Capítulo 11

Classificação de Áudio aplicada à Saúde

Detecção de Problemas Respiratórios

Marcelo Matheus Gauy
Larissa Cristina Berti
Flaviane Romani Fernandes Svartman
Beatriz Raposo de Medeiros
Celso Ricardo Fernandes de Carvalho
Marcelo Gomes de Queiroz
Marcelo Finger

Publicado em: 16/04/2026

11.1 Introdução

Classificação de áudio é um tipo de tarefa de aprendizado de máquina que consiste em atribuir uma classe a cada um dos áudios de um conjunto de dados. Um exemplo clássico de conjunto de dados para esta tarefa é o AudioSet (Gemmeke et al., 2017). O AudioSet consiste em aproximadamente 5000 horas de áudios do Youtube distribuídos em 527 classes (anotadas). Alguns exemplos de classe presentes no AudioSet são fala, respiração, tosse, ruídos de barcos e aviões, música clássica, entre outros. Modelos de aprendizado de máquina e, em particular, de aprendizado profundo, são utilizados para processar o sinal de áudio, seja ele em forma de onda ou espectrograma (Brigham; Morrow, 1967) e classificá-lo em uma das 527 classes. Trata-se, portanto, de uma tarefa de classificação, via aprendizado supervisionado, clássica.

O sucesso das técnicas de aprendizado de máquina (especialmente às de aprendizado profundo) tornou possível que pesquisadores buscassem usar essas técnicas na classificação de áudios de pessoas com problemas de saúde que possam afetar sua voz e fala (Fagherazzi et al., 2021). Dentre estes, destacam-se especialmente as condições de saúde que afetam o sistema respiratório de alguma maneira (Landry et al., 2025). Este capítulo tratará prioritariamente de modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios. Algumas especificidades distinguem este tipo de problema de classificação de uma tarefa clássica de classificação de áudio. Essas especificidades são dificuldades típicas no uso de aprendizado de máquina no contexto da saúde. A principal delas é a dificuldade, tanto por motivos práticos quanto por motivos éticos, de coletar dados de pacientes em hospitais. Modelos complexos de classificação necessitam, normalmente, de grande volume de dados para atingirem desempenho satisfatório. Em geral, é um grande desafio coletar áudios de pacientes no volume demandado por modelos do estado da arte de aprendizado profundo. Além do volume de dados coletados, a diversidade do conjunto de dados coletados também é um limitante importante em muitos casos. Esta diversidade vai desde possíveis comorbidades em pacientes, a outros fatores como variações da doença ao

longo do tempo. Além disso, o tratamento de áudios em situações reais pode levar a grandes variações nos tipos de ruído presentes, o que pode fazer com que modelos de excelente performance no conjunto de dados original falhem uma vez que ocorram alterações nestes ruídos.

O problema de lidar com o baixo volume de dados é tratado, em geral, através de técnicas clássicas de transferência de aprendizado. Modelos de classificação de áudio são pré-treinados em grandes volumes de áudio, como o AudioSet e, posteriormente, refinados para o conjunto de dados de menor volume em questão (Huang et al., 2022; Kong et al., 2020; Niizumi et al., 2022). A Subseção Tratamento de conjuntos pequenos de dados apresentará uma visão detalhada desta e outras técnicas para lidar com conjuntos pequenos de dados. O problema da baixa diversidade de dados presentes no conjunto de áudios de pacientes é mais complexo. No momento, não existem técnicas satisfatórias para tratar deste problema de forma geral. Isto torna a maior parte dos modelos de aprendizado profundo aplicados à saúde pouco úteis, uma vez que foram retirados de seu (restrito) contexto original (Cohen et al., 2021). Uma discussão sobre esse problema e como potencialmente lidar com ele será feito na Seção Considerações finais e impactos para a sociedade.

Descreveremos nas próximas subseções alguns problemas clássicos de saúde que afetam a voz e fala e podem ser detectados com modelos de classificação de áudio. Neste capítulo, focaremos principalmente na língua portuguesa, embora muitas das técnicas presentes aqui não sejam específicas para o português.

11.1.1 Detecção de COVID-19 através da voz e tosse

Durante a pandemia, diversas iniciativas de pesquisa (em nível internacional) foram criadas com o objetivo de detectar COVID-19 de forma automática, com aprendizado de máquina, através da voz, tosse e sons da respiração de pacientes (Brown et al., 2020a; Despotovic et al., 2021; Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Orlandic et al., 2021; Pinkas et al., 2020). Esses trabalhos apontaram a possibilidade de detectar COVID-19 através de sinais de áudio, atingindo entre 70% e 90% de acurácia na detecção da doença. A maioria dos conjuntos de dados utilizados foram construídos por contribuição colaborativa (*'crowdsourcing'*). As técnicas de aprendizado de máquina variam desde modelos mais simples com *features* construídos manualmente até o uso de redes neurais recorrentes.

Os trabalhos citados acima têm foco principal na língua inglesa, uma vez que os áudios coletados são de falantes do inglês. Trabalhos similares para o português foram feitos no escopo do projeto SPIRA-BM (Finger et al., 2025), um projeto temático da FAPESP de número 23/00488-5 cujas informações podem ser encontradas em https://bv.fapesp.br/pt/auxilios/113158. O SPIRA-BM é uma sequência de outro projeto FAPESP (número 20/06443-5) chamado SPIRA (Aluísio et al., 2022). O SPIRA-BM, bem como o projeto SPIRA, seu predecessor, possuem uma natureza ligeiramente diferente, uma vez que o foco do SPIRA-BM está na insuficiência respiratória (sintoma típico de COVID-19), e não na detecção da COVID-19 em si. Inclusive, o projeto SPIRA-BM busca detectar não apenas a insuficiência respiratória proveniente de COVID-19, mas também a proveniente de outras fontes. A próxima subseção tratará dos resultados obtidos pelo projeto até meados de 2025 para o caso de insuficiência respiratória.

11.1.2 Insuficiência Respiratória e seus efeitos na voz e fala

A insuficiência respiratória (IR) é um dos sintomas causados pela COVID-19 e a principal responsável pela hospitalização de pacientes de COVID-19. Por esse motivo, o projeto SPIRA reuniu médicos, fisioterapeutas, linguistas, fonoaudiólogos e cientistas da computação com o intuito de investigar a possibilidade de sua detecção automática através da voz e fala de pacientes. O projeto SPIRA-BM expandiu o escopo do projeto SPIRA original para além da IR proveniente de COVID-19, incluindo IR proveniente de outras fontes. No escopo desses projetos foram realizadas duas coletas de dados, uma na época da pandemia (*corpus* SPIRA que pode ser encontrado no Github) e outra um pouco depois das vacinas (*corpus* SPIRA 2.0 que não foi tornado público por questões de privacidade dos pacientes). Esses *corpora* consistem de sinais de áudio (voz e fala) de pacientes com IR e controles saudáveis e permitem a construção de modelos de aprendizado de máquina para a detecção da IR através do sinal de áudio. Mais detalhes sobre esses e outros *corpora* serão descritos na Seção *Corpora* de áudios. O projeto concluiu, através de análises dos áudios presentes, que os pacientes faziam pausas mais numerosas e mais longas que os controles, com uma distribuição, muitas vezes, em desacordo com o que seria esperado de um falante do português (Fernandes-Svartman et al., 2022). Em termos de parâmetros de voz, foi verificado que parâmetros relacionados à frequência fundamental são diferentes entre pacientes e controles (Berti et al., 2023).

Mais ainda, foram construídos diversos modelos de aprendizado profundo (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021; Matheus Gauy et al., 2026) que buscam a detecção da IR nos *corpus* SPIRA e SPIRA 2.0. Em (Casanova et al., 2021b), os autores construíram uma rede convolucional, chamada de SpiraNet, que recebe o MFCC-grama[1]. De fato, ambos MFCC-grama e espectrograma foram testados como entrada da rede convolucional construída. O desempenho da rede ao receber MFCC-grama foi substancialmente superior. Com o MFCC-grama, a rede atingiu aproximadamente 87% de acurácia no *corpus* SPIRA. Posteriormente, a rede convolucional foi substituída por um Transformer (Vaswani et al., 2017b) que recebia MFCC-grama, chamado de MFCC-Transformer (Gauy; Finger, 2021). Esta alteração da rede levou a uma performance de aproximadamente 96% no mesmo *corpus* SPIRA. Posteriormente, o grupo SPIRA buscou usar técnicas de transferência de aprendizado que permitissem aumentar a complexidade dos modelos usados e, também, permitissem o uso do espectrograma direto, sem passar pelo MFCC. Assim, os pesquisadores usaram redes convolucionais e redes baseadas no Transformer que foram pré-treinadas no AudioSet. Esses novos modelos atingiram performance acima de 98.5% (Matheus Gauy et al., 2026), tanto no caso de IR proveniente de COVID-19 (*corpus* SPIRA), quanto no caso em que a IR provém de diversas causas potenciais (*corpus* SPIRA 2.0). Uma observação importante é que os modelos não apresentam o mesmo desempenho caso sejam retirados de seu contexto original de treinamento, com os modelos treinados no *corpus* SPIRA, não obtendo um bom desempenho no SPIRA 2.0 e vice versa (Gauy et al., 2023). Apesar destas dificuldades, as técnicas propostas pelo projeto serão de grande valor quando treinadas em *corpus* mais diversos. Como exemplo, trabalhos do projeto SPIRA foram vencedores no desafio de detecção de COVID-19 ComParE 2021 (Casanova et al., 2021a). Além disso, usando técnicas de interpretabilidade de modelos de aprendizado profundo, foi possível verificar que os modelos do projeto SPIRA dão atenção principalmente à voz e fala dos

[1]Em linhas gerais, o MFCC é o resultado do cálculo do espectro dos áudios, via FFT, duas vezes. Na primeira, obtemos o espectrograma. Uma segunda aplicação da FFT sobre o espectrograma, produz o MFCC-grama. Mais detalhes são apresentados na Subseção Pré-processamento de áudios.

pacientes e não para os ruídos de fundo presentes nos áudios (Silva et al., 2024).

