Lab #3
Speech Emotion Recognition

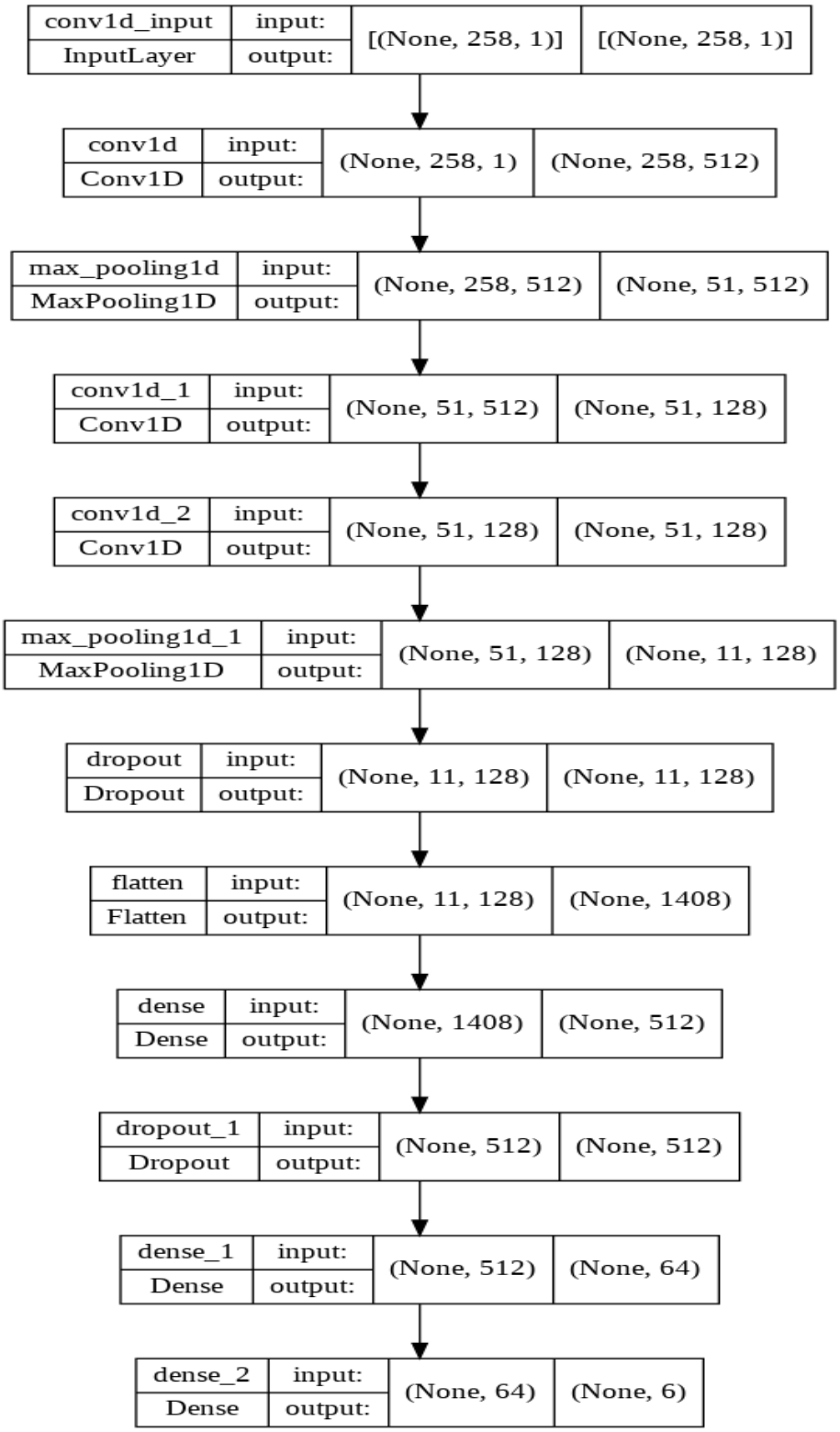
Mariam Saeed **Nada Hassan** **Mohamed Metwalli**

