Evaluation part:

Confusion Matrices of the Best Model after Data Mitigation:

The following images depict the confusion matrices of the best model for the test folds ranging from one to ten. Based on these matrices, it can be observed that the highest recognition accuracy is achieved in test K-fold seven, while the lowest precision in recognizing class I occurs in K-fold nine. The recognition performance in the remaining folds is approximately consistent. Therefore, the non-uniform distribution of data is noticeable specifically in these two folds.

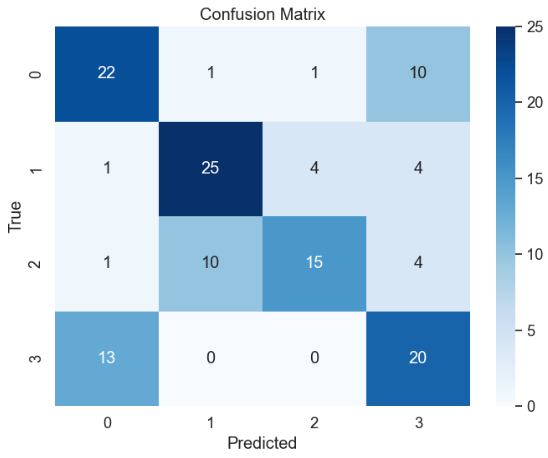


Figure 1 K-fold one