Considerando o sucesso obtido na detecção de IR pelo projeto, bem como o sucesso na detecção do COVID-19, foi criado o projeto temático SPIRA-BM que busca aprofundar a compreensão dos modelos de IR, potencialmente construindo modelos capazes de adicionalmente detectar a causa da IR, e também amplia o projeto para considerar outros problemas respiratórios, como asma e tabagismo. Na próxima subseção, discutiremos um pouco sobre esses problemas.

11.1.3 Outros problemas respiratórios

Inspirado no sucesso da detecção de IR, o projeto SPIRA-BM propôs estudar outros problemas respiratórios e seus efeitos na voz e na fala. Além da insuficiência respiratória, o projeto SPIRA-BM tem outras duas linhas gerais propostas. São elas: asma e tabagismo. Neste capítulo apresentamos esses problemas, o método de coleta de dados, os objetivos e seus impactos. Os resultados são parte do projeto temático SPIRA-BM (veja no link https://bv.fapesp.br/pt/auxilios/113158).

Consideremos o caso da asma: trata-se de uma doença crônica das vias aéreas, caracterizada por limitação ao fluxo aéreo expiratório que está associado a sintomas respiratórios como sibilância, dispneia, aperto no peito e tosse que variam ao longo do tempo quanto à sua frequência, ocorrência e intensidade. Esses sintomas são, frequentemente, desencadeados por fatores como exercício, exposição a alérgenos e/ou irritantes, alterações climáticas ou infecções respiratórias virais. O Brasil tem a oitava maior prevalência de asma no mundo, atingindo 12% da população brasileira. Na fase de criança e adolescência, a asma é mais prevalente em meninos enquanto, na fase adulta, tem prevalência maior entre mulheres. Em termos das tarefas de classificação de áudio que podemos considerar estão: i) a classificação entre asma não controlada e asma controlada; ii) a detecção de disfunção de prega vocal; iii) a previsão da probabilidade de exacerbação da asma dentro de um período de dois dias (estudo longitudinal). Os dois primeiros são problemas clássicos de classificação (binária) de áudio, com os dados de pacientes com asma controlada ou não controlada e com ou sem disfunção da prega vocal sendo coletados pelo projeto. A terceira tarefa é um estudo longitudinal, em que os pacientes com asma serão acompanhados ao longo do tempo e realizarão coletas de dados periódicas, com o intuito de permitir uma análise do risco de exacerbação da asma em um futuro próximo. Tal quadro de exacerbação, muitas vezes, leva à hospitalização. Estas tarefas, bem como outras análogas que possam ser consideradas, possuem o potencial, de auxiliar médicos e enfermeiros no tratamento preventivo da asma, se antecipando, por exemplo, a uma possível exacerbação dos sintomas.

No caso do tabagismo, busca-se: i) classificar o sujeito entre fumante ou não fumante; ii) estimar o nível de carga tabagística (quantidade de cigarros por dia vezes o número de anos como fumante); iii) determinar o nível de monóxido de carbono exalado pelos pacientes através do sinal de áudio. As tarefas ii) e iii) podem ser formuladas como problemas de regressão e, se necessário, quantizadas em problemas de classificação. Tarefas de regressão tendem a ser mais difíceis que tarefas de classificação, em particular, para modelos de análise de áudio. Um exemplo desta dificuldade pode ser encontrado em um trabalho do projeto SPIRA-BM que busca detectar o nível de saturação de oxigênio dos pacientes com insuficiência respiratória através do sinal de áudio (Matheus Gauy et al., 2026). Segundo Karelitz et al. (2017), a medida do monóxido de carbono no ar exalado (COex) é uma avaliação simples e de baixo custo, além de ser um método não invasivo e objetivo que avalia a exposição recente a cigarro ou fumaça de cigarro. É bastante útil para

detectar fumantes diários e mostra correlação com o consumo diário de cigarros. Dessa forma, estamos propondo o desenvolvimento de um método rápido e barato que auxilie os médicos e enfermeiros em programas de cessação de tabagismo.

11.1.4 Organização do capítulo

A Seção *Corpora* de áudios apresentará uma descrição dos diversos *corpora* existentes atualmente, bem como suas limitações. Incluiremos os *corpora* construídos no escopo do projeto SPIRA e diversos outros *corpora*. Será feita uma divisão entre os *corpora* para COVID-19 e IR. Também serão brevemente apresentadas algumas características dos dados em processo de coleta de asma e tabagismo. A Seção Métodos de detecção de problemas respiratórios através de áudios descreverá as técnicas de classificação de áudios para problemas respiratórios usados em geral. Trataremos das estratégias usuais de pré-processamento dos áudios e dos modelos de aprendizado profundo usados, descrevendo suas arquiteturas e técnicas para tratar de conjuntos pequenos de dados. A Seção Considerações finais e impactos para a sociedade apresentará os potenciais impactos práticos do desenvolvimento de modelos de aprendizado de máquina para a detecção de problemas respiratórios e das limitações existentes nos modelos atuais e possíveis estratégias para o seu aprimoramento.

11.2 *Corpora* de áudios

Nesta seção, descreveremos os *corpora* existentes para detecção de problemas respiratórios. Ressaltamos que não temos como objetivo fazer uma descrição completa, escolhendo apenas alguns dos *corpora* principais. Para revisões mais detalhadas dos *corpora* existentes indicamos, por exemplo, (Shuja et al., 2021) (para a COVID-19) e (Kapetanidis et al., 2024; Xia et al., 2022) (para uma descrição um pouco mais geral). Ainda nesta seção, descreveremos os *corpora*: SPIRA (também chamado de primeira fase do projeto SPIRA), SPIRA 2.0 (também chamado de segunda fase do projeto SPIRA) e alguns *corpora* focados na COVID-19. Por último, comentaremos brevemente sobre *corpora* em construção no escopo do projeto SPIRA-BM para a insuficiência respiratória, asma e tabagismo.

11.2.1 *Corpus* de detecção de insuficiência respiratória

Os dois principais *corpora* já coletados no escopo do projeto SPIRA para a detecção de insuficiência respiratória são o SPIRA e o SPIRA 2.0.

11.2.1.1 *Corpus* SPIRA

Este *corpus* foi coletado pelo projeto SPIRA durante a pandemia da COVID-19 e consiste de dados de pacientes com insuficiência respiratória (causada pela COVID-19) e controles saudáveis. Os dados dos pacientes de insuficiência respiratória foram coletados em dois hospitais universitários da cidade de São Paulo. Os dados dos controles foram coletados através de um aplicativo na internet para a coleta dos áudios. A coleta consistiu em 4 áudios de cada paciente ou controle. O primeiro foi o consentimento do falante para o estudo. Usualmente, esse tipo de consentimento é feito por escrito, mas excepcionalmente na pandemia, foi permitido o consentimento por fala. O segundo áudio coletado foi uma frase elaborada por linguistas, para ser longa o bastante para conter pausas naturais, mas simples o bastante para evitar que dificuldades de leitura levassem os pacientes a ter dificuldades de locução. A frase proposta foi: 'O amor ao próximo ajuda a enfrentar o

coronavírus com a força que a gente precisa'. A função do segundo áudio é verificar se os pacientes com insuficiência respiratória necessitam fazer pausas adicionais, não previstas gramaticalmente em português brasileiro, para recuperar o ar, durante a fala. O terceiro áudio consiste em uma parlenda conhecida pelos falantes brasileiros como 'batatinha quando nasce...'. O objetivo desse áudio é averiguar propriedades da fala quando as pausas estão predeterminadas. O último áudio consiste em cantar 'Parabéns para você'.

