UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELOS DE DECISÃO E SAÚDE

DANILO RANGEL ARRUDA LEITE

DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO DA TIPOLOGIA DOS SINAIS VOCAIS COM BASE NO DEEP LEARNING

João Pessoa/PB

2022

DANILO RANGEL ARRUDA LEITE

DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO DA TIPOLOGIA DOS SINAIS VOCAIS COM BASE NO DEEP LEARNING

Tese apresentada ao Programa de Pós-graduação em Modelos de Decisão e Saúde – Nível Doutorado, do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba – UFPB, como requisito para obtenção do título de Doutor.

Área de Concentração: Modelos de Decisão

Orientador: Prof. Dr. Leonardo Wanderley Lopes
Orientador: Prof. Dr. Ronei Marcos de Moraes

João Pessoa/PB

2022

Catalogação na publicação
Seção de Catalogação e Classificação

L533d Leite, Danilo Rangel Arruda.
Desenvolvimento de um modelo de classificação da tipologia dos sinais vocais com base no Deep Learning / Danilo Rangel Arruda Leite. - João Pessoa, 2022.
82 f.

Orientação: Ronei Marcos de Moraes, Leonardo Wanderley Lopes.
Tese (Doutorado) - UFPB/CCEN.

1. Inteligência Artificial. 2. Espectrogramas. 3. Tipologia da voz. 4. Deep Learning. 5. Grad-CAM. I. Moraes, Ronei Marcos de. II. Lopes, Leonardo Wanderley. III. Título.

UFPB/BC CDU 004.8(043)

Elaborado por Gracilene Barbosa Figueiredo - CRB-15/794

DANILO RANGEL ARRUDA LEITE

DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO DA TIPOLOGIA DOS SINAIS VOCAIS COM BASE NO DEEP LEARNING

Aprovado em / /

BANCA EXAMINADORA

_____

Prof. Dr. Leonardo Wanderley Lopes - Orientador
Universidade Federal da Paraíba – UFPB)

_____

Prof. Dr. Ronei Marcos de Moraes - Orientador
Universidade Federal da Paraíba – UFPB)

_____

Profa. Dra. Liliane dos Santos Machado – Examinadora interna
(Universidade Federal da Paraíba – UFPB)

_____

Profa. Dra. Anna Alice Figueiredo de Almeida Queiroz – Examinadora interna
(Universidade Federal da Paraíba – UFPB)

_____

Profa. Dra. Suzete Élida Nóbrega Correia – Examinadora externa à instituição
(Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB)

_____

Profa. Dra. Ana Cristina Côrtes Gama – Examinadora externa à instituição
(Universidade Federal de Minas Gerais – UFMG)

Dedico este trabalho primeiramente a Deus por me mostrar os caminhos da busca pela sabedoria, à minha mãe Edelzuith Arruda Leite, ao meu pai, Francisco de Assis Leite (*in memorian*), à minha esposa Gabriela Oliveira pelo apoio, incentivo e amor, e às minhas filhas Maria Luiza e Marina, pelo carinho e motivação.

AGRADECIMENTOS

À vida por sempre me surpreender com desafios inesperados, apresentando mais uma vez, o desafio da construção do conhecimento em uma nova estrada, com novos caminhos, mais longos, mais emocionantes, mais desafiadores.

Agradecer a Deus, pelas mãos de Maria, por essa graça alcançada.

Aos meus pais Edelzuith Arruda Leite quem me ensina a ser forte e enfrentar a vida e Francisco de Assis Leite (*in memorian*) minha referência de humildade e honestidade que juntos sempre estiveram ao meu lado nas minhas escolhas, estarão orgulhosos pela conquista.

À minha esposa Gabriela Oliveira, pelo carinho, pelo amor e ajuda incondicional, assim como me manter motivado e não me deixar desfalecer quando todo não parecia dar certo. Sem ela, não teria começado esse desafio. Ser minha fortaleza!

As minhas filhas Maria Luiza e Marina por serem o motivo da minha constante busca pelo conhecimento e motivação para ser um ser humano melhor. A quem retirei muita atenção, paciência e acompanhamento, agradeço a preocupação manifestada diariamente através de perguntas no cotidiano.

Aos meus familiares por darem suporte aos meus pais quando não pude estar presente e desta forma me incentivaram a continuar.

Aos meus colegas da GEP/HULW, bem como UNIESP pelo incentivo, apoio e companheirismo nas horas mais difíceis desta caminhada. Em particular a pessoa do Professor Fabio Sampaio.

Aos professores e colegas do Programa de Pós-Graduação em Modelos de Decisão e Saúde pela excelência do ensino e aprendizados.

Aos meus orientadores Professores Leonardo Wanderley Lopes e Ronei Marcos de Moraes, pelo exemplo de profissionais, acadêmicos, pesquisadores e pessoas. Especialmente, pela oportunidade de me receber como orientando.

Ao professor Sergio Ribeiro, por ter me aceitado na seleção de doutorado e iniciado essa caminhada.

Danilo Leite

“Não há saber mais ou saber menos: há saberes diferentes.”

(Paulo Freire)

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.”

(Arthur Schopenhauer)

“A inteligência artificial só será um problema se não reconhecermos o valor do ser humano.”

(Raidalva de Castro)

RESUMO

A voz é um dos principais meios de comunicação do ser humano, sua emissão deve ser agradável, sem esforços e conforme aos interesses profissionais, sociais e pessoais do interlocutor. Qualquer alteração na sua emissão, pode ser classificada como distúrbio de voz. Diagnosticar o distúrbio no seu estágio inicial, pode ser crucial para evitar situações de morbidade mais sérias, pois fornece ao paciente a oportunidade de um tratamento sem complicações, oferecendo uma qualidade de vida melhor. Na prática clínica tradicional, são necessários diversos exames médicos para detectar um distúrbio de voz, como a observação das pregas vocais por meio de laringoscopia, para visualizar possíveis alterações morfológicas, ou a análise acústica, úteis para evidenciar possíveis alterações funcionais. Esses exames são muitas vezes invasivos e demorados podendo causar desconforto ao paciente durante o procedimento. A análise acústica tem sido indicada como uma ferramenta auxiliar que utiliza procedimentos não invasivo, de baixo custo, utilizando técnicas de processamento digital de sinal de voz colaborando no diagnóstico de patologias da voz. Dentre as possibilidades de análise acústica, a espectrografia é um recurso de grande relevância, a partir dela podem ser visualizadas informações como presença de ruído em média e altas frequências, intensidade, instabilidade dos harmônicos, quebras de sonoridade, entre outras. Diante do exposto, esse estudo construiu um modelo inteligente utilizando uma *Deep Neural Network* (DNN) pré-treinada para classificar imagens espectrográficas da tipologia do sinal da voz da vogal sustentada "é" de acordo com a proposta de Titze (1975) e Sprecher et al. (2010). A classificação proposta por Titze (1995), mais utilizada em procedimentos de pesquisa, categoriza os sinais em Tipo I, II e III. Sprecher et al. (2010) propuseram a inclusão do sinal Tipo IV à classificação original feita por Titze (1995). Também foi utilizado o Grad-CAM para marcar no espectrograma as partes mais relevantes utilizadas pelo modelo na classificação. Nesse sentido, uma classificação automática utilizando a proposta de Titze(1975) e Sprecher et al. (2010) pode ser útil como medida de resultado de tratamento, uma vez que a classificação reflete a intensidade do desvio vocal e a presença de alteração laríngea. A construção desse modelo de classificação automática para classificar a tipologia do sinal, poderá auxiliar o clínico no processo de tomada de decisão no seguimento do tratamento. A arquitetura desenvolvida na metodologia resultou em uma Acurácia Global do Teste de 0.94, *Precision* de 0.94, F1Score 0.94, *kappa* 0.91, sensibilidade e especificidade 0.94 e 0.98, respectivamente. O modelo construído pode ser utilizado como ferramenta na etapa de pré-processamento antes de calcular qualquer medida de perturbação, bem como contribuir para potencializar a eficiência nas análises espectrográficas, auxiliando o clínico na sua tomada de decisão.

Palavras-chave: Espectrogramas, Tipologia da Voz, *Deep Learning*, Grad-CAM.

ABSTRACT

The voice is one of the main means of communication of the human being, its emission must be pleasant, effortless and in accordance with the professional, social and personal interests of the interlocutor. Any change in its emission can be classified as a voice disorder. Diagnosing the disorder at its early stage can be crucial to avoid more serious morbidity situations, as it provides the patient with the opportunity for an uncomplicated treatment, offering a better quality of life. In traditional clinical practice, several medical tests are necessary to detect a voice disorder, such as observation of the vocal folds by means of laryngoscopy, to visualize possible morphological alterations, or acoustic analysis, useful to evidence possible functional alterations. These exams are often invasive and time-consuming and may cause discomfort to the patient during the procedure. Acoustic analysis has been indicated as an auxiliary tool that uses non-invasive, low-cost procedures, using digital voice signal processing techniques, collaborating in the diagnosis of voice pathologies. Among the possibilities of acoustic analysis, spectrography is a resource of great relevance, from which information such as the presence of noise in medium and high frequencies, intensity, instability of harmonics, breaks in sound, among others, can be viewed. Given the above, this study built an intelligent model using a pre-trained Deep Neural Network (DNN) to classify spectrographic images of the voice signal typology of the sustained vowel "é" according to the proposal of Titze (1975) and Sprecher et al. al. (2010). The classification proposed by Titze (1995), most used in research procedures, categorizes signals into Type I, II and III. Sprecher et al. (2010) proposed the inclusion of the Type IV signal to the original classification made by Titze (1995). Grad-CAM was also used to mark in the spectrogram the most relevant parts used by the model in the classification. In this sense, an automatic classification using the proposal by Titze (1975) and Sprecher et al. (2010) may be useful as a treatment outcome measure, since the classification reflects the intensity of the vocal deviation and the presence of laryngeal alteration. The construction of this automatic classification model to classify the signal typology may help the clinician in the decision-making process following the treatment. The architecture developed in the methodology resulted in an Overall Test Accuracy of 0.94, Precision of 0.94, F1Score of 0.94, kappa of 0.91, sensitivity and specificity of 0.94 and 0.98, respectively. The built model can be used as a tool in the pre-processing stage before calculating any disturbance measure, as well as contributing to enhance the efficiency of spectrographic analyses, helping the clinician in his decision making.

Keywords: Spectrograms, voice typology, Deep Learning, Grad-CAM

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Adam	*Adaptative Momentum Estimation* (Estimação adaptativa do *momentum*)
ANN	*Artificial Neural Networks* (Redes Neurais Artificiais)
CNN	*Convolutional Neural Networks* (Redes Neurais Convolucionais)
DNN	*Deep Neural Network* (Redes Neurais Profundas)
FN	Falso Negativo
FP	Falso Positivo
F0	Frequência Fundamental
GPU	Graphics Processing Units
HNR	Proporção Harmônica para Ruído
LIEV	Laboratório Integrado de Estudos da Voz
MDVP	Multi-Dimensional Voice Program
MEEI	*Massachusetts Eye and Ear Infirmary*
ML	*Machine Learning*
ResNet	*Residual Networks*
STFT	*Short-Time Fourier Transform*
VGG	*Visual Geometry Group* (Grupo de Geometria Visual)
SVD	*Saarbrucken Voice Database*

SUMÁRIO

1 INTRODUÇÃO

A voz é um dos elementos fundamentais na vida do ser humano, é um dos principais meios de sua comunicação (LAVAN *et al.*, 2019), é um fenômeno que envolve grandes variações anatomofuncionais e comportamentais e sua emissão deve ser agradável, sem esforço e seguindo os interesses do interlocutor. Nesse sentido, Akbulut *et al.* (2020) caracteriza o distúrbio de voz como um processo patológico causado por fatores anatômicos ou funcionais que afeta a produção vocal e/ou o funcionamento da laringe.

A presença de distúrbio na voz pode causar mudanças significativas nos padrões vibratórios, afetando a qualidade da produção normal da voz, podendo influenciar negativamente na qualidade de vida de um indivíduo, limitando sua comunicação no seu trabalho, entre outros (LOPES *et al.*, 2017; VIEIRA, 2014). Pacientes com distúrbios na voz podem apresentar diferentes sintomas, a rouquidão, ardor na garganta, fadiga vocal e pigarro estão entre os mais relatados (MARTINEZ e RUFINER, 2000). Problemas de voz podem prejudicar a qualidade de vida, o que pode resultar em depressão ou outras dificuldades psicológicas. Portando, os distúrbios da voz podem ter um impacto significativo na saúde mental do paciente.

Distúrbios da voz são condições médicas que envolvem tom anormal, intensidade sonora ou qualidade do som produzido pela laringe e, portanto, afetam a produção da fala. Estão frequentemente associados a várias doenças neurológicas, como a doença de Parkinson, doenças cancerígenas, câncer de laringe, entre outras. Diagnosticar o distúrbio no seu estágio inicial, pode ser crucial para evitar situações de morbidade mais sérias (VAVREK *et al.*, 2021), pois fornece ao paciente a oportunidade de um tratamento sem complicações, oferecendo uma qualidade de vida melhor.

Nesse contexto, um processo de avaliação da voz eficiente ou bem-sucedido permite ao fonoaudiólogo diagnosticar distúrbio de voz, determinar a efetividade das várias abordagens de tratamento e, até mesmo, formular um prognóstico mais preciso (LEITE; MORAES; LOPES, 2020a; LOPES *et al.*, 2017), evitando, assim, a degeneração progressiva para uma doença mais grave. Desse modo, é fundamental identificar as alterações na voz em um estágio inicial e fornecer ao paciente a oportunidade de superar qualquer problema e melhorar sua qualidade de vida.

A detecção de distúrbios da voz é um processo complexo e demorado que requer equipamentos tecnológicos e profissionais treinados para realizar o diagnóstico. A precisão de uma avaliação completa da voz, permite que o clínico ter uma maior eficiência no diagnóstico

dos distúrbios vocais, planejar melhor suas abordagens terapêuticas e definir melhor o prognóstico do tratamento.

Nas últimas décadas, houve um aumento no número de pesquisas referentes ao estudo da produção da voz e suas alterações patológicas, com a intenção de melhorar os procedimentos de análise vocal. Como consequência desses estudos a avaliação da qualidade vocal tornou-se um processo multidimensional e multidisciplinar devendo contemplar as seguintes análises: uma anamnese detalhada, análise perceptiva auditiva, análise acústica, avaliação aerodinâmica e o exame visual laríngeo. Feito esse processo multidimensional, é possível diagnosticar o desvio vocal, determinar a efetividade das várias abordagens de tratamento e, até mesmo, formular um prognóstico mais preciso (LOPES *et al.*, 2018; PISHGAR *et al.*, 2018). Dos procedimentos que são fundamentais para a rotina clínica do fonoaudiólogo, são considerados o julgamento perceptivo-auditivo, a avaliação aerodinâmica, a análise acústica, o exame visual laríngeo e a autoavaliação do paciente em relação ao seu distúrbio vocal.

Na prática clínica tradicional, são necessários diversos exames médicos para detectar um distúrbio de voz, como a observação das pregas vocais por meio de laringoscopia, para visualizar possíveis alterações morfológicas, ou a análise acústica, úteis para evidenciar possíveis alterações funcionais. Esses exames são muitas vezes invasivos e demorados podendo causar desconforto ao paciente durante o procedimento (LEITE; MORAES; LOPES, 2020a; LOPES *et al.*, 2017). A laringoscopia, por exemplo, é a principal forma de exame realizada para o diagnóstico dos distúrbios da voz. No entanto, é imprescindível que a técnica seja administrada por um profissional especialista e experiente, que possua equipamento médico adequado, muitas vezes caros e invasivos, a fim de proporcionar uma visualização precisa e abrangente das pregas vocais (VERDE *et al.*, 2022).

A análise perceptiva-auditiva é a avaliação mais utilizada em ambientes clínicos, considerado o principal padrão de referência utilizado pelo fonoaudiólogo na realização de avaliações vocais. Contudo, por ser tratar de uma avaliação que depende da percepção do especialista, está sujeita a equívocos e variações, mesmo entre avaliadores treinados, pois é influenciada por fatores como experiência prévia do avaliador, como também seu estado emocional (ALHUSSEIN; MUHAMMAD, 2018).

A análise acústica, por sua vez, tem sido indicada como uma ferramenta auxiliar que utiliza procedimentos não invasivo, de baixo custo, utilizando técnicas de processamento digital de sinais de voz colaborando no diagnóstico de patologias da voz (QUEIROZ *et al.*, 2017). A análise acústica pode compreender a análise descritiva de padrões visuais, como a

espectrografia de faixa larga (240hz e 300hz) para visualização de formantes e a espectrografia de faixa estreita para visualização de harmônicos (45 e 60hz) (LOPES *et al.*, 2020a).

Dentre as possibilidades de análise acústica, a espectrografia é um recurso de grande relevância, a partir dela podem ser visualizadas informações como presença de ruído em média e altas frequências, intensidade, instabilidade dos harmônicos, quebras de sonoridade, entre outras e independe do grau de aperiodicidade do sinal avaliado (BINDER, 2016; LOPES *et al.*, 2016). O espectrograma é uma forma visual de representar a intensidade do sinal de um sinal ao longo do tempo em várias frequências presentes em uma forma de onda específica. Assim, em virtude das possibilidades de informações fornecidas através de espectrogramas, gerou interesse em sistemas automáticos que possam auxiliar os especialistas em voz na classificação automática dos tipos de sinais de voz utilizando imagens espectrográficas (MIRAMONT *et al.*, 2020).

Na literatura, há duas descrições disponíveis para classificação do sinal vocal, baseadas no traçado espectrográfico de faixa estreita: a de Yanagihara (1967) e Titze (1995). Para Yanagihara (1967), mais utilizada no contexto clínico, classifica os espectrogramas em Tipo I, II, III e IV, essa classificação ocorre de acordo com a presença de ruído nos traçados e da regularidade dos harmônicos. A classificação proposta por Titze (1995), mais utilizada em procedimentos de pesquisa, fundamenta-se no modelo de dinâmica não linear da produção vocal e categoriza os sinais em Tipo I, II e III. Sprecher *et al.* (2010), por sua vez, propuseram a inclusão do sinal Tipo IV à classificação original feita por Titze. O sinal Tipo IV corresponderia à ausência de estrutura periódica no traçado, cuja fonte de produção é o próprio ruído turbulento transglótico dissipado no trato vocal e não o movimento vibratório das pregas vocais (LOPES *et al.*, 2020a). Nesse sentido, uma classificação automática utilizando a proposta de Titze (1995) e Sprecher *et al.* (2010) pode ser útil como medida de resultado de tratamento, uma vez que a classificação reflete a intensidade do desvio vocal e a presença de alteração laríngea. A construção de um modelo de classificação automática para classificar a tipologia do sinal, pode auxiliar o clínico no processo de tomada de decisão no seguimento do tratamento.

Nesse cenário, os avanços tecnológicos estão proporcionando cada vez mais um melhor suporte para auxiliar os profissionais da voz para detectar precocemente e classificar os distúrbios de voz automaticamente, com mais eficiência e com baixo custo. Tecnologias como 5G, computação de borda, computação em nuvem, Internet das Coisas (IoT) e a Inteligência Artificial (IA), podem ou estão proporcionando ao paciente um maior cuidado com sua saúde

em qualquer hora, em qualquer lugar de forma não invasiva, de baixo custo e em tempo real (VERDE *et al.*, 2022).

Nesse contexto, as imagens espectrográficas do sinal vocal alinhada a técnicas de Inteligência Artificial (IA), podem proporcionar a construção de ferramentas para classificar o desvio vocal, bem como a tipologia do sinal da voz, de forma não invasiva e com menor custo. Técnicas de IA, especificamente Aprendizado de Máquina (AM), estão sendo cada vez mais utilizadas para constituir excelentes ferramentas para processamentos de dados e imagens, reduzindo, assim, o tempo de resposta e otimizando processos decisórios e previsões em diversas áreas, como a da saúde (RAČIĆ *et al.*, 2021; TRINH; DARRAGH, 2019).

Diversos estudos comparam espectrogramas de grupos de pessoas sem disfonia e com disfonia utilizando o espectrograma com técnicas de AM (ARIAS-VERGARA *et al.*, 2021; GUMELAR *et al.*, 2020; LEITE; MORAES; LOPES, 2021; SAKASHITA; AONO, 2018; TRINH; DARRAGH, 2018), analisando as diferenças acústicas entre os espectrogramas. No entanto, não foi encontrado na literatura, propostas de modelos de aprendizado profundo (*Deep Learning*) utilizando imagens espectrográficas para classificar automaticamente a tipologia do sinal da voz da vogal sustentada /e/ com base na proposta de Titze (1995) e Sprecher *et al.* (2010), bem como a utilização do GRAD-CAM para marcar na imagem do espectrograma as partes mais relevantes utilizadas pelo modelo na classificação e potencializar a eficiência nas análises espectrográficas, visto que, para fazer essa classificação, é necessária uma avaliação criteriosa com base no julgamento visual feito por fonoaudiólogos experientes, e por consequência contribuir na produção do conhecimento.

O *Deep Learning* (DL), é uma subárea do AM que está tornando a IA cada vez mais eficiente, e vem sendo cada vez mais utilizada no processamento de imagens médicas (AGGARWAL, 2018). Nas últimas décadas, o DL provou ser uma ferramenta muito poderosa devido à sua capacidade de lidar com grandes quantidades de dados.

Nesse sentido, DL vem superando as técnicas tradicionais, principalmente no reconhecimento de padrões e tem como principais características as várias camadas ocultas na sua arquitetura. No DL, a Rede Neural Convolucional (*Convolutional Neural Network* - CNN) (ARAVINDA *et al.*, 2021), faz parte da classe das Redes Neurais Profundas (*Deep Neural Networks* - DNN) mais comumente utilizada para analisar imagens. Nos últimos anos, o DL vem demonstrando bons resultados em tarefas de classificação e, desde então, o uso dessa tecnologia vem crescendo rapidamente, alcançando um desempenho surpreendente na classificação de sinais acústicos, objetos, palavras, no processamento digital de imagens,

principalmente nas imagens médicas entre outros (GUMELAR *et al.*, 2020c; MAHAJAN; CHAUDHARY, 2019a).

Dentro deste campo, neste trabalho, para classificar a tipologia do sinal da voz através de imagens espectrográficas, foi desenvolvido um modelo inteligente de classificação utilizando uma abordagem de DL com base na CNN VGG-16. Para analisar sua eficiência na classificação da tipologia do sinal da voz, comparamos seus resultados com mais dois modelos de DL, a saber: VGG-19 e ResNet50.

Portanto, nesse estudo, foi desenvolvido um modelo inteligente para classificação automática de imagens espectrografias da tipologia da voz, utilizando métodos de DL. Esse modelo tem como objetivo auxiliar os profissionais da voz no processo de triagem, dando mais rapidez e eficiência nos procedimentos de avaliação e monitoramento dos pacientes, podendo ser utilizado em ambientes *off-line* (gravação seguido de análise) e *on-line* (gravação e análise em tempo real) por meio de tecnologias modernas baseadas em nuvem, assim como auxiliar no treinamento de novos profissionais, sejam eles acadêmicos ou profissionais. Esse modelo poderá determinar a qualidade vocal do indivíduo indicando parâmetros acústicos que compõem o sinal da voz, bem como indicar quais medidas poderão ser utilizadas no processo de apoio a tomada de decisão, tato para o monitoramento quanto no tratamento do paciente.

