



# Model Report

Model Name	Four Dataset Model	One Dimensional Model	Spectrogram Model
Links to Notebooks	<a href="https://www.kaggle.com/codemaster101/emo-audio-and-text-based-emotion-recognition-with-cnn">https://www.kaggle.com/codemaster101/emo-audio-and-text-based-emotion-recognition-with-cnn</a>	<a href="https://github.com/mariamgabbar/emotion-recognition-cnn-1d-model">https://github.com/mariamgabbar/emotion-recognition-cnn-1d-model</a>	<a href="https://github.com/arisa-ai/Audio-and-text-based-emotion-recognition/blob/master/G">https://github.com/arisa-ai/Audio-and-text-based-emotion-recognition/blob/master/G</a>
Training Datasets***	Ravdess, Savee, Tess, Crema-D	Crema-D	Iemocap
Preprocess	Noise adding, stretching, shifting and pitching for augmentation	Noise reducing, length adjusting	Add delta spectrogram to get a 3d spectrogram
Input Features	1. Zero crossing rate (zcr) 2. Root-mean-square (rms) 3. Mel-frequency cepstral coefficient (mfcc)	1. Zero Crossing rate 2. <i>Energy</i> 3. Mel-frequency cepstral coefficient 4. <i>Spectral Centroid</i> 5. <i>Spectral Bandwidth</i> 6. <i>Spectral Flatness</i> 7. <i>Spectral Rolloff maximum frequencies</i> 8. <i>Spectral Rolloff minimum frequencies</i>	3D Spectrogram image
Input Sample Number	48648	7442	6738
Input Labels	1. angry 2. disgust 3. fear 4. happy	1. angry 2. disgust 3. fear 4. happy	1. neutral 2. sad 3. surprised 4. angry

- |             |            |              |
|-------------|------------|--------------|
| 5. neutral  | 5. neutral | 5. happy     |
| 6. sad      | 6. sad     | 6. excited   |
| 7. surprise |            | 7. fear      |
|             |            | 8. disgusted |

## Model architecture

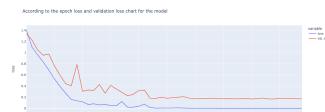
Layer (type)	Output Shape	Param #
conv1d_5 (Conv1D)	(None, 2376, 512)	3072
batch_normalization_6 (Batch Normalization)	(None, 2376, 512)	2048
max_pooling1d_5 (MaxPooling1D)	(None, 1188, 512)	0
conv1d_6 (Conv1D)	(None, 1188, 512)	1311232
batch_normalization_7 (Batch Normalization)	(None, 1188, 512)	2048
max_pooling1d_6 (MaxPooling1D)	(None, 594, 512)	0
conv1d_7 (Conv1D)	(None, 594, 256)	655616
batch_normalization_8 (Batch Normalization)	(None, 594, 256)	1024
max_pooling1d_7 (MaxPooling1D)	(None, 297, 256)	0
conv1d_8 (Conv1D)	(None, 297, 256)	196864
batch_normalization_9 (Batch Normalization)	(None, 297, 256)	1024
max_pooling1d_8 (MaxPooling1D)	(None, 149, 256)	0
conv1d_9 (Conv1D)	(None, 149, 128)	98432
batch_normalization_10 (Batch Normalization)	(None, 149, 128)	512
max_pooling1d_9 (MaxPooling1D)	(None, 75, 128)	0
Flatten_1 (Flatten)	(None, 9600)	0
dense_2 (Dense)	(None, 512)	4915712
batch_normalization_11 (Batch Normalization)	(None, 512)	2048
dense_3 (Dense)	(None, 7)	3591
Total params:	7,193,223	