Uma observação importante é que o projeto SPIRA coletou áudios de pacientes e controles de ambientes distintos. Em particular, os ruídos de fundo presentes nos áudios foram bastante diferentes e muito pronunciados no caso dos pacientes (pois foram gravados em hospitais). Para evitar que esses ruídos introduzissem vieses nos modelos de aprendizado de máquina, foram coletados alguns áudios adicionais no início de cada coleta nos hospitais. Esses áudios contêm apenas o ruído ambiente presente, sem a voz ou fala de nenhum paciente. O intuito dessa coleta adicional foi permitir que os ruídos ambientes fossem adicionados aos áudios de controle, dificultando a tarefa da rede em classificar os áudios baseada apenas no ruído presente. A Subseção Pré-processamento de áudios tratará de maneiras de lidar com os diferentes tipos de ruído presentes nos áudios de pacientes e controles.

Ao todo, foram coletados mais de 6000 áudios de controle e 536 áudios de pacientes de insuficiência respiratória. Foi realizada uma triagem dos áudios de pacientes, para evitar, principalmente, que áudios contendo voz e fala dos coletores fossem mantidos no conjunto final. A triagem também foi feita, no caso dos estudos (Berti et al., 2023; Fernandes-Svartman et al., 2022) para excluir os dados em que fossem detectadas pausas ocasionadas por problemas de letramento. Como o número de controles total é muito maior que o de pacientes, também é feito um balanceamento do conjunto de dados, selecionando os controles (e pacientes) de forma a manter distribuições similares de homens e mulheres entre ambos[2]. Para a construção de modelos de aprendizado de máquina, os áudios remanescentes foram divididos em conjuntos de treino, validação e teste. A distribuição final dos dados está expressa na Tabela 11.1. O *corpus* original pode ser encontrado no Github. O conjunto de dados de controle completo (sem a triagem - mais de 6000 áudios e aproximadamente 18 horas de Português Brasileiro) pode ser encontrado no Zenodo.

Tabela 11.1: *Corpus* SPIRA

Dados		Treino	Validação	Teste
Controle	Masculino	59	8	22
	Feminino	84	8	26
	Duração média (s)	8.15	7.75	7.77
Pacientes	Masculino	83	8	28
	Feminino	66	8	32
	Duração média (s)	13.18	10.78	9.43
Áudios totais		292	32	108
Duração total (s)		3110	296	983

Fonte: (Casanova et al., 2021b)

[2]Segundo os autores, foi observado também um desbalanceamento das faixas etárias dos pacientes e controles, mas o tamanho do conjunto de dados tornou o balanceamento desse fator impraticável.

11.2.1.2 *Corpus* SPIRA 2.0

Este *corpus* possui uma estrutura similar ao anterior. A principal diferença é que ele foi coletado após o auge da pandemia e posterior às doses principais de vacinação. Assim, os dados de insuficiência respiratória não são dominados por pacientes de COVID-19, sendo provenientes de diversas causas. Outra diferença crucial é que tanto os dados dos controles quanto pacientes foram coletados em hospitais. Os dados foram coletados em 4 hospitais do estado de São Paulo. São eles: a Beneficência Portuguesa, o Hospital da Unimar, ambos em São Paulo, e a Santa Casa de Marília (*SC*) e CEES-Marília em Marília. Esta diferença torna o tratamento de ruído bem menos difícil que no *corpus* SPIRA. Apesar disso, os ruídos ambiente dos hospitais foram novamente coletados. Quanto aos outros áudios coletados de cada falante, novamente, o consentimento do estudo pelo falante foi em forma de áudio. A frase foi ligeiramente alterada para uma similar, que permitisse comparação posterior: 'O amor ao próximo ajuda a enfrentar *essa fase* com a força que a gente precisa'. A parlenda (com pausas predeterminadas) foi mantida. A canção 'Parabéns para você' foi substituída por uma vogal sustentada (vogal a, reproduzida por alguns segundos). A função da vogal sustentada é avaliar parâmetros de voz dos falantes, em oposição aos parâmetros de fala[3].

Uma desvantagem desse *corpus* sobre o anterior é seu tamanho. Por ter sido coletado após o auge da pandemia, a quantidade de pacientes nos hospitais que sofrem de insuficiência respiratória é bastante menor que durante a pandemia. Uma vantagem, porém, é que a insuficiência respiratória presente provém de diversas causas, não estando restrita apenas à COVID-19. Outra vantagem é que tanto pacientes e quanto controles foram coletados nos mesmos ambientes. Esse fato torna o tratamento do ruído bem mais simples.

Ao todo, foram coletados 116 áudios de controle[4] e 26 dados de pacientes com insuficiência respiratória de causas diversas. A Tabela 11.2 apresenta a distribuição dos dados de pacientes e controles por gênero usada em (Gauy et al., 2023). Novamente, existe um desbalanceamento. Como se trata de poucos dados, o balanceamento não é possível. Em (Gauy et al., 2023; Matheus Gauy et al., 2026), foi feito apenas o balanceamento dos conjuntos de validação e teste, mantendo o conjunto de treino com todos os dados remanescentes. Esse conjunto de dados não é público, como diversos outros conjuntos de dados de saúde. Observe que o *corpus* SPIRA 2.0 possui a limitação de ser excessivamente pequeno, não chegando a 20 minutos de duração total.

Tabela 11.2: *Corpus* SPIRA 2.0

Dados		
Controle	Masculino	36
	Feminino	80
	Duração média (s)	7.41
Pacientes	Masculino	14
	Feminino	12
	Duração média (s)	8.14
Áudios totais		144

[3]Aqui, temos uma distinção entre parâmetros de voz, que correspondem às características acústicas da elocução humana, e parâmetros de fala, que envolvem a produção de linguagem natural.

[4]Segundo os autores, foram removidos do conjunto de dados, aqueles pacientes que apresentaram sintomas de COVID longa, uma vez que eles poderiam enviesar os resultados (Robotti et al., 2021).

Fonte: (Gauy et al., 2023)

11.2.1.3 Limitações atuais dos *corpus* de insuficiência respiratória

Além das limitações específicas já citadas em cada um dos *corpora*, temos limitações gerais, que estão presentes em diversos *corpora* de saúde. Estas limitações são relativas a baixa diversidade do *corpus* de áudios coletado, em parte pelo número total de áudios, em parte pela limitada distribuição espacial e temporal da coleta realizada. Os áudios coletados estão concentrados em São Paulo, e em um período bem específico de tempo, não contemplando, por exemplo, variações ao longo do tempo (ou de localização geográfica) na distribuição das doenças. Isto impede que modelos *end-to-end* de aprendizado de máquina, eficazes na prática, sejam construídos. Um exemplo desta dificuldade pode ser observada pelos dois *corpora* coletados. Modelos treinados no *corpus* SPIRA apresentam um desempenho pior que uma previsão aleatória, se testados no SPIRA 2.0 e vice-versa (Gauy et al., 2023).

11.2.2 *Corpus* de detecção de COVID-19

Durante a pandemia, diversos pesquisadores no mundo coletaram dados de voz, fala e tosse de pacientes para a detecção de COVID-19. Enquanto a maior parte da pesquisa no Brasil focou na tarefa de detecção de insuficiência respiratória (dados apresentados na Subseção *Corpus* de detecção de insuficiência respiratória), pesquisadores no resto do mundo focaram principalmente na detecção de COVID-19 diretamente. Diversos *corpora* foram construídos por meio de coleta colaborativa (crowdsourcing) e foram apresentados em (Bhattacharya et al., 2023; Brown et al., 2020a; Despotovic et al., 2021; Orlandic et al., 2021). Embora existam problemas inerentes à coleta de dados de forma colaborativa (Garcia-Molina et al., 2016), essa é uma forma bastante eficaz de construir *corpus* de volume maior. Neste capítulo, apresentaremos dois *corpora* principais: o Coswara (Bhattacharya et al., 2023) (língua inglesa) e o ComParE (Schuller et al., 2021) (multilingual). Outros *corpora* foram construídos e podem ser encontrados nos artigos originais (Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Pinkas et al., 2020; Watase et al., 2023).

11.2.2.1 *Corpus* Coswara

Este *corpus* foi construído durante a pandemia, entre abril de 2020 e fevereiro de 2022 (Bhattacharya et al., 2023). Ao todo, foram coletados áudios de 2635 indivíduos, dos quais 1819 são negativos para o Coronavírus, 674 são positivos para o Coronavírus e 142 contraíram e se recuperaram da COVID-19. Os áudios foram gravados através de um aplicativo online (https://coswara.iisc.ac.in/). Os participantes informaram o seu consentimento para a coleta de dados e responderam um questionário contendo informações demográficas (incluindo idade, gênero, localização, se são fumantes ou não e se a gravação foi feita com máscara), estado de saúde atual (incluindo sintomas como tosse, febre, diarréia, dificuldades de respiração, entre outros, e comorbidades como pneumonia e asma) e teste de COVID-19. Após responder o questionário, os participantes gravaram nove áudios de diferentes tipos. Dois deles tratam da respiração. Um é superficial, definido como alguns ciclos de respiração rápida sem forçar os pulmões. O outro é profundo, um tipo de respiração mais lenta que trabalha mais os pulmões. Os dois áudios para a tosse envolvem uma tosse superficial induzida, que não força o pulmão, e uma tosse profunda, que envolve mais esforço pulmonar. Três tipos de vogais sustentadas foram gravados, com as vogais (como em *boot*), (como em

beet) e (como em *bat*), sendo escolhidas. Para a fala, foi solicitado que os participantes contassem de 1 a 20, em velocidade normal (primeiro) e rápida (segundo). Assim, foram totalizados nove áudios de cada paciente. O tempo de coleta médio por participante foi 7 minutos. O total de horas gravadas foi de aproximadamente 65. Importante observar que uma fração significativa ($> 95\%$) dos participantes que reportaram teste positivo de COVID-19 foi gravada em hospitais e centros de saúde, tornando estas anotações de alta qualidade.