1.1 JUSTIFICATIVA

Nos últimos anos, houve um aumento no interesse dos pesquisadores e profissionais da voz por métodos de análise acústica para avaliar a qualidade vocal, bem como na construção de ferramentas tecnológicas de classificação automática de distúrbios vocais (LOPES *et al.*, 2018; PISHGAR *et al.*, 2018). Nesse sentido, sistemas computacionais estão sendo cada vez mais aplicados para contribuir no diagnóstico mais assertivo e classificação dos distúrbios de voz com métodos não invasivos e com menor custo (MIRAMONT *et al.*, 2020; ROY; SAYIM; AKHAND, 2019; WU *et al.*, 2018).

O *Praat e Multi-Dimensional Voice Program* (MDVP) (LOVATO *et al.*, 2016; PAUL BOERSMA; DAVID WEENINK, 2021) são os principais sistemas computacionais utilizados na prática clínica e de pesquisa para avaliar parâmetros característicos, incluindo a frequência fundamental (F0), proporção harmônica para ruído (HNR) e espectrograma. No entanto, esses sistemas podem ser utilizados apenas por especialistas, pois exigem uma avaliação criteriosa dos valores extraídos, que pode ser realizada apenas pelo clínico.

Por essas razões, a construção de sistemas automáticos, não invasivo, de baixo custo e

inteligente, poderiam desempenhar um papel decisivo no diagnóstico precoce de distúrbio de voz e auxiliar no processo de tomada de decisão clínica (VERDE *et al.*, 2022), possibilitando iniciar o tratamento dando mais esperança de cura ao paciente. Nesse sentido, comparado aos métodos tradicionais de análise, a IA tem mostrado resultados superiores na resolução de problemas computacionais para detectar e classificar problemas na voz (RAČIĆ *et al.*, 2021b; TRINH; DARRAGH, 2019).

Estudos utilizaram o espectrograma com técnicas de AM para classificar grupos de pessoas sem disfonia e com disfonia (ARIAS-VERGARA *et al.*, 2021; GUMELAR *et al.*, 2020; LEITE; MORAES; LOPES, 2021; SAKASHITA; AONO, 2018; TRINH; DARRAGH, 2018), analisando as diferenças acústicas entre os espectrogramas, e obtiveram bons resultados na classificação de distúrbios da voz. No entanto, não foi encontrado na literatura pesquisada, estudos utilizando CNNs para classificar imagens espectrográficas da tipologia do sinal da voz para auxiliar o clínico no processo de tomada de decisão.

O uso de modelos de CNNs alinhado a técnica de Grad-CAM (Selvaraju *et al.*, 2016) na classificação da tipologia da voz podem ser uma ferramenta útil no auxílio de diagnóstico médico como técnica complementar, podendo contribuir, significativamente, na redução da quantidade de exames invasivos a serem realizados pelo paciente. Além disso, a aplicação pode beneficiar positivamente o acompanhamento de tratamentos e terapias vocais.

A CNN executando a classificação e o Grad-CAM marcando as partes mais importantes no espectrograma utilizados para classificar a imagem, e assim, auxiliando o profissional a tomar sua decisão no tratamento do paciente.

Considerando que o tipo de sinal pode ser um critério relevante para conduzir o tipo de análise acústica a ser realizada e que, além disso, há possibilidade de utilização da classificação de Titze (1995) e Sprecher *et al.* (2010) para categorização dos sinais de pacientes avaliados clinicamente. Este estudo propõe a construção de um modelo de inteligente utilizando DL para classificar imagens espectrográficas da tipologia do sinal da voz da vogal sustentada “é” baseado na proposta de Titze (1995) e Sprecher *et al.* (2010).

Por fim, o modelo proposto poderá auxiliar o profissional da voz nos procedimentos de avaliação e monitoramento dos distúrbios de voz assim como contribuir na produção do conhecimento e treinamento de novos profissionais, sejam eles acadêmicos ou profissionais (BEHLAU; MURRY, 2012; HOSSAIN; MUHAMMAD, 2016; LOPES *et al.*, 2018; MIRAMONT *et al.*, 2020).

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um modelo inteligente de classificação da tipologia dos sinais vocais utilizando o *Deep Learning* (DL) a partir de imagens espectrográficas.

1.2.2 Objetivos Específicos

- Desenvolver um modelo utilizando a CNN pré-treinada VGG-16 como base, para classificar espectrogramas da tipologia do sinal da voz da vogal sustentada “é”;
- Analisar e avaliar os resultados obtidos do modelo proposto na classificação da tipologia do sinal de voz, a partir das imagens espectrográficas;
- Comparar os resultados do modelo desenvolvido com as CNNs VGG-19 e ResNet50;
- Verificar a acurácia, concordância, sensibilidade, especificidade, precision, F1-score, valor preditivo positivo e negativo do modelo baseado na imagem espectrografia na classificação da tipologia do sinal da voz.
- Promover a viabilidade da utilização do modelo no formato de um sistema computacional para ser utilizado nas clínicas e universidades.

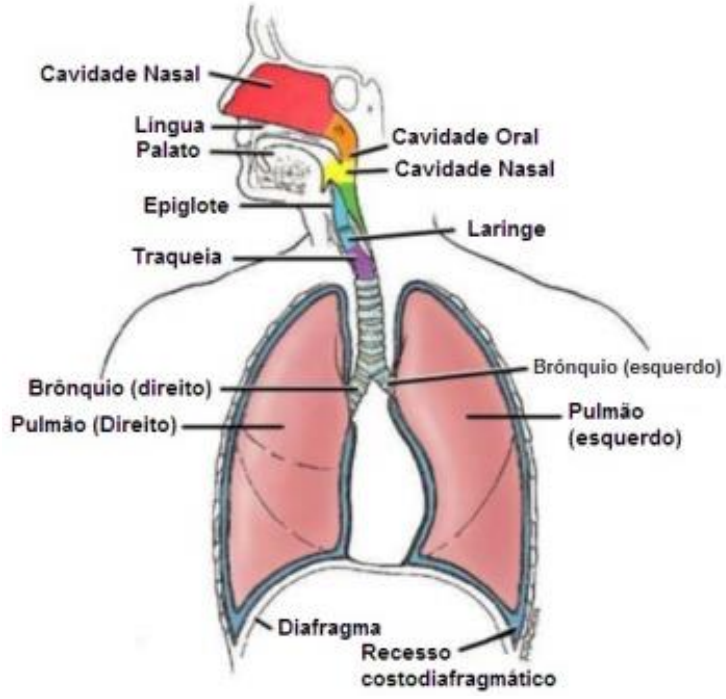
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo apresenta-se a fundamentação teórica que norteou a construção deste estudo. Para isto, discutem-se os seguintes temas: processo de produção da voz, distúrbios de voz, tipologia da voz e aprendizado de máquina.

2.1 PROCESSO DE PRODUÇÃO DA VOZ

A produção da voz é um processo complexo que inclui diversos componentes estruturais e funcionais. A anatomia do sistema de produção da fala humana é mostrada na Figura 1. O aparato vocal compreende três cavidades: nasal, oral e faríngea. As cavidades faríngeas e orais são geralmente agrupadas em uma unidade chamada de trato vocal, e a cavidade nasal costuma ser chamada de trato nasal (SANTOS, 2015).

Figura 1 - O sistema de produção da fala humana



Fonte: Santos (2015)

O trato vocal estende-se da abertura das pregas vocais, ou glote, através da faringe e boca até os lábios. O trato nasal se estende desde o véu até as narinas. O processo da fala começa quando o ar é expelido dos pulmões pela força muscular que fornece a fonte de energia (sinal de excitação). Em seguida, o fluxo de ar é modulado de várias maneiras para produzir sons de

fala diferentes. Os pulmões, brônquios e traqueia produzem o “ar”, matéria-prima da produção vocal; a laringe (onde se encontram as pregas vocais) produz a energia da fala e, a faringe, fossas nasais e boca são responsáveis pela ressonância (SANTOS, 2015).

A modulação é realizada principalmente no trato vocal (principal estrutura ressonante), por meio de movimentos de diversos articuladores, como véu, dentes, lábios e língua. Os movimentos dos articuladores modificam a forma do trato vocal, que cria diferentes frequências ressonantes e, consequentemente, diferentes sons da fala (SANTOS, 2015; VIEIRA, 2014).

2.2 DISTÚRBIOS DA VOZ

Os distúrbios da voz são processos patológicos que afetam diretamente a produção vocal, apresentando-se de diferentes formas, incluindo a presença de sintomas sensoriais e auditivos, desvios da qualidade vocal e a presença de alterações funcionais e/ou estruturais da laringe, que pode envolver fatores comportamentais e/ou orgânicos associados à sua gênese e manutenção, podendo afetar diferentes faixas etárias (RAMIG, 1998; BIREME, 2017). Esses distúrbios podem ter impacto negativo na qualidade de vida dos pacientes, comprometendo aspectos sociais, emocionais e laborais (LOPES *et al.*, 2016).

No que se refere às possíveis causas, podem ser listados: uso inapropriado da voz, estresse vocal, uso de drogas e as patologias na laringe. Pacientes com distúrbios da voz podem apresentar diferentes sintomas, sendo a rouquidão, ardor na garganta, fadiga vocal e pigarro os mais referidos e associados ao uso intenso da voz, a infecções de vias aéreas superiores, ao estresse e ao tabagismo (FERREIRA *et al.*, 2009). A manifestação de um distúrbio de voz é multidimensional, assim, sua avaliação precisa abranger diferentes aspectos, incluindo a avaliação perceptiva da voz, o exame visual laríngeo, a análise acústica, a avaliação aerodinâmica e a autoavaliação vocal (DEJONCKERE *et al.*, 2001).

O uso inapropriado da voz refere-se a comportamentos de produção vocal que não atingem um desempenho vocal eficiente. Assim, a voz destoa do que seria considerada uma emissão acústica considerada normal. O que pode influenciar este comportamento é a falta de conhecimento vocal por parte do indivíduo. Alguns dos fatores que estão relacionados ao uso inadequado da voz, são (RAYMOND, 2006):

- Aumento de tensão ou esforço: para chegar certo resultado vocal (por exemplo, aumentar a intensidade da voz), é possível que haja tensão nos músculos ligados ao sistema de produção vocal. Em alguns casos, este comportamento pode ser refletido em dores ao falar. Contudo, um paciente pode ser considerado disfônico quando a sua produção vocal

requer mais esforço que o suficiente;

- Uso inadequado de frequência: a frequência fundamental é um parâmetro que difere para vozes masculinas, femininas e infantis. Uma alteração vocal pode ser percebida quando a frequência fundamental do indivíduo difere da média do gênero/idade;
- Uso inadequado da voz: as pregas vocais podem ser tencionadas pelo uso de muita tensão ao falar. Isso pode causar problemas nos músculos da garganta e afetar a voz. O abuso vocal também pode causar um distúrbio de voz. Exemplos de abuso vocal incluem falar demais, gritar ou tossir. Fumar e limpar constantemente a garganta também são um abuso vocal.
- Distúrbios psicogênicos: os aspectos psicológicos que podem contribuir para o surgimento de afonia ou disfonia geralmente surgem na infância. Distúrbios emocionais causados por traumas físicos ou por conflitos familiares, e até mesmo a relação com outras crianças podem ser fatores de ocorrência de distúrbios da voz que podem acompanhar o indivíduo ao longo dos anos.

Exemplos de distúrbios da voz, segundo Rosen, Sataloff e Sataloff, (2020), incluem:

- Laringite: ocorre quando pregas vocais incham, fazendo a voz soar rouca, ou mesmo fazer com que o locutor não consiga falar nada. Ou você pode não conseguir falar nada. A laringite aguda acontece repentinamente, geralmente por causa de um vírus no trato respiratório superior. Geralmente dura apenas algumas semanas. O tratamento consiste em descansar a voz e beber bastante líquido. Laringite crônica é quando o inchaço dura por muito tempo. As causas comuns incluem tosse crônica, uso de inaladores para asma e doença do refluxo gastroesofágico (DRGE). O tratamento da laringite crônica depende da causa.
- Paresia ou paralisia das pregas vocais: as pregas vocais podem estar paralisadas ou parcialmente paralisadas (paresia). Isso pode ser causado por uma infecção viral que afeta os nervos das pregas vocais, uma lesão em um nervo durante uma cirurgia, derrame ou câncer. Se uma ou ambas as cordas vocais estiverem paralisadas em uma posição quase fechada, você pode ter respiração ruidosa ou difícil. Se eles estiverem paralisados em uma posição aberta, você pode ter uma voz fraca e ofegante. Algumas pessoas vão melhorar com o tempo. Em outros casos, a paralisia é permanente. A cirurgia e a terapia de voz podem ajudar a melhorar a voz.
- Disfonia espasmódica: este é uma alteração de base neurológica que causa espasmos nas

cordas vocais. Pode fazer a voz soar tensa, rouca, trêmula ou espasmódica. Às vezes, a voz pode parecer normal. Outras vezes, a pessoa pode não conseguir falar. O tratamento pode incluir terapia da fala e injeções de toxina botulínica nas pregas vocais.

A disfonia e todos os sintomas desencadeados por ela acarretam uma série de consequências, tais como, alterações psicoemocionais, depressão e frustração, que interferem direta e indiretamente no desempenho social, profissional e nas atividades de vida diária do indivíduo (LOPES *et al.*, 2016). Diante disto verifica-se a importância de realizar classificação da tipologia da voz utilizando análise espectrografia da voz como ferramenta auxiliar.

2.3 ANÁLISE ESPECTROGRÁFICA DA VOZ

Na acústica qualitativa, a espectrografia é o principal método de análise utilizado em pesquisas e no âmbito clínico (SILVA, 2022). Consiste em um gráfico tridimensional originado pelo sinal sonoro que indica o tempo apresentado no eixo horizontal, frequência exibidos no eixo vertical e a intensidade que é dada pelo contraste das cores do traçado (LOPES *et al.*, 2020).

O espectrograma é uma ferramenta que pode proporcionar a análise de diversos atributos e características do sinal da voz, fundamentado pela transformada de Fourier de curta duração (*Short-Time Fourier Transform* – STFT), que consiste em um cálculo matemático que converte a forma da onda (amplitude pelo tempo) no espectro (frequência pela amplitude), formando um espectro, conforme Equação 1.

Um espectrograma é uma ferramenta que pode proporcionar a análise de diversos atributos e características do sinal da voz, pode ser formado, também, a partir do cálculo da transformada de Fourier de curta duração (*Short-Time Fourier Transform* – STFT) aplicado ao sinal digital do som, conforme Equação 1.

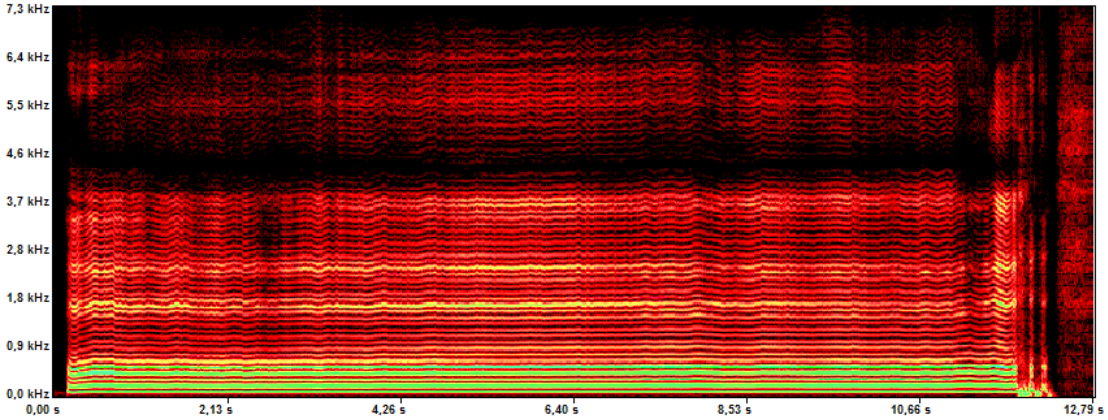
$$X[n,\lambda] = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \qquad (1)$$

onde x[n+m] representa o sinal analisado, como uma função de uma variável de tempo discreta *n*, deslocando em *m* posições, λ representa o espectro (contínuo) de frequências e *X* representa a energia do sinal, em uma função de *n* e de λ.

O eixo na STFT mostra tempo e frequência, e a escala de cores da imagem do espectrograma representa a amplitude da frequência (Figura 2). A base para a representação da

STFT é conhecida como uma soma de uma série de sinusoides (senos e cossenos) (ALASKAR, 2018).

Figura 2 - Espectrograma extraído de um paciente com presença de distúrbio de voz



Fonte: pesquisador do estudo (2022)

O espectrograma pode ser definido como: um gráfico que mostra a intensidade por meio do escurecimento ou coloração do traçado, as faixas de frequência no eixo vertical e o tempo no eixo horizontal. Sua representação mostra estrias horizontais, denominadas harmônicas. O espectrograma demonstra visualmente as características acústicas da emissão, contudo essas informações exigem interpretação por parte do avaliador (VALENTIM; CÔRTES; GAMA, 2010).

A espectrografia acústica de faixa estreita é a mais utilizada na rotina clínica, pois permite colher informações sobre a fisiologia da produção vocal ao glótico. Tem como principais vantagens, avaliação vocal independente do grau de desvio, utilização simultâneo com o JPA para aumento da confiabilidade e monitoramento terapêutico de curto e longo prazo (LOPES *et al.*, 2020a). Além disso, os dados que podem ser obtidos através do espectrograma são a regularidade do traçado, detecção de regiões incremento de energia e visualização da presença de ruído (LOPES *et al.*, 2020b).

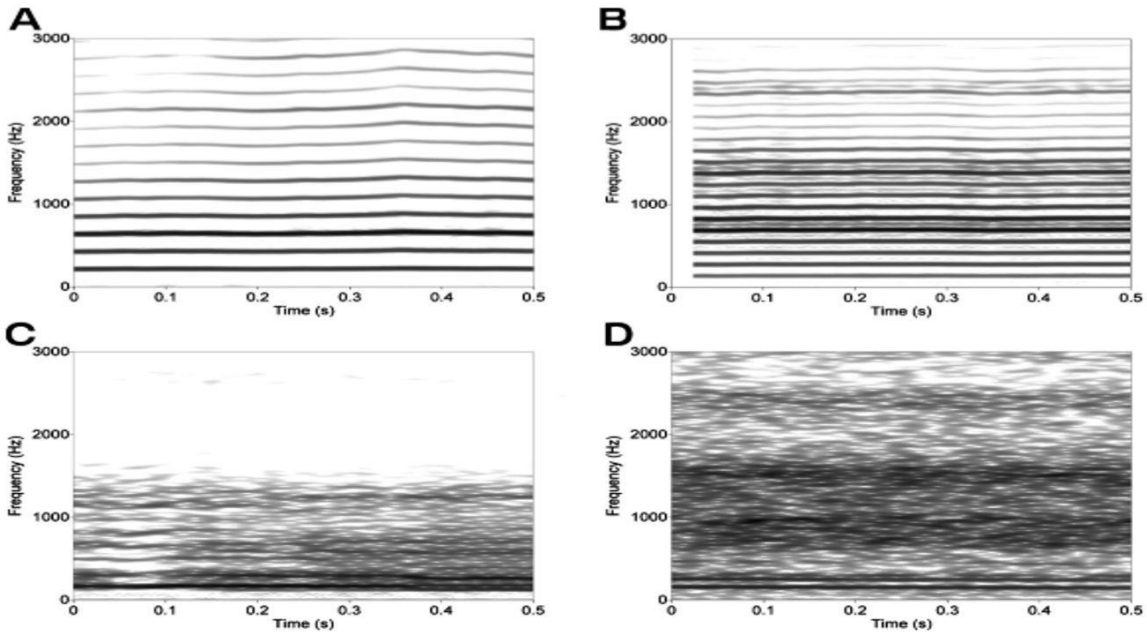
A espectrografia de banda estreita de vogal sustentada é utilizada em procedimentos de avaliação de distúrbios vocais e vocais profissionais. Vários estudos relacionam as características desse tipo de traçado espectrográfico a desvios na qualidade vocal e alterações no funcionamento laríngeo (LOPES *et al.*, 2018; LOPES; ALVES; MELO, 2017; VALENTIM; CÔRTES; GAMA, 2010).

Espectrogramas são mais utilizados para classificar sinais de voz nos diferentes tipos,

como proposto em Yanagihara (1967) e Titze (1995), independentemente do esquema de três ou quatro tipos empregado. A classificação do sinal de voz em três tipos, de acordo com seu grau de periodicidade, é uma tarefa conhecida como digitação de sinal, ou seja, é uma etapa relevante de pré-processamento antes de calcular qualquer medida de perturbação (MIRAMONT *et al.*, 2020).

A classificação realizada por Yanagihara (1967) a mais utilizada no contexto clínico, para caracterização do desvio observado no sinal e sua relação com a intensidade do desvio vocal, no plano perceptivo-auditivo, classifica os sinais em tipo I, II, III **e** IV (Figura 3), utilizando como critérios a regularidade dos harmônicos, a presença de ruído em diferentes faixas de frequência e a relação entre a estrutura harmônica e o ruído presente no traçado.

Figura 3 - Espectrogramas extraídos com base em amostras de voz. Figuras (A), (B), (C) e (D) classificadas, respectivamente, como sinais do 1, 2, 3 e 4. Espectrogramas (A) e (D) são de vozes femininas, enquanto os (B) e (C) de vozes masculinas



Fonte: Sprecher et al. (2010)

A classificação proposta por Titze (1995), por sua vez, é mais utilizada em procedimentos de pesquisa, para determinação do tipo de análise a ser realizada (linear vs. não linear, análise de padrão visual vs. extração de medidas). É baseada no modelo de dinâmica não linear da produção vocal e caracteriza a mudança quantitativa no comportamento do sinal, advinda de mudança no padrão vibratório das pregas vocais.

Contudo, Sprecher *et al.* (2010), propuseram uma modificação no modelo de

caracterização de Titze (1995) quanto a divisão do sinal tipo III em mais um tipo de sinal, o tipo IV. O tipo IV são caracterizados por predomínio de ruído em relação à estrutura harmônica, em toda a faixa de frequência (SPRECHER *et al.*, 2010).

A classificação dos tipos de sinais, de acordo Titze (1995) e Sprecher *et al.* (2010), são:

- **Sinal Tipo I:** quase periódico, sem modulações ou sub-harmônicos, série rica de harmônicos definidos até 4 kHz, regularidade no traçado, ausência ou presença de ruído de baixa amplitude, em relação à estrutura harmônica;
- **Sinal Tipo II:** oscilam entre quase periódicos e aperiódicos, com presença de modulações e bifurcações, presença de sub-harmônicos claramente definidos, intermitentes (com duração ≤ 1s) e com intensidade próxima à intensidade de f0;
- **Sinal Tipo III:** aperiódicos, caóticos e com dimensão finita, presença de energia de ruído entre os harmônicos em baixas frequências, maior definição nos primeiros harmônicos de baixa frequência, em detrimento da substituição dos harmônicos de alta frequência por ruído difuso, concentração de energia abaixo de 1,5 kHz;
- **Sinal Tipo IV:** aperiódicos e caóticos, com dimensão infinita e predomínio de ruído em relação à estrutura harmônica, em toda a faixa de frequência.

Nenhum desses autores teve como objetivo desenvolver um roteiro de análise espectrográfica para diferentes condições vocais. Além disso, não fornecem descrições do comportamento dos harmônicos e dos componentes do ruído ao longo do tempo (que é uma das dimensões do espectrograma), aspecto importante para a caracterização de um sinal vocal desviado. Além disso, o modelo de classificação de Yahagihara (1967), não aborda a caracterização espectrográfica de sinais produzidos por indivíduos sem desvio vocal.

