2020

# Facial Emotion Recognition using Machine Learning

ALEKSY RUSZAŁA

ABERTAY UNIVERSITY |

# Table of Contents

## Introduction

Over the years, Machine Learning had started to play more and more important role in the world of business and had revolutionized the information technology sector. Machine learning relies on the technological memory of certain patterns and behaviours that the machine has already "experienced" before. ML make it possible to analyse new data, as well as look for solutions that have worked in other, previously known to machine situations (Vishwanathan, 2008).

Facial expression recognition is currently a constantly evolving technology mechanism in the Machine Learning, used in many industries such as automotive, marketing, PR or in a game sector (Gupta, 2018). Face has emerged as an area of special communication significance, as it reflects internal emotional states and attitudes of the person, and expresses intended messages and comments related to the interactive situation.

Effective recognition of emotions has always been a challenge for researchers. Tasks such as recognizing emotions or objects are simple for humans, but they pose a challenge to computer systems. Effective detection of human faces is one of the most serious problems of image analysis. As a result, it is even more difficult to locate facial features effectively and efficiently in the analysed image and translate them into the expressed emotion.

Many researchers had tried to find the best way to create a “state-of-the-art” facial emotions recognition, study conducted by A. Saravanan et al. in 2019 (K.S.Gayathri, 2019) using various models on the FER2013 dataset resulted in 0.60 accuracy using Adam optimizer with modified hyperparameters. Those result were achieved with the use of only single dataset. Research done by A. Mollahosseini in 2016 (Mahoor, 2016) used a combination of 7 different datasets. The results showed that the architecture performed best on MMI and FER2013 datasets, the FER2013 itself resulted in the highest accuracy of 0.66. The highest accuracy of 80.9 resulted from the studies conducted by Tan et al. in 2017 (Qiao, 2017). It involved use of 2068 images from combination of different datasets.

Further research and development regarding effective facial expression recognition can have a significant effect for many sectors. For instance, marketers will be able to use it to make more effective analyses and profiling on unprecedented nowadays scale, it will allow safety organizations to analyse facial expressions of people entering places like banks to detect potential threats.

The aim of this paper is to introduce a new approach to create an effective facial emotions recognition application that will be able to distinguish between seven basic facial expressions: joy, surprise, neutral, anger, fear, disgust and sadness.

The project objective is to achieve the best performance for facial emotion recognition model and aims to evaluate the increase or decrease in performance of the model created by the algorithm based on implementation of various methods, techniques such as data preparation, image pre-processing, learning with pre-train algorithm and more.

## Methodology

Research showed that use of combination of different datasets seems to be more effective (K.S.Gayathri, 2019). Two different datasets were chosen in this project; Cohn-Kanade and MUG datasets (see example in figure 2&3), based on the best quality of the images compering to other available datasets for this project. Datasets together contained number 1009 of images of people with different facial expressions (see figure 1).

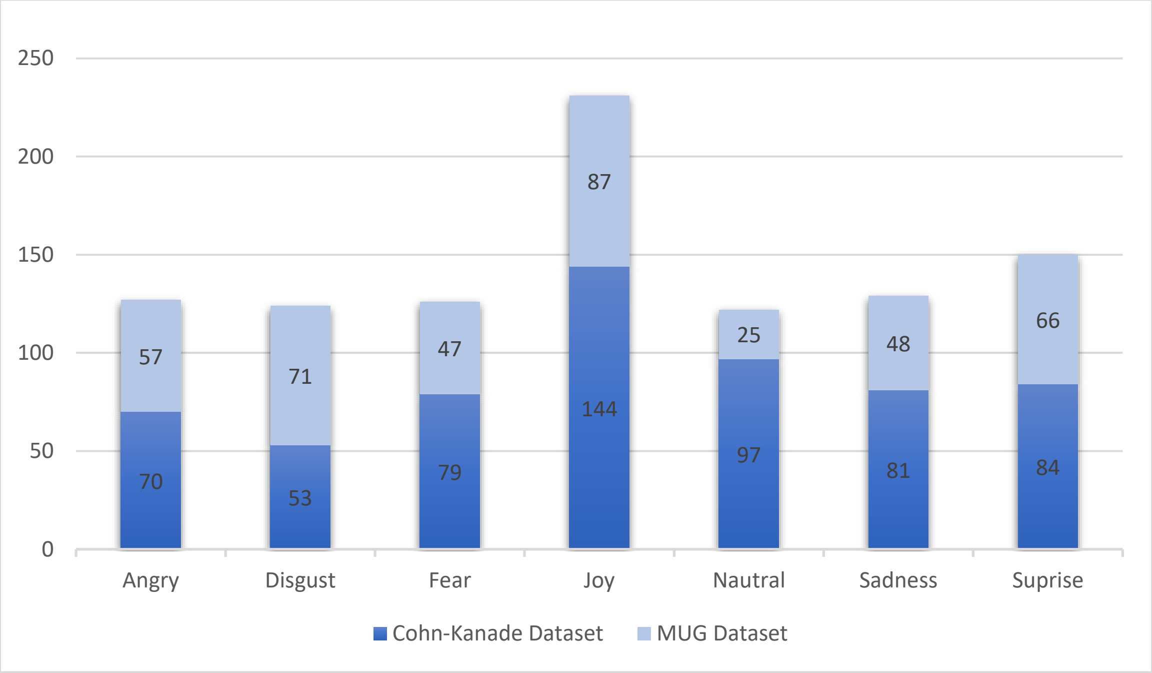


*Figure 1, the number of images in the datasets from in each category.*

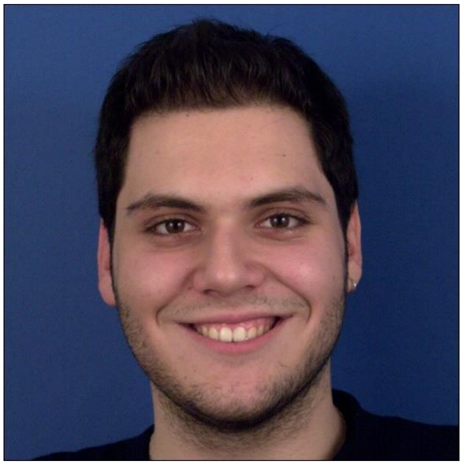


*Figure 2, example of image from MUG dataset.*

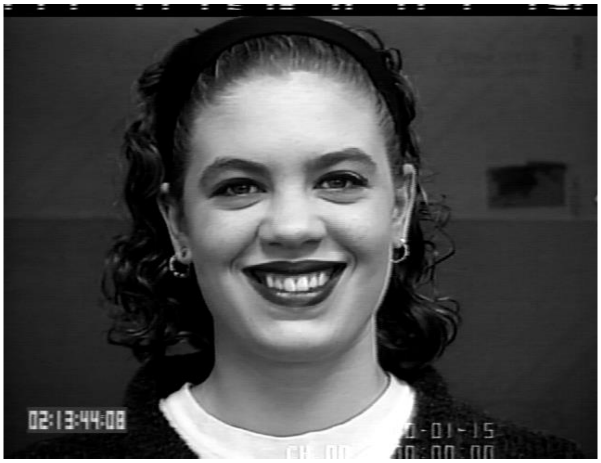


*Figure 3, example of image from Cohn-Kanade dataset.*

Framework “Dlib” was used in order identify people faces from images and then to produce the detector points. This created a set of points on each picture from the datasets corresponding to the key locations on the face, which allowed to calculate the distances between key landmarks (see figure 4):

- Left lip & Right lip
- Width & Height of the lips
- Left eyebrow & eye
- Right eyebrow & eye.

To allow the model to recognize and distinguish seven basic facial expressions: joy, surprise, neutral, anger, fear, disgust, and sadness use in this study. The processed and normalized data has been saved in a form of numeric dataset that will be understandable by machine learning program.

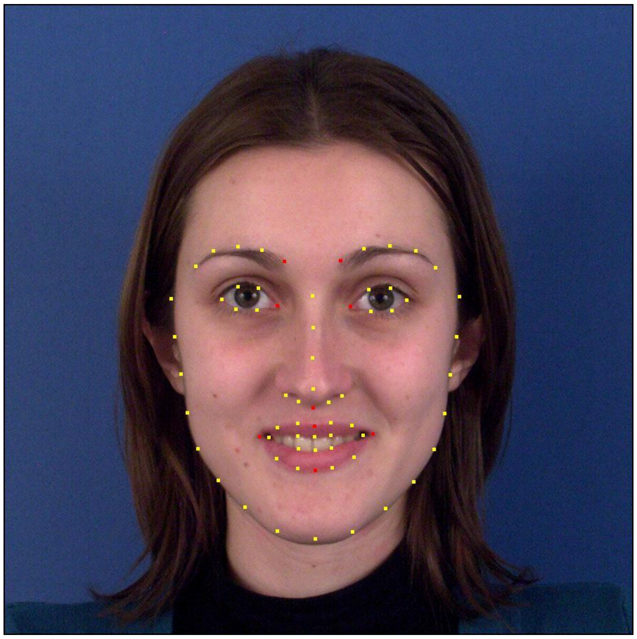


*Figure 4, example of the image from dataset containing detector points*

The model uses a supervised algorithm as it was best suited for this project, as knowing what the output values should be allowed to evaluate the behaviour of the algorithm. The application of this type of algorithm allowed to supervise it and evaluate the results to the expected precision of emotions, performance and verify problems or deviations. The whole set is discrete, and each input image was allocated to one of seven types of emotions, this data properties allowed the use of a classification algorithm.

## Training & Testing

### Dataset

Both original Cohn-Kanade and MUG datasets combined were divided into train and testing dataset in ratio 80/20. Based on a set of photos from several reports, two smaller collections were created to for the validation data, which together gave 14 images, two per category (Bayburt, 2019) (Clara R. Freeman, 2018).

