

Disciplina de Visualização de Dados para a Área de Saúde

Projeto - Análise de condições clínicas respiratórias com base em anotações médicas e áudios de tosse

Brunna Raphaelly Amaral da Silva - 144566

Carolina Vieira Campos - 263081

Gabriel Bianchin de Oliveira - 230217

Taciana Alessandra Gomes Cruz - 107132

INTRODUÇÃO

Neste trabalho fizemos a classificação de áudios de tosse comparando diferentes modelos construídos que vão desde transfer learning, staking ensemble e visual transformers.

DATASET

INSPEÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

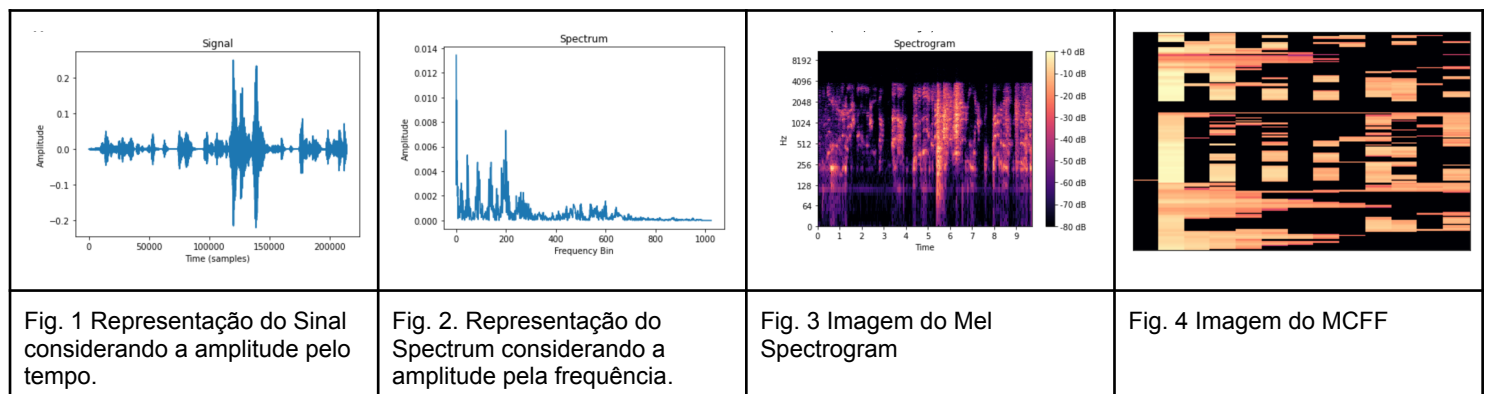
O dataset possui 20.000 registros tanto de áudio quanto de metadados referentes aos áudios. Os áudios de tosse possuem durações diferentes entre eles e em alguns áudios é possível perceber a presença de ruídos ambientes tais como pessoas conversando e aparelho de televisão ligados, assim como áudios sem tosse ou ainda áudios apenas com silêncio.

O dataset possui a priori as classes COVID-19, healthy, symptomatic e NaN(classe ausente). Desses, filtramos os relevantes para a nossa pergunta, sendo: COVID-19, healthy, symptomatic. A Tabela abaixo mostra a quantidade de áudios em cada uma das categorias.

COVID-19	1.010
healthy	8.562
symptomatic	1.742

Inicialmente os áudios sem tosse foram excluídos utilizando uma feature presente no dataset que classifica com um grau de certeza a presença de uma tosse ou não no áudio. Esta feature apresenta valores de 0.0 até 1.0. Levando isso em consideração, excluímos das análises os áudios que possuem graus iguais ou menores de certeza de serem tosse igual a 0.5. Dentre os áudios com indicação de tosse acima de 0.5, algumas amostras possuem a análise feita por especialistas, indicando condições presentes naquele áudio.

É possível a transformação do áudio em imagens utilizando o Sinal (amplitude quanto ao tempo) , o Spectrum (amplitude quanto a frequência) e também pelo Mel Spectrogram e do Mel-frequency cepstral coefficients (MFCCs), conforme mostram as figuras abaixo.



Neste trabalho iremos explorar o uso da representação do áudio em Mel Spectrogram e MFCC.

O dataset foi dividido entre conjunto de treinamento, validação e teste. Separou-se inicialmente o conjunto de teste com 20% exemplos e dentro os exemplos restantes separamos 80% para treino e 20% para validação, sendo assim ficando com 20% para o teste, 16% para validação e 64% para treinamento. Utilizou-se os pesos para cada classe como forma de lidar com o desbalanceamento dos dados.

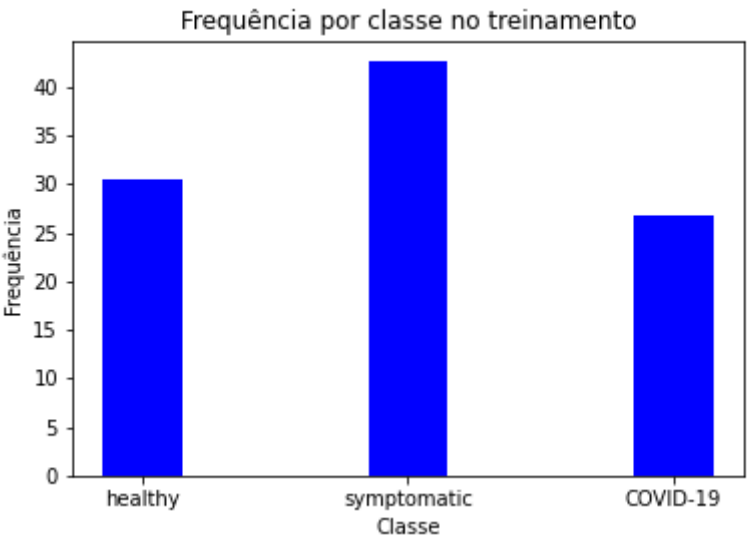
Com o conjunto de treinamento já definido, elaboramos inicialmente duas vertentes de treinamento para a construção dos modelos. A primeira abordagem considera apenas os dados anotados por especialistas no conjunto original de treinamento e a segunda abordagem considera dados adicionais no treinamento. Os dados adicionais agregados no conjunto de treinamento são providos da inclusão dos áudio que possuam 50% de existência de tosse avaliada pela variável *cough_detect* e que não foram avaliados por especialistas. Já nos conjuntos de validação e teste, todos os dados foram verificados por especialistas.

Efetuamos em seguida, o mapeamento das classes que até então eram categóricas, para os valores 0, 1 e 2. Logo, a classe healthy passou a ser 0, a classe symptomatic passou a ser 1 e a classe COVID-19 passou a ser 2.