Assim, considerando que o tipo de sinal pode ser um critério relevante para conduzir o tipo de análise acústica a ser realizada, considerando a importância da análise espectrográfica para o contexto clínico, este estudo propõe a construção de um modelo de decisão automático, utilizando a DNN, para classificar imagens espectrográficas de acordo com a tipologia e os critérios propostos por Titze (1995) e Sprecher et al. (2010).

2.4 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (*Machine Learning* - ML) é uma subárea da Inteligência Artificial (IA), busca aprender automaticamente relacionamentos e padrões significativos a partir de exemplos e observações. É um recurso utilizado para aumentar o nosso conhecimento sobre determinada doença e aperfeiçoar nossas habilidades, auxiliar a construção do

diagnóstico clínico, direcionar tratamentos, monitorar pacientes, possibilitar a humanização no relacionamento com o paciente, possibilitar a diminuição dos erros médicos e auxiliar no processo de tomada de decisão médica.

Segundo Faceli (2015), AM é o processo de indução de uma hipótese, através de uma função, método ou hipótese, utilizando-se de dados históricos ou experiência passada. Para Wernick *et al.* (2010), pode-se utilizar a AM em diversas áreas para resolver problemas, tais como:

- Detectar e classificar sinais de vozes saudáveis e patológicas;
- Direcionar e/ou personalizar tratamentos;
- Possibilitar a redução nos atrasos no diagnóstico;
- Reconhecimento de fala e rostos;
- Classificação de imagens.

Aprendizado de Máquina pode resolver problemas que não podem ser gerenciados manualmente, em virtude da enorme quantidade de dados que devem ser processados ou a complexidade do problema. Quando aplicado a grandes conjuntos de dados, o AM pode encontrar relacionamentos tão sutis que nenhuma quantidade de análise manual jamais os descobriria. E quando muitas dessas relações "fracas" são combinadas, elas se tornam fortes preditores (BRINK; RICHARDS; FETHEROLF, 2016; RANSCHAERT; MOROZOV; ALGRA, 2019).

A literatura científica apresenta várias definições para AM com base no contexto em que é usado e nas várias entidades envolvidas no processo de aprendizado. Russell (2013), por exemplo, diz que a Aprendizado de Máquina, tem várias dimensões, uma delas é ideia que os agentes inteligentes, devem usar as percepções não apenas para agir, mas também para melhorar as habilidades do próprio modelo inteligente, para agir em um futuro.

Existem várias teorias e definições sobre o AM, uma das mais simples é a de Mitchell (1997) que diz “a capacidade de melhorar o desempenho na realização de alguma tarefa por meio de experiência”. Arthur Samuel foi o pioneiro no ramo de aprendizado de máquina (BRINK; RICHARDS; FETHEROLF, 2016). Ele escreveu em 1959 que o aprendizado de máquina é o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados (SHI; IYENGAR, 2020; XING; YANG, 2016). Logo, “AM visa desenvolver um conjunto de algoritmos para detectar padrões nos dados existentes e, em seguida, usa esses padrões descobertos para fazer previsões sobre novos dados e, assim, se adaptarem às mudanças em seus arredores” (SHI; IYENGAR, 2020).

Tom Mitchell (1997) explica que, “AM é um programa de computador que aprende com a experiência E em relação a alguma tarefa T e alguma medida de desempenho P, se seu desempenho em T, medido por P”.

Em uma definição mais moderna, o AM é definido como um conjunto de métodos que detectam automaticamente os padrões nos dados disponíveis e, em seguida, utilizam os padrões descobertos para prever dados futuros ou permitir a tomada de decisões em condições incertas (MALIK *et al.*, 2019; SHI; IYENGAR, 2020; XING; YANG, 2016).

Assim, o AM pode ser utilizado para a resolução de problemas específicos nas diversas áreas de conhecimento, como no auxílio aos especialistas da área médica no diagnóstico de enfermidade e em processos de tomada de decisão, tem a precisão de classificar automaticamente qualquer tipo de informação, descobrir padrões e fornecer o conteúdo de que precisa, de acordo com os dados disponíveis (ALBON, 2018; WERNICK *et al.*, 2010). Portanto, espera-se que uma máquina aprenda e preveja os resultados futuros com base nas mudanças observadas na estrutura externa, nos dados/entradas alimentados que precisariam ser respondidos e no programa/função que foi criado para executar.

Noutra perspectiva, encontrar padrões e classificar sinais de voz, vêm sendo um tema de grande interesse pelos pesquisadores. Estudos de reconhecimento de padrões em sinais de voz, está baseado em um ou conjunto de algoritmos capazes de encontrar características semelhantes em sinais para serem utilizadas na classificação em diferentes categorias (BRINK *et al.*, 2017).

Um dos maiores desafios para construir modelos de classificação que detectam automaticamente os padrões nos dados de forma eficiente, é encontrar características de entrada do algoritmo e usá-las para classificar os dados. Nas últimas décadas, as técnicas de AM para o reconhecimento de padrões em imagens, têm sido utilizados em sistemas inteligentes de saúde para detectar problemas na voz utilizando imagens de espectrograma (SAKASHITA; AONO, 2018), especialmente no diagnóstico e prognóstico da disfonia (LIU; POLCE; JIANG, 2019; TRINH; DARRAGH, 2019), onde, tradicionalmente, a precisão do diagnóstico de um paciente depende da experiência do especialista em voz. No entanto, esse conhecimento é construído ao longo de muitos anos de observações dos sintomas de diferentes pacientes e diagnósticos confirmados (AKBULUT *et al.*, 2020).

Mesmo assim, a precisão ainda não pode ser garantida. Com os avanços das tecnologias de computação, agora é relativamente fácil adquirir e armazenar muitos dados e imagens, e, posteriormente, utilizá-los para diagnosticar patologias (RANSCHAERT; MOROZOV;

ALGRA, 2019). Akbulut *et al.* (2020) dizem que técnicas de AM podem ser uma boa alternativa para auxiliar os profissionais nos trabalhos manuais complexos na detecção e classificação automática de distúrbios de voz.

2.4.1 Componentes do aprendizado de máquina

Para construir bons modelos de AM e obter bons resultados em tarefas de classificação, alguns componentes são necessários para sua construção, são eles:

- Conjunto de dados – os dados que serão utilizados para o aprendizado ou avaliação do modelo. Por exemplo, o conjunto de vozes do Laboratório Integrado de Estudo da Voz (LIEV) utilizando nesse trabalho;
- Dados de treinamento – uma fatia do conjunto de dados que serão apresentados ao algoritmo para criar o modelo;
- Dados de validação - uma fatia do conjunto de dados utilizados para uma avaliação imparcial do modelo;
- Dados de teste - uma fatia do conjunto de dados utilizados para uma avaliação final do modelo, simulando previsões reais que o modelo realizará, permitindo verificar o desempenho do modelo;
- Hiperparâmetros – parâmetros atribuídos ao algoritmo antes de iniciar o processo de treinamento. Por exemplo, número de camadas e/ou neurônios de uma Rede Neural Artificial (RNA), taxa de aprendizagem entre outros;
- Características – conjunto de informações representado por um vetor de características de partes ou padrões de um objeto em uma imagem que ajudam a identificá-la. São partes da imagem com algum significado para o algoritmo. Por exemplo, cor, intensidade, área em *pixel* etc.
- Validação cruzada - método de remostragem usado para avaliação de modelo para evitar o teste de um modelo no mesmo conjunto de dados no qual ele foi treinado. Divide o conjunto de dados em vários subconjuntos. É uma técnica eficiente para treinar modelos com um conjunto de dados pequeno.

2.4.2 Técnicas de aprendizado de máquina

Muitas técnicas de AM foram aplicadas por pesquisadores na solução desse problema de classificação. As metodologias de Aprendizado de Máquina são classificadas de acordo com os tipos de tarefas de aprendizado, conforme mostra a Figura 4.

Figura 4 - Hierarquia do aprendizado

Fonte: Faceli (2015)

Na hierarquia do aprendizado, o que vem no topo é o aprendizado indutivo, processo pelo qual são utilizadas as generalizações a partir dos dados. No segundo nível há dois tipos de aprendizado, o supervisionado (preditivo) e o não supervisionado (descritivo). Ainda existem mais três tarefas de aprendizados que não se enquadram na estrutura apresentada, são elas: aprendizado semi-supervisionado (ZHU; GOLDBERG, 2009), aprendizado ativo (SETTLES, 2012) e o aprendizado por reforço (SUTTON; BARTO, 1998).

No campo da voz, a maioria das inovações ou abordagens utilizando AM, concentram-se principalmente na técnica de aprendizado supervisionado (HEGDE *et al.*, 2019), devido à disponibilidade, mesmo que limitada, de bases de dados já rotuladas, principalmente, para classificar vozes saudáveis e patológicas (AL-NASHERI *et al.*, 2017, 2018).

2.4.3 Aprendizagem supervisionada

Aprendizado supervisionado é um algoritmo que aprende por meio de dados de treinamento previamente rotulados por especialistas, para criar modelos preditivos mais eficientes para prever resultados com dados de entrada não conhecidos. Portanto, no aprendizado supervisionado, a máquina é treinada usando dados rotulados armazenados em uma base de dados (ALBON, 2018; XING; YANG, 2016).

Nessa técnica, os dados são divididos em um conjunto de dados de treinamento e um conjunto de dados de teste. O conjunto de dados de treinamento é usado para treinar a máquina, enquanto o conjunto de dados de teste atua como novos dados para prever resultados ou para ver a precisão do modelo (ALBON, 2018).

O aprendizado supervisionado tende a ter um desempenho melhor quando com comparado ao não supervisionado, visto que, os dados do conjunto de dados que alimenta o modelo, já estão rotulados com os resultados do problema a ser resolvido, facilitando o aprendizado dessa técnica. Quanto mais dados forem disponibilizados para o processo de treinamento do modelo, mais ele aprende.

Depois do modelo treinado, ele pode ser utilizado para prever resultados futuros, possibilitando uma avaliação da qualidade vocal mais precisa. Nesse sentido, os algoritmos de aprendizado supervisionado estão se tornando uma ferramenta valiosa no processo de tomada de decisão para os profissionais da voz na detecção e classificação de vozes com ou sem distúrbios vocais.

O aprendizado supervisionado pode ser utilizado para criar modelos de aprendizado de máquina para resolver dois tipos de problemas, são eles:

- Regressão - um problema de regressão é utilizado para prever uma saída a partir de uma faixa contínua de valores possíveis;
- Classificação - um problema de classificação é quando a variável de saída é uma categoria, por exemplo a intensidade do desvio vocal entre leve e moderada (leve, moderada e intenso).

O aprendizado supervisionado é utilizado para induzir modelos preditivos por meio da observação de um conjunto de objetos rotulados. Diante disto, pode-se citar algumas vantagens dessa técnica: permite coletar dados e produzir saída de dados das experiências anteriores; ajuda a otimizar os critérios de desempenho, através da experiência; resolve vários tipos de problemas de computação do mundo real; auxiliar no processo de tomada de decisão médica.

2.4.4 Aprendizado de máquina não supervisionado

O aprendizado de máquina não supervisionado, não existe resultado pré-definido ou dados rotulados para que o modelo os utilize como referência no aprendizado. O modelo busca características semelhantes no conjunto de dados, separa os dados e agrupa-os em subconjuntos de dados. O modelo busca similaridades entre os dados sem a necessidade de intervenção humana (JO, 2021).

Enquanto os algoritmos de aprendizado de máquina supervisionado utilizam dados rotulados ou resultados conhecidos, o não supervisionados não precisa conhecer previamente seus resultados ou ter seus dados rotulados. Conhecer essa distinção ajuda a entender por que

os métodos de aprendizado de máquina não supervisionados não podem ser aplicados a problemas de regressão ou classificação porque não tem o valor/resposta para os dados de saída.

Modelos de aprendizado não supervisionados são utilizados para as três principais tarefas: agrupamento (*Clustering*), associação e redução de dimensionalidade.

- Agrupamento – é o processo de agrupar instâncias em *clusters*. O agrupamento é definido como o processo de segmentação de um grupo de instâncias em subgrupos com características semelhantes;
- Associação - é um método que usa diferentes regras para encontrar relacionamentos entre instâncias em um determinado conjunto de dados;
- Redução de dimensionalidade - é uma técnica usada quando o número de recursos, ou dimensões, ou atributos em um determinado conjunto de dados é muito alto. Ele reduz o número de entradas de dados para um tamanho gerenciável, ao mesmo tempo em que preserva ao máximo a integridade do conjunto de dados.

Comparando o não supervisionado com algoritmos de aprendizado supervisionado, os algoritmos não supervisionados são mais aptos a realizar tarefas complexas, principalmente quando não dispõem de dados já classificados ou rotulados. No entanto, os modelos de aprendizado supervisionado produzem resultados mais precisos, pois um especialista diz explicitamente ao sistema o que procurar nos dados fornecidos. Já no aprendizado não supervisionado, os resultados podem ser bastante imprevisíveis (JO, 2021).

2.4.5 Aprendizagem por transferência

Treinar uma CNN do zero pode ser um processo demorado e custoso quando não dispor de recursos computacionais suficientes. Para tentar sanar esse problema, foi desenvolvida uma variação da aprendizagem supervisionada chamada aprendizado por transferência. A utilização dessa técnica é especialmente vantajosa, quando se precisa resolver problemas de classificação de imagens, mas não se dispõe de recursos suficientes (dados rotulados e recursos computacionais) para treinar um novo modelo em escala real a partir do zero (DING *et al.*, 2021).

O aprendizado por transferência é uma técnica que reutiliza o conhecimento de uma CNN treinada em um grande conjunto de dados de imagens para a resolução de um problema geral, denominado domínio de origem, como ponto de partida para a resolução de outro problema, domínio de destino. Com essa técnica, é possível potencializar o aprendizado de um

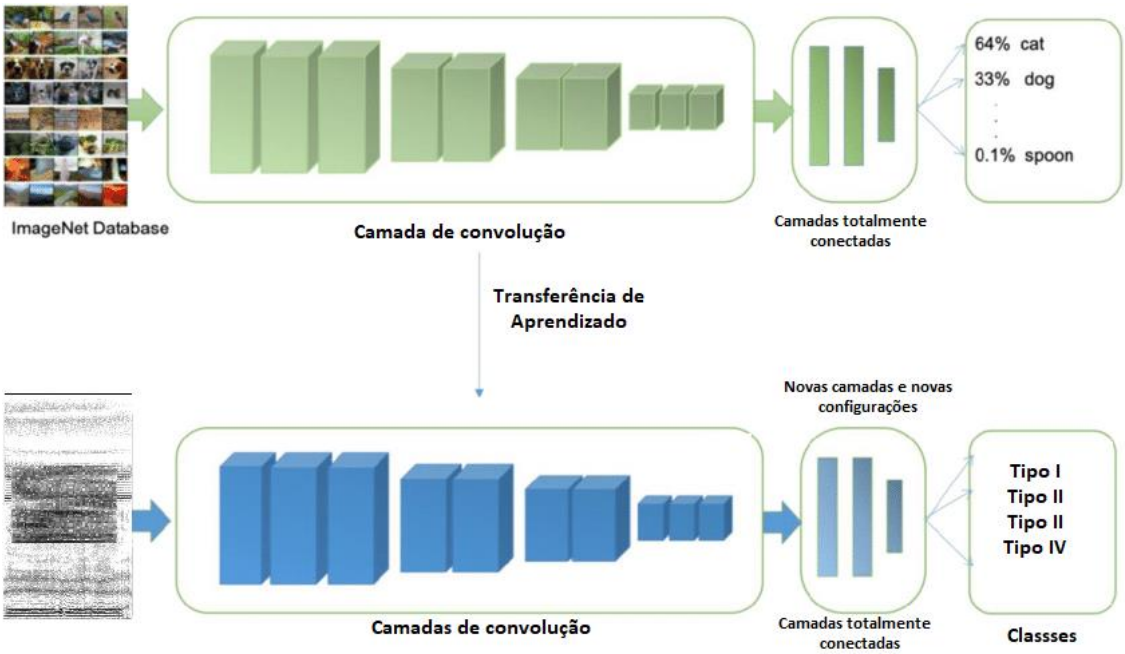
novo modelo para resolver um novo problema com menos iterações do que aquelas que seriam necessárias para treinar uma CNN do zero (KIM, 2020). Alguns exemplos dessas redes pré-treinadas são VGG 16/19 e o ResNet50, que serão abordados posteriormente.

Uma vez que essas CNNs pré-treinadas são treinadas em conjuntos de dados amplos e genéricos, os recursos aprendidos podem atuar como um modelo genérico em muitos problemas de classificação, mesmo que esses novos problemas possam envolver classes completamente diferentes daquelas da tarefa original. Para utilizar tal técnica, é recomendado que a CNN pré-treinada tenha excelente generalização no seu problema original (GUMELAR *et al.*, 2020). As principais vantagens do aprendizado por transferência, são:

- Não requer um poder computacional robusto;
- Possibilidade de utilizar um conjunto de dados pequenos;
- Tempo de treinamento menor;
- Possibilidade de ajustes finos, por exemplo, adição de mais camadas e parâmetros ou treinadas somente as camadas totalmente conectadas.

A Figura 5, ilustra a aprendizagem por transferência pré-treinada no *ImageNet* e posteriormente treinada com imagens de espectrogramas da tipologia da voz. *ImageNet* é um conjunto de dados que consiste em mais de 14 milhões de imagens pertencentes a quase 1000 classes diferentes (SIMONYAN; ZISSERMAN, 2015; VAVREK et al., 2021a).

Figura 5 - Aprendizagem por transferência: uma rede pré-treinada no *ImageNet*, depois com imagens de espectrogramas da tipologia da voz



Essas redes pré-treinadas são divididas em duas partes, base convolucional e classificador:

- A base convolucional, comumente composta por pilhas de camadas convolucionais e *pooling*, visa gerar características a partir das imagens apresentadas a rede. Esse processo é chamado de extração de recursos.
- O classificador, muitas vezes composto por camadas totalmente conectadas, classifica a imagem com base nas características extraídas pela base convolucional.

O fluxo de trabalho padrão de aprendizagem por transferência comumente utilizado, é (GUMELAR *et al.* (2020).

- Selecionar uma CNN pré-treinada que resolva um problema semelhante ao que se pretende ser resolvido;
- Substituir a camada de classificação por uma nova para ser treinada no novo conjunto de dados de imagens;
- Congelar a base convolucional e treinar a rede neural no novo conjunto de dados.

Para enfrentar os desafios na classificação da tipologia da voz imposto pelos poucos dados de imagens disponíveis para este estudo, bem como a limitação de recursos computacionais, o aprendizado por transferência (*transfer learning* - TL) foi a opção utilizada para classificação dos espectrogramas.

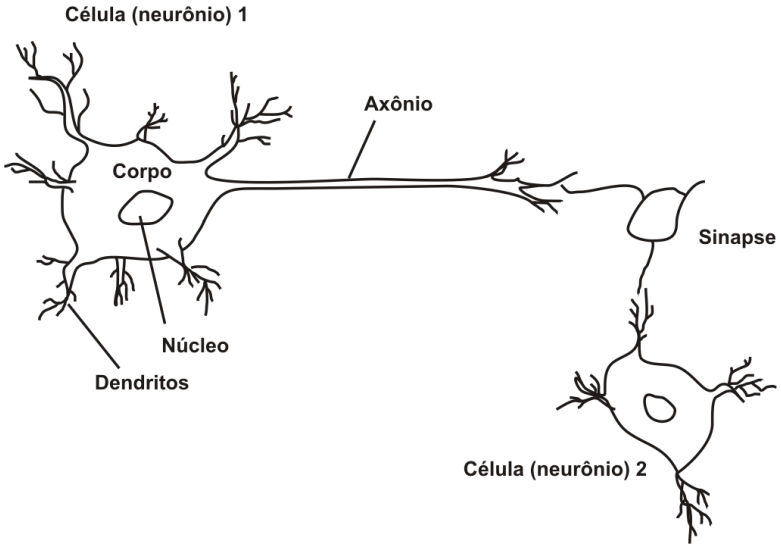
2.4.6 Redes Neurais

Nas últimas décadas, o campo de AM trouxe uma variedade de notáveis avanços em algoritmos de aprendizado de máquina supervisionada, bem como técnicas mais eficientes de pré-processamento. Entre esses avanços, a evolução das redes neurais artificiais (RNAs) vem tendo maior destaque com uma arquitetura de camadas cada vez maior e mais profunda. Possuem processamento altamente paralelo e distribuído apresentando a capacidade de aprender e armazenar conhecimento experimental, sendo amplamente utilizadas na resolução de problemas de reconhecimento de padrões.

As Redes Neurais Artificiais (RNA) são técnicas computacionais que utilizam um modelo de algoritmo baseado na estrutura neural de organismos inteligentes. Essas redes adquirem conhecimento através da experiência sobre um modelo de soluções existentes. Do mesmo modo que os neurônios naturais, as redes de neurônios artificiais se comunicam através de sinapses.

A sinapse é a região onde dois neurônios entram em contato e através da qual os impulsos nervosos são transmitidos entre eles, ou seja, onde há a troca de informações. A Figura 6 ilustra a estrutura de um neurônio biológico.

Figura 6 - Neurônio biológico



Fonte: Pesquisador do estudo (2022)

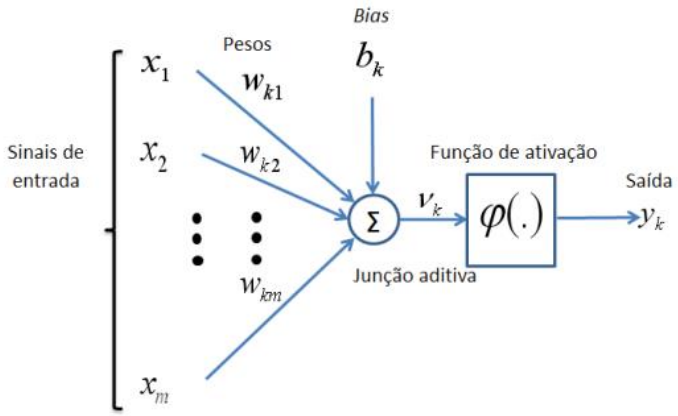
A RNA é um modelo computacional de aprendizado de máquina supervisionado de inspiração no sistema neural biológico. São sistemas paralelos compostos por unidades de processamento denominados neurônios ou *perceptrons* artificiais, distribuídos em camadas e intensamente conectados com coeficientes (pesos) que constituem a estrutura neural. Essas camadas estão encadeadas para mapear os dados de entrada para fazer predições de acordo com as classes definidas, cada camada faz um processamento da informação e passa para a camada seguinte, ou seja, as saídas de algumas camadas tornam-se entradas para as próximas camadas.

As vantagens das redes neurais se destacam em sua habilidade para executar computação distribuída e na sua generalização. Segundo Haykin (2001), a generalização é a capacidade de apresentar saídas coerentes para entradas que não estavam presentes durante o processo de treinamento (aprendizagem). Essas duas características fazem com que as RNAs sejam capazes de solucionar problemas de maior grau de complexidade, bem como capacidade de trabalhar com problemas não lineares. Algumas aplicações clássicas das RNAs são:

- Reconhecimento de padrões em imagens;
- Processamento de voz;
- Processamento de sinais;
- Diagnósticos médicos.