## First Implementation and Feature extraction

Based on the combined sets of images a dataset was generated containing a list of features described by a specific emotion label. The data was created using “Dlib” library that produced a detector points on the images to generate features. Prepared data was imported into the project and the first model has been created using classification and ML.NET framework. The pipeline created used the maximum entropy classification algorithm which has been loaded with previously generated data containing basic facial features. The model had been trained, saved, and then evaluated. The metrics interpreted during each test had been used for comparison with subsequent changes during the project and served as a reference. Prediction of facial emotions is based on the generated model. The implementation also assumed testing on validation data to assess the model for under / overfitting.

## Cross-Validation

To create a more stable training a cross-validation approach has been used to train and test sets for the model. Use of cross-validation is more effective in assessing actual model accuracy. This method involves splitting the data set into subsets and creating test and training datasets from them. There are different variants of cross-validation, this project used K-fold cross-validation divided into 5 K-fold. By dividing the whole dataset into 5 parts the ratio will be 80/20, where in this case 4K is the dataset to train and 1K is used for test. This is the most popular and most useful validation method which allowed to create an estimate performance and precision of the model (Tensorflow, 2019).

## Improvements

Six main different steps have been implemented including different possible solution to improve the accuracy of the model. After implementation of each technique new test was conducted and the results had been analysed.

## Data preparation and image pre-processing:

The first step included data preparation and image pre-processing, doing so can have a significant effect on the accuracy of the model (Pintelas, 2006). Research showed that performance of the model can be negatively affected by poor data, which is not representative or noisy (Scikit-Learn & Keras, 2019). All the images that the algorithm was not able to produce correct detector on were edited using contrast enhancement or/and using sharpen tools (see figure 5) and checked again, all images that once again were wrongly detected were removed from the dataset. Data was transformed and catalogize further to create the same amount of images in each group of the emotions. In addition, all images have been reviewed and sorted into the best fitting categories based on the emotions on the face.

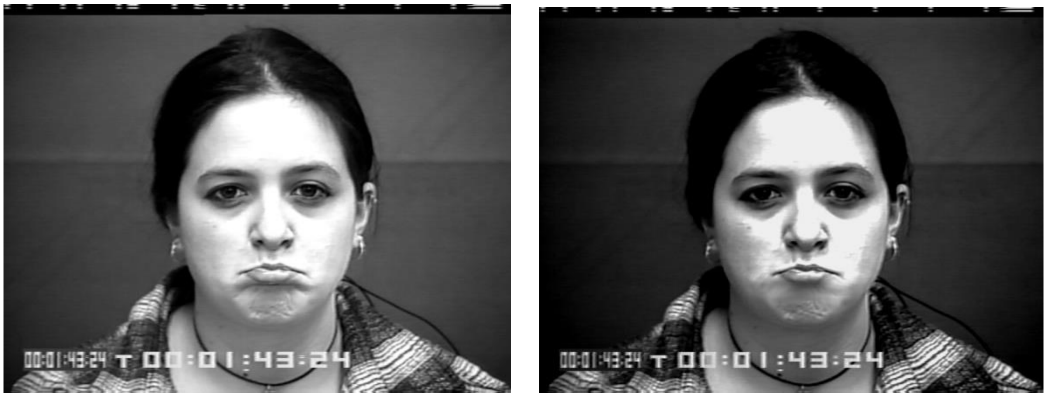


*Figure 5, example of image pre-processing before & after.*

## Shuffle the dataset:

The second step taken to improve the algorithm was to shuffle the dataset. After feature mining, data has a particular order to label order, this can negatively affect model learning when using some algorithms. Many sources state that it is good practice to shuffle the dataset to avoid this problem (ODSC - Open Data Science, 2019), although it does not work in some algorithms, it is an easy preventive solution. The whole sense of mixing the data is the fact that the algorithm can make any generalization on the whole set of data. To prevent algorithm to load data in order, before conducting tests, the rows in the whole dataset has been mixed and split for training and testing sets.

## Adding more features:

Next, the decision was made to create more features. The standard features had only calculated few main distances between key landmarks on the face, for example the distance between left lip & right lip and its width & height (see figure 6), which are not able to assess whatever or not mouth or eyes are actually open, whilst its distance can have a huge impact when assessing facial expression, such us "surprise" as commonly the eyes of the person in this case are wide open.  Therefore, an additional feature was added to also calculate the distance between top and bottom lip and the inner eye height value. An extra parameter was also added to measure the distance between the edges of the lips, nose, and chin (see figure 6). The aim of adding extra features was to allow the algorithm to learn based on the bigger amount of information and to increase the difference between images that had similar parameters.

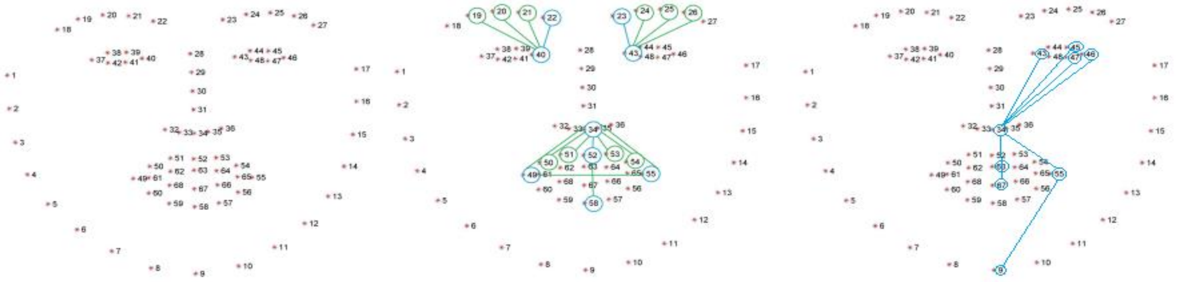


*Figure 6, first face shows points detector without any features, second the basic features implemented at the beginning of the project, third face the extra added features.*

## Size of the datasets:

Use of the large datasets can largely contribute to achieving very good results, a great example of it is Google Translate, in which case a trillion rows of dataset were used. However, the key and most important requirement, in that case, is the quality of the set. The model trained on a reliable and on more trusted is more likely to achieve better results. Although, it can be very difficult to uphold high data quality with larger data sets because it can be more difficult to control and maintain. The use of smaller datasets but of better quality may in effect be better than the use of uncertain quality but of a large set of images (Developers, 2020).

## Change of the algorithm:

Currently, there are many different classification algorithms available that differ from other algorithms in the way how they work, their purpose, features, or limitations (Microsoft, 2020). Each of them has a different principle of functioning, so it is obvious that their results will vary depending on the situation. For the needs of the project, some of the restrictions such as learning time or scalability have been omitted. It allowed to test, based on the same data, different algorithms using the Net.ML.AUTO library and then compare the results and choose the best algorithm for this project.

### Pre-train algorithm:

For the needs of this project, a program had also been created using DNN and ResNet50, which uses learning algorithms based on the pre-train model. This algorithm learns from the provided photos and tries to recognize and classify data based on its pre-train model. In this case, no manual extraction of features was needed because the algorithm tries to do it itself and decide what features to generate. This method is simpler to implement. Creating this project aimed to compare the effectiveness of the used teaching method with the learning using pre train models.

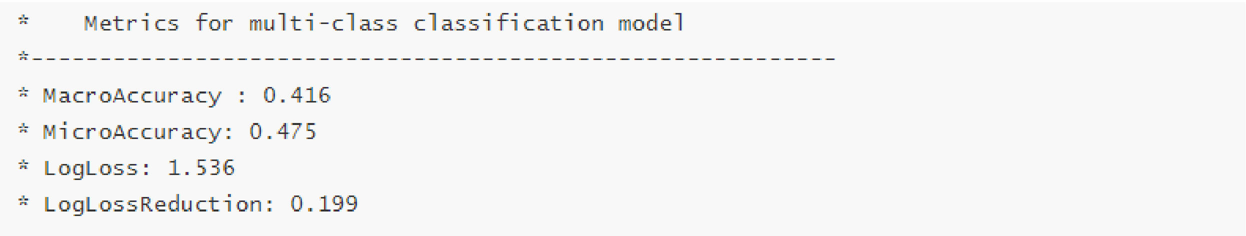
### Testing underfitting of the model:

A new set of validation data was created based on 14 photos of different emotions and described with an appropriate label. Emotions on the faces clearly define each of the categories, which was very important to verify how the learned model copes with their predictions. Based on the results, it will be possible to assess whether the model is underfitted and how it will change after entire study.

## Results

### Test 1

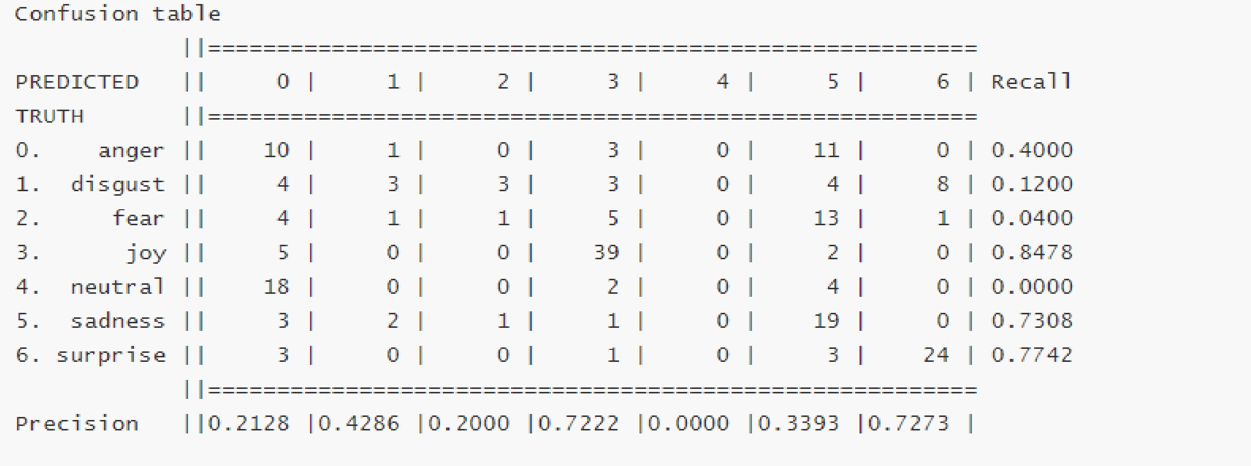
Table one shows the results from first test using both datasets before any data or image pre-processing. The model was created using classification and ML.NET framework. The pipeline created used the maximum entropy classification algorithm which has been loaded with previously generated data with the basic facial features.



```
*    Metrics for multi-class classification model
*-----------------------------------------------------------
* MacroAccuracy : 0.416
* MicroAccuracy: 0.475
* LogLoss: 1.536
* LogLossReduction: 0.199
```