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 849, 64)	256
conv1d_1 (Conv1D)	(None, 849, 64)	12352
max_pooling1d (MaxPooling1D)	(None, 424, 64)	0
conv1d_2 (Conv1D)	(None, 424, 128)	24704
conv1d_3 (Conv1D)	(None, 424, 128)	49280
max_pooling1d_1 (MaxPooling1D)	(None, 212, 128)	0
conv1d_4 (Conv1D)	(None, 212, 256)	98568
conv1d_5 (Conv1D)	(None, 212, 256)	196864
conv1d_6 (Conv1D)	(None, 212, 256)	196864
max_pooling1d_2 (MaxPooling1D)	(None, 106, 256)	0
conv1d_7 (Conv1D)	(None, 106, 128)	393728
conv1d_8 (Conv1D)	(None, 106, 128)	786944
conv1d_9 (Conv1D)	(None, 106, 128)	786944
max_pooling1d_3 (MaxPooling1D)	(None, 53, 128)	0
conv1d_10 (Conv1D)	(None, 53, 512)	786944
conv1d_11 (Conv1D)	(None, 53, 512)	786944
conv1d_12 (Conv1D)	(None, 53, 512)	786944
max_pooling1d_4 (MaxPooling1D)	(None, 26, 512)	0
flatten (Flatten)	(None, 13312)	0
dense_1 (Dense)	(None, 4096)	54538048
dense_1_1 (Dense)	(None, 4096)	16781312
dense_2 (Dense)	(None, 6)	24582

Layer (type:depth-idx)	Output Shape	Param #
ModifiedAlexNet	[1, 9]	..
Sequential: 1-1	[1, 9, 5, 12]	..
Conv1D: 1-1	[1, 64, 49, 109]	23,296
ReLU: 2-2	[1, 64, 49, 109]	..
MaxPool1d: 2-3	[1, 64, 24, 54]	..
Conv1D: 3-4	[1, 192, 24, 54]	307,392
ReLU: 4-5	[1, 192, 24, 54]	..
MaxPool1d: 5-6	[1, 192, 12, 27]	..
Conv1D: 6-7	[1, 384, 11, 26]	663,936
ReLU: 7-8	[1, 384, 11, 26]	..
Conv1D: 8-9	[1, 256, 11, 26]	884,992
ReLU: 9-10	[1, 256, 11, 26]	..
MaxPool1d: 10-11	[1, 256, 11, 26]	..
Conv1D: 11-12	[1, 256, 11, 26]	590,080
ReLU: 12-13	[1, 256, 11, 26]	..
Sequential: 1-2	[1, 9]	..
Linear: 1-14	[1, 9, 560]	..
Linear: 2-15	[1, 9]	2,313
Softmax: 1-3	[1, 9]	..
Total params:	2,472,089	
Trainable params:	2,472,089	

Training epoch number

Training / Validation Results

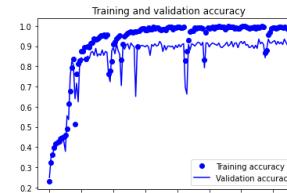
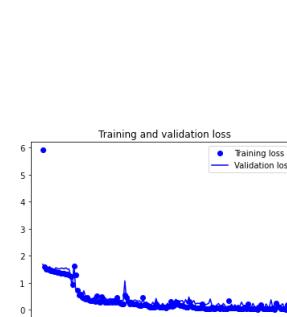


```
loss: 8.5794e-04 -
accuracy: 0.9997 -
val_loss: 0.1734 -
val_accuracy: 0.9625
```