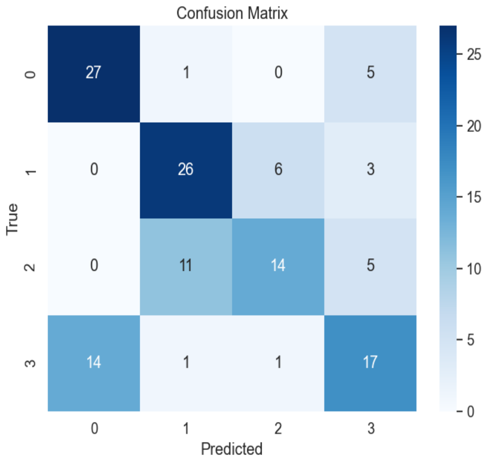


Figure 2 K-fold two

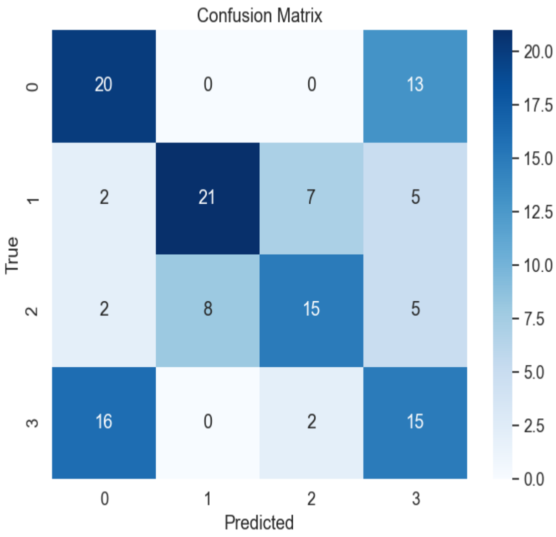


Figure 3 K-fold three

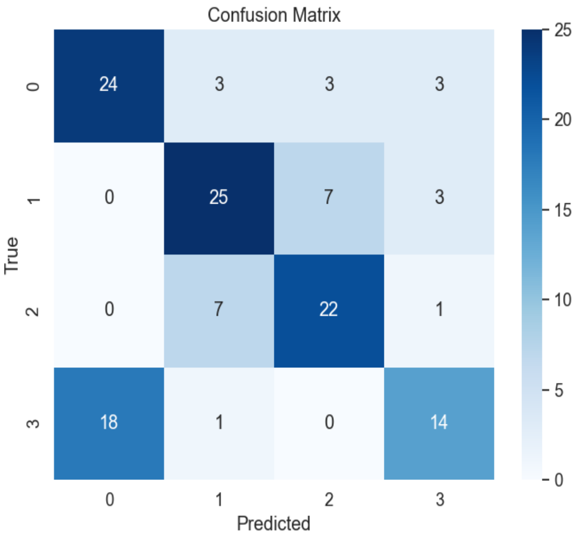
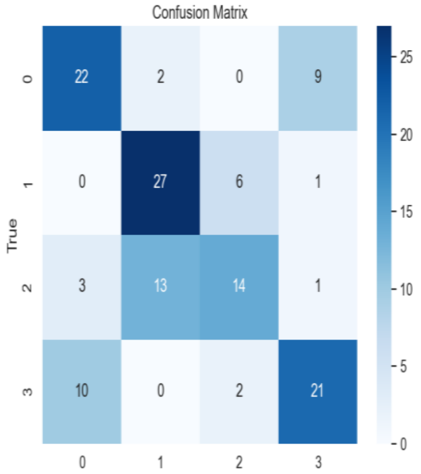
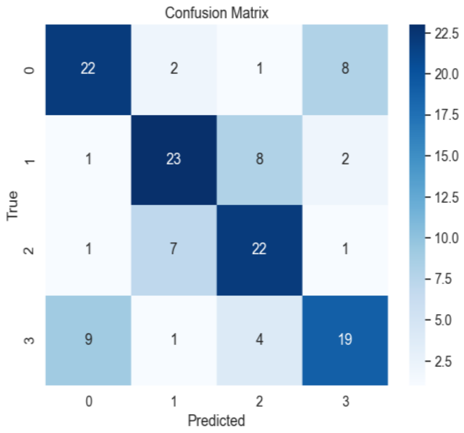


Figure 4 K-fold four



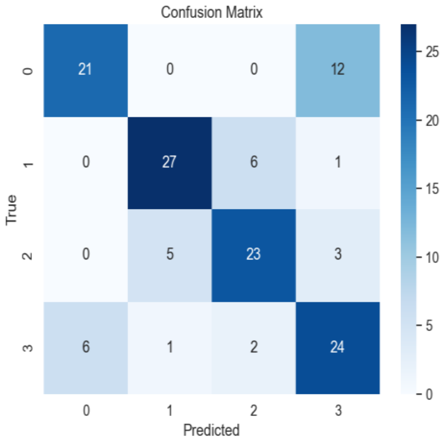
	0	1	2	3
0	22	2	0	9
1	0	27	6	1
2	3	13	14	1
3	10	0	2	21

Figure 5 K-fold five



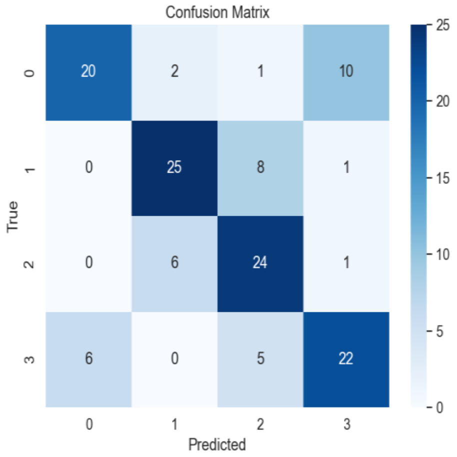
	0	1	2	3
0	22	2	1	8
1	1	23	8	2
2	1	7	22	1
3	9	1	4	19

Figure 6 K-fold six



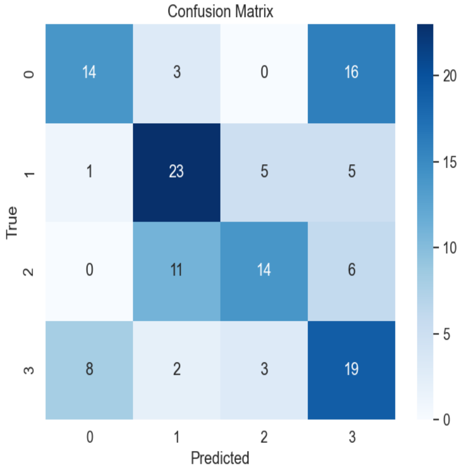
	0	1	2	3
0	21	0	0	12
1	0	27	6	1
2	0	5	23	3
3	6	1	2	24