First Model 1D convolution

- **Feature Extraction**
  - :Total of 258 features extracted using the means of
    - **Zero-crossing rate:** which is a measure of the number of times in a given time interval/frame that the amplitude of the speech signals passes .through a value of zero
    - **energy:** which is the total magnitude of the signal, .i.e. how loud the signal is
    - **mfcc(n_mfcc=100):** which is a replication of the human hearing system intending to artificially implement the ear's working principle. 100 features .were used
    - **chroma shift:** which computes a chromagram from a waveform that captures harmonic and melodic characteristics of music
    - **Mel spectrogram**: which is a spectrogram where the frequencies are converted to the mel scale
    - **roll of:** which is This is a measure of the amount .of the right-skewness of the power spectrum
    - **centroids:** spectral centroid is the location of the center of mass of the spectrum. Since the audio files are digital signals and the spectral centroid is a measure that can be useful in the characterization of the spectrum of the audio file .signal
    - **contrast:** In an audio signal, the spectral contrast is the measure of the energy of frequency at each timestamp. is a way to measure that energy .variation over time
    - **bandwidth:** is the difference between the upper and lower frequencies in a continuous band of .frequencies

- **tonnetz:** which represents a lattice diagram represents tonal space

Structure of The Model



- ## Improvement of The Model
  - We started with the model in the pdf assignment with pure data which was about 7442 samples of data and features 258
  - Then, we used augmentation on the data such as noise, stretch and pitch we got over 29,000 samples of data which helped in training the model better
  - Then, we made some modifications to the layers and we noticed that adding more layers wasn't improving the accuracy so we removed some layers and added dropout layers to the model
  - We ran the model a couple of times with different number of epochs and we found that 200 epochs gave us the best results
- ## Results
  - **confusion matrix**

The rows represents the true values or observations
The columns respresent the model's predictions

	angry	disgust	fear	happy	neutral	sad
angry	1186	50	46	157	49	7
disgust	44	1027	36	92	152	134
fear	47	87	937	108	91	247
happy	126	109	80	1112	107	35
neutral	5	134	37	45	982	94
sad	2	116	92	21	133	1204

  - **accuracy**

```
Training Accuracy:  0.9529679417610168
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/preprocessing
  y = column_or_1d(y, warn=True)
Testing Accuracy:  0.7219796180725098
```

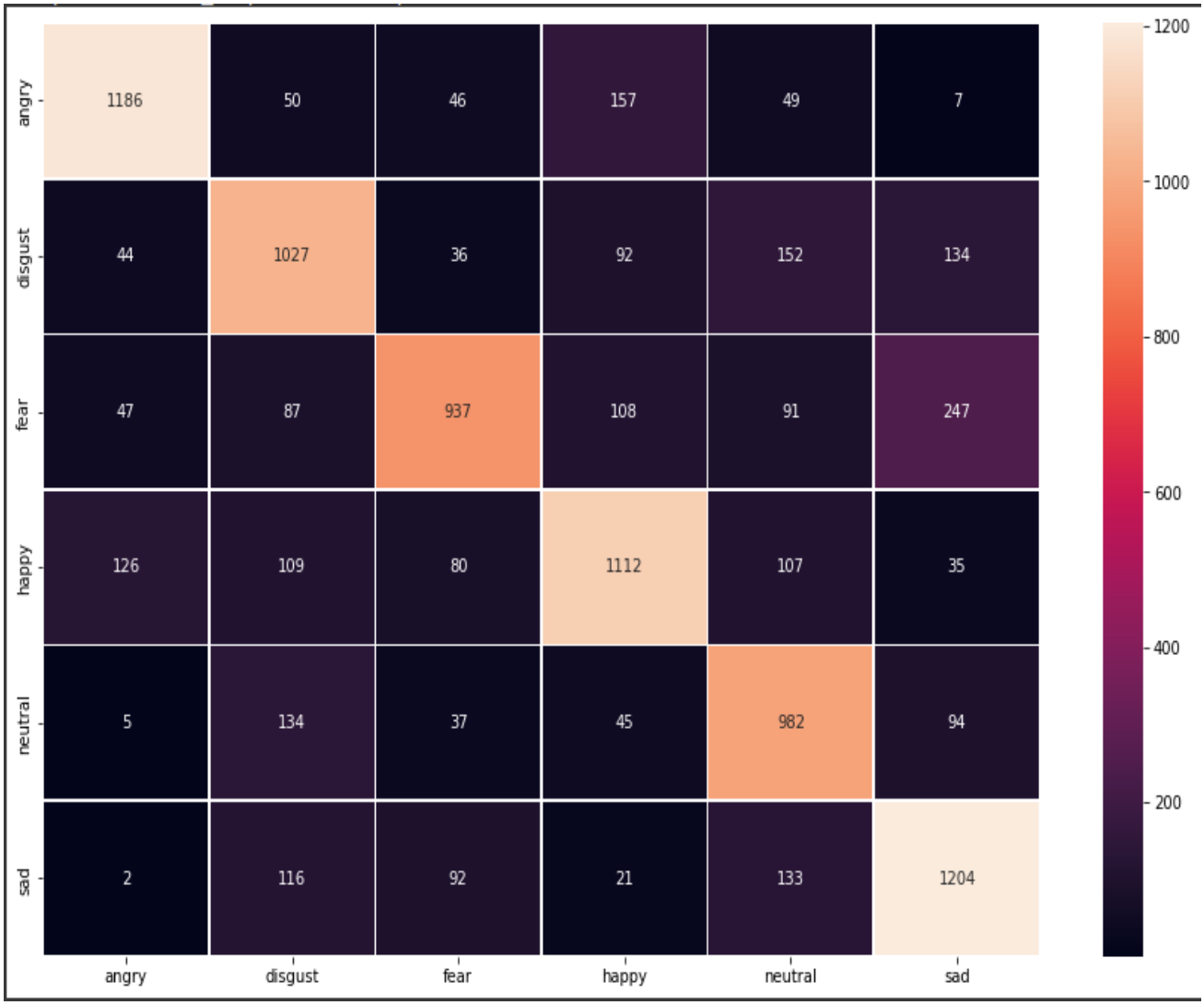
- **f-score, precision and recall**

```
Percision:  0.7251878711572747
Recall:  0.7227134395174205
F-score:  0.7215634677727829
```

