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

Na primeira etapa do trabalho foi explorado o entendimento das features do dataset e foi elaborada a seleção dos dados pertinentes para a resolução da pergunta definida além do baseline do projeto.

DATASET

INSPEÇÃO DOS DADOS

O dataset possui 20.000 tanto de registros 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 do ambiente tais como pessoas conversando e aparelho de televisão ligados, assim como áudios sem tosse (apenas 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

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.

Outro ponto relevante de filtragem dos dados foi a identificação que alguns exemplos do dataset não possuem avaliação por médicos especialistas e outros possuem de 1 avaliação ao máximo de 3 avaliações. Dessa forma, removemos também áudios que não foram avaliados por especialistas. A avaliação do especialista se mostrou relevante para o problema que estamos tratando, pois através dela é constatada a veracidade do diagnóstico e é também anotada features extras sobre o áudio. Após a seleção dos dados a base ficou com 7.036 registros.

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 MCFF, conforme mostram as fuguras abaixo.

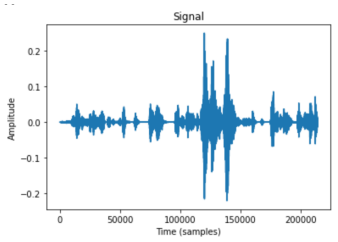


Fig. 1 Representação do Sinal considerando a amplitude pelo tempo.

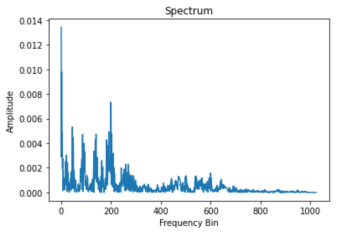


Fig. 2. Representação do Spectrum considerando a amplitude pela frequência.

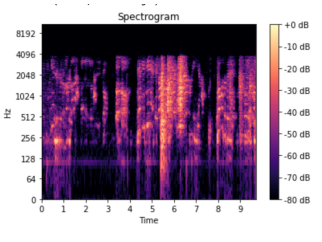


Fig. 3 Imagem do Mel Spectrogram

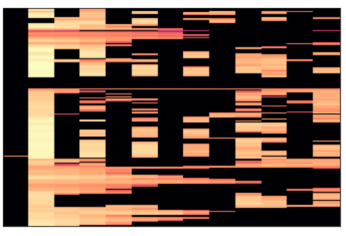


Fig. 4 Imagem do MCFF

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

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.

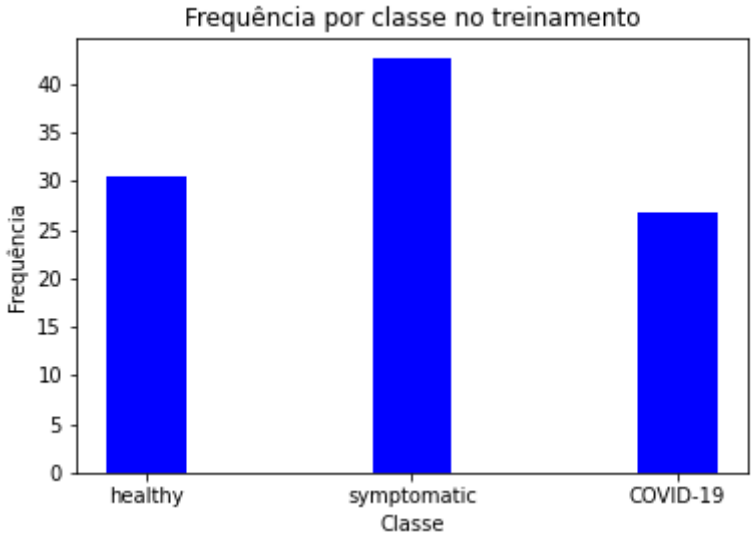
Com o conjunto de treinamento já definido, elaboramos inicialmente duas vertentes de treinamento para a construção dos modelos, uma considerando apenas os dados anotados por especialistas no conjunto original de treinamento e a outra considerando dados adicionais no treinamento. Os dados adicionais agregados no conjunto de treinamento são providos da inclusão dos áudio que possuam 50% de existência de tosse avaliada pela variável *cough_detec*t e que não foram avaliados 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.

