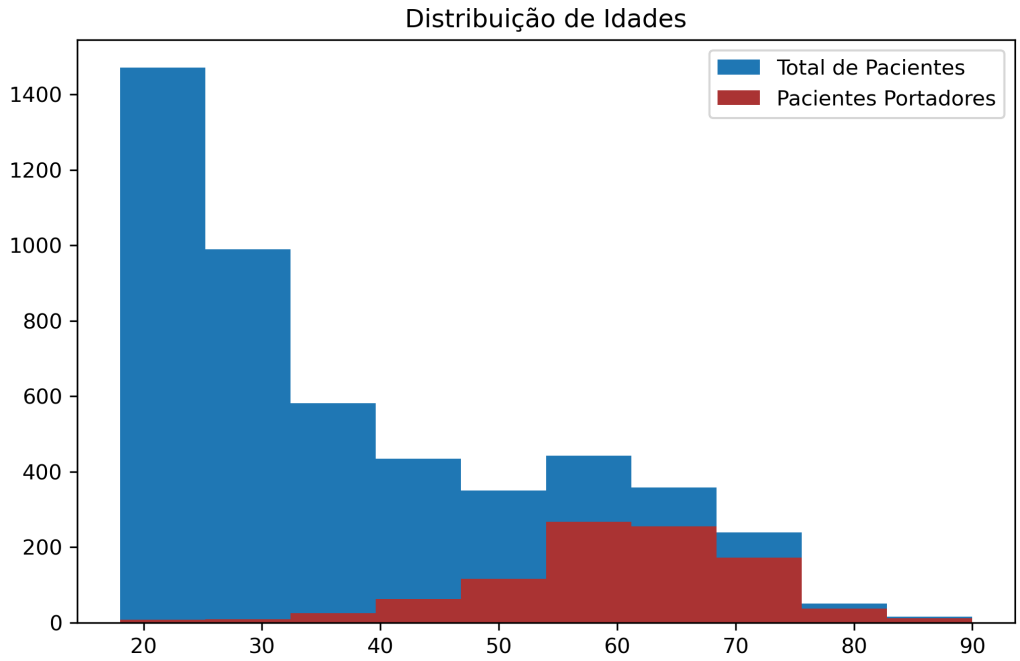
Fonte: Autoria Própria

A Figura 17 apresenta a distribuição de pacientes por idade. É possível observar que a maior parte dos portadores são idosos, enquanto que o grupo de controle é composto em sua maioria por pessoas com idade inferior a 40 anos. Esse é um ponto de atenção, pois o modelo

[5] Acesso em: www.tensorflow.org/

pode utilizar erroneamente características relacionadas a idade que não possuem relação com a doença como critérios de classificação.

Figura 17 – Distribuição da idade dos pacientes



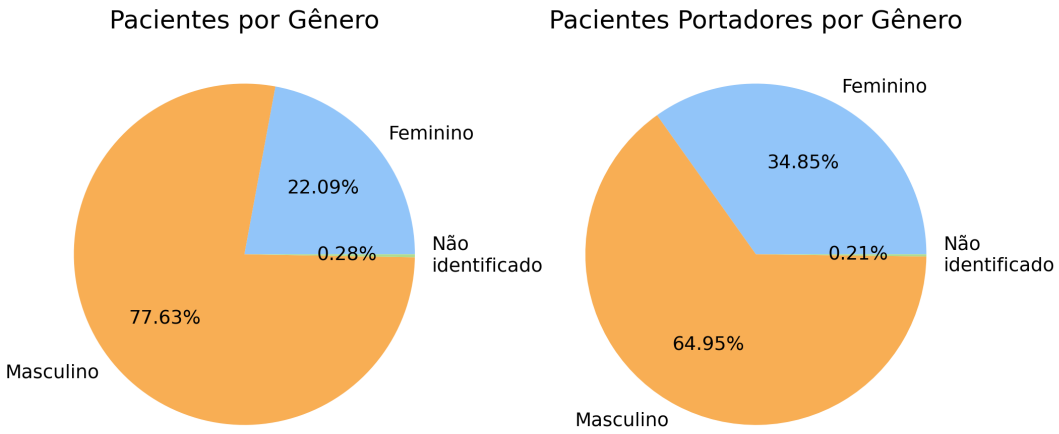
Fonte: Autoria Própria

A Figura 18 apresenta a proporção de pacientes por gênero. Um percentual inexpressivo optou por não identificar seu gênero. Observa-se que a maioria dos pacientes são do gênero masculino, havendo uma maior prevalência de pessoas do gênero feminino entre os portadores. Diferentemente do que ocorre com a distribuição de idades, esse desbalanceamento é mais improvável de acarretar problemas, pois ele ocorre de forma semelhante para portadores e não portadores da DP, embora atributos relacionados ao tom da voz possam ser utilizados erroneamente pelo classificador, uma vez que a voz feminina tende a ser mais aguda.

A Figura 19 apresenta a proporção de pacientes pelo seu grau de familiariedade com o uso de *smartphones*. É possível observar que a maior parte dos pacientes tem facilidade no uso. Contudo, quando se analisa apenas os que não têm facilidade, constata-se que a grande maioria é de portadores da DP. Esse é um ponto de atenção, pois como as gravações são realizadas utilizando esses aparelhos, dificuldade em utilizá-los pode acarretar em perda de qualidade ou até mesmo inutilização do áudio, o que explicaria os problemas encontrados na Subseção 4.1.2. Isso também poderia elevar o desbalanceamento dos dados.

A Figura 20 apresenta a distribuição de pacientes por status empregatício. Observa-se que a maior parte dos aposentados e dos impossibilitados para o trabalho são portadores da DP. Essa

Figura 18 – Proporção entre pacientes por gênero



Fonte: Autoria Própria

característica é esperada, uma vez que os portadores são em geral idosos e que a própria DP pode ser a causa da incapacitação para o trabalho. Essa variável não deve ter maiores consequências na construção dos modelos.

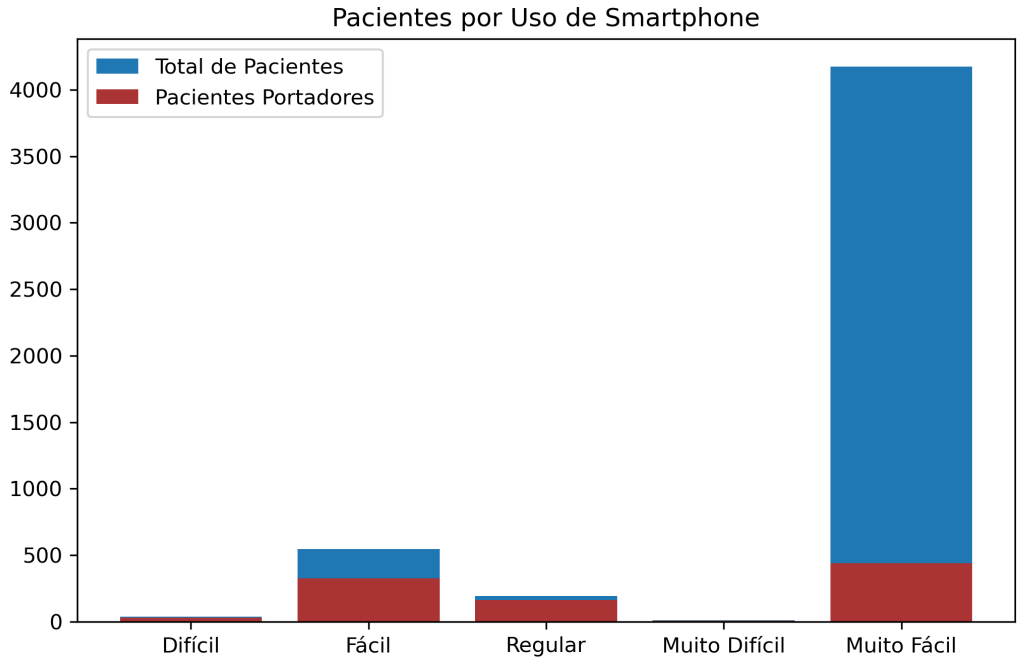
A Figura 21 apresenta a distribuição de pacientes por nível de escolaridade. Observa-se que grande parte dos pacientes possui nível universitário ou superior e não há aparente relação entre o nível de escolaridade e a presença da DP, de forma que essa variável não trará nenhum impacto significativo à pesquisa.

A Figura 22 apresenta a proporção de pacientes portadores e saudáveis que são cuidadores. O esperado era que nenhum paciente portador fosse simultaneamente cuidador, mas observa-se que há cerca de 4% dos cuidadores nessa situação. Esse é um ponto de atenção grave que também foi descrito por Karaman et al. (2021). As amostras provenientes desses pacientes devem ser removidas.

4.2.2 Filtragem dos dados

Para a filtragem dos dados, além dos critérios apresentados na Subseção 4.1.1, também foram acrescentados 4 novos critérios. O primeiro deles foi a remoção dos pacientes portadores da DP que também eram cuidadores de outros portadores, conforme apresentado na Subseção

Figura 19 – Distribuição de pacientes por uso de *smartphone*



Fonte: Autoria Própria

4.2.1. Embora seja possível que tal fato se verifique na realidade, considerou-se que o mais provável é que esses dados sejam inconsistências da base.

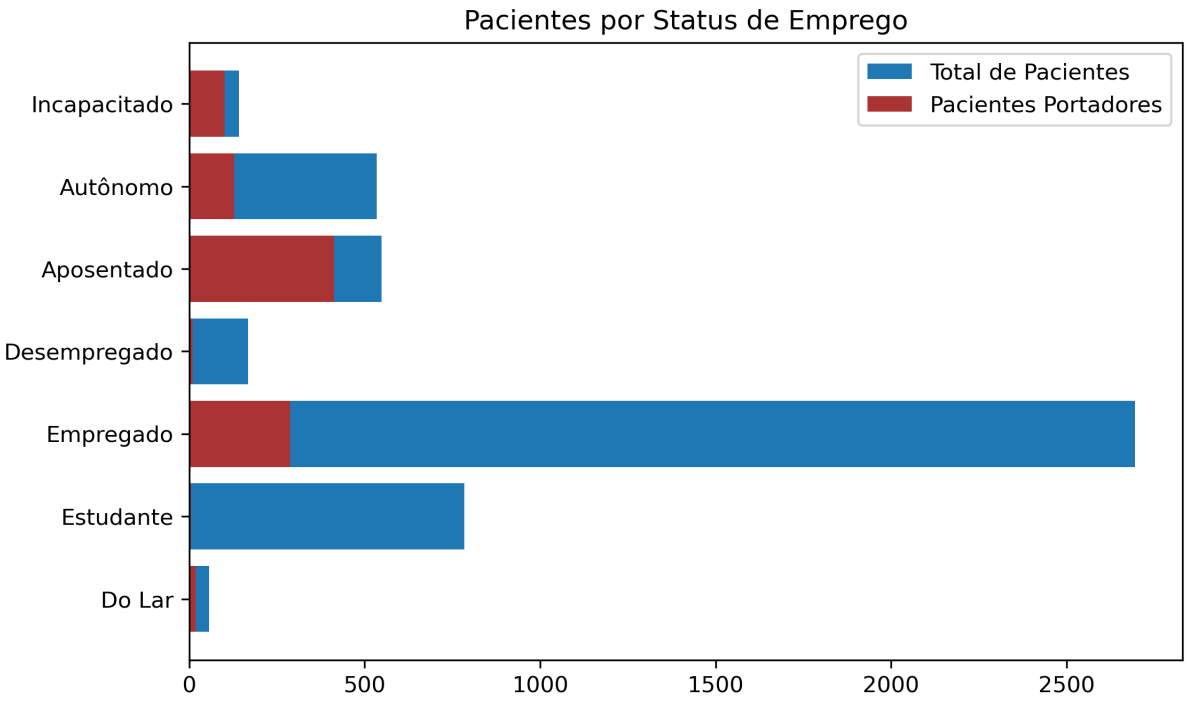
O segundo critério utilizado foi a realização de procedimentos cirúrgicos relacionados à DP. Considerou-se que uma vez que esses procedimentos tendem a atenuar os sintomas motores, é possível que o mesmo ocorra com as imparidades vocais, o que poderia impossibilitar a identificação da DP por meio da voz.

Os dois últimos critérios não estavam relacionados com os pacientes, mas com as gravações. O terceiro consistia na remoção de todas as amostras para as quais não seja possível extrair os atributos de áudio ou espectrogramas, de acordo com os métodos apresentados nas Subseções 4.1.2 e 4.1.3. Já o quarto critério aplicado estava relacionado com o consumo de medicamentos de combate aos sintomas da DP. Amostras provenientes de pacientes que tomaram a medicação antes da gravação foram removidas, por motivos análogos aos do segundo critério.

Com isso, restaram 28764 gravações, das quais aproximadamente 36% correspondiam a portadores da DP, o que também provocou uma melhoria no balanceamento dos dados. Os critérios de seleção das amostras são apresentados na forma de uma árvore de decisão na Figura 23.

Os dados foram então divididos em 3 conjuntos de dados distintos e sem sobreposição

Figura 20 – Distribuição de pacientes por status empregatício



Fonte: Autoria Própria

de elementos. O conjunto de treinamento ficou com 80% das amostras (23011 instâncias). O conjunto de validação ficou com 10% (2877 instâncias), de forma semelhante ao conjunto de teste (2876 instâncias). A divisão foi feita de forma que cada conjunto mantivesse a mesma proporção entre portadores e não portadores da DP.

4.3 Modelo de Random Forest

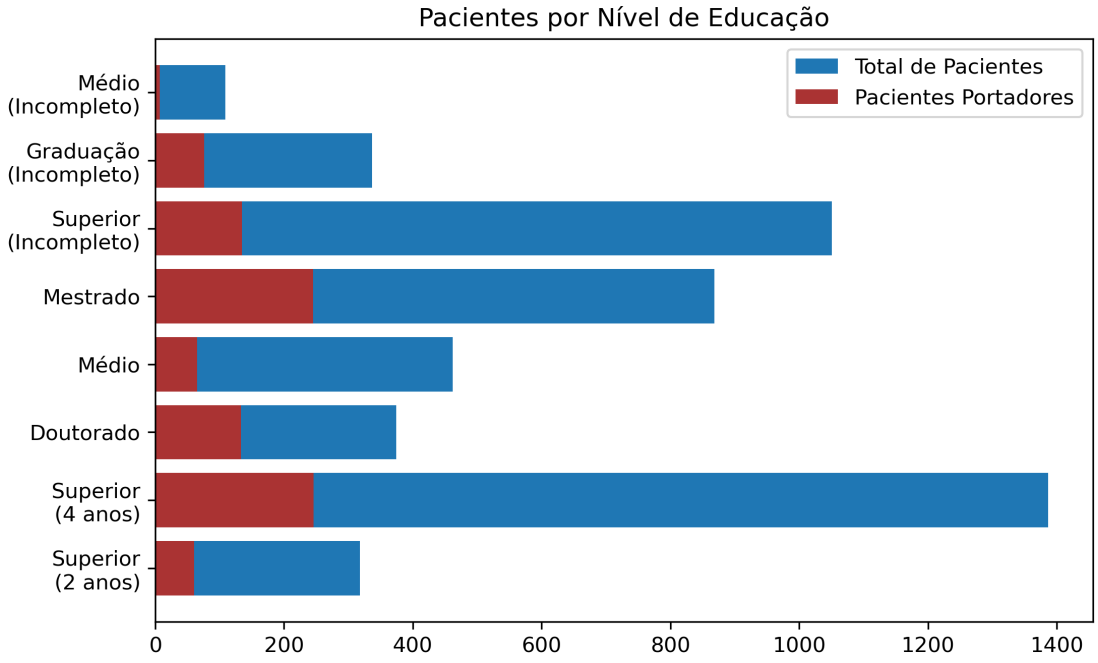
Esta seção traz o processo de treinamento e validação de um modelo de *Random Forest* criado com os atributos de áudio. Ela apresenta a análise estatística realizada nos atributos de áudio, a hiperparametrização do modelo e seus resultados preliminares. Todos os experimentos computacionais realizados nesta seção utilizaram uma máquina virtual no ambiente *Google Colab*. As especificações da máquina utilizada são na Tabela 25.

