

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"



FACULDADE DE ENGENHARIA DE BAURU



PROGRAMA DE PÓS GRADUAÇÃO EM

ENGENHARIA ELÉTRICA

JOSÉ NARCISIO SALVADOR JUNIOR

**IDENTIFICAÇÃO DE PATOLOGIAS EM PREGAS VOCAIS POR MEIO
DE CLASSIFICADOR *BAG OF FEATURES* E REDES NEURAIS
CONVOLUCIONAIS**

**Bauru
2022**

JOSÉ NARCISIO SALVADOR JUNIOR

**IDENTIFICAÇÃO DE PATOLOGIAS EM PREGAS VOCAIS
POR MEIO DE CLASSIFICADOR *BAG OF FEATURES* E
REDES NEURAIS CONVOLUCIONAIS**

Dissertação apresentada à Faculdade de Engenharia de Bauru/FEB da Universidade Estadual Paulista “Júlio de Mesquita Filho”/UNESP para obtenção do título de Mestre em Engenharia Elétrica

Área de concentração: Automação

Linha de Pesquisa: Mecatrônica

Orientador: Prof. Dr. José Alfredo Covolan
Ulson

Coorientadora: Profa Dra Kelly C A Silverio –
Departamento de Fonoaudiologia da
FOB/USP

Bauru

2022

Salvador Junior, José Narcisio.

Identificação de patologias em pregas vocais por meio de classificador *bag of features* e redes neurais convolucionais/ José Narcisio Salvador Junior, 2020
70 f. : il.

Orientador: José Alfredo Covolan Ulsoon

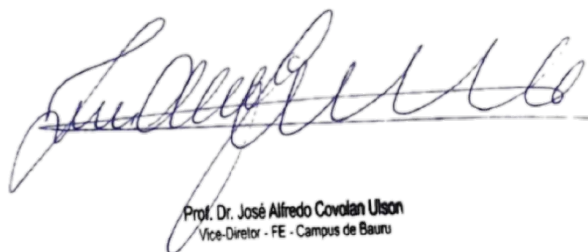
Dissertação (Mestrado)-Universidade Estadual Paulista. Faculdade de Engenharia, Bauru, 2022

1. Classificação de Imagens. 2. Bag of Features. 3. Redes Neurais Convolucionais. 4. Patologia de Laringe. I. Universidade Estadual Paulista. Faculdade de Arquitetura, Artes e Comunicação. II. Título.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE JOSÉ NARCISIO SALVADOR JUNIOR, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, DA FACULDADE DE ENGENHARIA - CÂMPUS DE BAURU.

Aos 05 dias do mês de outubro do ano de 2021, às 10:00 horas, no(a) Via sistemas de videoconferência e outras ferramentas para comunicação a distância, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de JOSÉ NARCISIO SALVADOR JUNIOR, intitulada **IDENTIFICAÇÃO DE PATOLOGIAS EM PREGAS VOCALIS POR MEIO DE CLASSIFICADOR BAG OF FEATURES E REDES NEURAIIS CONVOLUCIONAIS**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. JOSÉ ALFREDO COVOLAN ULSON (Orientador(a) - Participação Virtual) do(a) Departamento de Engenharia Eletrica / Faculdade de Engenharia de Bauru UNESP, Prof. Dr. ADRIANO DE SOUZA MARQUES (Participação Virtual) do(a) Engenharia da Computação / Instituto Federal de São Paulo - IFSP - Câmpus Birigui, Prof. Dr. BRUNO ALBUQUERQUE DE CASTRO (Participação Virtual) do(a) Professor Substituto do Departamento de Engenharia Elétrica / Faculdade de Engenharia de Bauru - UNESP. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: aprovado _____. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. JOSÉ ALFREDO COVOLAN ULSON



Prof. Dr. José Alfredo Covolan Ulson
Vice-Diretor - FE - Câmpus de Bauru

Dedico aos meus pais, professores e amigos, em especial do Ministérios das Universidades Renovadas, que tanto me incentivam e auxiliam a buscar o conhecimento e sobretudo a Verdade.

AGRADECIMENTOS

Agradeço a Deus que em sua divina providência permitiu que eu chegasse até aqui, mesmo em meio a dificuldades. A Ele não só o agradecimento, mas todo o louvor - *Ad maiorem Dei Gloriam*

Agradeço aos meus pais, Narcisio e Regina, por ter desde cedo me incentivado à vida de estudos e a dedicação em cada uma das atividades.

Agradeço a minha noiva, Geizi, por todo o apoio, sobretudo no final da pesquisa, e todo cuidado em me ajudar de maneira integral a passar por essa etapa.

Agradeço aos amigos do Ministério das Universidades Renovadas por serem exemplo de jovens que buscam em tudo conciliar fé e razão. Especialmente a um grande amigo e irmão do coração, Tiago Cabana, por ter sido e ser aquele que me auxilia em em várias dimensões da vida, inclusive a acadêmica.

Agradeço ao professor José Alfredo por ter me aceitado como aluno mesmo tendo vindo de outra instituição, como aluno especial e posteriormente como aluno regular do programa de mestrado do departamento de Engenharia Elétrica. Agradeço-o ainda pela paciência e motivação em cada uma das etapas, sobretudo as do fechamento da pesquisa.

Agradeço a toda equipe de fonoaudiologia da Universidade de São Paulo (FOB-USP), nominalmente agradecendo a Prof. Dra. Kelly Silvério pelo apoio à pesquisa possibilitando um estudo multidisciplinar.

Agradeço a todos que, de uma alguma maneira, contribuíram para a realização dessa dissertação.

Agradeço, por fim, a todo o departamento de Pós Graduação de Engenharia Elétrica e à Coordenação de Aperfeiçoamento Pessoal de nível Superior.

“A educação é obra do coração”.

- Dom Bosco

RESUMO

Patologias de laringe, como cistos e nódulos, são caracterizadas como alterações benignas e têm seu diagnóstico dado por meio de análise visual e perceptivo-auditiva em telelaringoscopias, exame endoscópico ambulatorial para inspeções laríngeas auxiliadas por vídeo. Porém, uma inspeção visual incorreta, que ocorre por similaridade vocal, acústica e também em suas propriedades anatômicas, pode levar o diagnóstico de uma das patologias benignas a prejudicar o diagnóstico de outra. Visto que há possibilidade de inversão na determinação diagnóstica entre essas duas patologias, pela equipe médica responsável, propõe-se uma ferramenta para auxílio diagnóstico que visa identificar e classificar imagens de pregas vocais saudáveis e não saudáveis - nódulos e cistos - advindas de telelaringoscopias e demais bancos de imagens, utilizando classificador inteligente Bag of Features e Redes Neurais Convolucionais - AlexNet, ResNet e DenseNet. Foi realizada segmentação de 45 imagens e aplicação de técnicas de *Data Augmentation*, passando a ter 300 imagens, divididas em três grupos entre teste e validação; sendo 100 imagens no grupo “normais”, 100 imagens no grupo “cistos” e 100 imagens no grupo “nódulos”. Ambos os grupos foram submetidos a 5 testes com 19 épocas cada um em todas as ferramentas utilizadas e alcançou-se o melhor resultados na CNN ResNet com acurácia de 97,33% de robustez e erro válido de 2,67% nos testes e classificação de pregas vocais saudáveis - normais e não saudáveis cistos e nódulos.

Palavras-chave: Classificação de Imagens, Bag of Features, Redes Neurais Convolucionais, Patologia de Laringe

ABSTRACT

Laryngeal pathologies, such as cysts and nodules, are characterized as benign changes and have their diagnosis given through visual and auditory-perceptual analysis in telelaryngoscopy. Outpatient endoscopic examination for video-assisted laryngeal inspections. However, an incorrect visual inspection, which occurs due to vocal and acoustic similarity and also in their anatomical properties, can lead to the diagnosis of one of the benign pathologies affecting the diagnosis of the other. Since there is a possibility of inversion in the diagnostic determination between these two pathologies, by the responsible medical team, a diagnostic aid tool is proposed, which aims to identify and classify images of healthy and unhealthy vocal folds - nodules and cysts - arising from telelaryngoscopy and other image banks, using Bag of Features smart classifier and Convolutional Neural Networks - AlexNet, ResNet and DenseNet. 45 images were segmented and Data Augmentation techniques were applied, which now has 300 images, divided into three groups between test and validation; 100 images in the “normal” group, 100 images in the “cysts” group and 100 images in the “nodules” group. Both groups were submitted to 5 tests with 19 times each in all the tools used and the best results were achieved in CNN ResNet with an accuracy of 97.33% of robustness and valid error of 2.67% in the tests and classification of healthy vocal folds - normal and unhealthy cysts and nodules.

Key-words: Image Classification, Bag of Features, Convolutional Neural Networks, Laryngeal Pathology

LISTA DE FIGURAS

Figura 1 - Anatomia da Laringe.....	19
Figura 2 - Pregas vocais localizadas no espaço interior da cartilagem tireóide.....	21
Figura 3 - Esquema e partes das pregas vocais.....	21
Figura 4 - Visão - Laringoscopia indireta.....	23
Figura 5a - Estroboscopia rígido.....	24
Figura 5b - Uso do estroboscópio em uma videolaringoscopia ou TLE.....	24
Figura 6 - Imagem capturada a partir de uma TLE de PPVV saudável.....	26
Figura 7a - posição da prega vocal na fonação.....	27
Figura 7b - posição da prega vocal na respiração.....	27
Figura 8 - Captura de uma TLE de PPVV com nódulos vocais bilaterais e simétricos.....	28
Figura 9 - Captura de uma TLE de PPVV com cisto.....	30
Figura 10 - Salto de camadas em ResNet.....	39
Figura 11 - Estruturas em camadas das CNNs DenseNet e ResNet.....	40
Figura 12 - Exemplo de Matriz de Confusão.....	41
Figura 13 - Fluxograma metodológico.....	44
Figura 14 - Captura de tela da telelaringoscopia.....	47
Figura 15 - Recorte e filtro escala em cinza.....	48
Figura 16 - Prega vocal com cisto submetida à rotação.....	49
Figura 17 - Prega vocal com nódulo submetida à transladação.....	49
Figura 18 - Prega vocal normal/saudável.....	50
Figura 19 - Image Batch Processor - Validação Classificador Bag of Features.....	52
Figura 20 - Matriz de Confusão DenseNet121.....	54
Figura 21 - Maiores perdas de treinamento DenseNet121.....	55
Figura 22 - Matriz de Confusão AlexNet.....	57
Figura 23 - Maiores perdas de treinamento AlexNet.....	57

Figura 24 - Matriz de Confusão AlexNet.....	59
Figura 25 - Maiores perdas de treinamento ResNet.....	60

LISTA DE TABELAS

Tabela 1 - Divisão em camadas das pregas vocais.....	26
Tabela 2 - Resultados da classificação BoF.....	51
Tabela 3 - Interações DenseNet121.....	53
Tabela 4 - Interações AlexNet.....	56
Tabela 5 - Interações ResNet34.....	58

LISTA DE ABREVIATURAS

TLE	Telelaringoscopia
PPVV	Pregas vocais
BoF	<i>Bag of Features</i>
BoW	<i>Bag of Words</i>
PPVV-S	Pregas vocais saudáveis
PPVV-NS	Pregas vocais não saudáveis
DA	Data Augmentation
CNN	<i>Convolutional Neural Network</i>
SURF	<i>Speeded Up Robust Feature</i>
RNA	Redes Neurais Artificiais

SUMÁRIO

1 INTRODUÇÃO	15
2 REVISÃO BIBLIOGRÁFICA	18
2.1 Histórico da laringoscopia e anatomia da laringe	18
2.1.1 A invenção do Laringoscópio	22
2.1.2 O estroboscópio Laríngeo	23
2.2 Diagnóstico Laríngeo - Pregas vocais saudáveis e patológicas	25
2.2.1 Pregas vocais saudáveis	25
2.2.2 Nódulos	28
2.2.3 Cistos	29
2.3 Extração de características e segmentação	32
2.3.1 Contraste e suavização	32
2.3.2 Data Augmentation e aplicação de ruído artificial	32
2.4 Classificador Bag of Features	34
2.5 Redes Neurais Convolucionais - CNN	36
2.6 Métrica de Validação - Matriz de Confusão	40
2.7 Trabalhos correlatos	41
3 MATERIAIS E MÉTODOS	44
3.1 Delimitação de uso de banco de dados	45
3.2 Segmentação das Imagens	47
3.3 Divisão dos dados - teste e validação	50
4 RESULTADOS E DISCUSSÃO	51
4.1 Resultados - Classificador Bag of Features	51
4.2 Resultados - CNNs	53
4.3 Discussão	60
5 CONCLUSÃO	63
REFERÊNCIAS	65
APÊNDICE	68

1 INTRODUÇÃO

Patologias na laringe podem causar problemas em dimensão social e profissional nas pessoas, afetando diretamente a voz, meio mais importante de comunicação, e a qualidade de vida.

O diagnóstico de patologias na laringe é dado por meio de visualização das pregas vocais dos pacientes, entre outras formas, nas videolaringoscopias, exames de vídeo que permitem a visualização da laringe, por otorrinolaringologistas e acompanhado de fonoaudiólogos por meio de um aparelho chamado vídeo-estroboscópio. Este processo de inspeção associa-se a análise perceptivo-auditiva (visualização das pregas vocais e análise sonora) dos vocábulos dos pacientes para o diagnóstico da alteração de voz, denominada de disfonia, dentre as quais se enquadram patologias como cistos e nódulos. Um exemplo de como ocorre o diagnóstico da disfonia é que, segundo Barata (2010, p. 5),

“há relação entre os parâmetros visuais e perceptivo-auditivos selecionados, analisados dos trechos de fonação expiratória e inspiratória durante o exame laríngeo com a hipótese médica diagnóstica de nódulo e cisto”.

Apesar de algumas formas de análises, nódulos, pólipos e cistos, por exemplo, são similares em suas propriedades auditivas. Somente a análise visual é capaz de diagnosticar a lesão laríngea, mas, segundo Sodr  (2016), muitas vezes, torna-se dif cil a diferencia  o entre n dulos e cistos, como no exemplo citado, quando se analisa as imagens das videolaringoscopias. Essa an lise pode ser aprimorada com o uso de ferramentas computacionais que s o capazes de auxiliar na conclus o do diagn stico, visto que, dependendo da patologia encontrada, a conduta terap utica para o seu esclarecimento, e mesmo o seu tratamento,   o processo cir rgico, o que   custoso tanto para o paciente, como para a equipe de sa de vocal.

Atualmente a intera  o entre Engenharia e cursos de Ci ncias da Sa de cresce cada vez mais e a  rea m dica depende direta ou indiretamente de novas

tecnologias, destacando crescimento nas publicações relacionadas a sistemas inteligentes na área de saúde, como, por exemplo, estudos em análise de imagens de laringes (pregas vocais) na área da Fonoaudiologia e Medicina. Aponta-se esse crescimento, segundo PINHEIRO (2017), visto que a análise de imagens médicas é um dos domínios que mais tem atraído a atenção de investigadores devido às potenciais contribuições na melhoria das condições da saúde humana que daí podem advir.

Considerando-se a necessidade de visualização de pregas vocais há a possibilidade de usar recursos de visão computacional para a detecção de patologias na laringe contribuindo para diminuir possíveis falhas, visto que imagens são frequentemente examinadas apenas por inspeção visual, tornando o processo extremamente subjetivo, demorado e pode induzir a erros.