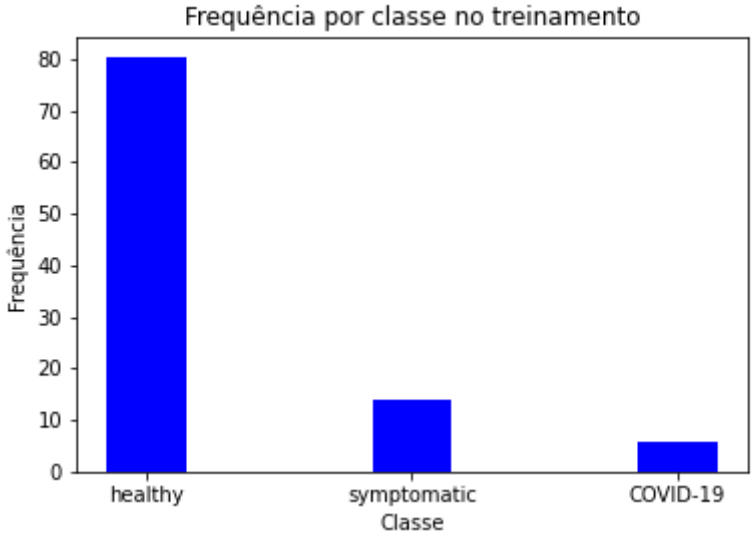
Os pesos das classes para cada uma dessas vertentes são descritos abaixo. Resolvemos utilizar os pesos para cada classe para que os modelos tivessem uma penalização maior para erros nas classes minoritárias. Calculamos os pesos a partir da divisão da classe com mais amostras pela classe em questão, de 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

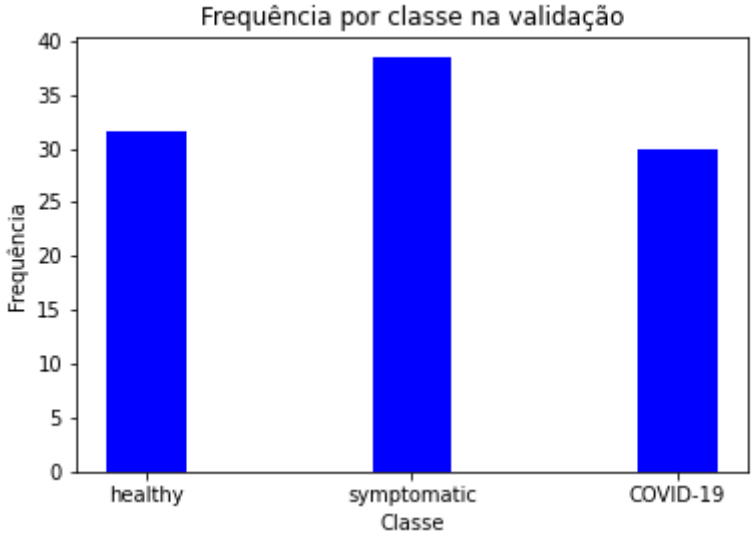
Para o experimento utilizando apenas os dados anotados por especialistas, a distribuição no treinamento é dada pela figura abaixo. A classe mais comum é symptomatic, com mais de 40% dos dados, seguida por healthy, com 30% dos dados, e COVID-19, com 27% dos dados.