- **report of all classes**

	precision	recall	f1-score	support
angry	0.84	0.79	0.82	1495
disgust	0.67	0.69	0.68	1485
fear	0.76	0.62	0.68	1517
happy	0.72	0.71	0.72	1569
neutral	0.65	0.76	0.70	1297
sad	0.70	0.77	0.73	1568
accuracy			0.72	8931
macro avg	0.73	0.72	0.72	8931
weighted avg	0.73	0.72	0.72	8931

- **most confusing classes**



According to classification report disgust and fear classes has the lowest f1-score

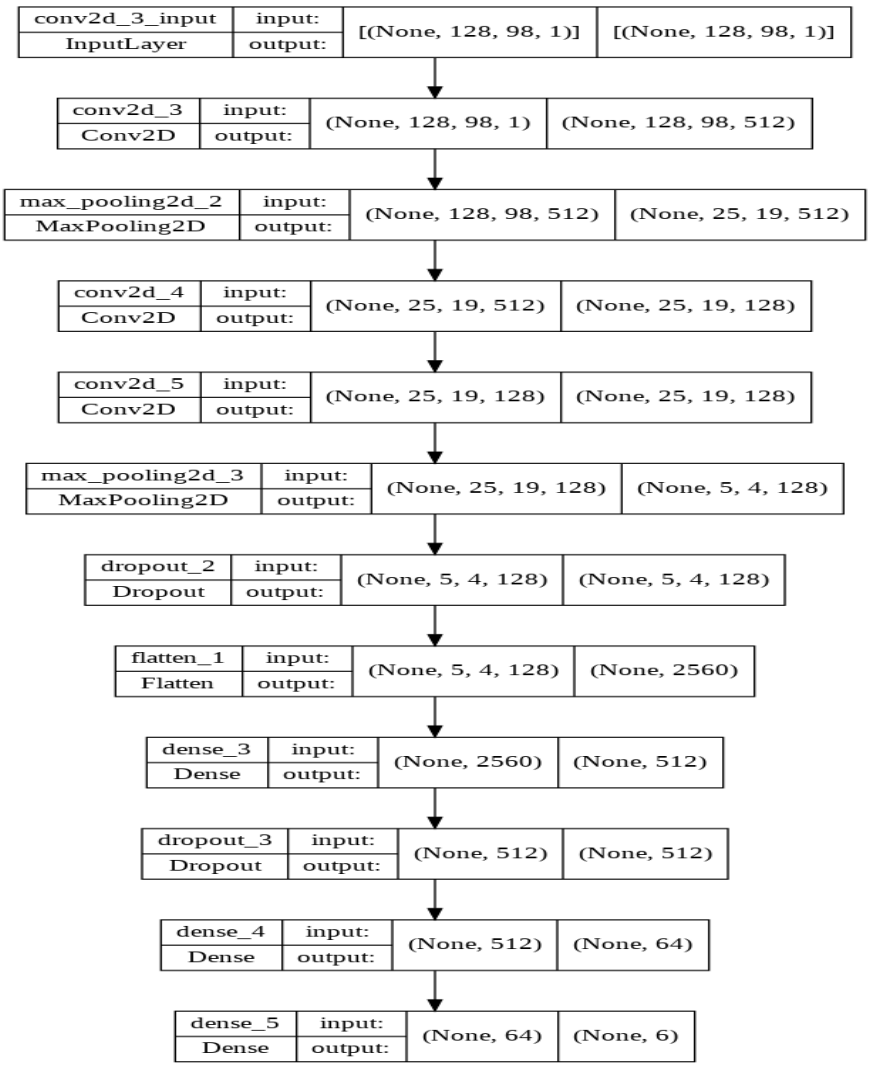
.so the model is confused about those classes