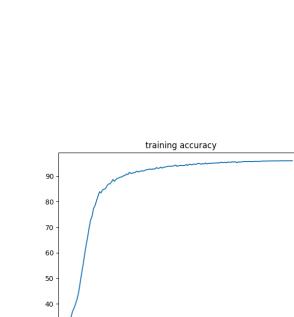
50

150

150

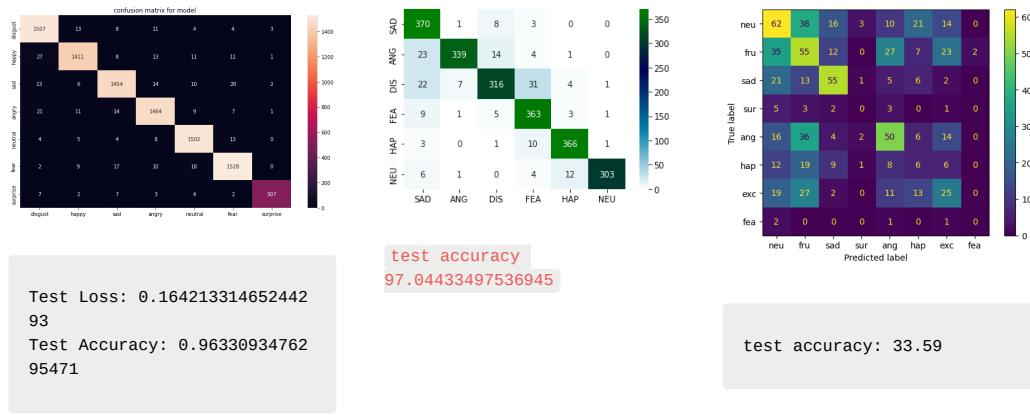


```
train accuracy: 99.757480
62133789
val accuracy: 98.08428883
552551
```

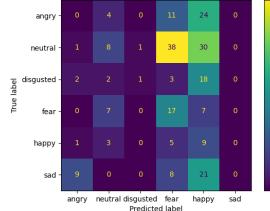


```
train accuracy: 96.074
train loss: 0.0054
```

Test Results  
(On training datasets)



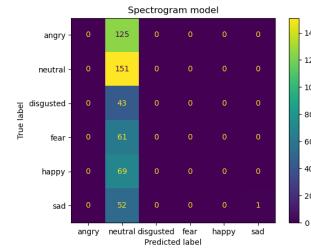
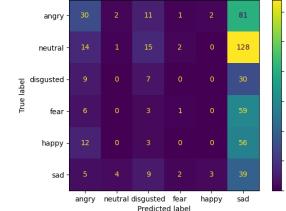
## Test Results on Emodb dataset \*\*\*



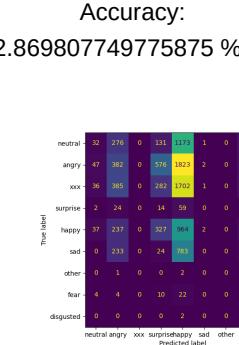
Accuracy: 6.54 %

Accuracy: 14.57 %

Accuracy: 28.41 %



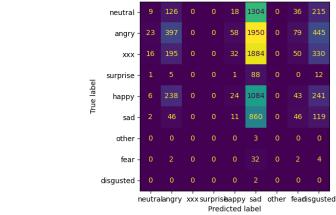
## Test Results on Iemocap dataset \*\*\*



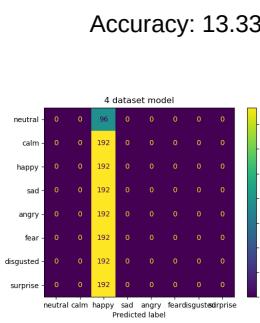
Accuracy: 12.869807749775875 %

Accuracy: 13.86592290068732 %

This model is already trained with IEMOCAP dataset, so no need for additional test



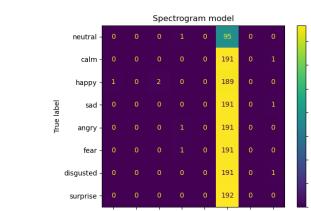
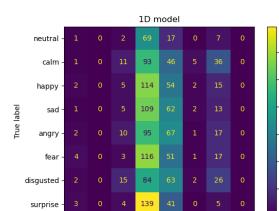
## Test results on Ravdess dataset



Accuracy: 13.33 %

Accuracy: 14.51 %

Accuracy: 13.40 %



## \*\*\*Notes on Datasets

### Ravdess

- 60 trials per actor x 24 actors = 1440 samples
- Calm, happy, sad, angry, fearful, surprise, and disgust
- Normal, strong or neutral expression