Delimita-se este trabalho em análise de imagens visto que o processo mais utilizado que conduz ao diagnóstico de patologias, segundo Behlau (2008), é por meio das videolaringoscopias (ou telelaringoscopias), onde a visualização é fundamental.

Diversos trabalhos atuais, como descritos posteriormente, usam técnicas de análise de voz para propor auxílio diagnóstico laríngeo, com a justificativa inicial que os exames usando estroboscópios são invasivos e desconfortáveis, recorrendo assim somente ao processamento acústico. Porém, o uso de classificação de imagens se torna uma ferramenta muito eficaz no processo de caracterização de patologias entre nódulos e cistos visto que vocal e acusticamente são patologias similares.

Logo, como afirma PINHEIRO (2017) “o processamento computacional e a análise de imagens são determinantes para obter informação estatisticamente relevante e fidedigna”, visto que se busca resultados elevados, acima de 95% de acerto, para a aplicação de ferramentas das engenharias nas áreas médicas.

Para realizar análise computacional de imagens, alguns passos importantes devem ser seguidos. Na maioria das vezes, antes de extrair qualquer tipo de informação a partir de uma imagem, é necessário o pré-processamento dos seus dados, de modo a corrigir eventuais desequilíbrios na sua iluminação, reduzir o ruído

e/ou realçar determinadas características como, por exemplo, os contornos e as bordas dos objetos presentes na imagem (TEEFFELEN; SHAEVITZ; GITAI, 2012).

Por fim, as ferramentas utilizadas para a classificação entre pregas vocais saudáveis e patológicas (com presença de nódulos e cistos) serão o classificador Bag Of Features (BoF) e Redes Neurais Convolucionais (CNN).

O objetivo desta dissertação é identificar e classificar imagens de pregas vocais saudáveis e não saudáveis - patologias laríngeas leves, nódulos e cistos - advindas de telelaringoscopias e demais bancos de imagens.]

Frente ao exposto, pretende-se pre processar as imagens, aplicando técnicas de segmentação de imagens para se verificar a possibilidade de identificação de patologias laríngeas por meio de visão computacional, para auxílio na tomada de decisão e esclarecimento de diagnóstico laringológico e apoio as terapias fonoaudiológicas.

Para isso, este trabalho está dividido em 4 capítulos, além desta introdução. No capítulo dois serão descritas as bases desta pesquisa e as delimitações de escolhas das técnicas na revisão bibliográfica. Posteriormente, no capítulo três será descrita a aplicação dos materiais e métodos escolhidos. No capítulo quatro os resultados do trabalho bem como as discussões frente ao alcance do método e em suma, no capítulo cinco, as conclusões. Finalizando completamente com referências e apêndices.

2 REVISÃO BIBLIOGRÁFICA

Para a compreensão da aplicação de classificadores inteligentes, redes neurais e demais ferramentas de visão computacional vinculados a inteligência artificial, tal como citado acima, necessita-se de um conhecimento mesmo que básico de conceitos e procedimentos da área médica de otorrinolaringologia e também de fonoaudiologia, que será exposto nas seções 2.1 e 2.2. Além disso, é necessário também compreensão sólida dos processos de segmentação de imagens, que consiste um dos passos mais importantes para a configuração de entrada de imagens nas técnicas utilizadas para a classificação de patologias, que será descrito na seção 2.3. Finalizando esse capítulo a seção 2.4 trará conceituações sobre o classificador Bag Of Features, a seção 2.5 sobre as Redes Neurais Convolucionais utilizadas e por fim a seção 2.6 os trabalhos correlatos que auxiliam o domínio e melhor ilustração dos temas propostos.

2.1 Histórico da laringoscopia e anatomia da laringe

A laringoscopia é um procedimento de observação da laringe que faz uso de um instrumento designado chamado laringoscópio. A seguir será demonstrado de maneira breve tanto o histórico da laringoscopia como conceitos básicos de anatomia laríngea, porém muito importantes para o entendimento das futuras análises computacionais.

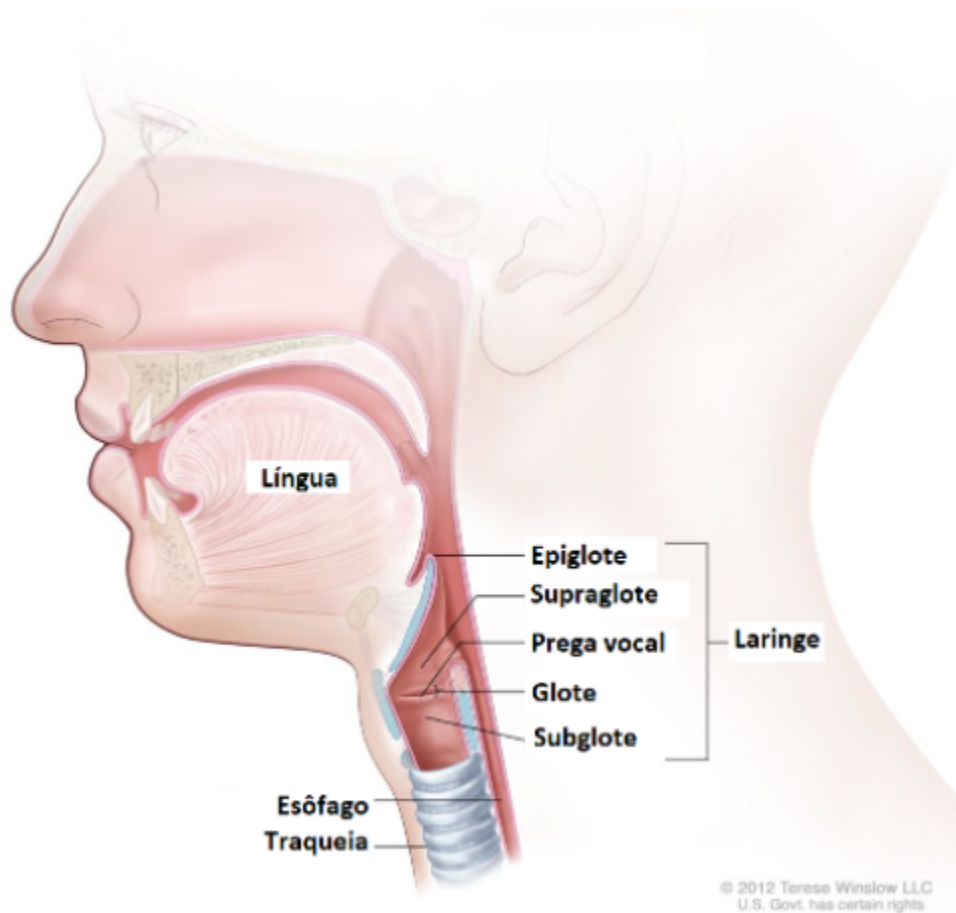
A laringe é geralmente examinada por dois métodos: laringoscopia indireta e laringoscopia direta. No primeiro caso, dentre outros aspectos e características o paciente recebe somente uma anestesia local, na língua, o que não acontece no segundo caso por se tratar de uma sedação geral. Na dissertação será somente

abordado conceitos sobre laringoscopia indireta, por se tratar da forma que foram realizados os exames nos quais serão processados para a classificação inteligente.

O início das observações e exames de laringes se deu, segundo Jahn, A., & Blitzer, A. (1996, apud AIDO, 2017, p 13.), no início de século XIX com o médico alemão Bozzini e alguns anos depois com o Professor de Voz Manuel Garcia em 1854. Mas desde o século XV vinha-se estudando a anatomia da laringe. Desde os primeiros estudos, a análise visual se caracteriza como o mais importante fator nos exames laríngeos, tendo como principal problema a luz para a correta visualização da laringe.

Para se entender o problema da visualização, na Figura 1 é demonstrado a anatomia básica de uma laringe:

Figura 1: Anatomia da laringe



A complexidade da visualização laringe, além da fonte de luz adequada, como mencionado acima, reside no fato de que o eixo longitudinal do órgão está localizado em um ângulo reto ao eixo da cavidade oral, pelo que ela não pode ser inspecionada de forma direta, com uma inspeção visual sem auxílio de outros equipamentos, sobretudo, para a visualização correta das pregas vocais, para diagnóstico de patologias nesta região. Isso se dá pela divisão laringe em três partes, segundo Behlau, M. et al (2001, p.42), supraglote, glote e infraglote. Essa divisão também é chamada de cavidades da laringe. Na região supraglótica inclui a epiglote, pregas vocais falsas e ventrículos (porção entre as pregas ventriculares acima e pregas vocais abaixo); a região glótica – inclui as pregas vocais verdadeiras e as comissuras anteriores e posterior; a região subglótica – inicia-se a um centímetro abaixo das pregas vocais verdadeiras e alonga-se até à borda inferior da cricóide.

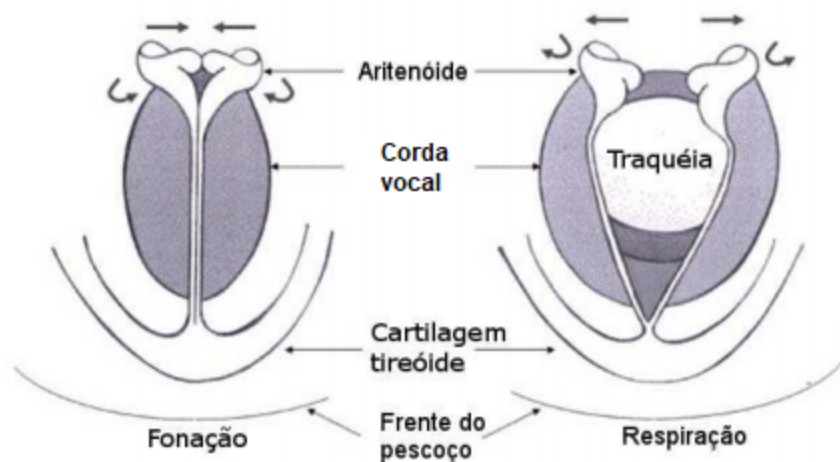
Dentre essas regiões a de maior foco é a segunda, ou seja, a região ou cavidade glótica, onde se encontra as pregas vocais, que serão analisadas como maior ênfase nesta dissertação. Porém, na fase de segmentação das imagens, segundo KUO; Chung-Feng; Jeffrey et al. (2013) a área da supraglote que inclui a pregas vocais falsas deve ser posta em evidência ou em separação, para não haver problematização nos resultados de classificação, caso as imagens utilizadas, segundo a equipe de fonoaudiologia, já não tenha ocultado (pela forma de fonação - sustentação de vogais) essa parte.

Sendo assim, segundo AIDOS (2017), as pregas vocais são pregas músculo membranosas que se conectam anteriormente à cartilagem tireóide e posteriormente às cartilagens aritenóides, conforme ilustrado na Figura 2. Ainda segundo o mesmo autor durante a laringoscopia

podem ser visualizadas em forma de triângulo com ápice inserido na parede anterior da cartilagem tireóide, e sua base, posteriormente nas cartilagens aritenóides (“V” invertido, dependendo da visualização-espelho), conforme

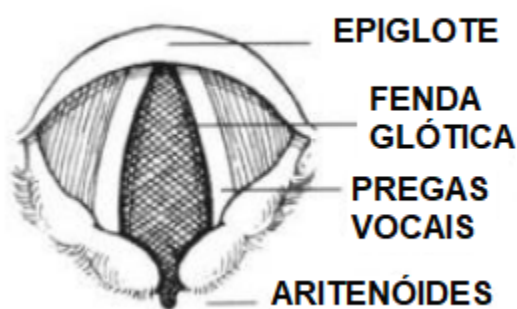
ilustrado na Figura 3 e o espaço entre ambas é denominado glote. (AIDOS, 2017, p. 19)

Figura 2 – Pregas vocais localizadas no espaço interior da cartilagem tireóide.



Fonte: Adaptado de <https://newt.phys.unsw.edu.au/jw/voice.html>

Figura 3 – Esquema e partes das pregas vocais



Fonte: Adaptado de AIDOS, 2017

A partir dessas figuras pode-se notar a dificuldade de visualização das pregas vocais para diagnósticos vocais laringológicos. Com isso, até chegar em novas ferramentas de análise perceptivo-auditiva ligada a processamento automático por Transformada de Wavelet Packet, Máquinas de Vetor Suporte (SVM, do inglês: support vector machine), Redes Neurais Artificiais, Algoritmos matemáticos combinados a outras tecnologias de processamento de imagem ou até

mesmo o processamento de imagens por classificadores inteligentes como proposto nesta dissertação, se tem um longo histórico de desenvolvimento na laringoscopia direta ou indireta, que será descrito resumidamente abaixo.

Inicialmente, remontando os séculos IX e X, segundo Urbina, A. P., & Trullén, A. P. (2006, apud AIDO, 2017, p. 21), com os médicos Rhazes (850-932) e Abulkassis de Córdoba (936- 1013), descreve-se a inspeção visual de órgãos internos a partir da reflexão da luz solar mediante o emprego de um espéculo (instrumento destinado a dilatar a entrada de certas cavidades orgânicas) com espelhos de vidro. Itens que foram empregados se fato, após muitos avanços com Levret, que:

Conseguiu uma iluminação das cavidades introduzindo um espelho fabricado com aço polido onde se podiam ver refletidas lesões polipóides da garganta ouvidos e fossas nasais. Pensa-se poder ter sido realizada, nesta altura, a primeira laringoscopia, no entanto com resultados nada satisfatórios. (Urbina, A. P., & Trullén, A. P. 2006, apud AIDO, 2017, p. 22)

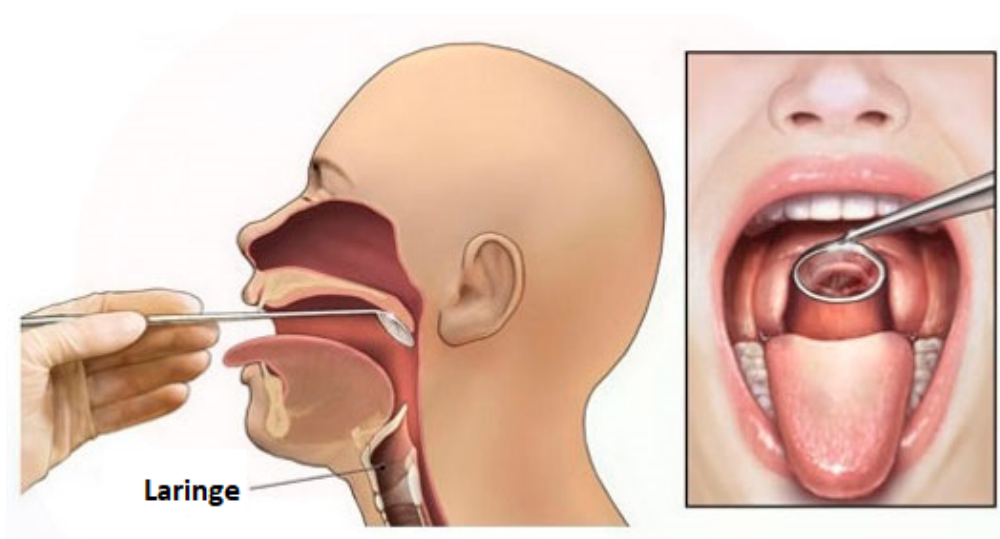
Mais tarde, segundo Aidos (2017), instrumentos criados por Archibald em 1789 e Bozzini em 1807 para visualização de cavidades corporais, sobretudo a cavidade oral, não foram satisfatórias para o emprego destinado.