Figure 7 K-fold seven



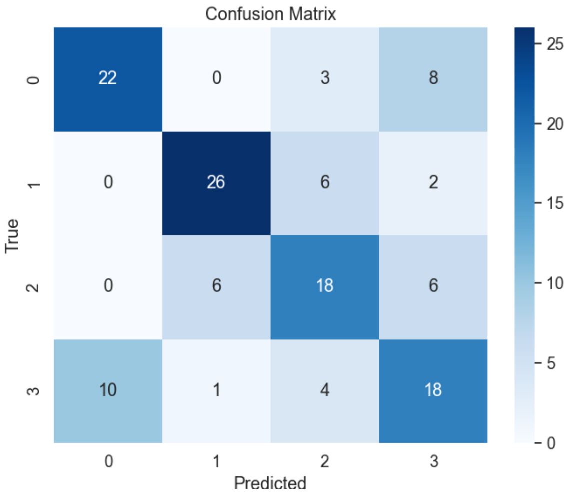
	0	1	2	3
0	20	2	1	10
1	0	25	8	1
2	0	6	24	1
3	6	0	5	22

Figure 8 K-fold eight



	0	1	2	3
0	14	3	0	16
1	1	23	5	5
2	0	11	14	6
3	8	2	3	19

Figure 9 K-fold nine



	0	1	2	3
0	22	0	3	8
1	0	26	6	2
2	0	6	18	6
3	10	1	4	18

Figure 10 K-fold ten

Performance Results of Each Fold:

The table below illustrates the overall performance consistency of the best model from part two.

Table 1 Performance details of best model two

		Micro			Macro		
Fold	**P**	**R**	**F**	**P**	**R**	**F**	**Accuracy**
1	0.6529	0.6529	0.6529	0.6714	0.6723	0.6718	0.6529
2	0.7120	0.7120	0.7120	0.7294	0.7193	0.7094	0.7120
3	0.6555	0.6555	0.6555	0.6727	0.6616	0.6626	0.6555
4	0.6786	0.6786	0.6786	0.7004	0.6874	0.6845	0.6786
5	0.6658	0.6658	0.6658	0.6858	0.6838	0.6676	0.6658
6	0.6863	0.6863	0.6863	0.7369	0.7147	0.6861	0.6863
7	0.7474	0.7474	0.7474	0.7672	0.7524	0.7572	0.7474
8	0.6262	0.6262	0.6262	0.6475	0.6332	0.6227	0.6262
9	0.6932	0.6932	0.6932	0.7208	0.6861	0.6929	0.6932
10	0.6907	0.6907	0.6907	0.6937	0.6917	0.6895	0.6907
Average	0.6809	0.6809	0.6809	0.7026	0.6902	0.6844	0.6809

The next table shows the performance details of the best model from part three after data mitigation.

Table 2 Performance details of best model three

		Micro			Macro		
Fold	**P**	**R**	**F**	**P**	**R**	**F**	**Accuracy**
1	0.6259	0.6259	0.6259	0.6413	0.6221	0.6243	0.6259
2	0.6412	0.6412	0.6412	0.6396	0.6357	0.6302	0.6412
3	0.5419	0.5419	0.5419	0.5609	0.5401	0.5455	0.5419
4	0.6488	0.6488	0.6488	0.6550	0.6497	0.6431	0.6488
5	0.6412	0.6412	0.6412	0.6410	0.6371	0.6330	0.6412
6	0.6564	0.6564	0.6564	0.6563	0.6571	0.6557	0.6564
7	0.7251	0.7251	0.7251	0.7344	0.7249	0.7263	0.7251
8	0.6946	0.6946	0.6946	0.7013	0.6955	0.6941	0.6946
9	0.5384	0.5384	0.5384	0.5619	0.5365	0.5364	0.5384
10	0.6461	0.6461	0.6461	0.6463	0.6442	0.6451	0.6461
Average	0.6360	0.6360	0.6360	0.6438	0.6343	0.6334	0.6360

In terms of the consistency across different folds, it is evident that the majority of folds exhibit a similar distribution of data, resulting in approximately the same performance. However, two folds stand out, yielding the best and worst results, identified as fold 7 and fold 9, respectively.