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/03006498-f51e-4cc3-bbd6-1906685fe591/03-01-08-02-02-02-14.wav>

## SAVEE

- 120 trials per actor x 4 male actors = 480 samples
- Anger, disgust, fear, happiness, sadness , surprise and neutral

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/41c9f4e5-7715-4719-a479-28d8866b7077/KL\\_su15.wav](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/41c9f4e5-7715-4719-a479-28d8866b7077/KL_su15.wav)

## TESS

- 200 words x 7 emotions x 2 female actresses = 2800 samples
- Anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9d329c77-1253-4193-bc6f-43dda165665d/YAF\\_white\\_ps.wav](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9d329c77-1253-4193-bc6f-43dda165665d/YAF_white_ps.wav)

## Crema-D

- 91 male and female actors, 12 sentences → 7442 samples
- Anger, Disgust, Fear, Happy, Neutral, and Sad
- Low, Medium, High, and Unspecified intensity

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/040b9615-9e04-41d9-980e-adfab6b96da2/1015\\_TIE\\_SAD\\_XX.wav](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/040b9615-9e04-41d9-980e-adfab6b96da2/1015_TIE_SAD_XX.wav)

## EmoDB

- In German
- 10 actors (5 male, 5 female) → 535 samples
- Anger, boredom, anxiety, happiness, sadness, disgust and neutral.

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/2b02ec37-66bb-492b-926a-dcb327315884/16b10Wa.wav>

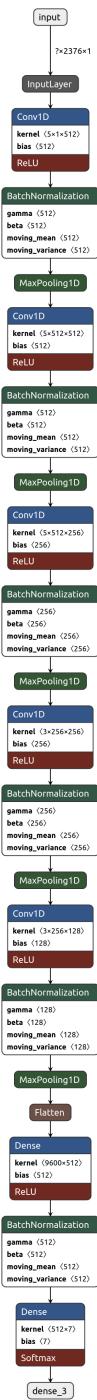
## IEMOCAP

- 10 actors (5 male, 5 female) in scripted and improvised situations → 10039 samples
  - anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state

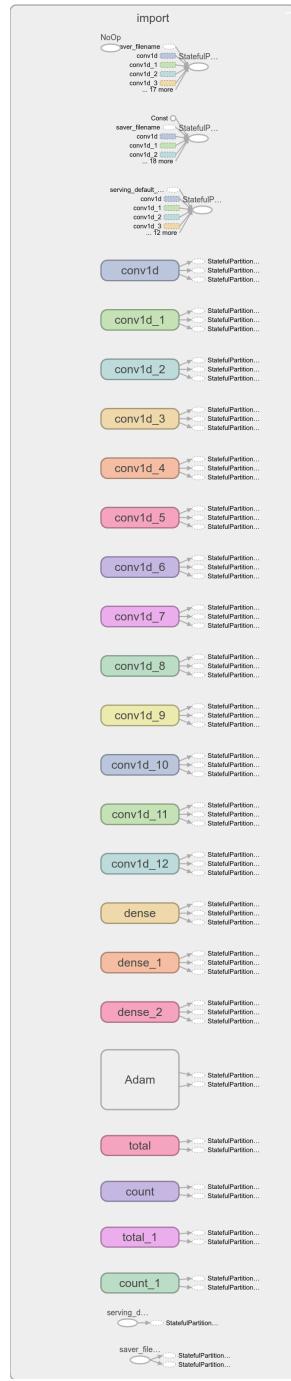
[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/76781a35-82df-420f-80f6-b2f17e9b5344/Ses01M\\_script01\\_1\\_M042.wav](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/76781a35-82df-420f-80f6-b2f17e9b5344/Ses01M_script01_1_M042.wav)

## Model block diagrams

## Four Dataset Model



## One Dimensional Model



## Spectrogram Model