*Table 1, metrics for multi-classification model*

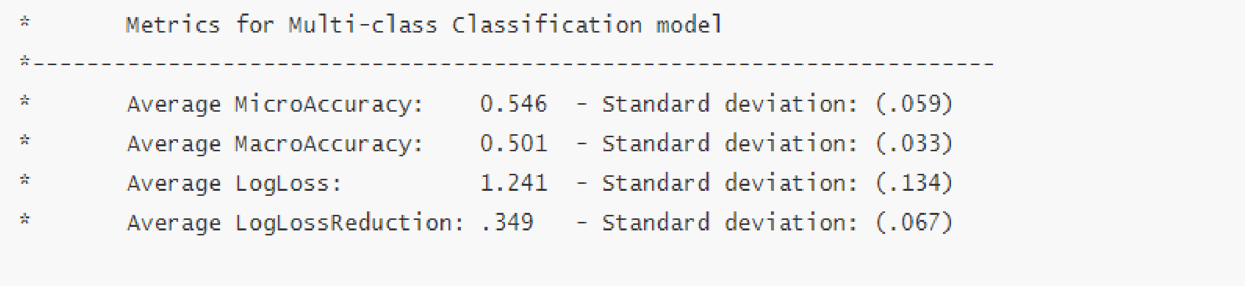
The results of the first test show that the model achieved 42% of macro accuracy and 47% micro accuracy. Log Loss results had been quite high – 1.536. Accuracy shows in how many percentages model had meet the predictions, although this metric is not always the best indicator due to its simplicity and its yes or no nature. Whilst, Log Loss results show the performance of the model, as it considers the uncertainty of the prediction based on how much it differs from the actual label, making it more valuable data.



```
Confusion table
            ||========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||========================================================
0.    anger ||    10 |     1 |     0 |     3 |     0 |    11 |     0 | 0.4000
1.  disgust ||     4 |     3 |     3 |     3 |     0 |     4 |     8 | 0.1200
2.     fear ||     4 |     1 |     1 |     5 |     0 |    13 |     1 | 0.0400
3.      joy ||     5 |     0 |     0 |    39 |     0 |     2 |     0 | 0.8478
4.  neutral ||    18 |     0 |     0 |     2 |     0 |     4 |     0 | 0.0000
5.  sadness ||     3 |     2 |     1 |     1 |     0 |    19 |     0 | 0.7308
6. surprise ||     3 |     0 |     0 |     1 |     0 |     3 |    24 | 0.7742
            ||========================================================
Precision   ||0.2128 |0.4286 |0.2000 |0.7222 |0.0000 |0.3393 |0.7273 |
```

*Table 2, metrics for confusion table*

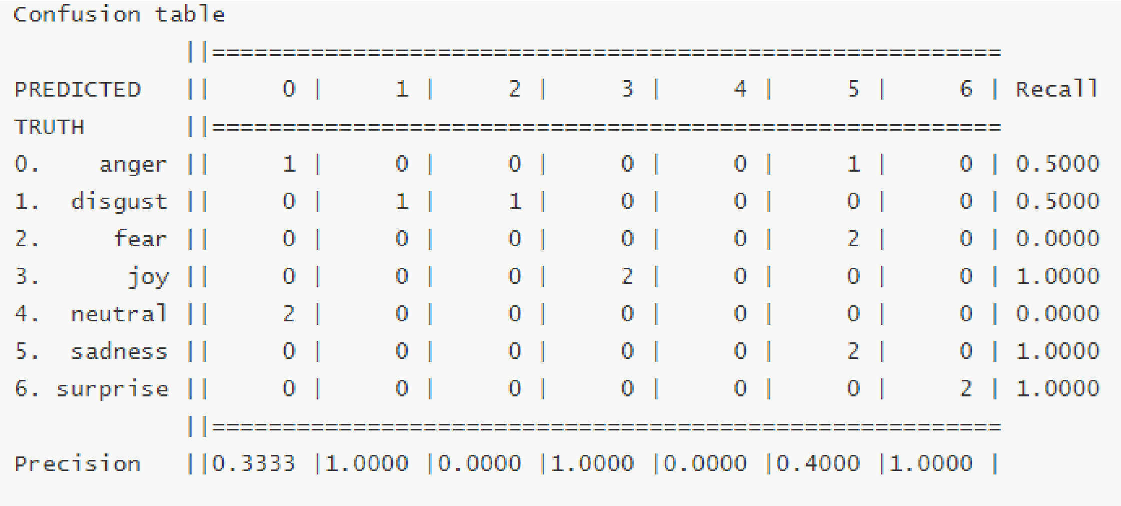
Analysing the data from the first test, it can be seen that the model had the biggest problem with recognizing emotions from the images in categories neutral, fear and disgust. In neutral category which contained 24 images, none of the facial emotions were correctly recognized and this emotion was mostly mistaken with anger. The model correctly recognized only 1 image from category fear which contained in total 25 pictures.  The table shows that the model mostly mistaken fear with sadness (13 of images label fear had been recognized as sadness). Disgust category was misinterpreted with surprise, only 3 images were guessed correctly while 8 was wrongly recognized as surprise out of the total of 25.



```
*       Metrics for Multi-class Classification model
*------------------------------------------------------------------------
*       Average MicroAccuracy:    0.546  - Standard deviation: (.059)
*       Average MacroAccuracy:    0.501  - Standard deviation: (.033)
*       Average LogLoss:          1.241  - Standard deviation: (.134)
*       Average LogLossReduction: .349   - Standard deviation: (.067)
```

*Table 3, cross-validation metrics*

To improve the accuracy of the results cross-validation has been done to create an estimate performance and precision of the model. Table below shows the results of the cross-validation test based on the standard features and without pre-processing the data and images. Based on the cross-validation test average micro accuracy is 0.55%, average macro accuracy is 0.50% and the average log loss is 1.241.



```
Confusion table
              ||========================================================
PREDICTED     ||      0 |      1 |      2 |      3 |      4 |      5 |      6 | Recall
TRUTH         ||========================================================
0.    anger   ||      1 |      0 |      0 |      0 |      0 |      1 |      0 | 0.5000
1.  disgust   ||      0 |      1 |      1 |      0 |      0 |      0 |      0 | 0.5000
2.     fear   ||      0 |      0 |      0 |      0 |      0 |      2 |      0 | 0.0000
3.      joy   ||      0 |      0 |      0 |      2 |      0 |      0 |      0 | 1.0000
4.  neutral   ||      2 |      0 |      0 |      0 |      0 |      0 |      0 | 0.0000
5.  sadness   ||      0 |      0 |      0 |      0 |      0 |      2 |      0 | 1.0000
6. surprise   ||      0 |      0 |      0 |      0 |      0 |      0 |      2 | 1.0000
              ||========================================================
Precision     ||0.3333 |1.0000 |0.0000 |1.0000 |0.0000 |0.4000 |1.0000 |
```

*Table 4, metrics for confusion table for validation data*

Validation data showed underfitting of the model. Neutral and fear category had zero recall in this test, whilst model achieved to correctly recall anger and disgust only once.

## Test 2

The first step to improve the model included data preparation and image pre-processing. The analyse of the work of the framework "Dlib" responsible for producing the detectors points showed that in total 51 images were wrongly detected (see figure 7 & 8). Out of 51 images 39 has been deleted due to its poor quality and the rest has been successfully improved by contrast enhancement and use of histogram tool to reduce any shadows and improve any imperfections.

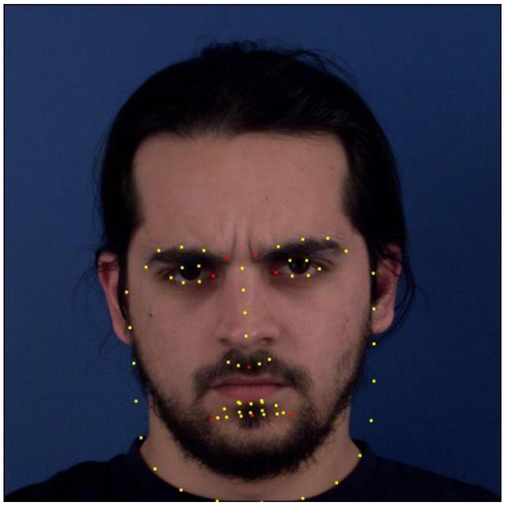


*Figure 7, example of badly marked image*

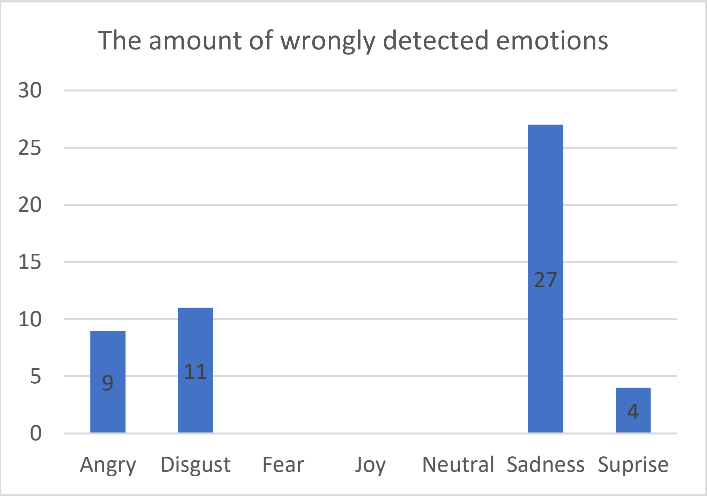
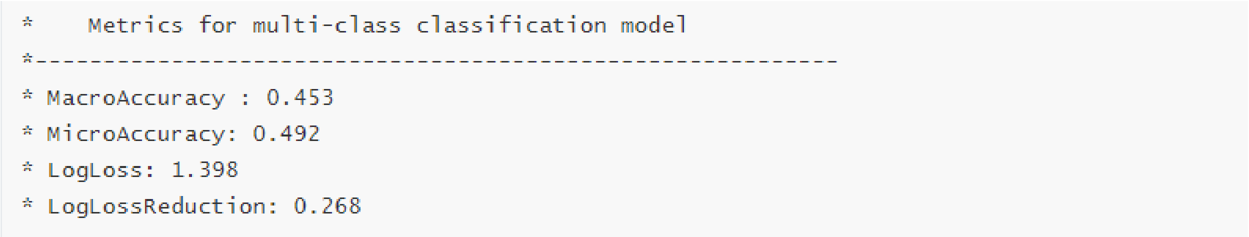