Outros equipamentos que trouxeram sua contribuição para a observação laríngea mas que não foram satisfatórios foram os de Benjamin Guy Babington (1794-1866) e Bennati (1798-1843), ora por não utilização de luz artificial ora pelas dimensões reduzidas que não favoreciam a visualização.

2.1.1 A invenção do Laringoscópio

Manuel Garcia é considerado o pai da laringoscopia indireta, possibilitando a partir de 1854, segundo Chavolla-Magaña RMA, C.-G. R., Pérez-, & MA., F. (2016, apud AIDOS, 2017, p. 29) a visualização indireta da laringe, através um pequeno espelho de dentista bem como um espelho de mão, conforme ilustra a Figura 4.

Figura 4 - Visão - Laringoscopia indireta



Fonte: https://pt.iliveok.com/health/laringoscopia_105671i15989.html

2.1.2 O estroboscópio Laríngeo

A técnica moderna de laringoscopia indireta não é diferente da utilizada há mais de 150 anos, mas passou por modernização principalmente quanto a iluminação e quanto a forma de visualização, possibilitando uma inspeção visual mais acertada.

Se trata do estroboscópio laríngeo, equipamento usado para os exames laringológicos, denominados videolaringoscopias ou telelaringoscopias. Esse equipamento consiste, segundo ALMEIDA (2017), em um instrumento “destinado à iluminação ou à observação periódica de um sistema vibrante, e com o auxílio do qual se podem reconhecer diferentes características do movimento”. Pode ser rígido ou flexível.

A primeira publicação da utilização do estroboscópio mecânico para o exame da vibração das pregas vocais foi feito em 1895 com Oertel. Somente 65 anos mais tarde é que foi desenvolvida a estroboscopia eletrônica. Porém, fontes de luzes, lentes e câmeras configuravam os principais problemas para o avanço dessa técnica. Sendo quebrada essas barreiras com telescópios rígidos, luzes de maior eficiência e câmeras que pudessem ser introduzidas nas cavidades laríngeas que a

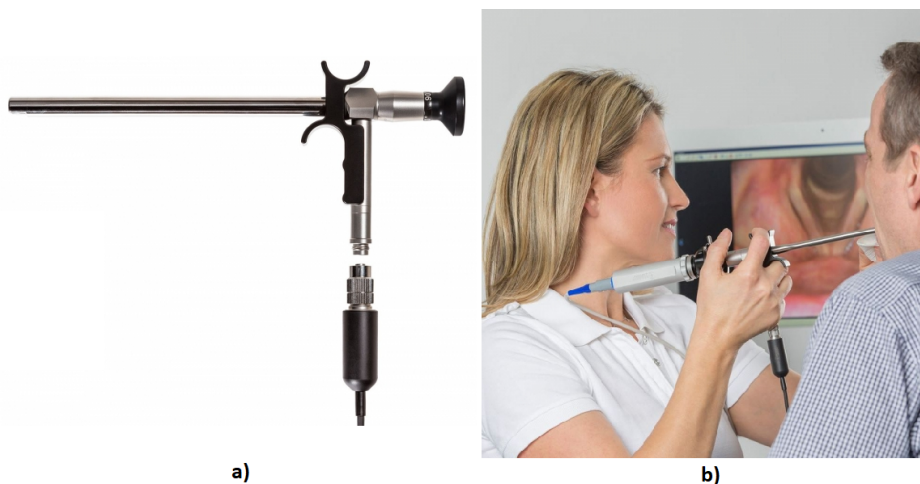
telelaringoestroboscopia (TLE) laríngea tornou-se uma importante ferramenta para o diagnóstico de lesões das pregas vocais (Mendes, 2008). A ascensão tecnológica inclui uma melhor iluminação com uma fonte de luz potente e transmissão da mesma por fibra ótica e microcâmeras para melhor visualização e conforto nos procedimentos dos exames, possibilitando a execução das TLEs.

A TLE, segundo Dailey e Poels:

“permite a visualização, em tempo real, do fechamento glótico, além da movimentação da onda mucosa durante a fonação. Permite a avaliação do ligamento vocal com a manobra de inspiração”. (Dailey SH et. al & Poels PP, apud, Mendes, 2008, p. 870)

Essa visualização é feita pelo estroboscópio laríngeo, demonstrado na Figura 5a, juntamente com o mesmo instrumento sendo utilizado em uma telelaringoscopia, na Figura 5b.

Figura 5 - a)Estroboscópio rígido b)Uso do estroboscópio em uma videolaringoscopia ou TLE.



Fonte: Adaptado de
http://www.dgsotorrinolaringologia.med.br/APOST_SEMIOL_TELELARINGOSC.html

As imagens a serem utilizadas como dados de entrada no classificador e redes neurais são de TLE. Mesmo as imagens já recortadas de vídeos são

provenientes desse exame. Isso é um ponto que facilita a alimentação de dados, visto que, segundo Pontes:

Todo paciente com queixa de alteração vocal deve ser avaliado no mínimo pela TLE. A disfonia pode ser o resultado de pequenas anormalidades na vibração da onda mucosa das pregas vocais ou secundárias a uma incompetência do fechamento glótico ou a uma associação entre estes fatores (Pontes, 1998, apud Mendes, 2008, p. 870)

Isso caracteriza um vasto banco de dados, se encontrado os endereços e formas corretas de utilização. Logo, é a partir das TLE que se obtém as imagens para o diagnóstico de laringe, sobretudo em caso de disfonias, que tem relação com as patologias benignas das pregas vocais. Com isso, vê-se a seguir a forma básica de se diagnosticar essas patologias benignas e também os casos saudáveis. Esse passo tem sua importância no processo, visto que cada uma dessas informações alimentam os inputs dos classificadores e redes neurais, a partir de uma boa segmentação seguindo essas características abaixo descritas.

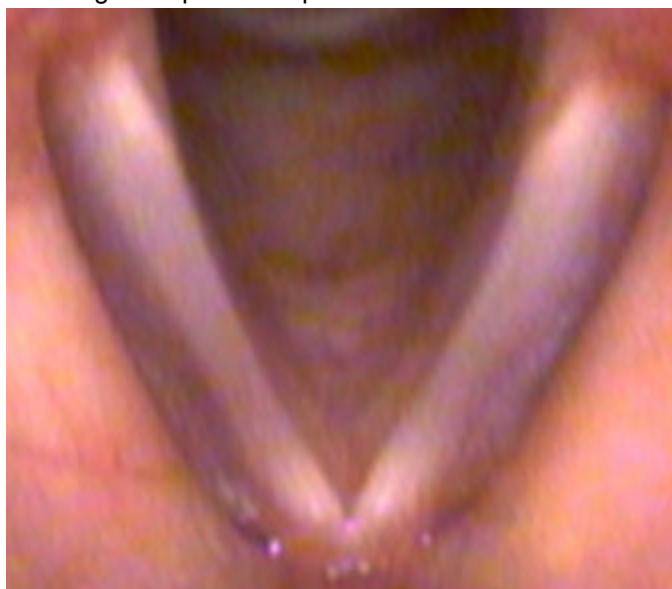
2.2 Diagnóstico Laríngeo - Pregas vocais saudáveis e patológicas

A seguir serão descritos os meios de diagnóstico laringeo para ambos os casos, pregas vocais saudáveis e não saudáveis - cistos e nódulos - para correta programação e acerto de dados de entrada na análise computacional do classificador e das redes neurais convolucionais.

2.2.1 Pregas vocais saudáveis

As PPVV saudáveis são aquelas que estão livres de alguma protuberância ou incisões nas mesmas, conforme demonstrado na Figura 6. Segundo Behlau (2001), o corpo das PPVVs são formados por camadas que são divididas em três seções - cobertura (epitélio), transição e corpo (músculo vocal), de acordo com o que é demonstrado na Tabela 1.

Figura 6: Imagem capturada a partir de uma TLE de PPVV saudável.



Fonte:

<https://www.voicedoctor.net/media/normal-vocal-cord/flexible-videoscope-exam/normal-young-female-vocal-cords>

Tabela 1 - Divisão em camadas das pregas vocais.

Cobertura	Transição	Corpo
A transição é formada pelas camadas intermediária e profunda da lâmina própria (ou ligamento vocal).	A cobertura é a mucosa da prega vocal, formada pelo epitélio e pela camada superficial da lâmina própria (espaço de Reinke). Como são frouxamente conectados ao ligamento e músculo vocal, apresentam grande mobilidade e elasticidade	O corpo é formado pelo músculo vocal, que pode ser considerado rígido no momento da fonação, tanto pela adução e tensão das pregas vocais (gerada pela contração dos músculos intrínsecos), como pela contração do próprio músculo vocal.

Fonte: Adaptado de

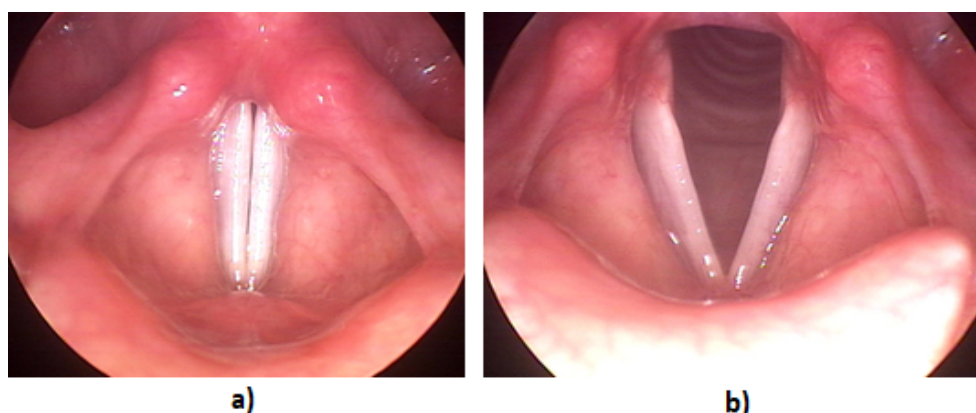
<https://idoc.pub/download/voz-o-livro-do-especialista-mara-behlaupdf-x4e6535g59n3>

As lesões que serão estudadas - cistos e nódulos - segundo CIELO (2011), não apresentam alterações histológicas nas camadas de transição e no músculo vocal (corpo da prega vocal), apenas na cobertura (espaço de Reinke). Em outras

palavras, as protuberâncias serão encontradas no contorno ou epitélio, das pregas vocais, facilitando análise computacional através de aplicação de filtro de segmentação “*Find Edges*”.

Na inspeção visual pela TLE se observa duas posições características das pregas vocais, conforme é visualizado na Figura 7, realizado por um exame utilizando estroboscópico. Se trata da posição durante a fonação (fala) e a respiração (estado de repouso vocal), onde se pode observar as três seções de camadas das PPVV.

Figura 7: a) posição da prega vocal na fonação. b) posição da prega vocal na respiração.



Fonte:

<http://www.tecnicavocalmaster.com.br/2016/08/aula-de-canto-e-tecnica-vocal-o-que-sao-cordas-vocais-pregas-vocais.html>

No momento em que se respira normalmente, as pregas vocais apresentam-se separadas e formam uma abertura com formato triangular entre elas. Quando as pregas vibram, no momento da fonação, elas ficam muito próximas uma da outra.

Todas as imagens utilizadas para análise de identificação e classificação são imagens capturadas somente no momento de repouso vocal, respiração, tendo as PPVV “abertas”, formando o espaço glótico e classe I de verificação em laringoscopia.

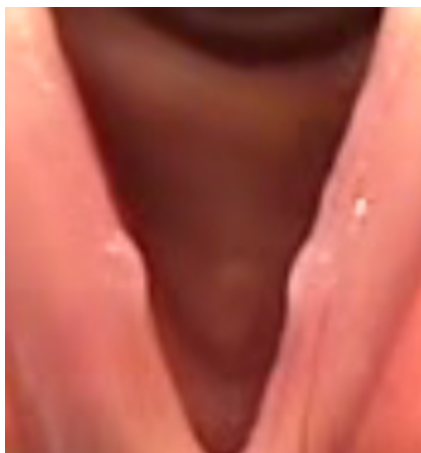
A seguir será exposto sobre o diagnóstico de pregas vocais patológicas, destacando duas patologias aplicadas nesta dissertação: Nódulos e Cistos - patologias laríngeas consideradas benignas.

2.2.2 Nódulos

Nódulos vocais, segundo CIELO (2011), são protuberâncias bilaterais quase sempre simétricas que se localizam na extremidade livre e superfície inferior próximo ao ponto médio da porção membranosa das pregas vocais. O mesmo autor ainda ressalta que podem se apresentar de tamanhos diferentes em decorrência de assimetrias anatômicas e/ou vibratórias entre as pregas vocais.

Na Figura 8 é demonstrado uma PPVV tendo os nódulos vocais como patologia. Assim como descrito por CIELO (2011), as lesões são bilaterais e simétricas.

Figura 8 - Captura de uma TLE de PPVV com nódulos vocais bilaterais e simétricos



Fonte: <https://www.voicedoctor.net/media/structural-injury>

Outras características dos nódulos, importantes para a inspeção visual e sobretudo a preparação para a aplicação de visão computacional por classificadores inteligentes, são as possibilidades de apresentarem, segundo

Behlau (2001), forma pontiaguda ou ainda espaiada, sendo denominados nesta última de espessamento nodular. Além do mais, segundo DO BRASIL; PERAZZO; YAMAZAKI (2012), quanto mais antigos os nódulos, mais fibrosos são e, portanto, se apresentam com coloração esbranquiçada.

Se tratando de imagens advindas de TLE - laringoscopia indireta - pode-se observar lesões brancas e brilhantes, com base alargada, simétricas, situadas na junção dos terços anterior e médio das pregas vocais.

Como diagnóstico diferencial, destaca PASSEROTTI (2005), que nódulos unilaterais ou assimétricos são geralmente cistos intracordais. Podem ser confundidos com os nódulos devido ao aspecto da prega vocal contralateral com lesão irritativa ou fibrosa por contato com o cisto.

Essas informações são primordiais e justificam a futura aplicação de visão computacional à TLE para não ocasionar uma inspeção visual duvidosa e um diagnóstico incorreto pela equipe médica.