Tabela 25 – Especificações do ambiente de treinamento da *Random Forest*

Parâmetro	Valor
RAM do sistema	12.7GB
Disco	107.7GB

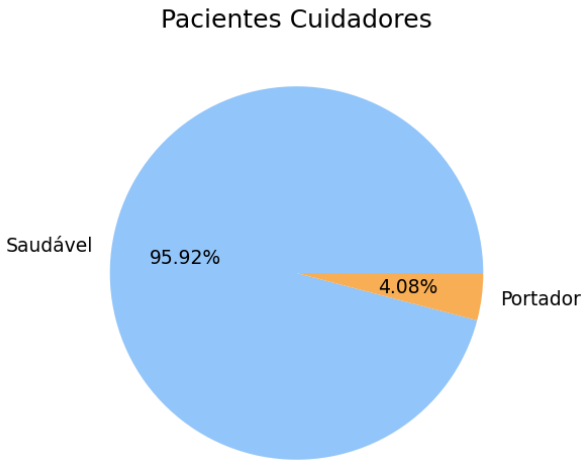
Fonte: Autoria Própria

Figura 21 – Distribuição de pacientes por nível de escolaridade



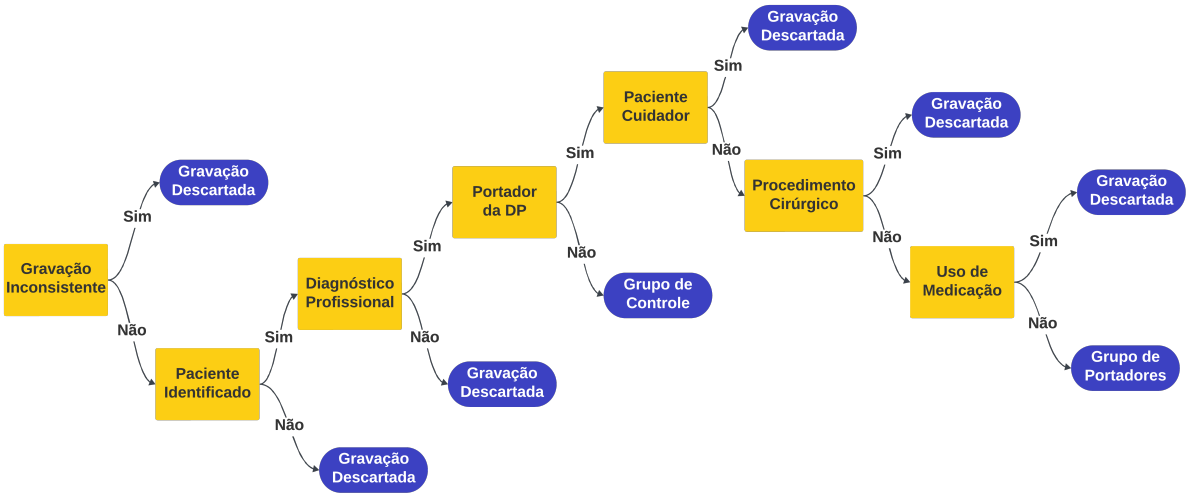
Fonte: Autoria Própria

Figura 22 – Proporção de pacientes cuidadores



Fonte: Autoria Própria

Figura 23 – Critérios de seleção das amostras



Fonte: Autoria própria

4.3.1 Modelo preliminar de *Random Forest*

Inicialmente, foi desenvolvido um modelo de *Random Forest* com todos os 92 atributos de áudio extraídos. As métricas obtidas desse modelo serviriam como linha de base do projeto. Os hiperparâmetros utilizados são exibidos na Tabela 26.

Tabela 26 – Hiperparâmetros do modelo preliminar *Random Forest*

Hiperparâmetro	Valor
Número de árvores	100
Função de impureza nodal	Gini
Número de amostras por folha	1
Número de atributos por *split*	9

Fonte: Autoria Própria

Os resultados de treino e validação do modelo preliminar são apresentados na Tabela 27. Todos os resultados de treinamento foram perfeitos e bem superiores aos de validação, o que é um claro indicativo de *overfitting*. O índice Kappa de validação ficou acima de 61%, indicando uma concordância substancial. A superioridade das performances da acurácia e precisão perante o *recall* e o F1-score se devem ao desbalanceamento do conjunto de dados e demonstram que o modelo é melhor em prever pacientes saudáveis do que portadores da DP.

4.3.2 Seleção de atributos de áudio

Como primeira tentativa para diminuir o *overfitting*, procurou-se diminuir o número de atributos utilizados no treinamento do modelo. Diminuindo-se a a quantidade de atributos,

Tabela 27 – Resultados de treinamento e validação do modelo preliminar de *Random Forest*

Métrica	Treino	Validação
Acurácia	100%	83.4%
Kappa	100%	62.0%
Precisão	100%	85.2%
Recall	100%	65.1%
F1-score	100%	73.8%

Fonte: Autoria Própria

espera-se uma diminuição na variância e, consequentemente, um aumento no viés, o que tende a melhorar a capacidade de generalização do modelo.

Para tal, foi realizado um teste t bicaudal com um critério de significância de 95%. Dessa forma, foram removidos todos os atributos com um valor-p inferior a 0.025, restando 64 atributos. A Tabela 28 apresenta os atributos selecionados. De forma geral, observa-se que quase todos os atributos se mostraram estatisticamente significativos, excetuando-se a métrica de desvio padrão.

O modelo preliminar foi então retreinado utilizando-se apenas os atributos de alta relevância estatística. A Tabela 29 apresenta os resultados obtidos. Todas as métricas foram incrementadas, mas com um aumento sempre inferior a meio ponto percentual. Além disso, como as métricas de treino ainda continuaram perfeitas. Conclui-se que a seleção de atributos não teve impacto significativo nos resultados.

A segunda tentativa para melhorar os resultados do modelo foi tentar remover atributos que tivessem alta correlação com outros atributos da base. Entre os atributos tradicionais, por exemplo, existem subgrupos de variáveis que tratam de uma mesma característica vocal. Dessa forma, espera-se que elas sejam fortemente correlacionadas. Isso impacta na variância do modelo, tendo relação direta com o *overfitting*.

Assim, para cada par de atributos com alta significância, foi feita uma análise de correlação. Quando o par estava fortemente correlacionado, isto é, uma correlação acima de 0.8, o atributo de maior valor-p era removido. Após esse processo, restaram 46 atributos. A Tabela 30 apresentado os atributos selecionados. Um dos atributos não lineares, o RPDE, foi descartado nessa etapa.

O modelo preliminar foi então retreinado com o novo subconjunto de atributos. Os resultados são apresentados na Tabela 31. Todas as métricas tiveram um decrescimento superior a 1%, o que indica que variáveis relevantes para a classificação estão sendo removidas, sendo esse provavelmente o caso do RPDE.

A última tentativa para melhorar os resultados do modelo por meio da manipulação dos atributos foi realizada aplicando-se uma PCA no conjunto de dados. Foram testadas as performances com todas as quantidades de componentes possíveis, isto é, entre 1 e 92 com-

Tabela 28 – Atributos com alta relevância estatística

Atributo	Métrica
F0semitoneFrom27.5Hz	Média, mediana, 20º percentil, 80º percentil
Loudness	Média, desvio padrão, mediana, 20º percentil, 80º percentil, média e desvio padrão da curva ascendente, média e desvio padrão da curva descendente
SpectralFlux	Média e desvio padrão
MFCC1	Média
MFCC3	Média
MFCC4	Média
Shimmer	Média e desvio padrão
Jitter	Média e desvio padrão
HNR	Média
LogRelF0-H1-A3	Média
F1frequency	Média e desvio padrão
F1bandwidth	Média e desvio padrão
F1amplitudeLogRelF0	Média
F2frequency	Média e desvio padrão
F2bandwidth	Desvio padrão
F2amplitudeLogRelF0	Média
F3frequency	Média e desvio padrão
F3bandwidth	Média e desvio padrão
F3amplitudeLogRelF0	Média
alphaRatioV	Média
hammarbergIndexV	Média
slopeV0-500	Média e desvio padrão
slopeV500	Média
spectralFluxV	Desvio padrão
MFCC1V	Média
MFCC3V	Média
MFCC4V	Média
alphaRatioUV	Média
hammarbergIndexUV	Média
slopeUV0-500	Média
slopeUV500	Média
loudness	Média
VoicedSegmentLengthSec	Média e desvio padrão
UnvoicedSegmentLength	Média e desvio padrão
equivalentSoundLevel_dBp	Média
hurst	Média
DFA	Média
RPDE	Média
PPE	Média

Fonte: Autoria Própria

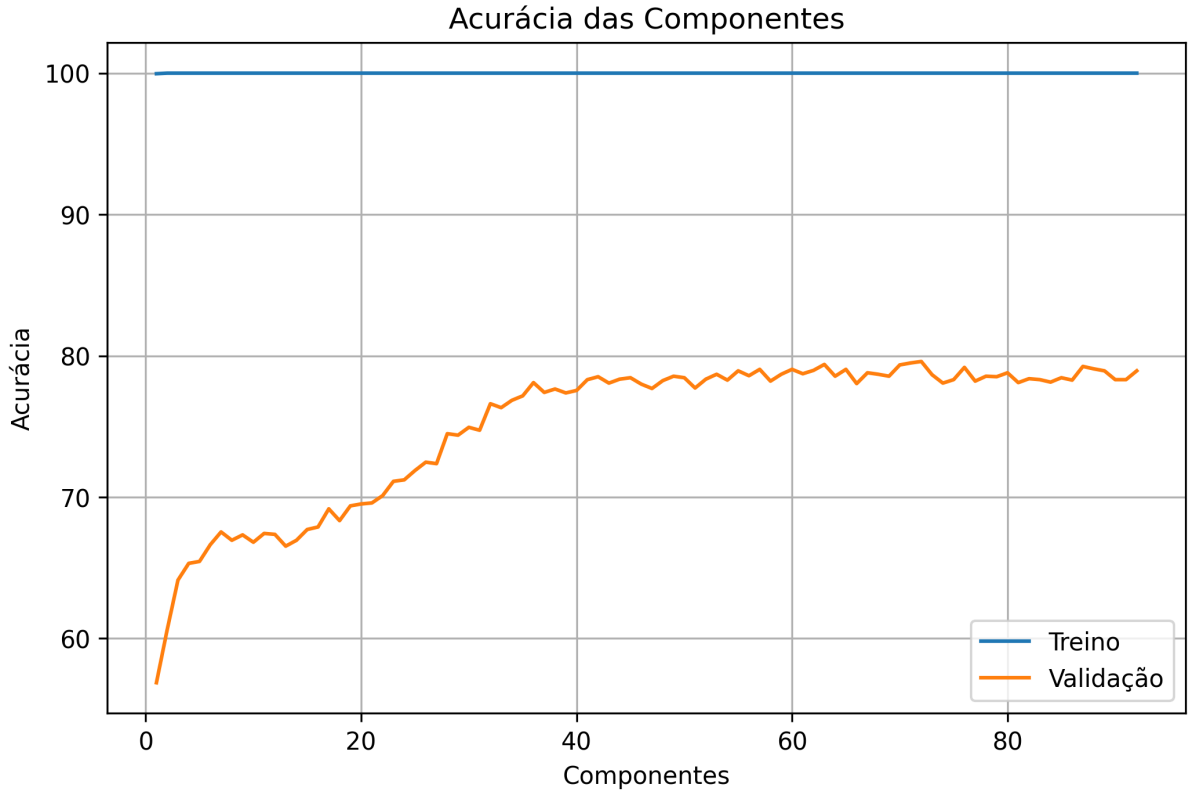
Tabela 29 – Resultados de treinamento e validação utilizando apenas atributos com alta significância

Métrica	Treino	Validação
Acurácia	100%	83.5%
Kappa	100%	62.3%
Precisão	100%	85.2%
Recall	100%	65.5%
F1-score	100%	74.1%

Fonte: Autoria Própria

ponentes. Para cada subconjunto de componentes, um modelo foi treinado e suas acurácias de treinamento e validação foram extraídas. A Figura 24 apresenta os resultados obtidos. De forma geral, a acurácia de validação nunca foi superior a 80% e por isso optou-se por abandonar essa abordagem.

Figura 24 – Evolução da acurácia de treinamento e validação do modelo utilizando PCA



Fonte: Autoria própria

Tabela 30 – Atributos com alta relevância estatística e sem alta correlação

Atributo	Métrica
F0semitoneFrom27.5Hz	80º percentil e pctlrange0-2
Loudness	Desvio padrão, 20º percentil, desvio padrão da curva ascendente, desvio padrão da curva descendente epctlrange0-2
SpectralFlux	Média
MFCC4	Média
Shimmer	Média e desvio padrão
Jitter	Média e desvio padrão
HNR	Média
LogRelF0-H1-A3	Média
F1bandwidth	Média e desvio padrão
F1amplitudeLogRelF0	Média
F2frequency	Média e desvio padrão
F2bandwidth	Desvio padrão
F3frequency	Média e desvio padrão
F3bandwidth	Média e desvio padrão
alphaRatioV	Média
slopeV0-500	Média e desvio padrão
slopeV500	Média
spectralFluxV	Desvio padrão
MFCC1V	Média
MFCC3V	Média
hammarbergIndexUV	Média
slopeUV0-500	Média
slopeUV500	Média
loudness	Média
VoicedSegmentLengthSec	Desvio padrão
UnvoicedSegmentLength	Média e desvio padrão
equivalentSoundLevel_dBp	Média
hurst	Média
DFA	Média
PPE	Média

Fonte: Autoria Própria

Tabela 31 – Resultados de treinamento e validação após análise de correlação

Métrica	Treino	Validação
Acurácia	100%	82.1%
Kappa	100%	58.8%
Precisão	100%	84.2%
Recall	100%	61.9%
F1-score	100%	71.4%

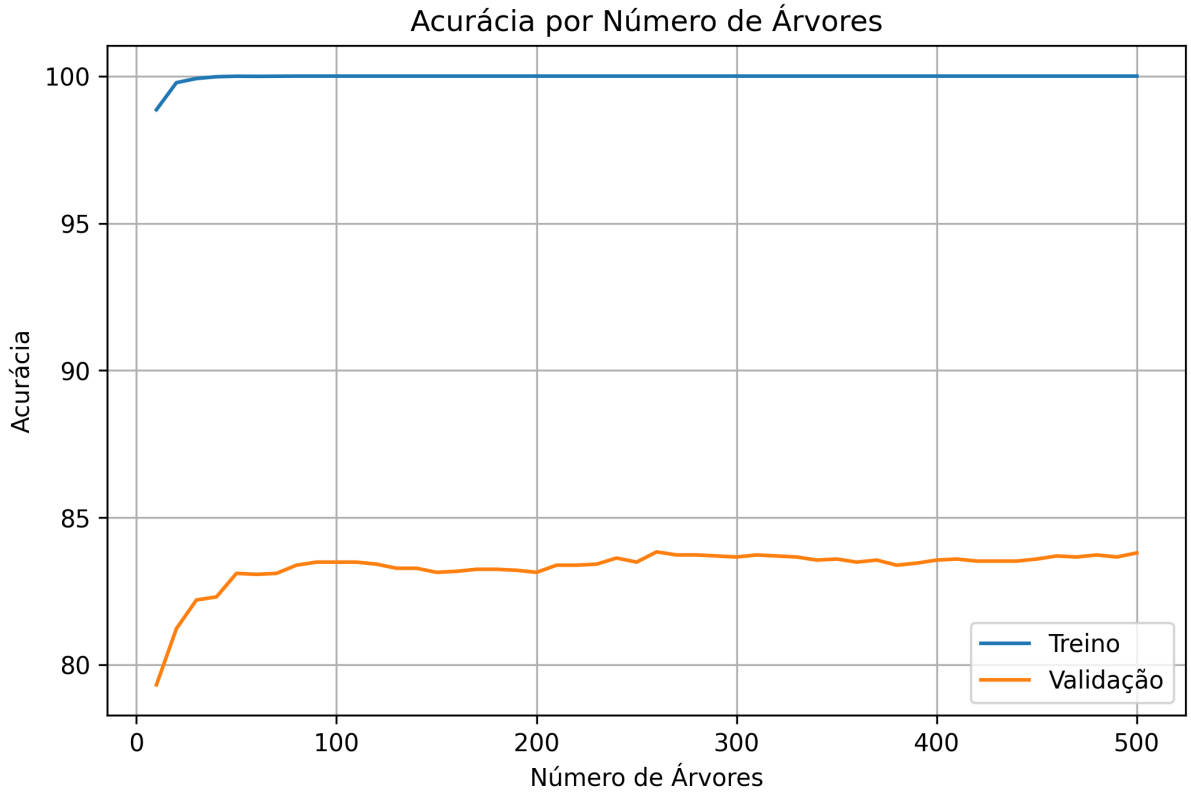
Fonte: Autoria Própria

4.3.3 Hiperparametrização do modelo de *Random Forest*

O processo de hiperparametrização do modelo foi iniciado considerando-se apenas os atributos com alta relevância estatística, uma vez que eles foram os que apresentaram o melhor resultado de acurácia. O primeiro hiperparâmetro trabalhado foi o número de árvores. Um número inferior de árvores aproxima a Random Forest de uma árvore de decisão simples, enquanto que um número elevado, em teoria, pode diminuir a variância do modelo.