The previous heatmap shows how the model predicts disgust and .fear

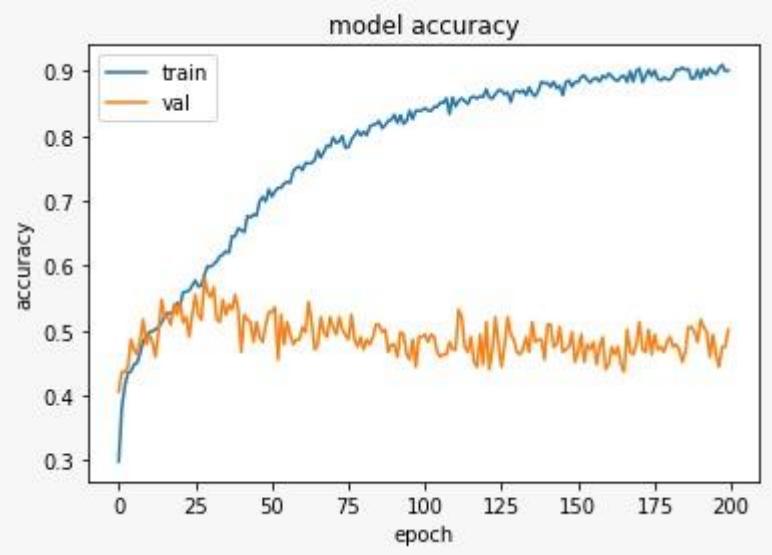
Second Model 2D convolution

- ## Feature Extraction
  - As each sound has a variable length the data needed :padding so the data was processed as follows
    - load the data into a dataframe
    - replace NaN values with the mean of audio (padding)
    - take only 50000 out of approximately 80000 columns
    - extract Mel spectrogram of each audio
    - reshape data to be suitable for input

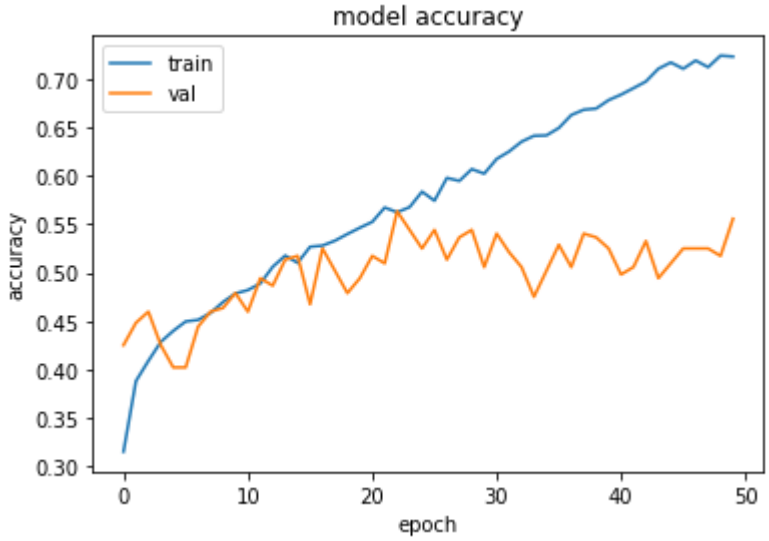
- ## Structure of The Model



- ## Improvement of The Model
  - First shrink the large size of data to include just 50,000 columns
  - Try larger model to be able to train across the large feature space
  - try large number of epochs to know the best range of epochs



  - The previous image shows that the best number of epochs is approximately 50



- **Results**
  - **confusion matrix**

The rows represents the true values or observations
The columns respresent the model's predictions

	angry	disgust	fear	happy	neutral	sad
angry	265	26	26	54	11	3
disgust	42	109	40	46	66	75
fear	26	40	174	62	32	76
happy	62	26	63	193	25	10
neutral	6	33	20	27	197	37
sad	2	47	45	13	54	200

  - **accuracy**

```
Training Accuracy:  0.895917534828186
```

```
/usr/local/lib/python3.7/dist-packages/sklea
  y = column_or_1d(y, warn=True)
Testing Accuracy:  0.5096282958984375
```