Os áudios coletados foram escutados por humanos para serem manualmente anotados entre qualidade excelente (nenhum ruído ambiente), moderados (pouco ruído ambiente) e ruim (ruído ambiente significativo). Aproximadamente 78% dos áudios foram considerados excelentes, com 12% considerados moderados e 10% considerados ruins. Os dados originais estão disponibilizados no Zenodo.

11.2.2.2 *Corpus* ComParE COVID-19

O *COMputational PARalinguistics challengE* de 2021 (COMPARE 2021) envolveu 4 tarefas, duas delas focadas na detecção de COVID-19. A primeira focou em dados de tosse e a segunda foi direcionada para a fala. A construção do *corpus* é similar ao Coswara e foi feita nos trabalhos (Brown et al., 2020a; Han et al., 2021). Novamente, foi usada a técnica de coleta colaborativa. Os dados foram coletados através de um aplicativo que podia ser acessado por diversas plataformas (móvel ou internet). Os participantes responderam um questionário com informações demográficas, médicas e o sintomas reportados. Para o desafio ComParE, foram incluídos apenas os áudios de tosse e fala dos participantes com a informação do teste de COVID-19. Para a primeira tarefa, focada em tosse, foram gravados 725 áudios de 343 participantes, totalizando uma hora e meia. Cada participante gravou entre uma e três tosses forçadas. Para a segunda tarefa, focada na fala, foram gravados 893 áudios de 366 participantes, totalizando pouco mais de três horas. Cada participante gravou de um a três áudios recitando a frase: *"I hope my data can help to manage the virus pandemic"* em um idioma (inglês, alemão, italiano, etc). O *corpus* ComParE não inclui áudios de português.

Para as tarefas de detecção de COVID no desafio ComParE 2021 foi proposta uma divisão entre treino, validação e teste. A Tabela 11.3 apresenta esta divisão. É importante observar que cada participante pode aparecer mais de uma vez nos conjuntos então a divisão deve ser respeitada.

Tabela 11.3: *Corpus* ComParE

Dados		Treino	Validação	Teste
Tosse	Covid positivo	71	48	39
	Covid negativo	215	183	169
	Total	286	231	208
Fala	Covid positivo	72	142	94
	Covid negativo	243	153	189
	Total	315	295	283

Fonte: (Schuller et al., 2021)

11.2.3 *Corpora* para detecção de outros problemas respiratórios

Trataremos brevemente de *corpora* para a detecção de problemas respiratórios como asma e tabagismo. Na literatura internacional, é comum usar o *corpus* Coswara para a detecção de asma (Looi et al., 2024). Além do Coswara, temos *corpora* privados construídos para detecção de asma (Balamurali et al., 2021; Xu et al., 2025) e infecções das vias aéreas (Balamurali et al., 2021).

No âmbito do Português brasileiro, o projeto SPIRA-BM estuda as condições de insuficiência respiratória, asma e tabagismo, como mencionado anteriormente. O *corpus* coletado para as três condições será similar ao SPIRA 2.0. Adicionalmente, são coletadas comorbidades de pacientes, bem como diversas informações demográficas, como idade, gênero. Também são coletados fatores como frequência respiratória, saturação do oxigênio no sangue (usualmente, via oxímetro). No caso de asma, os participantes respondem um questionário típico para avaliar a situação atual da doença. No caso do tabagismo, é coletado o nível de monóxido de carbono exalado. Para mais informações, veja (Finger et al., 2025).

11.3 Métodos de detecção de problemas respiratórios através de áudios

Nesta seção, apresentaremos as principais técnicas usadas na construção de modelos de aprendizado de máquina para a detecção de problemas respiratórios através do sinal de áudio (voz e fala) de pacientes. Na Subseção Pré-processamento de áudios, discorreremos sobre as etapas de pré-processamento necessárias para evitar a introdução de vieses no modelo (como tratamento de ruído, uniformização da duração dos áudios, divisão em treino, validação e teste) e para preparação de *features* que podem ser passados aos modelos (como o espectrograma e MFCC-grama). Na Subseção Modelos de aprendizado profundo para classificação de áudios, apresentaremos os principais modelos de classificação de áudio existentes, destacando as técnicas usadas para tratar com poucos dados (Subseção Tratamento de conjuntos pequenos de dados) e as arquiteturas usualmente consideradas baseadas em redes convolucionais e redes do tipo Transformer (Subseção Arquiteturas).

11.3.1 Pré-processamento de áudios

Para o treinamento de modelos de aprendizado de máquina em geral, e, em particular, no caso de áudio, é essencial um pré-processamento adequado dos dados. Em termos das *features* inseridas nos modelos, a maioria dos modelos (em especial, os de aprendizado profundo) realiza um treinamento *end-to-end*, ou seja, o modelo recebe como entrada o áudio 'cru' e tem como saída a classe daquele áudio recebido. Mesmo assim, existem variações nos formatos de entrada dos áudios que são passados para os modelos. Um meio usado por alguns modelos é receber o áudio em forma de onda. Um exemplo de modelo com essa característica é o Wav2Vec e suas variantes (Baevski et al., 2020). Como alternativa à forma de onda, podemos passar ao modelo, como entrada, o espectrograma do áudio. O espectrograma é obtido através de uma transformada (rápida) de Fourier (Brigham; Morrow, 1967) que para cada janela de tempo, calcula um vetor de frequências relativas ao áudio. Ao realizar este cálculo para cada uma das janelas, obtemos uma representação do áudio no espaço de frequências, chamada de espectrograma. O espectrograma se assemelha a uma imagem (bidimensional) e os modelos de classificação de áudio que usam espectrograma

como entrada são muito similares a modelos de processamento de imagem (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Kong et al., 2020). Adicionalmente, podemos fazer uma transformada de Fourier adicional sobre o espectrograma para obter o 'espectro' do espectrograma, chamado de MFCC-grama. O MFCC-grama é muito usado como entrada para os modelos de classificação de áudio (Gauy; Finger, 2021; Gourisaria et al., 2024; Vimal et al., 2021), apresentando muitas vezes desempenho superior ao espectrograma, em particular, nos casos em que o tamanho do *corpus* de áudios analisado seja pequeno.

Além do processamento do áudio para tornar a entrada compatível ao modelo de classificação, seja para espectrograma, MFCC-grama ou mesmo em forma de onda, devemos evitar que vieses sejam introduzidos nos modelos por características do processo de coleta de dados usado (entre outros motivos). Temos três fatores principais que introduzem viés nos áudios: ruído, duração dos áudios e distribuição demográfica. Eles serão descritos nas próximas subseções.

11.3.1.1 Formas de viés: ruído

O ruído é uma das principais fontes de viés. Em particular, como os *corpora* são usualmente pequenos e coletados em poucos locais, é bem possível que eles possuam ruído característico daqueles locais de coleta, o que provavelmente vai gerar vieses nos modelos de classificação. Um exemplo claro em que o ruído é determinante é o *corpus* SPIRA apresentado na Subseção *Corpus* SPIRA. No *corpus* SPIRA, os dados de controle e pacientes foram coletados em ambientes distintos e com ruídos claramente distintos. Assim, um modelo treinado diretamente naquele *corpus* para distinguir pacientes de controles, provavelmente, não faria mais do que classificar o tipo de ruído presente (Casanova et al., 2021b). Há duas técnicas gerais para tratar a presença de ruído. A primeira é eliminar o ruído presente nos áudios, por exemplo, através de um filtro que elimina frequências abaixo de um certo limiar. Embora esta técnica seja muito eficaz em diversos problemas, para a detecção de problemas respiratórios, ela pode resultar na remoção de parte do sinal relevante para a classificação dos áudios (como a respiração do paciente). Assim, uma técnica alternativa pode ser considerada: coletar o ruído presente nos ambientes e artificialmente acrescentá-lo nos áudios antes do modelo recebê-los. Esta foi a técnica usada no *corpus* SPIRA (Casanova et al., 2021b; Gauy; Finger, 2021). Observe que acrescentar ruído no áudios diminui a relação sinal-ruído. Assim, determinar a técnica mais adequada depende da realização de experimentos empíricos com ambas as técnicas.