Os neurônios são unidades menores que compõem as RNA. São unidades de processamento fundamentais de uma rede neural. O neurônio é constituído por um conjunto de conexões conhecidas como pesos sinápticos, um somador e uma função de ativação (HAYKIN, 2001). Cada neurônio possui um conjunto de entradas e de saídas ligadas a outros neurônios, exceto os neurônios de entrada e de saída que possuem um ou outro, funciona como uma unidade autônoma de processamento, cuja tarefa individual é converter um sinal de entrada em outro sinal de saída. A Figura 7 exibe uma representação do neurônio artificial apresentado por Rosenblatt (1958), conforme extraído de Haykin (2001).

Figura 7 - Modelo matemático de um neurônio artificial



Fonte: Haykin (2001)

m – é o número de sinais de entrada do neurônio;
x_m – é a entrada do neurônio;
w_{km} – é o peso da sinapse;
b – é o viés ou bias;
v_k – é a saída intermediária;
ϕ(.)– é a função de ativação, do k-ésimo neurônio;
y_k - é a saída ativada.

O vetor x_m (x_1, x_2... x_m) é apresentado como a entrada do neurônio, o vetor w_{km} (w_{k1}, w_{k2}, ..., w_{km}) corresponde ao conjunto de pesos sinápticos (representam o aprendizado da rede) que ponderam cada elemento presente em x_m, sendo este acrescido com o valor da polarização (bias) b. O *bias* tem o efeito de aumentar ou diminuir o grau de liberdade da função de ativação, conforme seu sinal muda de positivo para negativo. Posteriormente, são inseridos no combinador linear. Para se obter o valor da saída intermediária v_k realiza-se o somatório da multiplicação entre as entradas s seus respectivos pesos numa função denominada soma que passa pela função de ativação ϕ (·) que é responsável por decidir se um neurônio deve ser ativado ou não. A resposta da função de ativação a este potencial interno corresponde a resposta

final da saída y_k do neurônio. Assim, pode-se descrever matematicamente o neurônio k através da equaçã 2:

$$y_k = \sum_{n=0}^{m} x_n . w_{kn} + b \qquad (2)$$

A função de ativação ou de transferência modula a amplitude do intervalo do sinal de saída do neurônio para algum valor finito, dentre as funções de transferência utilizadas em RNA, destacam-se:

- Sigmóide - também conhecida como função logística, é uma função contínua que tem a forma de um S que faz a transição gradual entre os dois extremos, variando de 0 a 1, representada pela Equação 3:

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (3)$$

- *Softmax* - utilizada para a classificação de várias classes, representada pela Equação 4

$$S(x) = \frac{e^{j}}{\sum_i e^{i}} \qquad (4)$$

- ReLU – tem como objetivo introduzir a não linearidade na rede. Representada pela Equação 5.

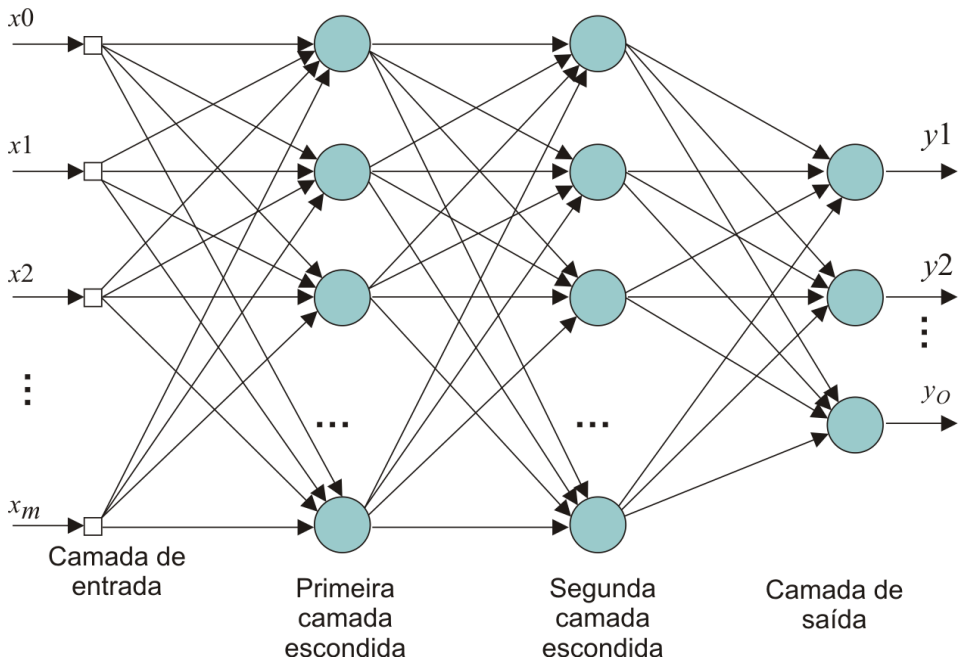
$$RelU(x) = \max(0, x) \qquad (5)$$

A principal habilidade de uma RNA é a capacidade de aprender por meio de treinamento e, com isso, melhorar seu desempenho, em termos de reprodução de uma saída desejada, partindo-se de um conjunto de entrada.

Uma RNA (*Multilayer Perceptron Neural Network* – MLP) é subdividida em camadas: a camada de entrada, que recebe como entrada os dados brutos que podem ser desde valores numéricos, até imagens e vídeos e os repassa para o restante da rede; uma ou mais camadas ocultas, são camadas intermediárias entre a camada de entrada e a camada de saída, onde serão aplicadas as funções que influenciarão na propagação da informação de camada para camada. Por fim, a camada de saída que recebe como entrada os dados processados e produz os resultados, é responsável por retornar uma resposta na forma de classificação ou previsão.

Essas camadas são o componente chave que permite que uma rede neural aprenda tarefas complexas e alcance um excelente desempenho (GOODFELLOW; BENGIO; COURVILLE, 2016). A camada de entrada recebe o vetor de dados x_m que serão apresentados para treinamento da rede. Cada um dos valores x_m de entrada está conectado a todos os pesos sinápticos presentes na primeira camada oculta. Uma rede neural profunda (*Deep Neural Network* - DNN) utiliza-se de múltiplas camadas compostas por neurônios ocultos totalmente conectados. A Figura 8 ilustra a estrutura de uma DNN organizada em camadas.

Figura 8 - Estrutura de uma DNN



Fonte: Pesquisador do estudo (2022)

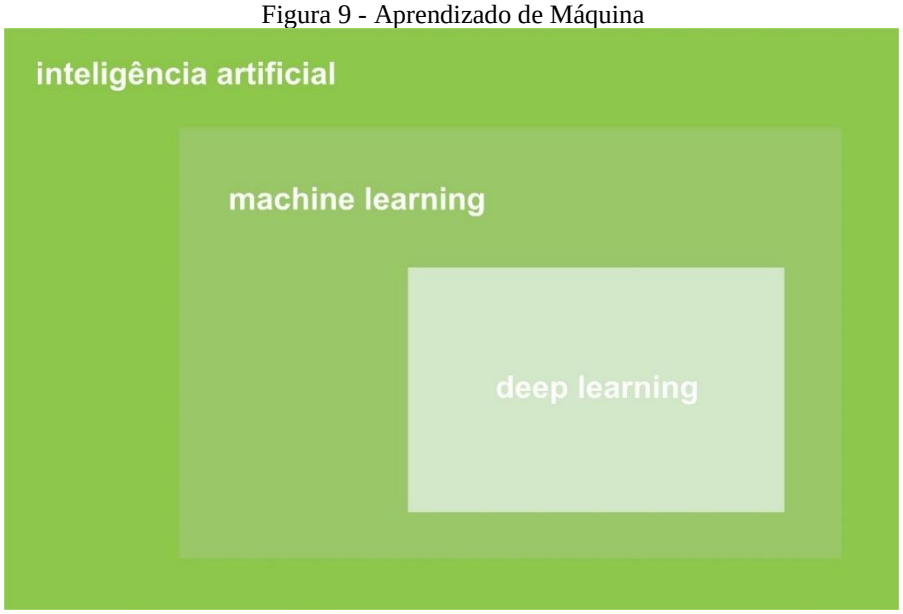
Dito isto, os avanços das RNAs vêm proporcionando ao AM uma maior capacidade de trabalhar com uma grande quantidade de dados e imagens com resultados expressivos (CHEN; CHEN, 2022). Dentre essas evoluções das RNAs, pode-se citar as Redes Neurais Profundas (RNP) ou *Deep Neural Network* (DNN) (DING *et al.*, 2021; GOODFELLOW; BENGIO; COURVILLE, 2016).

2.4.7 *Deep Learning*

O Aprendizado Profundo (*Deep Learning* - DL) também conhecido como aprendizado hierárquico ou aprendizado de máquina profundo é uma subárea do AM, baseados em Redes Neurais Artificiais empilhadas, tem como característica uma quantidade maior de camadas,

parâmetros e operações de convolução que ocorrem em uma ou mais camadas na sua arquitetura (HE *et al.*, 2016; VAVREK *et al.*, 2021). A Figura 9 ilustra a relação entre inteligência artificial (IA), *Machine Learning* (ML) e *Deep Learning* (DL). Conforme pode-se observar, o *Machine Learning* e a *Deep Learning* são subconjuntos da Inteligência Artificial.

Figura 9 - Aprendizado de Máquina



Fonte: Pesquisador do estudo (2022)

O processo de aprendizado é profundo porque a estrutura da DL consiste em várias camadas: em camadas de entrada, convolução, *pooling*, totalmente conectadas e de saída ordenada hierarquicamente (PANWAR *et al.*, 2020). Cada camada contém unidades que convertem os dados de entrada em informações que a próxima camada pode usar para executar uma determinada tarefa preditiva.

O aprendizado profundo usa uma cascata de várias camadas de unidades de processamento não linear para extração e transformação de recursos. As camadas inferiores próximas à entrada de dados aprendem recursos simples, enquanto as camadas superiores aprendem recursos mais complexos derivados de recursos da camada inferior (MICROSOFT, 2022). Isso significa que o aprendizado profundo é adequado para analisar e extrair conhecimento útil de grandes quantidades de dados e dados coletados de diferentes fontes.

Nos últimos anos, o DL vem evoluindo acentuadamente. O aumento do poder computacional, proporcionou melhorias significativas na sua arquitetura, desenvolvendo novo modelos de DL e obtendo resultados expressivos especialmente no reconhecimento de padrões (AGGARWAL, 2018; CHEN; CHEN, 2020). O DL vem demonstrando excelentes desempenhos em tarefas de classificação e, desde então, o uso dessa tecnologia vem crescendo

rapidamente, alcançando um desempenho surpreendente na classificação de sinais acústicos e de áudio, reconhecimento automático de fala e no processamento digital de imagens, principalmente nas imagens médicas (GUMELAR *et al.*, 2020; MAHAJAN; CHAUDHARY, 2019).

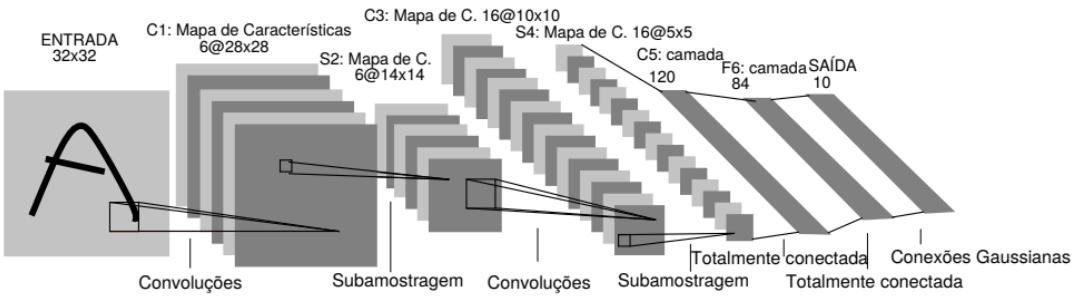
Nesse sentido, DL vem superando as técnicas tradicionais de AM, principalmente no reconhecimento de padrões. O desempenho de ponta do DL em relação às abordagens tradicionais de aprendizado de máquina, permite a construção de modelos para reconhecimento de fala, detecção e classificação de sinais de vozes saudáveis e/ou patológicas, tradução automática, processamento de imagens médicas, bioinformática, robótica e muito mais (VAVREK *et al.* 2021).

Alguns modelos de DL se tornaram populares pelo seu bom desempenho e alta aplicabilidade em problemas reais, como por exemplo as Redes Neurais Recorrentes Profundas (*Deep Recorrent Neural Networks* - DRNNs), utilizadas com dados sequenciais para tarefas de previsão e as Redes Neurais Convolucionais ou ConvNet (*Convolutional Neural Network* - CNN), lida com dados topológicos como imagens, vídeo e áudio. As CNNs, nos últimos vem recebendo grande destaque, principalmente, a partir dos concursos de classificação de imagens realizados a partir de 2012 (AGGARWAL, 2018). Desde então, as CNNs vem sendo comumente utilizadas em tarefas de classificação de imagens médicas.

2.4.7.1 Rede Neural Convolucional

As CNNs foram inicialmente propostas por LeCun *et al.* (1998) como modelo LeNet, como ilustrado na Figura 10. Foi utilizada para reconhecimento de escrita manual de símbolos numéricos em documentos.

Figura 10 - Arquitetura LeNet

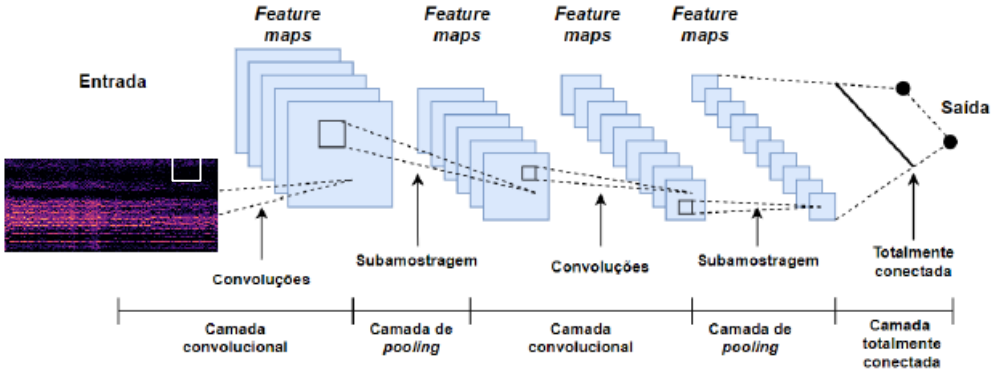


Fonte: Adaptada de LeCun *et al.* (1998)

As CNNs receberam destaque com os resultados obtidos por Krizhevsky *et al.* (2012) na competição ILSVRC para classificar objetos presentes em imagens. As CNNs têm como objetivo subdividir continuamente imagens com o intuito de extrair características, realizando o aprendizado por meio dos métodos de propagação de sinais e retropropagação de erros.

Nas CNNs as camadas iniciais são utilizadas para extrair características das imagens em que os neurônios não são totalmente conectados com os da camada seguinte. As camadas finais são responsáveis por interpretar as informações obtidas por meio das camadas iniciais e gerar uma previsão, sendo totalmente conectadas. A Figura 11 ilustra uma arquitetura de uma CNN (VAVREK *et al.* 2021).

Figura 11 - Exemplo simples de arquitetura de uma CNN



O objetivo da CNN é reduzir as imagens para que sejam mais fáceis de processar sem perder recursos valiosos para uma previsão mais eficiente. Sua arquitetura básica consiste em três tipos de camadas: camada convolucional, camada de *pooling* (subamostragem) e camada totalmente conectada (*Full Connection* - FC).

A camada convolucional é responsável por identificar e extrair características da imagem de entrada por meio de filtros convolucionais de tamanhos reduzidos e repassa para a próxima camada na forma de mapa de características. Dada a entrada de uma imagem bi-dimensional *x*, considere sua divisão em uma entrada sequencial $x = (x_1, x_2 \ldots x_n)$. Para compartilhar o peso, a camada convolucional é definida como:

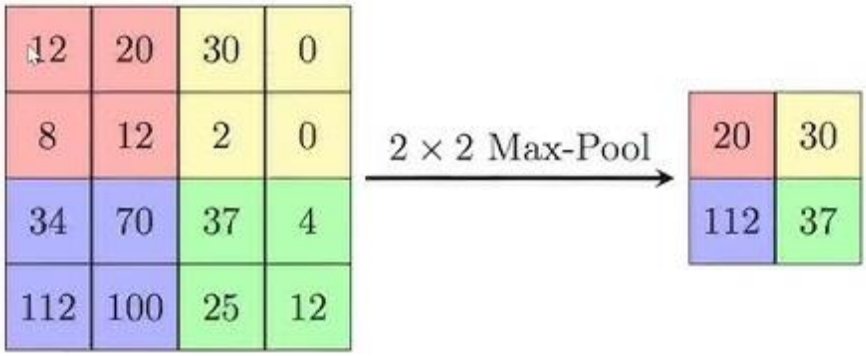
$$y_j = f(\sum_i K_{ij} \otimes x_i + b_j)$$

Em que y_j denota a j-ésima saída para a camada convolucional e K_{ij} denota o *kernel* convolucional com o i-ésimo mapa de entrada x_i. $\otimes$ denota o operador de convolução discreta

e b_j denota o *bias*. Além disso, *f* denota uma função de ativação tipicamente não linear, já que permite o aprendizado de características mais complexas dos dados.

A camada de *pooling* (subamostragem) tem como objetivo reduzir a dimensão do mapa de características, assim como tornar a rede menos sensível às distorções e deslocamentos da imagem. Ela simplifica a informação, resumindo os dados do mapa de características da imagem em um único valor e os repassa para uma camada totalmente conectada. Os parâmetros considerados são o tamanho da janela (ou filtro) e deslizamento utilizado. A forma mais comumente utilizada é um filtro 2 × 2 com deslizamento 2, que reduz a largura e altura do mapa de características pela metade. A Figura 12 ilustra a aplicação do *max pooling* em um mapa de características de uma imagem utilizando o filtro 2x2.

Figura 12 - Aplicação de *max pooling* em uma imagem 4x4 utilizando um filtro 2x2



Fonte: Panwar *et al.* (2020)

Por fim, as camadas totalmente conectadas e uma camada softmax. Essas camadas são posicionadas no final da rede para classificação dos mapas de características obtidos pelas camadas anteriores. A camada totalmente conectada inicia o processo de classificação, é formada por uma estrutura de conexão entre os neurônios, que prevê que cada neurônio de saída tem ligação a todos os neurônios da camada anterior por um peso exclusivo entre eles, e um bias compartilhado. A camada softmax (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) é responsável por transformar as saídas dos neurônios em uma quantidade de neurônios igual ao número de classes, tendo como saída a probabilidade normalizada para cada classe (ABED MOHAMMED *et al.*, 2020; TRINH; DARRAGH, 2018; TRINH; DARRAGH, 2019; VAVREK *et al.*, 2021b).

Além das camadas de convolução, *pooling* e totalmente conectadas, há a camada de *dropout*. *Dropout* tem o objetivo de reduzir o *overfitting* (sobreajuste) da rede, desativando alguns neurônios aleatoriamente durante o treinamento e transferência de aprendizado. Para

isso, deve-se informar a porcentagem de neurônios que serão desativados. O *overfitting* ocorre quando a rede decora os dados utilizados no aprendizado e não é capaz de aplicar as relações aprendidas em dados novos, ou seja, nos dados de treinamento tem um desempenho excelente, contudo, nos dados de teste o resultado é ruim. Esse problema já foi investigado em diversas pesquisas, ressaltando a necessidade de utilizar técnicas como o *dropout* (SRIVASTAVA *et al.*, 2014).

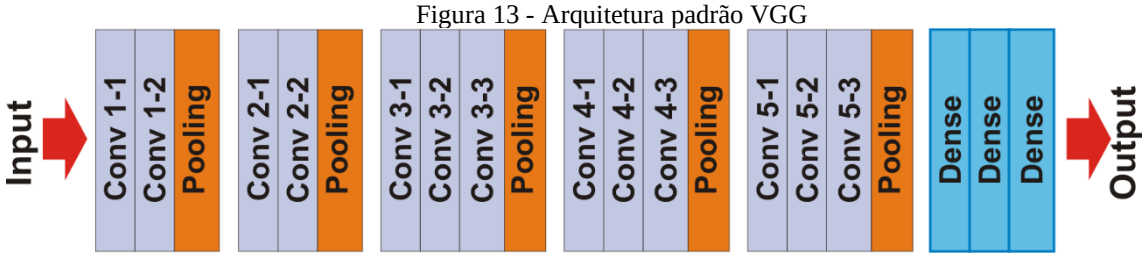
Neste trabalho, é empregado a CNN VGG-16 para classificar a tipologia da voz e os resultados obtidos são comparados com mais duas CNNs pré-treinadas, são elas: VGG-19 e a ResNet50. Essas CNNs vêm recebendo destaque nas pesquisas referentes a classificação de patológica da voz por meio da análise de imagens (DING *et al.*, 2021; GUMELAR *et al.*, 2020b).

2.4.7.2 Arquitetura VGG

VGG é uma arquitetura de CNN proposta por *Visual Geometry Group* da Universidade de Oxford (SIMONYAN; ZISSERMAN, 2015). São um conjunto de redes da mesma família que varia de acordo com o número de camadas na sua configuração. Por ser uma arquitetura mais simples entre CNNs pré-treinadas, vem tendo destaque em tarefas de classificação de imagens médicas, bem como sendo objeto de estudo na classificação de patologias da voz (VAVREK *et al.*, 2021). As duas configurações mais utilizadas, são: VGG-16 e VGG-19.

O VGG foi treinado em mais de um milhão de imagens do banco de dados *ImageNet*. A rede pode classificar imagens com aproximadamente 1000 categorias e tem como principal características utilizar tamanho de filtro em janelas de *pixels* com dimensão 3x3. Na sua arquitetura padrão, a entrada para a camada conv1 é de uma imagem RGB com tamanho fixo de 224 x 224, depois cinco blocos de convolução, seguido por 3 camadas densas totalmente conectadas, como ilustrado na Figura 13 (GUMELAR *et al.*, 2020a; VAVREK *et al.*, 2021a).

Figura 13 - Arquitetura padrão VGG

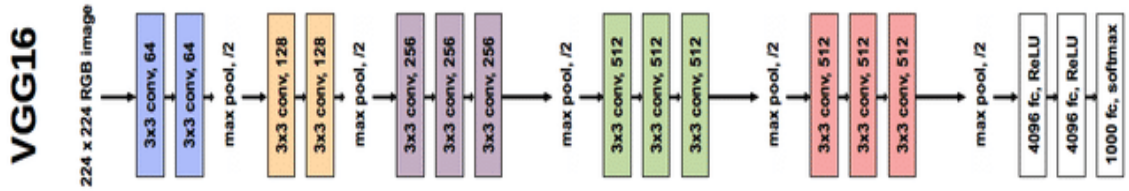


Fonte: Pesquisador do estudo (2022)

2.4.7.2.1 VGG-16

O VGG-16 faz parte da arquitetura VGG, o número 16 identifica o número de camadas desta arquitetura. Sua configuração inicia no primeiro bloco com duas com camadas convolucionais de tamanho 3 × 3, com 64 filtros nessa camada, o segundo bloco também possui duas camadas (128 filtros por camada). Os próximos três blocos contêm três camadas de tamanho 3 × 3. O número de filtros por camada é 256, 512 e 512, respectivamente (GUMELAR *et al.*, 2020a; SIMONYAN; ZISSERMAN, 2015; VAVREK *et al.*, 2021a). A quantidade de filtros aumenta em valor de 2 após cada camada de sub amostragem (*pooling*). As camadas *pooling*, que totalizam cinco ao todo, realiza o processo sobre áreas 2x2 com passo do mesmo tamanho, sem sobreposição. Além da diminuição dos parâmetros, em cada uma das camadas totalmente conectadas, para aumentar a não linearidade do modelo, foi adicionado à função de ativação ReLU. Por fim, a camada final dessa *softmax* responsável pela função de classificação. A Figura 14 ilustra a arquitetura do VGG-16.