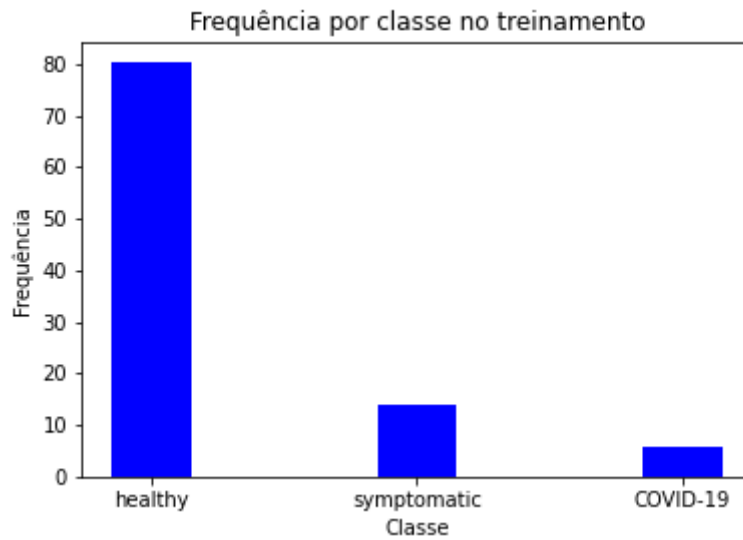
Os pesos das classes para cada uma dessas vertentes são descritos abaixo. Resolvemos utilizar os pesos para cada classe para que os modelos tivessem uma penalização maior para erros nas classes minoritárias. Calculamos os pesos a partir da divisão da classe com mais amostras pela classe em questão, de modo que o peso da classe majoritária fosse igual a 1 e as outras classes tivessem pesos maiores que 1.

healthy - com o peso 1.3982300884955752 symptomatic - - com o peso 1.0 COVID-19 -- com o peso 1.5906040268456376	healthy - com o peso 1.0 symptomatic - com o peso 5.826048171275647 COVID-19 - com o peso 13.807610993657505
Tab. 1 Pesos do treinamento com dados de especialistas	Tab. 2 Pesos com com dados de especialistas e dados adicionais

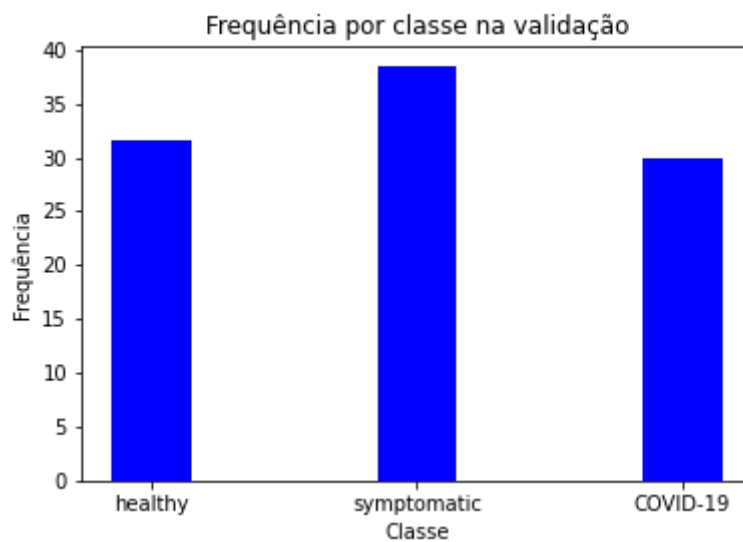
Para o experimento, utilizando apenas os dados anotados por especialistas no conjunto de treinamento, a distribuição neste conjunto é dada pela Figura abaixo. A classe mais comum é symptomatic, com mais de 40% dos dados, seguida por healthy, com 30% dos dados, e COVID-19, com 27% dos dados.



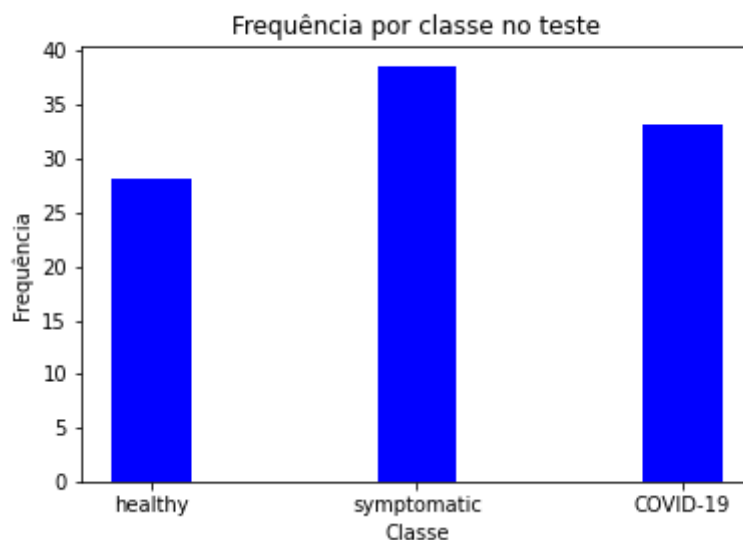
Para o experimento com dados adicionais, a distribuição muda no treinamento. A classe mais frequente é healthy, com 80% dos dados, seguido por symptomatic, com mais de 10% dos dados, e COVID-19, com pouco menos de 10% dos dados.



Já no conjunto de validação (mesmo conjunto utilizado tanto para os experimentos com apenas os dados de especialistas e dados adicionais, com o objetivo dos resultados serem comparáveis), a distribuição segue próxima ao treinamento com os dados dos especialistas.



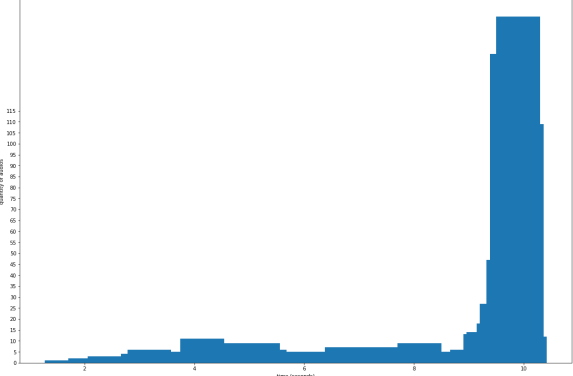
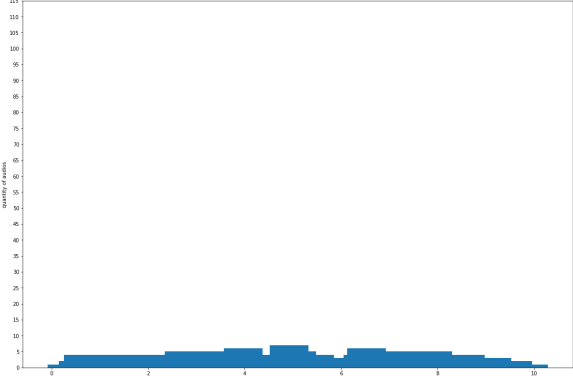
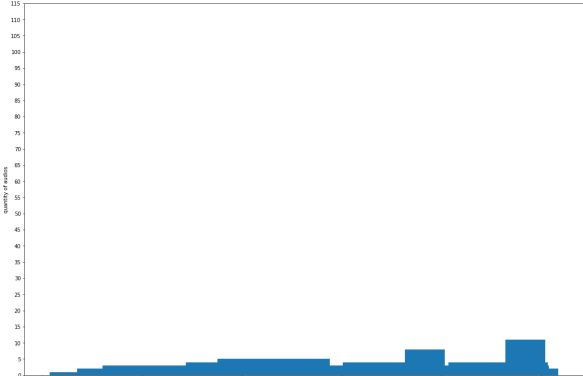
Já para o conjunto de teste, a distribuição segue com mais de 35% dos dados sendo da classe symptomatic, mais de 30% de COVID-19 e cerca de 27% como healthy.