11.3.1.2 Formas de viés: duração

O ruído não é a única forma de viés comumente presente em *corpus* de áudio para detecção de problemas respiratórios. Também é muito comum que vieses existam na própria duração dos áudios. Por exemplo, no caso de insuficiência respiratória, pacientes usualmente gravam áudios mais longos que controles, visto que fazem pausas (para respiração, provavelmente) mais longas e mais numerosas (Fernandes-Svartman et al., 2022). Assim, se os modelos de classificação receberem os áudios com a duração total, eles tenderão a focar excessivamente na duração dos áudios e não no conteúdo em si. Uma solução simples e eficaz para esse problema é quebrar os áudios em trechos de mesmo tamanho. Inclusive, podemos usar janelas de um tamanho fixo (por exemplo 4 segundos) com um salto menor (por exemplo, de 1 segundo), de forma que um áudio de 8 segundos se transforme em 5 áudios de 4 segundos. Este método é chamado de janelamento (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021) e, além de prevenir o viés pela duração dos áudios, é uma técnica

simples e eficaz de aumento de dados. Em particular, ao lidar com conjuntos pequenos de dados, técnicas de aumentação são bastante importantes. Outra técnica de aumentação de dados para o caso de sinal de áudio é a inserção de ruído de diversos tipos nos áudios. Uma revisão da literatura sobre as muitas técnicas de aumentação de dados para problemas de classificação de áudio pode ser encontrada em (Ferreira-Paiva et al., 2022).

11.3.1.3 Formas de viés: distribuição demográfica

Além do ruído e da duração dos áudios, existe uma outra fonte bastante comum de vieses em modelos de classificação de áudio para detecção de problemas respiratórios. Essa fonte corresponde a diferentes características demográficas (por exemplo, sexo e idade) entre as diferentes classes. Por exemplo, podemos ter um *corpus* em que o número de pessoas do sexo feminino com a doença (asma, por exemplo) seja maior que o número de pessoas do sexo masculino. No conjunto de controle, porém, a relação se inverte. Nestes casos, os modelos de classificação teriam foco excessivo no sexo do paciente ao fazer a classificação. A solução para este viés é realizar uma divisão balanceada do *corpus* em conjuntos de treino, validação e teste, de forma que (pelo menos) a proporção de pacientes e controles nos diferentes grupos demográficos (especialmente gênero) seja similar em cada um dos conjuntos (de treino, validação e teste). A divisão em treino, validação e teste deve ser cuidadosa ao considerar técnicas como janelamento, por exemplo. Deve-se evitar, a todo custo, que áudios do mesmo participante caiam em conjuntos diferentes. Em geral, o conjunto de teste deve ser grande o bastante para permitir que os resultados medidos a partir dele sejam representativos, e o mais independente possível dos outros conjuntos. Sobre o conjunto de treino, existem menos restrições, e podemos usar diversas técnicas de aumentação de dados, por exemplo, para tentar melhorar os resultados na validação e, consequentemente, no teste.

11.3.1.4 Capacidade de generalização dos modelos

Por fim, gostaríamos de ponderar que a construção de modelos de aprendizado de máquina sobre conjuntos pequenos de dados é inerentemente problemática. É muito comum que modelos apresentem desempenho excelente no conjunto de teste mas que não generalizem na prática, quando testados em dados novos coletados por outros meios. Isto tende a ocorrer independentemente do cuidado ao pré-processar os dados e é uma característica de modelos de aprendizado de máquina, em especial, dos modelos de aprendizado profundo, por se tratarem de modelos caixa preta que encontram o caminho mais fácil para aprender a tarefa para a qual foram treinados (Cohen et al., 2021). Um exemplo desse tipo de problema, para o caso de insuficiência respiratória, é o trabalho (Gauy et al., 2023). Nesse trabalho, modelos treinados no *corpus* SPIRA foram testados no *corpus* SPIRA 2.0 e vice versa. Em nenhum dos casos, os modelos foram capazes de apresentar generalizações de forma eficaz. A Seção Considerações finais e impactos para a sociedade irá expandir esta discussão, apresentando os impactos para a sociedade dos modelos construídos e como, potencialmente, recuperar a capacidade de generalização dos modelos.

11.3.2 Modelos de aprendizado profundo para classificação de áudios

Nesta seção, apresentaremos os principais modelos de aprendizado profundo para classificação de áudio, usados no contexto de detecção de problemas respiratórios. Na Subseção Tratamento de conjuntos pequenos de dados, discutiremos sobre os cuidados necessários ao

serem considerados conjuntos pequenos de dados. Trataremos principalmente da técnica de realizar um pré-treinamento sobre o modelo em um conjunto de áudios relacionado e de grande volume, seguido de um rápido refinamento na tarefa em questão. Também discutiremos brevemente técnicas de aumentação de dados e regularização dos modelos. Na Subseção Arquiteturas, apresentaremos as arquiteturas comumente usadas pelos modelos, como as redes convolucionais e redes do tipo Transformer. Discutiremos as características das redes profundas de classificação de áudio, o tipo de entrada que recebem e a forma como são treinadas.

11.3.2.1 Tratamento de conjuntos pequenos de dados

Modelos do estado da arte de aprendizado profundo necessitam de grandes volumes de dados para seu treinamento. Quando tratamos de problemas de saúde, porém, atingir um grande volume de dados é extremamente difícil, visto que a coleta de dados é bastante custosa. Os *corpora* apresentados na Seção *Corpora* de áudios, em geral, não chegam a 100 horas de áudio, mesmo quando usam coleta colaborativa (*crowdsourcing*) e não chegam a 10 horas, quando coletados diretamente em hospitais. O contraste com conjuntos de dados gerais é nítido. O AudioSet, por exemplo, possui 5000 horas de áudios anotados em 527 classes. O *corpus* CommonVoice (Ardila et al., 2020), possui mais de 3000 horas, apenas de inglês, e mais de 30000 horas, incluindo mais de 100 idiomas.

A discrepância observada acima, sugere o uso de técnicas de transferência de aprendizado para permitir o uso de modelos de aprendizado profundo de complexidade maior. No contexto de processamento de linguagens naturais, o conhecido modelo BERT (Devlin et al., 2019) usou de uma técnica de pré-treinamento não supervisionado chamada de *masked language modelling*, que consistia em apagar uma proporção (15%) das palavras em uma frase para (pré-)treinar um modelo de forma não supervisionada para reconstruir a parte apagada. Posteriormente, o modelo foi refinado de maneira rápida e eficaz, mesmo em conjuntos com pequeno volume de dados, para outras tarefas de forma supervisionada. Esse método foi adaptado para o contexto de imagens e áudio alguns anos depois (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Liu et al., 2020a, 2020b). O espectrograma dos áudios (ou as imagens) é visto como blocos de pixels (16×16 por exemplo) e aproximadamente 70% dos blocos são apagados. O modelo é treinado para reconstruir a parte apagada. Note que, como áudios e imagens possuem muito mais informação local, é necessário apagar uma fração muito maior dos mesmos de forma a induzir os modelos a compreenderem a estrutura global (e não apenas local) dos dados. Para permitir esta alteração, é necessário usar um modelo mais complexo (baseado na rede Transformer), tanto para a codificação (*Encoder*) quanto para a decodificação (*Decoder*) (Baade et al., 2022; He et al., 2022; Huang et al., 2022). Além do pré-treinamento não supervisionado, podemos também realizar um pré-treinamento supervisionado para classificação de áudio como feito em (Kong et al., 2020). Neste trabalho, as redes foram pré-treinadas de forma supervisionada no AudioSet e depois refinadas para outras tarefas, muitas vezes com poucos áudios. Qualquer que seja o método de pré-treinamento usado, a ideia é que a transferência de aprendizado permita que modelos complexos de aprendizado profundo sejam usados mesmo em conjuntos com pequeno volume de dados.

Outro cuidado que pode ser benéfico para o tratamento de *corpora* pequenos é a aumentação dos dados. Técnicas de aumentação dos dados ajudam a evitar que modelos se especializem no conjunto de treino sem serem eficazes para fora dele (Ferreira-Paiva et al., 2022). Isto é chamado de *overfitting* e costuma ocorrer com modelos mais complexos.

No caso de áudio, a inserção de ruído é uma das principais técnicas de aumentação dos dados. Ela pode ser feita de diversas formas, como adicionar ruído gaussiano (branco), ou adicionar ruído de fundo comuns e não relacionados à tarefa em questão. Outra técnica é qualquer forma de segmentação dos áudios em áudios menores, como o janelamento, a remoção de partes silenciosas no começo e no final ou a translação do áudio por um fator aleatório. Também podemos alterar a velocidade de reprodução do áudio, a escala melódica dos áudios e ajustar o volume. Diversos tipos de filtros podem ser aplicados, para remover frequências mais baixas ou altas por exemplo. Alternativamente, podemos mascarar algumas frequências ou trechos do áudio aleatoriamente. Uma revisão das técnicas de aumentação de dados pode ser encontrada em (Ferreira-Paiva et al., 2022).