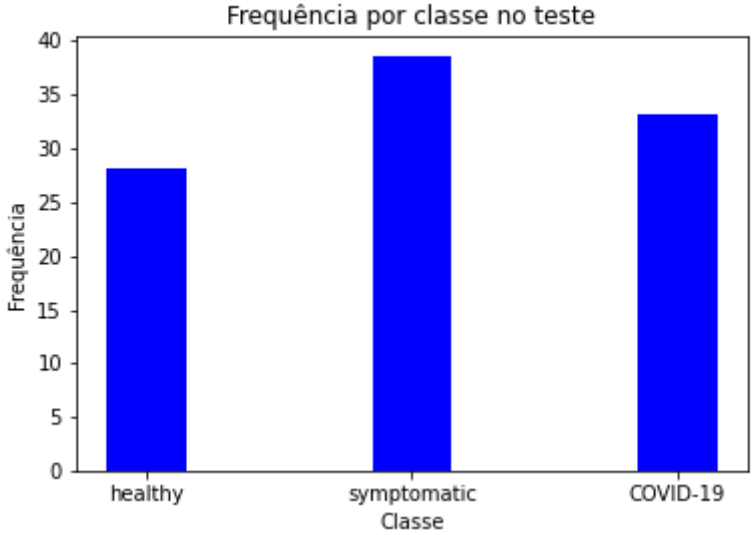
Para o experimento com dados adicionais, a distribuição muda no treinamento. A classe mais frequente é healthy, com 80% dos dados, seguido por symptomatic, com mais de 10% dos dados, e COVID-19, com pouco menos de 10% dos dados.



Já no conjunto de validação (utilizada tanto para os experimentos com apenas os dados de especialistas e dados adicionais), 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 sympotimatic, mais de 30% de COVID-19 e cerca de 27% como healthy.

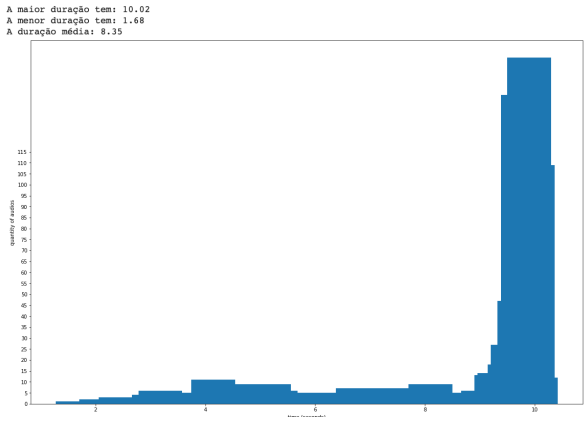
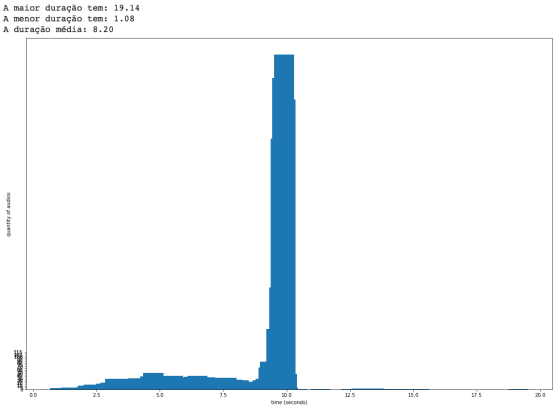
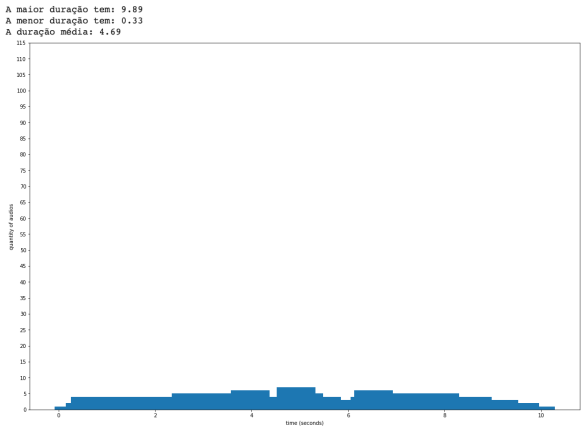
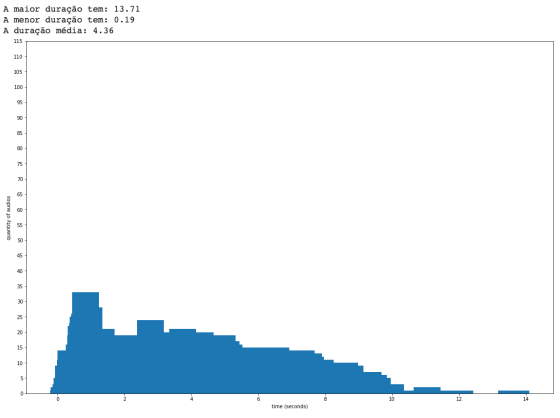