Mantendo o número de amostras por folha em 1, o número de atributos por *split* em 9 e usando a Gini como função de impureza nodal, foram testados todos os números de árvores entre 10 e 500 em intervalos de 10 em 10. Para cada teste, um modelo foi treinado e validado e as acurácias de treinamento e validação foram recolhidas. A Figura 25 apresenta os resultado obtidos. Observa-se um crescimento acelerado na acurácia de validação entre 10 e 100 árvores, seguido de uma estabelização por volta de 250. O pico de acurácia foi de 83.8%, obtido com 260 árvores.

Figura 25 – Evolução da acurácia de treinamento e validação por número de árvores do modelo

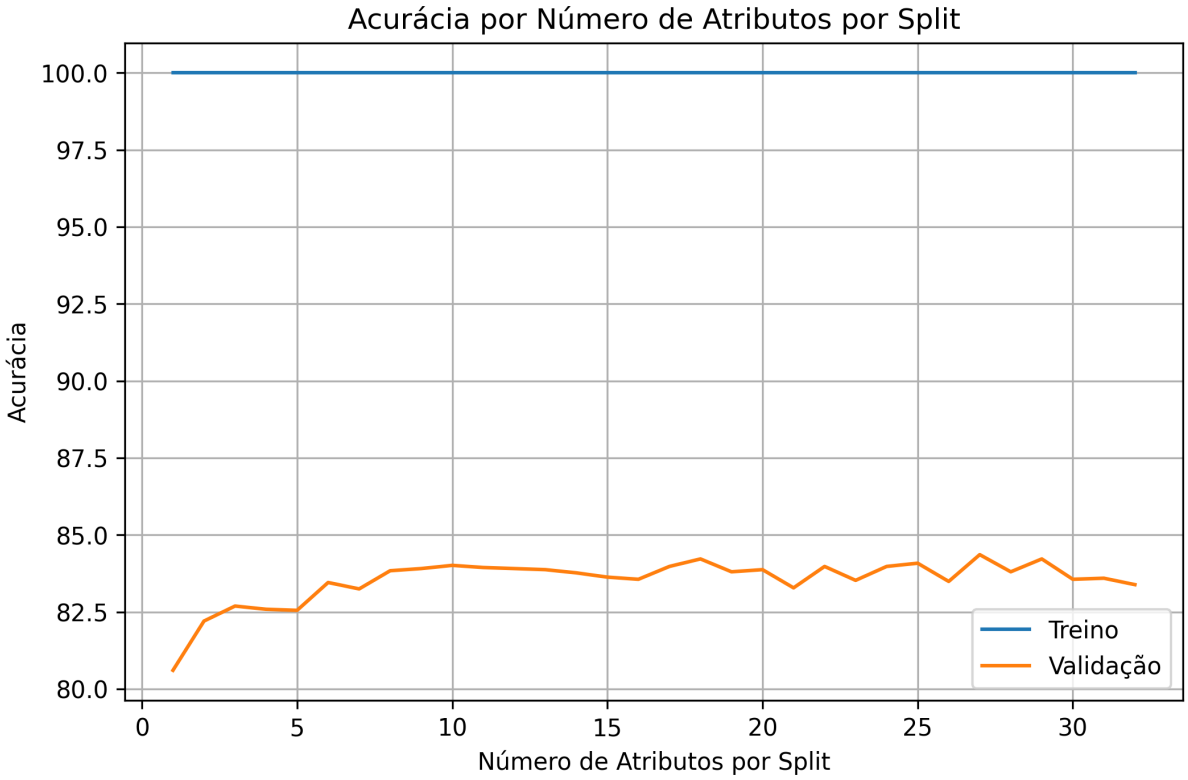


Fonte: Autoria própria

O segundo hiperparâmetro parametrizado foi o número de atributos por *split*. Isso foi feito mantendo-se o número de árvores em 260 (melhor resultado encontrado na etapa anterior), enquanto que os demais hiperparâmetros não sofreram alteração. Foram testados

todos os números de atributos entre 1 e 32. Os resultados são apresentados na Figura 26. O comportamento da curva de acurácia foi semelhante ao detectado anteriormento: um crescimento acelerado seguido de uma estabilização com leves oscilações. O maior valor de acurácia foi de 84.4%, obtido com 27 atributos.

Figura 26 – Evolução da acurácia de treinamento e validação por número de atributos por *split* do modelo

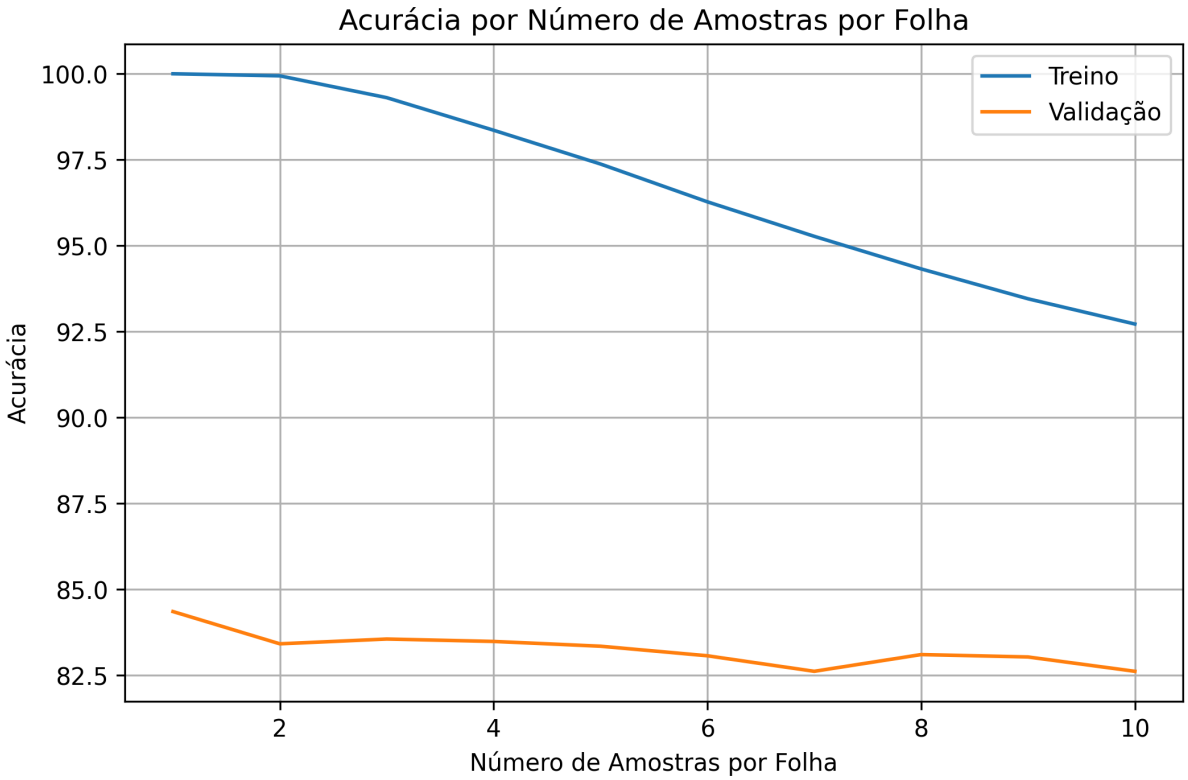


Fonte: Autoria própria

Por fim, o último hiperparâmetro parametrizado foi o número de amostras por nó folha. Isso foi feito preservando-se os melhores resultados encontrados nas duas etapas anteriors, isto é, mantendo-se o número de atributos por *split* em 26 e o número de árvores em 260. Foram testadas todas as quantidades de amostras entre 1 e 10. Os resultados são apresentados na Figura 28. Observa-se que o aumento no número de amostras por nó folha leva a uma diminuição na acurácia de treino. Contudo, isso não foi acompanhado por uma melhoria na acurácia de validação.

A Tabela 32 apresenta os hiperparâmetros da versão final do modelo, treinado apenas com atributos de alta relevância estatística. A Tabela 33 apresenta os resultados de treinamento e validação. Todas as métricas de validação foram incrementadas em pelo menos 1 ponto percentual em comparação com o modelo preliminar. O maior crescimento se deu no Recall,

Figura 27 – Evolução da acurácia de treinamento e validação por número de amostras por nó folha do modelo



Fonte: Autoria própria

cerca de 2.1%, o que indica que o modelo final melhorou sua performance na classificação de amostras provenienetes de portadores da DP. A Figura 28 apresenta graficamente o modelo final.

Tabela 32 – Hiperparâmetros do modelo final de *Random Forest*

Hiperparâmetro	Valor
Número de árvores	260
Função de impureza nodal	Gini
Número de amostras por folha	1
Número de atributos por *split*	27

Fonte: Autoria Própria

4.4 Modelo de CNN

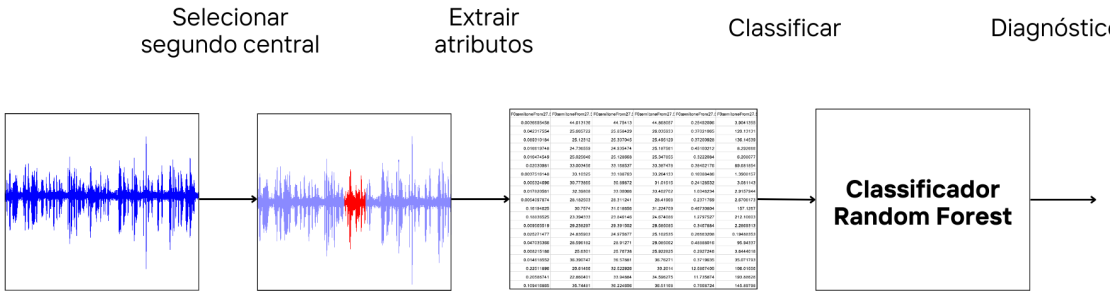
Esta seção traz o processo de treinamento e validação de um modelo de CNN criado com os arrays numpy e tensores gerados a partir de espectrogramas de áudios. A Subseção 4.4.1 apresenta a arquitetura e os resultados de um modelo preliminar de CNN que foi desenvolvido

Tabela 33 – Resultados do modelo final de *Random Forest*

Métrica	Treino	Validação
Acurácia	100%	84.4%
Kappa	100%	64.3%
Precisão	100%	86.2%
Recall	100%	67.2%
F1-score	100%	75.6%

Fonte: Autoria Própria

Figura 28 – Modelo Final de Random Forest



Fonte: Autoria própria

com o intuíto de servir como linha de base. Já a Subseção 4.4.2 aborda o processo de hiperparametrização realizado principalmente por meio de inserção de camadas convolucionais e densas no modelo. Todos os experimentos computacionais realizados nesta seção utilizaram uma máquina virtual no ambiente *Google Colab*. As especificações da máquina utilizada são na Tabela 34.

4.4.1 Modelo preliminar de CNN

Inicialmente, uma CNN simples foi desenvolvida utilizando os arrays numpy dos espectrogramas gerados na Subseção 3.1.3. A arquitetura da rede é composta por uma camada de entrada, que reescalonava os elementos do array de um intervalo entre 0 e 255 para 0 e 1,

Tabela 34 – Especificações do ambiente de treinamento da CNN

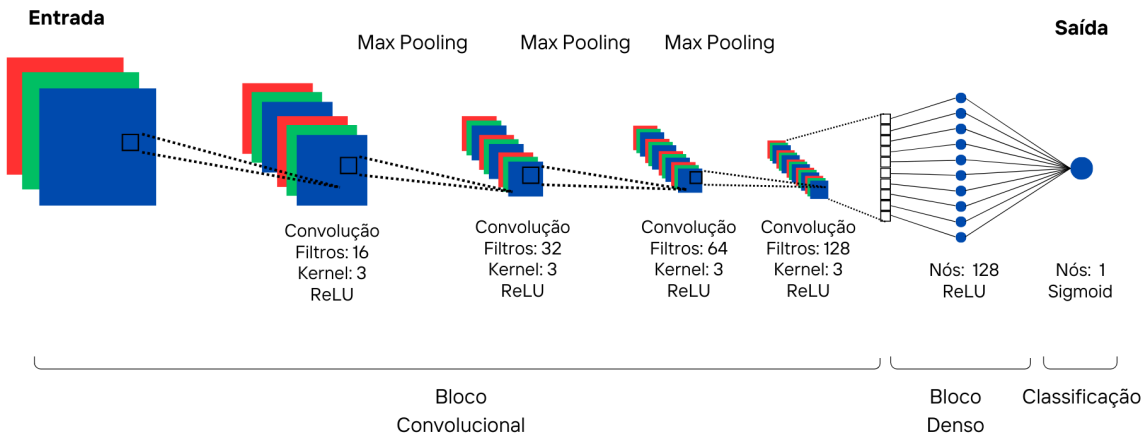
Parâmetro	Valor
RAM do sistema	12.7GB
Disco	112.6GB
RAM da GPU	15GB

Fonte: Autoria Própria

seguida de um bloco convolucional e de um bloco denso.

O bloco convolucional era composto por 4 camadas convolucionais intercaladas por 4 camadas de agrupamento pelo valor máximo. O bloco denso era composto por 2 camadas, sendo a mais externa a responsável pela classificação. A ReLU foi utilizada como função de ativação de todas as camadas, com exceção da última, que utiliza a função Sigmoid. Já a função de *loss* utilizada foi a entropia cruzada binária. O total de parâmetros da rede é de 7323041 e sua arquitetura é apresentada pela Figura 29.