Além das técnicas acima, existem dois outros métodos gerais para tratar de conjuntos com pequeno volume de dados. O primeiro consiste em aplicar algum tipo de regularização ao modelo. Uma técnica de regularização comum é o acréscimo na função de perda de um termo que controle o tamanho dos pesos da rede, segundo alguma norma como a euclideana. Outra técnica de regularização é o uso do *dropout* (Srivastava et al., 2014). Esta consiste em remover uma proporção aleatória dos neurônios artificiais de uma camada da rede, forçando a mesma a construir um modelo mais robusto. O segundo método, em muitos casos imprescindível, é usar redes menos complexas. Redes de maior complexidade, com maior quantidade de parâmetros, possuem uma tendência mais forte ao *overfitting*. Uma solução é diminuir a complexidade do modelo construído, reduzindo o número de parâmetros, por exemplo, e evitando assim que a rede memorize o conjunto de treino.

Existe um problema comumente presente em conjuntos com pequeno volume de dados que não possui solução até o momento. Trata-se da tendência de que a distribuição dos dados do *corpus* seja diferente da que será encontrada na prática no momento de aplicação do modelo (Cohen et al., 2021). Este problema é bastante comum para problemas de saúde, visto que a população e as doenças passam por transformações ao longo do tempo. Além disso, a distribuição das doenças é diferente em localizações diferentes e conjuntos com pequeno volume de dados normalmente foram coletados em poucos locais. Na Seção Considerações finais e impactos para a sociedade, discutiremos as implicações dessa dificuldade prática e como poderíamos tentar contorná-la.

11.3.2.2 Arquiteturas

Em termos de arquiteturas comumente usadas em modelos de aprendizado profundo para detecção de problemas respiratórios, as redes convolucionais e as redes baseadas no Transformer são as mais comuns. As redes recorrentes tiveram diminuída a sua relevância com o advento do Transformer, assim como ocorreu em outros domínios. Em geral, os áudios são passados para as redes em forma de onda, ou mais comumente, como espectrograma. Em alguns casos, é usado o MFCC-grama. Tanto espectrograma quanto MFCC-grama são entradas parecidas com imagens, de forma que os métodos para análise de áudio usados se assemelham aos métodos usados para imagens.

11.3.2.3 Redes convolucionais: *Pretrained Audio Neural Networks*

Dentre as arquiteturas que usam redes convolucionais, as *Pretrained Audio Neural Networks* ou PANNs, propostas em (Kong et al., 2020), se destacam. Mais precisamente, as PANNs são redes pré-treinadas de forma supervisionada em grandes volumes de áudios (usualmente no AudioSet) e não necessariamente tem uma arquitetura definida. No caso de modelos de detecção de problemas respiratórios as PANNs baseadas em redes convolucionais CNN6,

CNN10 e CNN14 foram bastante utilizadas. Além de serem pré-treinadas, elas possuem a vantagem de serem 3 modelos similares de redes convolucionais com grau crescente de complexidade. Isto permite evidenciar o efeito do baixo volume de dados presente nos *corpora* considerados. A CNN6 é uma rede convolucional com 6 camadas, enquanto a CNN10 possui 10 camadas e a CNN14 possui 14 camadas. Todas as três recebem o espectrograma dos áudios como entrada. As camadas convolucionais da CNN6 usam kernels 5×5, enquanto que a CNN10 e CNN14 usam kernels 3×3. Cada uma dessas camadas é seguida de uma camada de *Batch normalization* (Ioffe; Szegedy, 2015) e a não linearidade escolhida é a *ReLu* (Nair; Hinton, 2010). Quatro dessas camadas estão presentes na CNN6 e entre duas delas uma camada de *average pooling* (Kong et al., 2019) está presente. Na CNN10 e CNN14, dois blocos convolucionais são aplicados em sequência antes do *average pooling* ser realizado. A CNN10 possui 4 pares desses blocos, enquanto a CNN14 tem 6 pares desses blocos. No último bloco convolucional, é usada uma camada de *Global pooling* (soma do *average* e *max pooling*) no lugar do *average pooling*. Uma penúltima camada totalmente conectada, intermediária entre os blocos convolucionais e a saída sigmoide (também totalmente conectada) com as 527 classes, está presente. Para prevenir o *overfitting*, a técnica de *dropout* é aplicada após cada camada. A CNN6 tem 4.8 milhões de parâmetros, a CNN10 tem 5.2 milhões e a CNN14 tem 80.7 milhões.

11.3.2.4 Redes convolucionais: SpiraNet

A SpiraNet, proposta em (Casanova et al., 2021b), também usa camadas convolucionais em sequência e recebe espectrograma ou MFCC-grama (mais eficaz para a detecção de insuficiência respiratória). Como a SpiraNet foi treinada do zero e não recebeu nenhum tipo de pré-treinamento, seu desempenho costuma ser inferior ao das PANNs. A tendência futura da área é o uso de modelos pré-treinados, de forma a facilitar o aprendizado da rede em *corpus* de poucas horas (ou mesmo minutos) de áudio.

11.3.2.5 Redes Transformer: *Masked Autoencoder*

Diversas arquiteturas recentes eficazes, baseadas no Transformer, foram construídas. Discutiremos principalmente um tipo de arquitetura chamada de *Masked Autoencoder* (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Niizumi et al., 2022). Esta arquitetura consiste em duas redes do tipo Transformer, uma para codificação e outra para decodificação. A construção da rede tem por objetivo permitir o pré-treinamento não supervisionado sobre grandes volumes de áudios (ou imagens). O codificador recebe o espectrograma dos áudios como blocos (usualmente 16×16). O codificador é alguma variação do *Vision Transformer* (Dosovitskiy et al., 2021). O decodificador é um Transformer clássico que vai reconstruir o espectrograma original, a partir da saída do codificador. O pré-treinamento consiste em apagar uma proporção alta (70% por exemplo) do espectrograma do áudio original e transmitir apenas a parte não apagada para o codificador. Como apenas uma fração menor dos dados é transmitida ao codificador, ele usualmente possui mais parâmetros que o decodificador. O decodificador deve reconstruir o espectrograma original a partir da saída. Esta arquitetura permite que o modelo aprenda a estrutura global dos áudios de forma auto-supervisionada. Este modelo pré-treinado é um dos mais eficazes classificadores de áudio existentes atualmente. Naturalmente, para as tarefas de detecção de problemas respiratórios, esta arquitetura também costuma se sobressair.

11.3.2.6 Redes Transformer: MFCC-Transformer

O MFCC-Transformer, usado em (Gauy; Finger, 2021), também usa um Transformer como base, tratando os tokens do Transformer como blocos de 1 frame pelo número de faixas de frequência. Ele recebe o MFCC-grama o que o torna mais eficiente do que se recebesse o espectrograma. Embora MFCC-Transfomer possua versões pré-treinadas no português brasileiro (Gauy; Finger, 2023), a técnica de pré-treinamento utilizada é menos eficaz (construída antes do surgimento do *Masked Autoencoder*) e o volume de áudios do português usados no pré-treinamento também é menor.

Um caminho potencial futuro é realizar um pré-treinamento adicional sobre um modelo pré-treinado. Esse pré-treinamento adicional poderia usar dados do português brasileiro, por exemplo, para preparar o modelo para melhor compreender o idioma (caso estejamos tratando de detectar problemas respiratórios de falantes do português). A técnica de pré-treinamento adicional foi proposta em (Niizumi et al., 2023, 2025), no modelo chamado *Masked Modeling Duo* ou M2D. Resumidamente, esse modelo usa duas redes em paralelo, uma que codifica os dados apagados e outra que codifica os dados não apagados e prevê o resultado codificado pelo outro modelo. A rede é treinada para maximizar o acordo entre os dois métodos. Esta técnica permite que o modelo seja pré-treinado de forma não supervisionada em um grande volume de dados e, posteriormente, que seja realizado um pré-treinamento adicional em uma tarefa mais próxima da tarefa alvo desejada. Em geral, esse pré-treinamento adicional leva a uma melhora de performance na tarefa final. No entanto, algumas dificuldades existem ao selecionar os parâmetros para o pré-treinamento adicional.

11.3.2.7 Outras arquiteturas

Além das redes apresentadas acima, existem diversos modelos de processamento de áudio cujo enfoque é o processamento de fala. Dentre essas, se destacam os modelos do tipo Wav2Vec (Baevski et al., 2020), que recebem o áudio em forma de onda e usam tanto camadas convolucionais (para extrair a informação do áudio), como camadas no estilo do Transformer. O Wav2Vec usa um quantizador para os áudios que permite extrair componentes discretas que compõe o mesmo (se assemelham a fones, ou seja, a sons da fala). Isto torna o modelo um transcritor de fala bastante potente. Outros modelos que têm foco no processamento de fala são o HuBERT (Hsu et al., 2021) e o XLS-R (Babu et al., 2021). Um tipo de modelo recente que foca em processamento multimodal e pode se tornar relevante para as tarefas de detecção de problemas respiratórios é o OmniVec (Srivastava; Sharma, 2024).