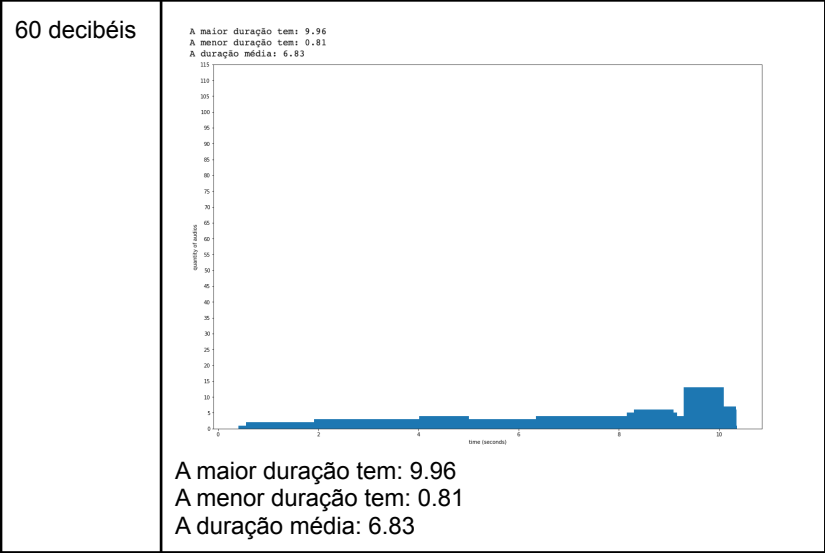


Análise da duração do áudio e do trim considerando as duas vertentes de treinamento

O trim do áudio é muito utilizado no processamento do áudio para eliminação de silêncio. Por isso, elaboramos também uma variação do trim do áudio com 28db, 50db, e 60db para identificar a influência sobre a quantidade de informação dos áudios que foram eliminadas. Abaixo podemos observar os gráficos da quantidade de áudios pelo tempo de duração.

Podemos inferir, através dos gráficos abaixo, que a variação da duração dos áudios com a influência do trim (28, 50 e 60) como ponto de corte é perceptível, modificando significativamente a distribuição dos dados originais da quantidade de áudios pelo tempo de duração. Entendemos também que o ponto de corte do trim levando em consideração 28 decibéis é o mais drástico dentre os analisados, pois ele apresentou a menor média de duração dos áudios.

	Vertente 1 : Treinamento com dados anotados por especialistas
Original	<div><div>A maior duração tem: 10.02 A menor duração tem: 1.68 A duração média: 8.35</div><div>A maior duração tem: 10.02 A menor duração tem: 1.68 A duração média: 8.35</div></div>
28 decibéis	<div><div>A maior duração tem: 9.89 A menor duração tem: 0.33 A duração média: 4.69</div><div>A maior duração tem: 9.89 A menor duração tem: 0.33 A duração média: 4.69</div></div>
50 decibéis	<div><div>A maior duração tem: 9.95 A menor duração tem: 0.55 A duração média: 6.30</div><div>A maior duração tem: 9.95 A menor duração tem: 0.55 A duração média: 6.30</div></div>



Experimentos

Utilizamos os seguintes experimentos na elaboração do trabalho.

- 21 redes diferentes do imageNet
- Variação dos trim dos áudios em 90 (default), 60 e 40 decibéis.
- Utilizando um algoritmo segmentatório que identifica os trechos de tosse no áudio
- Transformers
- Stacking Ensemble
- Aumentação dos dados

A seguir iremos descrever brevemente cada um deles além de exibir um quadro comparativo com base na acurácia balanceada na validação de cada modelo.

CONSIDERAÇÕES SOBRE AS REDES DO IMAGENET

Fixamos todas as redes analisadas com o early stopping para loss de validação com patience igual a 10 épocas e com o Reduce learning rate com a diminuição no learning rate em fator igual a 0,1 após 5 épocas sem melhorias. Neste trabalho, avaliou-se todos os modelos com base na loss e na acurácia balanceada. Para avaliarmos as redes, independente do experimento, utilizamos o mesmo conjunto de validação.

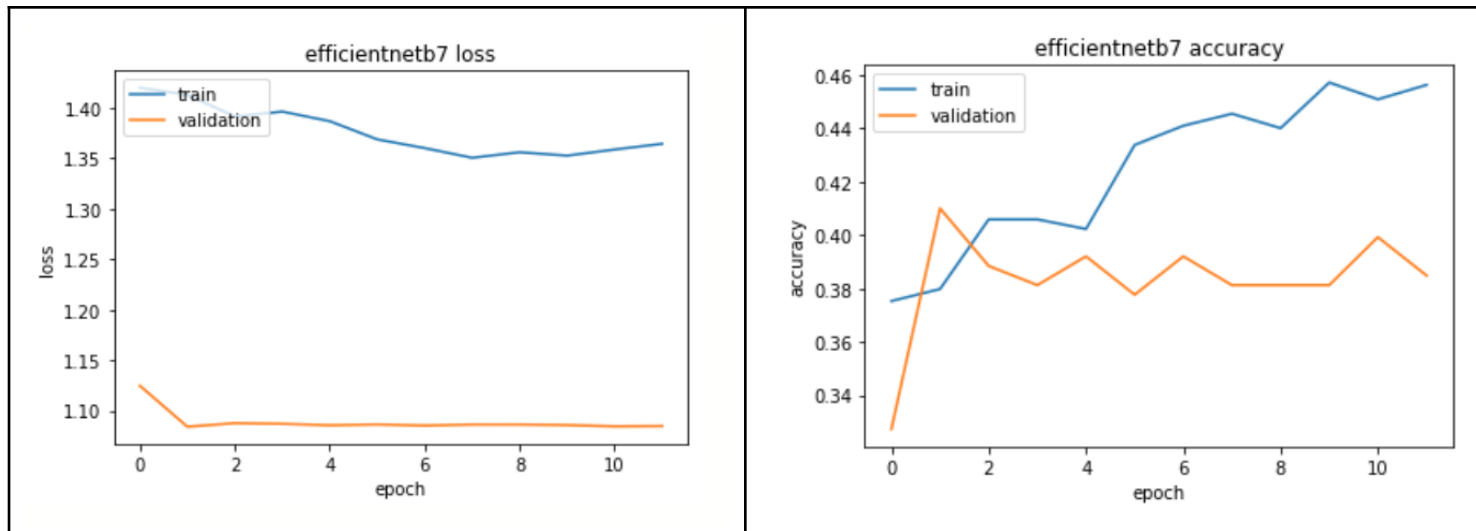
Analisamos o desempenho de 21 redes diferentes da imageNet, que são apresentadas na Tabela a seguir.