2.2.3 Cistos

A lesão cística é definida como cavidade fechada delimitada por uma cápsula revestida por tecido epitelial que pode ser de retenção mucosa epidermóide (Behlau, 2001). A literatura caracteriza suas formas anatômicas e fisiológicas da seguinte maneira, principalmente segundo PASSEROTTI (2005):

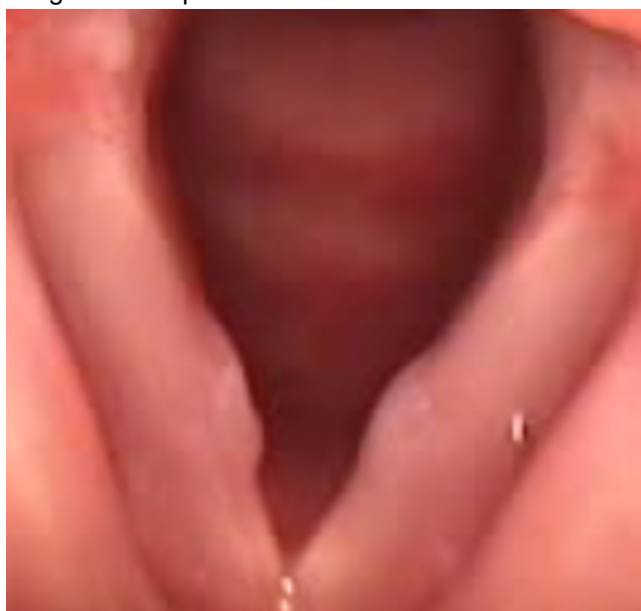
O primeiro apresenta uma parede fina de revestimento constituído por tecido epitelial cuboidal ou cilíndrico e tem em seu interior acúmulo de material PAS-positivo, o que comprova a sua origem mucóide. É frequentemente translúcido e apoiado na margem livre das pregas vocais. Pode ser decorrente de glândula anômala congênita posicionada fora do local habitual e que inflama devido a fonotrauma. Esse tipo de cisto é considerado de origem adquirida e, portanto, não é tido como pertencente ao grupo das alterações estruturais mínimas de cobertura. Já o cisto epidermóide caracteriza-se por apresentar um epitélio de revestimento malpighiano, pluriestratificado, tendo em seu interior acúmulo de produto de descamação epitelial, como queratina e cristais de colesterol. Nas pregas vocais, essa lesão costuma se localizar no plano subepitelial ou submucoso (entre o epitélio e o ligamento/músculo vocal). Acredita-se que se originam de fonotrauma repetido causando microfissuras e, posteriormente, favorecendo a invaginação do epitélio para o interior das pregas vocais. Outra possibilidade é de seja congênito em que um núcleo

de inclusão epidermóide anômalo sofreria processo de evolução cística. No entanto, Hirano et al. afirmam que o cisto pode estar localizado exclusivamente na espessura do espaço de Reinke, ou parcialmente infiltrada no ligamento vocal. O cisto pode ainda estar rompido, formando uma abertura em forma de fístula puntiforme (cisto aberto) ou em sulco (sulco-bolsa). (PASSEROTTI, 2005, p. 7)

As informações acima citadas, se complementam com as que se seguem, segundo o mesmo autor, e pode ser visualizado através da Figura 9, formando assim a base de reconhecimento das patologias pelo classificador, uma vez que as características de Nódulo e Cistos diferem em na maioria dos casos em tamanho, posição e cor:

“Clinicamente, a lesão típica do cisto vocal é descrita como uma esfera de coloração amarela esbranquiçada localizada em plano submucoso, provocando abaulamento na borda da prega vocal, que pode ser visualizada por transparência do epitélio mucoso. No entanto, a grande maioria dos cistos epidermóides se apresenta como um pequeno espessamento inflamatório da prega vocal, acompanhado de hiperemia de mucosas adjacentes e ectasia vascular”. (PASSEROTTI, 2005)

Figura 9 - Captura de uma TLE de PPVV com cisto



Fonte: <https://www.voicedoctor.net/media/structural-injury>

No último ponto de diferenciação, acima citado, quanto a coloração, pode ser aplicado, além de outras técnicas, uma limiarização de tons de cinza, aplicado a uma grade pré definida de seleção para identificação correta de patologias. Essas e outras informações terão como base o que será descrito na próxima seção, sobre extração de características e segmentação, para correta classificação das imagens.

2.3 Extração de características e segmentação

A segmentação de imagem desempenha um papel chave nos sistemas automáticos visuais e em geral a é uma das mais difíceis tarefas em processamento de imagem. Esse passo no processamento determina o eventual sucesso ou falha de toda a análise (GONZALEZ; WOODS, 2002).

Entendendo o que é exposto por Gonzalez e Woods se pode definir de maneira básica no que consiste a segmentação de imagens, segundo Hirata Jr (1997) - segmentação sendo um processo de encontrar uma partição dos pontos da imagem de forma a separar os objetos ou partes de interesse que a compõem - e segundo Roncero (2005) - A segmentação consiste da subdivisão de uma imagem em suas partes constituintes ou objetos, onde um objeto neste contexto refere-se a uma componente convexa.

Ainda segundo Roncero (2005), A segmentação pode ser vista como um problema de classificar N elementos em K regiões e elementos em uma mesma região K têm propriedades similares entre si e distintas das propriedades dos elementos de outras regiões. Assim, a segmentação também pode ser modelada como um problema de otimização combinatória, no qual procura-se uma região ótima de acordo com algum critério de similaridade entre os elementos de uma mesma região.

Na classificação de imagens, muitas técnicas apresentam limitações quando se tratam de uma análise macro (análise da imagem como um todo). Uma forma muito usada para superar essas limitações é o uso de segmentação de imagem,

anterior à fase de classificação, o que será aplicado nas imagens de TLE onde se extraem os as partes mais relevantes para a aplicação desejada, sendo a aplicação em questão a classificação de patologias de laringe.

Desta forma a segmentação constituirá a primeira, mais importante e mais complexa etapa do processamento da imagem, com já exposto por Gonzalez e Woods, acima, visto que “a percepção das partes interessantes de uma cena é um passo preliminar para o reconhecimento e a interpretação da imagem”. (RONCERO, 2005).

Para isso, será apontado a seguir as técnicas a serem utilizadas nesta dissertação, a saber, suavização, contraste (segmentação) e a ampliação de banco de dados por meio de técnicas de *Data Augmentation* (DA).

2.3.1 Contraste e suavização

Os filtros lineares de suavização são aqueles que utilizam um algoritmo com uma função linear para realizar o procedimento de suavização, e são utilizados para borramentos e redução de ruídos. O borramento é aplicado em tarefas de pré-processamento, como a remoção de pequenos detalhes da imagem antes da extração de objetos (grandes) e conexão de pequenas descontinuidades em linhas ou curvas (GONZALEZ, WOODS, 2010)

2.3.2 *Data Augmentation* e aplicação de ruído artificial

Para a classificação correta de imagens, sobretudo no que tange à área médica e demais aplicações práticas, é necessário uma alta quantidade de dados, um grande conjunto e treinamento e validação, a fim de que os resultados tenham melhor eficácia possível no auxílio diagnóstico. Apesar de todos os esforços para construir uma base de dados extensa e diversificada, vários fatores podem contribuir negativamente com esse objetivo, como coleta de dados de exames

médicos, bancos disponíveis sem restrição na internet e delimitação de área de análise. Logo, o processo de captura de dados manual pode ser exaustivo.

Quando isso não é possível, tendo um banco de dados relativamente baixo, para aumentar o desempenho de tais sistemas, o aumento artificial de dados é amplamente utilizado, aplicando-se as técnicas de Data Augmentation (DA) - Aumento de Dados. Segundo Shorten (2019), o DA na análise de dados são técnicas usadas para aumentar a quantidade de dados adicionando cópias ligeiramente modificadas de dados já existentes ou dados sintéticos recém-criados a partir de dados existentes.

A classificação de imagens com um banco de dados baixo pode gerar um overfitting e a aplicação do DA atua como um regularizador e ajuda a reduzir esse problema, que, entre outras palavras, segundo o mesmo autor, se refere ao fenômeno quando uma rede aprende uma função com variância muito alta, como modelar perfeitamente os dados de treinamento ou, segundo O'SHEA (2015) é basicamente quando uma rede não consegue aprender com eficácia devido a uma série de razões, inclusive baixo número de imagens.

Como o processo de DA está ligado com a sobreamostragem, vários ruídos artificiais podem ser aplicados abrangendo um conjunto de técnicas que aumentam o tamanho e a qualidade dos conjuntos de dados de treinamento, de modo que melhores modelos de aprendizado profundo podem ser construídos com eles, alcançando as cópias ligeiramente modificadas como descrito acima.

Três técnicas de DA serão brevemente descritas nas subseções abaixo, de acordo com o artigo *A survey on Image Data Augmentation for Deep Learning* de Connor Shorten: Injeção de ruídos, Rotação horizontal e vertical e transladação:

A. Injeção de ruído

A injeção de ruído consiste em inserir uma matriz de valores aleatórios normalmente extraídos de uma distribuição gaussiana na imagem original (Claro et

al, 2020). Adicionar ruído às imagens pode ajudar a Rede Neural Convolucional (do inglês - *Convolutional Neural Networks* - CNNs) a aprender recursos mais robustos.

B. Rotação

Os aumentos de rotação são feitos girando a imagem para a direita ou esquerda em um eixo entre 1° e 359° . A segurança dos aumentos de rotação é fortemente determinada pelo parâmetro de grau de rotação. Ligeiras rotações, como entre 1 e 20 ou - 1 a - 20, podem ser úteis em tarefas de reconhecimento de dígitos, mas à medida que o grau de rotação aumenta, o rótulo dos dados não é mais preservado após a transformação. (Claro et al, 2020).

C. Translação

Deslocar as imagens para a esquerda, direita, para cima ou para baixo pode ser uma transformação muito útil para evitar viés posicional nos dados. Por exemplo, se todas as imagens em um conjunto de dados estiverem centralizadas, o que é comum em conjuntos de dados de reconhecimento de rosto, isso exigiria que o modelo também fosse testado em imagens perfeitamente centradas. Como a imagem original é traduzida em uma direção, o espaço restante pode ser preenchido com um valor constante, como 0 s ou 255 s, ou pode ser preenchido com ruído aleatório ou gaussiano. Esse preenchimento preserva as dimensões espaciais do pós-aumento da imagem. (Claro et al, 2020).

2.4 Classificador *Bag of Features*

O *Bag of Features* (BoF) - pacote de características, em uma tradução literal, é um modelo metodológico em visão computacional com fundamentos lançados em uma técnica clássica para a classificação de texto chamada de Bag of Words (BoW). (Sebastiani, 2002)

A classificação do BoW, de acordo com Caruana e Niculescu-Mizil (2006), é feita a partir de dicionário de palavras, pré agrupadas e criando uma representação em histogramas de forma alternativa.

No contexto de visão computacional, o método possui forma similar, porém o agrupamento é feito a partir de características (features) retiradas de imagens, ao invés de texto. (CHRISTMANN, 2017)

O BoF utilizado é o que está disponível no Toolbox do MATLAB, com alterações em funções principais para criar o dicionário de recursos visuais, fazendo com que os pontos de interesse sejam detectados, assim computando os recursos descritivos de imagens das pregas vocais por meio de algoritmo de recursos robustos de velocidade e histograma de gradientes orientados.

O patch de Visão Computacional (Vision Computer) do MATLAB possibilita gerenciamento das coleções de imagens e divisão em conjuntos de treinamento e validação. É possível então construir um pacote de palavras visuais para uso na classificação de categorias de imagens.

As principais funções utilizadas para no classificador BoF, bem como para a extração de características e delimitação de espaço são as seguintes, resumidamente descritas abaixo.

A principal função é a 'bag = bagOfFeatures(imds)', que retorna um pacote de características do objeto (da imagem). O objeto de saída do pacote é gerado usando amostras da entrada (imagens de entrada). Por padrão, o vocabulário visual é criado a partir de recursos do detector de recurso robusto acelerado (speeded up robust feature - SURF), como mencionado acima, extraídos de imagens em imds.

Uma função que aperfeiçoa a extração de características é Função extratora de recurso personalizado, especificada o par separado por vírgula que consiste em 'CustomExtractor' é um identificador de função. Esta função personalizada extrai recursos do objeto bagOfFeatures de saída para aprender o vocabulário visual do objeto.

Por fim, a função 'PointSelection' consiste em um método de seleção para locais de ponto de coleta que pode ser definida como "grade" (padrão) ou "detector". Existem dois estágios para extração de recursos. No primeiro é selecionado um método para escolher as localizações dos pontos, (SURF 'Detector' ou 'Grid'), com a propriedade 'PointSelection'. A segunda etapa extrai os recursos. A extração de recursos usa um extrator SURF para ambos os métodos de seleção de pontos.

Quando se define PointSelection como 'Detector', os pontos de recurso são selecionados usando um detector de recurso robusto acelerado (SURF). Caso contrário, os pontos são selecionados em uma grade pré definida com espaçamento definido por 'GridStep'. Esta propriedade se aplica apenas quando você não está especificando um extrator customizado com a propriedade CustomExtractor.

Quando se escolhe a opção 'Grid' - em grade - tem-se o tamanho do passo da grade em pixels, especificado como o par separado por vírgulas que consiste em 'GridStep' e um vetor 1 por 2 [x y]. Esta propriedade se aplica apenas quando se define PointSelection como 'Grid' e não está especificando um extrator personalizado com a propriedade CustomExtractor. As etapas nas direções x e y definem o espaçamento de uma grade uniforme. As interseções das linhas de grade definem locais para extração de recursos.

2.5 Redes Neurais Convolucionais - CNN

Redes Neurais Artificiais (RNAs) são sistemas de processamento computacional que são fortemente inspirados pelo modo como os sistemas nervosos biológicos (como o cérebro humano) operam (O'SHEA, 2015). As RNAs são compostas principalmente por um grande número de nós computacionais interconectados (referidos como neurônios), dos quais trabalham entrelaçados em uma forma distribuída de aprender coletivamente a partir da entrada, a fim de otimizar sua saída final.

As Redes Neurais Convolucionais (CNNs), segundo O'SHEA (2015), são análogas às RNAs tradicionais na medida em que são compostos de neurônios que

se otimizam por meio do aprendizado. Cada neurônio ainda receberá uma entrada e realizará uma operação (como um escalar produto seguido por uma função não linear) - a base de inúmeras RNAs.

A única diferença notável entre CNNs e RNAs tradicionais é que CNNs são usados principalmente no campo de reconhecimento de padrões em imagens. Isso permite uma aplicação direta em problemas com dados desta natureza além de contribuir para a redução de parâmetros comparando com as RNAs.

Além dos ótimos resultados com imagens, as CNN acumulam casos de sucesso na classificação e apoio diagnóstico na área médica, com diversas contribuições como: Identificar moléculas com potencial para tratamento de câncer, diagnóstico preciso e rápido de doenças cardiovasculares por MRI (não invasivo), interpretação de imagens médicas (tomografia, raio-X) usando-se uma base mundial de imagens associadas a termos típicos de diagnósticos, detecção prematura de retinopatia diabética (uma das maiores causas de cegueira), detecção de tumores em geral etc.

As CNNs utilizadas são: AlexNet, DenseNet e ResNet. Ambas serão descritas nas subseções abaixo.

A. AlexNet

Alex Krizhevsky usou uma nova forma de processar e classificar altíssimos números de imagens em diversas classes diferentes que ganhou protagonismo a partir de 2012, despertando interesse de empresas como Google (busca de imagens), Amazon (recomendação de produtos), Instagram (sistema de busca), Facebook (marcação de usuários em fotos), etc. A técnica utilizada fez com que as CNNs alcançassem um processamento paralelo de dados. Segundo KRIZHEVSKY (2014), há duas formas óbvias de se realizar o paralelismo de CNNs, sendo a primeira em toda a dimensão do modelo, onde diferentes trabalhadores treinam diferentes partes do modelo - paralelismo de modelos - e a segunda na dimensão de

dados, onde diferentes trabalhadores treinam em diferentes exemplos de dados - paralelismo de dados.