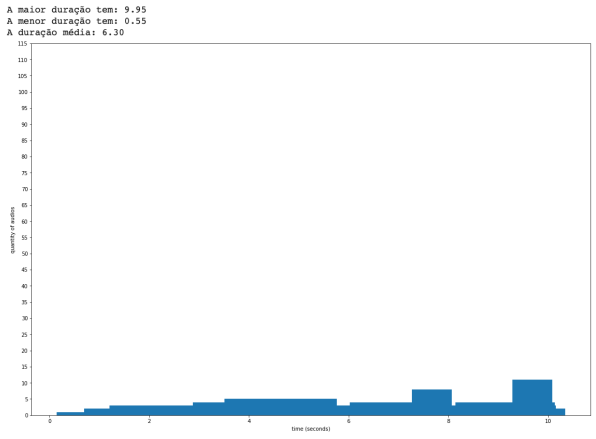
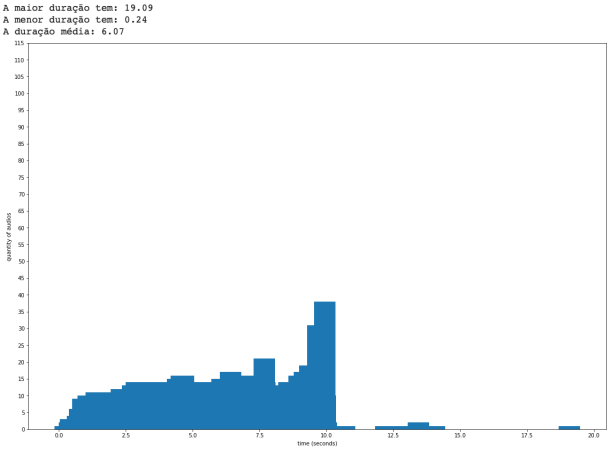
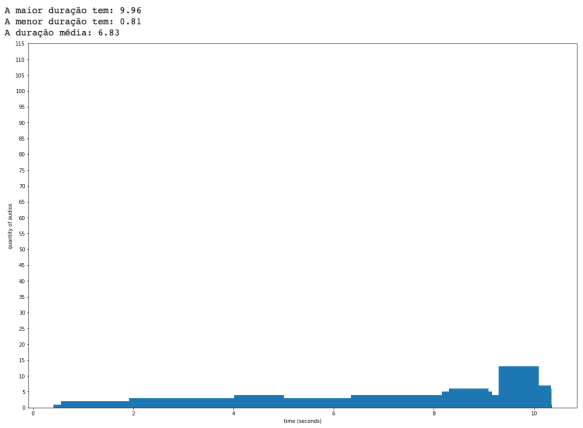
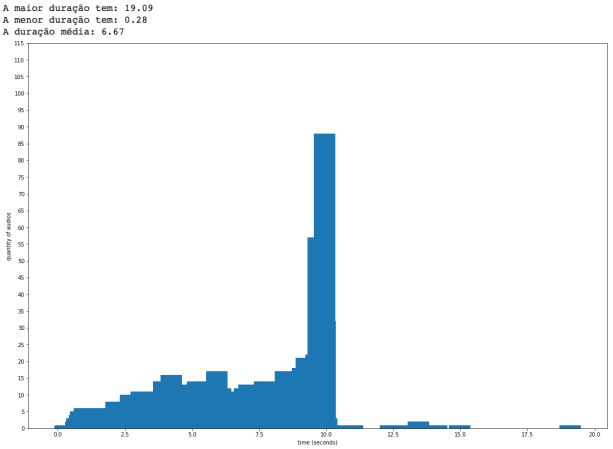
Análise da duração do áudio e do trim considerando as duas vertentes de treinamento

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.