ResNet50	MobileNet
ResNet101	MobileNetV2
ResNet152	DenseNet121
EfficientNetB0	DenseNet169
EfficientNetB1	DenseNet201
EfficientNetB2	InceptionV3
EfficientNetB3	InceptionResNetV2
EfficientNetB4	Xception
EfficientNetB5	VGG16
EfficientNetB6	VGG19
EfficientNetB7	

Experimento comparativo entre os áudios originais do *Especialista* e *Especialista + adicional* sobre as imagens de representação do MEL Spectrogram e do MFCC.

Redes	Mel Spectrogram		MFCC	
	Especialista	Especialista + adicional	Especialista	Especialista + adicional
ResNet50	37,99	38,72	35,14	30,34
ResNet101	37,43	39,70	33,01	38,52
ResNet152	36,07	34,51	40,55	35,50
EfficientNetB0	31,74	35,40	34,72	36,98
EfficientNetB1	39,28	42,02	35,65	35,67
EfficientNetB2	37,47	40,77	37,62	33,80
EfficientNetB3	32,13	39,76	31,71	37,19
EfficientNetB4	34,30	41,28	38,82	32,86
EfficientNetB5	37,52	37,07	38,52	36,78
EfficientNetB6	34,00	44,23	33,28	41,81
EfficientNetB7	41,34	38,39	41,42	43,49
MobileNet	35,73	34,75	37,49	34,30
MobileNetV2	34,97	32,75	38,50	39,43
DenseNet121	33,24	37,49	33,16	36,14
DenseNet169	38,24	38,52	36,45	34,24
DenseNet201	36,72	37,63	35,70	30,18
InceptionV3	40,30	36,76	34,39	39,53
InceptionResNetV2	32,23	36,91	34,89	33,01
Xception	37,11	33,82	40,41	31,23
VGG16	33,80	37,42	34,66	38,71
VGG19	40,17	40,99	35,89	38,40

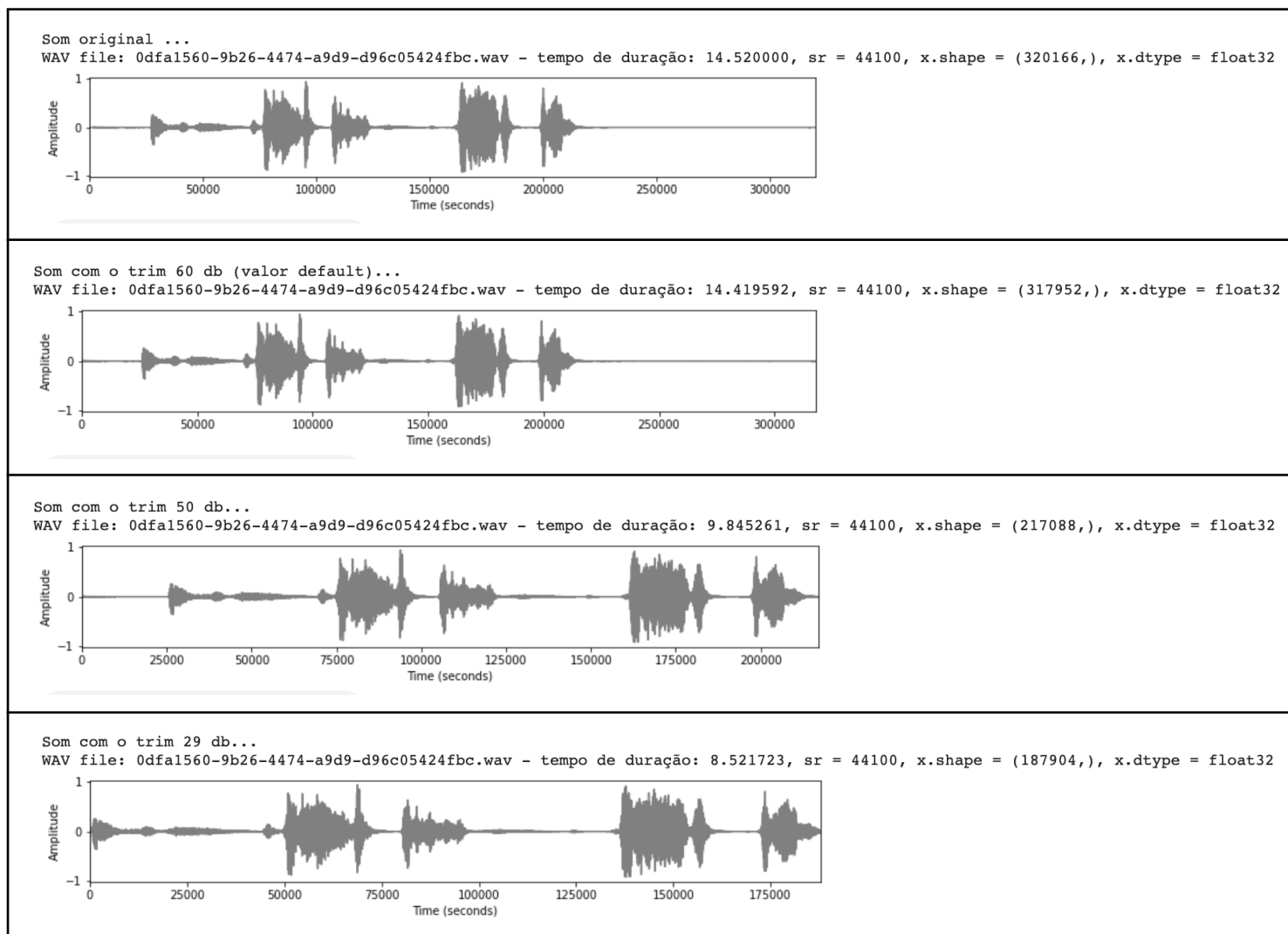
Com base na tabela acima temos a que a maior acurácia balanceada é referente ao MEL Spectrogram com a EfficientNetB6 considerando os dados do especialistas+adicional, com 44,23% de acurácia balanceada no conjunto de validação, e em segundo lugar temos a representação do MFCC do especialista + adicional, com 43,49%. Porém, considerando que teor da problemática trabalhada traz mais credibilidade para os registros vistoriados pelos especialistas, temos que a maior acurácia balanceada obteve o valor de 41,42% com as imagens do tipo MFCC pelo EfficientNetB7 e em segundo lugar obteve 41,34% de acurácia balanceada com as imagens do tipo MEL Spectrogram, valor este também obtido pelo EfficientNetB7. Além dos melhores valores, destacados em cores verde e amarelo, estão em destaque os resultados que atingiram acima de 40% de acurácia balanceada na validação. Abaixo temos as curvas de avaliação pelo MFCC do EfficientNetB7 só com os dados dos especialistas.