Figura 29 – Arquiterura do modelo preliminar de CNN



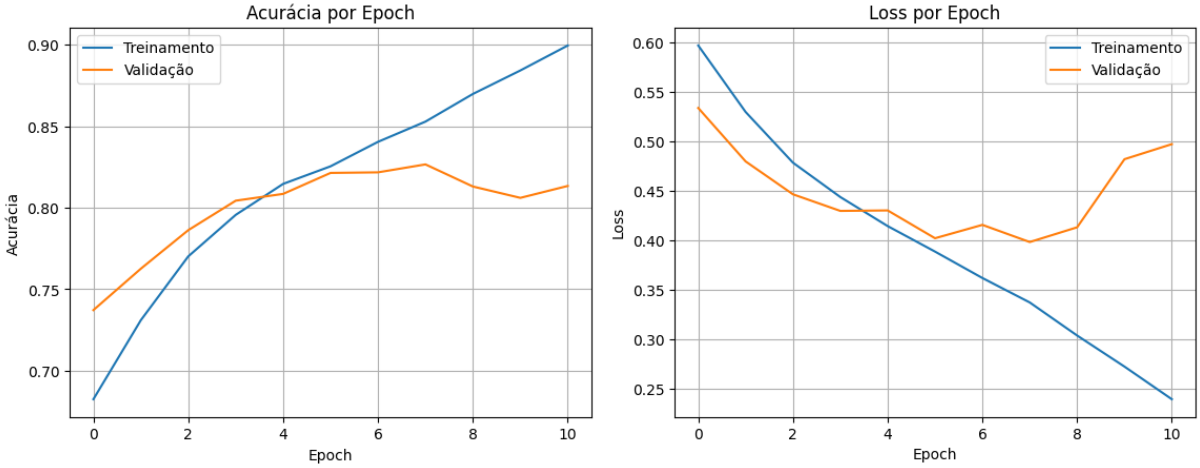
Fonte: Autoria própria

O modelo foi treinado com um batch de tamanho 21. Esse valor foi escolhido por ser um divisor perfeito da quantidade de amostras de validação. A quantidade máxima de *epochs* utilizada foi 50, mas esse número foi inferior na prática, uma vez que utilizou-se um mecanismo de parada de treinamento precoce. Ele consistia em monitar a *loss* de validação. Se ela tivesse seu menor valor superado 3 vezes consecutivas, o treinamento era encerrado.

A Figura 30 apresenta os resultados de acurácia e *loss* por *epoch* do modelo. O treinamento teve um total de 11 *epochs*. Observa-se que a acurácia de treinamento cresceu de forma constante até o fim do treinamento, alcançando um valor máximo de 90%. O inverso ocorreu com a *loss* de treinamento, que caiu de forma constante, chegando a um valor mínimo de cerca de 0.25. A acurácia de validação atingiu um pico de 82.66% na *epoch* 7, mesmo ponto onde a *loss* atingiu seu valor mínimo, e terminou com cerca de 81.33%.

Figura 30 – Evolução da acurácia e *loss* de treinamento e validação por *epoch*

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

Os resultados do modelo preliminar são apresentados na Tabela 35. Todas as métricas pioraram em comparação com o modelo preliminar de *Random Forest*. Em especial, a precisão teve uma queda de 10 pontos percentuais. O índice Kappa está abaixo de 60%, levando a uma concordância moderada.

Tabela 35 – Resultados do modelo preliminar de CNN

Métrica	Validação
Acurácia	81.3%
Kappa	59.1%
Precisão	75.2%
Recall	71.9%
F1-score	73.5%

Fonte: Autoria Própria

4.4.2 Hiperparametrização do modelo de CNN

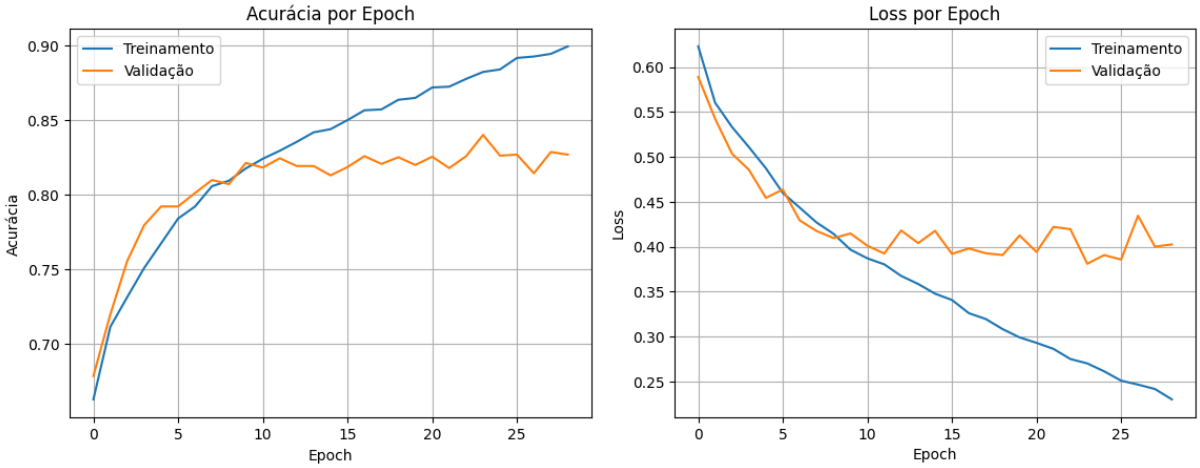
Para melhorar a performance do modelo, o primeiro hiperparâmetro trabalhado foi a quantidade de camadas densas. Isso foi feito considerando-se que a acurácia de treinamento ainda

estava longe de 100%, ou seja, distante de um *overfitting*. Também foi alterada a "paciência" do monitoramento da *loss* de validação. Agora, o treinamento do modelo só seria interrompido caso a *loss* de validação tivesse seu menor valor superado 5 vezes seguidas.

A Figura 31 apresenta a evolução da acurácia e *loss* do modelo. A acurácia de validação teve um pico de 84% na *epoch* 23, mesmo ponto em que a *loss* atingiu seu valor mínimo, cerca de 0.38. As próximas 5 epochs tiveram um valor de *loss* superior ao mínimo e por isso o treinamento foi encerrado na *epoch* 28 com uma acurácia de 82.69%. A acurácia de treinamento evoluiu constantemente, mas ainda se manteve inferior a 90%.

Figura 31 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com 2 camadas densas

(a) Acurácia. (b) *Loss*.



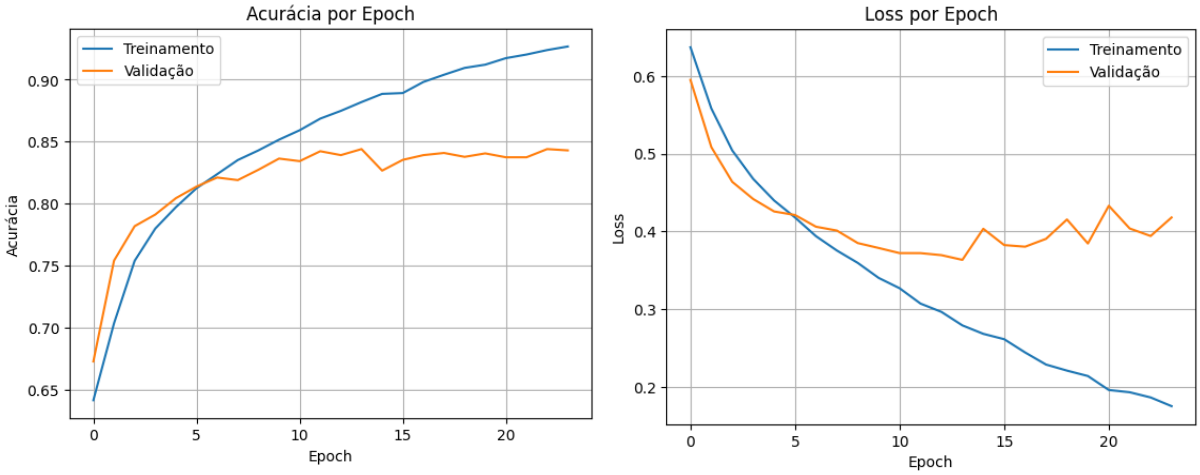
Fonte: Autoria Própria

Como o modelo com 2 camadas densas ainda se manteve relativamente distante de um *overfitting*, optou-se por aumentar ainda mais o número de camadas densas. Dessa vez, foram adicionadas 3 novas camadas. Além disso, para evitar um crescimento demasiado precoce da acurácia de treinamento, foram inseridas camadas de *dropout* entre cada par de camadas densas e entre cada camada de agrupamento e convolucional. Cada camada de *dropout* removia aleatoriamente 20% das entradas da camada posterior. Por fim, a paciência do monitoramento da *loss* de validação foi aumentada para 10.

A Figura 32 apresenta a evolução da acurácia e *loss* do modelo. A acurácia de validação chegou a um pico de 84.39%. em dois momentos distintos. O primeiro ocorreu na *epoch* 13, sendo este o mesmo ponto em que se obteve o menor valor para a *loss* de validação, cerca de 0.36. O segundo pico ocorreu na *epoch* 22 e o treinamento foi finalizado na *epoch* 23, com uma acurácia final de 84.29%. A acurácia de treinamento ultrapassou a barreira dos 90%, chegando ao valor máximo de 92.67%. A *loss* de treinamento continua caindo de forma constante.

Figura 32 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com 5 camadas densas

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

Por fim, foi inserida uma nova camada no bloco denso, além de duas novas camadas no bloco convolucional, sendo uma convolucional e outra de agrupamento pelo valor máximo, preservando a estrutura original. Dessa forma, o modelo final ficou com 6 camadas densas e 5 convolucionais. A arquitetura do modelo é apresentada na Figura 33.

A Figura 34 apresenta a evolução da acurácia e *loss* do modelo. Dessa vez, o treinamento não foi encerrado precocemente, pois a *loss* de validação chegou em seu valor mínimo apenas na *epoch* 43. A acurácia de validação chegou a um pico de 85.4% na última *epoch*, sendo esta a primeira vez em que o pico de acurácia não ocorreu de forma simultânea ao vale de *loss*.

Embora a acurácia de treinamento tenha chegado ao seu valor máximo apenas na última *epoch*, ela não cresce mais de forma constante. Comportamento semelhante ocorre com a *loss* de treinamento, que agora não cai mais de forma constante. Esse é um indicativo de que o modelo está se aproximando de um *overfitting*. Por isso, optou-se por encerrar o treinamento nesse ponto.

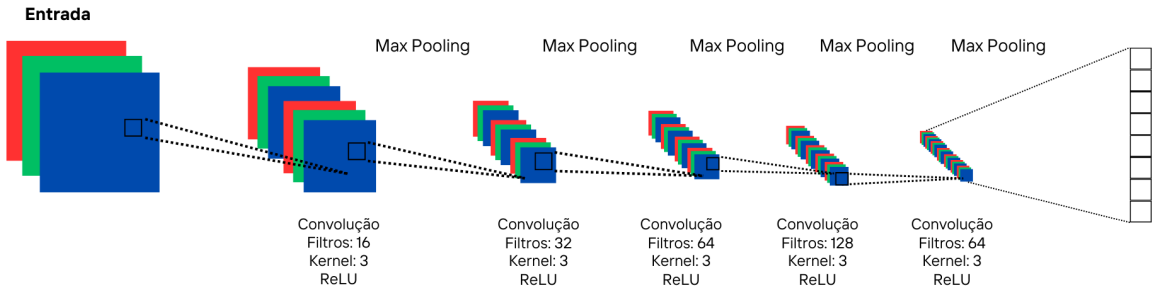
A Tabela 36 apresenta os resultados do modelo. Com exceção da precisão, que caiu 2.2 pontos percentuais, todas as métricas melhoraram em comparação com o modelo final de *Random Forest*. O aumento mais expressivo se deu no *recall*, que foi incrementado em mais de 6 pontos. Isso indica que o modelo melhorou bastante na classificação de pacientes portadores da DP. A Figura 35 apresenta graficamente o modelo final.

4.5 Aumento de dados

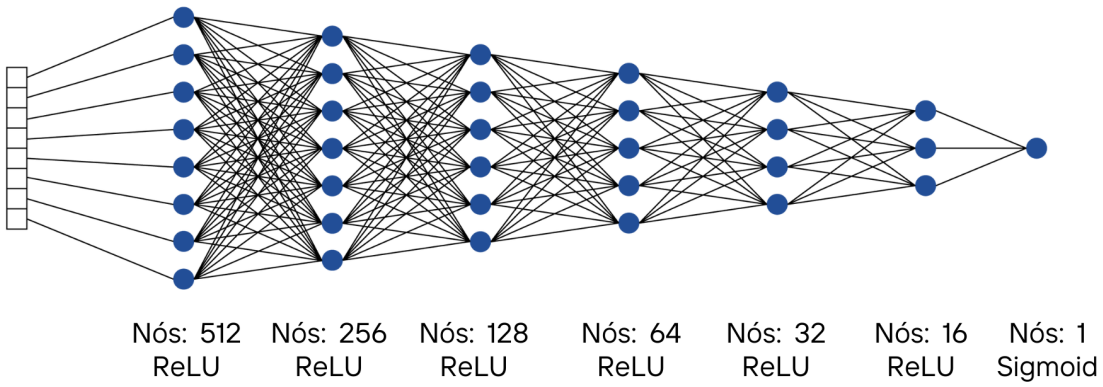
Esta seção traz o processo de aumento de dados (também citado pelo termo em inglês *data augmentation*) realizado para melhorar a performance do modelo final de CNN. Além disso,

Figura 33 – Arquitetura do modelo final

(a) Bloco convolucional.



(b) Bloco denso.

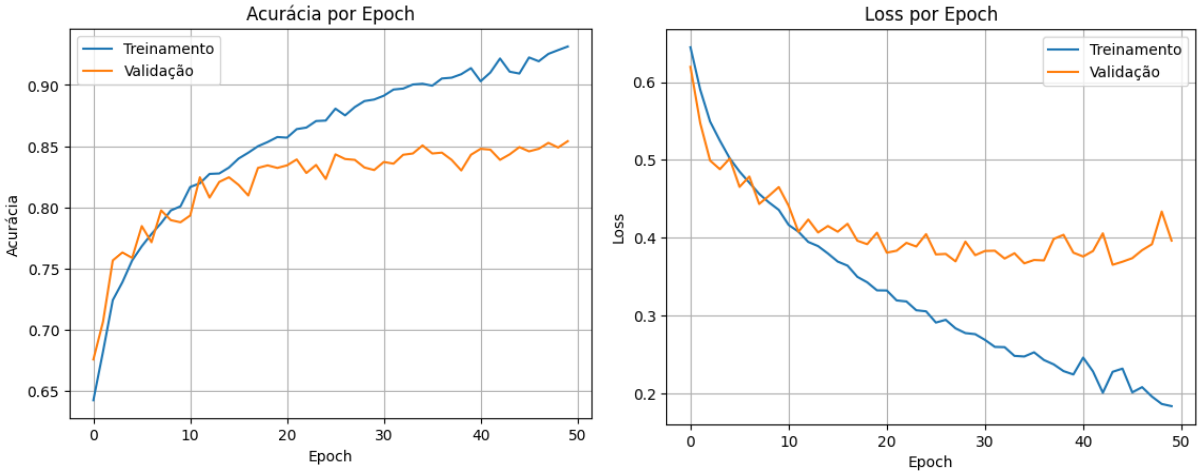


Fonte: Autoria Própria

também é apresentado a mudança realizada nas amostras de teste e validação. Os experimentos foram realizados na mesma máquina apresentada na Seção 4.4.