*Figure 8, the amount of wrongly detected emotions*

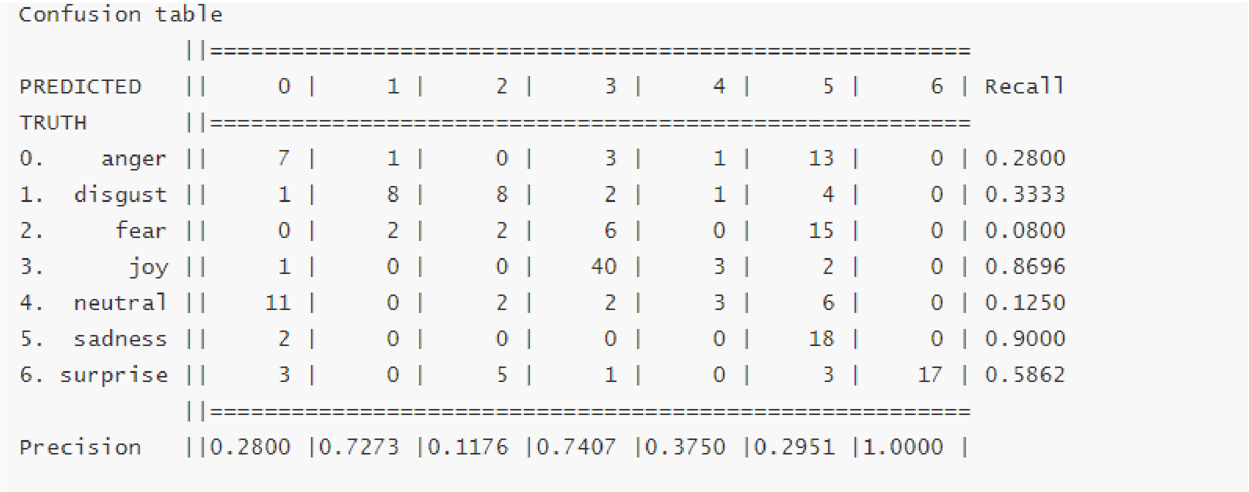
New test has been run including the improved images with a total amount of 970 pictures from two datasets, to evaluate if the improvements had any influence on the model performance and accuracy.



```
*    Metrics for multi-class classification model
*-----------------------------------------------------------
* MacroAccuracy : 0.453
* MicroAccuracy: 0.492
* LogLoss: 1.398
* LogLossReduction: 0.268
```

*Table 5, metrics for multi-classification model*

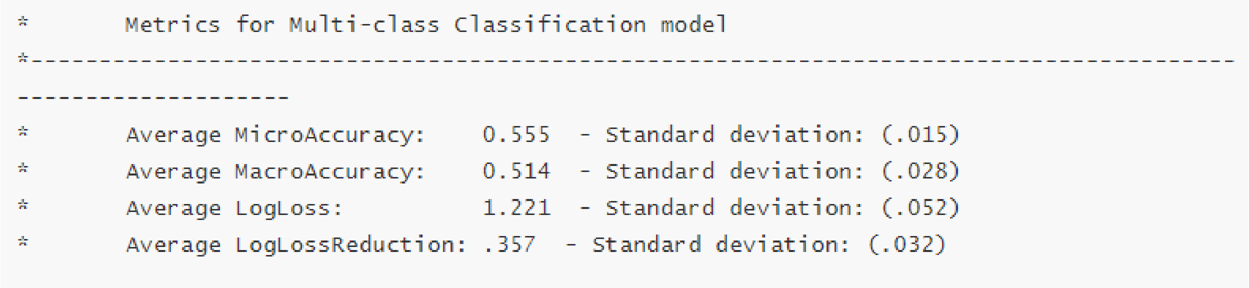
Second test of the model shows improvement in both macro and micro accuracy as well as in log loss metric.



```
Confusion table
            ||========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||========================================================
0.    anger ||     7 |     1 |     0 |     3 |     1 |    13 |     0 | 0.2800
1.  disgust ||     1 |     8 |     8 |     2 |     1 |     4 |     0 | 0.3333
2.     fear ||     0 |     2 |     2 |     6 |     0 |    15 |     0 | 0.0800
3.      joy ||     1 |     0 |     0 |    40 |     3 |     2 |     0 | 0.8696
4.  neutral ||    11 |     0 |     2 |     2 |     3 |     6 |     0 | 0.1250
5.  sadness ||     2 |     0 |     0 |     0 |     0 |    18 |     0 | 0.9000
6. surprise ||     3 |     0 |     5 |     1 |     0 |     3 |    17 | 0.5862
            ||========================================================
Precision   ||0.2800 |0.7273 |0.1176 |0.7407 |0.3750 |0.2951 |1.0000 |
```

*Table 6, metrics for confusion table*

The confusion table also shows improvement in both recall and precision metrics. The model correctly recognized 12% of neutral facial expressions comparing to the 0% recall from the first test, although the model still highly confused neutral emotion with anger. There also have been a small improvement in recall of fear emotion from 4% to 8% and the model correctly recognized 33% of disgust emotions compering to 12% from the previous test. Overall recall has increased in five categories and decreased in two created more balanced results. In addition, results show increase in precision for most of the categories, a slight decreased precision can be seen for fear and sadness category.



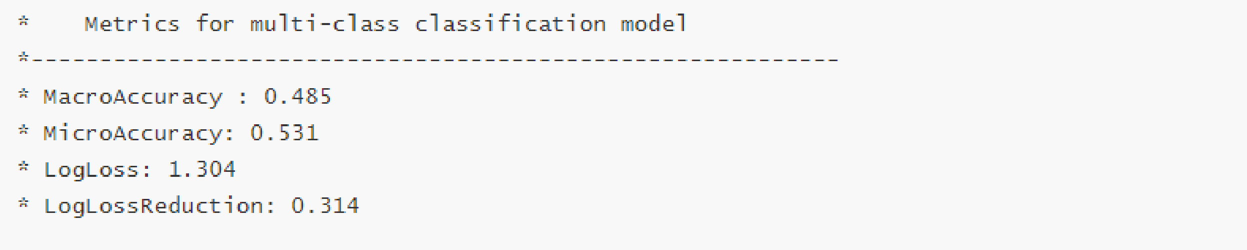
```
*       Metrics for Multi-class Classification model
*------------------------------------------------------------------------------------------
--------------------
*       Average MicroAccuracy:    0.555  - Standard deviation: (.015)
*       Average MacroAccuracy:    0.514  - Standard deviation: (.028)
*       Average LogLoss:          1.221  - Standard deviation: (.052)
*       Average LogLossReduction: .357  - Standard deviation: (.032)
```

*Table 7, cross-validation metrics*

Cross-validation metrics shows improvement in both micro and macro accuracy as well as in log loss metric. In addition, the standard deviation for each metric has decreased. Standard deviation measures the variability of the results, it shows whether the spread of the results around average is small or large.

## Test 3

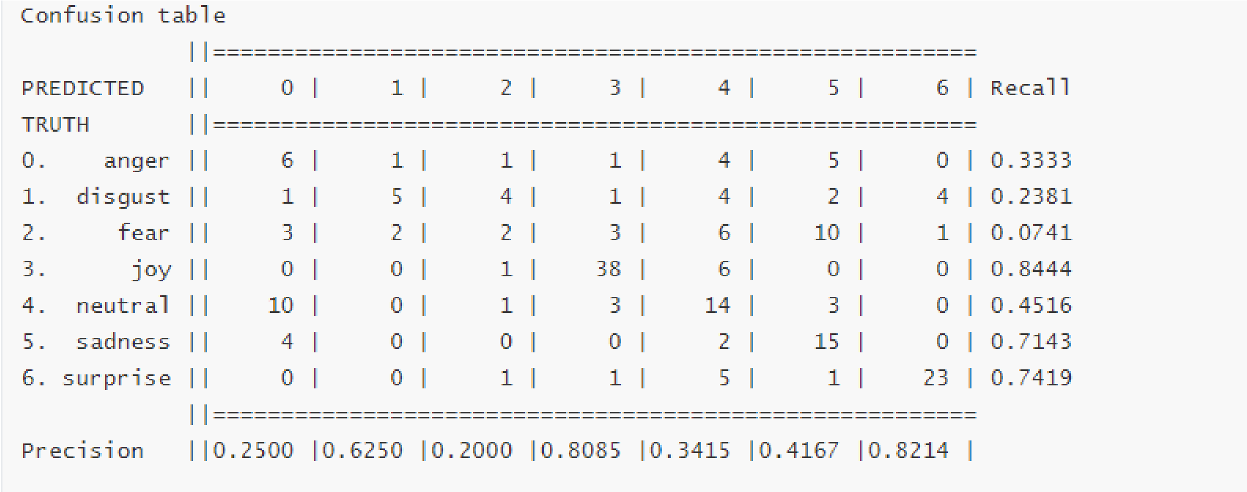
In addition to data preparation, all images have been reviewed and sorted into the best fitting categories based on the emotions on the face. Another test has been run to evaluate the changes in model performance.



```
*    Metrics for multi-class classification model
*------------------------------------------------------------
* MacroAccuracy : 0.485
* MicroAccuracy: 0.531
* LogLoss: 1.304
* LogLossReduction: 0.314
```