A partir dos gráficos abaixo podemos inferir que a vertente 2 possui mais áudios longos chegando a 19.14 segundos de duração que na vertente 1 com a maior duração sendo 10.02.

Podemos inferir também 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	Vertente 2 : Treinamento considerando dados adicionais
Original	 A maior duração tem: 10.02 A menor duração tem: 1.68 A duração média: 8.35	 A maior duração tem: 19.14 A menor duração tem: 1.08 A duração média: 8.20
28 decibéis	 A maior duração tem: 9.89 A menor duração tem: 0.33 A duração média: 4.69	 A maior duração tem: 13.71 A menor duração tem: 0.19 A duração média: 4.36

50 decibéis	 A maior duração tem: 9.95 A menor duração tem: 0.55 A duração média: 6.30	 A maior duração tem: 19.09 A menor duração tem: 0.24 A duração média: 6.07
60 decibéis	 A maior duração tem: 9.96 A menor duração tem: 0.81 A duração média: 6.83	 A maior duração tem: 19.09 A menor duração tem: 0.28 A duração média: 6.67

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.

BASELINE

Propusemos inicialmente a execução do mesmo conjunto pré-processado sobre 21 redes diferentes da **imageNet** para avaliar seu desempenho e iremos considerar o nosso baseline, a melhor avaliada dentre elas.

As redes analisadas são:

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

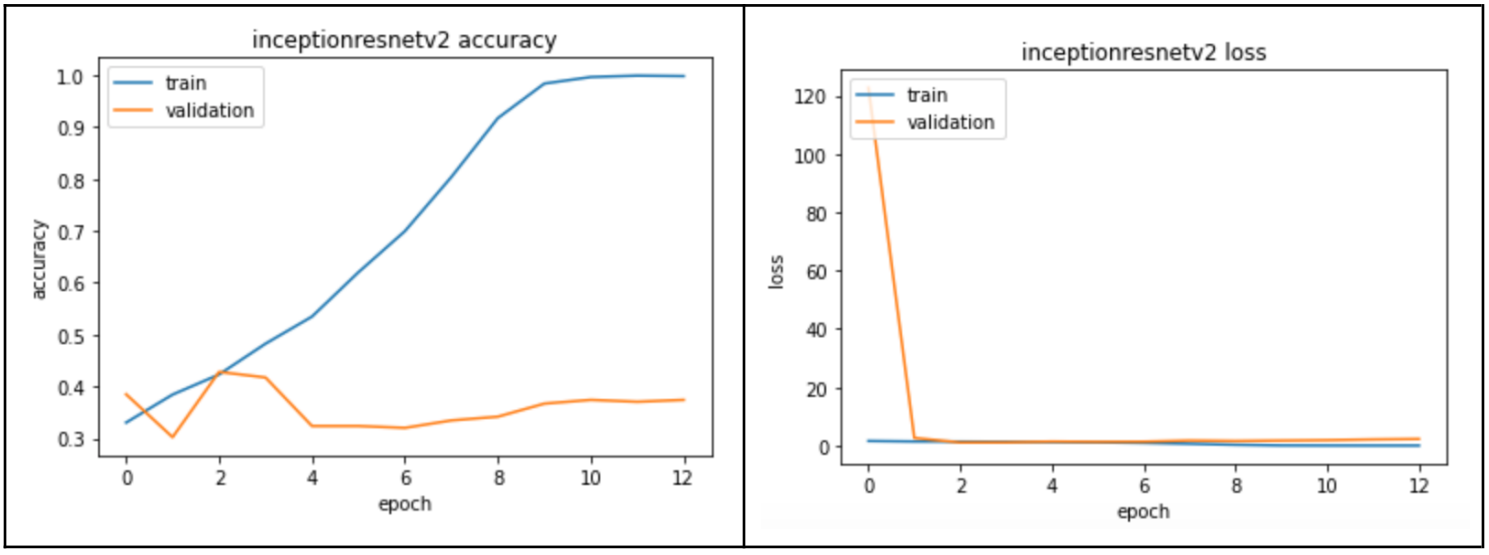
Experimento com os dados anotados por especialistas

Obtivemos como resultado que o **InceptionResNetv2** obteve a melhor acurácia balanceada na validação, com 42,63%.