Figura 34 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com 6 camadas densas

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

Tabela 36 – Resultados do modelo final de CNN

Métrica	Validação
Acurácia	85.4%
Kappa	67.4%
Precisão	84.0%
Recall	73.4%
F1-score	78.4%

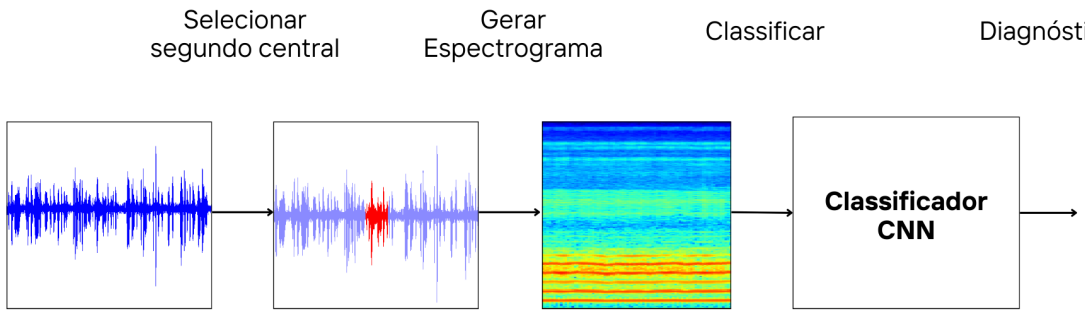
Fonte: Autoria Própria

4.5.1 Amostras de treinamento

A quantidade de amostras de treinamento é um fator crítico de performance de um modelo de CNN. O modelo final apresentado na Subseção 3.5.2 possui 3623649 parâmetros. Esse número supera em mais de 100 vezes a quantidade de amostras de treinamento (23011 amostras). Um processo de *data augmentation* foi realizado para melhorar essa relação e, consequentemente, a performance do modelo.

O processo realizado consiste em utilizar outros instantes das gravações para obter novos espectrogramas. Isso foi feito partindo-se do pressuposto de que cada instante de uma gravação poderia ser tratada de forma independente dos demais. O primeiro teste foi feito adicionando-se espectrogramas extraídos do período entre os instantes 3.5 e 4.5 segundos à base de treinamento. Esses espectrogramas foram extraídos apenas de gravações em que os espectrogramas do segundo central já fazia parte da base de treinamento. As amostras de validação não foram alteradas. A Figura 36 apresenta graficamente o processo.

Figura 35 – Modelo Final de CNN



Fonte: Autoria própria

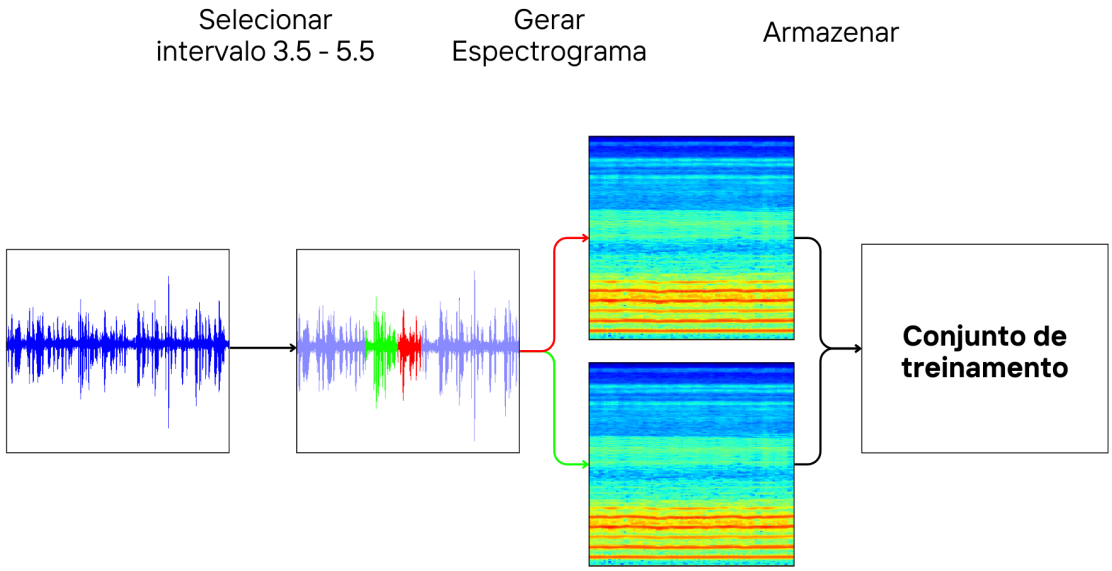
Com isso, a base de treinamento teve a sua quantidade de amostras dobradas, chegando a 46022. Um novo processo de treinamento e validação foi realizado com a nova base, sendo a mudança do tamanho do *batch* de 21 para 128 a única alteração realizada no modelo. A Figura 37 apresenta a evolução de acurácia e *loss* do modelo. Observa-se que a acurácia de validação se manteve de forma mais constante acima da faixa de 85%, sendo que seu valor máximo foi atingindo na última *epoch*, cerca de 85.8%. A acurácia e *loss* de treinamento voltaram a ter um comportamento de crescimento e queda constante.

O resultado do modelo é apresentado na Tabela 37. A métrica mais afetada foi o *recall*, que subiu mais de 5 pontos percentuais, chegando ao valor de 78.7%. Isso indica que o modelo melhorou sua performance na classificação de portadores da DP. A precisão foi a única métrica que piorou, caindo aproximadamente 3 pontos percentuais. De forma geral, a quantidade de verdadeiros e falsos positivos aumentou, sendo que os verdadeiros positivos tiveram um crescimento substancialmente maior.

Como os resultados obtidos foram considerados satisfatórios, optou-se por repetir o processo. Foram adicionadas amostras correspondentes aos espectrogramas extraídos do período entre os instantes 5.5 e 6.5 segundos. A Figura 38 apresenta graficamente o processo realizado.

Um novo treinamento foi realizado utilizando-se a nova base de treinamento, que agora contava com 69033 amostras. O tamanho do *batch* foi mais uma vez alterado para 256, sendo

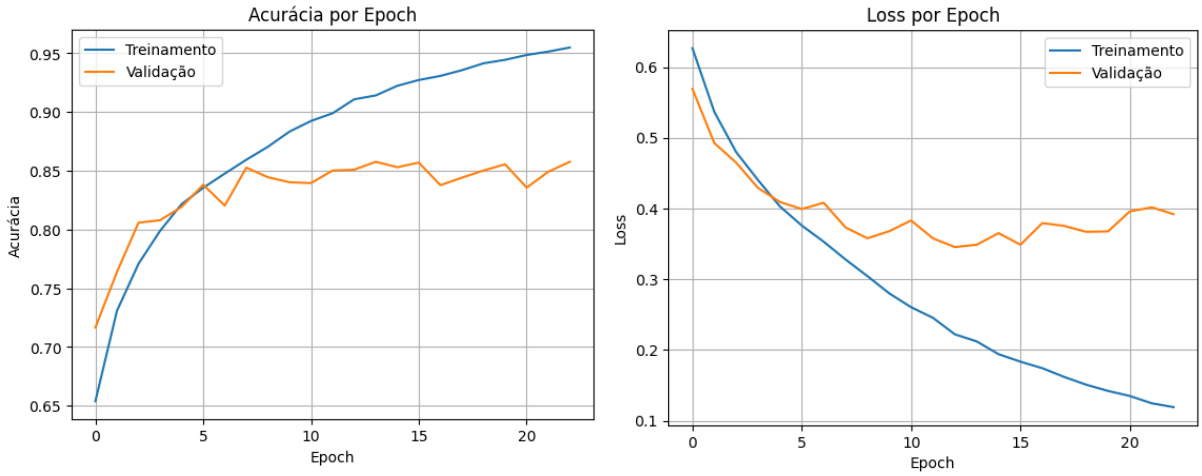
Figura 36 – Data augmentation - 2 amostras



Fonte: Autoria própria

Figura 37 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com o dobro de amostras de treinamento

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

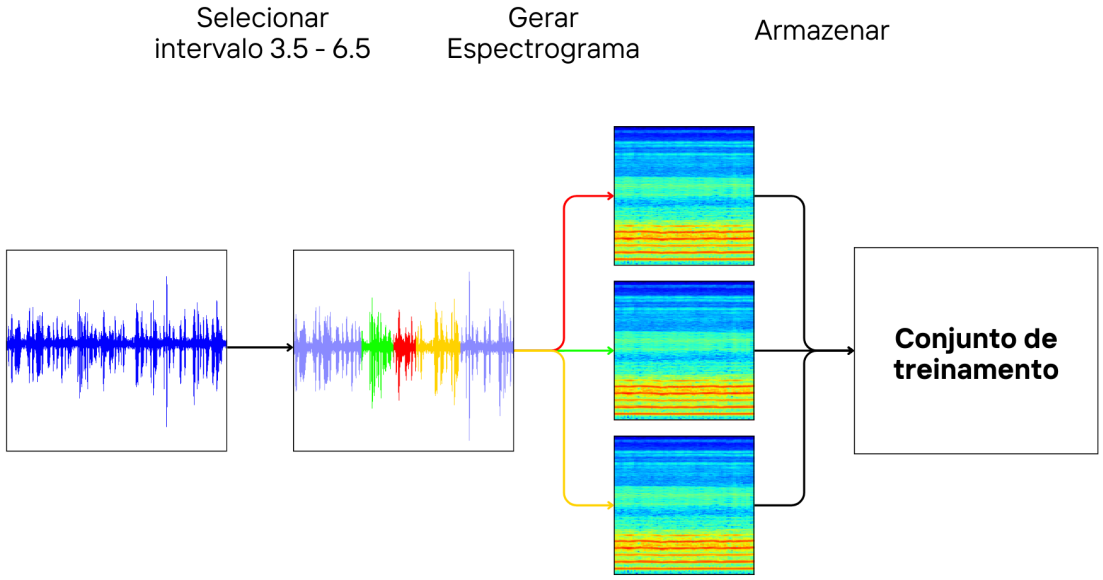
esse o valor máximo possível considerando as limitações de RAM da máquina utilizada. Os resultados de acurácia e *loss* são apresentados na Figura 39. É possível observar que a acurácia de validação se consolidou acima de 85%, chegando a um pico de 86.9% na última *epoch*. Além

Tabela 37 – Resultados do modelo com o dobro de amostras de treinamento

Métrica	Validação
Acurácia	85.7%
Kappa	68.9%
Precisão	81.2%
Recall	78.7%
F1-score	79.9%

Fonte: Autoria Própria

Figura 38 – Data augmentation - 3 amostras



Fonte: Autoria própria

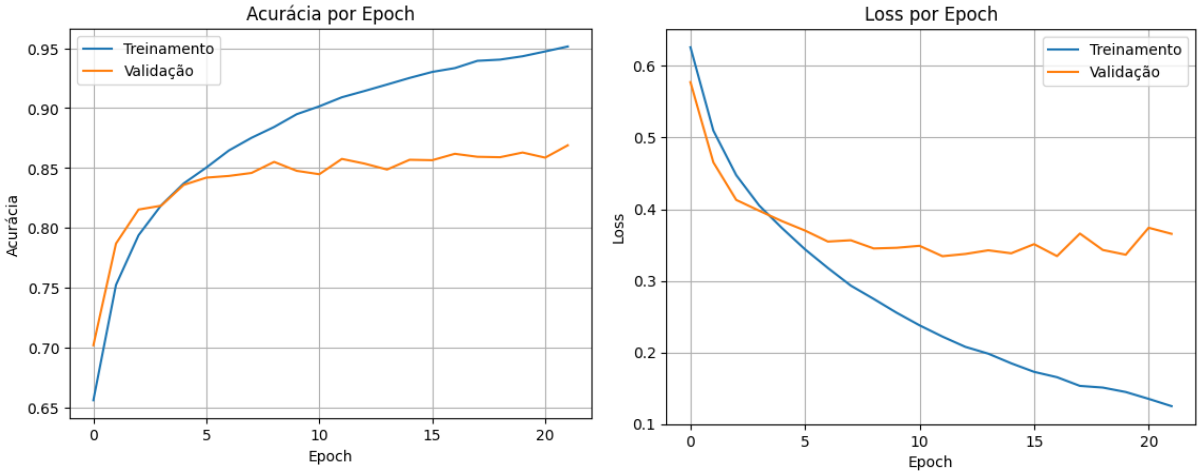
disso, mesmo com a *loss* de validação tendo o seu valor mínimo superado 10 vezes consecutivas, a acurácia de validação aparenta ter ainda mais margem para crescer, caso o treinamento não seja interrompido precocemente. A acurácia e *loss* de treinamento continuam apresentando um comportamento de crescimento e queda constante.

Os resultados do modelo são apresentados na Tabela 38. Pela primeira vez, o índice Kappa e o F1-score ultrapassaram, respectivamente, as barreiras de 70% e 80%. O *recall* recuou 2.5 pontos percentuais, mas isso foi compensado por uma aumento de quase 5 pontos na precisão.

Por fim, tentou-se extrair o máximo de espectrogramas de cada gravação. Foi possível obter 6 espectrogramas de cada áudio. Todos com a duração de 1 segundo, iniciando-se no instante 1.5 segundos até o instante 7.5 segundos. Os instantes fora desse intervalo foram

Figura 39 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com o triplo de amostras de treinamento

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

Tabela 38 – Resultados do modelo com o triplo de amostras de treinamento

Métrica	Validação
Acurácia	86.9%
Kappa	70.9%
Precisão	85.9%
Recall	76.2%
F1-score	80.8%

Fonte: Autoria Própria

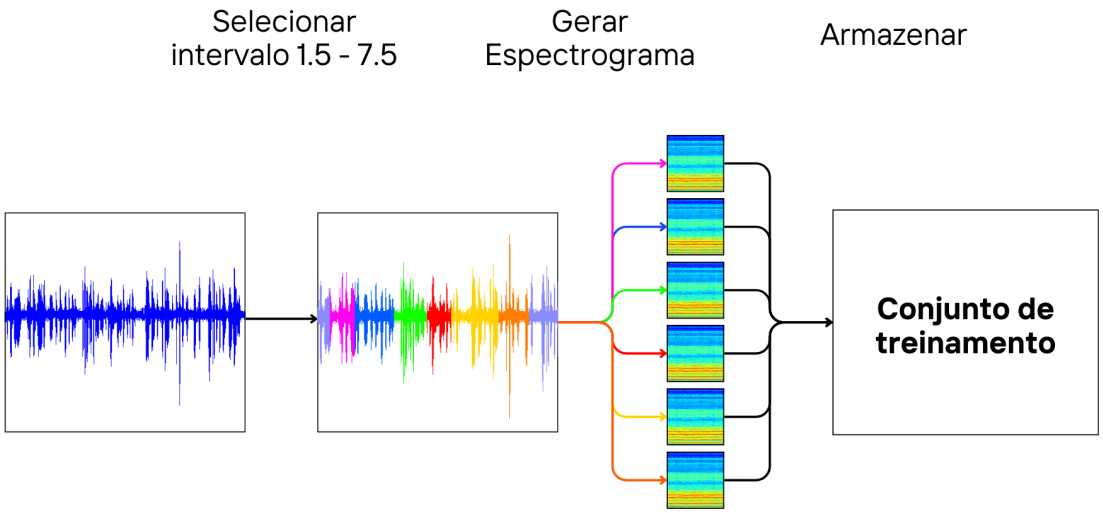
descartados por não possuírem áudio com a qualidade mínima necessária à extração. A Figura 40 exemplifica o processo de formação do novo conjunto de treinamento.

4.5.2 Amostras de validação

Levando em consideração os achados na etapa anterior, optou-se por verificar se o instante central das gravações de fato era o mais adequado para a validação do modelo. Para isso, o modelo final de CNN apresentado na subseção anterior foi validado com cada um dos 6 instantes disponíveis.

Os resultados são apresentados na Tabela 39. Os espectrogramas extraídos entre os instantes 1.5 e 2.5 segundos obtiveram os melhores resultados de acurácia, índice Kappa e F1-score. Além disso, esse mesmo conjunto obteve o segundo melhor resultado para a precisão e o terceiro para o *recall*. Por isso, esse conjunto foi escolhido para substituir o instante central (4.5-5.5 segundos) como conjunto de validação. A Figura 41 apresenta a formação final dos

Figura 40 – Novo conjundo de treinamento



Fonte: Autoria própria

novos conjuntos de validação e teste.