Com base nas curvas podemos perceber que não houve overfitting dos dados e que a rede apresentou o melhor modelo para esse problema logo na primeira época.

Experimento com os TRIMs do áudio

Um áudio é composto por vários níveis de frequência. Abaixo temos uma representação do áudio 0dfa1560-9b26-4474-a9d9-d96c05424fbc.wav considerando a amplitude da onda pelo tempo (segundos).



Podemos analisar que o trim é como um ponto de corte considerando um determinado valor de decibéis passado. É preciso levar em consideração que valores baixos de decibéis podem extrapolar os níveis considerados de silêncio podendo remover também fragmentos significativos do áudio.

O default como limiar de corte para exclusão de silêncio é 90 decibéis. Logo, consideramos a exclusão de três pontos 90 (default), 60, e 40 decibéis.

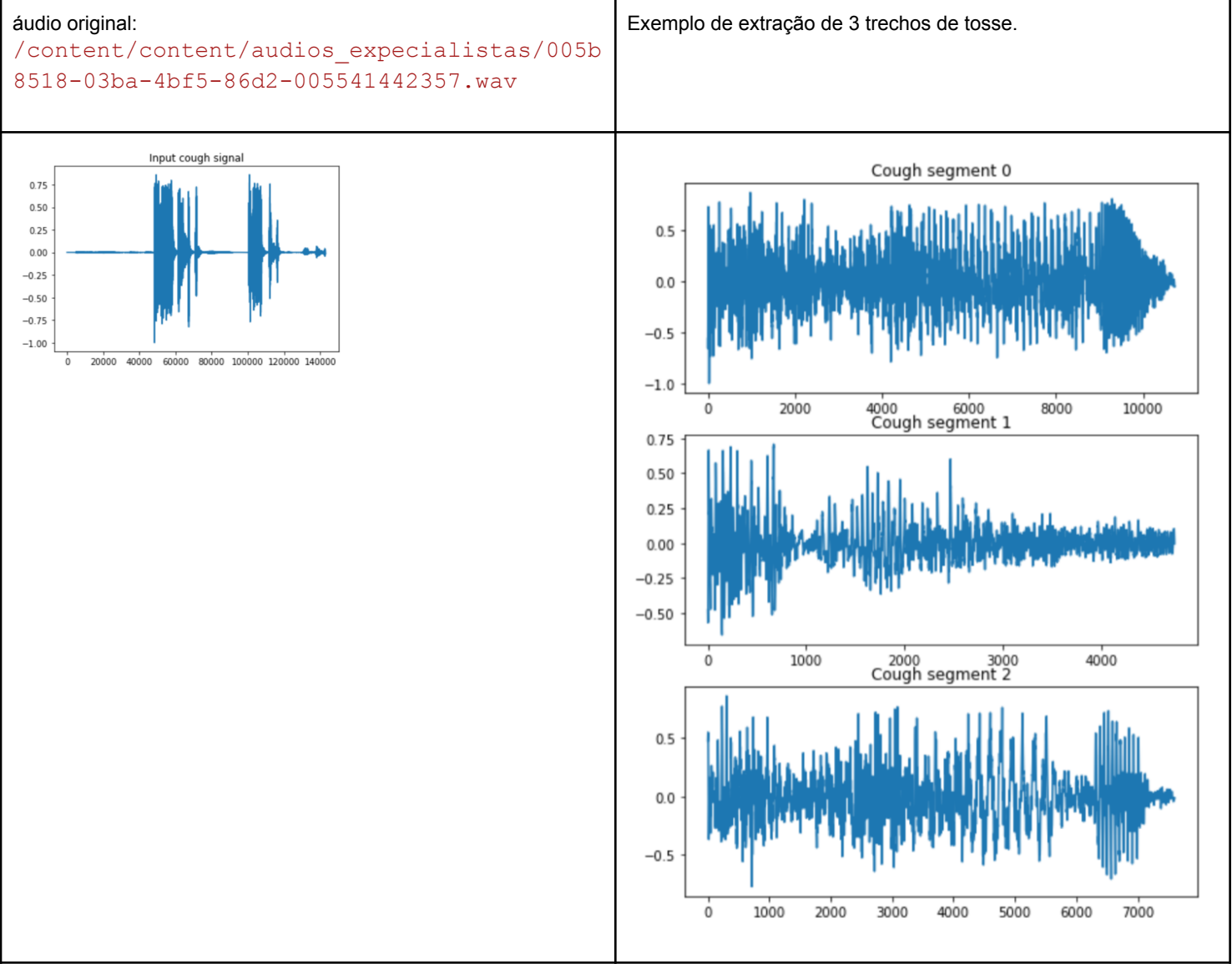
	Mel Spectrogram			MFCC		
Redes	trim 40	trim 60	trim 90	trim 40	trim 60	trim 90
ResNet50	40,37	37,36	32,60	36,73	35,33	38,25
ResNet101	38,79	37,60	36,72	34,01	35,73	35,27
ResNet152	37,64	35,47	31,47	37,78	34,41	34,81
EfficientNetB0	31,67	32,62	34,97	34,49	35,76	34,07
EfficientNetB1	38,37	36,29	36,22	31,93	36,29	39,64
EfficientNetB2	33,54	35,39	35,12	36,08	37,83	31,51
EfficientNetB3	34,37	34,16	34,23	34,13	40,62	31,25
EfficientNetB4	32,24	34,54	40,48	37,14	32,08	33,06
EfficientNetB5	36,82	35,09	32,21	32,99	33,20	34,38
EfficientNetB6	35,60	34,71	40,37	35,31	37,27	32,33
EfficientNetB7	36,45	38,35	38,36	29,75	40,18	33,23
MobileNet	38,28	40,04	41,52	36,56	35,05	29,83
MobileNetV2	34,48	35,14	37,85	32,46	33,26	34,92
DenseNet121	36,06	33,14	36,45	38,63	33,46	31,84
DenseNet169	37,45	32,12	39,23	35,43	37,36	33,78
DenseNet201	38,48	42,31	34,48	35,21	36,51	30,90
InceptionV3	38,45	34,39	33,71	35,94	37,26	36,42
InceptionResNet V2	33,18	32,03	33,34	33,16	36,51	30,88
Xception	40,17	31,49	34,58	37,64	33,73	37,69
VGG16	38,39	36,73	36,16	36,04	30,71	37,09
VGG19	28,47	36,64	36,13	39,65	33,31	40,87