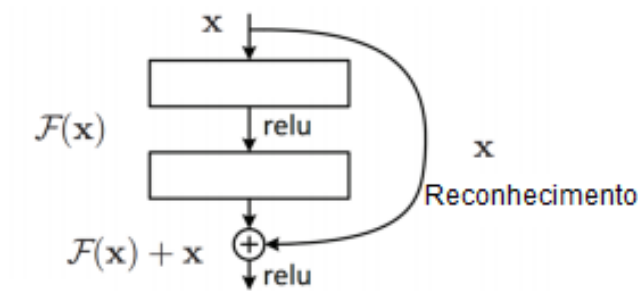
AlexNet competiu no Desafio de Reconhecimento Visual em Grande Escala ImageNet. De acordo com KRIZHEVSKY, SUTSKEVER, HINTON (2017), a rede alcançou um erro entre os 5 primeiros, de 15,3%, mais de 10,8 pontos percentuais abaixo do segundo colocado. O resultado primário do artigo original foi que a profundidade do modelo era essencial para seu alto desempenho, que era caro do ponto de vista computacional, mas tornou-se viável devido à utilização de unidades de processamento

B. ResNet

Uma rede neural residual (ResNet) (Microsoft Research Asia), segundo os próprios criador, HE, et al. (2016), é uma rede neural artificial (RNA) de um tipo que se baseia em construções conhecidas de células piramidais no córtex cerebral. As ResNets fazem isso utilizando conexões de salto ou atalhos para saltar algumas camadas, conforme ilustrado na Figura 10. Os modelos ResNet típicos são implementados com saltos de camada dupla ou tripla que contêm não linearidades (ReLU) e normalização de lote entre eles.

Na competição de Visão Computacional *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) 2016 o desempenho da ResNet chegou a 96,4% na classificação de mais de 1 milhão de imagens em 1000 categorias, com 8 unidades gráficas de processamento durante 2 a 3 semanas, utilizando 152 camadas.

Figura 10 - Salto de camadas em ResNet.



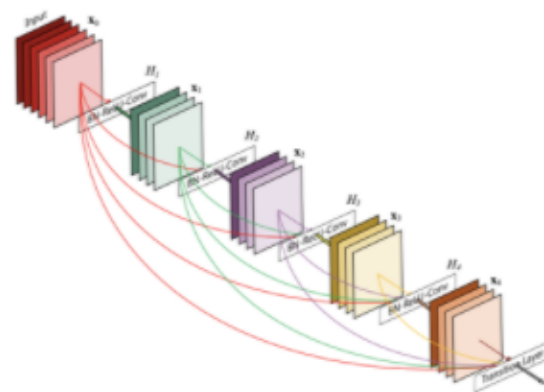
Fonte: Adaptado de Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 771.

C. DenseNet

DenseNet, segundo HUANG, Gao et al (2017), é uma das novas descobertas em redes neurais para reconhecimento visual de objetos. O DenseNet é bastante semelhante ao ResNet, com algumas diferenças fundamentais. O ResNet usa um método aditivo (+) que mescla a camada anterior (identidade) com a camada futura, enquanto o DenseNet concatena (.) A saída da camada anterior com a camada futura.

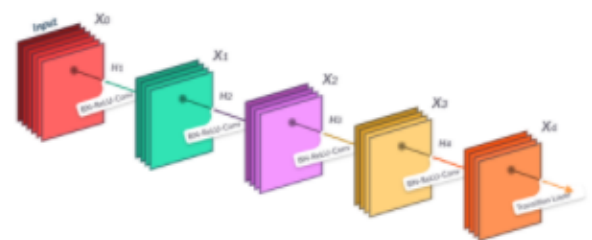
O DenseNet foi desenvolvido especificamente para melhorar a precisão diminuída causada pelo gradiente de desaparecimento em redes neurais de alto nível. Em termos mais simples, devido ao caminho mais longo entre a camada de entrada e a camada de saída, a informação desaparece antes de chegar ao seu destino. Na Figura 11 vê-se um bloco denso de 5 camadas com uma taxa de crescimento de $k = 4$ e a estrutura ResNet padrão.

Figura 11 - Estruturas em camadas das CNNs DenseNet e ResNet



Estrutura da
DenseNet

$$a^{[l]} = g([a^{[0]}, a^{[1]}, a^{[2]}, \dots, a^{[l-1]}])$$



Estrutura da ResNet

$$a^{[l]} = g(z^{[l+1]} + a^{[l]})$$

Fonte: Adaptado de Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700

2.6 Métrica de validação - Matriz de Confusão

Uma matriz de confusão compara as previsões de um modelo com o padrão verdadeiro (Silva, 2018). A diagonal da matriz representa classes que foram corretamente previstas (do inglês, True Positive - TP), enquanto elementos fora da diagonal representam indivíduos que foram classificados de forma errada (do inglês, False Positive - FP). A partir da matriz de confusão, é possível estabelecer métricas para avaliar se o algoritmo está ou não alcançando bons resultados, de acordo com a acurácia alcançada e o mínimo necessário para a aplicação. A Figura 12 mostra um exemplo de matriz de confusão que será utilizadas nas Seções seguintes para fazer os cálculos das métricas

A acurácia (Ac) é calculada de acordo com a equação abaixo (1):

$$Ac (\%) = \frac{\Sigma TP}{\Sigma FP + \Sigma TP} 100 \quad (1)$$

Logo, o erro válido (Er) é calculado pela equação (2):

$$Er (\%) = (1 - Ac)100$$

Fig 12 - Exemplo de matriz de confusão

Real	Nódulos	23 (TP1)	0 (FP1)	0 (FP2)
	Cistos	0 (FP3)	25 (TP2)	2 (FP4)
	Normais	0 (FP5)	0 (FP6)	25 (TP3)
		Nódulos	Cistos	Normais
		Predito		

Fonte: O autor

2.7 Trabalhos correlatos

Os mais recentes estudos para diagnósticos e classificação de patologias envolvem redes neurais, inteligência artificial, transformada *wavelet* e *support vector machine*, além de patentes para diagnóstico desenvolvidas na última década (ARROYAVE; BONILLA; TREJOS, 2012). Na literatura há poucos trabalhos relacionados à auxílio diagnóstico de laringe por meio de visão computacional, área da Inteligência Artificial e aplicação direta de Redes Neurais Artificiais, então a seguir serão expostos os trabalhos correlatos que auxiliam o entendimento da extração de características para fins de classificação.

Muitos artigos de visão computacional nesta área de classificação de imagens podem dar suporte a esse trabalho. Alguns deles, não exatamente na área de ciências da saúde, mas em áreas como ciências ambientais, como o exemplo do projeto “Estudo de descritores de texturas e cores para a classificação automática de imagens no contexto das ciências ambientais” de Pierre Pereira Prado (2018). Neste trabalho é proposta a classificação de texturas, para identificação de

espécies arbóreas e imagens aéreas de acordo com dois bancos de imagens – BARKTEX e UC MERCED – por meio de Rede Neural ARTMAP, simulado no software MATLAB. De acordo com Prado (2018), a pesquisa teve fins de inventário de espécies arbóreas e de usos e cobertura do solo por inspeção de imagens, tendo-se obtido descritores que permitiram taxas de acerto entre 87 e 100% em um determinado caso e entre 70 e 74% em outro.

Relacionando a áreas de ciências da saúde e classificação de imagens, o trabalho de Mozart Lemos de Siqueira intitulado “Reconhecimento automático de padrões em imagens ecocardiográficas” propôs um método hierárquico para localização automática de estruturas cardíacas – coração e cavidades, usando o software MATLAB para os principais scripts de programação, apesar de iniciar o trabalho em linguagem C. Foram realizados testes quanto à invariância, à rotação das estruturas e consequentemente das imagens – “desde os primeiros resultados, ficou claro que o método era capaz de suportar a rotação de estruturas” (Siqueira, 2010). Isso garante o processo de reconhecimento de padrões de imagens independente da posição final da mesma a ser processada.

O artigo “Análise vocal e laríngea na hipótese diagnóstica de nódulos e cistos” de Lívia Fernandes Barata, citado anteriormente, dá apoio ao presente projeto com informações a serem usadas como apoio e instrução dos algoritmos computacionais. Em uma análise e classificação computacional de patologias na laringe como nódulos e cistos, por exemplo, todos os nódulos foram simétricos em localização (100%) e a maioria foi simétrica em tamanho e os cistos tendem à simetria em localização (75%), mas não foram simétricos em tamanho (100%) (BARATA, 2010).

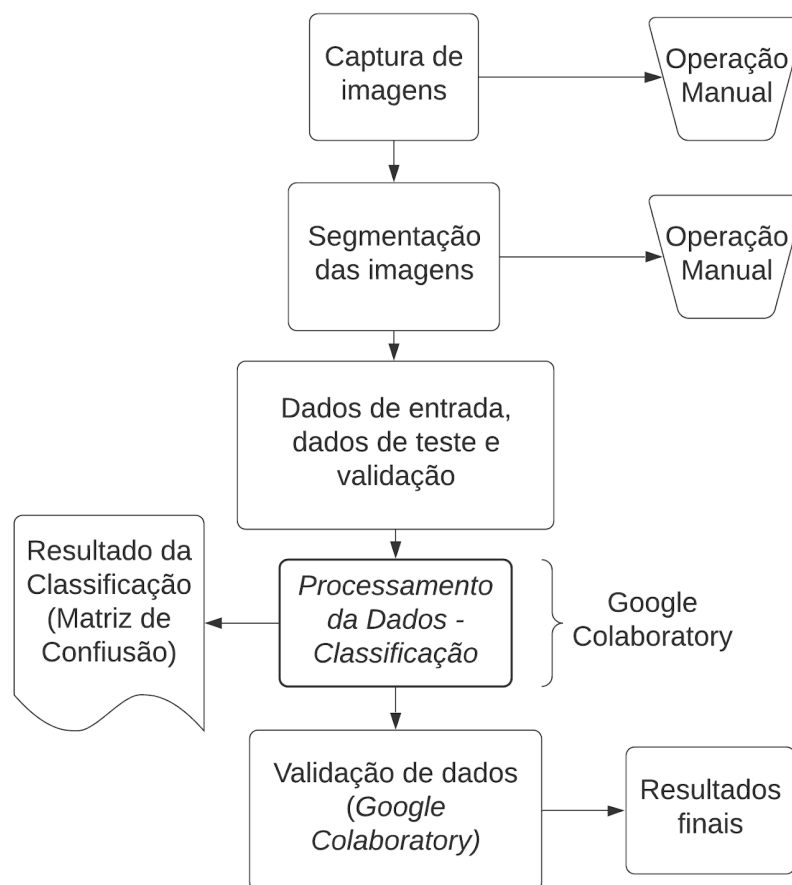
Os trabalhos pesquisados de identificação de patologias de laringe trazem uma versão de pesquisa acústica, com aplicação da Transformada de Wavelet para análise de sinais digitais – os áudios dos exames. Dentre os trabalhos pesquisados destacam-se os de Crovato (2007) e Barizão (2018) que, após as simulações e verificações de resultados, alcançaram a robustez e eficiência exigidas para um processo de bioengenharia.

As técnicas típicas de diagnóstico de patologias de voz são consideradas invasivas pois utilizam instrumentos introduzidos dentro da boca ou nariz até obter uma visualização das pregas vocais. Essas técnicas causam desconforto no paciente que necessitam de anestésico local para evitar a reação de náusea, além de utilizar equipamentos sofisticados e de alto custo, mas permitem a visualização direta da laringe e das pregas vocais. Algumas dessas técnicas básicas são a laringoscopia indireta por espelho circular, laringoscopia direta tradicional e laringoscopia indireta com fibra ótica (BEHLAU, 2001).

3 MATERIAIS E MÉTODOS

Para atingir os objetivos propostos propõe-se o emprego de abordagens de processamento de imagens digitais e segmentação baseado no que é proposto em PINHEIRO (2017), KUO (2013) e RONCERO (2005), técnicas de classificação com o classificador *Bag Of Features*, baseando-se em LEEZA e FAROOQ (2019) e em CHRISTMANN (2017), aplicação de Redes Neurais Convolucionais segundo HE (2016), KRIZHEVSKY (2017) e HUANG (2017) além de testes para validação do método, utilizando toolbox do *MATLAB* e Google Colaboratory conforme é ilustrado no esquema da Figura 13, abaixo.

Figura 13: Fluxograma metodológico



fonte: O autor

3.1 Delimitação de uso de banco de dados

Uma etapa anterior ao processo de captura de imagens foi realizada, executando a verificação de dados com auxílio de equipe de saúde médica e fonoaudiológica. Delimitou-se a análise de elementos de videolaringoscopias classificadas como Grau I de visualização da laringe - glote bem visível - de acordo com a tabela de categorias descrita por Cormack e Lehane, em 1984, como demonstra Melhado, V., & Fortuna, A. (2004, apud AIDOS, 2017, p.41). As TLE e imagens selecionados foram todas de pregas vocais femininas com diagnóstico normal (saudável) e patologias benignas (cistos e nódulos)

As imagens são extraídas de TLE, disponíveis no bancos de dados com licenças abertas ao uso sem fins comerciais VoiceDoctor¹, divididas em 3 classes, a saber *Normal vocal cord*, *Behavioral injury* e *Structural injury*- ou imagens já capturadas e devidamente rotuladas para cada patologia ou com diagnóstico saudável da Clínica Otorrino Cóser²

Para o banco VoiceDoctor em que se tem a maior parte dos dados em vídeo foi realizado um corte a partir de captura de tela nos instantes em que a se tinha melhor visualização da prega vocal, saudável ou patológica, de maneira manual, já que para o caso analisado quando a segmentação de imagens envolvia processos automáticos de seleção e corte, era necessário controlar o ambiente onde se retira a imagem, tornando-se inviável, se tratando de imagens capturadas pelos vídeos de telelaringoscopias ou banco de imagem já disponível.

Outros fatores que levaram à escolha do processo manual, para a seleção foram:

- Controle da luminosidade: Conforme a aplicação envolvida a existência de sombras tendem a dar uma falsa impressão acerca do tamanho real da região a ser segmentada.

¹ <https://www.voicedoctor.net/media>

² <https://www.clinicacoser.com/veja-fotos-de/laringoscopia/>.

- Bordas das regiões são muitas vezes irregulares e imprecisas.
- Precisão da técnica aplicada depende da qualidade da distinção entre os diferentes elementos da imagem.
- Escolha da melhor estratégia e adequação à aplicação que se deseja

Apesar de toda análise médica de laringe contar com o processo de um exame de telelaringoscopia, como citado na introdução, em decorrência de fechamento parcial de centros médicos, centros de pesquisas e demais atividades vinculadas à pesquisa, devido à Pandemia do COVID-19, e não alimentação dos bancos de dados, foram analisadas 60 amostras, dos dois bancos de dados, das quais 45 foram utilizadas. As demais 15 amostras não se adequaram a aplicação por não gerar visualização completa das patologias ou pelo fato de não se enquadrar no Grau 1 de abertura da glote.