Tabela 39 – Resultados do modelo com o triplo de amostras de treinamento

Métrica	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5
Acurácia	88.0%	87.8%	87.9%	86.9%	87.6%	86.8%
Kappa	73.6%	73.1%	73.5%	70.9%	72.9%	71.2%
Precisão	85.3%	84.9%	84.4%	85.9%	84.0%	82.9%
Recall	80.4%	80.3%	81.4%	76.2%	81.1%	79.8%
F1-score	82.8%	82.5%	82.8%	80.8%	82.5%	81.3%

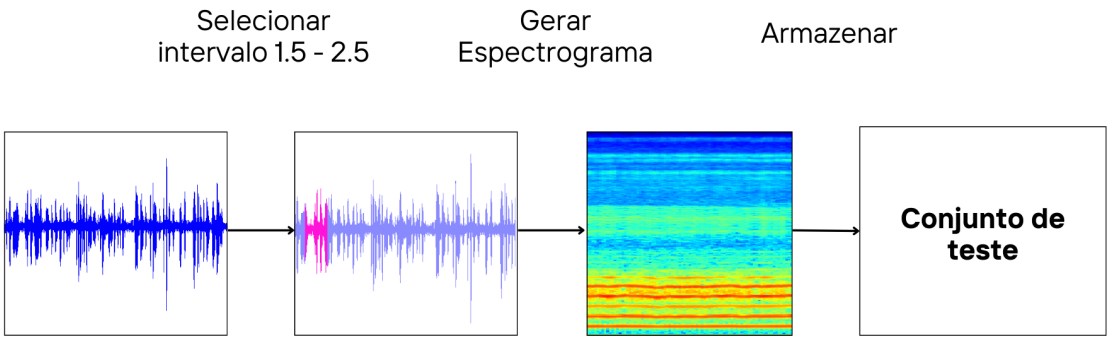
Fonte: Autoria Própria

4.5.3 Treinamento com base completa

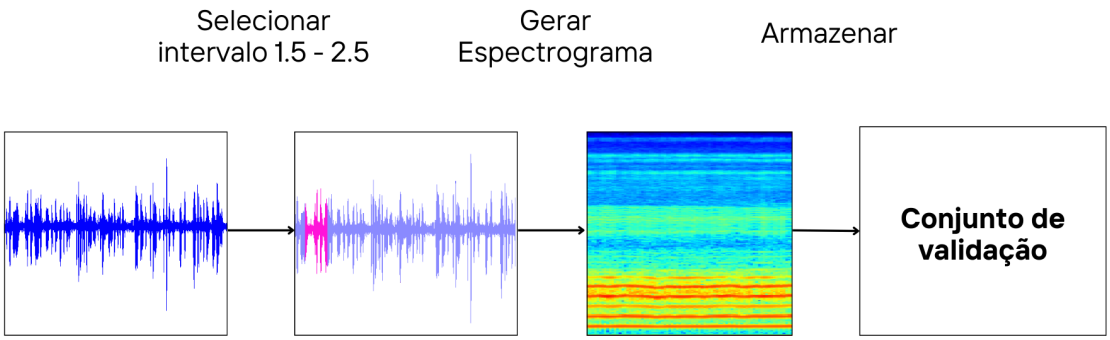
Utilizando os novos conjuntos de treinamento e validação gerados na etapa anterior, o modelo de CNN foi novamente treinado com algumas alterações. A primeira foi do monitoramento da *loss* de validação e, consequentemente, da parada precoce. Agora o modelo seria treinado por 30 *epochs*, independentemente dos resultados obtidos durante o processo. A segunda alteração foi criar um mecanismo que salvasse a configuração da *epoch* que obtivesse o melhor resultado para a acurácia de validação, independentemente dele ter ocorrido antes do final do treinamento.

Figura 41 – Novos conjuntos

(a) Validação



(b) Teste



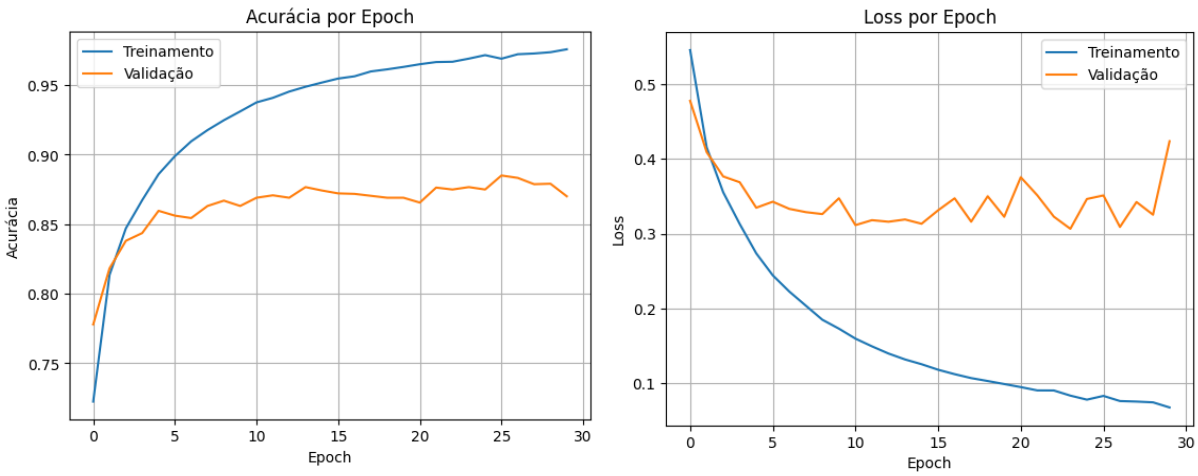
Fonte: Autoria Própria

A Figura 42 apresenta a evolução de acurácia e *loss* de treinamento e validação do modelo. Na última *epoch*, o modelo atingiu uma acurácia de validação de 87%. Contudo, como o melhor resultado ocorreu anteriormente, na *epoch* 25, essa foi a configuração salva para o

modelo. O valor máximo obtido para acurácia de validação foi de 88.49%.

Figura 42 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* com nova base

(a) Acurácia (b) *Loss*



Fonte: Autoria Própria

A Tabela 40 apresenta os resultados obtidos pelo modelo. Todas as métricas ultrapassaram os valores máximos obtidos previamente. Destaque para o índice Kappa e o *recall*, que aumentaram em quase 4 pontos percentuais.

Tabela 40 – Resultados do modelo com nova base

Métrica	Validação
Acurácia	88.5%
Kappa	74.6%
Precisão	86.5%
Recall	80.6%
F1-score	83.4%

Fonte: Autoria Própria

4.6 Comitê de classificadores

Esta seção traz o processo de treinamento e validação de um comitê de classificadores que mescla as características dos algoritmos de *Random Forest* e CNN apresentados nas seções anteriores a fim de se obter melhores resultados que os obtidos pelos modelos individualmente. Os experimentos foram realizados na mesma máquina apresentada na Seção 4.4.

4.6.1 Modelo preliminar de comitê

Um comitê de classificadores foi utilizado com o objetivo de melhorar os resultados obtidos pelos modelos de *Random Forest* e CNN. Isso foi feito utilizando as saídas do modelo de CNN como entradas do modelo de *Random Forest*. De forma mais específica, os espectrogramas de cada um dos 6 instantes considerados válidos (instantes entre 1.5 e 75 segundos das gravações) foram classificados utilizando o melhor modelo de CNN treinado na Seção 3.5.

A classificação de cada instante foi tratado como um novo atributo para ser utilizado no treinamento de um modelo de *Random Forest*. Optou-se por utilizar esses novos atributos em conjunto com o conjunto de 64 atributos com alta relevância estatística apresentados na Subseção 3.3.2. Assim, a nova base ficou composta com 70 atributos. A Figura 43 apresenta o esquema completo do modelo proposto.

O modelo de *Random Forest* foi treinado e validado com essa nova base. Os hiperparâmetros do modelo foram configurados de forma idêntica aos do modelo preliminar apresentado na Subseção 3.3.1. A Tabela 41 apresenta os resultados obtidos. A acurácia se aproximou de 91%. O indíce Kappa está bem próximo de 80%, o que colocaria o modelo na faixa de concordância quase perfeita. O *recall* teve um aumento bem expressivo, quase 4 pontos percentuais.

Tabela 41 – Resultados do modelo de comitê

Métrica	Validação
Acurácia	90.8%
Kappa	79.9%
Precisão	89.1%
Recall	84.8%
F1-score	86.9%

Fonte: Autoria Própria

4.6.2 Hiperparametrização do modelo de comitê

De forma análoga ao apresentado na Subseção 3.3.3, o modelo de *Random Forest* foi hiperametrizado com o objetivo de melhorar a performance do modelo preliminar. A Tabela 42 apresenta a configuração final dos hiperparâmetros do modelo. Com exceção do número de atributos por *split*, todos os hiperparâmetros tiveram seus valores preservados, quando comparados com a configuração do modelo de *Random Forest* apresentado na Tabela 32.

A Tabela 43 apresenta os resultados finais de validação do modelo. Com exceção do *recall*, todas as métricas cresceram. Contudo, a variação foi pouco relevante, sempre inferior a 1 ponto percentual.

Tabela 42 – Hiperparâmetros do modelo final de comitê

Hiperparâmetro	Valor
Número de árvores	260
Função de impureza nodal	Gini
Número de amostras por folha	1
Número de atributos por *split*	3

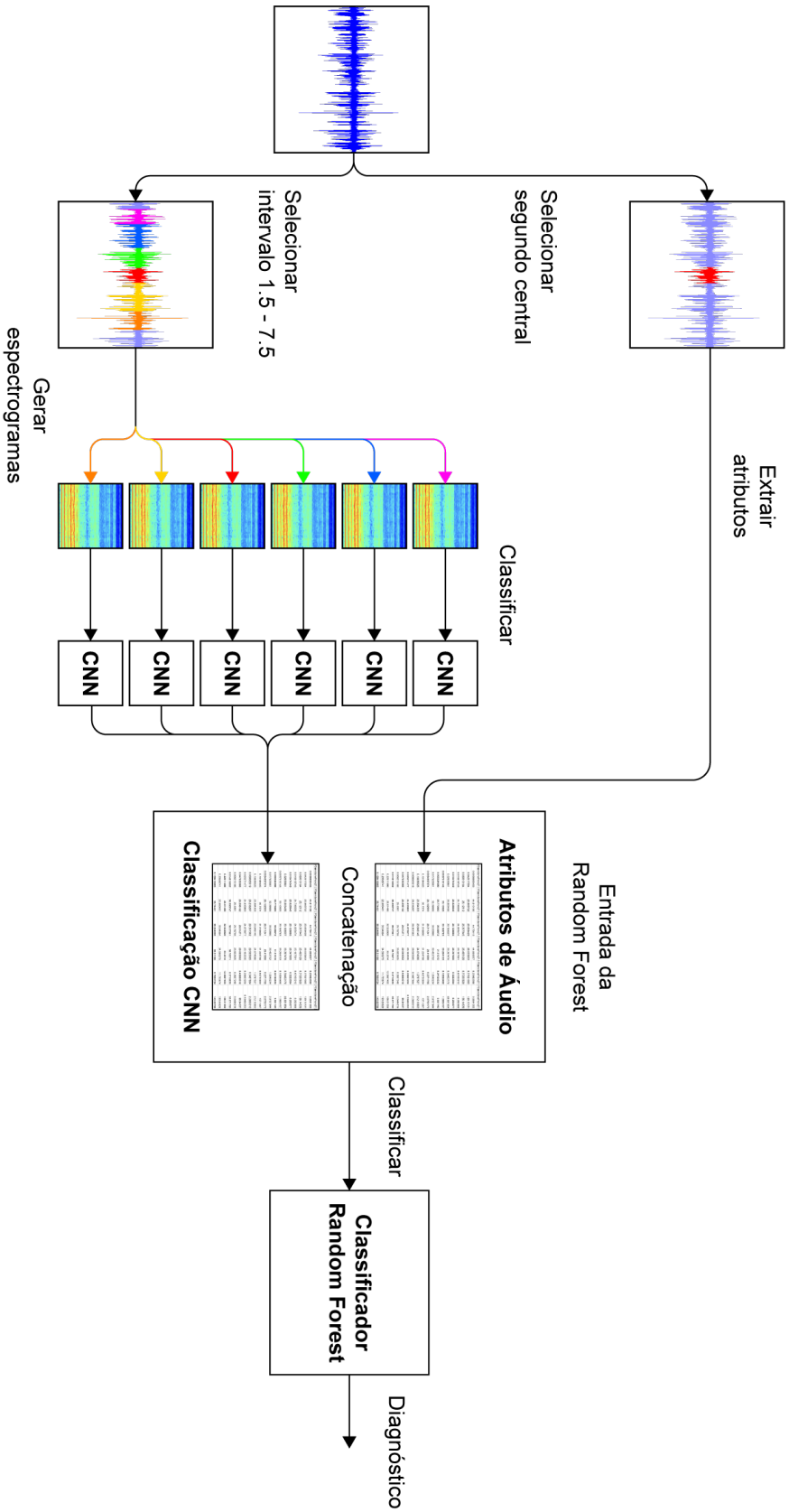
Fonte: Autoria Própria

Tabela 43 – Resultados do modelo final de comitê

Métrica	Validação
Acurácia	90.9%
Kappa	80.0%
Precisão	89.6%
Recall	84.4%
F1-score	87.0%

Fonte: Autoria Própria

Figura 43 – Modelo proposto



Fonte: Autoria própria

5 AVALIAÇÃO DO MODELO

Este capítulo apresenta os principais resultados obtidos pelos modelos quando avaliados utilizando os dados do conjunto de teste. A Seção 5.1 trata do processo de treinamento e teste do melhor modelo da referência com os conjuntos de dados utilizados na pesquisa. A Seção 5.2 apresenta os resultados dos modelos propostos, além de abordar as discussões levantadas pela pesquisa.

5.1 Comparabilidade de resultados

Esta seção apresenta todas as etapas realizadas nos modelos propostos por Karaman et al. (2021) que possibilitaram a comparabilidade entre os seus resultados e os resultados dos modelos propostos. De forma mais específica, a Subseção 5.1.1 trata do modelo com arquitetura DenseNet121, enquanto que a Subseção 5.1.2 apresenta o modelo com arquitetura ResNet50V2. Por fim, a Subseção 5.1.3 apresenta o modelo com arquitetura SqueezeNet1_1.

5.1.1 Arquitetura DenseNet121

Como os critérios de seleção de amostras utilizados nesta pesquisa divergiram dos critérios utilizados por Karaman et al. (2021), os seus modelos precisaram ser retreinados com os novos conjuntos de dados propostos de forma a permitir a comparabilidade dos resultados encontrados.

Inicialmente, as gravações de áudio dos conjuntos de treinamento, validação e teste foram transformadas em espectrogramas utilizando as mesmas técnicas apresentadas por Karaman et al. (2021). A principal distinção entre esse método e o utilizado nesta pesquisa é que toda a duração (10 segundos) da gravação é usada na geração do espectrograma.