*Table 8, metrics for multi-classification model*

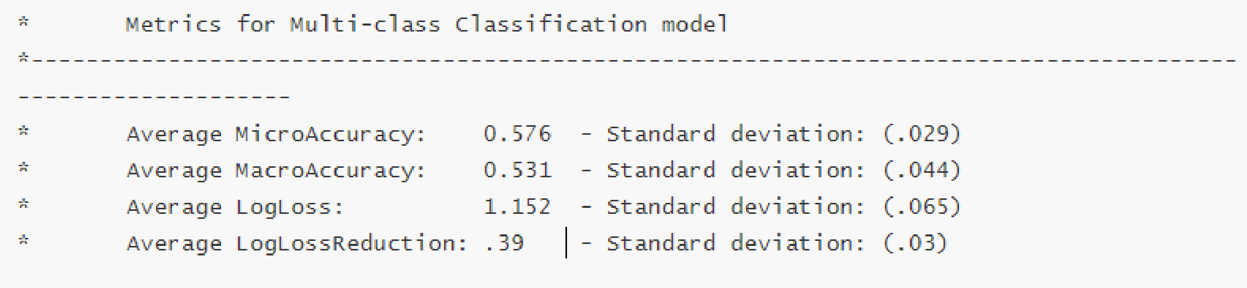
Again, implemented improvement resulted in increase of macro and micro accuracy for the model and decrease is log loss. Macro accuracy increased by 3% comparing to the results from test two, micro accuracy increased by 4% and log loss decreased from 1.398 to 1.304.



```
Confusion table
            ||========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||========================================================
0.    anger ||     6 |     1 |     1 |     1 |     4 |     5 |     0 | 0.3333
1.  disgust ||     1 |     5 |     4 |     1 |     4 |     2 |     4 | 0.2381
2.     fear ||     3 |     2 |     2 |     3 |     6 |    10 |     1 | 0.0741
3.      joy ||     0 |     0 |     1 |    38 |     6 |     0 |     0 | 0.8444
4.  neutral ||    10 |     0 |     1 |     3 |    14 |     3 |     0 | 0.4516
5.  sadness ||     4 |     0 |     0 |     0 |     2 |    15 |     0 | 0.7143
6. surprise ||     0 |     0 |     1 |     1 |     5 |     1 |    23 | 0.7419
            ||========================================================
Precision   ||0.2500 |0.6250 |0.2000 |0.8085 |0.3415 |0.4167 |0.8214 |
```

*Table 9, metrics for confusion table*

The confusion table results shows overall decrease in recall and precision for most of the categories, although the implemented improvement had a significant effect on recognition of neutral facial emotions which had increased from 12% to 45% (based on the results from test 2). None of the categories had been highly affected, the decrease in recall and precision has not been highly significant.



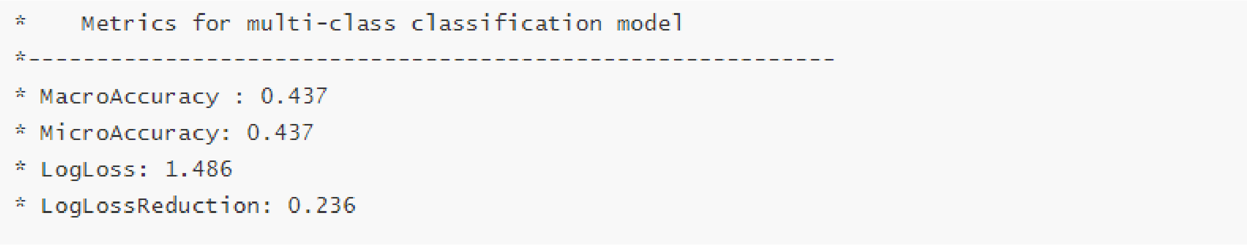
```
*       Metrics for Multi-class Classification model
*------------------------------------------------------------------------------------------
--------------------
*       Average MicroAccuracy:    0.576  - Standard deviation: (.029)
*       Average MacroAccuracy:    0.531  - Standard deviation: (.044)
*       Average LogLoss:          1.152  - Standard deviation: (.065)
*       Average LogLossReduction: .39    - Standard deviation: (.03)
```

*Table 10, cross-validation metrics*

Cross-validation results shows increased in micro and macro accuracy in both cases by 2% and decreased in log loss from 1.22 to 1.15, although standard deviation had increased slightly.

## Test 4

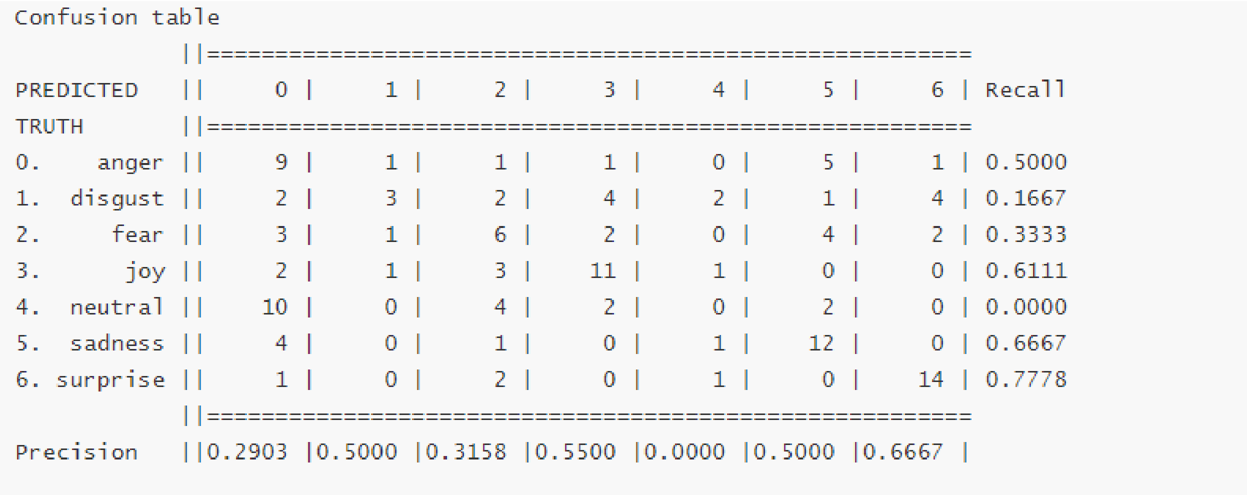
The next step taken to improve the model performance included a transformation and further categorization of the data in order to create the same amount of data (images) in each group of emotions to improve the accuracy of the model.



```
*    Metrics for multi-class classification model
*-----------------------------------------------------------
* MacroAccuracy : 0.437
* MicroAccuracy: 0.437
* LogLoss: 1.486
* LogLossReduction: 0.236
```

*Table 11, metrics for multi-classification model*

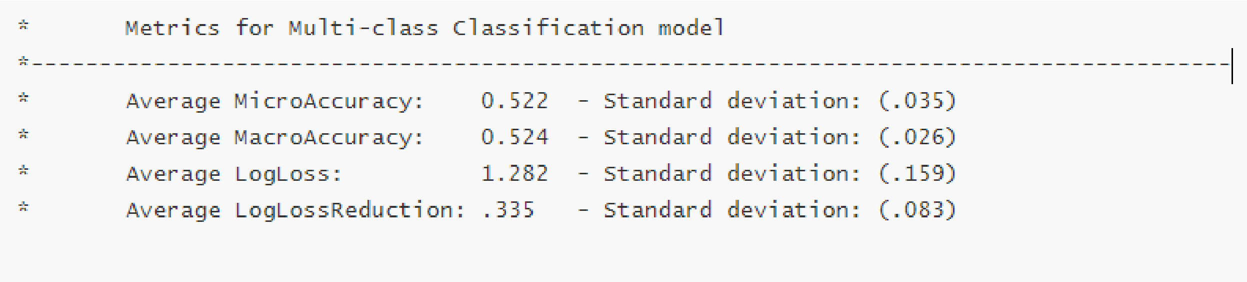
Metrics table shows significant decrease in macro and micro accuracy as well as increase in log loss metric comparing to the achievement from the test 3.



```
Confusion table
            ||========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||========================================================
0.    anger ||     9 |     1 |     1 |     1 |     0 |     5 |     1 | 0.5000
1.  disgust ||     2 |     3 |     2 |     4 |     2 |     1 |     4 | 0.1667
2.     fear ||     3 |     1 |     6 |     2 |     0 |     4 |     2 | 0.3333
3.      joy ||     2 |     1 |     3 |    11 |     1 |     0 |     0 | 0.6111
4.  neutral ||    10 |     0 |     4 |     2 |     0 |     2 |     0 | 0.0000
5.  sadness ||     4 |     0 |     1 |     0 |     1 |    12 |     0 | 0.6667
6. surprise ||     1 |     0 |     2 |     0 |     1 |     0 |    14 | 0.7778
            ||========================================================
Precision   ||0.2903 |0.5000 |0.3158 |0.5500 |0.0000 |0.5000 |0.6667 |
```

*Table 12, metrics for confusion table*

For most of the emotions recall and precision has also decreased, the model again had not correctly recognized any of the images with neutral facial expression and a high decrease in recall can be seen in joy category, from 84% to 61%. Although the implemented improvement had significantly increased the correct recognition of fear emotion, from 0.7% to 33%. The precision has fall to 0% in neutral category, although again a high increase of the precision can be seen in fear category which had increased from 20% to 31%.



```
*       Metrics for Multi-class Classification model
*-------------------------------------------------------------------------------------
*       Average MicroAccuracy:    0.522  - Standard deviation: (.035)
*       Average MacroAccuracy:    0.524  - Standard deviation: (.026)
*       Average LogLoss:          1.282  - Standard deviation: (.159)
*       Average LogLossReduction: .335   - Standard deviation: (.083)
```

*Table 13, cross-validation metrics*

Based on cross-validation average micro and macro accuracy had decreased and the log loss metric had increased comparing to the results from test 3. Although the recall has significantly increased in fear category by implementation of this improvement, overall, the results are much worst then before. Due to this fact, previous datasets without the change in the size of dataset have been used in the further tests.