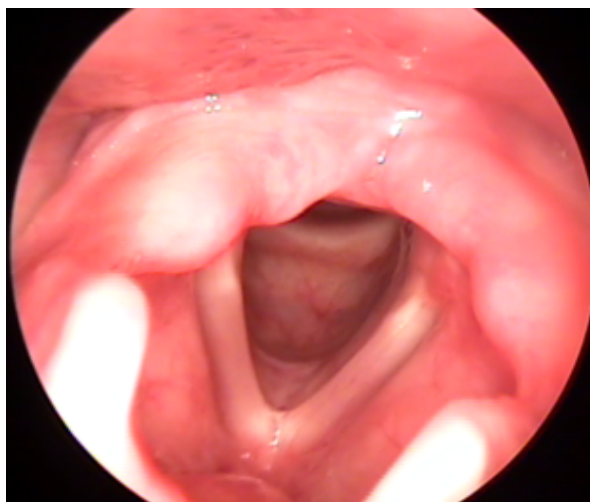
Sendo assim, as 45 imagens se dividem da seguinte forma: 17 com diagnóstico saudável (PPVV-S) e 17 com diagnóstico de cistos e 11 com diagnóstico de nódulos. Devido ao número de amostras com presença de patologia ser baixo, haja visto que a quantidade total de amostras engloba duas categorias de patologia, aplicou-se técnicas de Data Augmentation para obter 100 imagens de cada tipo - normais, cistos e nódulos, totalizando 300 imagens no banco de dados, entre testes e validação.

Para aplicações de classificações de imagens o número exato para treinamento se dá, muitas vezes de maneira empírica, porém, sabendo que quanto mais imagens se têm maior é a chance de uma classificação mais correta, visto que se trata de aprendizagem máquina.

Essas imagens foram selecionadas garantindo a identificação para uma das duas categorias estabelecidas - PPVV-S ou PPVV-NS - conforme a Figura 14, uma captura de uma prega vocal normal, saudável, sem patologia.

Como descrito acima, todos os vídeos e as imagens já estavam rotuladas para cada condição dos três grupos - normais, nódulos e cistos. Sem o diagnóstico prévio não é possível treinar corretamente os classificadores e para tal execução contou-se com apoio de equipe especializada da Universidade de São Paulo (USP), do departamento de Fonoaudiologia da Faculdade de Odontologia de Bauru.

Figura 14 - Captura de tela da telelaringoscopia



Fonte: Adaptado de
www.voicedoctor.net/media/normal-vocal-cord/flexible-videoscope-exam/normal-young-female-vocal-cords

3.2 Segmentação das Imagens

A partir das imagens segmentadas, o banco de dados aumentado e a divisão de grupos de teste e validação, seguiu-se com a análise computacional para realizar a alimentação de entrada do classificador e das CNN, para treinamento e definição de categorias. Logo, para realizar análise computacional de imagens, alguns passos importantes foram seguidos. Antes extrair qualquer tipo de informação a partir de uma imagem, foi necessário um pré-processamento dos dados, além das simples capturas, de modo a corrigir eventuais desequilíbrios na sua iluminação, reduzir o ruído e/ou realçar determinadas características como, por exemplo, os contornos e as bordas das pregas vocais.

O recorte da imagem foi o primeiro passo de processamento ainda realizado de forma manual nos softwares *PhotoShop®* e *ImageJ®*. Esse redimensionamento foi realizado de maneira unitária para cada figura para garantir a visualização somente do formato anatômico em “V” que se forma quando as pregas vocais estão abertas. No mesmo software (ImageJ) foi aplicado o primeiro filtro, de escala cinza com módulo de 2,5 de intensidade para, posteriormente, obter os histogramas da imagem no classificador. A figura 15 demonstra o resultado do primeiro processamento manual.

Figura 15- Recorte e filtro escala em cinza



Fonte: O autor

Todas as imagens de ambas as categorias - Normais, Cistos e Nódulos foram submetidas a esse procedimento.

De forma automatizada, após esse processo realizou-se o redimensionamento de todas as imagens no software *MATLAB*, para que os pontos a serem selecionados por meio do classificador de imagens, usando *BoF* e as *CNN* fossem melhores identificados. O tamanho geral da imagem passou a variar entre 220 a 270 pixels de largura e 290 a 350 pixels de comprimento. Imagens pequenas e de fácil processamento.

Assim como descrito na seção acima, somente 45 imagens (contando as duas categorias e incluindo dois tipos de anomalias) o processamento e classificação

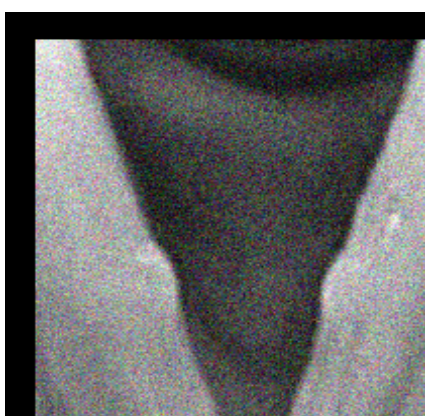
ficariam abaixo do esperado e até mesmo enviesado. Para resolver esse problema, utilizou-se técnicas de DA (aumento de dados) para se alcançar o número de 100 imagens para cada tipo de anomalia e prega vocal salvável. Os ruídos aplicados foram os de rotação, com demonstrado na Figura 16, de transladação, como ilustrado na Figura 17 e por fim o ruído gaussiano como esboçado na Figura 18.

Figura 16 - Prega vocal com cisto submetida à rotação



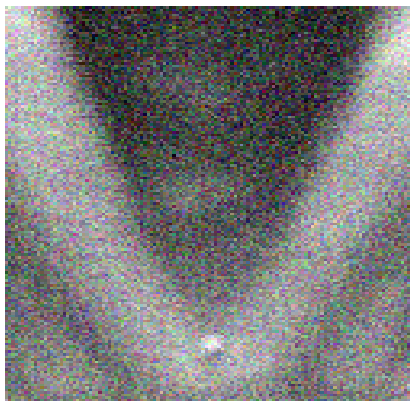
Fonte: O autor

Figura 17 - Prega vocal com nódulo submetida à translação



Fonte: O autor

Figura 18 - Prega vocal normal/saudável submetida a um filtro gaussiano



Fonte: O autor

3.3 Divisão dos dados - teste e validação

As 300 imagens (100 - normais / 100 - nódulos / 100 cistos) foram divididas em dois grupos para serem realizados os testes, tanto do classificador BoF quanto das CNNs, e a validação das ferramentas de classificação. De cada classe, 75% das imagens ficaram no grupo de teste e 25% no grupo de validação das ferramentas.

Os softwares utilizados para o teste e a validação serão detalhados no próximo capítulo de resultados e discussão.

4 RESULTADOS E DISCUSSÃO

As 75 imagens de cada classe foram processadas diversas vezes pelo classificador Bag of Features e as CNNs até se obter, empiricamente, um número de épocas ideal para testes, utilizado como critério de parada. Por se tratar de imagens com baixo número de pixels, a quantidade de interações determinada para a classificação foi de 19 épocas em cada teste, fazendo com que se obtenha melhores resultados de Perda de Treino, Perda Válida e Acurácia para selecionar a de maior contribuição para ser carregada no pacote de visão computacional do MATLAB - *Image Batch Processor* e *Google Colaboratory*.

Após a definição do número de épocas (interações) definiu-se 5 treinos para cada uma das CNNs e o mesmo para o BoF.

4.1 Resultados - Classificador *Bag of Features*

Os resultados da detecção de patologias nas imagens no classificador BoF são mais simples e objetivos, sendo expostos na Tabela 2 a seguir.

Tabela 2 - Resultados da classificação BoF

Conhecido	Predito		
	Nódulos	Cistos	Normais
Nódulos	94%	0	6%
Cistos	0	94%	6%
Normais	0	0	100%
Acurácia média 96%			

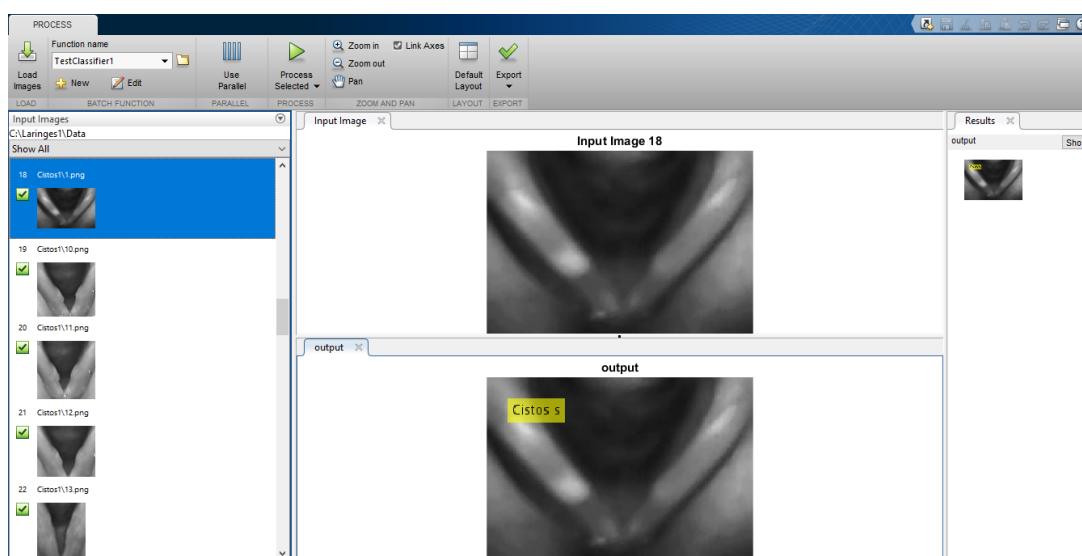
Fonte: O autor

Por meio da Tabela 2 são demonstrados os resultados em porcentagem a respeito do acerto na classificação de imagem em comparação às três classes. 94%

das imagens com rótulos de nódulos foram classificadas corretamente como imagens com presença de patologia - nódulos e os outros 6% foram erroneamente classificados como PPVVs normais. O mesmo aconteceu com as imagens rotuladas com presença de cisto, com os mesmos dados. Todas as imagens rotuladas como Normais foram corretamente classificadas como PPVVs saudáveis, alcançando classificação de 100%. A acurácia média, ou seja, precisão de classificação foi de 96% de acerto.

A acurácia do classificador BoF demonstra altos níveis de acerto e sua validação pode ser demonstrada pelo aplicativo *Image Batch Processor* (IBP) do *toolbox* de visão computacional do MATLAB. As imagens de validação são inseridas no IBP e rotuladas de acordo com o alcance dos testes. Os acertos e erros são demonstrados em rótulos para cada uma das imagens. A interface do IBP pode ser visualizada na Figura 19 a seguir.

Figura 19 - *Image Batch Processor* - Validação Classificador *Bag of Features*



Fonte: O autor

4.2 Resultados - CNNs

As três CNNs escolhidas para os testes foram redes que se destacaram na competição de visão computacional nos últimos anos. Os modelos de cada uma são: DenseNet121, ResNet34 e AlexNet. Os resultados trazem maior detalhamento de informações contando com perdas de treinamento e perda válida e uma matriz de confusão numérica para visualização dos resultados além dos dados mais custosos à rede para serem classificados..

A Tabela 3 demonstra os resultados das interações da DenseNet121.

Tabela 3 - Interações DenseNet121

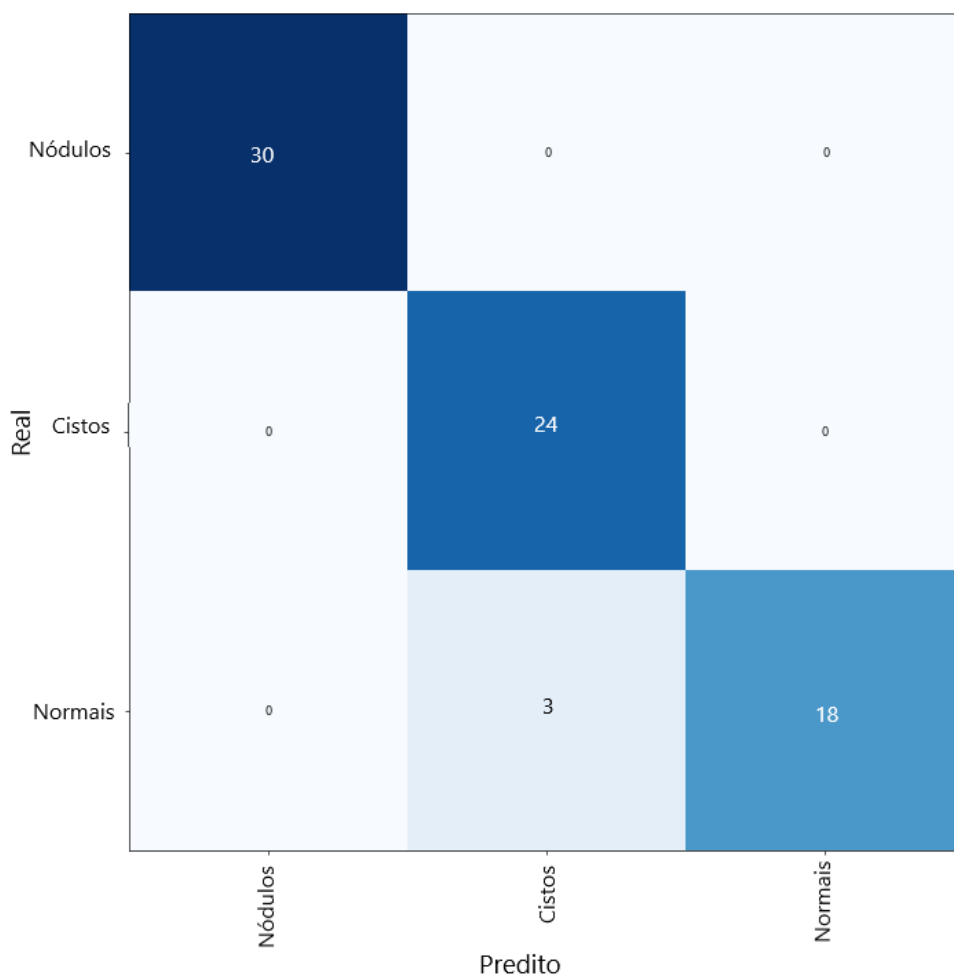
Nº de Épocas	Perda de Treinamento	Perda Válida	Acurácia	Tempo
0	4,53%	3,59%	98,67%	00:03
1	3,42%	4,41%	98,67%	00:03
2	3,68%	7,14%	97,33%	00:03
3	4,95%	1,26%	97,33%	00:03
4	4,23%	1,28%	96,00%	00:03
5	3,76%	1,58%	96,00%	00:03
6	3,74%	2,33%	94,67%	00:03
7	3,40%	1,89%	93,33%	00:03
8	3,43%	2,28%	96,00%	00:03
9	3,50%	2,57%	96,00%	00:03
10	3,22%	1,85%	96,00%	00:03
11	2,99%	1,35%	96,00%	00:03
12	2,87%	1,31%	96,00%	00:03
13	2,77%	1,09%	96,00%	00:03
14	2,64%	1,02%	97,33%	00:03
15	2,58%	9,76%	96,00%	00:03
16	2,62%	9,28%	96,00%	00:03
17	2,49%	8,94%	96,00%	00:03
18	2,44%	9,30%	96,00%	00:03

Fonte: O autor

Como já descrito acima, as imagens extraídas de TLE e dos bancos de dados utilizados têm baixo número de pixels, fazendo com que o tempo de resposta a cada interação da CNN seja somente de três segundos.