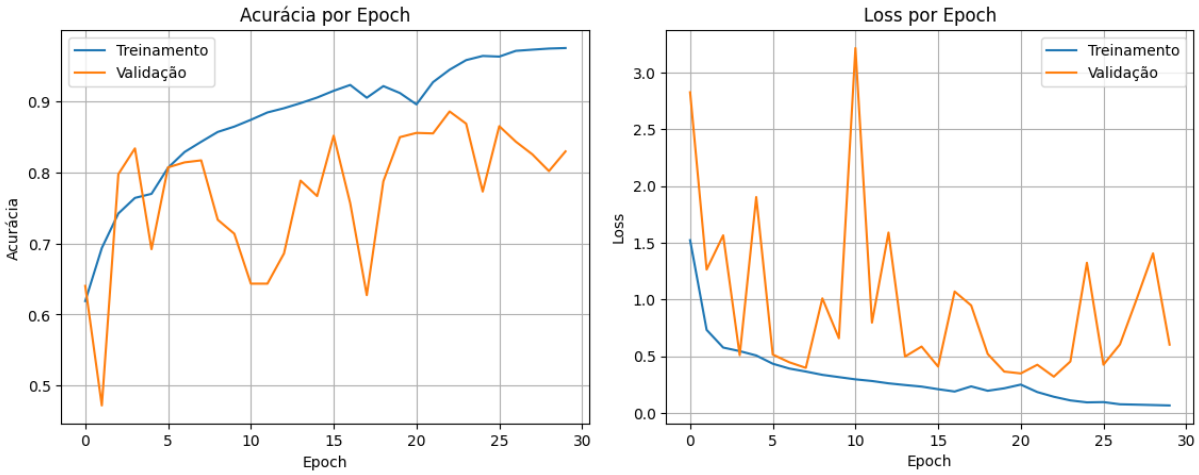
O primeiro modelo abordado foi o que faz uso da arquitetura DenseNet121. O modelo utilizado passou por um processo de aprendizado transferido, tendo sido previamente treinado com o conjunto de dados ImageNet. O modelo foi treinado de acordo com o reportado por Karaman et al. (2021), com uma única diferença no número de *epochs* utilizado. O modelo original foi treinado com 24 *epochs* e isso foi alterado para 30. Essa mudança não traz nenhum prejuízo ao modelo e pode até alavancar os resultados obtidos.

Além disso, como o tamanho do *batch* não foi informado, utilizou o tamanho 32, que era o tamanho máximo possível para a quantidade de RAM disponível na máquina utilizada na pesquisa. A Figura 44 apresenta a evolução da acurácia e *loss* de treinamento e validação do modelo. Observa-se que os resultados de validação oscilam bastante, enquanto que os resultados de treinamento apresentam um comportamento mais linear. A acurácia de treinamento chegou a

um ápice de 97.47% na última *epoch*, um claro indicativo de *overfitting*. Já acurácia de validação atingiu seu valor máximo na *epoch* 22, cerca de 88.5%.

Figura 44 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* do modelo DenseNet121

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

A Tabela 44 apresenta os resultados de validação e treinamento do modelo. Observa-se que todos os resultados de teste são inferiores aos de validação, sendo o índice Kappa o mais afetado, com uma queda acima de 3 pontos percentuais. As outras métricas sofreram uma queda na faixa de 2 pontos percentuais. Uma possível explicação para essa queda de performance é a oscilação que o modelo apresentou durante o seu treinamento.

Tabela 44 – Resultados do modelo DenseNet121

Métrica	Validação	Teste
Acurácia	88.5%	86.9%
Kappa	75.4%	72.0%
Precisão	81.5%	79.5%
Recall	87.9%	85.6%
F1-score	84.6%	82.5%

Fonte: Autoria Própria

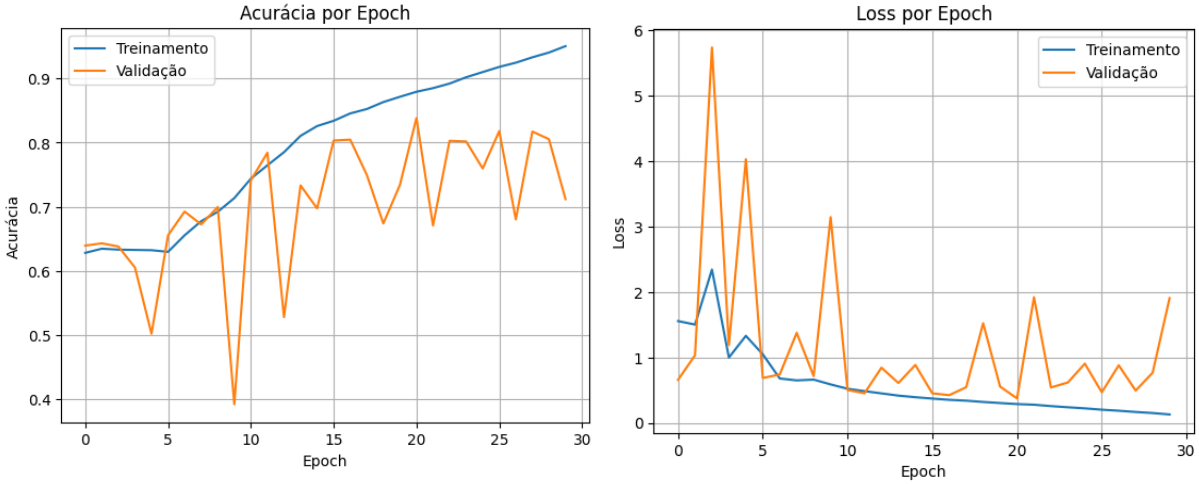
5.1.2 Arquitetura ResNet50V2

De forma análoga ao apresentado na Subseção 5.1.1, um modelo com arquitetura ResNet50V2 foi retreinado. A evolução de acurácia e *loss* de treinamento e validação do modelo é apresentada na Figura 45. Assim como no caso do modelo com arquitetura DenseNet121,

a acurácia e *loss* de validação oscilam bastante. Tanto a acurácia de treinamento quanto a de validação atingiram picos menores, alcançando respectivamente os valores 95.5% e 83.8%.

Figura 45 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* do modelo ResNet50V2

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

A Tabela 45 apresenta os resultados de validação e treinamento do modelo. Observa-se que todos os resultados de teste e validação pioraram em relação ao modelo DenseNet121. Contudo, com exceção do *recall*, todos os resultados de teste são superiores aos de validação. Isso pode ser um indicativo de que esse modelo é melhor em generalizar as suas classificações.

Tabela 45 – Resultados do modelo ResNet50V2

Métrica	Validação	Teste
Acurácia	83.8%	84.3%
Kappa	64.5%	65.6%
Precisão	78.6%	80.0%
Recall	75.5%	75.3%
F1-score	77.0%	77.6%

Fonte: Autoria Própria

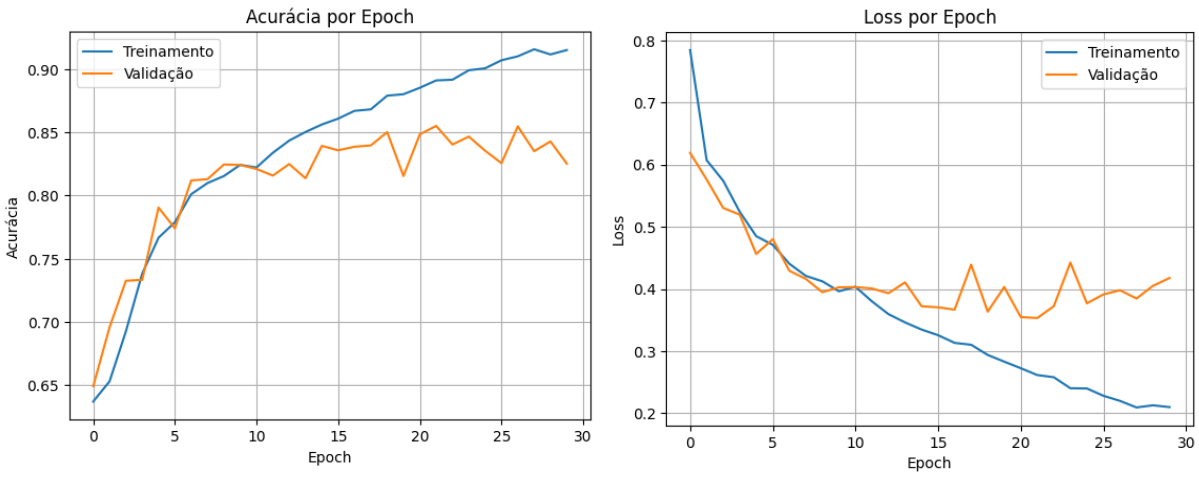
5.1.3 Arquitetura SqueezeNet1_1

O modelo de arquitetura SqueezeNet1_1 foi o último a ser retreinado e seguiu o mesmo processo dos modelos anteriores. A Figura 46 apresenta a evolução de acurácia e *loss* de treinamento e validação do modelo. A acurácia e *loss* de validação apresentam baixa oscilação quando comparadas com as dos outros modelos. O valor máximo atingido para a acurácia de validação foi de aproximadamente 85.4%. Já acurácia de treinamento atingiu um pico de 92.4%,

sendo inferior a ambos os modelos anteriores, sendo esse o modelo mais estável e mais distante de um *overfitting*.

Figura 46 – Evolução da acurácia e *loss* de treinamento e validação por *epoch* do modelo ResNet50V2

(a) Acurácia. (b) *Loss*.



Fonte: Autoria Própria

A Tabela 46 apresenta os resultados de validação e teste do modelo. Observa-se que todos os resultados de teste e validação pioraram em relação ao modelo DenseNet121. Contudo, com exceção do *recall*, todos os resultados de teste são superiores aos de validação. Isso pode ser um indicativo de que esse modelo é melhor em generalizar as suas classificações.

Tabela 46 – Resultados do modelo SqueezeNet1_1

Métrica	Validação	Teste
Acurácia	85.4%	85.1%
Kappa	67.7%	67.2%
Precisão	81.8%	81.4%
Recall	76.2%	76.0%
F1-score	78.8%	78.6%

Fonte: Autoria Própria

5.2 Resultados e discussões

Esta seção apresenta os principais resultados obtidos pelos modelos propostos quando avaliados com o conjunto de teste e as discussões levantadas pela pesquisa. Esses resultados são destrinchados e comparados com os principais resultados encontrados na literatura na Subseção 5.2.1. Já a Subseção 5.2.2 trata da validade da utilização dos atributos de áudio para diagnóstico da DP, enquanto que a Subseção 5.2.3 trata da aplicabilidade da pesquisa.

5.2.1 Resultados

O modelo final de *Random Forest* apresentado na Seção 4.6, o modelo final de CNN pós processo de *data augmentation* apresentado na Seção 4.5 e o comitê de classificadores apresentado na Seção 3.6 foram avaliados por meio do conjunto de teste. A Tabela 47 apresenta os resultados dos modelos propostos, além dos resultados dos modelos de Karaman et al. (2021) treinados com a nova base, de acordo com o apresentado na Seção 5.1. Os resultados são apresentados em ordem crescente de acurácia.

Tabela 47 – Resultados de teste dos principais modelos

Modelo	Acurácia	Kappa	Precisão	Recall	F1-score
Random Forest	83.1%	61.2%	86.0%	63.5%	73.0%
ResNet50V2 (KARAMAN et al., 2021)	83.8%	64.5%	78.6%	75.5%	77.0%
SqueezeNet1_1 (KARAMAN et al., 2021)	85.1%	67.2%	81.4%	76.0%	78.6%
DenseNet121 (KARAMAN et al., 2021)	86.9%	72.0%	79.5%	**85.6%**	82.5%
CNN (Data Augmentation)	89.3%	76.3%	88.9%	80.3%	84.4%
Comitê de classificadores	**91.6%**	**81.5%**	**91.5%**	84.5%	**87.9%**

Fonte: Autoria Própria

O comitê de classificadores apresentou os melhores resultados para todas as métricas, com exceção do *recall*, que ficou cerca de 1 ponto percentual abaixo do modelo DenseNet121. Destaque para a acurácia e precisão, que superaram a barreira dos 90%. O índice Kappa atingiu o valor de 81.1%, o que indica uma concordância quase perfeita. Esse nível de concordância não foi atingido por nenhum dos outros modelos. Observa-se também que a CNN com *data augmentation* ocupa a segunda colocação em todas as métricas, com exceção do *recall*. Isso indica que o processo de *data augmentation* realizado não apenas é válido, mas pode elevar ainda mais os resultados de modelos mais robustos, como a DenseNet121. Outra característica marcante do modelo com CNN e do comitê de classificadores é que em ambos os resultados de teste foram superiores aos de validação. Isso não se verifica nos outros modelos e indica que eles generalizam bem as suas classificações.

A Tabela 48 apresenta os resultados de teste dos modelos de acordo com o reportado por Karaman et al. (2021). O índice Kappa e o F1-score foram omitidos por não estarem presentes no trabalho original. Os resultados estão ordenados por acurácia. O comitê de classificadores ainda apresenta os melhores resultados para acurácia e precisão. O modelo DenseNet121 ultrapassou a CNN proposta por meio ponto percentual. O modelo SqueezeNet1_1 ocupa agora a última posição, apresentando um resultado pior que a *Random Forest* para todas as métricas, com exceção do *recall*. Todos os modelos propostos por Karaman et al. (2021) reportaram acurácia e precisão inferiores a 90%. A ResNet50V2 e a DenseNet121 apresentam *recall* na faixa de 90%.

Tabela 48 – Resultados de teste dos principais modelos e reportados por Karaman et al. (2021)

Modelo	Acurácia	Precisão	Recall
SqueezeNet1_1 (KARAMAN et al., 2021)	72.8%	72.2%	74.0%
Random Forest	83.1%	86.0%	63.5%
ResNet50V2 (KARAMAN et al., 2021)	88.5%	87.3%	90.0%
CNN (Data Augmentation)	89.3%	88.9%	80.3%
DenseNet121 (KARAMAN et al., 2021)	89.8%	88.4%	91.5%
Comitê de classificadores	91.6%	91.5%	84.5%

Fonte: Autoria Própria

5.2.2 Validade dos atributos de áudio

Wang et al. (2020) argumenta que os resultados obtidos por pesquisas que fizeram uso de atributos de áudio em bases menores, como é o caso da base de Little et al. (2009), não se verificam quando submetidos a uma base de dados mais robusta, como a do estudo mPower. As pesquisas de Sakar e Kursun (2009), Das (2010), Melo e Gouveia (2023) e Govindu e Palwe (2023) reportaram acurácias acima de 90 enquanto que Wang et al. (2020) atingiu um resultado inferior a 60% utilizando um conjunto de atributos semelhante.

Contudo, foi encontrada uma acurácia de 83.1% com o modelo de *Random Forest* aqui proposto treinado utilizando um conjunto de atributos de áudio. Embora esse resultado seja inferior aos resultados reportados por pesquisas que utilizaram a base de Little et al. (2009), ele é bem superior ao resultado reportado por Wang et al. (2020). Dessa forma, conclui-se que os atributos de áudio possuem validade para o problema de classificação de sinais de voz aplicados no dianóstico da DP.

Além disso, a diferença de performance obtida entre as bases de Little et al. (2009) e do estudo mPower pode ser explicada pela baixa qualidade das gravações do estudo mPower. Enquanto que a primeira base apresentava os atributos previamente calculados e validados, obtidos de gravações em um ambiente extremamente controlado e com um equipamento de gravação robusto, a segunda é composta por gravações de áudio em estado bruto, obtidas por pacientes voluntários que gravavam as fonações sustentadas por meio de uma aplicativo em seus celulares. A falta de rigor na gravações é verificada na inutilização de algumas amostras, como apresentado na Seção 4.1, e pode facilmente ser responsável pelo decrescimento nas acurácias reportadas.

5.2.3 Aplicabilidade

A Tabela 49 apresenta a sensibilidade e especificidade do comitê de classificadores, conforme o verificado no conjunto de teste, além da prevalência da DP, como reportado por BRASIL. Ministério da Saúde (2017). Também são apresentados seus respectivos símbolos estatísticos. Essas medidas são usadas em conjunto no teorema de Bayes para verificar a real

funcionalidade do modelo. Ou seja, a probabilidade de um indivíduo ser portador da DP dado um diagnóstico positivo do modelo.

Tabela 49 – Medidas para aplicação do teorema de Bayes.