## Test 5

Before conducting the next test, the datasets had been shuffled to see if this will affect the performance and the accuracy of the model. The dataset had been shuffled every time before conducting the test. In total seven tests had been done to evaluate the results (see figure 9). The results are based on cross-validation results of each test.

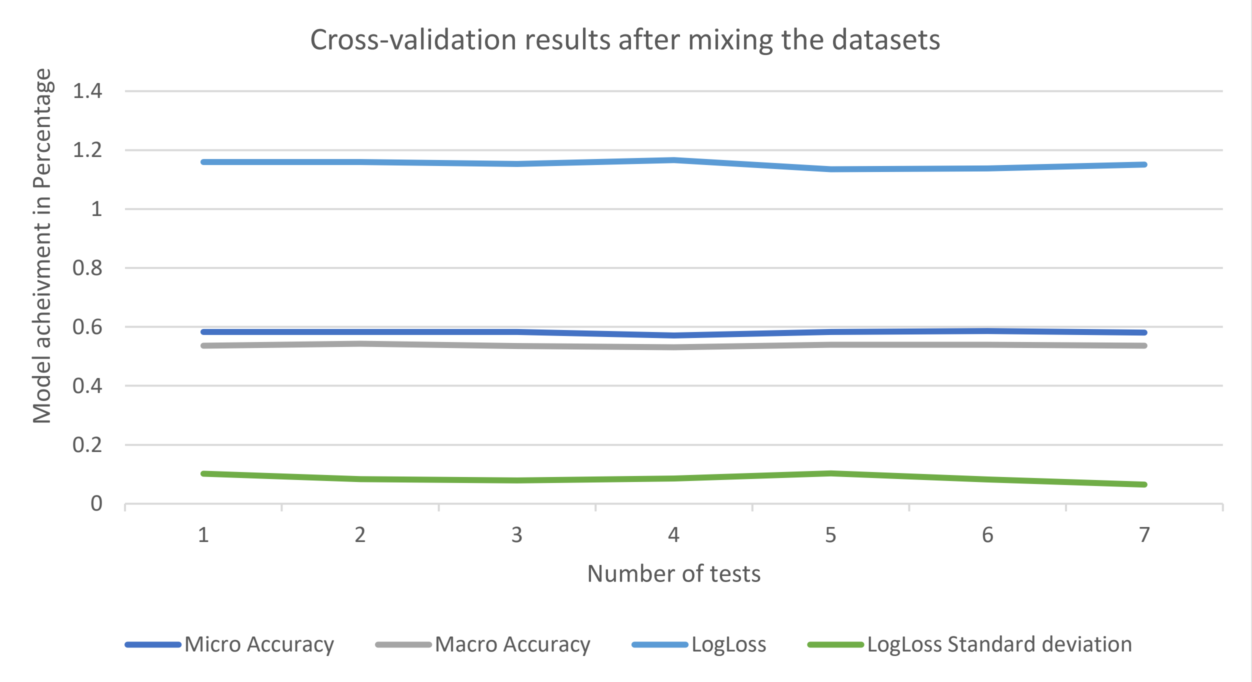
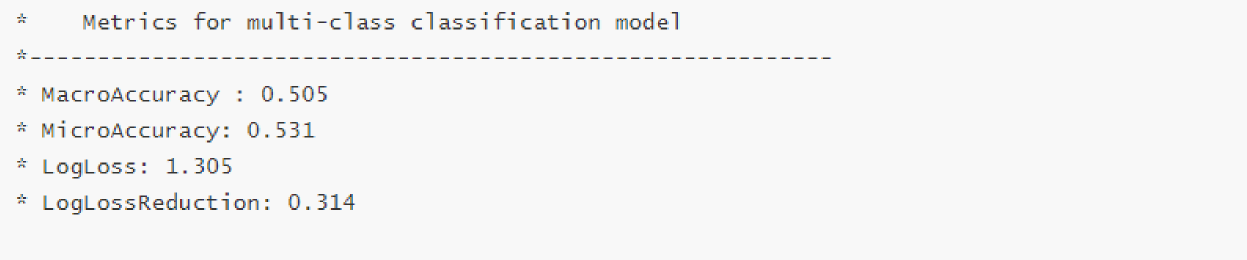


*Figure 9, cross-validation results after mixing the datasets*

The results of cross-validation from seven tests with shuffled data conducted show that mixing the data has no significant effect on model performance or its accuracy. On average based on seven tests, micro accuracy was 0.58%, macro accuracy was 0.53% and log loss metric on average was 1.15.

## Test 6

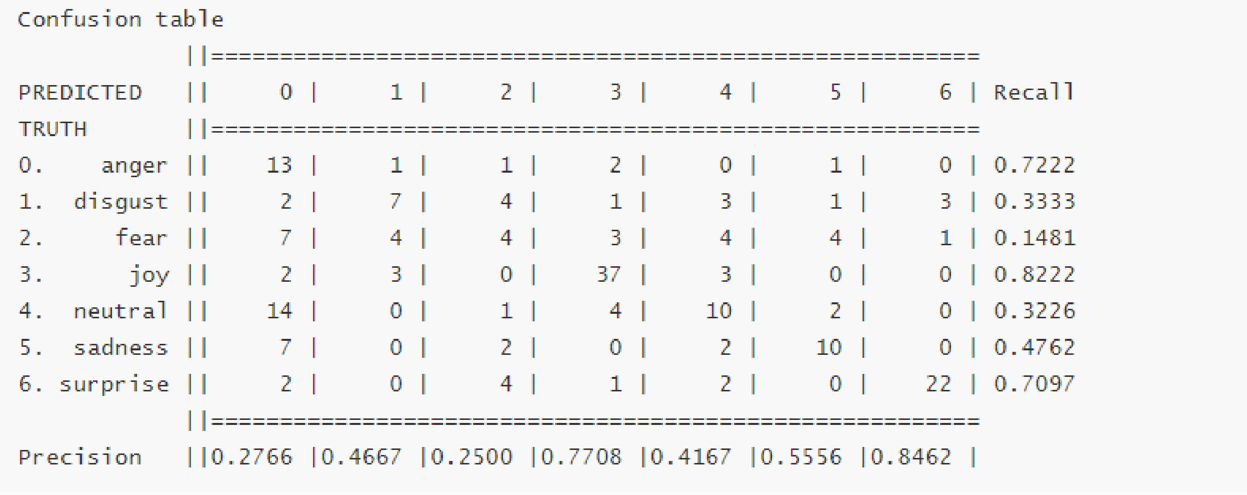
To improve the model performance more features has been added to also calculate the distance between top and bottom lip and the inner eye height value and to measure the distance between the edges of the lips, nose, and chin. The aim of adding extra features was to allow the algorithm to learn based on the larger data to increase the difference between images that had similar parameters.



```
*    Metrics for multi-class classification model
*-----------------------------------------------------------
* MacroAccuracy : 0.505
* MicroAccuracy: 0.531
* LogLoss: 1.305
* LogLossReduction: 0.314
```

*Table 14, metrics for multi-classification model*

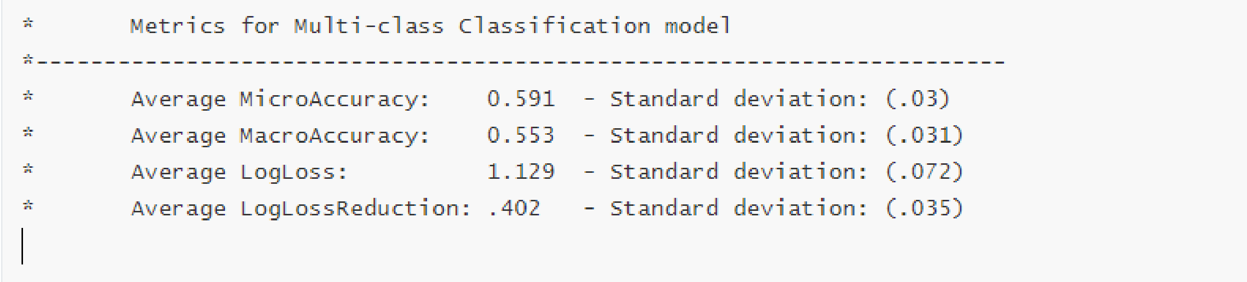
The results from test 6 show decrease in macro accuracy, micro accuracy remained the same (53%) compering to average results from test 5.  Log loss had increased to 1.30 from 1.15.



```
Confusion table
             ||========================================================
PREDICTED    ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH        ||========================================================
0.    anger  ||    13 |     1 |     1 |     2 |     0 |     1 |     0 | 0.7222
1.  disgust  ||     2 |     7 |     4 |     1 |     3 |     1 |     3 | 0.3333
2.     fear  ||     7 |     4 |     4 |     3 |     4 |     4 |     1 | 0.1481
3.      joy  ||     2 |     3 |     0 |    37 |     3 |     0 |     0 | 0.8222
4.  neutral  ||    14 |     0 |     1 |     4 |    10 |     2 |     0 | 0.3226
5.  sadness  ||     7 |     0 |     2 |     0 |     2 |    10 |     0 | 0.4762
6. surprise  ||     2 |     0 |     4 |     1 |     2 |     0 |    22 | 0.7097
             ||========================================================
Precision    ||0.2766 |0.4667 |0.2500 |0.7708 |0.4167 |0.5556 |0.8462 |
```

*Table 15, metrics for confusion table*

The recall and precision metric had been quite balanced in test 6. Adding more features resulted in model improvement in correct recognition of a fear emotion, the recall reach 14% what is the highest result achieved yet, although sadness recall had decreased mostly and gradually compering to previous tests, however it still maintains high percentage of recall overall.