A Acurácia considerada é a da última interação, alcançando 96% de precisão. A perda válida alcançou 9,3% e isso pode ser melhor ilustrado na matriz de confusão da Figura 20, quando os dados passam pelo processo de validação e os erros dos testes aparecem de forma quantitativa, onde a classificação apesar do alto nível de acurácia tem uma variação grande no momento de classificar as três classes.

Figura 20 - Matriz de Confusão DenseNet121

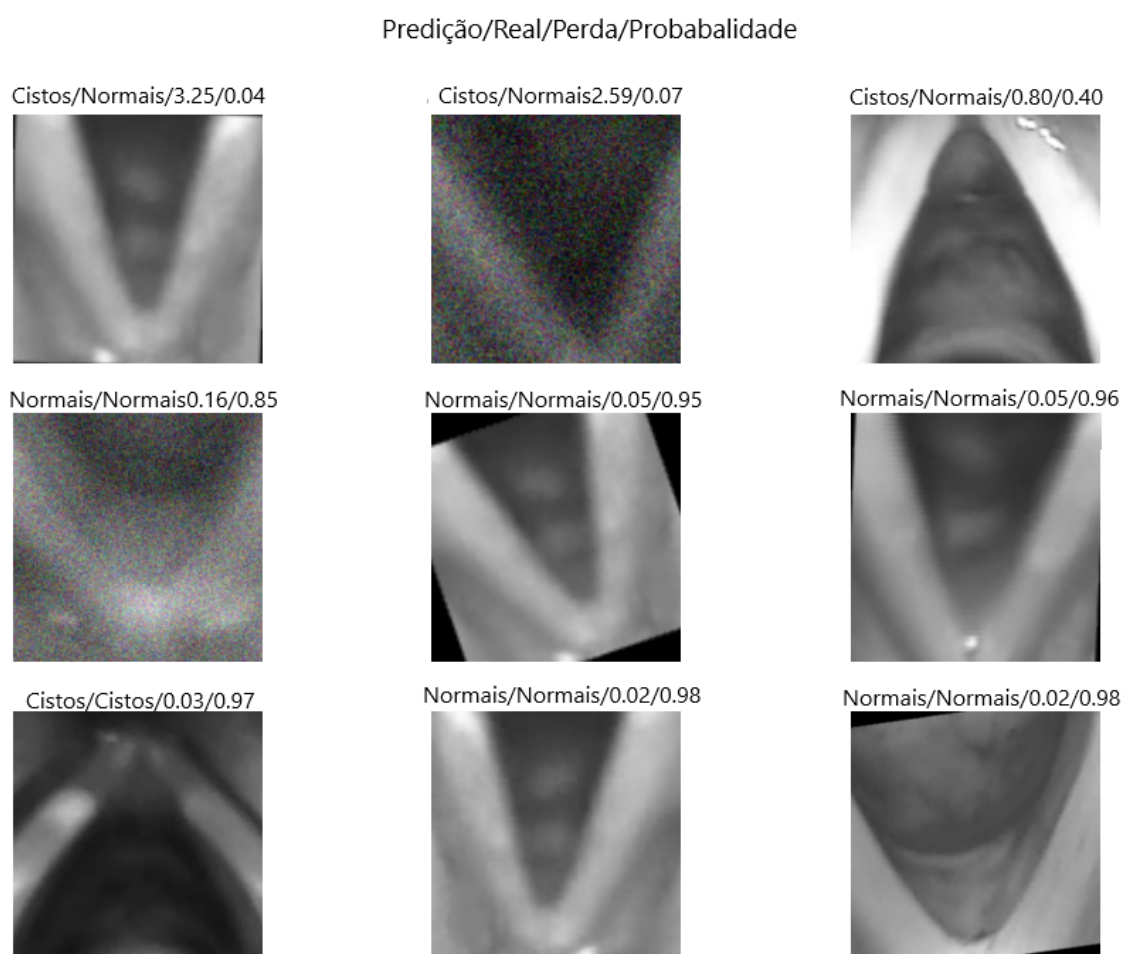


Fonte: O autor

Analisando os dados percebe-se os erros acima comentados. De 25 imagens a serem classificadas como “nódulo” foram acrescentadas erroneamente a essa classe mais 5 dados. Quanto à segunda classe, 96% das imagens rotuladas como “cisto” (24 imagens) foram corretamente classificadas. Por fim, somente 18 das 25 imagens (72%) rotuladas como “normais” foram corretamente classificadas.

Nessa validação de dados, as imagens que mais remeteram perdas à CNN DenseNet foram as que estão ilustradas na Figura 21.

Figura 21 - Maiores perdas de treinamento DenseNet121



Fonte: O autor

A tabela 4 demonstra os resultados das interações de treinamento da AlexNet.

Tabela 4 - Interações AlexNet

Nº de Épocas	Perda de Treinamento	Perda Válida	Acurácia	Tempo
0	1,01%	3,65%	88,00%	00:03
1	1,35%	3,57%	88,00%	00:03
2	1,29%	3,36%	88,00%	00:03
3	1,20%	3,37%	88,00%	00:03
4	1,27%	3,09%	90,67%	00:03
5	1,35%	2,54%	89,33%	00:03
6	1,35%	2,13%	94,67%	00:03
7	1,38%	2,04%	93,33%	00:03
8	1,45%	2,94%	86,67%	00:03
9	1,50%	1,82%	93,33%	00:03
10	1,50%	1,81%	93,33%	00:03
11	1,52%	2,35%	93,33%	00:03
12	1,53%	2,90%	90,67%	00:03
13	1,49%	2,66%	90,67%	00:03
14	1,45%	2,37%	92,00%	00:03
15	1,45%	2,35%	92,00%	00:03
16	1,47%	2,31%	92,00%	00:03
17	1,50%	2,23%	92,00%	00:03
18	1,47%	2,25%	92,00%	00:03

Fonte: O autor

A Acurácia da CNN AlexNet foi 92% de precisão. A perda válida alcançou 2,25% trazendo mais acertos na classificação que serão demonstrados na matriz de confusão da Figura 22.

Figura 22 - Matriz de confusão AlexNet

Real	Nódulos	22	0	0
	Cistos	0	22	2
	Normais	0	4	25
		Nódulos	Cistos	Normais
		Predito		

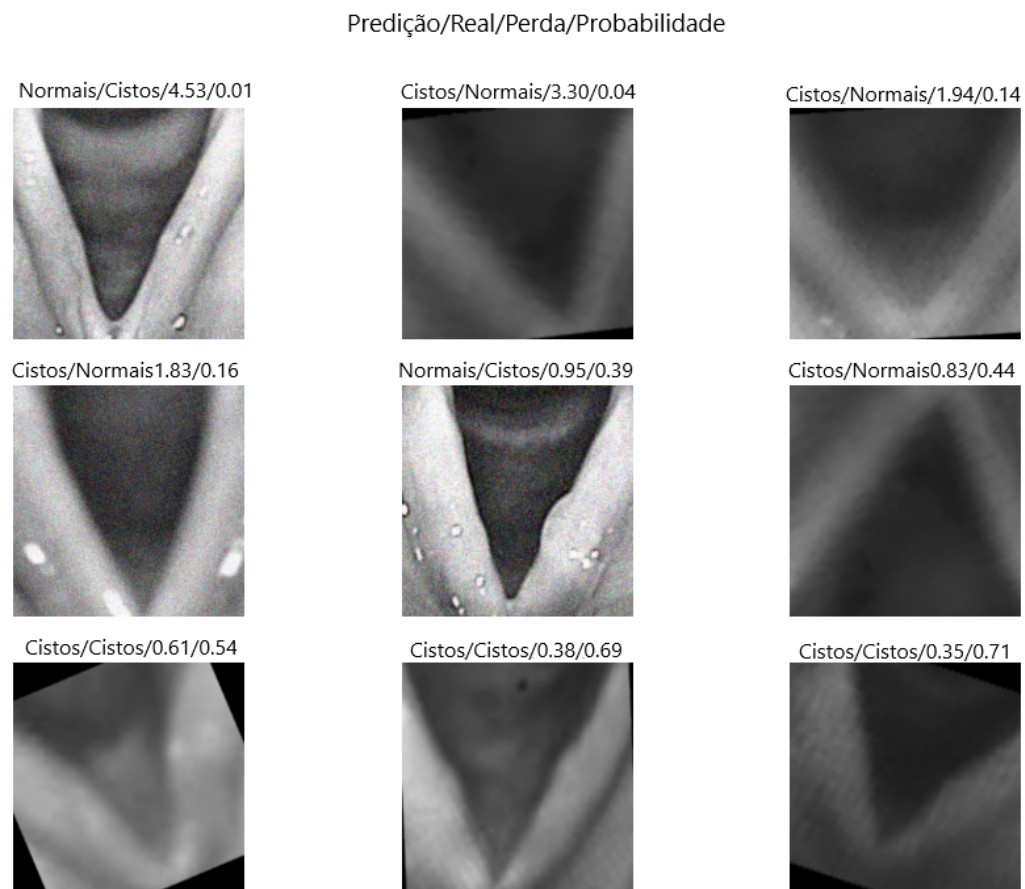
Fonte: O autor

A única classe corretamente classificada nesta aplicação foi a de imagens de PPVVs normais. De 25 imagens a serem classificadas como “nódulo”, 22 foram incluídas nessa classe, totalizando um acerto de 88%. Quanto à segunda classe, de imagens de Cistos, o mesmo ocorreu. A acurácia é a média dos valores da diagonal da matriz de confusão.

Nessa validação de dados, as imagens que mais remeteram perdas ou dificuldade na classificação à CNN AlexNet foram as que estão ilustradas na Figura 23.

A CNNs que obteve melhor desempenho foi a Resnet34. Na tabela 5, abaixo, são demonstrados os dados de treinamento.

Figura 23 - Maiores perdas de treinamento AlexNet



Fonte: O Autor

Tabela 5 - Interações ResNet34

Nº de Épocas	Perda de Treinamento	Perda Válida	Acurácia	Tempo
0	1,94%	1,28%	38,67%	00:32
1	1,56%	9,30%	61,33%	00:06
2	1,32%	6,65%	68,00%	00:03
3	1,08%	4,69%	74,67%	00:03
4	9,00%	3,00%	86,67%	00:03
5	7,97%	2,32%	89,33%	00:03
6	7,20%	2,16%	89,33%	00:03
7	6,64%	1,98%	92,00%	00:03
8	5,93%	1,63%	94,67%	00:03
9	5,43%	1,28%	96,00%	00:03
10	5,06%	1,13%	96,00%	00:03

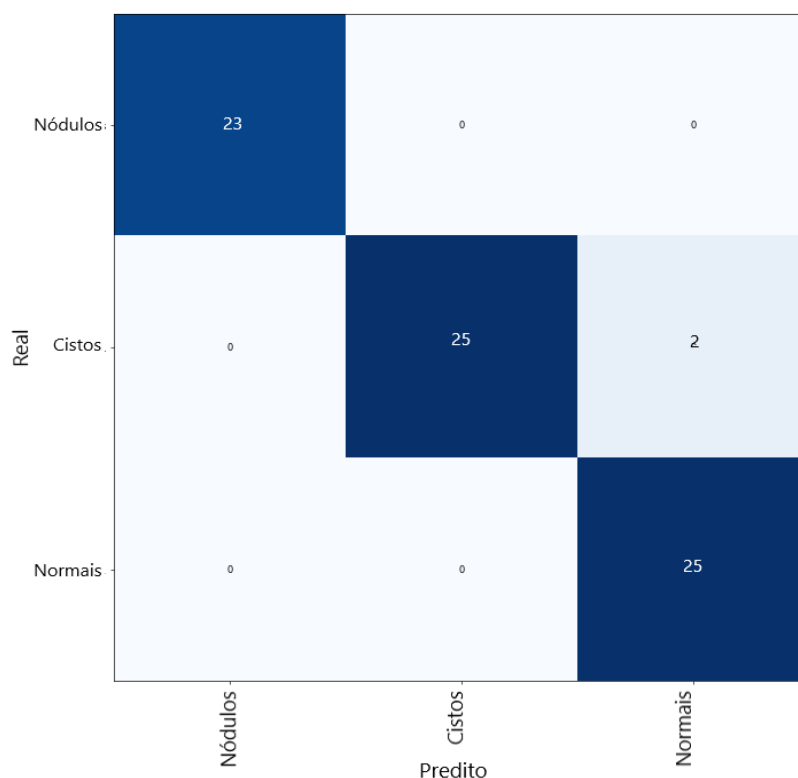
11	4,69%	8,56%	97,33%	00:03
12	4,36%	7,57%	97,33%	00:03
13	4,01%	6,72%	98,67%	00:03
14	3,74%	6,22%	98,67%	00:03
15	3,55%	5,91%	98,67%	00:03
16	3,17%	6,27%	97,33%	00:03
17	2,98%	3,02%	97,33%	00:03
18	2,55%	2,67%	97,33%	00:03

Fonte: O Autor

Com os menores erros de treinamento e perda de validação a ResNet34 alcançou 97,33% de acurácia e uma classificação de somente 2 imagens erradas, conforme ilustra a Figura 24, na matriz de confusão.

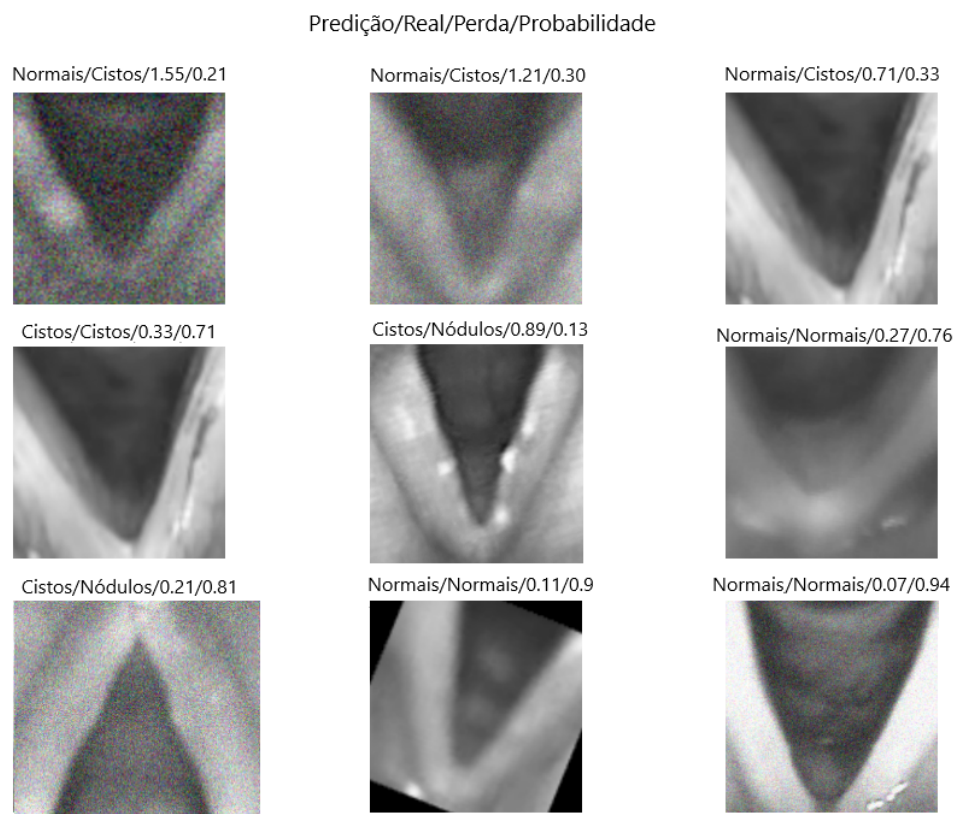
As imagens com maiores perdas na classificação da ResNet24 são demonstradas na Figura 25.

