Pattern Recognition

SER: Speech Emotion Recognition

REPORT - Assignment 3

Team Members:

1.	Maram Mohamed Ghazal	5467
2.	Alaa Abdelmajied Kamel	5591
3.	Lina Hazem	5613
4.	Mariam Fekry	5614
5.	Mariam Matar	5653

Speech Emotion Recognition

Introduction

SER It is a system through which various audio speech files are classified into different emotions such as happy, sad, anger and neutral by computer. SER can be used in areas such as the medical field or customer call centers. In this assignment, we use Convolutional Neural Network to recognize emotion from the audios in Crema dataset.

Discussion

The first step in this assignment is loading the audio dataset (Crema Dataset) using librosa library. The dataset consists of 6 different emotions which are **SAD**, **ANGER**, **DISGUST**, **FEAR**, **HAPPY**, and **NEUTRAL** each emotion is given a unique label starting from 0 to 5. After that, the dataset is splitted into 70% training and validation and 30% testing and 5% of the training and validation data for validation in case of 2D CNN model, while for the 1D CNN model, the dataset is splitted into 80% training and validation and 20% testing and 10% of the training and validation data for validation.

The assignment is divided into two parts, one for the 1D CNN model and the other for the 2D CNN model.

- **1D CNN MODEL**

The process of generation:

1. Data Augmentation

   The training data is augmented using different methods including:

   - Gaussian white noise
   - Shifting.

2. Features Extraction

   Different types of features extraction are used including:

   - Zero crossing rate
   - Energy
   - Root mean square
   - Chroma short time frequency transform
   - Mel Frequency Cepstral Coefficients

These types are used to extract different features from the training, validation, and testing datasets. After that, the data dimensions are altered to enter the model.

3. Model Architecture

   - 1st Trial

     validation accuracy = 46%

   - 2nd Trial

     validation accuracy = 49%

   - 3rd Trial

     validation accuracy = 50%

   - Final Trial

     - 1st trial

     validation accuracy = 51%

     test accuracy = 50.3%

     - 2nd trial

     validation accuracy = 50.5%

     test accuracy = 51.2%

4. Outputs of the Final Trial

- <u>1st Trial</u>

```
              precision    recall  f1-score   support

           0       0.54      0.57      0.55       270
           1       0.62      0.72      0.67       257
           2       0.54      0.39      0.45       263
           3       0.45      0.29      0.35       257
           4       0.39      0.44      0.41       225
           5       0.46      0.62      0.53       217

    accuracy                           0.50      1489
   macro avg       0.50      0.51      0.49      1489
weighted avg       0.50      0.50      0.50      1489

val accuracy: 51.00671052932739
train accuracy: 60.83629131317139
test accuracy 50.3693754197448
```