Rede	Acurácia Balanceada
ResNet50	32,74
ResNet101	37,84
ResNet152	37,89
EfficientNetB0	38,52
EfficientNetB1	36,94
EfficientNetB2	33,01
EfficientNetB3	38,37
EfficientNetB4	37,87
EfficientNetB5	33,58
EfficientNetB6	33,71
EfficientNetB7	35,00
MobileNet	38,29
MobileNetV2	32,44
DenseNet121	34,14
DenseNet169	34,99
DenseNet201	35,89
InceptionV3	38,38
InceptionResNetV2	**42,63**
Xception	33,33
VGG16	33,33
VGG19	33,33

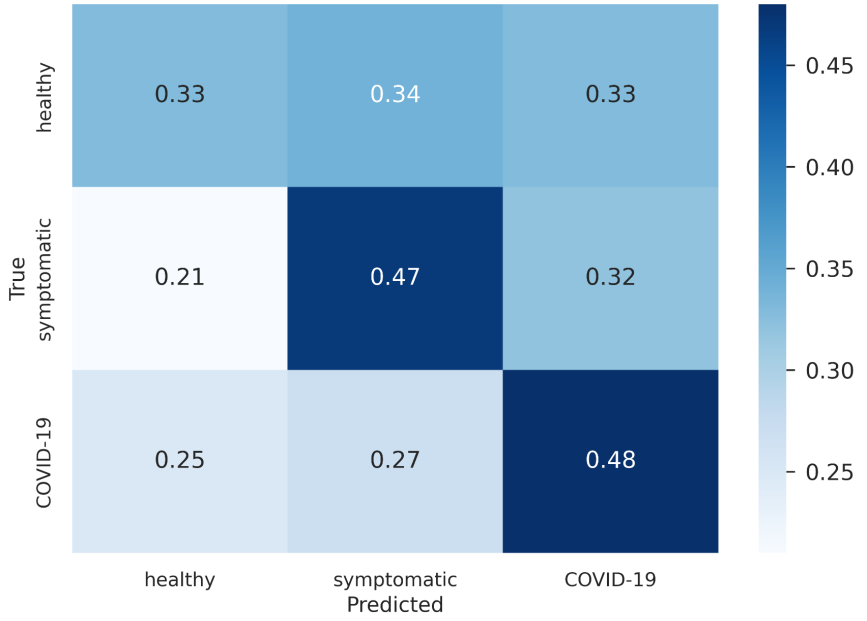
Abaixo temos a arquitetura do **InceptionResNetv2**, junto com os gráficos da acurácia e da loss de treino e validação.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_resnet_v2 (Functio (None, 1536)              54336736
_________________________________________________________________
dense_8 (Dense)              (None, 3)                 4611
=================================================================
Total params: 54,341,347
Trainable params: 54,280,803
Non-trainable params: 60,544
```



No gráfico da acurácia vemos que a partir da época 3 houve overfitting fazendo o conjunto de validação se distanciar do de treinamento. Quanto ao gráfico da loss a partir da época 1 a loss de treinamento e validação apresentam valores sem nenhuma diferença estatística significativa.

A figura abaixo mostra a matriz de confusão na validação obtida a partir das predições da **InceptionResNetv2**. Neste conjunto, são 88 dados da classe healthy, 107 dados da classe symptomatic e 83 dados da classe COVID-19. Os resultados foram normalizados por linha. Como resultado, podemos perceber que quase metade dos dados das classes symptomatic e COVID-19 foram classificados corretamente, enquanto apenas 33% dos