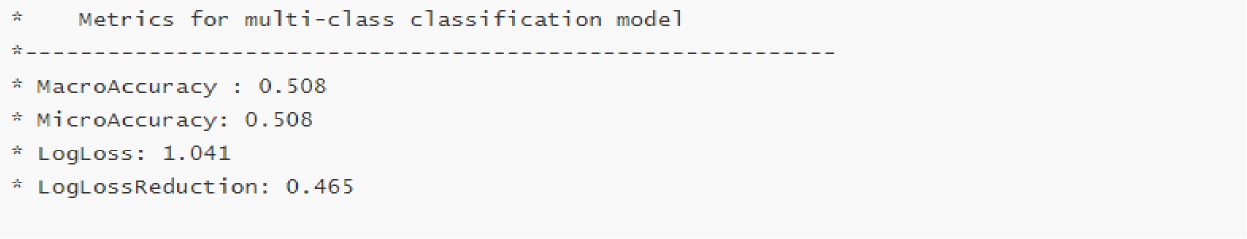
```
*       Metrics for Multi-class Classification model
*------------------------------------------------------------------------
*       Average MicroAccuracy:    0.591  - Standard deviation: (.03)
*       Average MacroAccuracy:    0.553  - Standard deviation: (.031)
*       Average LogLoss:          1.129  - Standard deviation: (.072)
*       Average LogLossReduction: .402   - Standard deviation: (.035)
```

*Table 16, cross-validation metrics*

Results for cross-validation metric shows that average micro and macro accuracy had actually increased after the implementation of new features and the log loss had decreased from 1.15 to 1.12, showing that further improvement do the model has been successful.

## Test 7

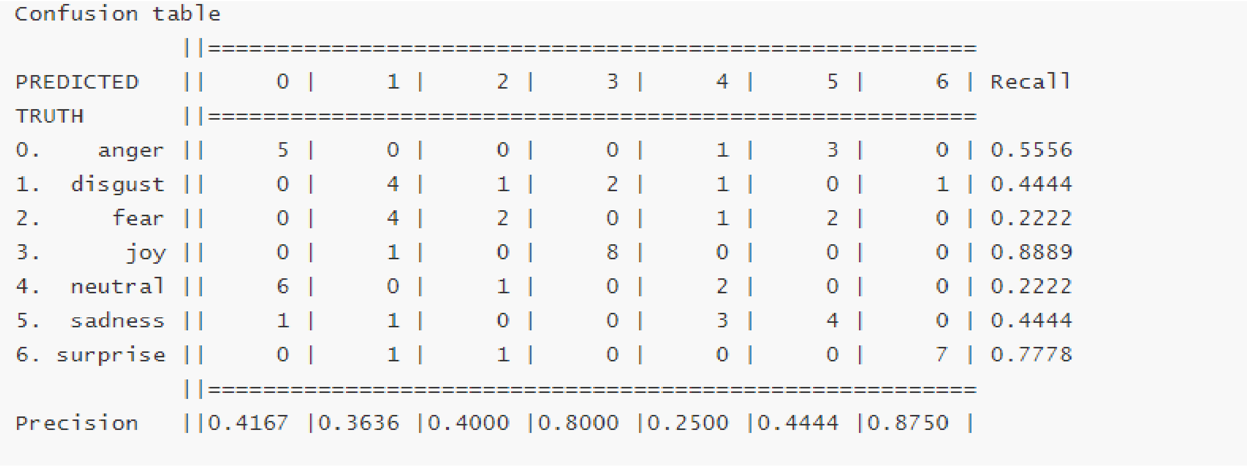
The next step taken to improve the model included decreasing the size of the dataset to 315 images, which again had been balanced so each category of emotion has the same amount of pictures in it, although this time each image had been carefully selected manually to make sure only the best quality of pictures are being tested.



```
*    Metrics for multi-class classification model
*------------------------------------------------------------
* MacroAccuracy : 0.508
* MicroAccuracy: 0.508
* LogLoss: 1.041
* LogLossReduction: 0.465
```

*Table 17, metrics for multi-classification model*

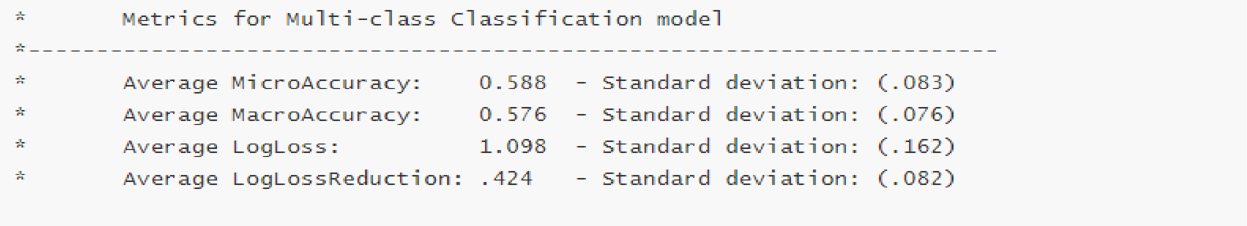
The metrics from multi-class classification model in test 7 showed decrease in macro and micro accuracy although there had been a significant decrease in log loss metric.



```
Confusion table
            ||=========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||=========================================================
0.    anger ||     5 |     0 |     0 |     0 |     1 |     3 |     0 | 0.5556
1.  disgust ||     0 |     4 |     1 |     2 |     1 |     0 |     1 | 0.4444
2.     fear ||     0 |     4 |     2 |     0 |     1 |     2 |     0 | 0.2222
3.      joy ||     0 |     1 |     0 |     8 |     0 |     0 |     0 | 0.8889
4.  neutral ||     6 |     0 |     1 |     0 |     2 |     0 |     0 | 0.2222
5.  sadness ||     1 |     1 |     0 |     0 |     3 |     4 |     0 | 0.4444
6. surprise ||     0 |     1 |     1 |     0 |     0 |     0 |     7 | 0.7778
            ||=========================================================
Precision   ||0.4167 |0.3636 |0.4000 |0.8000 |0.2500 |0.4444 |0.8750 |
```

*Table 18, metrics for confusion table*

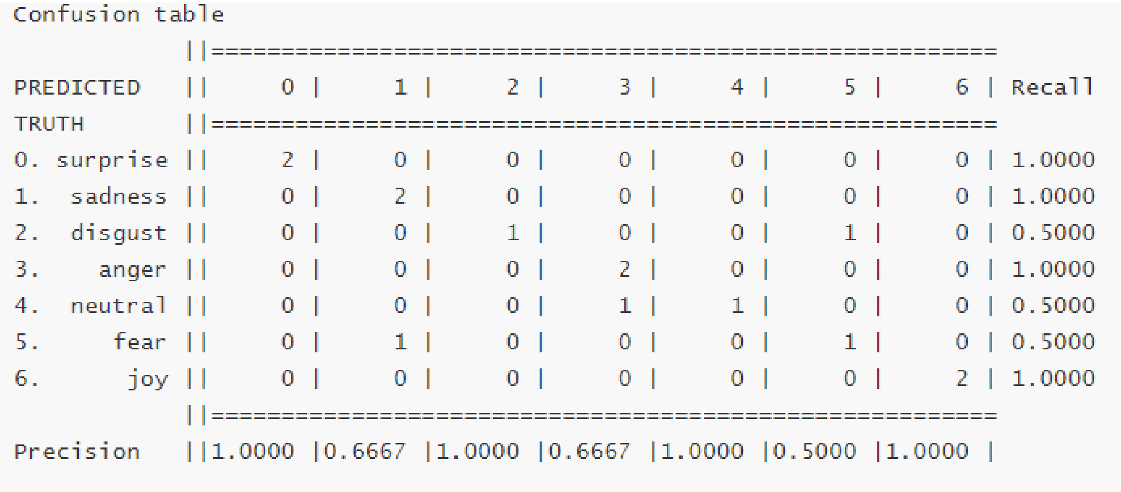
Recall and precision results show further increase in recognition of a fear emotion to 22% / 40%, there had been a quite high decrease in recall of an angry emotion by model, although the precision had increased. Overall, the metrics for confusion table are still well balanced by this point.



```
*       Metrics for Multi-class Classification model
*-------------------------------------------------------------------------
*       Average MicroAccuracy:    0.588  - Standard deviation: (.083)
*       Average MacroAccuracy:    0.576  - Standard deviation: (.076)
*       Average LogLoss:          1.098  - Standard deviation: (.162)
*       Average LogLossReduction: .424   - Standard deviation: (.082)
```

*Table 19, cross-validation metrics*

The cross-validation metrics shows that average micro accuracy had decrease but only by less than 1% making it no significant difference, whilst average macro accuracy increased from 55% to 57%. Again, decrease can be seen in average Log Loss metric which when from 1.12 to 1.09.



```
Confusion table
              ||========================================================
PREDICTED     ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH         ||========================================================
0. surprise   ||     2 |     0 |     0 |     0 |     0 |     0 |     0 | 1.0000
1.  sadness   ||     0 |     2 |     0 |     0 |     0 |     0 |     0 | 1.0000
2.  disgust   ||     0 |     0 |     1 |     0 |     0 |     1 |     0 | 0.5000
3.    anger   ||     0 |     0 |     0 |     2 |     0 |     0 |     0 | 1.0000
4.  neutral   ||     0 |     0 |     0 |     1 |     1 |     0 |     0 | 0.5000
5.     fear   ||     0 |     1 |     0 |     0 |     0 |     1 |     0 | 0.5000
6.      joy   ||     0 |     0 |     0 |     0 |     0 |     0 |     2 | 1.0000
              ||========================================================
Precision     ||1.0000 |0.6667 |1.0000 |0.6667 |1.0000 |0.5000 |1.0000 |
```

*Table 20, metrics for confusion table*

Test conducted on validation dataset, aiming to evaluate whatever or not the model is underfitted, showed improvement in recall although still model was not able to correctly predict all images from categories: disgust, neutral and fear, showing continuous problem with those categories.

