Mosaic Problem Statement 1

Team Name-Weldright
Members-

- **Prachi Kumar**
- **Raunak Pandey**
- **Samarth Tankasali**

Mosaic PS-1 2023

PROBLEM STATEMENT

We were provided with audios of lung sounds which could belong to one of four cases, each representing a lung disease/disorder.

OUR UNDERSTANDING

The problem statement can be understood as that of image recognition from that of audio recognition. Using the **librosa** library we can extract various features from the audio(as images) and then feed the images to a CNN model.
We tried using **ensemble deep learning** in order to use various features and improve accuracy.

DATA PREPROCESSING

We were provided with multiple audio files along with a corresponding text file which included information regarding each respiratory cycle and observation made in that cycle. Each text file contained many information such as patient id,recording index etc. separated by '_'.we extracted patient id and mode from the file name.

An observation can be made looking at each text file that multiple respiratory cycles were there in one audio file. So our next task was to extract those parts of the audio file in which respiratory cycle has occurred by utilizing it's corresponding start and end time which is given in the text file.

Also in order to make length of each files the same we padded the shorter audios and clipped the longer ones. Finally the processed audio files were stored in a folder separately.

A new dataframe was also created that contained audio start,end times, patient id ,method of collection of audio and also the filename for each audio.

After preprocessing the data, the notebook defines the neural network model.

MODEL

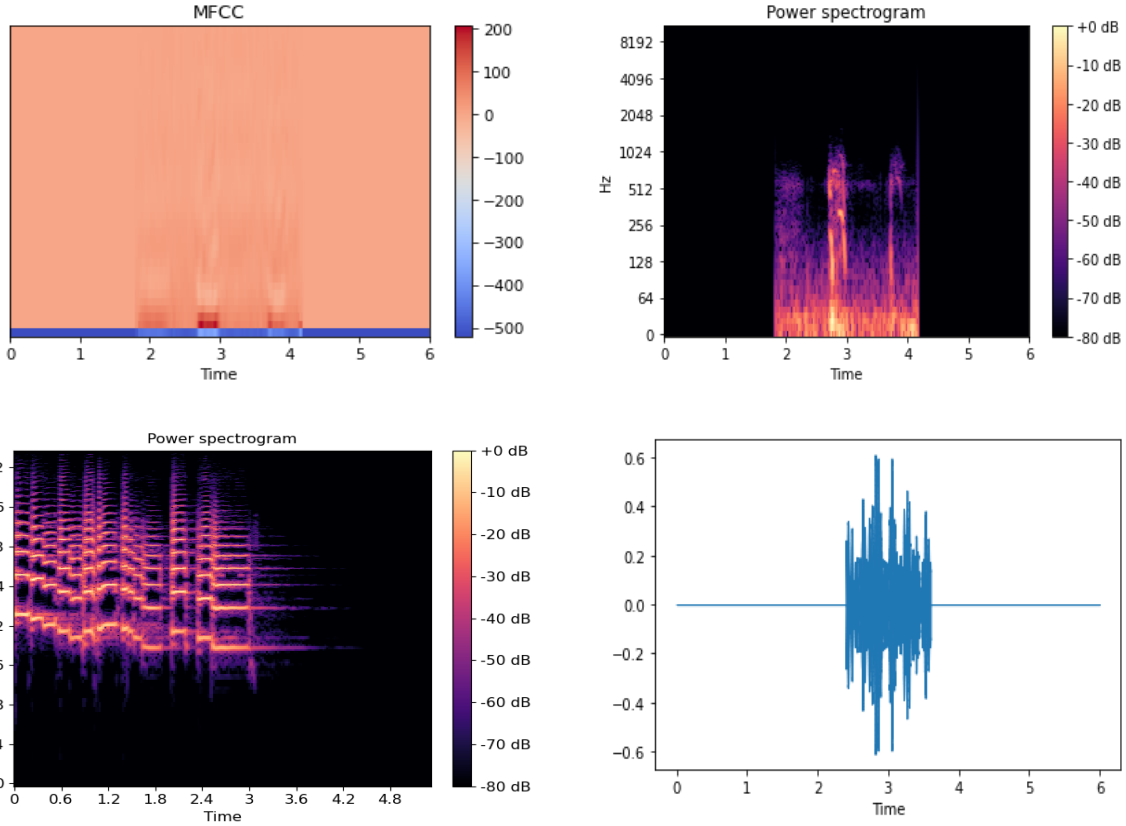
Now using the mentioned dataframe we first created training and testing data that is we separated the dataframe into training and testing parts. We would use now these to train our model.
Before proceeding to train the neural network first we performed label encoding and stored the labels as a dictionary.

Feature extraction

Now we extracted features from the arrays and using librosa library.
We have converted original audio into its

- Short time fourier Transform
- Mel Spectogram
- MFCC

The information about these transformations are stored in a numpy array which is later fed into a model.