Com base na tabela acima podemos verificar que a rede DenseNet201 com o MEL Spectrogram e considerando o trim de 60 obteve o melhor resultado com 42,31%. Em segundo lugar ficou a rede MobileNet com o MEL Spectrogram com o trim de 90 com 41,52%. Assim como no experimento anterior, destacamos os resultados que obtiveram valores acima de 40% de acurácia balanceada na validação.

Experimento considerando a Segmentação dos trechos de áudio

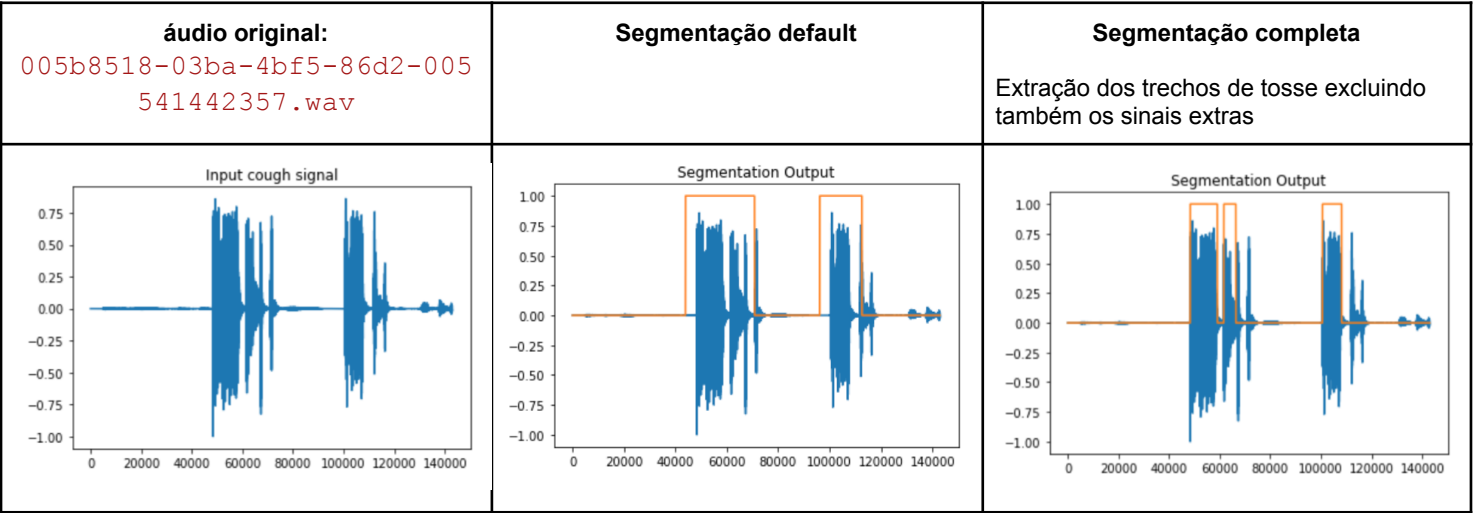
Elaboramos um experimento que tem como objetivo a identificação e a extração pontual dos trechos da tosse, desta forma efetuando a eliminação de ruídos externos como sons de televisão ao fundo, pessoas falando, entre outros. O código utilizado como base para esse experimento é de domínio público: <https://c4science.ch/diffusion/10770/browse/master/src/segmentation.py>

Essa extração tomou como referência que o sinal de destino é composto por uma respiração seguida por duas tosses e um ruído de pigarro.



Levando em consideração que a duração do áudio no dataset é variada, a identificação da tosse no decorrer do áudio pode apresentar um trecho ou vários trechos de tosse. Por padrão (default), o comprimento mínimo da tosse é definido como 200 ms (0,2 segundos) e o "enchimento" (amostras de sinais extras antes e depois de cada tosse detectada) é 200 ms (0,2 segundos). Esses limites são baseados na fisiologia dos sons da tosse.

É possível também excluir dos trechos da tosse identificados **sinais extra** como ruído de respiração e ruído de limpeza de garganta. Abaixo conseguimos observar os trechos relevantes no áudio identificados pela máscara amarela no resultado da segmentação.



Dessa forma, efetuamos dois tipos de abordagens, uma considerando a concatenação dos trechos das tosses identificados para cada áudio considerando apenas a extração das tosses **com** os sinais extras, referenciados como *Segmentação default* e a concatenação dos trechos da tosse **sem** os sinais extras, referenciados como *Segmentação completa*.

Abaixo temos a tabela com os resultados da acurácia balanceada para cada rede.

	MEL Spectrogram		MFCC	
Redes	Seg. default	Seg. completa	Seg. default	Seg. completa
ResNet50	33,70	37,69	33,81	34,26
ResNet101	34,91	35,97	34,41	39,90
ResNet152	33,88	42,1	35,95	37,56
EfficientNetB0	32,89	37,16	30,77	34,40
EfficientNetB1	38,90	33,37	32,60	34,78
EfficientNetB2	34,46	40,48	34,22	35,63
EfficientNetB3	36,26	34,17	36,54	38,13
EfficientNetB4	34,31	34,59	35,66	36,79
EfficientNetB5	32,93	30,80	42,11	32,96
EfficientNetB6	40,89	35,42	31,45	34,47
EfficientNetB7	34,07	36,74	30,65	43,18
MobileNet	40,65	37,78	33,29	36,70
MobileNetV2	32,85	35,18	35,71	33,50
DenseNet121	39,67	30,95	37,98	33,68
DenseNet169	37,67	37,35	37,35	36,91
DenseNet201	36,61	35,02	37,82	35,61
InceptionV3	35,78	40,11	35,63	35,94
InceptionResNetV2	32,75	40,04	37,03	36,07
Xception	34,76	37,54	32,51	33,15
VGG16	35,46	38,61	39,47	35,40
VGG19	35,91	42,77	37,83	36,43

Com base nesses valores temos que a melhor representação foi a MFCC com o EfficientNetB7, com 43,18% de acurácia balanceada na validação, e em segundo lugar em o MEL Spectrogram com o VGG19, com 42,77% de acurácia balanceada na validação. Novamente, destacamos os resultados com acurácia balanceada acima de 40%.