dados da classe healthy foram classificados corretamente. Dentre os erros, os piores são de dados das classes symptomatic e COVID-19 classificados como healthy, onde são pacientes que possuem ou podem possuir a COVID-19 mas foram classificados como saudáveis.



Experimento considerando os dados adicionais

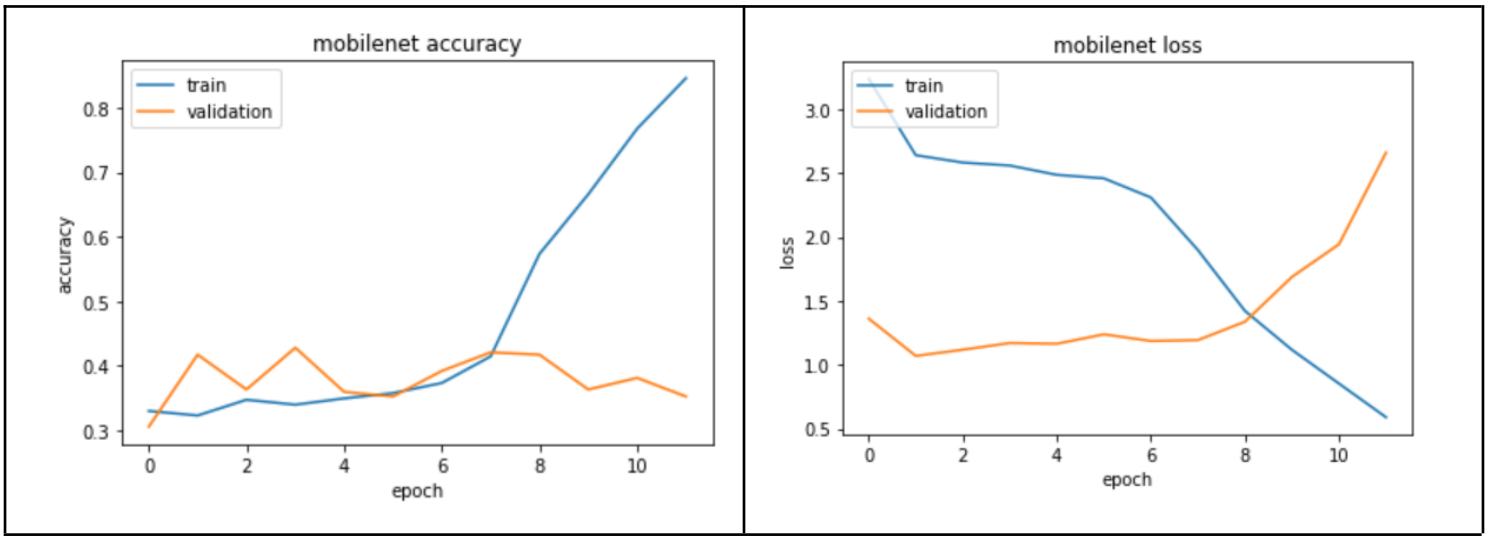
Tomando como base a tabela abaixo, que sinaliza a acurácia balanceada em cada uma das redes analisadas, temos que a rede **MobileNet** obteve a melhor acurácia balanceada com 42,50%. Em segundo lugar tivemos a **EfficientNetB3** com 41,26%.

A **MobileNet** foi configurada com o *early stopping* e com o *Reduce learning rate* que reduz a taxa de aprendizado quando não tem melhora na loss da validação.

Rede	Acurácia Balanceada na Validação
ResNet50	36,10
ResNet101	35,30
ResNet152	32,80
EfficientNetB0	38,37
EfficientNetB1	39,75
EfficientNetB2	37,82
EfficientNetB3	41,26
EfficientNetB4	38,30
EfficientNetB5	41,01
EfficientNetB6	39,68
EfficientNetB7	36,86
MobileNet	**42,50**
MobileNetV2	32,60
DenseNet121	32,63
DenseNet169	40,78
DenseNet201	38,78
InceptionV3	33,33
InceptionResNetV2	35,17
Xception	36,97
VGG16	33,33
VGG19	33,33

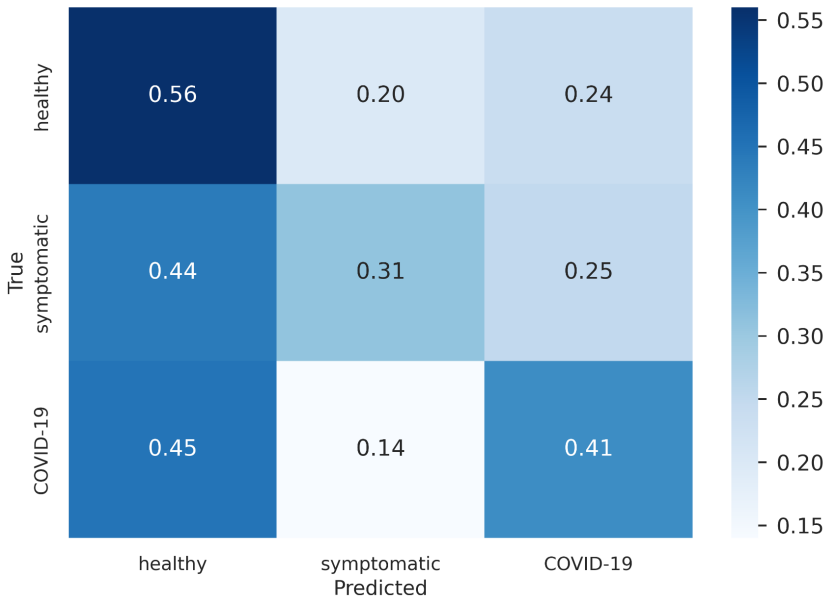
Abaixo temos a arquitetura do **MobileNet**, junto com os gráficos da acurácia e da loss de treino e validação.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
mobilenet_1.00_224 (Function (None, 1024)              3228864
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 3075
=================================================================
Total params: 3,231,939
Trainable params: 3,210,051
Non-trainable params: 21,888
```