Figura 14 - Modelo VGG-16



Fonte: Leonardo *et al.* (2018)

A VGG-16, por ser uma CNN mais simples e exigir um custo computacional menor, quando comparada a outras CNN, foi a CNN selecionada como base para este estudo.

2.4.7.2.2 VGG-19

O conceito para o VGG-19 é o mesmo do VGG-16, exceto pela quantidade de camadas na sua configuração. O VGG-19 consiste em 19 camadas (16 camadas de convolução, 3 camadas totalmente conectadas). A Figura 15 ilustra um comparativo entre as 2 CNNs.

A Figura 15 ilustra a configuração geral das CNNs VGG 16 e 19 que usam o mesmo princípio, mas variam apenas pela quantidade de camadas.

Figura 15 - Arquitetura VGG-16 e VGG-19

Arquiteturas	
VGG-16	**VGG-19**
conv3-64	conv3-64
conv3-64	conv3-64
maxpool	
conv3-128	conv3-128
conv3-128	conv3-128
maxpool	
conv3-256	conv3-256
conv3-256	conv3-256
conv3-256	conv3-256
	conv3-256
maxpool	
conv3-512	conv3-512
conv3-512	conv3-512
conv3-512	conv3-512
	conv3-512
maxpool	
conv3-512	conv3-512
conv3-512	conv3-512
conv3-512	conv3-512
	conv3-512
maxpool	
FC-4096	
FC-4096	
FC-1000	
soft-max	

Fonte: Pesquisador do estudo (2022)

2.4.7.3 ResNet50

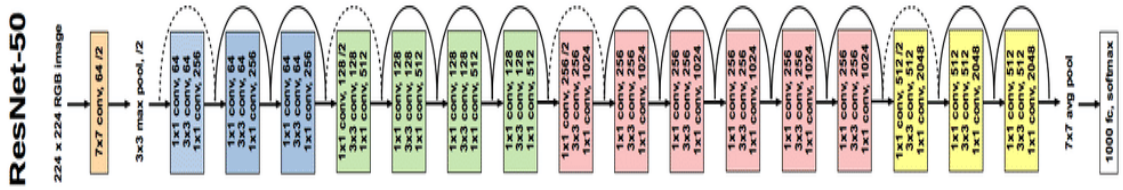
Resnets50 (HE *et al.*, 2016b; MAHAJAN; CHAUDHARY, 2019b) são um tipo de CNNs chamadas Redes Residuais. Tem muito mais camadas quando comparadas com VGG-16 ou VGG-19. São CNNs com 50 camadas de profundidade. Uma rede neural residual (ResNet) é uma rede neural artificial (ANN) de um tipo que empilha blocos residuais uns sobre os outros para formar uma rede.

Resnet introduziu conexões residuais entre camadas, o que significa que a saída de uma camada é uma convolução de sua entrada mais sua entrada. Além disso, as camadas em um Resnet também usam a normalização em lote, que também foi incorporada ao VGG. Foi introduzida pela primeira vez por Kaiming He, Xiangyu Zhang, Shaoqing Ren e Jian Sun em

seu artigo de pesquisa de visão computacional de 2015 intitulado "*Deep Residual Learning for Image Recognition*" (HE *et al.*, 2016b). O ResNet tem muitas variantes que funcionam com o mesmo conceito, assim como o VGG, têm diferentes números de camadas. Resnet50 é usado para denotar a variante que pode funcionar com 50 camadas de rede neural (MAHAJAN; CHAUDHARY, 2019c).

Essa arquitetura de rede obteve sucesso quando venceu a competição ILSVRC 2015, com um erro de apenas 3,57%. Tem como principal característica o uso de uma conexão residual para resolver o problema da degradação ao treinar uma rede muito profunda. A ResNet tem menor perda de convergência sem *overfitting* e a conexão residual pode acelerar a convergência das camadas profundas. A Figura 16 mostra a sua arquitetura.

Figura 16 - Internal architecture of the ResNet50 CNN



Fonte: Leonardo *et al.* (2018)

Embora as redes neurais profundas permitam um desempenho superior em tarefas visuais, como classificação de objetos, classificação de imagens, detecção e incorporação de recursos visuais, a falta de uma explicação mais intuitiva e compreensível torna difícil a interpretação de como a rede chegou aquele resultado ou aquela decisão. Consequentemente, quando essas redes falham, falham de uma forma que pode levar a problemas preocupantes e graves, como um diagnóstico errado, mostrando uma predição equivocada e deixando a dúvida de como o sistema chegou a essa conclusão e prejudicando a tomada de decisão.

Ao construir modelos profundos, ao contrário de sistemas baseados em regras, comumente sacrificamos o grau de interpretabilidade nos módulos para obter maior desempenho por meio de maior abstração (mais camadas) e integração mais estreita (treinamento de ponta a ponta) (ARAVINDA *et al.*, 2021; CHAKRABORTY *et al.*, 2022; SELVARAJU *et al.*, 2016). Nesse sentido, há necessidade de construir modelos de DL mais "transparentes", com a capacidade de mostrar como ele tomou a decisão para prever tal resultado e avançar para sua integração significativa em nossas vidas cotidianas.

Para tanto, para entendermos e mitigar esse problema, utilizamos neste trabalho a técnica de mapeamento de ativação de classe ponderada por gradiente (*Gradient Weighted Class Activation Mapping* - Grad-CAM), proposta por Selvaraju *et al.* (2016). Essa técnica

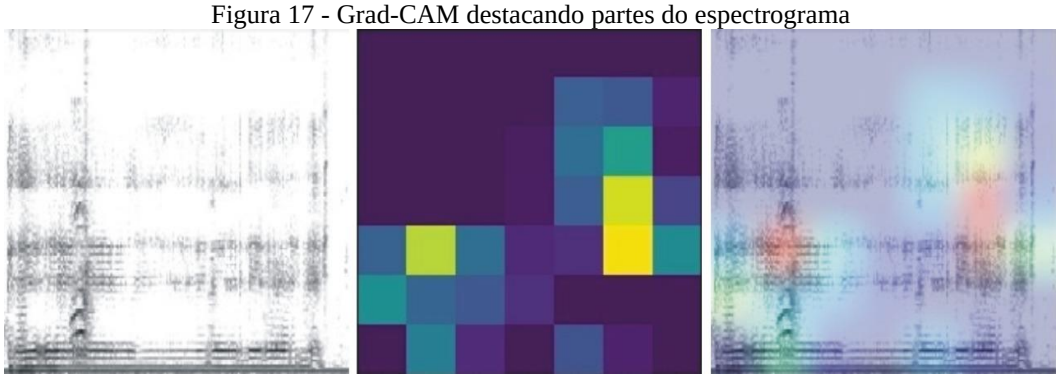
auxilia mostranto quais partes da imagem o modelo utilizou para classificar o espectrograma, destacando nas imagens as regiões de entrada que são "importantes" para as previsões do modelo - ou explicações visuais.

2.4.8 *Gradient Weighted Class Activation Mapping* (Grad-CAM)

Com os avanços no desenvolvimento de novos modelos de DL e, consequentemente, o aumento da complexidade das redes convolucionais com adição de mais camadas tornando-se cada vez mais profundas, o problema da interpretabilidade dos resultados obtidos por elas vem ganhando importância entre os pesquisadores. Redes neurais e convolucionais não tem uma estrutura naturalmente interpretáveis como regressões lineares e árvores de decisão.

Nesse sentido, Selvaraju *et al.*(2016) propôs a técnica Grad-CAM para entender como foi o processo de tomada de decisão feito pela CNN para classificar a imagem. Essa técnica destaca as partes mais importantes de uma imagem que foram utilizadas para classificação, como apresentado na Figura. 17. Muito útil para entender o processo de aprendizado da CNN e visualizar a representação interna.

Figura 17 - Grad-CAM destacando partes do espectrograma



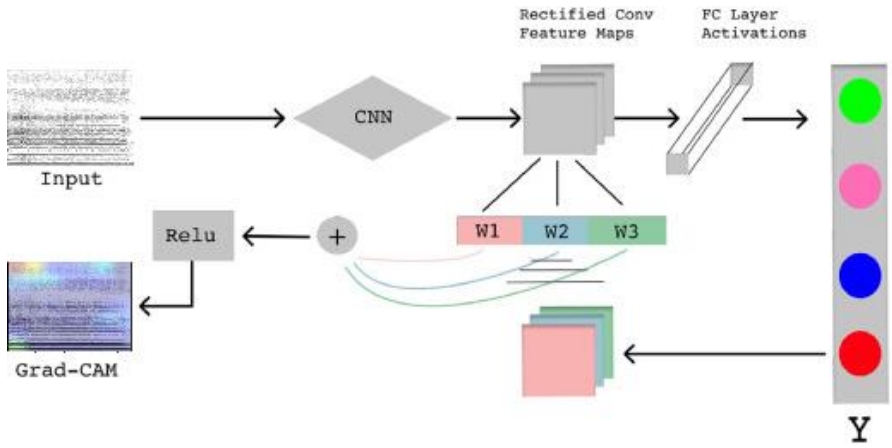
Pesquisador do estudo (2022)

Grad-CAM utiliza mapas de recursos produzidos pela última camada convolucional de uma CNN. Os autores do Grad-CAM argumentam que “podemos esperar que as últimas camadas convolucionais tenham o melhor compromisso entre semântica de alto nível e informações espaciais detalhadas” (SELVARAJU *et al.*, 2016).

Grad-CAM utiliza as informações do gradiente, fluindo para camada convolucional para gerar um mapa de calor, destacando as regiões mais importantes da imagem para tomada a decisão de qual classe a imagem pertence. Demostrando como o modelo tomou a decisão para suas previsões. A Figura 18 mostra o processo de Grad-CAM utilizando o mapa de características gerado durante o processo de treinamento do modelo CNN e os gradientes

utilizados para prever a classe alvo.

Figura 18 - Um diagrama explicativo simples do Grad-CAM.



Fonte: Panwar *et al.* (2020)

Grad-CAM não necessita que a rede seja inteiramente convolucional, permitindo que ela contenha camadas completamente conectadas. Isso dispensa a necessidade de retreinar um modelo sem suas camadas completamente conectadas. A técnica cria uma visualização levando em consideração os gradientes em relação a uma classe da saída da rede.

2.5 MÉTODO DE ANÁLISE

A viabilidade de modelos de aprendizado de máquina depende da precisão do modelo. A capacidade de discriminação e confiabilidade são dois aspectos para medir a precisão desses modelos. Para tanto, existem várias técnicas para avaliar e validar o desempenho de algoritmos e classificadores de aprendizado de máquina. Para tal, diversas medidas de avaliação foram criadas para auxiliar nessa tomada de decisão (MYTHILI; VIJAYA, 2018).

Nesse sentido, umas das formas de avaliar e validar um conjunto de dados que foram classificados é por meio de uma matriz de confusão. Esta matriz é uma ferramenta importante que determina se o valor classificado corresponde ou não ao valor real, de forma que todos os casos em cada categoria são contabilizados e os totais são exibidos na matriz. A validação representa a capacidade que um método de teste possui para identificar corretamente a classe de um indivíduo dentro de uma população.

Para avaliar a precisão dos resultados obtidos através dos classificadores, deve-se levar em consideração que os estudos de precisão no campo dos distúrbios da voz podem ser direcionados à triagem, classificação do diagnóstico e monitoramento da eficácia do tratamento

oferecido (LOPES *et al.*, 2017). Dependendo do propósito do estudo, pode-se focar nas medidas de sensibilidade (rastreamento) ou especificidade (classificação do diagnóstico). Essas medidas estão relacionadas com os resultados de classificação e diagnóstico verdadeiro.

O teste é considerado positivo (desvio) ou negativo (saudável), e o desvio presente ou ausente. O teste está correto quando ele é positivo na presença do desvio (Verdadeiro Positivo-VP) ou negativo na ausência do desvio (Verdadeiros Negativo-VN). Além disso, o teste está errado quando ele é positivo na ausência do desvio (Falso Positivo-FP), ou negativo quando o desvio está presente (Falso Negativo-FN).

Acurácia (ACC) é definida como a relação entre o número de casos corretamente classificados e todos os casos expostos ao classificador:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{6}$$

Sensibilidade (SEN), é a capacidade do teste em identificar corretamente o desvio entre aqueles que o possuem. É definida pela relação entre o número de casos corretamente classificados com a presença do distúrbio e a quantidade total de casos com o distúrbio (LEITE; MORAES; LOPES, 2020b; LOPES; VIEIRA; BEHLAU, 2020):

$$SEN = \frac{TP}{TP+FN} \tag{7}$$

Especificidade (ESP) é a capacidade do teste em excluir corretamente aqueles que não possuem o desvio. É definida pela relação entre o número de casos corretamente classificados como saudável e a quantidade total de casos saudáveis (ANISHA, 2020):

$$ESP = \frac{VN}{VN + FP} \tag{8}$$

Precision (P), calcula a fração de detecção positiva correta da tipologia;

$$P = \frac{TP}{TP+FP} \tag{9}$$

F1-Score (F1) é definido como a média harmônica de *precision* e sensibilidade;

$$F1 = 2 * \frac{(precision * sen)}{precision + recall} \quad (10)$$

Além disso, pode-se aplicar o coeficiente de *Kappa*, que é uma medida ponderada robusta que leva em consideração concordâncias e discordâncias entre duas fontes de informação é o Coeficiente *Kappa* (MORAES; MACHADO, 2014):

$$kappa = \frac{P(O) - P(E)}{1 - P(E)} \quad (11)$$

em que:

P(O): proporção observada de concordâncias (soma das respostas concordantes dividida pelo total);

P(E): proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total).

Nesta etapa, utilizou-se a seguinte classificação para as métricas acurácias, sensibilidade, especificidade e f1-Score e Precision na avaliação dos modelos: excelente (> 90%), bom (80% –90%), aceitável (70% –80%), ruim (60% –70%) e sem capacidade de discriminação aceitável (<60%) (IYER; HOSMER; LEMESHOW, 1991). Para o kappa, empregou-se a seguinte interpretação para os valores: falta de concordância (<0); acordo ruim (0-0.19); acordo de luz (0.20-0.39); acordo moderado (0,40-0,59); acordo substantivo (0,60-0,79); acordo quase perfeito (0,80-1,00) (LANDIS; KOCH, 1977).

3 ESTADO DA ARTE

Na literatura, vários grupos de pesquisa propuseram diferentes abordagens para a identificação da presença de distúrbios de voz por meio da análise da gravação de voz usando técnicas de inteligência artificial, como, por exemplo, aprendizado de máquina ou redes neurais profundas.

Nos últimos anos, houve um aumento crescente no número de pesquisas utilizando aprendizagem de máquina para esse objetivo. Alguns dos trabalhos publicados (AMARA; FEZARI; BOUROUBA, 2016; CHEN; CHEN, 2022; FANG *et al.*, 2019; MOHAMMED *et al.*, 2020), são baseados na análise de características de voz extraídas do sinal de voz, os quais como dados unidimensionais (1D). Outros (ALHUSSEIN; MUHAMMAD, 2018; MUHAMMAD *et al.*, 2018; TRINH; DARRAGH, 2019; WU *et al.*, 2018), em vez disso, são baseados em métodos e técnicas que transformam o sinal de voz em uma imagem, ou seja, um sinal bidimensional (2D), e processa-o através de processamento de imagem baseado em IA.

Duas publicações recentes listam e descrevem as principais técnicas de AM que têm sido amplamente utilizadas para detecção dos distúrbios de voz (HEGDE *et al.*, 2019; ISLAM; TARIQUE; ABDEL-RAHEEM, 2020). Segundo os autores, os classificadores mais usados neste campo, são o *Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Convolutional Neural Network (CNN), Probabilistic Neural Network (PNN), Generalised Regression Neural Network (GRNN), K-Means Clustering, Decision Tree Algorithm e o Linear Discriminant Analysis (LDA),* entre outros (HEGDE *et al.*, 2019; ISLAM; TARIQUE; ABDEL-RAHEEM, 2020).

No entanto, estudos que utilizam DNN para detectar e classificar a tipologia do sinal da voz utilizando imagens espectrográficas ainda carece de pesquisas, bem como experimentos. Nesse sentido, dentre as pesquisas encontradas na literatura com o intuito de investigar AM na avaliação dos distúrbios da voz, Trinh e Darragh (2018), propuseram uma estrutura da *Convolutional Neural Networks* (CNNs) utilizando a biblioteca *keras* (CHRIS ALBON, 2018) para classificar distúrbio de voz. Testaram o modelo em duas bases de dados: *The Spanish Parkinson's Disease Dataset* (SPDD) *Saarbrucken Voice Database* (SVD). Obtiveram uma precisão na classificação de 0.99% na base de dados SVD e 0.96 na base de dados SPDD. Nesse trabalho, os espectrogramas foram extraídos usando a biblioteca librosa (MCFREE *et al.*, 2015), para gerar as imagens espectrográficas, utilizaram o STFT.

Fang *et al.* (2019a) criaram uma abordagem baseada em DL para a detecção de voz alteradas. Neste trabalho, amostras normais e disfônicas de oito distúrbios clínicos comuns da

voz são coletadas em um hospital universitário. Extraíram o MFCC das amostras de voz contendo som de vogal sustentado por uma duração de três segundos e, em seguida, usados em três algoritmos de aprendizado de máquina, a saber, DNN, SVM e GMM, usando uma validação cruzada de cinco vezes. Para avaliar o desempenho desses classificadores, os autores utilizam o banco de dados de distúrbios vocais do MEEI. Os resultados mostram que o melhor resultado foi para o DNN, alcançando uma acurácia de 0.99.

Chen *et al.* (2020) propõem um *Random Forest* (RF) difusa para reconhecimento de emoções de fala, os resultados experimentais mostraram que as precisões de reconhecimento para o RF foram de 0.87. Para a rede neural de retropropagação obteve 0.74%.

Gumelar *et al.* (2020) utilizaram o VGG-16 para classificar imagens de vozes normais e disfônicas. Para investigar seu desempenho, utilizaram o conjunto de dados do *Pathological Voice Disorder* (PVD). O experimento mostrou que a precisão da detecção da alteração da voz chegou a 92,03%.

Leite *et al.* (2020), utilizaram imagens espectrográficas do sinal da voz, geradas a partir da STFT, para classificar a intensidade do desvio vocal, avaliaram e compararam eficiência do modelo de classificação RF com o *Naive Bayes* (NB) *e Support Vector Machine* (SVM) utilizando a vogal /e/ sustentada. O RF obteve o melhor resultado, com acurácia de 78% e *Kappa* 0,41. Neste trabalho, utilizaram 198 amostras de uma base de dados proveniente do Laboratório Integrado de Estudos da Voz (LIEV) localizado na Universidade Federal da Paraíba - UFPB.

No trabalho de Vavrek *et al.* (2021), foi testado o modelo VGG-16 simples para classificar patologia da voz utilizando imagens espectrográficas geradas a partir do STFT. Os resultados obtidos nesse estudo Para acurácia, especificidade e sensibilidade, foram: VGG16 simples – 0.79, 0.77, 0.80; VGG16 ensemble – 0.82, 0.84, 0.79. Utilizaram a base de dados reduzida com 506 amostras do SVD. Nestes estudo, utilizou a vogal sustentada /a/.

Demircan e Örnek, (2020), utilizaram a base de dados Berlim DB – EmoDB. Duas abordagens diferentes foram apresentadas sobre o reconhecimento de emoções a partir da fala, primeiro, as imagens do espectrograma foram classificadas usando *AlexNet*, que é a arquitetura da CNN. Em seguida, a classificação foi conduzida via DNN usando os recursos MFCC extraídos manualmente. Nesse experimento, os autores obtiveram uma acurácia superior a 88%.

Chen *et al.* (2020) por sua vez, propôs uma DNN e comparou seus resultados com o SVM e RF. O DNN obteve melhor resultado na discriminar voz patológica de voz saudável, seus resultados para sensibilidade, especificidade, *precision*, acurácia e *F1score*, foram 0.97,

0.99, 0.99, 0.98 e 0.98, respectivamente. Sua base de dados era composta por 208 gravações (151 vozes patológicas e 57 saudáveis).

Por fim, Changwei *et al.*(2020) utilizaram o espectrograma da voz para classificar entre a voz normal e a voz patológica. Sua amostra foi de 53 casos de voz normal e 133 casos de voz patológica do banco MEEI. A precisão da classificação foi de 0.96.

Wu *et al.*, (2018) propôs um sistema para detecção de voz disfonia utilizando a rede neural convolucional (CNN), alcançando uma precisão geral de 0.71. No estudo de Alhussein e Muhammad (2018) utilizou um banco com 315 instâncias de dado e utilizou CNN obtendo uma acurácia de 0.97, ambos utilizando o espectrograma como entrada para classificar voz disfonia e voz normal.

Trinh e Darragh (2019) propuseram uma estrutura da CNN utilizando a biblioteca *Keras* para testar a abordagem. Obtiveram uma precisão de classificação acima de 0.95. *Keras* é uma biblioteca *python* utilizado em aprendizado profundo (CHOLLET, 2016). Neste trabalho, os espectrogramas foram extraídos utilizando o *Librosa*. As imagens geradas foram utilizadas na entrada da CNN para classificação, o banco de dados utilizado foi *The Saarbrucken Voice Database* (SVD) (Pützer & Barry, s.d.) com aproximadamente 1800 vozes. *Librosa* é uma biblioteca *python* utilizado para análise de música e áudio (MCFEE *et al.*, 2015).

Miramont *et al.* (2020), propõem uma abordagem de reconhecimento de padrões para a classificação automática de sinais de voz com base em uma máquina de vetores de suporte linear de várias classes e usando parâmetros bastante conhecidos como *Jitter*, *Shimmer*, razão harmônica/ruído e pico de proeminência cepstral em combinação com medidas de dinâmica não linear. Também foram encontradas diferenças estatisticamente significativas entre todos os tipos para todos os recursos. Precisões acima de 0.82 foram estimadas em todos os conjuntos de dados intra e entre dados usando validação cruzada.

No Quadro 1 são apresentadas de forma sucinta as informações das pesquisas que foram descritas neste Capítulo.