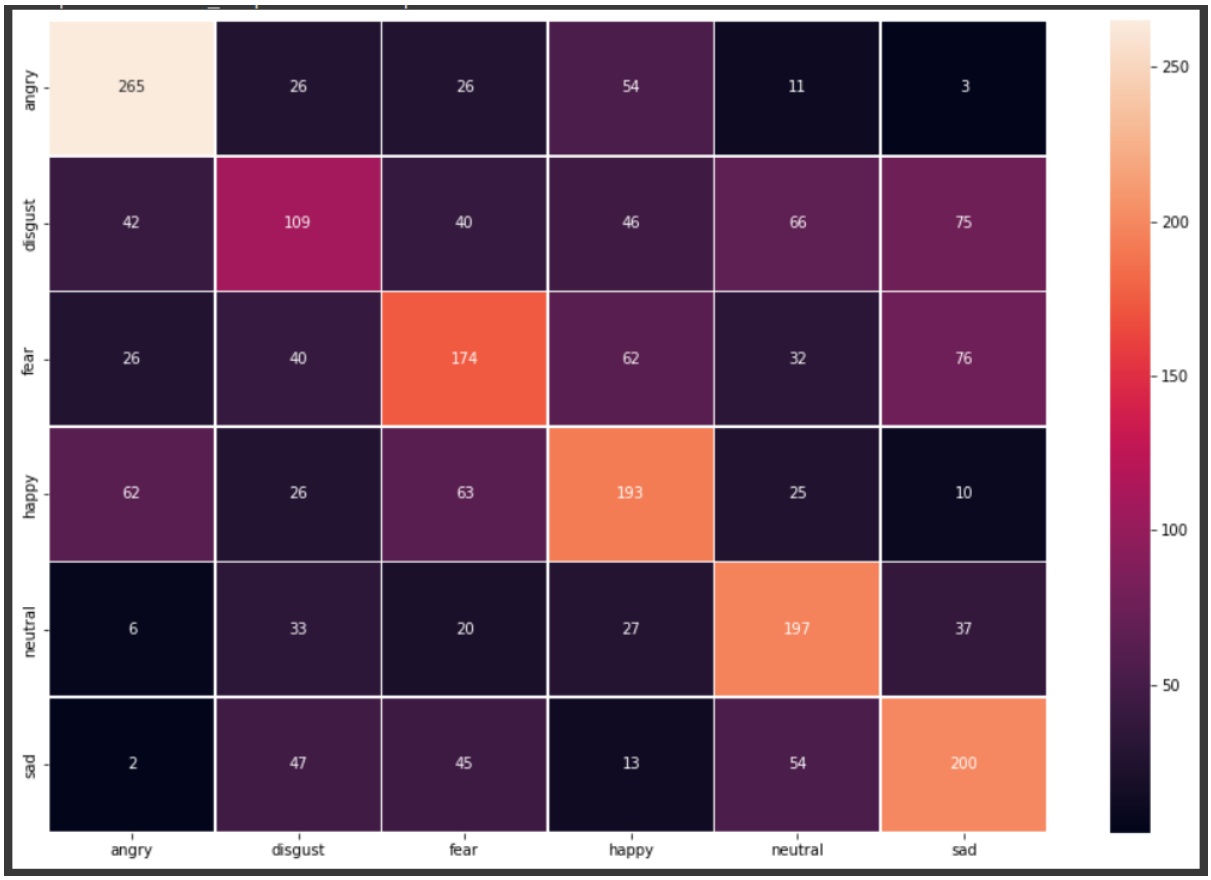
  - **f-score, precision and recall**

```
Percision:  0.5028906174789683
Recall:  0.5133230282614288
F-score:  0.5055335823870692
```

- **report of all classes**

	precision	recall	f1-score	support
angry	0.66	0.69	0.67	385
disgust	0.39	0.29	0.33	378
fear	0.47	0.42	0.45	410
happy	0.49	0.51	0.50	379
neutral	0.51	0.62	0.56	320
sad	0.50	0.55	0.52	361
accuracy			0.51	2233
macro avg	0.50	0.51	0.51	2233
weighted avg	0.50	0.51	0.50	2233

- **most confusing classes**



According to the classification report disgust has the lowest f1-
.score so the model is confused about those classes
.The previous heatmap also shows how the model predicts disgust

Comparing models

The first model was better as it considered more data and even when it was confused about some of the classes , it managed to .predict most of the samples