Figura 24 - Matriz de confusão ResNet34



Fonte: O Autor

Figura 25 - Maiores perdas de treinamento ResNet



Fonte: O Autor

4.3 Discussão

Algumas vantagens do BoF na aplicação para detecção de patologia e classificação em categorias distintas é que não há necessidade de carregar as imagens direto no software, somente transformando-as em uma matriz de dados que serão analisados e mais facilmente processados. Além disso, ainda destaca-se o seu uso pela conexão do *Toolbox* no MATLAB com o aplicativo *Image Batch Processor*, sendo uma interface prática para validação tanto da equipe envolvida com a programação e execução do classificador como também uma possível equipe médica e de fonoaudiólogos para executar validações para auxílio diagnóstico. Como desvantagem pode destacar-se que, apesar de bons resultados - acurácia de 96% na classificação de três categorias de imagens de características similares, outros fatores do classificador BoF precisam ser analisados por métricas e ferramentas específicas, visto que os treinos retornam somente número de

características extraídas, interações e acurácia. Sendo assim, ainda não se pode afirmar uma classificação exata, pois, segundo HEMMERLING (2018) em sistemas voltados às áreas médicas não importa somente a taxa de acerto; é necessário levar em conta a sensibilidade do sistema. Outra desvantagem é que o uso das ferramentas PointSelec e Grid limitam a aplicação para imagens mais padronizadas, o que torna difícil a aplicação nas imagens advindas de TLE.

Pretende-se continuar usando esse classificador mas recorrendo à novas funcionalidades e aplicações, visto que a classificação ainda é feita por extração macro de características. Esse é um ponto a ser melhor analisado, visto que ainda não se recorreu fortemente ao uso de ferramentas mais complexas do classificador segundo NOWAK, et al (p 491, 2006) de descrição, caracterização, distribuições resultantes das imagens e como classificar imagens com base no resultado. Apesar de não ter sido ilustrado, há a possibilidade também de recorrer à análise dos histogramas de uma amostra de imagens.

Os experimentos com o BoF demonstraram alcançaram acurácia próximo aos que são esperados na área médica - segundo Spatti (2018), com robustez maior que 95%.

Em relação às aplicações das CNNs, obteve-se resultados mais robustos, mediante a aplicação de redes atuais e com vasta aplicações em diversas áreas, inclusive com vários sucessos quando aplicadas à reconhecimento de padrões e classificações nas áreas médicas. Por se tratar de redes indicadas para a análise de imagens o resultado obtido alcança níveis esperados e auxilia a alcançar o objetivo proposto da dissertação.

Como visto no capítulo de resultados preliminares, cada uma das CNNs apresentaram níveis elevados de acerto com uma média geral, entre as três, de 95,11%. Taxa de acerto nos testes e validações considerada mínima para aplicações de engenharia na área médica.

A ResNet34 destacou-se com a maior acurácia de 97,33% e com sua baixa taxa de perda válida, fazendo com que a matriz de confusão fosse demonstrada como visto na Figura 25.

Assim como na aplicação dos classificador BoF pretende-se aprimorar a sensibilidade, principalmente da ResNet34 para alcançar melhores valores e propor uma interface de validação mais usual para aplicações cotidianas na área da saúde.

5 CONCLUSÃO

Patologias de laringe podem afetar diretamente as relações sociais trazendo dificuldades na comunicação. Diversas patologias podem ser diagnosticadas nas pregas vocais, mas em especial cistos e nódulos podem ser confundidos na inspeção visual e gerar um desgaste tanto na equipe médica como no paciente quando erroneamente se submete um caso à cirurgia que poderia ter sido resolvido por terapia fonoaudiológica.

Para auxiliar na inspeção visual e no diagnóstico laríngeo, Neste trabalho foi proposto a identificação e classificação de patologias benignas de laringe por meio do processamento de imagens a partir de um classificador inteligente - Bag of Features e redes neurais convolucionais AlexNet, DenseNet e ResNet34 disponíveis no *Google Colaboratory*. As imagens rotuladas foram obtidas a partir de banco de dados específico e o número de imagens foi ampliado por meio da técnica de *data augmentation*.

A classificação que alcançou maior desempenho foi a CNN ResNet34 com acurácia de 97,33% e erro de validação de 2,67%.

Neste cenário, as principais contribuições deste trabalho foram:

- A possibilidade de identificação e classificação de patologias de laringe com alto grau de confiabilidade a partir de imagens médicas;
- Eliminação da subjetividade do processo manual de identificação e classificação de patologias de laringe, bem como redução no tempo de diagnóstico.
- Disponibilidade de mais uma abordagem para o diagnóstico de patologias de laringe.

Por fim os possíveis trabalhos futuros são listados abaixo:

- Buscar-se-á maior banco de dados e outras redes convulsionais a fim de reforçar o aprendizado de redes e assim, aumentar a confiabilidade destes sistemas auxílio de diagnósticos;

- Explorar-se-á as possibilidades de automatização no processo de seleção de imagens usando os próprios algoritmos para a seleção de imagens, redimensionamento, segmentação e outros processos realizados manualmente e
- Investigar-se-á a aplicação da Rede Neural Convolucional RestNet34, bem como outras redes e classificadores, em um sistema de auxílio diagnóstico com reconhecimento da patologia diretamente em vídeo, possibilitando inferência dos resultados de classificação no momento do exame laringológico.

REFERÊNCIAS

AIDOS, João Rafael Meneses da Silva. **História e desenvolvimento da laringoscopia: revisão bibliográfica**. 2017. Tese de Doutorado.

ALMEIDA, Náthalee Cavalcanti de. **Sistema inteligente para diagnóstico de patologias na laringe utilizando máquinas de vetor de suporte**. 2010.

BARATA, Livia Fernandes et al. Análise vocal e laríngea na hipótese diagnóstica de nódulos e cistos Vocal and laryngeal analyses in diagnostic hypotheses of nodules and cysts. **Revista da Sociedade Brasileira de Fonoaudiologia**, v. 15, n. 3, p. 349-54, 2010.

Behlau, M., Azevedo, R., & Madazio, G. (2001). **Anatomia da Laringe e Revista Brasileira de Otorrinolaringologia**, v. 74, n. 6, p. 869-875, 2008.

Behlau M. Voz: **o livro do especialista**. Rio de Janeiro: Revinter; 2001 p.86-176.

CLARO, Maíla et al. Utilização de Técnicas de Data Augmentation em Imagens: Teoria e Prática. **Sociedade Brasileira de Computação**, 2020.

CARUANA, Rich; NICULESCU-MIZIL, Alexandru. An empirical comparison of supervised learning algorithms. In: Proceedings of the **23rd international conference on Machine learning**. 2006. p. 161-168.

CHRISTMANN, Guilherme HG et al. **Implementação de um modelo Bag of Features para classificação de frutas e vegetais**. Universidade Federal de Santa Maria (UFSM) - Santa Maria - RS - Brasil., 2017. Disponível em: <<http://rodrigoguerra.com/wp-content/uploads/2019/08/30228-168039-2-PB.pdf>> Acesso em 01 de setembro de 2020.

CIELO, Carla Aparecida et al. Lesões organofuncionais do tipo nódulos, pólipos e edema de Reinke. **Revista CEFAC**, v. 13, n. 4, p. 735-748, 2011.

DO BRASIL, Osíris de Oliveira Camponês; PERAZZO, Paulo; YAMASAKI, Rosiane. **Nódulos vocais—diagnóstico e prevenção**. Revista Científica, v. 1, n. 1, p. 21-5, 2012.

FARÍAS, Fga Patricia. Tópicos iberoamericanos en voz cantada y hablada Manuel García. Su influencia en el nacimiento de la laringología (Vol. 1). **Comité científico**. 2012.

Fisiologia da Produção Vocal. In Revinter (Ed.), O Livro do Especialista (Vol. Gonzalez, R. C. and Woods, R. E. **Digital Image Processing**. Prentice-Hall, Upper Saddle River, NJ, second ed., 2002.

GONZALEZ, Rafael C. WOODS, Richard C. **Processamento Digital de Imagens**. 3ªed. São Paulo: Pearson Prentice Hall, 2010

HE, Kaiming et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 770-778.

HEMMERLING, Daria. **Voice pathology distinction using autoassociative neural networks**. In: 25th European signal processing conference (EUSIPCO). IEEE, 2017. p. 1844-1847.

HIRATA JR, Roberto. **Segmentação de imagens por morfologia matemática**. 1997. Tese de Doutorado. Instituto de Matemática e Estatística da Universidade de São Paulo, 07.03. 97.

HUANG, Gao et al. Densely connected convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2017. p. 4700-4708.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, n. 6, p. 84-90, 2017.

KRIZHEVSKY, Alex. **One weird trick for parallelizing convolutional neural networks**. arXiv preprint arXiv:1404.5997, 2014.

KUO, Chung-Feng Jeffrey et al. Using image processing technology and mathematical algorithm in the automatic selection of vocal cord opening and closing images from the larynx endoscopy video. **Computer methods and programs in biomedicine**, v. 112, n. 3, p. 455-465, 2013.

LEEZA, Mona; FAROOQ, Humera. **Detection of severity level of diabetic retinopathy using Bag of features model**. IET Computer Vision, v. 13, n. 5, p. 523-530, 2019.

MENDES NETO, José Arruda et al. **Comparação entre a telelaringoscopia e a laringoscopia de suspensão no diagnóstico das lesões benignas das pregas vocais**. 1, pp. 1-37). Rio de Janeiro. 2017

MONTANARI, Raphael. **Detecção e classificação de objetos em imagens para rastreamento de veículos**. 2016. Tese de Doutorado. Universidade de São Paulo.

NOWAK, Eric; JURIE, Frédéric; TRIGGS, Bill. Sampling strategies for bag-of-features image classification. In: **European conference on computer vision**. Springer, Berlin, Heidelberg, 2006. p. 490-503.

O'SHEA, Keiron; NASH, Ryan. **An introduction to convolutional neural networks**. arXiv preprint arXiv:1511.08458, 2015.

PASSEROTTI, Gustavo Haruo. **DOENÇAS BENIGNAS DA LARINGE**. Seminário de Otorrinolaringologia. Disponível em: <
http://otorrinousp.org.br/imageBank/seminarios/seminario_44.pdf > acesso em 01 de setembro de 2020.

PINHEIRO, Catarina Isabel Cartaxo. **Algoritmos de Segmentação para Aplicações Biomédicas**. 2017. Tese de Doutorado.

RONCERO, Valeriana Gomes. **UM ESTUDO DE SEGMENTAÇÃO DE IMAGENS BASEADO EM UM MÉTODO DE COMPUTAÇÃO EVOLUCIONÁRIA**. 2005. Tese de Doutorado. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO.

Sebastiani, F. . **Machine learning in automated text categorization**. ACM computing surveys (CSUR), 34(1):1–47., 2002.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M. A survey on image data augmentation for deep learning. **Journal of Big Data**, v. 6, n. 1, p. 60, 2019.

SILVA, Rodrigo Emerson Valentim da. **Um estudo comparativo entre redes neurais convolucionais para a classificação de imagens**. 2018.

SODRÉ, Bruno Ribeiro et al. **Reconhecimento de padrões aplicados à identificação de patologias de laringe**. 2016. Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná.

SPATTI, Danilo Hernane. **Desenvolvimento tecnológico e inovação em saúde: novas perspectivas para sistemas inteligentes**. Seminário de Pós Graduação em Engenharia Elétrica. 60 p. 2018.

VAN TEEFFELLEN, Sven; SHAEVITZ, Joshua W.; GITAI, Zemer. **Image analysis in fluorescence microscopy: bacterial dynamics as a case study**. *Bioessays*, v. 34, n. 5, p. 427-436, 2012.

Xavier SD, Freitas SV, Behlau MS. **Phono-Articulatory Analysis Before, During and After a Desintoxication Program for Alcoholists**. In: Behlau MS. *Voz do Especialista*. Rio de Janeiro: Revinter; 2001 p 8.

APÊNDICE

Programação Base

- Classificador *Bag of Fetures* - MATLAB

```
imgSets = imageSet('C:\Laringes\Data','recursive');
[imgSets.Count]
minSetCount = min([imgSets.Count]);
trainingSets = partition(imgSets,minSetCount,'randomize');
[trainingSets.Count]

%Vocabulario visual a partir dos dados de treinamento
bag = bagOfFeatures(trainingSets,'VocabularySize',150,'PointSelection','Detector');

img = read(imgSets(1), randi(imgSets(1).Count));
featureVector = encode(bag,img);
figure
subplot(3,2,1);
imshow(img);
subplot(3,2,2);
bar(featureVector);
title('Ocorrencia Visual');
xlabel('Indice Visual');
ylabel('Frequencia');

img = read(imgSets(2), randi(imgSets(2).Count));
featureVector = encode(bag,img);
subplot(3,2,3);
imshow(img);
subplot(3,2,4);
```

```

bar(featureVector);
title('Ocorrencia Visual');
xlabel('Indice Visual');
ylabel('Frequencia');

```

```

%treinar classificador para discriminar entre categorias
categoryClassifier = trainImageCategoryClassifier(trainingSets, bag);

```

```

%Avaliar a performace do classificador
%uso dos dados de teste
%Matriz de confusão
confMatrix = evaluate(categoryClassifier, trainingSets);

```

- Redes Neurais Convolucionais - Alexnet, Resnet e Densinet - *Google Colaboratory*

```

### Notebook para a classificação de imagens

```

```

# Importante as bibliotecas necessárias
from fastai.vision import *
from fastai.metrics import error_rate, accuracy

```

```

import warnings
warnings.filterwarnings('ignore')

```

```

### Viculando o Google Drive

```

```

from google.colab import drive

```

```

drive.mount('/content/drive')
## Definindo caminho
path = '/content/drive/MyDrive/Mestrado/Data/dataset2'; path

## Adequando as imagens e o dataset e separando entre treinamento e validação
tfms = get_transforms(do_flip=False)

data = (ImageList.from_folder(path) .split_by_rand_pct(valid_pct=0.25)
        .label_from_folder() .transform(tfms, size=224) .databunch())

data.show_batch(rows=3, figsize=(5,5))

data.classes

learn = cnn_learner(data, models.densenet121, metrics = [accuracy])

###learn = cnn_learner(data, models.densenet121, metrics = [accuracy])

##Disponível em: https://pytorch.org/docs/stable/torchvision/models.html

learn.fit_one_cycle(19)

interp = ClassificationInterpretation.from_learner(learn)
interp.plot_top_losses(9, figsize=(15,11))

interp.plot_confusion_matrix(figsize=(12,12), dpi=60)

```