TRANSFORMERS

Transformers é uma modelagem que normalmente trata problemas relacionados a dados textuais, porém recentemente surgiu uma variação na qual pode-se também lidar com imagens. Avaliamos neste trabalho também a performance do Transformers visual sobre as imagens do MEL Spectrogram e do MFCC dos áudios da tosse.

MEL Spectrogram	MFCC
42,41	35,90

Consideramos a execução do Transformers com os dados originais do Especialista considerando a variação quanto a representação da imagem do áudio entre MEL Spectrogram e MFCC. Dentre eles, a que obteve a melhor acurácia balanceada foi o MEL Spectrogram, com 42,41% de acurácia balanceada na validação.

STACKING ENSEMBLE

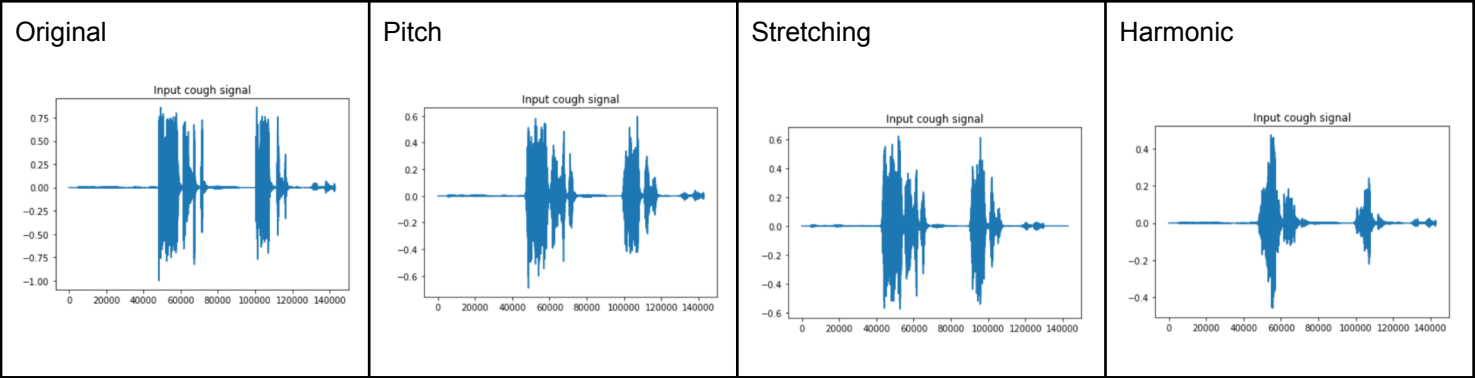
Stacking ensemble é um técnica onde a combinação de vários modelos contribuem para uma predição combinada final. Efetuamos também teste com o Stacking Ensemble considerando a combinação, (EfficientnetB1 + EfficientNetB7 + DenseNet169 + InceptionV3 + VGG19) Essa combinação de modelos são os top 5 considerando os dados originais do Especialista para o Mel Spectrogram. Para este experimento, deixamos todas as redes aprenderem, com os pesos descongelados, e concatenamos a saída das redes em uma camada densa. A Tabela abaixo apresenta os resultados obtidos.

MEL Spectrogram	MFCC
28,85	33,03

Desta forma, o melhor resultado foi obtido por imagens do tipo MFCC, que atingiu 33,03% de acurácia balanceada no conjunto de validação.

AUMENTAÇÃO DOS DADOS

A aumentoção para dados de áudios considera que as variações no som não modificam a classe original do áudio, porém trazem variabilidade para os dados durante o treinamento dos modelos. Sendo assim, para o nosso problema, consideramos as seguintes variações para os áudios: pitch (processo de mudar o tom sem afetar a velocidade do áudio), stretching (processo de mudar a velocidade ou duração do áudio sem afetar o pitch, como o efeito do som parecer mais lento) e harmonic (processo de somar um tom puro e outros tons puros harmonicamente relacionados). A seguir podemos visualizar a variação do sinal com base nessas transformações.



Foi elaborado um generator on the fly para geração do áudio tomando como base as transformações acima, dessa forma, pudemos verificar os seguintes valores de acurácia balanceada.

Realizamos o teste com os top 5 modelos de acurácia balanceada do MEL Spectrogram com os dados de especialista originais, sendo eles: EfficientnetB1, EfficientNetB7, DenseNet169, InceptionV3 e VGG19. Além destes modelos, utilizamos o Transformers para imagens.

Redes	MEL Spectrogram
Transformers	30,06
EfficientnetB1	31,65
EfficientNetB7	32,54
DenseNet169	28,20
InceptionV3	31,21
VGG19	34,73

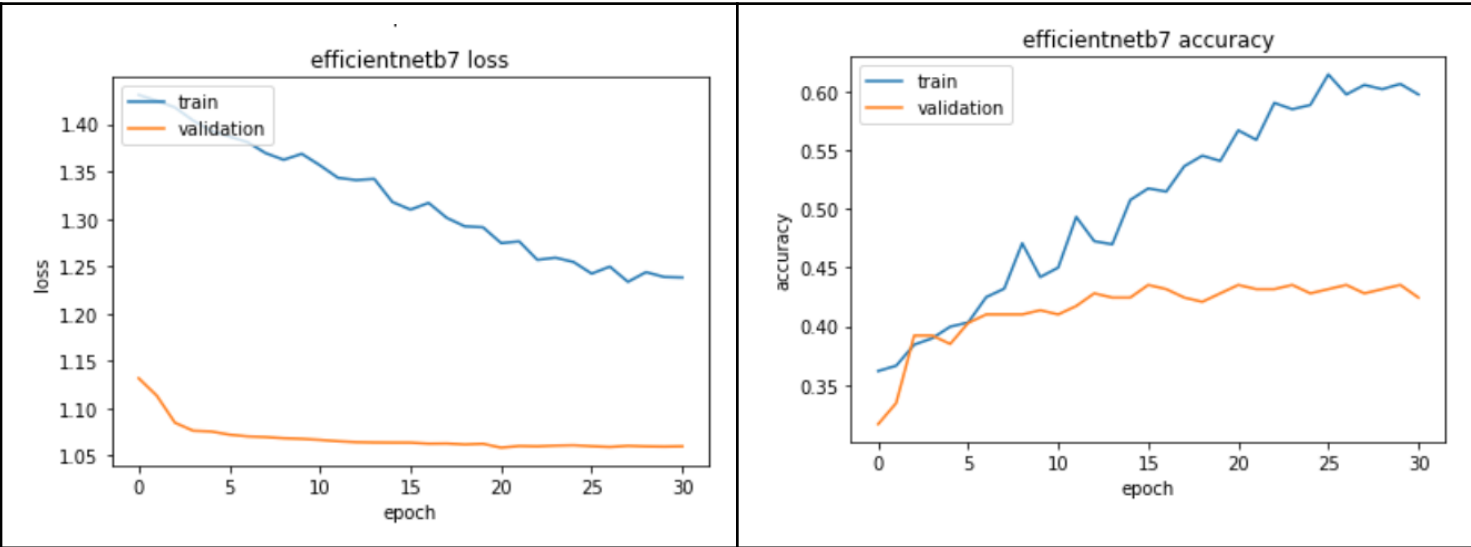
Dentre eles o melhor resultado ficou com o VGG19, com 34,73% de acurácia balanceada na validação.