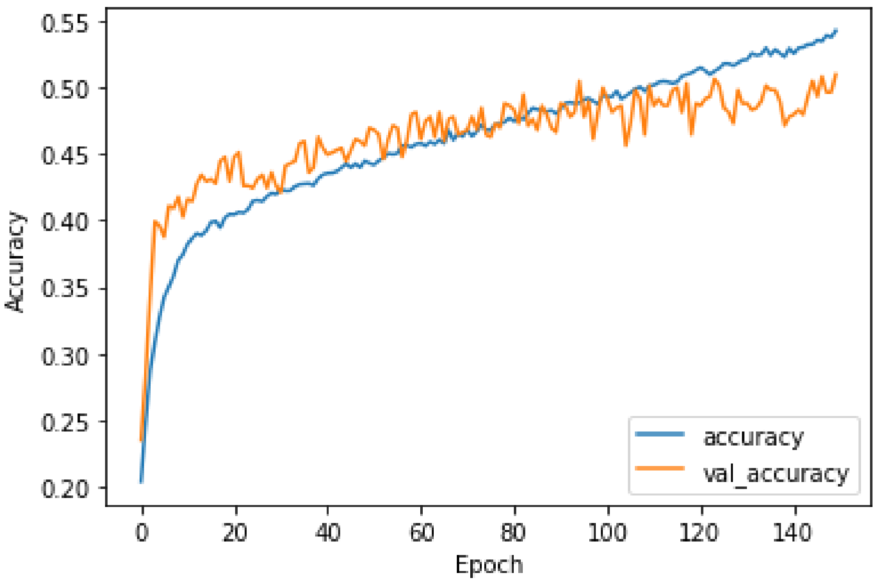


Figure 1. Training Accuracy vs Validation Accuracy

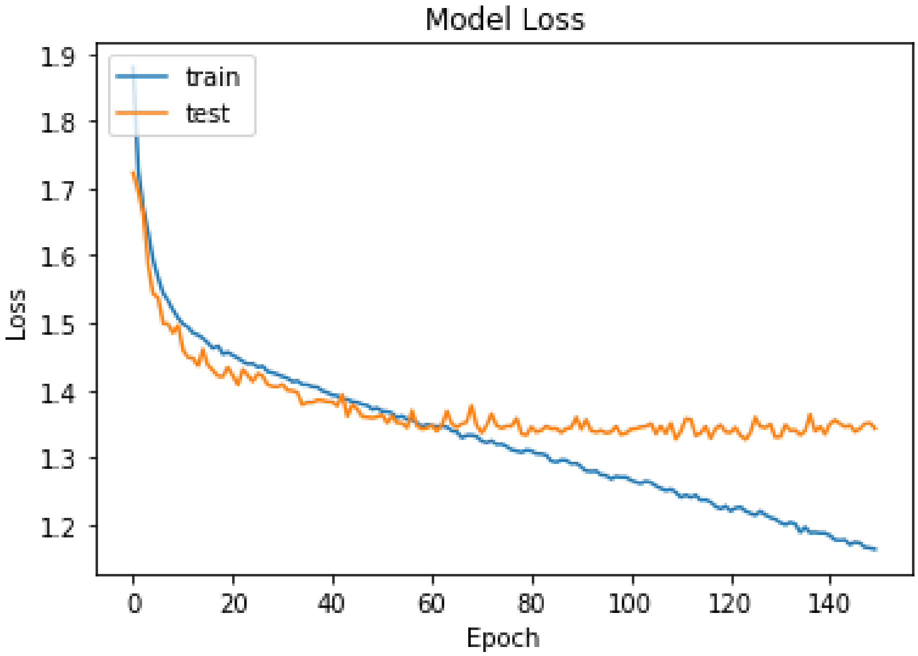


Figure 2. Training Loss vs Validation Loss

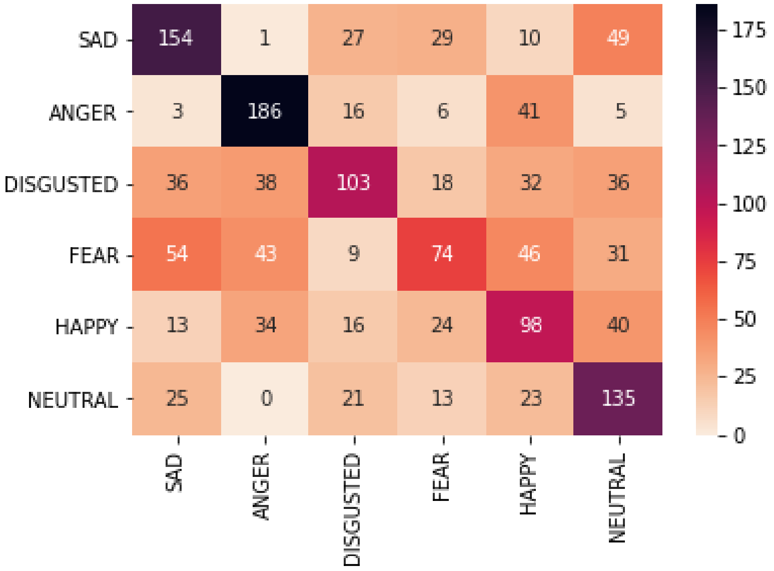


Figure 3. The Confusion Matrix

- The best classified emotion is ANGER, while the worst classified emotion is FEAR.