Medida	Símbolo	Valor
Sensibilidade	$P(+\|D)$	84.5%
Especificidade	$P(-\|D^C)$	95.6%
Prevalência	$P(D)$	0.2%

Fonte: Autoria Própria

Aplicando o teorema com esses valores, encontra-se que a probabilidade do indivíduo ser portador da DP dado um diagnóstico positivo é de aproximadamente 3.7%. Isso é um problema usual de testes de doenças raras e poderia sugerir falsamente de que o modelo não possui aplicabilidade. Contudo, levando-se em consideração que o processo de *data augmentation* aumentou consideravelmente a performance do modelo de CNN, conforme apresentado na Seção 4.5, é possível argumentar que cada segundo de gravação pode ser tratado de forma independente dos demais. Ou seja, os segundos seriam condicionalmente independentes.

Partindo-se desse pressuposto, conclui-se que o mesmo pode ser dito a respeito de gravações distintas. Dessa forma, um mesmo indivíduo poderia efetuar múltipas gravações de voz para serem submetidas ao modelo. Um sequência de diagnósticos positivos aumentaria a probabilidade do paciente de fato ser portador da DP. Nesse contexto, o teorema de Bayes para múltiplos testes, sendo n a quantidade de testes realizada, pode ser obtido pela fórmula apresentada na seguinte equação:

$$P(D|+) = \frac{P(+|D)^n P(D)}{P(+|D)^n P(D) + [1 - P(-|D^C)]^n [1 - P(D)]} \tag{11}$$

A Tabela 50 apresenta a probabilidade de um indivíduo ser portador da DP dado n diagnósticos positivos. Observa-se que com apenas 2 testes, a probabilidade aumenta mais de 10 vezes. Com 3 testes, tem-se uma probabilidade 93.4% e acima de 4 a probabilidade é aproximadamente 100%. Dessa forma, aplicando-se 3 ou mais testes, fica evidente a aplicabilidade do modelo no diagnóstico da DP.

Tabela 50 – Probabilidade de indivíduo ser portador dado n testes com diagnóstico positivo.

n	Probabilidade
1	3.7%
2	42.5%
3	93.4%
4	99.6%

Fonte: Autoria Própria

6 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo o desenvolvimento de um modelo de Aprendizado de Máquina capaz de classificar sinais de voz provenientes de portadores e não portadores da DP. Isso foi realizado por meio de comitê de classificadores que mescla as características de uma CNN, treinada com espectrogramas, e uma *Random Forest*, treinada com atributos de áudio. O modelo proposto atingiu uma acurácia de 91.6% no conjunto de teste, superando em quase dois pontos percentuais o melhor resultado previamente reportado para a mesma base de dados.

Além disso, também foi apresentado como o modelo poderia ser utilizado na prática para o diagnóstico da DP por meio de múltiplos testes. Um paciente que receba três diagnósticos positivos do modelo tem uma chance de 93.4% de ser portador da DP. Embora não se possa nem se deseje afirmar que a ferramenta é capaz de substituir profissionais da área da saúde, ela pode auxiliar no processo de diagnóstico da DP, que até os dias atuais ainda não conta com um biomarcador definitivo que ateste a sua presença.

Também foi demonstrado que os atributos de áudio se apresentam como uma boa alternativa para a construção de classificadores da DP. O modelo de *Random Forest* apresentado que fazia uso desses atributos atingiu uma acurácia de cerca de 83% no conjunto de teste. Esse valor foi semelhante ao obtido pelo modelo inicial de CNN, cerca de 85.4% no conjunto de validação. É provável que um processo de *data augmentation* semelhante ao realizado no modelo de CNN tenha a capacidade de melhorar os resultados da *Random Forest*. Além disso, uma base de gravações com uma qualidade mais elevada também ajudaria a melhorar o modelo, uma vez que os atributos de áudio parecem ser mais sensíveis a variações nas gravações do que os espectrogamas.

Por fim, vale salientar que a CNN atingiu uma acurácia de 89.3% no conjunto de teste após um processo de *data augmentation*. A técnica aqui apresentada consistia em transformar uma única gravação em 6 entradas que eram tratadas como amostras de treinamento distintas. O modelo utilizado possui uma arquitetura simples com um bloco convolucional seguido de um bloco denso. É provável que uma arquitetura mais robusta, como as encontradas na referência, associada com essa técnica apresente resultados ainda mais elevados. A utilização de um comitê de classificadores também se mostra uma alternativa viável, misturando as características passíveis de serem encontradas nos espectrogramas com as dos atributos de áudio.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDUL, Z. K.; AL-TALABANI, A. K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, v. 10, p. 122136–122158, 2022. Citado na página 20.

ALZUBAIDI, L. et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Jornal of Big Data*, 2021. Citado na página 37.

AMATO, F. et al. Machine learning- and statistical-based voice analysis of parkinson's disease patients: A survey. *Expert Systems with Applications*, v. 219, 2023. ISSN 0957-4174. <https://www.sciencedirect.com/science/article/pii/S0957417423001525>. Citado na página 40.

AREFIN, A. S. *Art of Programming Contest*. Bangladesh: Gyankosh Prokashoni, 2006. ISBN 9843233824. Citado na página 33.

BHATTACHARYA, I.; BHATIA, M. Svm classification to distinguish parkinson disease. *Proceedings of the A2CWiC'10*, Article 14, p. 1–6, 2010. <https://doi.org/10.1145/1858378.1858392>. Citado 3 vezes nas páginas 10, 42 e 47.

BOT, B. M. et al. The mpower study, parkinson disease mobile data collected using researchkit. *Scientific Data*, v. 3, 2016. ISSN 2052-4463. <https://doi.org/10.1038/sdata.2016.11>. Citado 6 vezes nas páginas 40, 47, 48, 50, 52 e 53.

BRASIL. Ministério da Saúde. Portaria conjunta nº 10, de 31 de outubro de 2017. *Lex* — Secretária de Atenção Especializada à Saúde, São Paulo, out. 2017. Citado na página 92.

CAFFO, B. *Statistical Inference for Data Science*. [S.l.]: Leanpub, 2016. Citado na página 30.

DAS, R. A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, v. 37, p. 1568–1572, 2010. ISSN 0957-4174. <https://www.sciencedirect.com/science/article/pii/S0957417409006137>. Citado 4 vezes nas páginas 10, 42, 43 e 92.

DONG, X. et al. A survey on ensemble learning. *Frontiers of Computer Science*, v. 14, n. 2, 2020. Disponível em: <https://doi.org/10.1007/s11704-019-8208-z>. Citado na página 38.

EYBEN, F.; WöLLMER, M.; SCHULLER, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In: *Proceedings of the 18th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2010. (MM '10), p. 1459–1462. ISBN 9781605589336. Disponível em: <https://doi.org/10.1145/1873951.1874246>. Citado 2 vezes nas páginas 48 e 53.

GOVINDU, A.; PALWE, S. Early detection of parkinson's disease using machine learning. *Procedia Computer Science*, v. 218, p. 249–261, 2023. ISSN 1877-0509. <https://www.sciencedirect.com/science/article/pii/S1877050923000078>. Citado 5 vezes nas páginas 10, 43, 44, 45 e 92.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*: Data mining, inference, and prediction. Stanford, CA: Springer, 2008. Citado 4 vezes nas páginas 26, 28, 29 e 34.

HESTERBERG, T. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 3, n. 6, p. 497–526, 2011. Citado na página 34.

HO, A. K. et al. Speech impairment in a large sample of patients with parkinson's disease. *Behavioural Neurology*, v. 11, p. 131–137, 1999. <https://doi.org/10.1155/1999/327643>. Citado 2 vezes nas páginas 10 e 19.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. d. O. Análise de componentes principais: Resumo teórico, aplicação e interpretação. v. 5, p. 83–90, jun. 2016. Disponível em: <https://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398>. Citado na página 38.

KARAMAN, O. et al. Robust automated parkinson disease detection based on voice signals with transfer learning. *Expert Systems with Applications*, v. 178, 2021. ISSN 0957-4174. <https://www.sciencedirect.com/science/article/pii/S0957417421004541>. Citado 12 vezes nas páginas 7, 10, 11, 25, 48, 49, 50, 55, 58, 87, 91 e 92.

KHERIF, F.; LATYPOVA, A. Chapter 12 - principal component analysis. In: MECHELLI, A.; VIEIRA, S. (Ed.). *Machine Learning*. Academic Press, 2020. p. 209–225. ISBN 978-0-12-815739-8. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128157398000122>. Citado na página 38.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, p. 159–174, 1977. <https://pubmed.ncbi.nlm.nih.gov/843571/>. Citado na página 31.

LENAIN, R. et al. Surfboard: Audio feature extraction for modern machine learning. *ArXiv*, 2020. ISSN 2005.08848. <https://doi.org/10.48550/arXiv.2005.08848>. Citado na página 19.

LITTLE, M. A. et al. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, v. 56, n. 4, p. 1015–1022, 2009. ISSN 0018-9294. <https://doi.org/10.1109/TBME.2008.2005954>. Citado 9 vezes nas páginas 10, 11, 16, 19, 40, 42, 50, 53 e 92.

LITTLE, M. A. et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, v. 6, 2007. <https://doi.org/10.1186%2F1475-925x-6-23>. Citado na página 11.

LITTLE, M. A. et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, v. 6, 2007. <https://doi.org/10.1186%2F1475-925x-6-23>. Citado na página 21.

LOPES, G. A. d. M. *Segmentação de voz em ambientes ruidosos utilizando análise de imagem do espectrograma*. 2013. Citado na página 22.

MASSANO, J.; CABREIRA, V. Doença de parkinsonn: Revisão clínica e atualização. *Acta Médica Portuguesa*, v. 32, 2019. <https://doi.org/10.20344/amp.11978>. Citado 6 vezes nas páginas 10, 11, 12, 14, 15 e 16.

MELO, M.; GOUVEIA, T. Classificação de sinais de voz para auxílio no diagnóstico dadoença de parkinson. *Revista Brasileira de Computação Aplicada*, 2023. Citado 6 vezes nas páginas 10, 44, 45, 46, 47 e 92.

MOHRI, M.; RASTAMIZADEH, A.; TALWALKAR, A. *Foudantions of Machine Learning*. London: MIT Press, 2018. ISBN 9780262018258. Citado 5 vezes nas páginas 26, 28, 29, 32 e 33.

NASRI, F. O envelhecimento populacional no brasil. *Einstein*, v. 24, 2008. <https://pesquisa.bvsalud.org/portal/resource/pt/lil-516986>. Citado 2 vezes nas páginas 12 e 16.

NGO, Q. C. et al. Computerized analysis of speech and voice for parkinson's disease: A systematic review. *Computer Methods and Programs in Biomedicine*, v. 226, 2022. ISSN 0169-2607. <https://www.sciencedirect.com/science/article/pii/S0169260722005144>. Citado na página 40.

OLIVEIRA, C. I. d. Modelos híbridos para classificar imagens histológicas: uma associação de deep features por transferência de aprendizado com comitê de classificadores. Universidade Estadual Paulista (Unesp), 2021. Citado na página 39.

OZKAN, V. A comparison of classification methods for telediagnosis of parkinson's disease. *Entropy*, v. 18, 2016. ISSN 1099-4300. <https://doi.org/10.3390/e18040115>. Citado na página 21.

PARKINSON, J. An essay on the shaking palsy. *J Neuropsychiatry Clin Neurosci*, v. 14, 1817. ISSN 0895-0172. <https://doi.org/10.1176/jnp.14.2.223>. Citado na página 14.

POLAT, K.; GüNEş, S. A new feature selection method on classification of medical datasets: Kernel f-score feature selection. *Expert Systems with Applications*, v. 36, n. 7, p. 10367–10373, 2009. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417409000840>. Citado na página 31.

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v. 6, p. 21–45, 2006. Disponível em: <https://api.semanticscholar.org/CorpusID:18032543>. Citado na página 39.

RIAD, R. et al. Vocal markers from sustained phonation in huntington's disease. *ArXiv*, 2020. ISSN 2006.05365. <https://doi.org/10.48550/arXiv.2006.05365>. Citado na página 10.

ROCHE. *Dia Mundial do Parkinson*. [S.l.], 2018. Available at <https://www.roche.com.br/pt/por-dentro-da-roche/dia-mundial-do-parkinson.html>. Citado na página 16.

RUAN, F. et al. Deep learning for real-time image steganalysis: a survey. *Journal of Real-Time Image Processing*, 2019. Citado na página 35.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, v. 8, n. 4, p. e1249, 2018. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249>. Citado 2 vezes nas páginas 38 e 39.

SAKAR, C. O.; KURSUN, O. Telediagnosis of parkinson's disease using measurements of dysphonia. *Journal of Medical Systems*, v. 34, p. 591–599, 2009. <https://doi.org/10.1007/s10916-009-9272-y>. Citado 4 vezes nas páginas 10, 41, 43 e 92.

SCHMIT, J. M. et al. Deterministic center of pressure patterns characterize postural instability in parkinson's disease. *Experimental Brain Research*, v. 168, n. 3, 2006. Https://doi.org/10.1007/s00221-005-0094-y. Citado na página 22.

SILVA, R. S. de Souza e. *CLASSIFICADOR PARA ESTEGANÁLISE COMO FERRAMENTA DA PERÍCIA FORENSE COMPUTACIONAL*. 2023. Citado 2 vezes nas páginas 35 e 36.

SOUZA, V. et al. Análise comparativa de redes neurais convolucionais no reconhecimento de cenas. *XI Computer on the Beach. Santa*, 2020. Citado na página 35.

STEIDL, E. M. d. S.; ZIEGLER, J. R.; FERREIRA, F. V. Doença de parkinson: Revisão bibliográfica. *Disciplinarum Scientia*, v. 8, 2007. ISSN 2177-3355. <https://periodicos.ufn.edu.br/index.php/disciplinarumS/article/view/921/865>. Citado 6 vezes nas páginas 11, 14, 15, 16, 17 e 18.

SUSMAGA, R. Confusion matrix visualization. *Intelligent Information Processing and Web Mining*, v. 218, p. 107–116, 2004. <https://link.springer.com/chapter/10.1007/978-3-540-39985-8_12>. Citado na página 29.

SUTTON, C. D. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, Elsevier, v. 24, p. 303–329, 2005. Citado na página 34.

TAI, Y. C. et al. A voice analysis approach for recognizing parkinson's disease patterns. *IFAC-PapersOnLine*, v. 54, p. 382–387, 2021. ISSN 2405-8963. <https://www.sciencedirect.com/science/article/pii/S2405896321016918>. Citado na página 11.

TEIXEIRA, J. P.; OLIVEIRA, C.; LOPES, C. Vocal acoustic analysis: Jitter, shimmer and hnr parameters. *Procedia Technology*, v. 9, 2013. ISSN 2212-0173. <https://www.sciencedirect.com/science/article/pii/S2212017313002788>. Citado 3 vezes nas páginas 19, 20 e 21.

TELES, V. de C.; ROSINHA, A. C. U. 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*, v. 12, p. 523–530, 2008. <https://arquivosdeorl.org.br//conteudo/acervo_port.asp?id=567>. Citado na página 18.

WANG, M. et al. Robust feature engineering for parkinson disease diagnosis: New machine learning techniques. *JMIR Biomed Eng*, v. 5, n. 1, p. e13611, Jul 2020. ISSN 2561-3278. Disponível em: <https://biomedeng.jmir.org/2020/1/e13611>. Citado 5 vezes nas páginas 21, 22, 48, 50 e 92.

WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, v. 48, p. 2839–2846, 2015. ISSN 0031-3203. <https://www.sciencedirect.com/science/article/pii/S0031320315000989>. Citado 2 vezes nas páginas 29 e 30.

YIU, T. *Understanding Random Forest*: How the algorithm works and why it is so effective. 2019. Disponível em: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>. Acesso em: 13 SET. 2021. Citado na página 34.

ZHOU, X. et al. Linear versus mel frequency cepstral coefficients for speaker recognition. In: *2011 IEEE Workshop on Automatic Speech Recognition Understanding*. [S.l.: s.n.], 2011. p. 559–564. Citado na página 20.