Baseando no gráfico da loss podemos observar o comportamento claro onde a curva da loss de treinamento vai decaindo enquanto a curva da loss de validação vai aumentando caracterizando um comportamento claro de **overfitting**.

A figura abaixo mostra a matriz de confusão da **MobileNet** no conjunto de validação. A rede teve uma taxa de acerto acima de 50% para dados da classe healthy, 31% dos dados da classe symptomatic e 41% dos dados da classe COVID-19. Mesmo possuindo o melhor valor de acurácia balanceada na validação, podemos perceber muitos erros para a classificação de dados das classes symptomatic e COVID-19 para a classe healthy, sendo erros preocupantes em situações médicas, onde são pacientes que possuem ou podem possuir COVID-19 mas são diagnosticados como saudáveis. Como uma possível causa deste tipo de erro, podemos atribuir ao grande nível de desbalanceamento no conjunto de treinamento deste experimento, onde existem muitos dados da classe healthy.



COMPARAÇÃO DOS RESULTADOS

Os melhores resultados levando em consideração a acurácia balanceada são descritos na tabela abaixo:

InceptionResNetv2 obteve a melhor acurácia balanceada com 42,63% na validação	**MobileNet** obteve a melhor acurácia balanceada com 42,50% na validação
Com anotações de especialistas	**Com os dados adicionais**

Ou seja, podemos perceber que a rede **InceptionResNetv2** obteve o melhor valor de acurácia balanceada com **42,63%** na validação. Dessa forma a consideramos o baseline desse projeto.

CONCLUSÃO

Nesta primeira etapa exploramos a seleção das features pertinentes para responder a pergunta do projeto, além do pré-processamento e da representação do áudio em imagem pelo Mel Espectrograma, após isso definimos o baseline com base no pré-processamento inicial dos dados. O baseline definido foi a rede **InceptionResNetv2**.