- <u>2nd Trial</u>

```
              precision    recall  f1-score   support

           0       0.54      0.68      0.60       262
           1       0.65      0.70      0.68       235
           2       0.52      0.42      0.46       269
           3       0.36      0.27      0.31       239
           4       0.46      0.45      0.46       260
           5       0.48      0.56      0.52       224

    accuracy                           0.51      1489
   macro avg       0.50      0.51      0.50      1489
weighted avg       0.50      0.51      0.50      1489

val accuracy: 50.503355264663696
train accuracy: 61.14118695259094
test accuracy 51.24244459368704
```

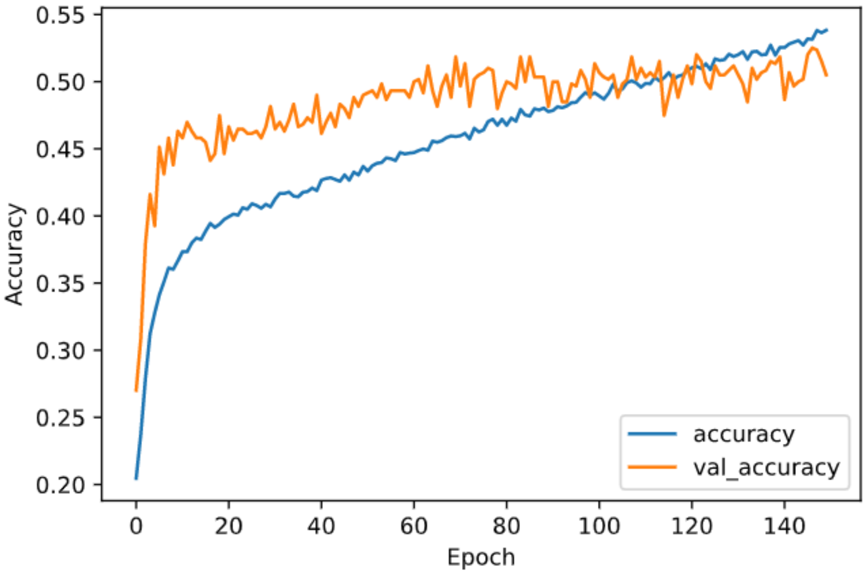


Figure 4. Training Accuracy vs Validation Accuracy

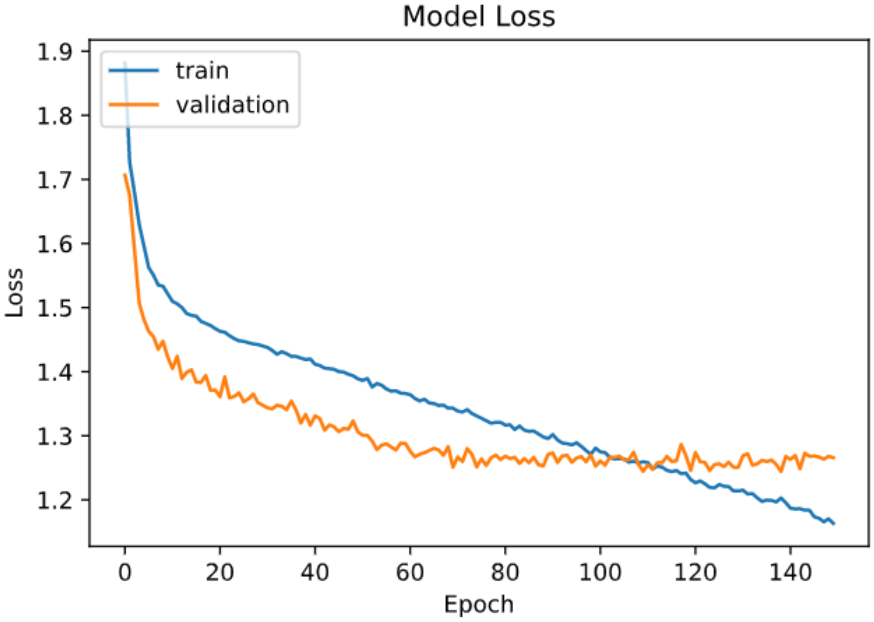
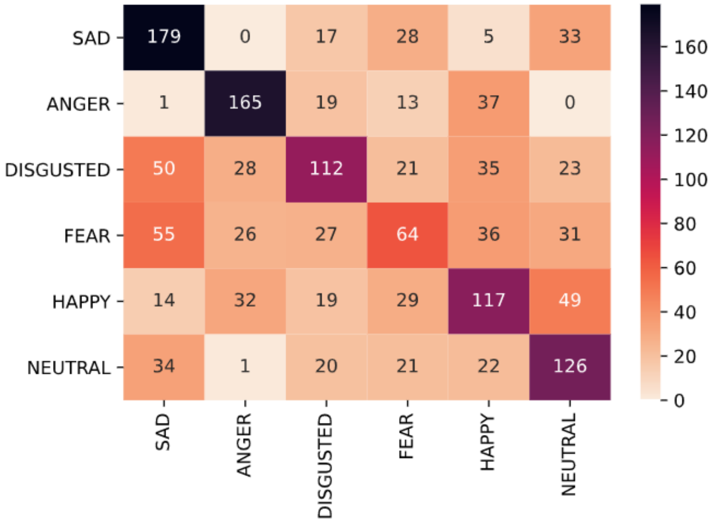


Figure 5. Training Loss vs Validation Loss



	SAD	ANGER	DISGUSTED	FEAR	HAPPY	NEUTRAL
SAD	179	0	17	28	5	33
ANGER	1	165	19	13	37	0
DISGUSTED	50	28	112	21	35	23
FEAR	55	26	27	64	36	31
HAPPY	14	32	19	29	117	49
NEUTRAL	34	1	20	21	22	126

Figure 6. The Confusion Matrix

- The best classified emotion is SAD, while the worst classified emotion is FEAR.

- **2D MODEL**

The process of generation:

1. Data Augmentation

   The training data is augmented using:

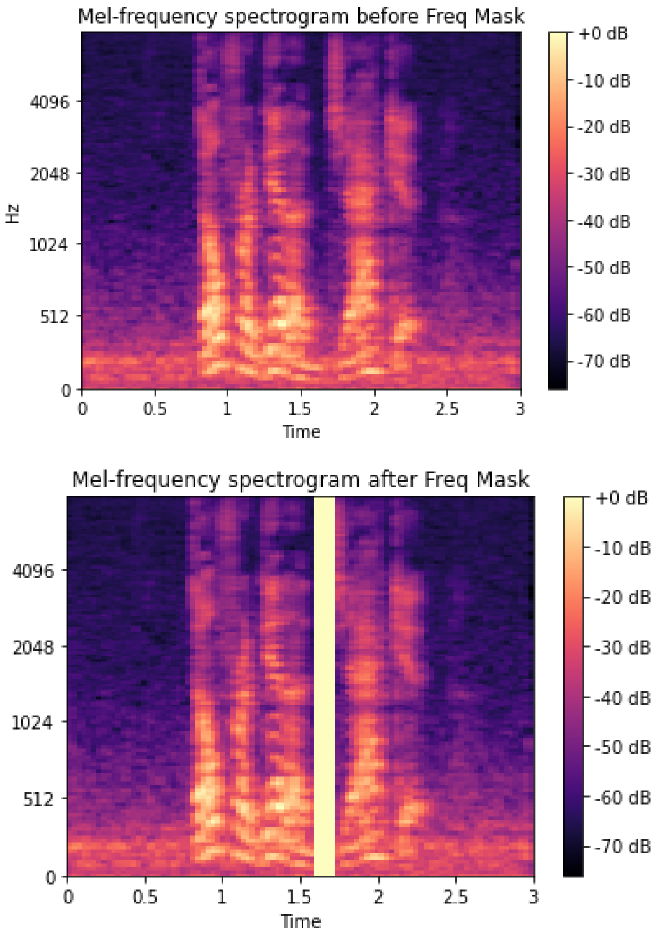
   - Frequency mask

2. Features Extraction

   Type of feature extraction used:

   - Mel Spectogram

Mel Spectogram sample from training set before and after frequency mask


Mel-frequency spectrogram before Freq Mask

Mel-frequency spectrogram after Freq Mask

3. Model Architecture

- **1st Trial:**

- Using relu as activation function + l1 as kernel_regularizer + RMSprop optimizer

validation accuracy: 45.59%

test accuracy 44.8%

- **2nd Trial (Final Model):**

- Using relu as activation function + l1_l2 as kernel_regularizer + RMSprop optimizer
- Using class weights to handle imbalance
- Trial to get out of stucking in local minimum
- Using ReduceLROnPlateau

validation accuracy: 56.3%

test accuracy:57.09%

4. Outputs of the Final Model

```
              precision    recall  f1-score   support

           0       0.49      0.72      0.59       394
           1       0.74      0.67      0.70       385
           2       0.61      0.40      0.48       391
           3       0.47      0.50      0.48       374
           4       0.53      0.57      0.55       354
           5       0.69      0.56      0.62       335
accuracy                             0.57      2233
macro avg          0.59      0.57      0.57      2233
weighted avg       0.59      0.57      0.57      2233

[[284   6  25  49   6  24]
 [  8 258  18  21  68  12]
 [ 95  34 156  48  32  26]
 [ 98  12  10 186  58  10]
 [ 22  36  19  61 203  13]
 [ 67   4  26  32  18 188]]

       val accuracy: 56.32184147834778
       train accuracy: 62.803155183792114
       test accuracy 57.09807433945365
```