11.3.3 Avaliação dos modelos de detecção de problemas respiratórios

Em termos de avaliação dos modelos, algumas métricas, como a acurácia, são comumente usadas. Estas são métricas que costumam ser relevantes para qualquer problema de classificação. As seguintes métricas são comuns (VP corresponde a verdadeiro positivo, VN corresponde a verdadeiro negativo, FP corresponde a falso positivo e FN a falso negativo):

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN}$$

$$Recall = \frac{VP}{VP + FN}$$

$$\text{Precisão} = \frac{VP}{VP + FP}$$

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

$$\text{G-médio} = \sqrt{\text{Especificidade} \times \textit{Recall}}$$

$$\text{F-measure} = 2 * \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$$

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Em geral, a acurácia é a métrica mais comum. Porém, muitos *corpora* possuem um desbalanceamento apresentando mais controles que pacientes. Isto torna medidas alternativas, que consideram a precisão, o *recall* e a especificidade, como o G-médio ou a *F-measure*, muito mais acuradas para a avaliação dos modelos. No caso de problemas de saúde, poderiam ser consideradas ainda variações dessas medidas que põem pesos maiores para falsos negativos do que falsos positivos por exemplo. Curvas ROC e a área abaixo dela também são medidas comumente usadas para problemas de classificação (Hanley; McNeil, 1982), e não é diferente no caso particular de problemas respiratórios.

A Tabela 11.4 apresenta um resumo geral dos resultados obtidos para os modelos de detecção de insuficiência respiratória no escopo do projeto SPIRA. A Tabela 11.4 foca nos dados em português brasileiro, não apresentando modelos treinados em dados de outros idiomas. Em termos de métricas, ela apresenta apenas a acurácia, a título de comparação entre os modelos. Outras métricas podem ser encontradas em (Matheus Gauy et al., 2026).

Tabela 11.4: Resumo de trabalhos para a detecção de insuficiência respiratória

Referência	*Corpus*	*Features*	Modelo	Acurácia	Fraqueza
(Casanova et al., 2021b)	SPIRA	MFCC	SpiraNet	87%	Varia significativamente dependendo do ruído
(Gauy; Finger, 2021)	SPIRA	MFCC	MFCC-Transformer	96.5%	Modelo não pré-treinado
(Gauy et al., 2023)	SPIRA	MFCC	pretrained MFCC-Transformers	97.4%	Técnica de auto-supervisionamento é ineficaz
(Gauy et al., 2023)	SPIRA 2.0	MFCC	pretrained MFCC-Transformers	95%	*Corpus* muito pequeno para uma análise completa
(Matheus Gauy et al., 2026)	SPIRA	Audio Spectrogram	Audio-MAE	99.9%	Dados de pré-treinamento apenas em inglês
(Matheus Gauy et al., 2026)	SPIRA 2.0	Audio Spectrogram	CNN6	98.6%	Dados de pré-treinamento apenas em inglês. *Corpus* pequeno como (Gauy et al., 2023)

Fonte: Retirada de (Matheus Gauy et al., 2026).

11.4 Considerações finais e impactos para a sociedade

Os resultados apresentados indicam a viabilidade da detecção de problemas respiratórios através da voz e fala de pacientes. Os modelos analisados, embora apresentem um grau de complexidade elevado para o tamanho do *corpus* de treinamento, se mostraram bastante eficazes nas tarefas. O uso de diversas técnicas para prevenir o *overfitting* é crucial para atingir esse objetivo, especialmente o uso de modelos pré-treinados em grandes volumes de

áudios. A possibilidade de identificar problemas respiratórios de forma automática, com uma simples gravação de um paciente em um celular, é de grande valia para a sociedade e uma grande ajuda para médicos, enfermeiros e outros agentes de saúde.

Naturalmente, cuidados são necessários ao usar modelos caixa preta para detecção de problemas de saúde. Uma dificuldade natural é que esses modelos, por serem caixa preta, não permitem uma interpretação fácil de seus resultados, e não fornecem justificativas para a tomada de decisão. Embora existam técnicas como o Grad-CAM (Selvaraju et al., 2017) para facilitar a interpretação desses modelos caixa preta, e muitos autores as usaram para mostrar que os modelos de saúde em muitos casos identificam fatores relevantes nos dados (Moujahid et al., 2022; Panwar et al., 2020; Silva et al., 2024; Sobahi et al., 2022), é um problema em aberto, em geral, produzir modelos de aprendizado profundo que apresentem resultados interpretáveis. Parte desse problema pode ser evitado, deixando claro que esses modelos devem ser usados para uma análise preliminar, como a triagem de pacientes e a opinião final sempre deve ser de um médico especialista na área. Mesmo assim, seria bastante valioso criar modelos mistos que apresentem, como saída, além da previsão caixa preta do modelo, diversos parâmetros de voz e fala característicos e os resultados da comparação desses parâmetros de fala e da voz de pacientes com os mesmos parâmetros da fala e da voz de pessoas saudáveis. As análises dos linguistas e fonoaudiólogos sobre os áudios de pacientes e controles para determinar fatores de voz e fala que distinguem um grupo do outro são cruciais para uma melhor compreensão do problema (Berti et al., 2023, 2025; Fernandes-Svartman et al., 2022).

Outro cuidado, ainda mais relevante no uso desses modelos, consiste no fato de que é muito difícil gerar modelos que generalizem para fora do contexto original de seu conjunto de treinamento. Em particular, quando temos *corpora* pequenos como os de saúde, esses normalmente apresentam características específicas que permitem que o modelo use fatores específicos dos dados coletados e não generalize para fora daquele conjunto. Um exemplo claro desse efeito para problemas respiratórios pode ser encontrado em (Gauy et al., 2023). Os autores demonstraram que modelos de aprendizado profundo treinados no *corpus* SPIRA (coletado durante a pandemia), apresentam desempenho péssimo se testados no *corpus* SPIRA 2.0 (coletado após a vacinação e envolvendo insuficiência respiratória proveniente de causas além da COVID-19). O mesmo ocorre se o processo inverso for feito (treino no SPIRA 2.0 e teste no SPIRA). Ressalta-se que os modelos considerados em (Gauy et al., 2023) possuem todos excelente desempenho nos conjuntos de teste extraídos do *corpus* no qual foram treinados. Uma vez retirados de seu contexto original, porém, esses modelos passam de acurácias acima de 95% para acurácias menores que a de um chute aleatório. Esta dificuldade prática de implementar modelos de aprendizado profundo para problemas de saúde é documentada de forma mais ampla em (Cohen et al., 2021).

Para o caso de detecção de insuficiência respiratória e, potencialmente, para outros problemas respiratórios, é conhecido que diversos fatores de voz e fala são bastante diferentes entre pacientes e controles. Por exemplo, a distribuição das pausas nos enunciados de pacientes é bastante diferente da distribuição das pausas nos enunciados dos controles. A duração das pausas na fala dos pacientes também é bem maior que a dos controles. Considerando que diversos desses fatores que podem ser extraídos dos áudios devem ser semelhantes não apenas na fala dos controles mas na fala de todos os falantes do português brasileiro (a distribuição das pausas, por exemplo, é prescrita pela gramática do português, na medida em que há lugares na sentença em que sua ocorrência resultaria em agramaticalidade), é um caminho promissor construir modelos que não tratem dos áudios diretamente mas desses fatores específicos de voz e fala que sabemos serem relevantes e

mais estáveis, independentemente da forma de coleta. Naturalmente, esse método estaria abandonando a estratégia *end-to-end* tão comum na construção de modelos de aprendizado profundo. As dificuldades de construir *corpus* de saúde de grande volume e de criar modelos de aprendizado profundo que generalizem para fora de seu contexto podem tornar esse método alternativo, ao tentar controlar o tipo de entrada que é fornecida aos modelos, uma estratégia necessária para criar modelos mais robustos na prática.

Referências

ALUÍSIO, S. M. et al. Detecting Respiratory Insufficiency Via Voice Analysis: The SPIRA Project. Em: **Practical Machine Learning for Developing Countries on the Tenth International Conference on Learning Representations. Proceeding**. [s.l.] ICLR, 2022.

ARDILA, R. et al. **Common Voice: A Massively-Multilingual Speech Corpus**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.520/>

BAADE, A.; PENG, P.; HARWATH, D. **MAE-AST: Masked Autoencoding Audio Spectrogram Transformer**. (H. Ko, J. H. L. Hansen, Eds.)Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022. **Anais**...ISCA, 2022. Disponível em: <https://doi.org/10.21437/Interspeech.2022-10961>

BABU, A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. **arXiv preprint arXiv:2111.09296**, 2021.

BAEVSKI, A. et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations**., 2020. Disponível em: <https://arxiv.org/abs/2006.11477>

BALAMURALI, B. T. et al. Deep Neural Network-Based Respiratory Pathology Classification Using Cough Sounds. **Sensors**, v. 21, n. 16, 2021.