Upon comparing the performance results of the optimal model from the previous section with the outcomes updated after applying the K-fold algorithm, a noticeable decrease of approximately 6 percentage points in the average F-measure is observed across test folds 1 to 10. In the earlier phase of learning, 70% of the data was allocated to the training set, with 15% for validation and another 15% for testing. The data was shuffled before sampling, reducing the likelihood of non-uniform selection from various emotional classes. However, the observed decline in performance may not be solely attributed to non-uniform selection but is more likely a consequence of the mismatch in the number of collected data points from different classes. Specifically, the "angry" class constitutes the majority of the data, while the "bored" class contributes the least, with the "focused" and "neutral" classes falling in between. Consequently, the K-fold procedure unveils and accentuates this imbalance.

Bias analysis:

Introduction:

The features selected for bias analysis include age categorized into three groups: young, middle-aged, and senior, as well as gender divided into female and male subsets. Consequently, 24 label categories have been derived based on gender, age, and emotional class. During the data collection phase of the project, the dataset was labeled based on emotion. Gender and age labels were manually assigned to enhance assessment precision. Subsequently, a CSV file was created with four columns, encompassing gender, age, emotional class, and the image file name. This CSV file was prepared to facilitate proper data engineering. To identify biases, the number of data points in each of the aforementioned 24 groups was computed, revealing imbalances in the data distribution.

Bias detection results:

After counting the amount of data in each of the aforementioned 24 groups, this initial analysis has revealed a significant imbalance, particularly in the age group of seniors.

The following table displays the performance results of the best model from part two, categorized by different genders and ages. The previously mentioned prepared CSV file was utilized in this step to differentiate between various types of data.

Table 3 Performance examination of different groups for the biased model

Group	Accuracy	P	R	F
Young	0.684	0.591	0.526	0.524
Middle-aged	0.727	0.609	0.565	0.555
Senior	0.666	0.5	0.312	0.380
Average	0.693	0.566	0.467	0.486
Male	0.687	0.595	0.552	0.545
Female	0.604	0.535	0.535	0.737
Average	0.645	0.565	0.543	0.641
Overall System AVG	0.673	0.566	0.498	0.548

From the above table, the most significant decrease is observed in the recall for the senior age group, which impacts the system's performance on this age group. Consequently, it has become evident that there is an imbalance in the senior age group. Subsequently, for further investigation, the distribution of data based on gender, age, and emotional class was extracted to add appropriate data where needed. Following the results of this process, given that the average distribution of data is around 50 images per group of age, gender, and emotional class, groups with fewer than 50 images were identified. These groups were then subjected to the step of removing imbalance to decrease the variance in the data.

Bias mitigation step:

To compensate for the lack of data in the identified groups, augmentation functionalities were applied to the images of these underrepresented groups. Within this function, there is a loop with a condition to surpass the threshold of 50 images. Within this loop, one image from the group is randomly selected, and among the implemented augmentation functions, one is randomly chosen. The selected image is then passed to this function as input. The augmentation functions include rotating the selected image by a random angle between -30 and 30 degrees, horizontally flipping the image, zooming in by cropping the image with a random zoom factor, and shifting the image along width and height.

Comparative Performance Analysis:

After addressing the data deficiency in specific groups through data augmentation, the test and train procedures were rerun, and the performance results for these particular data groups are as follows.

Table 4 Performance examination of different groups for the unbiased model

Group	Accuracy	P	R	F
Young	0.583	0.589	0.587	0.582
Middle-aged	0.625	0.629	0.613	0.602
Senior	0.630	0.606	0.608	0.585
Average	0.613	0.608	0.603	0.590
Male	0.585	0.598	0.590	0.583
Female	0.642	0.629	0.632	0.621
Average	0.614	0.613	0.611	0.602
Overall System AVG	0.614	0.617	0.610	0.605

Upon examining the F-score, which reflects both precision and recall, it can be acknowledged that the performance has improved, and the bottleneck observed within the senior age group has been alleviated. The improved results in the second trained model with mitigated data are expected, given the increase in data within groups previously lacking sufficient samples. The augmentation function has been applied randomly, contributing to the diversification of the data.