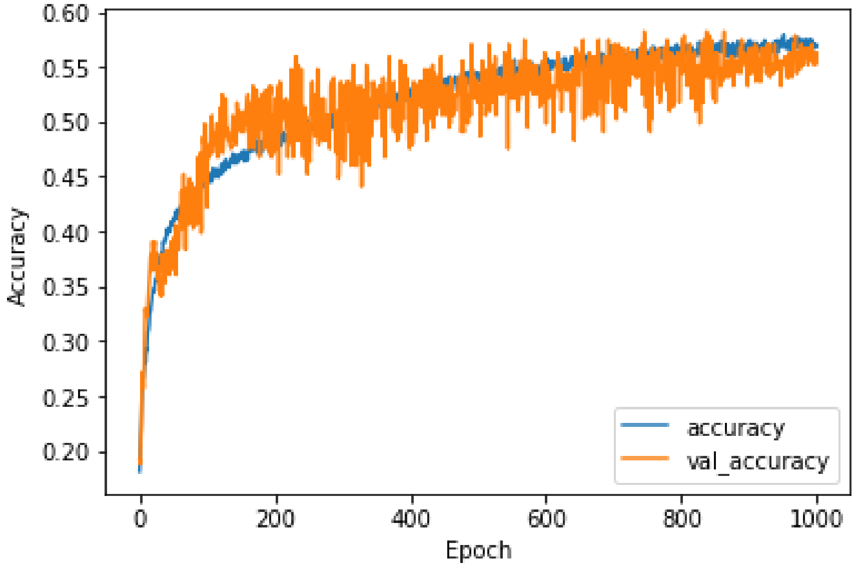


Figure 7. Training Accuracy vs Validation Accuracy

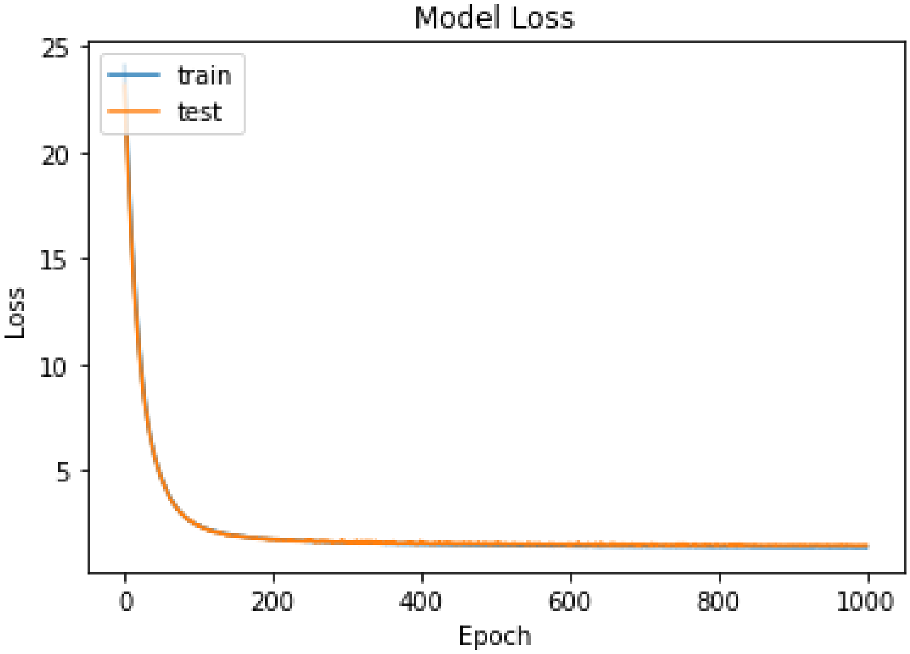
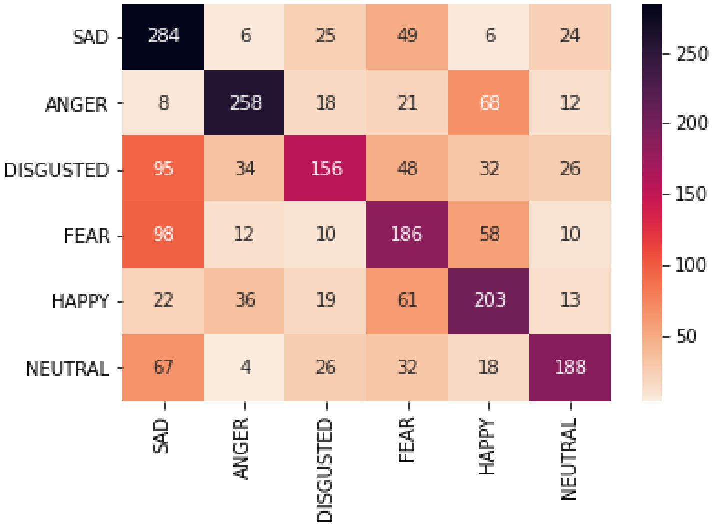


Figure 8. Training Loss vs Validation Loss



	SAD	ANGER	DISGUSTED	FEAR	HAPPY	NEUTRAL
SAD	284	6	25	49	6	24
ANGER	8	258	18	21	68	12
DISGUSTED	95	34	156	48	32	26
FEAR	98	12	10	186	58	10
HAPPY	22	36	19	61	203	13
NEUTRAL	67	4	26	32	18	188

Figure 9. The Confusion Matrix

- The best classified emotion is SAD, while the worst classified emotion is DISGUSTED.