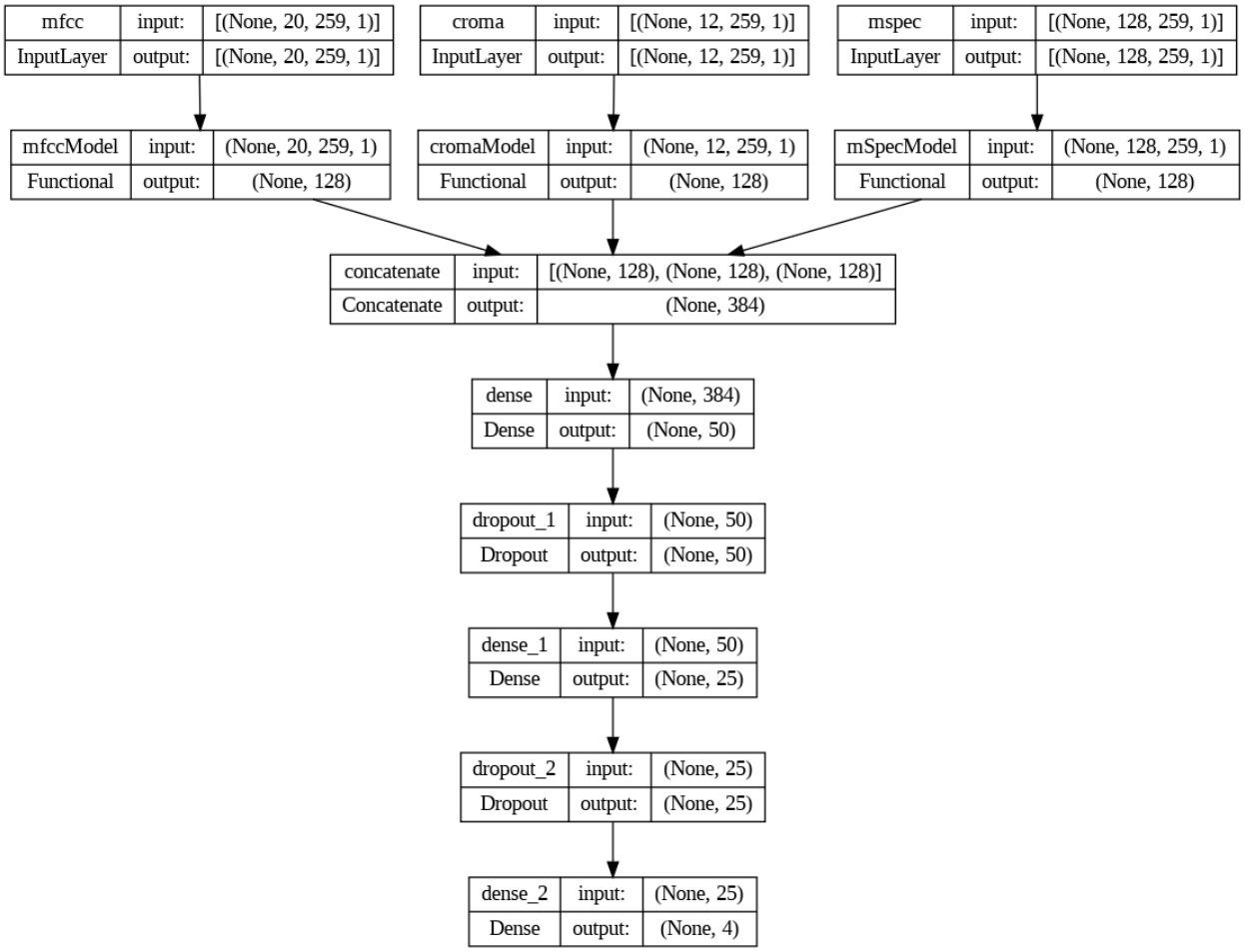


Coming to model architecture, we have approached the problem statement using the idea of ensemble(deep) neural network.

Model Architecture

Here the seperate features are trained on three sperate CNNs and the results are in turn put through another neural network which finally classifies the audios as one of four categories: both ,weezels ,crackles ,normal.

Each layers input is an array of size equal to the size of the transformation performed by librosa library. The outputs of these three layers is then given to final network which inturn givers 4 outputs which are the estimated likeliness of an audio sample belonging to a particular class.



The neural network defined above is a convolutional neural network (CNN) that takes a input of 3D array and produces a fixed-length output vector. The network architecture is composed of several convolutional layers, followed by batch normalization, activation, and max pooling layers.

The first convolutional layer has 32 filters with a kernel size of 5 and a stride of (2,3), which means that the filters move horizontally by two and vertically by three pixels. The subsequent batch normalization layer normalizes the output of the convolutional layer to accelerate the training process. The activation function used is ReLU, which helps to introduce non-linearity in the network and improve its expressive power. The first max pooling layer reduces the spatial size of the output by a factor of 2.

The second convolutional layer has 64 filters with a kernel size of 3 and a stride of (2,2). The same batch normalization and activation functions are applied before a second max pooling operation is performed.

The third convolutional layer has 96 filters with a kernel size of 2 and no stride. The activation and batch normalization functions are again applied, followed by max pooling with a pool size of 2.

The fourth and final convolutional layer has 128 filters with a kernel size of 2 and no stride. The output is batch normalized, activated, and then globally max pooled to produce a fixed-length feature vector.

The same architecture is there for all the 3 inputs. The outputs of these 3 models is then fed to a final model.The final model is modelled like a neural network that takes three different inputs (MFCC, chroma, and mel spectrogram) and concatenates them before passing through several dense layers with dropout regularization. The output layer has four nodes, corresponding to four output classes. The resulting model is named "Net".

The resulting model, can be used for classification of lung sounds.

The next step is to train the model on the training data for a specified number of epochs.

Finally we saved the model and then we are now using it to make predictions.

ACCURACY

The accuracy of model came out to be around [1]**75%-79%.**

CONCLUSION

The model is successful in detecting lung sound defects about **78%** of the time. The ensemble model can be improved by providing additional features to the ensemble. However that would come at cost of computational resources.

[1] When trained for different times and epochs.