Quadro 1 – Resumo das pesquisas

Título	**Metodologia**	**Modelo(s)**	**Resultados**
Detection of pathological voice using cepstrum vectors: A Deep Learning approach (FANG *et al.*, 2019)	Testaram três algoritmos AM com banco de dados de distúrbios vocais do MEEI. Utilizou 60 amostras de vozes normais e 402 patológicas	DNN, SVM e GMM,	Para Acurácia: GMM – 0.94 SVM – 0.90 DNN – 0.99
Pathological speech classification	Testaram o modelo em duas bases de dados: *The Spanish Parkinson's*	Convolutional Neural Networks (CNN)	Para acurácia SVD – 0.99

using a Convolutional Neural Network (TRINH; DARRAGH, 2018)	*Disease Dataset* (SPDD) e *Saarbrucken Voice Database* (SVD). Utilizaram o STFT para transformar os sinais de áudio em espectrograma		SPDD – 0.96
Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando Random Forest (LEITE; MORAES; LOPES, 2020)	Utilizaram 198 amostras de indivíduos com desvios vocais classificados com intensidade entre leve e moderada. A vogal /ɛ/ sustentada foi selecionada para este estudo. Utilizaram o STFT para gerar as imagens	Random Forest (RF), Naive Bayes (NB) e Support Vector Machine (SVM)	Para kappa, acurácia e AUC RF – 0.41; 0.78; 0.70 NB – 0.38; 0.71; 0.59 SVM – 0.30; 0.65; 0.50
Deep convolutional neural network for detection of pathological speech (VAVREK *et al.*, 2021)	Utilizaram a base de dados reduzida com 506 amostras do *Saarbrucken Voice Database* (SVD). Utilizou a vogal sustentada /a/. Utilizaram o STFT para transformar os sinais de áudio em espectrograma. Utilizaram 60% dos dados para treinamento, 20% para validação e 20% para teste.	VGG16	Para acurácia, especificidade e sensibilidade VGG16 – 0.79; 0.77; 0.80 VGG16 ensemble – 0.82; 0.84; 0.79
Comparison of the Effects of Mel Coefficients and Spectrogram Images via Deep Learning in Emotion Classification (DEMIRCAN; ÖRNEK, 2020)	Utilizou a base de dados Berlin DB – EmoDB. Duas abordagens diferentes foram apresentadas sobre o reconhecimento de emoções a partir da fala. Primeiro, as imagens do espectrograma foram classificadas usando AlexNet, que é a arquitetura da CNN. Em seguida, a classificação foi conduzida via DNN usando os recursos MFCC extraídos manualmente.	AlexNET	Acurácia de 0.88
Multi-channel spectrograms for speech processing applications using Deep Learning methods (ARIAS-VERGARA *et al.*, 2021)	Espectrogramas Mel, *Gammatone spectrograms e Continuous Wavelet Transform* combinados para formar espectrogramas de três canais. Utilizou CNN para realizar a classificação binária	CNN	Precision – 0.86 Recall – 0.83 F1 – 0.84

Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction (CHEN *et al.*, 2020)	Utilizou as bases de dados Berlin DB – EmoDB e CASIA-*Chinese Emotional Speech Corpus*. Propôs um RF múltiplo aleatório difusa de duas camadas para reconhecimento de emoções de fala. Uma classificação separada de certas emoções que são difíceis de reconhecer até certo ponto	RF	Acurácia de 0.87
Pneumonia Detection Using Deep Learning Based on Convolutional Neural Network (RAČIĆ *et al.*, 2021)	Uso do algoritmo baseado em CNN para processar e classificar imagens de raios-X de tórax. Utilizou a base de dados do *Guangzhou Women and Children's Medical Center*. O conjunto de dados contém 5856 imagens de raio-X de tórax. 80% das imagens são utilizadas treinamento, 10% validação e 10% para teste.	CNN	Acurácia de 0.90
Enhancing Detection of Pathological Voice Disorder Based on Deep VGG-16 CNN (GUMELAR *et al.*, 2020)	Neste experimento, utilizaram o VGG-16 para classificar vozes normais e patológicas. Ao investigar o desempenho do o VGG-16, utilizaram o conjunto de dados do *Pathological Voice Disorder* (PVD), que compreende centenas de arquivos de som PVD amostrados. O experimento mostrou que a precisão da detecção da patologia da voz chega a 92,03%.	VGG-16	Acurácia de 0.92
Classification of Normal and Pathological Voices Using Convolutional Neural Network (CHANGWEI *et al.*, 2020)	Neste artigo, utilizaram o espectrograma da voz para classificar entre a voz normal e a voz patológica. Sua amostra foi de 53 casos de voz normal e 133 casos de voz patológica do banco de dados Massachusetts *Eye and Ear Infirmary* (MEEI) para experimentos.	CNN	Acurácia de 0.96

Fonte: Pesquisador do estudo(2022)

Assim, pode-se concluir nos trabalhos correlatos dos últimos anos, que a busca por novas técnicas ou construção de ferramentas para classificação automática utilizando AM, tem

sido amplamente estudada e obtendo bons resultados. A base de dados mais utilizada para o treinamento dos modelos é o SVD. Por outro lado, não foi encontrado trabalhos na literatura que utilizem uma arquitetura de CNN para classificar espectrogramas da tipologia do sinal da voz da vogal sustentada “é”, o que pode contribuir para potencializar a eficiência nas análises espectrográficas, visto que, para fazer essa classificação, é necessária uma avaliação criteriosa com base no julgamento visual feito por fonoaudiólogos experientes

Nesse sentido, este trabalho se mostra relevante e pode contribuir para que outros estudos possam iniciar suas pesquisas utilizando CNN e, assim, auxiliar o profissional da voz na construção de ferramentas que possam melhorar o processo de triagem, os procedimentos de avaliação e monitoramento dos pacientes, tanto *off-line* com gravação seguida de análise, quanto on-line com gravação e análise em tempo real.

Por fim, diante da importância da classificação da tipologia do sinal da voz para avaliação da qualidade da voz, para prevenção de agravos relacionados às alterações vocais, bem como para promover melhores informações para avaliação clínica no tratamento ou reabilitação de pessoas que possuem desvios/patologias vocais, este estudo se justifica pela relevância de apontar uma proposta promissora para potencializar a análise acústica utilizando a espectrografia, como ferramenta de apoio ao profissional da voz, por meio de sua tridimensionalidade (frequência/intensidade/tempo).

4 METODOLOGIA

Toda parte computacional, foi desenvolvida utilizando a linguagem de programação *Python* versão 3.8.5 (PYTHON, 2019), as bibliotecas *Keras* versão 2.4.3 (ALBON, 2018) para criar a CNN e *librosa* versão 0.8.0 (MCFEE *et al.*, 2015) utilizada para gerar os espectrogramas do sinal de voz. *Librosa* é uma biblioteca *python* bastante utilizada para processamento e análise de música e áudio. Fornece implementações de uma variedade de funções comuns usadas em todo o campo de recuperação de informações do áudio. Por fim, para treinar e executar o sistema desenvolvido, utilizou-se um *desktop* com processador Intel Xeon W-2133 de 3.60GHz com 16G de memória RAM com duas GPUs NVidia Quadro P1000.

4.1 TIPO DE ESTUDO

Esta pesquisa se trata de um estudo Retrospectivo, pois o pesquisador estuda os pacientes a partir de um desfecho (FREITAS, 2017), realizado por meio de investigação documental, transversal e de acurácia diagnóstica, avaliado e aprovado pelo comitê de Ética em Pesquisa da Instituição de origem, com o parecer de número 52492/12.

4.2 AMOSTRA DO ESTUDO

Participaram deste estudo, a população de pacientes que procurou o Laboratório de Estudos da Voz (LIEV) do Departamento de Fonoaudiologia (UFPB) por demanda espontânea ou encaminhada por otorrinolaringologista e/ou cirurgião de cabeça e pescoço, no período compreendido entre abril de 2012 e abril de 2017.

Até junho de 2021, o LIEV contava com 1.700 pacientes, que se submeteram à triagem vocal, onde foram coletados dados sociodemográficos, anamnese específica relacionada à queixa vocal, aplicação de protocolos de autoavaliação vocal e gravação da voz. Todas as informações dos pacientes estão contidas em prontuários devidamente enumerados e arquivados, assim como em arquivos digitalizados, especificamente quanto aos sinais de voz.

4.3 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

Foram incluídas no estudo amostras de adultos entre 18 e 65 anos, evitando, assim, o período da muda vocal e da presbifonia, respectivamente, submetidos à avaliação laringológica (exame laríngeo visual), incluindo um relatório otorrinolaringológico.

Foram excluídos do estudo indivíduos com distúrbios cognitivos ou neurológicos que impossibilitaram o uso de procedimentos de gravação, usuários profissionais de voz e

indivíduos que já haviam sido submetidos à terapia de voz formal ou que haviam sido submetidos a cirurgia na região da cabeça ou pescoço nos últimos 18 meses. Também aplica critérios de exclusão relacionados à qualidade dos sinais de voz. As amostras que apresentaram as seguintes características durante a gravação foram excluídas do estudo: duração inferior a cinco segundos, presença de recorte de pico no sinal acústico e relação sinal-ruído (SNR) abaixo de 30dB SPL (DELIYSKI, 2005).

Os indivíduos foram selecionados conforme os seguintes critérios de elegibilidade:

- Ter realizado gravação da vogal /ɛ/ sustentada com duração mínima de seis segundos;
- Apresentar laudo otorrinolaringológico referente ao exame visual laríngeo para confirmação do diagnóstico de distúrbio de voz;
- Não apresentar comprometimento cognitivo ou neurológico que impedisse a gravação da voz;
- Não ter realizado terapia vocal ou tratamento cirúrgico na laringe previamente;
- Pacientes que procuraram o serviço por demanda espontânea ou encaminhados pelo otorrinolaringologista e foram avaliados antes de realização da terapia vocal.

Desse modo, todos os indivíduos participantes apresentavam queixa vocal e receberam confirmação do diagnóstico do distúrbio de voz por meio do exame visual laríngeo. Como critério de exclusão, foram excluídos da pesquisa indivíduos com uso profissional da voz, fosse cantada ou falada, uma vez que o treinamento vocal pode criar ajustes glóticos e supraglóticos que modificam o sinal acústico, mesmo na presença de lesão laríngea ou desvio da qualidade vocal e tempo de gravação menor que 6 segundos.

4.4 PROCEDIMENTOS

Os procedimentos de coleta de dados para amostras de voz foram realizados de acordo com o protocolo de avaliação de voz de rotina para avaliações iniciais de pacientes no laboratório. Todos os dados foram registrados digitalmente (sinais de voz) ou em forma escrita no relatório médico (relatório de exame de anamnse e laríngeo) no banco de dados laboratorial. Os dados relacionados ao julgamento auditivo-perceptivo também foram acessados nos registros do banco de dados.

As amostras de voz investigadas foram coletadas durante a primeira sessão de avaliação de voz clínica antes do início da terapia de voz. Todos os sujeitos completaram uma breve

entrevista de histórico médico e foram encaminhados para exame visual laríngeo por um otorrinolaringologista, que deverá apresentar um relatório escrito no prazo de 15 dias após a sessão para avaliação de voz clínica. Os indivíduos submetidos ao exame laríngeo visual no departamento no prazo de 30 dias após a sessão para avaliação de voz clínica foram isentos de encaminhamento para exame mais aprofundado e seus relatórios internos com o diagnóstico laríngeo foram utilizados.

Primeiro, os sujeitos preencheram uma breve anamnese para coleta de dados pessoais e uma descrição subjetiva de suas queixas e sintomas de voz usando um questionário desenhado pelo nosso laboratório. O software *Fonoview* versão 4.5, desktop Dell *all-in-one*, microfone cardioide unidirecional, da marca Senheiser, modelo E-835, localizado em um pedestal e acoplado a um pré-amplificador Behringer, modelo U-Phoria UMC 204, foram usados para gravação de voz.

As coletas de vozes ocorreram em uma cabine de gravação com tratamento acústico e ruído inferior a 50 dB NPS, com taxa de amostragem de 44000 Hz, janelamento de 40 ms, tempo de atualização de 2,5 ms, faixa dinâmica de amplitude de 60 dB, limite de frequência de 7500 Hz e intervalo de tempo mínimo de 3 segundos (DELIYSKI, 2005). Durante a gravação de voz, todos os sujeitos estavam de pé, e o microfone foi colocado a uma distância padrão de 10 cm e em um ângulo de 45º para a boca do alto-falante (HECKMAN; PINTO; SAVELYEV, 1967). A distância entre a boca do alto-falante e o microfone foi medida usando uma régua de 10 cm antes de gravar cada tarefa. Os sujeitos foram posicionados no limite traseiro da cabine acústica, e o pedestal foi fixado na frente deles para manter uma distância constante entre a boca do orador e o microfone durante a gravação de voz. Assim, o movimento do alto-falante durante a gravação foi impedido. Foi instruído que os sujeitos não devem levantar a cabeça durante a gravação. A vogal sustentada “é” em *pitch* e *loudness* no padrão habitual de fala foi utilizada como amostra. Nesse estudo, além de ser a mais comumente utilizada na avaliação vocal no Brasil, optou-se a utilização da vogal “é” por ter características como: é uma vogal oral aberta, não arredondada e que possui a posição mais neutra e intermediária no trato vocal para o Português Brasileiro (PONTES *et al.*, 2009).

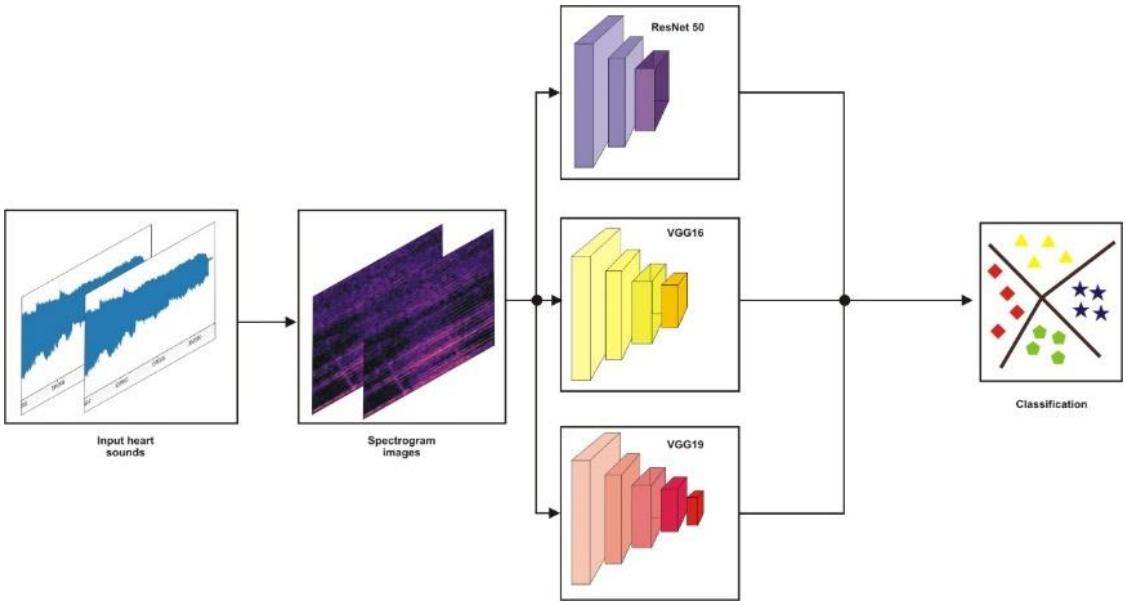
Na sequência, foram gerados os espectrogramas e traçado espectrográfico referente ao sinal foi salvo em formato .png, para posterior análise. Cinco juízes, acadêmicos do curso de Fonoaudiologia, foram treinados a realizar inspeção visual do traçado espectrográfico e classificá-lo em Tipo I, II, III ou IV, conforme as recomendações de Titze (1995) e Sprecher et al. (2010). O treinamento foi realizado por um fonoaudiólogo especialista em voz, com mais de

15 anos de experiência em avaliação vocal e inspeção visual de traçado espectrográfico de banda estreita. Para o treinamento, foram utilizados como estímulos-âncora quatro espectrogramas, correspondendo aos tipos de sinais estudados (SPRECHER et al., 2010) (Figura 3). Na sequência, os acadêmicos foram treinados a reconhecer a tipologia dos sinais em 120 espectrogramas, correspondentes a sinais vocais de pacientes com diferentes distúrbios de voz e indivíduos vocalmente saudáveis, previamente selecionados no banco de dados e utilizados em estudo anterior (LOPES et al., 2017).

Após o treino, os cinco acadêmicos receberam os 923 arquivos dos espectrogramas de faixa estreita em formato .png, juntamente com os quatro estímulos-âncora de cada sinal, citados anteriormente. Eles foram orientados a fazer a inspeção visual do traçado e classificar o sinal de acordo com a tipologia e os critérios propostos por Titze (1995) e Sprecher et al. (2010). Assim, os sinais deveriam ser classificados Tipos I, II, III e IV, como descritos na seção análise espectrográfica da voz.

Feito isto, a partir do conjunto de dados de vozes, foram utilizados 923 espectrogramas, gerados a partir do conjunto de dados para treinar e avaliar o modelo de CNN proposto, como será descrito no próximo tópico. Uma visão geral dos componentes utilizados para construir o modelo de classificação proposto nesse estudo, é ilustrado na Figura 19.

Figura 19 -Componentes utilizados no sistema de classificação



Fonte: Pesquisador do estudo (2022)

A arquitetura do modelo é composta por quatro componentes principais, as vozes coletadas, a construção dos espectrogramas, os modelos de classificação e os resultados.

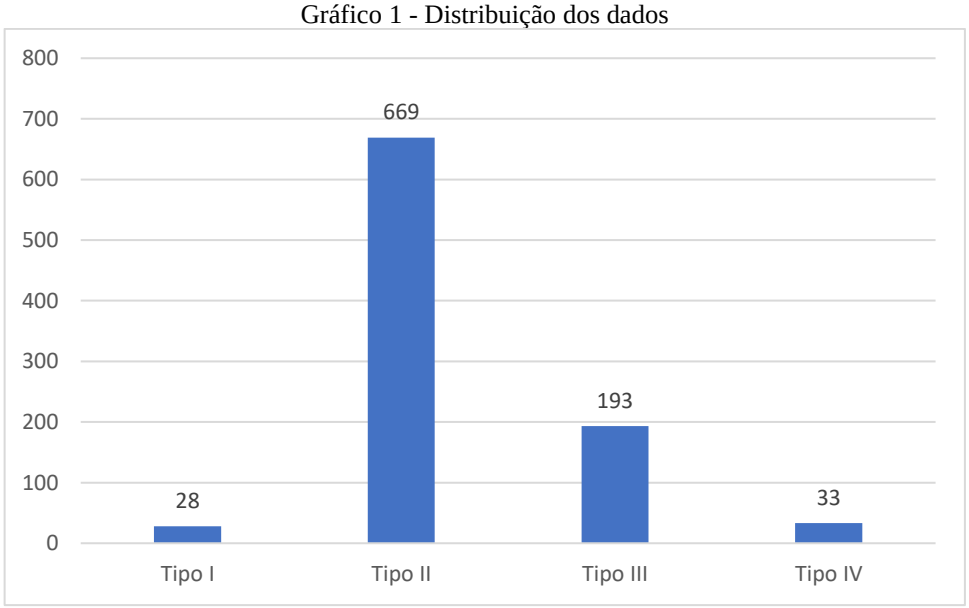
4.5 CONJUNTO DE DADOS DE VOZES

Para treinar os modelos, utilizou-se os dados de vozes coletados pelo LEIV que recebe pacientes para diagnóstico e terapia vocal. Os dados sobre sexo, idade, queixas vocais e o diagnóstico laríngeo dos sujeitos foram inseridos em um banco de dados. Os procedimentos de coleta de dados para as amostras de voz foram conduzidos de acordo com o protocolo de avaliação da voz de rotina para as avaliações iniciais dos indivíduos no laboratório. Os dados foram coletados digitalmente (sinais de voz) e por escrito em prontuário (anamnese e laudo do exame laríngeo) no banco de dados do laboratório.

Muitas tarefas de classificação usando aprendizado profundo melhoraram a precisão da classificação usando uma grande quantidade de dados de treinamento. No entanto, é difícil coletar dados de áudio e construir um grande banco de dados balanceados. Nesse sentido faz-se necessário utilizar técnicas para minimizar esse problema.

Nesse cenário, o conjunto de dados de vozes coletados e utilizados nesse estudo, inicialmente era composto por 1.700 arquivos de vozes, no entanto foram selecionados apenas os arquivos que atenderam aos critérios de elegibilidade e arquivos que não continham erros de leitura. Feito isso, a dimensão dos dados foi reduzida para 923 arquivos de vozes de indivíduos com e sem desvio, classificados com o tipo de voz sem alteração (Tipo I – 28 amostras), com desvio vocal leve (Tipo II - 669 amostras), moderada (Tipo III - 193 amostras) e intenso (Tipo IV – 33 amostras). O gráfico 1 mostra a distribuição dos dados.

Gráfico 1 - Distribuição dos dados



Fonte: Pesquisador do estudo (2022)

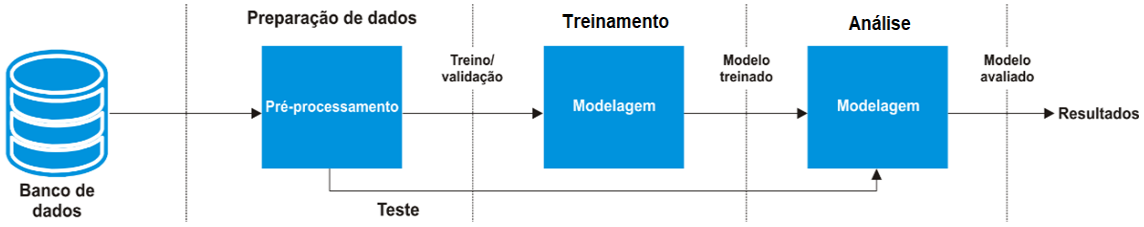
Devido ao pequeno número de amostra das classes I e IV para gerar as imagens espectrográficas para o treinamento, validação e teste, bem como pelo desbalanceamento dos dados, houve a necessidade de expandir o número de imagens durante a geração de cada espectrograma e, assim não prejudicar a precisão do modelo e reduzir o *overfitting* (GAO *et al.*, 2020; RAČIĆ *et al.*, 2021).

Os impactos dos dados altamente desbalanceados são implícitos, ou seja, não geram um erro imediato quando construído e executado, mas os resultados podem ser ilusórios (KRISHNENDU, 2020; KUHN; JOHNSON, 2020), podem apresentar dificuldade na diferenciação entre as classes. Se o grau de desequilíbrio de classe para a classe majoritária for acentuado, a tendência é produzir modelos (ou regras) de classificação que favorecem a classe com maior probabilidade de ocorrência, ou seja, a classe majoritária, podendo resultar em um baixo índice de reconhecimento para o grupo minoritário (GAO *et al.*, 2020; KRISHNENDU, 2020), uma vez que o modelo tem maior probabilidade de prever a maioria das amostras pertencentes à classe majoritária (GAO *et al.*, 2020). Portanto, além da expandir o número de imagens, também se fez uso da validação cruzada estratificada com 8 *k-folds* para minimizar esse problema. Em seguida, serão descritas as técnicas utilizadas.

4.6 FASES DO EXPERIMENTO

Este experimento inclui três diferentes fases: preparação dos dados, treinamento, evolução e avaliação de desempenho. As três diferentes fases executadas durante este trabalho, são mostrados na Figura 20 e descritas na sequência. A metodologia proposta é baseada em um esquema em cascata em que, primeiro foram classificadas as vozes com e sem desvio vocal e, em seguida, foi identificada a tipologia do sinal da voz.

Figura 20 - Fases executadas durante este trabalho



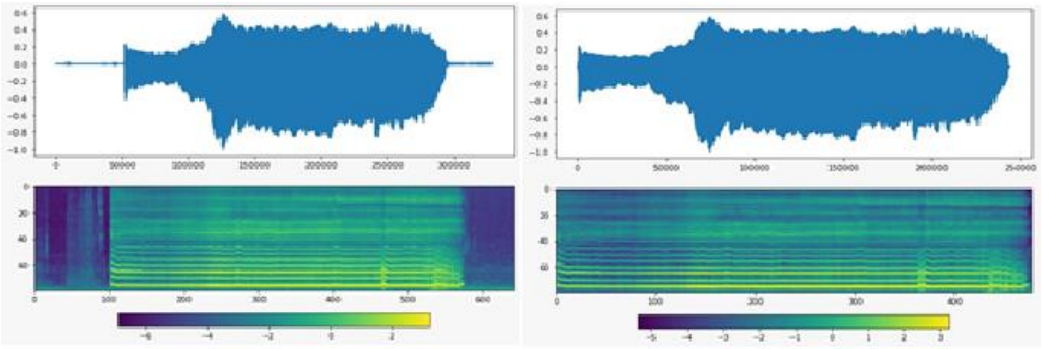
Fonte: Pesquisador do estudo (2022)

4.7 PREPARAÇÃO DOS DADOS

O conjunto de dados foi composto por 923 amostras de vozes classificada com a

tipologia do sinal da voz. Antes de converter os sinais de áudio em espectrogramas, foram removidos os espaços em branco ou silenciosos ou partes da gravação com amplitude máxima de 20db, conforme ilustrado na Figura 21.