COMPARAÇÃO DOS RESULTADOS

Com base nos resultados abaixo temos que a abordagem que obteve a melhor acurácia balanceada foi a Segmentação Completa utilizando a rede do EfficientNetB7 e a representação do MFCC com 43,18%.

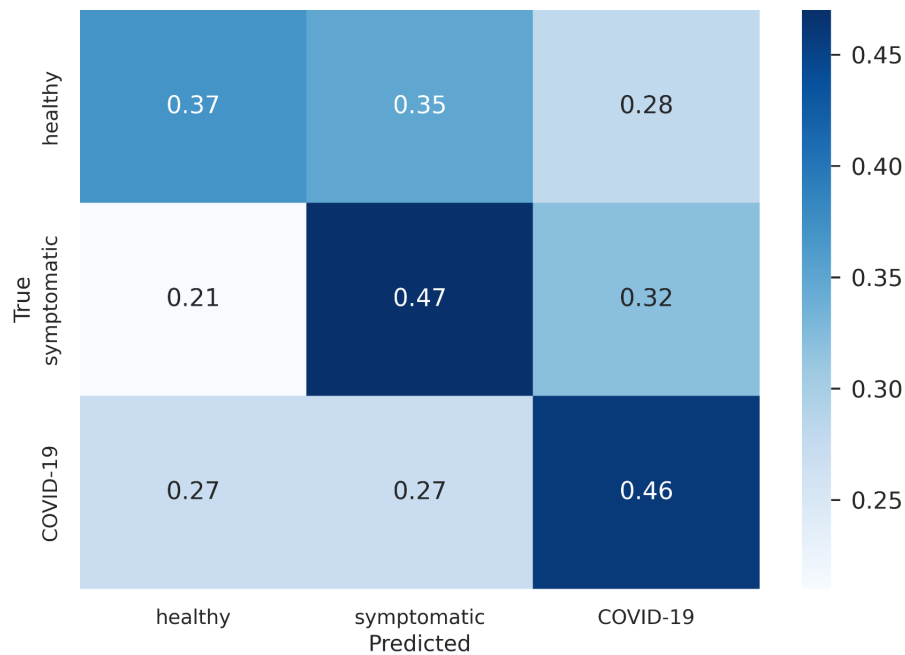
	MEL Spectrogram	MFCC
Transfer Learning com as redes do imageNet	41,34 [EfficientNetB7]	41,42 [EfficientNetB7]
Trim	42,31 [DenseNet201- trim de 60]	40,87 [VGG19 - trim 90]
Segmentação	42,77 [Seg. Completa - VGG19]	43,18 [Seg. Completa - EfficientNetB7]
Transformers	42,41	35,90
Stacking Ensemble	28,85	33,03
Aumentação dos dados	34,73	-

Abaixo podemos visualizar as curvas de avaliação do melhor modelo encontrado.

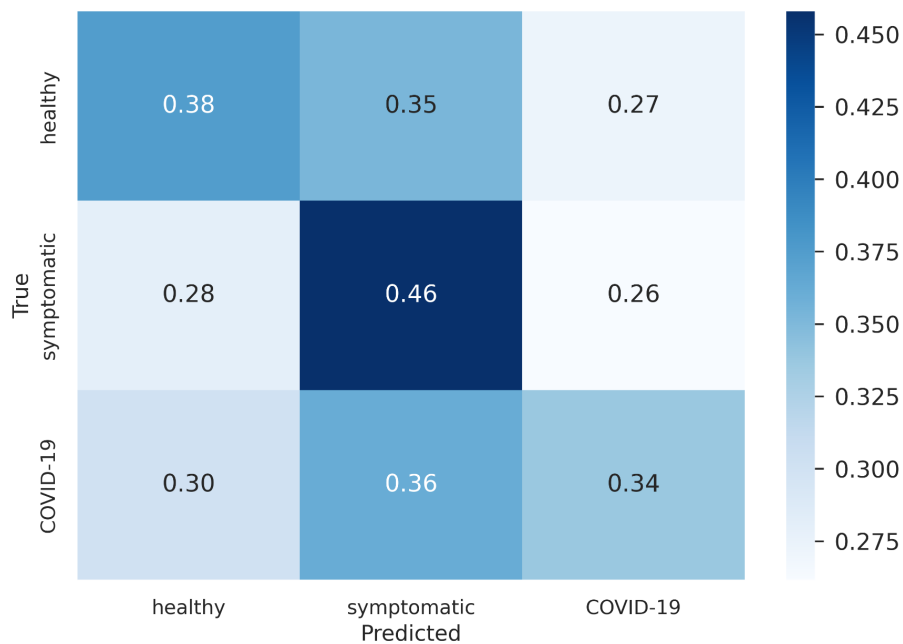


Com base nas curvas não foi identificado presença de overfitting dos dados.

A matriz de confusão do melhor modelo é apresentada na Figura abaixo.



Após a escolha do melhor modelo, re-treinamos com os mesmos parâmetros e avaliamos no conjunto de teste. Como resultado, obtivemos 39,75% de acurácia balanceada no conjunto de teste. A Figura abaixo mostra a matriz de confusão no conjunto de teste.



CONCLUSÃO

Neste trabalho, exploramos desde a seleção das features pertinentes para responder a pergunta do projeto, o pré-processamento e da representação do áudio em imagem pelo Mel Espectrograma e MFCC. Treinamento e avaliação de diversos modelos tais como transfer learnings, transformers, stacking ensemble, e aumento de dados para áudio e o melhor valor de acurácia balanceada obtida foi de 43,19% com a segmentação dos trechos de tosse, MFCC e a rede EfficientNetB7. No conjunto de teste, o valor foi igual a 39,75% de acurácia balanceada. O resultado final nos leva a concluir que a problemática de encontrar padrão em áudios de tosse não é trivial. Pois mesmo com o uso de técnicas consideradas estado da arte como o caso do transformers para imagens não foi possível obter resultados significantes.

Acreditamos que para ter um resultado mais favorável seria preciso aumentar significativamente os áudios relacionados ao COVID-19. A tentativa de aumento sintética dos dados provavelmente não gerou áudios que agregassem novas informações a ponto do modelo melhorar com o aprendizado.