BERTI, L. C. et al. Fundamental frequency related parameters in Brazilians with COVID-19. **The Journal of the Acoustical Society of America**, v. 153, n. 1, p. 576–585, 2023.

BERTI, L. C. et al. **Acoustic Characteristics of Voice and Speech in Post-COVID-19**. Healthcare. **Anais**...MDPI, 2025.

BHATTACHARYA, D. et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. **Scientific data**, v. 10, n. 1, p. 397, 2023.

BRIGHAM, E. O.; MORROW, R. The fast Fourier transform. **IEEE spectrum**, v. 4, n. 12, p. 63–70, 1967.

BROWN, C. et al. **Exploring automatic diagnosis of COVID-19 from crowdsourced**

respiratory sound data. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. **Anais**...2020.

CASANOVA, E. et al. **Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021**. Proc. Interspeech 2021. **Anais**...a2021.

CASANOVA, E. et al. **Deep Learning against COVID-19: Respiratory Insufficiency Detection in Brazilian Portuguese Speech**. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais**...Online: Association for Computational Linguistics, ago. b2021.

COHEN, J. P. et al. Problems in the deployment of machine-learned models in health care. **Cmaj**, v. 193, n. 35, p. E1391–E1394, 2021.

DESPOTOVIC, V. et al. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. **Computers in Biology and Medicine**, v. 138, p. 104944, 2021.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.)Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais**...Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/n19-1423>

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. International Conference on Learning Representations. **Anais**...2021.

ERDOĞAN, Y. E.; NARIN, A. COVID-19 detection with traditional and deep features on cough acoustic signals. **Computers in Biology and Medicine**, v. 136, p. 104765, 2021.

FAGHERAZZI, G. et al. Voice for health: the use of vocal biomarkers from research to clinical practice. **Digital biomarkers**, v. 5, n. 1, p. 78–88, 2021.

FERNANDES-SVARTMAN, F. et al. **Temporal prosodic cues for COVID-19 in Brazilian Portuguese speakers**. Proc. Speech Prosody 2022. **Anais**...2022.

FERREIRA-PAIVA, L. et al. **A survey of data augmentation for audio classification**. Congresso Brasileiro de Automática-CBA. **Anais**...2022.

FINGER, M. et al. **SPIRA-BM: Biomarkers for Respiratory Conditions by Audio Analysis via Artificial Intelligence**. Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/35566>

GARCIA-MOLINA, H. et al. Challenges in data crowdsourcing. **IEEE Transactions on**

Knowledge and Data Engineering, v. 28, n. 4, p. 901–911, 2016.

GAUY, M. M. et al. **Discriminant Audio Properties In Deep Learning Based Respiratory Insufficiency Detection In Brazilian Portuguese**. Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine, AIME 2023, Portorož, Slovenia, June 12–15, 2023, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2023. Disponível em: <https://doi.org/10.1007/978-3-031-34344-5_32>

GAUY, M. M.; FINGER, M. Acoustic models of Brazilian Portuguese Speech based on Neural Transformers. **arXiv preprint arXiv:2312.09265**, 2023.

GAUY, M.; FINGER, M. **Audio MFCC-gram Transformers for respiratory insufficiency detection in COVID-19**. Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17793>

GEMMEKE, J. F. et al. **Audio set: An ontology and human-labeled dataset for audio events**. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2017.

GOURISARIA, M. K. et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. **Discover Internet of Things**, v. 4, n. 1, p. 1, 2024.

HAN, J. et al. **Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data**. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2021.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HASSAN, A.; SHAHIN, I.; ALSABEK, M. B. **COVID-19 detection system using recurrent neural networks**. 2020 International conference on communications, computing, cybersecurity, and informatics (CCCI). **Anais**...IEEE, 2020.

HE, K. et al. **Masked autoencoders are scalable vision learners**. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. **Anais**...2022.

HSU, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 29, p. 3451–3460, 2021.

HUANG, P.-Y. et al. Masked autoencoders that listen. **Advances in Neural Information Processing Systems**, v. 35, p. 28708–28720, 2022.

IOFFE, S.; SZEGEDY, C. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. International conference on machine learning. **Anais**...PMLR, 2015.

KAPETANIDIS, P. et al. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. **Sensors**, v. 24, n. 4, p. 1173, 2024.

KARELITZ, J. L.; MICHAEL, V. C.; PERKINS, K. A. Analysis of agreement between expired-air carbon monoxide monitors. **Journal of smoking cessation**, v. 2, n. 12, p. 105–112, 2017.

KONG, Q. et al. **Cross-Task Learning for Audio Tagging, Sound Event Detection and Spatial Localization: DCASE 2019 Baseline Systems**. [s.l.] DCASE2019 Challenge, 2019.

KONG, Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020.

LAGUARTA, J.; HUETO, F.; SUBIRANA, B. COVID-19 artificial intelligence diagnosis using only cough recordings. **IEEE Open Journal of Engineering in Medicine and Biology**, v. 1, p. 275–281, 2020.

LANDRY, V. et al. Audio-based digital biomarkers in diagnosing and managing respiratory diseases: a systematic review and bibliometric analysis. **European Respiratory Review**, v. 34, n. 176, 2025.

LIU, A. T. et al. **Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders**. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, a2020.

LIU, A. T.; LI, S.-W.; LEE, H. Tera: Self-supervised learning of transformer encoder representation for speech. **arXiv preprint arXiv:2007.06028**, b2020.

LOOI, Z. Q. et al. **MeLoDicA AI-Machine Learning Based Detection of Asthma via Vocal Audio Analysis**. 2024 IEEE Conference on Artificial Intelligence (CAI). **Anais**...IEEE, 2024.

MATHEUS GAUY, M. et al. Contrasting deep learning audio models for direct respiratory insufficiency detection versus blood oxygen saturation estimation. **Intelligence-Based Medicine**, v. 13, p. 100331, 2026.

MOUAWAD, P.; DUBNOV, T.; DUBNOV, S. Robust detection of COVID-19 in cough sounds: using recurrence dynamics and variable Markov model. **SN Computer Science**, v. 2, n. 1, p. 34, 2021.

MOUJAHID, H. et al. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. **Intelligent Automation & Soft Computing**, v. 32, n. 2, 2022.

NAIR, V.; HINTON, G. E. **Rectified linear units improve restricted boltzmann machines**. Icml. **Anais**...2010.

NIIZUMI, D. et al. **Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation**. HEAR: Holistic Evaluation of Audio Representations. **Anais**...PMLR, 2022.

NIIZUMI, D. et al. **Masked modeling duo: Learning representations by encouraging both networks to model the input**. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2023.

NIIZUMI, D. et al. Towards Pre-training an Effective Respiratory Audio Foundation Model. **arXiv preprint arXiv:2505.15307**, 2025.

ORLANDIC, L.; TEIJEIRO, T.; ATIENZA, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. **Scientific Data**, v. 8, n. 1, p. 156, 2021.

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.

PINKAS, G. et al. SARS-CoV-2 Detection From Voice. **IEEE Open Journal of Engineering in Medicine and Biology**, v. 1, p. 268–274, 2020.

ROBOTTI, C. et al. Machine Learning-based Voice Assessment for the Detection of Positive and Recovered COVID-19 Patients. **Journal of Voice**, 2021.

SCHULLER, B. W. et al. **The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates**. Interspeech 2021. **Anais**...2021.

SELVARAJU, R. R. et al. **Grad-cam: Visual explanations from deep networks via gradient-based localization**. Proceedings of the IEEE international conference on computer vision. **Anais**...2017.

SHUJA, J. et al. COVID-19 open source data sets: a comprehensive survey. **Applied Intelligence**, v. 51, n. 3, p. 1296–1325, 2021.

SILVA, D. P. P. DA et al. Interpretability analysis of deep models for COVID-19 detection. **Artificial Intelligence in Health**, v. 1, n. 3, p. 114–126, 2024.

SOBAHI, N. et al. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. **Biocybernetics and Biomedical Engineering**, v. 42, n. 3, p. 1066–1080, 2022.

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.

SRIVASTAVA, S.; SHARMA, G. **OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning**. 2024 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR). **Anais**...2024.

VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., 2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

VIMAL, B. et al. **MFCC Based Audio Classification Using Machine Learning**. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). **Anais**...2021.

WATASE, T. et al. Severity Classification Using Dynamic Time Warping–Based Voice Biomarkers for Patients With COVID-19: Feasibility Cross-Sectional Study. **JMIR Biomedical Engineering**, v. 8, n. 1, p. e50924, 2023.

XIA, T.; HAN, J.; MASCOLO, C. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. **Experimental Biology and Medicine**, v. 247, n. 22, p. 2053–2061, 2022.

XU, S. et al. Automated Lightweight Model for Asthma Detection Using Respiratory and Cough Sound Signals. **Diagnostics**, v. 15, n. 9, 2025.