Figura 221 – (A) Sinal da voz com espaço em brando ou silencioso. (B) removido espaço em branco ou silencioso



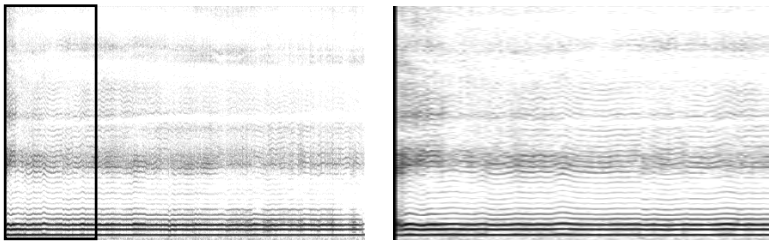
Fonte: Pesquisador do estudo (2022)

Após esse procedimento, para converter os sinais de áudio em imagens espectrográficas, utilizou-se a transformada de Fourier de curta duração (*Short-Time Fourier Transform - STFT*), com taxa de amostragem de 44.000 Hz, janelamento de 40ms, tempo de atualização de 2,5ms, faixa dinâmica de amplitude de 50 dB, limite de frequência de 5000 Hz (LOPES; ALVES; MELO, 2017). As imagens foram geradas com resolução de 500x300 *pixels*.

No pré-processamento dos dados, durante a geração de cada imagem espectrográfica, houve a necessidade de aumentar artificialmente a quantidade das imagens para minimizar o desbalanceamento dos dados e, assim, não prejudicar a previsão do modelo. Aumentar o número das imagens é o processo utilizado para fazer pequenas alterações nas imagens originais para ampliar sua diversidade sem coletar novos dados. É uma técnica utilizada para expandir um conjunto de dados de imagens de forma artificial, quando não há mais dados disponíveis de determinado tipo ou condições de fazer novas coletas (JAITLY; HINTON, 2010; KRISHNENDU, 2020; RAČIĆ *et al.*, 2021).

As técnicas para aumentar artificialmente os dados, podem incluir inversão horizontal e vertical, rotação, recortes, cisalhamento, conversão de imagens em tons de cinza, alteração nos padrões de cores etc. (KRISHNENDU, 2020; PIELAWSKI; WÄHLBY, 2020). Neste trabalho, optou-se por fazer recorte nos espectrogramas para aumentar número de amostra e treinamento os modelos. Na Figura 22, pode-se observar um espectrograma de voz do tipo III, onde a imagem A é um espectrograma original e a B é recorte da imagem A:

Figura 22 – (A) Espectrograma Original, (B) Recorte do espectrograma original



Fonte: Pesquisador do estudo (2022)

A técnica para aumentar artificialmente a quantidade de imagens é uma ótima maneira de expandir o tamanho ou dimensão do seu conjunto de dados. Pode-se criar imagens transformadas de seu conjunto de dados originais. Esse tipo de processamento de imagens, tende a melhorar o desempenho e a previsão do modelo quando não há dados suficientes (GAO *et al.*, 2020; KRISHNENDU, 2020; RAČIĆ *et al.*, 2021).

Dito isto, antes do treinamento dos modelos, foram feitos recortes das imagens originais para criar um subconjunto aleatório das imagens originais. Portanto ampliando o conjunto de dados de imagens de 923 para 1.697, conforme Tabela 1.

Tabela 1 - Conjunto de dados original e ampliado

	Base original	**Base ampliada**
Tipo I	28	196
Tipo II	669	669
Tipo III	193	579
Tipo IV	33	231
Total	923	1.675

Fonte: Pesquisador do estudo(2022)

Para as imagens dos tipos I e IV, foram realizados 6 recortes e para o tipo III, 2 recortes de cada espectrograma. Ressalta-se que nessa primeira parte de aumento artificial da base de dados de imagens, não houve a necessidade de ampliar os espectrogramas do Tipo II, pois poderia desbalancear ainda mais a base. Por conseguinte, na segunda parte de aumento artificial dos dados, para expandir ainda mais a quantidade de imagens em tempo real, ou seja, enquanto o modelo está no processo de treinamento, utiliza-se o gerador automático de dados de imagem da biblioteca *keras* (*ImageDataGenerato*r) com as seguintes configurações: *brightness_range* = 0.2 e *zoom_range*=0.10. Para padronizar todo o conjunto de dados de imagens na entrada dos modelos, as imagens foram dimensionadas para 224x224.

Por fim, todas as imagens foram separadas e incluídas em pastas em uma das classes possíveis (Tipo I, Tipo II, Tipo II e Tipo IV) e, por último, os dados foram divididos em conjuntos de treino, validação e teste.

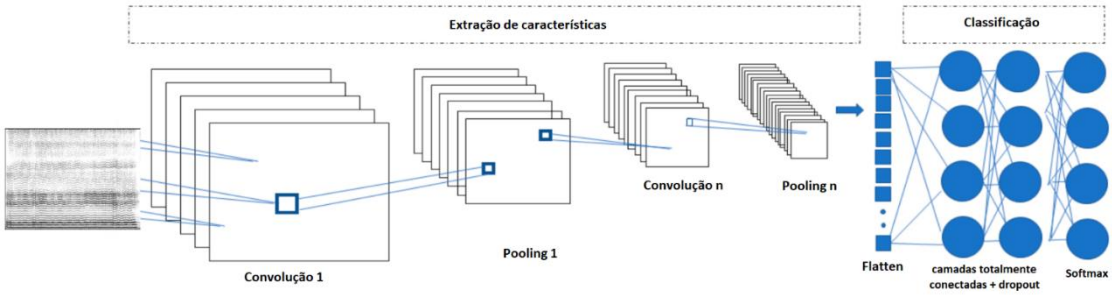
4.8 TREINAMENTO

4.8.1 Modelo proposto

Para treinar e analisar o modelo proposto, o conjunto de dados foi dividido aleatoriamente da seguinte forma: dados de treinamento, 70%; dados de validação, 20%; e os dados de teste, 10%. Para obter uma divisão aleatória do conjunto de dados e tentar manter a distribuição de classes em cada subconjunto, foi utilizada a validação cruzada estratificada com 10 *k-folds* para treinar e avaliar os modelos. O conjunto de teste foi utilizado para avaliar o desempenho final do modelo e validação para comparar os diferentes modelos.

A configuração do VGG-16, assim como dos modelos que foram comparados, analisados e testados, foi composto pela configuração base de cada CNN pré-treinada, utilizando na camada de entrada o tamanho de 224 × 224 *pixels*. A Figura 23 destaca a arquitetura que consiste em uma camada de entrada seguida por um conjunto de camadas padrão de cada modelo, ponderadas acomodando uma união de camadas convolucionais com não linearidade de retificação aplicada e camada *max-pooling*.

Figura 23 - A configuração de rede do modelo VGG-16 para tipologia.



Fonte: Pesquisador do estudo (2022)

Nas camadas convolucionais, filtros de tamanho 2 × 2, foram aplicados à imagem. Todas as camadas de *pooling* consiste basicamente em reduzir o tamanho dos dados de entrada. Normalmente, após uma camada convolucional utiliza-se uma camada de *pooling*, com isso as próximas camadas de convolução receberão uma outra forma de representação dos dados, possibilitando a rede aprender diversas representações, evitando com isso o *overfitting*.

As camadas totalmente conectadas são o conjunto final de camadas usadas para processar os resultados antes da classificação das imagens. Finalmente, a camada final dessa arquitetura é a camada *softmax*, camada responsável por uma função de classificação

multiclasse.

Após algumas tentativas de experimentação, determinou-se os hiperparâmetros mais adequados, conforme resumo mostrado na Tabela 2, com os quais se obteve os melhores resultados para os modelos analisados. O hiperparâmetros foram: Adam (KINGMA; BA, 2014) como *optimizer*, com a taxa de aprendizado de 0,001, tamanho do lote (*Batch Size*) de 32 e 20 épocas.

Durante a transferência de aprendizagem, utiliza-se a CNN base, removendo as camadas superiores (camadas densas), e congelando as outras camadas, o que significa que seus pesos permanecerão fixos durante o processo de treinamento. Foram adicionadas duas camadas densa com 4096 neurônios, duas camadas de *dropout*, para evitar *overfitting*, com valor de 0,5. Por fim, uma camada *softmax*. Esses hiperparâmetros foram fixos para todos os modelos avaliados neste estudo.

Tabela 2 - Valores de hiperparâmetros usados nos experimentos.

Hiperparâmetros	**Valores**
Batch Size	32 Amostras
Taxa de aprendizado	0,001
Treinamento, validação e teste	70%, 20%, 10%
Optimizer	Adam
Tamanho de entrada	224x224 *pixels*
Dropout	0.5
Camada densa	4096

Fonte: Pesquisador do estudo (2022)

Dentre os hiperparâmetros apresentados na Tabela 2, a taxa de aprendizado foi definida com o valor padrão de 0,001. O *ReduceLROnPlateau* (CHOLLET, 2016) foi usado para reduzir a taxa de aprendizado quando uma métrica acurácia de validação parar de melhorar ou atingir um platô durante o treinamento. Sendo assim, foram utilizados os seguintes parâmetros para o *ReduceLROnPlateau*:

Tabela 3 - Parâmetros do *ReduceLROnPlateau*

Parâmetros	Valores
Factor	0.4
Patience	3
Verbose	1
min_lr (limite mínimo)	0.00003
Monitor	val_accuracy

Fonte: Pesquisador do estudo (2022)

Argumentos:

- Factor: fator pelo qual a taxa de aprendizagem será reduzida. *new_lr = lr * factor*.
- Patience: número máximo e épocas sem melhoria na métrica monitorada
- Verbose: Ativar ou desativar mensagens
- Min_lr: limite mínimo da taxa de aprendizagem
- Monitor: seleciona a métrica que será monitorada;

5 RESULTADOS

Todas as métricas obtidas de todos os modelos treinados foram compiladas na Tabela 4:

Tabela 4 - Resultados obtidos.

Modelo	Acurácia	Kappa	Sensibilidade	Especificidade	f1-score	*Precision*
VGG-16	**0.94**	**0.91**	**0.94**	**0.98**	**0.94**	**0.94**
VGG-19	0.88	0.83	0.88	0.96	0.88	0.89
ResNet50	0.77	0.66	0.77	0.92	0.75	0.80

Fonte: Pesquisador do estudo (2022)

O modelo VGG-16, obteve os melhores resultados em todas as medidas de desempenho. O VGG-16 obteve uma acurácia de 0.94 e *kappa* de 0.91, ou seja, com grau de concordância quase perfeita. Para f1-score e *precision*, obteve-se 0.94 e 0.94, respectivamente. Esses resultados demonstram quão preciso o modelo foi em relação aos casos positivos previstos, quantos deles são realmente positivos. Para sensibilidade e especificidade, o VGG-16 obteve

0.94 e 0.98, respectivamente. Por outro lado, o VGG-19 vem logo em seguida obtendo os seguintes resultados, acurácia de 0.88, *kappa* 0.82, sensibilidade 0.88, especificidade 0.96, f1-score 0.88 e *precision* 0.90.

Por fim, a CNN ResNet50 obteve 0.77, 0.66, 0.92, 0.92, 0.85 e 0.84 para acurácia, *kappa*, sensibilidade, especificidade, f1-score e *precision*, respectivamente.

Vale ressaltar que os quatro modelos obtiveram bons resultados em todas as métricas de desempenho, ou seja, os modelos demonstraram sua eficiência na classificação dos espectrogramas. Nesse cenário, para construir o sistema de classificação, utiliza o modelo com melhor resultado, portanto, o VGG-16. Sendo assim, o modelo desenvolvido utilizou a CNN VGG-16 como modelo base para classificar os espectrogramas. A tabelas 5, apresenta a matriz de confusão de cada CNN.

Tabela 5 – Matriz de confusão do modelo VGG-16

VGG-16	*Verdadeito*	Tipo I	19	2	0	0
		Tipo II	0	63	1	0
		Tipo III	0	7	48	0
		Tipo IV	0	0	0	23
VGG-19	*Verdadeito*	Tipo I	18	1	1	1
		Tipo II	0	61	3	0
		Tipo III	0	11	45	0
		Tipo IV	1	0	1	20
ResNet 50	*Verdadeito*	Tipo I	19	1	0	1
		Tipo II	0	62	2	0
		Tipo III	0	29	27	0
		Tipo IV	3	0	2	17
		Classe	**Tipo I**	**Tipo II**	**Tipo III**	**Tipo IV**
			Previsto			

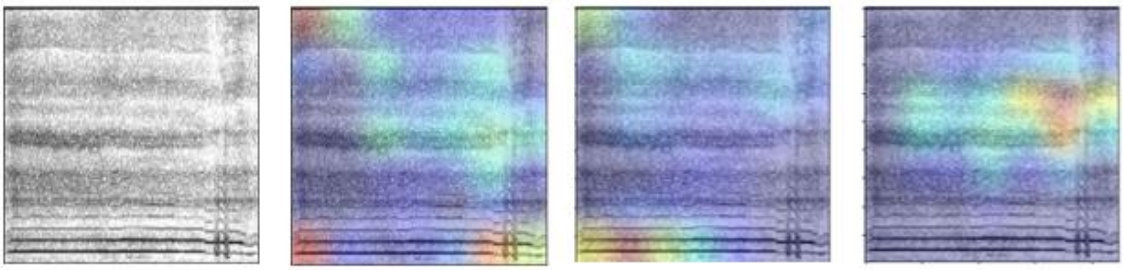
Fonte: Pesquisador do estudo (2022)

Para construir sistemas inteligentes e avançar para sua integração significativa em nossas vidas cotidianas, deve construir modelos "transparentes" que tenham a capacidade de explicar como foi feito o processo de previsão (SELVARAJU *et al.*, 2016). O objetivo da transparência e das explicações é identificar os modos de falha, ajudando o pesquisador a concentrar seus esforços em soluções mais relevantes.

Sendo assim, para tornar mais evidente o processo de tomada de decisão do sistema construído e para demonstrar como ele chegou à decisão, bem como sua eficiência, aplica o Grad-CAM (SELVARAJU *et al.*, 2016). ao classificador usando algumas amostras da base de

dados de imagens. A Figuras 24 demonstra como foi a decisão na classificação do espectrograma do tipo 3 para os modelos VGG-16, VGG-19 e o ResNet50, respectivamente.

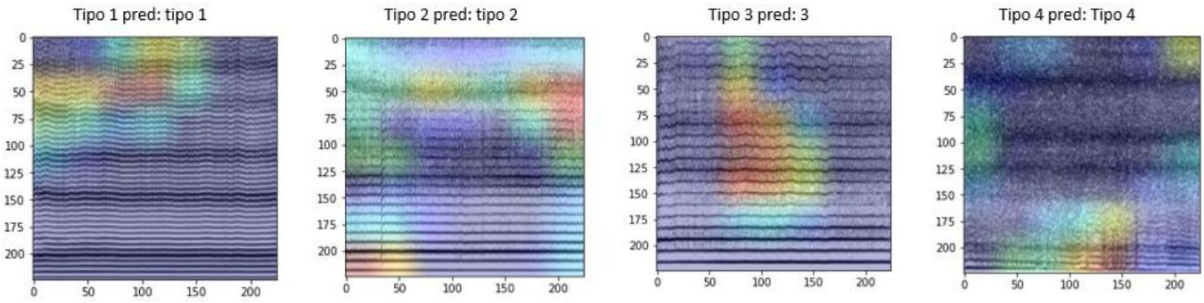
Figura 24 - Decisão na classificação do espectrograma do tipo 3



Fonte: Pesquisador do estudo (2022)

Observando tanto o espectrograma quanto seu mapa de ativação correspondente, na figura 25 pode-se observar que o VGG-16 pode capturar as características em bandas de alta frequência para fala normal (Tipo I) e pode dar pesos maiores para as características em bandas de baixa frequência para vozes alteradas (Tipos II, III e IV). Sendo assim, o VGG-16 possui excelente capacidade de aprendizado para extrair características e alcança melhores desempenhos na detecção da tipologia da voz.

Figura 25 - VGG-16



Fonte: Pesquisador do estudo (2022)

Por fim, as técnicas utilizadas nesse trabalho, foram úteis quando as classes estão desbalanceadas (TSIHRINTZIS; SOTIROPOULOS; JAIN, 2019). Os bons resultados de sensibilidade e especificidade evidenciaram a confiabilidade dos resultados para procedimentos de triagem ou confirmação do diagnóstico. Nesse contexto, sensibilidade e a especificidade são altamente influenciadas pela heterogeneidade do espectro ou condição patológica investigada. O VGG-16 é uma arquitetura de CNN, pré-treinada, simples e amplamente usada para classificação de imagens.

6 DISCUSSÕES

Na prática clínica tradicional, são necessários diversos exames médicos para detectar distúrbios de voz. Esses exames são muitas vezes invasivos e demorados, causando um desconforto ao paciente. A avaliação clínica da disfonia muitas vezes depende de uma combinação de técnicas perceptivo-auditivas e de medida acústica. A análise acústica, por ser uma técnica não invasiva, é um dos métodos mais abordados e utilizados no contexto clínico e de pesquisa para avaliação da emissão vocal. O espectrograma é uma das principais ferramentas utilizadas na análise acústica.

Nesse contexto, a classificação da tipologia do sinal da voz pode ser uma estratégia importante para monitorar a qualidade vocal e avaliar o resultado da intervenção clínica (terapêutica ou medicamentosa) ou cirúrgica em pacientes com distúrbios de voz. Essa tarefa, conhecida como tipologia de sinal, tornou-se uma etapa relevante de pré-processamento para determinar a adequação de uma análise de sinal à perturbação, que depende muito da suposição de que o sinal avaliado seja quase periódico.

Contudo, necessita de uma avaliação criteriosa com base no julgamento visual feito por fonoaudiólogos experientes. A classificação automática da tipologia da voz utilizando imagens espectrográficas, pode ser uma excelente ferramenta para ser utilizada na etapa do processamento do sinal, antes do processo de extração de medidas.

Observando essa problemática e para potencializar a eficiência nas análises espectrográficas, foi desenvolvido um modelo de DL utilizando o VGG-16 como base para classificar automaticamente a tipologia da voz utilizando imagens espectrográficas. Os resultados obtidos neste estudo, mostram que o modelo desenvolvido obteve bons resultados na classificação. Em alguns experimentos analisados neste trabalho, foi possível observar que os resultados obtidos são comparáveis aos apresentados em pesquisas existentes utilizando diferentes propostas de DNN, no entanto foi não encontrado estudos que utilizassem CNNs pré-treinados para classificar espectrogramas da tipologia do sinal da voz da vogal sustentada /e/.

Nesse cenário, Vavrek *et al.* (2021) utilizaram um ensamble de VGG-16, na sua configuração padrão, para classificar espectrogramas de 1012 indivíduos patológicos e indivíduos saudáveis. Os espectrogramas foram gerados utilizando o STFT e as gravações das vozes eram da vogal sustenta /a/. Nesse estudo, foram testados várias de configurações de hiperparâmetros, bem como o método Ensemble. Os melhores resultados obtidos foram para o *Ensemble* com acurácia, sensibilidade e especificidade de 0.82, 0.84, 0.79, respectivamente.

Apesar dos bons resultados obtidos pelos autores, utilizando uma metodologia de

divisão dos dados similares são utilizadas no presente estudo, divisão dos dados em treinamento, validação e teste, bem como a estratificação dos dados, nossos resultados foram bem melhores com a arquitetura proposta. Vavrek *et al.* (2021) não utilizou o GRAD-CAM para analisar os resultados da sua CNN.

Em Mohammed *et al.* (2020), utilizaram o DL ResNet34 para classificar pacientes com voz normal ou patológica. Nesse estudo, ele optou em utilizar apena 1 segundo do sinal da voz da vogal. Os espectrogramas foram gerados utilizando o STFT e as gravações das vozes eram da vogal sustenta /a/. Os autores obtiveram uma precisão de 0.95, bem como 0.94 e 0.96 para F1-Score e Sensibilidade, o banco de dados utilizado foi *The Saarbrucken Voice Database* (SVD) do com aproximadamente 2000 vozes. Para divisão dos dados, os autores utilizaram 80% dos dados para treinamento e 20% para validação, também utilizaram a validação cruzada com 8 *fold* para tentar manter a distribuição de classes em cada subconjunto.

No estudo de Mohammed *et al.* (2020), os autores ressaltam a necessidade de investigar mais modelos DL e destaca que os métodos de aprendizado de máquina mais utilizados para detecção de patologias vocais, são Rede Neural Artificial (ANN), *Random Forest* (RF) *e Support Vector Machine* (SVM). Corroborando com a importância do presente experimento.

Em outra pesquisa, Trinh e Darragh (2019), construíram e testaram uma CNN, utilizando a biblioteca *Keras*, para classificar 103 espectrogramas da voz de pacientes patológicos e saudáveis. Os espectrogramas foram gerados utilizando o STFT e as gravações das vozes eram da vogal sustenta /a/. Obtiveram uma precisão de classificação de 0.97. O resultado de acuraria ficou compatível com o nosso, porém, eles não avaliaram a capacidade do modelo em identificar corretamente as vozes disfonias entre aqueles que a possuem, sensibilidade, nem a capacidade do teste em excluir corretamente aqueles que não possuem vozes disfonias, especificidade. As imagens geradas nesse estudo, foram extraídas do banco de dados SVD.

Quando se compara os resultados obtidos com os de Vavrek *et al.* (2021), Mohammed *et al.* (2020) e Trinh e Darragh (2019), pode-se observar que a metodologia utilizada para classificar a tipologia do sinal através de imagens espectrográficas, demonstrou ser bastante eficiente, pois os resultados foram semelhantes ou melhores a eles, ou seja, uma acurácia excelente. Em nenhum dos estudos listados foi utilizado o GRAD-CAM para verificar qual parte do espectrograma o modelo utilizou para classificar. O quadro 2 mostra uma síntese dos 3 estudos citados e comparados.

Quadro 2: Síntese dos estudos comparados

Autores	**Metodologia**	**Modelo(s)**	**Resultados**
Vavrek *et al.* (2021)	**Vogal sustentada:** /a/ STFT **Divisão dos dados:** Treinamento: 60% Validação: 20% Teste: 20% Com estratificação dos dados	VGG16 E*nsemble* *Dropout* = 0,5 3 camadas densas com 128 neurônios	**Disfonia × saudável:** Acurácia: 0.82 Sensibilidade: 0.84 especificidade: 0.79
Mohammed *et al.* (2020)	**1 segundo da vogal** **Vogal sustentada:** /a/ STFT **Divisão dos dados:** Treinamento: 80% Validação: 20% Teste: 20% Validação cruzada com 10 *fold*	ResNet34 *Dropout*: 0.3	**Disfonia × saudável:** Acurácia: 0.95 F1-score: 0.94 Sensibilidade: 0.96
Trinh e Darragh (2019)	**Vogal sustentada:** /a/ STFT **Divisão dos dados:** Treinamento: 80% Validação e teste: 20% Com estratificação dos dados		**Disfonia × saudável:** Acurácia: 0.97 F1-score: 0.94 Sensibilidade: 0.96

Fonte: Pesquisador do estudo (2022)

Sendo assim, os resultados do modelo desenvolvido podem auxiliar em futuras pesquisas na construção de ferramentas que utilizem uma abordagem de Aprendizado Profundo. A construção dessa ferramenta poderá auxiliar profissional da voz nos procedimentos de avaliação e monitoramento dos desvios da voz, assim como contribuir para na produção do conhecimento e treinamento de novos profissionais, sejam eles acadêmicos ou profissionais.

Por fim, o modelo desenvolvido pode ser utilizado no processo de avaliação vocal, classificando espectrogramas da tipologia do sinal da voz da vogal sustentada /e/, de forma automática potencializando a análise espectrográfica, dando suporte a decisão ao fonoaudiólogo.

7 CONCLUSÕES

Vários estudos, modelos e técnicas de aprendizado de máquina e aprendizado profundo foram desenvolvidos ou estão em desenvolvimento, com o objetivo de auxiliar o fonoaudiólogo na detecção e classificação de distúrbios vocais. A construção de sistemas automáticos, não invasivo, de baixo custo e inteligentes, poderiam desempenhar um papel decisivo no diagnóstico precoce de patologias na voz e no processo de tomada de decisão clínica, possibilitando iniciar o tratamento dando mais esperança de cura ao paciente. A maioria desses estudos concentram-se principalmente nas vogais sustentada /a/, /i/, e /u/, bem como nas bases de dados SVD e MEEI visando alcançar alta precisão.

No entanto, isso pode dificultar a generalização para um cenário do mundo real quando dispor de dados, normalmente, desbalanceados ou insuficientes para treinar o modelo. Diante disto, é necessário utilizar técnicas para mitigar esses e outros problemas. Da mesma forma, utilizar técnicas possa demonstrar e entender como o modelo tomou a decisão para classificar o espectrograma.

Nesse cenário, este estudo analisou e comparou os resultados do modelo desenvolvido com base no DNN VGG-16 com os modelos VGG-19 e ResNet50 para classificar a tipologia do sinal da voz proposta por Titze (1995) e Sprecher *et al.* (2010). Para gerar os espectrogramas, utilizou-SE a base de dados de vozes proveniente do Laboratório Integrado de Estudos da Voz (LIEV).

O modelo desenvolvido obteve 0.98 em acuraria e kappa de 0.92, ou seja, com grau de concordância quase perfeita. Para f1-score e *precision*, obteve-se 0.95 e 0.96 respectivamente. Esses resultados demonstram quão preciso o modelo foi em relação aos casos positivos previstos, quantos deles são realmente positivos. Para sensibilidade e especificidade, o modelo obteve 0.95 e 0.98, respectivamente. Assim, nossos resultados mostraram que a DNN VGG-16 é um modelo eficiente e pode ser utilizado para construir sistemas inteligentes para classificação de imagens espectrográficas da tipologia da voz.

Da mesma forma, o GRAD-CAM demostrou ser uma técnica que pode ser utilizada para tornar transparente a decisão que a DNN tomou para classificar o espectrograma, destacando as partes da imagem mais discriminativas como regiões mais brilhantes no mapa de calor.

A metodologia utilizada no presente estudo mostrou-se eficiente para lidar com a base de dados de imagens desbalanceados e insuficientes para treinar um DNN. Sendo assim, o modelo desenvolvido, pode ser utilizando na construção de sistemas inteligentes que poderão ser utilizados no processo de triagem dando mais rapidez e automatizando a classificação de

espectrogramas da tipologia, indicando quais medidas poderão ser utilizadas nos procedimentos de avaliação e monitoramento dos pacientes, tanto off-line (gravação seguido de análise) quanto on-line (gravação e análise em tempo real) utilizando de tecnologias modernas baseadas em nuvem e 5G.

Como limitação encontrada nesse estudo, pode citar o desbalanceamento dos dados, recursos computacionais e a quantidade limitada de imagens originais disponíveis para o treinamento dos modelos.

Como contribuição apresentada neste estudo, contribuir para potencializar a eficiência nas análises espectrográficas, visto que, para fazer essa classificação, é necessária uma avaliação criteriosa com base no julgamento visual feito por fonoaudiólogos experientes, e por consequência contribuir na produção do conhecimento e treinamento de fonoaudiólogos, sejam eles acadêmicos ou profissionais e evidenciar que, apesar da baixa quantidade dos dados para algumas classes e o do grau de desbalanceamento dos dados, a engenharia utilizada na metodologia de construção do modelo para classificar a tipologia da voz, foi eficiente.

Como trabalhos futuros, adicionar uma funcionalidade que possa indicar quais medidas acústicas são as mais indicadas para o tratamento do paciente após a classificação da tipologia da voz, assim como aumentar a base de dados para avaliar e comparar outros modelos de *Deep* CNN e testar outras técnicas para dar mais eficiência ao modelo.

REFERÊNCIAS

AGGARWAL, Charu C. *et al.* Neural networks and *Deep Learning*. **Springer**, v. 10, p. 978-3, 2018.

AKBULUT, S. *et al.* Basics of Voice Disorders. In: **Phoniatrics I**. Springer, Berlin, Heidelberg, 2020. p. 193-238.

ALASKAR, H. *Deep Learning*-based model architecture for time-frequency images analysis. **International Journal of Advanced Computer Science and Applications**, v. 9, n. 12, 2018.

ALBON, C. **Machine learning with python cookbook: Practical solutions from preprocessing to *Deep Learning***. O'Reilly Media, Inc., 2018.

ALHUSSEIN, M; MUHAMMAD, G. Voice pathology detection using *Deep Learning* on mobile healthcare framework. **IEEE Access**, v. 6, p. 41034-41041, 2018.

AMARA, F.; FEZARI, M.; BOUROUBA, H. An improved GMM-SVM system based on distance metric for voice pathology detection. **Appl. Math**, v. 10, n. 3, p. 1061-1070, 2016.

ANISHA, C. D.; ARULANAND, N. Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers. *In*: **2020 International Conference on Innovative Trends in Information Technology (ICITIIT)**. IEEE, 2020. p. 1-6.

ARAVINDA, C. V. *et al.* A demystifying convolutional neural networks using Grad-CAM for prediction of coronavirus disease (COVID-19) on X-ray images. *In:* **Data Science for COVID-19**. Academic Press, 2021. p. 429-450.

ARIAS-VERGARA, T. *et al.* Multi-channel spectrograms for speech processing applications using *Deep Learning* methods. **Pattern Analysis and Applications**, v. 24, n. 2, p. 423-431, 2021.

BEHLAU, M.; MURRY, T. International and intercultural aspects of voice and voice disorders. **Battle DE. Communication disorders in multicultural and international populations. 4th ed. St. Louis: Elsevier/Mosby**, p. 174-207, 2012.

BINDER, J. R. Phoneme perception. **Neurobiology of language**, Academic Press, 2016. p. 447-461.

BIREME. Descritores em Ciências da Saúde: DeCS. **Descritores em Ciências da Saúde:** DeCS, 2017.

BOERSMA, P.; WEENINK, D. Praat: doing phonetics by computer [computer program] **Version**, v. 5, n. 3, p. 74, 2021.

BRINK, H.; RICHARDS, J.; FETHEROLF, M. **Real-world machine learning**. Simon and Schuster, 2016.

CHAKRABORTY, T. *et al.* Generalizing Adversarial Explanations with Grad-CAM. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern**

Recognition. 2022. p. 187-193.

CHANGWEI, Z. *et al.* Classification of normal and pathological voices using convolutional neural network. *In:***2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)**. IEEE, 2020. p. 325-329.

CHEN, L. *et al.* Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. **Information Sciences**, v. 509, p. 150-163, 2020.

CHEN, L; CHEN, J. Deep Neural Network for Automatic Classification of Pathological Voice Signals. **Journal of Voice**, v. 36, n. 2, p. 288.e15-288.e24, 2022.

CHEN, L; CHEN, J: Deep Neural Network for Automatic Classification of Pathological Voice Signals. **Journal of Voice**, [*s. l.*], p. 1–10, 2020.

CHOLLET, F. **Keras documentation: The Python *Deep Learning*** library. 2016.

DELIYSKI, D. Endoscope motion compensation for laryngeal high-speed videoendoscopy. **Journal of Voice**, v. 19, n. 3, p. 485-496, 2005.

DEJONCKERE PH, BRADLEY P, CLEMENTE P, CORNUT G, BUCHMAN LC, FRIEDRICH G, et al. **A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques**. Eur Arch Otorhinolaryngol. 2001;258(2):77-82. . PMid:11307610.

DEMIRCAN, S; ÖRNEK, H. Comparison of the effects of mel coefficients and spectrogram images via *Deep Learning* in emotion classification. **Traitement du Signal**, 2020.

DING, H. *et al.* Deep connected attention (DCA) ResNet for robust voice pathology detection and classification. **Biomedical Signal Processing and Control**, [*s. l.*], v. 70, p. 102973, 2021.

FACELI, K. **Inteligência Artificial:** Uma abordagem de aprendizado de Máquina. Rio de Janeiro: LTC, 2015.

FANG, S. *et al.* Combining acoustic signals and medical records to improve pathological voice classification. **APSIPA Transactions on Signal and Information Processing**, v. 8, 2019a.

FANG, S. *et al.* Detection of pathological voice using cepstrum vectors: A *Deep Learning* approach. **Journal of Voice**, [*s. l.*], v. 33, n. 5, p. 634–641, 2019.

FERREIRA LP, SANTOS JG, LIMA MFB. Sintoma vocal e sua provável causa: levantamento de dados em uma população. Rev. **CEFAC**. 2009;11(1):110-8.

FREITAS, R. **Metodologia Científica - Um guia prático para profissionais da saúde**. 1. ed. Petrolina: [s.n.], 2017.

GAO, L. *et al.* Handling imbalanced medical image data: A deep-learning-based one-class classification approach. **Artificial intelligence in medicine**, [*s. l.*], v. 108, p. 101935, 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. ***Deep Learning***. MIT press, 2016.

GUMELAR, A. B. *et al.* Enhancing detection of pathological voice disorder based on deep VGG-16 CNN. *In:* **2020 3rd International Conference on Biomedical Engineering (IBIOMED),** [*s. l.*], p. 28–33, 2020.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

HE, K. *et al.* Deep residual learning for image recognition. *In:* **Proceedings of the IEEE conference on computer vision and pattern recognition**, [*s. l.*], v. 2016-December, p. 770–778, 2016.

HECKMAN, J. J; PINTO, R.; SAVELYEV, P. A. Recommended Protocols for Instrumental Assessment of Voice: American Speech- Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. **Angewandte Chemie International Edition, 6(11), 951–952.**, [*s. l.*], p. 1–19, 1967.

HEGDE, S. *et al.* A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. **Journal of Voice**, [*s. l.*], v. 33, n. 6, p. 947.e11-947.e33, 2019.

HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. John Wiley & Sons, 2013.

HOSSAIN, M.; MUHAMMAD, G. Healthcare big data voice pathology assessment framework. **IEEE Access**, v. 4, p. 7806-7815, 2016.

HOWARD, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv**, 2017.

ISLAM, R.; TARIQUE, M.; ABDEL-RAHEEM, E. A survey on signal processing based pathological voice detection techniques. **IEEE Access**, [*s. l.*], v. 8, p. 66749-66776, 2020.

JAITLY, N.; HINTON, G. E. Vocal tract length perturbation (VTLP) improves speech recognition. In: **Proc. ICML Workshop on *Deep Learning* for Audio, Speech and Language**. [*s. l.*], v. 28, 2010.

JO, T. **Machine Learning Foundations:** supervised, unsupervised, and Advanced Learning. Springer International Publishing., 2021.

KAR, K. **Mastering Computer Vision with TensorFlow 2. x: Build Advanced Computer Vision Applications Using Machine Learning and *Deep Learning* Techniques**. Packt Publishing, Limited, [*S. l.: s. n.*], 2020.

KIM, S. ***Deep Learning* with R**, François Chollet, Joseph J. Allaire, Shelter Island, NY: Manning. 2020.

KINGMA, D. P.; BA, J. **Adam: A method for stochastic optimization**, 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KUHN, M.; JOHNSON, K. **Feature engineering and selection:** a practical approach for predictive models. CRC Press, [*S. l.: s. n.*], 2020.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, p. 159-174, 1977.

LAVAN, N. *et al.* Breaking voice identity perception: Expressive voices are more confusable for listeners. **Quarterly Journal of Experimental Psychology**, v. 72, n. 9, p. 2240-2248, 2019.

LECUN, Y. *et al.* Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Taipei, Taiwan, v. 86, n. 11, p. 2278-2324, 1998.

LEITE, D. R. A.; MORAES, R. M. de; LOPES, L. W. Machine learning classifiers: study to categorize healthy and pathological voices. *In*:, 2021, Bogotá. **14th International Conference Advances in Quantitative Laryngology, voice and speech research (AQL)**. Bogotá: AQL Local Organizing Committee, 2021a.

LEITE, D. R. A.; MORAES, R. M. de; LOPES, L. W. Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando Random Forest. **Journal of Health Informatics**, v. 12, 2021.

LEONARDO, M. M. *et al.* Deep feature-based classifiers for fruit fly identification (Diptera: Tephritidae). In: **31º Conference on Graphics, Patterns and Images**, 2018. p. 41-47.

LIU, B.; POLCE, E.; JIANG, J. An objective parameter to classify voice signals based on variation in energy distribution. **Journal of Voice**, v. 33, n. 5, p. 591-602, 2019.

LOPES, L. *et al.* Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses. **Journal of Voice**, v. 31, n. 3, p. 382, 2016.

LOPES, L. *et al.* Accuracy of Acoustic Analysis Measurements in the Evaluation of Patients With Different Laryngeal Diagnoses. **Journal of Voice**, [*s. l.*], v. 31, n. 3, p. 382.e15-382.e26, 2017.

LOPES, L. *et al.* Accuracy of traditional and formant acoustic measurements in the evaluation of vocal quality. In:**CoDAS**. Sociedade Brasileira de Fonoaudiologia, 2018.

LOPES, L. *et al.* Classificação espectrográfica do sinal vocal: relação com o diagnóstico laríngeo e a análise perceptivo-auditiva. **Audiology - Communication Research**, [*s. l.*], v. 25, 2020.

LOPES, L. *et al.* Effectiveness of recurrence quantification measures in discriminating subjects with and without voice disorders. **Journal of Voice**, v. 34, n. 2, p. 208-220, 2020.

LOPES, L. *et al.* Evidence of Internal Consistency in the Spectrographic Analysis Protocol. **Journal of Voice**, 2020a.

LOPES, L. *et al.* Evidence of Internal Consistency in the Spectrographic Analysis Protocol. **Journal of Voice**, [*s. l.*], v. 0, n. 0, 2020b.

LOPES, L.; ALVES, G.; MELO, M. Content evidence of a spectrographic analysis protocol. **Revista CEFAC**, v. 19, n. 4, p. 510–528, 2017.

LOPES, L.; VIEIRA, V.; BEHLAU, M. Performance of Different Acoustic Measures to Discriminate Individuals With and Without Voice Disorders. **Journal of Voice**, [*s. l.*], 2020.

LOVATO, A. *et al.* Multi-Dimensional Voice Program (MDVP) vs Praat for Assessing Euphonic Subjects: A Preliminary Study on the Gender-discriminating Power of Acoustic Analysis Software. **Journal of Voice**, [*s. l.*], v. 30, n. 6, p. 765.e1-765.e5, 2016.

MAHAJAN, A.; CHAUDHARY, S. Categorical image classification based on representational deep network (RESNET). In: **2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)**. IEEE, 2019. p. 327-330.

MALIK, S. *et al.* Data Driven Approach for Eye Disease Classification with Machine Learning. **Applied Sciences**, v. 9, n. 14, p. 2789, 2019.

MARTINEZ, C. E.; RUFINER, H. L. Acoustic analysis of speech for detection of laryngeal pathologies. In: **Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society**. IEEE, 2000. p. 2369-2372.

MCFEE, B. *et al.* librosa: Audio and music signal analysis in python. In: **Proceedings of the 14th python in science conference**. 2015. p. 18-25.

MICROSOFT. **Aprendizado profundo e o aprendizado de máquina - Microsoft**. [*S. l.*], 2022. Disponível em: https://docs.microsoft.com/pt-br/azure/machine-learning/concept-deep-learning-vs-machine-learning. Acesso at: 17 Apr. 2022.

MIRAMONT, J. M. *et al.* Voice Signal Typing Using a Pattern Recognition Approach. **Journal of Voice**, 2020.

MITCHELL, T. M. **Machine Learning**. [S.l.]: [s.n.], 1997.

MOHAMMED, Mazin. *et al.* Voice pathology detection and classification using convolutional neural network model. **Applied Sciences**, v. 10, n. 11, p. 3723, 2020.

MORAES, R. M.; Machado, L. S. (2014) Psychomotor Skills Assessment in Medical Training Based on Virtual Reality Using a Weighted Possibilistic Approach. Knowledge-Based Systems, v. 70, p. 97-102. DOI: https://doi.org/10.1016/j.knosys.2014.05.006

MUHAMMAD, G. *et al.* Edge computing with cloud for voice disorder assessment and treatment. **IEEE Communications Magazine**, v. 56, n. 4, p. 60-65, 2018.

MYTHILI, J.; VIJAYA, M. S. Pathology Voice Detection and Classification Using Ensemble Learning. **International Journal of Engineering Science Invention (IJESI),** v. 7, n. 8, p. 1-8, 2018.

PANWAR, H. *et al.* A *Deep Learning* and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. **Chaos, Solitons &**

Fractals, v. 140, p. 110190, 2020.

PIELAWSKI, N.; WÄHLBY, C. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. **PLOS ONE**, [*s. l.*], v. 15, n. 3, p. 2020.

PISHGAR, M. *et al.* Pathological voice classification using mel-cepstrum vectors and support vector machine. **arXiv preprint arXiv:1812.07729**, 2018.

PONTES, L. *et al.* Transfer function of Brazilian Portuguese oral vowels: a comparative acoustic analysis. [*s. l.*], v. 75, n. 5, p. 680–684, 2009.

PYTHON. Python.org. **Python**, 2019. Disponivel em: <https://www.python.org/>. Acesso em: 10 jun. 2020.

RAČIĆ, L. *et al.* Pneumonia Detection Using *Deep Learning* Based on Convolutional Neural Network. **2021 25th International Conference on Information Technology, IT 2021**, [*s. l.*], n. February, p. 17–20, 2021.

RANSCHAERT, E. R.; MOROZOV, S.; ALGRA, P. R. (Ed.). **Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks**. Springer, 2019.

RAMIG LO, VERDOLINI K. **Treatment efficacy: voice disorders**. J Speech Lang Hear Res. 1998;41(1):101-6. . PMid:9493749.

RAYMOND. H. Colton, J. K. Casper, and R. Leonard, **Understanding voice problems: A physiological perspective for diagnosis and treatment**. Wolters Kluwer Health, 2006.

ROSEN, D. C.; SATALOFF, J. B.; SATALOFF, R. T. **Psychology of voice disorders**. Plural Publishing, 2020.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

ROY, S.; SAYIM, M. I.; AKHAND, M. A. H. Pathological Voice Classification Using *Deep Learning*. In: **2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)**. IEEE, 2019. p. 1-6.

RUSSEL, S.; NORVIG, Peter. Inteligência Artificial. Tradução de Regina Célia Smille. **Rio de Janeiro: Campus Elsevier**, 2013.

SAKASHITA, Y.; AONO, M. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. **Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge**, 2018.

SANTOS, Mikaelle Oliveira. **Análise acústica de desvios vocais infantis utilizando a transformada wavelet**. 2016. 80 f. Dissertação (Mestrado em Engenharia Elétrica) – Curso de Pós-Graduação em Engenharia Elétrica, – Instituto Federal da Paraíba, João Pessoa (PB), 2015.

SELVARAJU, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-

based localization. In: **Proceedings of the IEEE international conference on computer vision**. 2016. p. 618-626.

SETTLES, B. Active Learning, volume 6 of Synthesis Lectures on Artificial Intelligence and Machine Learning. **Morgan & Claypool**, 2012.

SHI, B; IYENGAR, S. S. **Mathematical Theories of Machine Learning-Theory and Applications**. Springer, 2020.

SILVA, A. C. F. **Validação do protocolo de análise espectrográfica da voz baseada na relação com outras variáveis**. 2022. 80 f. Dissertação (Mestrado em Modelos de Decisão e Saúde) – Curso de Pós-graduação em Modelo de Decisão e Saúde – Universidade Federal da Paraíba(PB), 2022.

SIMONYAN, K; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR), 2014.

SPRECHER, A. *et al.* Updating signal typing in voice: addition of type 4 signals. **The Journal of the Acoustical Society of America**, v. 127, n. 6, p. 3710-3716, 2010.

SRIVASTAVA, N. *et al.* Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, v. 15, n. 1, p. 1929-1958, 2014.

SUTTON, R. S. *et al.* **Introduction to reinforcement learning.** 1998.

TITZE, I. R. **Workshop on acoustic voice analysis: Summary statement**. National Center for Voice and Speech, [*s. l.*], 1995.

TRINH, N.; DARRAGH, O. Pathological speech classification using a convolutional neural network. 2018.

TRINH, N.; DARRAGH, O. Pathological speech classification using a convolutional neural network. 2019.

TSIHRINTZIS, G. A.; SOTIROPOULOS, D. N.; JAIN, L. C. Machine learning paradigms: Advances in data analytics. In: **Machine Learning Paradigms**. Springer, Cham, 2019. p. 1-4.

VALENTIM, A. F.; CÔRTES, M. G.; GAMA, A. C. C. Análise espectrográfica da voz: efeito do treinamento visual na confiabilidade da avaliação. **Revista da Sociedade Brasileira de Fonoaudiologia**, v. 15, n. 3, p. 335-342, 2010.

VAVREK, L. *et al.* Deep convolutional neural network for detection of pathological speech. **SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings**, [*s. l.*], p. 245–249, 2021.

VERDE, L. *et al.* A *Deep Learning* Approach for Voice Disorder Detection for Smart Connected Living Environments. **ACM Transactions on Internet Technology**, [*s. l.*], v. 22, n. 1, 2022.

WERNICK, M. N. *et al.* Machine learning in medical imaging. **IEEE signal processing**

magazine, v. 27, n. 4, p. 25-38, 2010.

WU, H. *et al.* A *Deep Learning* method for pathological voice detection using convolutional deep belief networks. **Interspeech 2018**, 2018.

WU, H. *et al.* Convolutional Neural Networks for Pathological Voice Detection. **Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS**, [*s. l.*], v. 2018-July, p. 1–4, 2018.

XING, F.; YANG, L. **Machine Learning and Medical Imaging**. [*S.l.*]: Elsevier, 2016.

YANAGIHARA, N. Significance of Harmonic Changes and Noise Components in Hoarseness. **Journal of speech and hearing research**, [*s. l.*], v. 10, n. 3, p. 531–541, 1967.

YANAGIHARA, N. Significance of harmonic changes and noise components in hoarseness. **Journal of Speech and Hearing Research**, v. 10, n. 3, p. 531-541, 1967.

YANG, Y. **Temporal Data Mining via Unsupervised Ensemble Learning**. Elsevier, 2016.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, v. 3, n. 1, p. 1-130, 2009.