## Test 8

To measure the performance and accuracy of the model based on different algorithms, an experiment using Net.ML.AUTO framework was conducted testing improved data from the latest test. After lasting 100 second experiment the results below had been obtained.

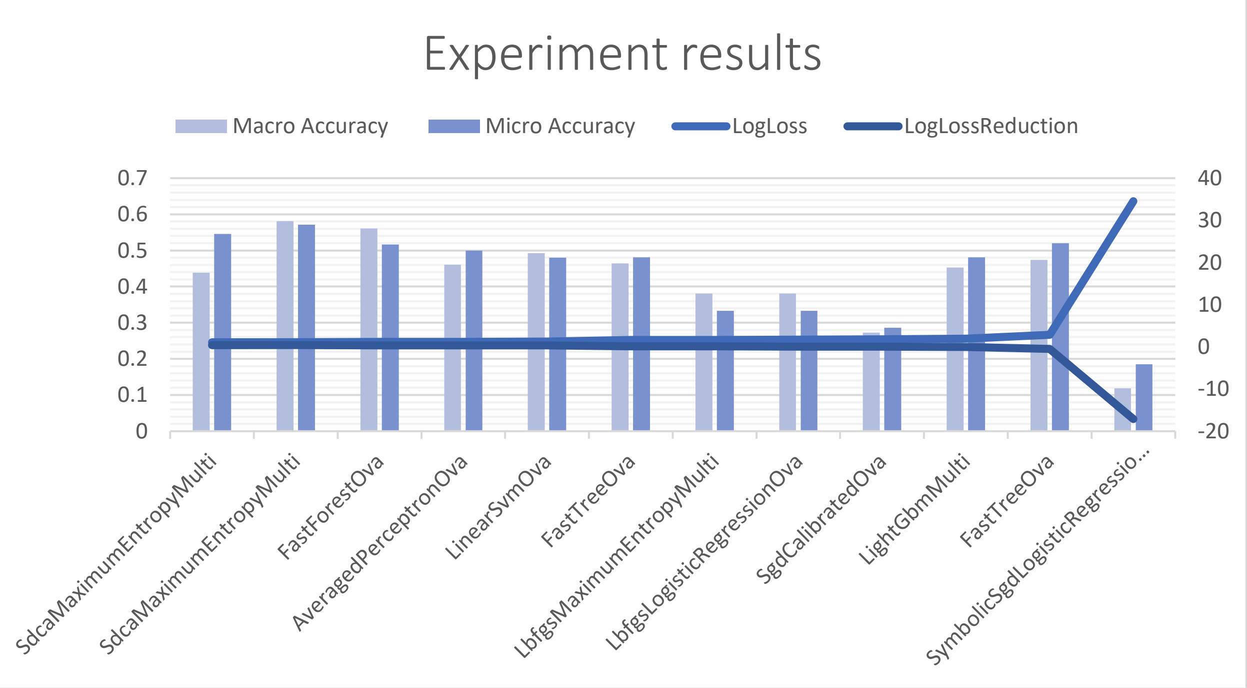
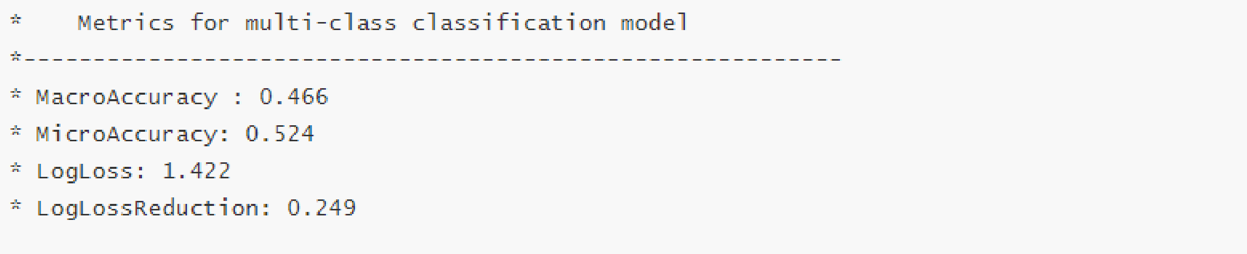


*Figure 9, experiment results*

The results show that the best algorithm is maximum entropy classification, which had been used from the beginning of this project, although decision tree binary classification and linear binary classification algorithm had performed quite well as well, resulting with close average metric to the algorithm used.

## Test 9

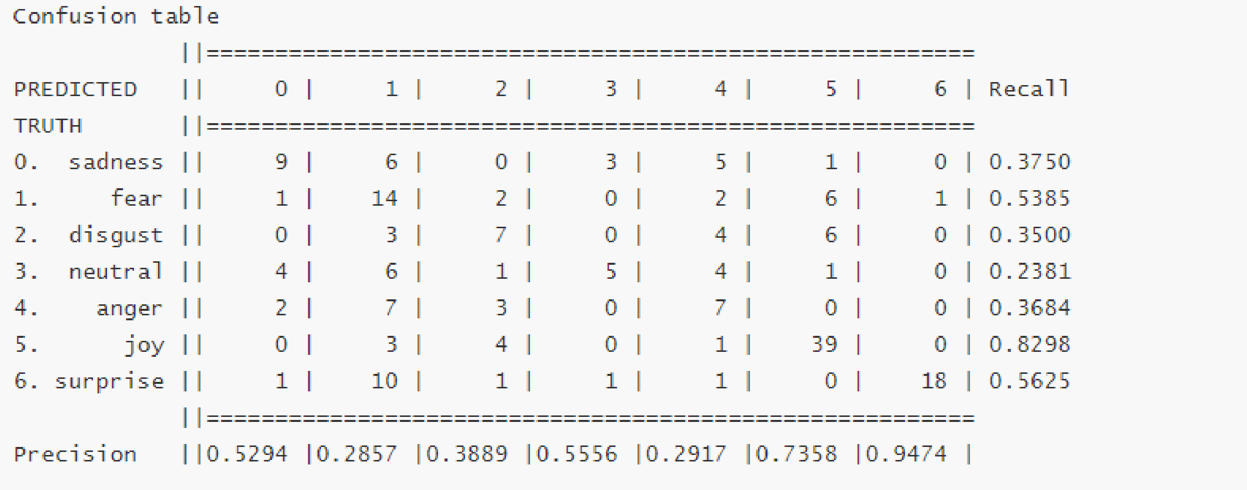
In test 9 different program had been used to recognized facial expressions using learning algorithms based on the pre-train model. The test aimed to compare the effectiveness of the used teaching method with the learning using pre-train model.



```
*    Metrics for multi-class classification model
*------------------------------------------------------------
* MacroAccuracy : 0.466
* MicroAccuracy: 0.524
* LogLoss: 1.422
* LogLossReduction: 0.249
```

*Table 21, metrics for multi-classification model*

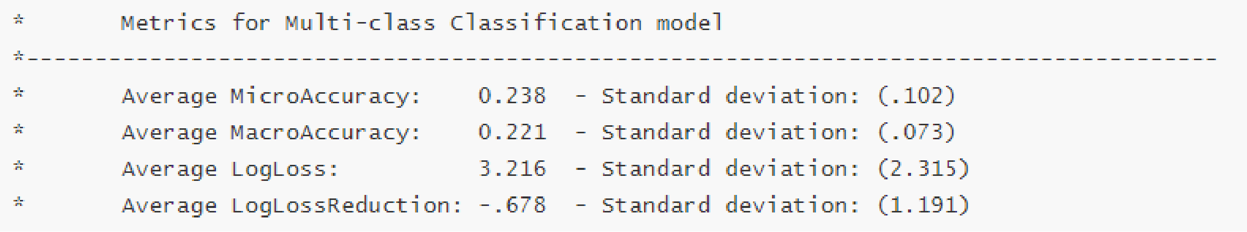
Results using pre-train teaching method shows that both macro and micro accuracy decreased, and the log loss metric had increased compering to the last results from test 7.



```
Confusion table
            ||========================================================
PREDICTED   ||     0 |     1 |     2 |     3 |     4 |     5 |     6 | Recall
TRUTH       ||========================================================
0.  sadness ||     9 |     6 |     0 |     3 |     5 |     1 |     0 | 0.3750
1.     fear ||     1 |    14 |     2 |     0 |     2 |     6 |     1 | 0.5385
2.  disgust ||     0 |     3 |     7 |     0 |     4 |     6 |     0 | 0.3500
3.  neutral ||     4 |     6 |     1 |     5 |     4 |     1 |     0 | 0.2381
4.    anger ||     2 |     7 |     3 |     0 |     7 |     0 |     0 | 0.3684
5.      joy ||     0 |     3 |     4 |     0 |     1 |    39 |     0 | 0.8298
6. surprise ||     1 |    10 |     1 |     1 |     1 |     0 |    18 | 0.5625
            ||========================================================
Precision   ||0.5294 |0.2857 |0.3889 |0.5556 |0.2917 |0.7358 |0.9474 |
```

*Table 22, metrics for confusion table*

Overall, recall and precision using this method is well balanced, although the results for most of the categories are lower than in previous test.



```
*       Metrics for Multi-class Classification model
*-------------------------------------------------------------------------------------
*       Average MicroAccuracy:    0.238  - Standard deviation: (.102)
*       Average MacroAccuracy:    0.221  - Standard deviation: (.073)
*       Average LogLoss:          3.216  - Standard deviation: (2.315)
*       Average LogLossReduction: -.678  - Standard deviation: (1.191)
```

*Table 23, cross-validation metrics*

Cross-validation test showed that pre-train method had achieved significantly worst results, the micro and macro accuracy had decreased to 23% & 22% and log loss had increased to 3.21 making it the worst result achieved in this project.

### Summary results

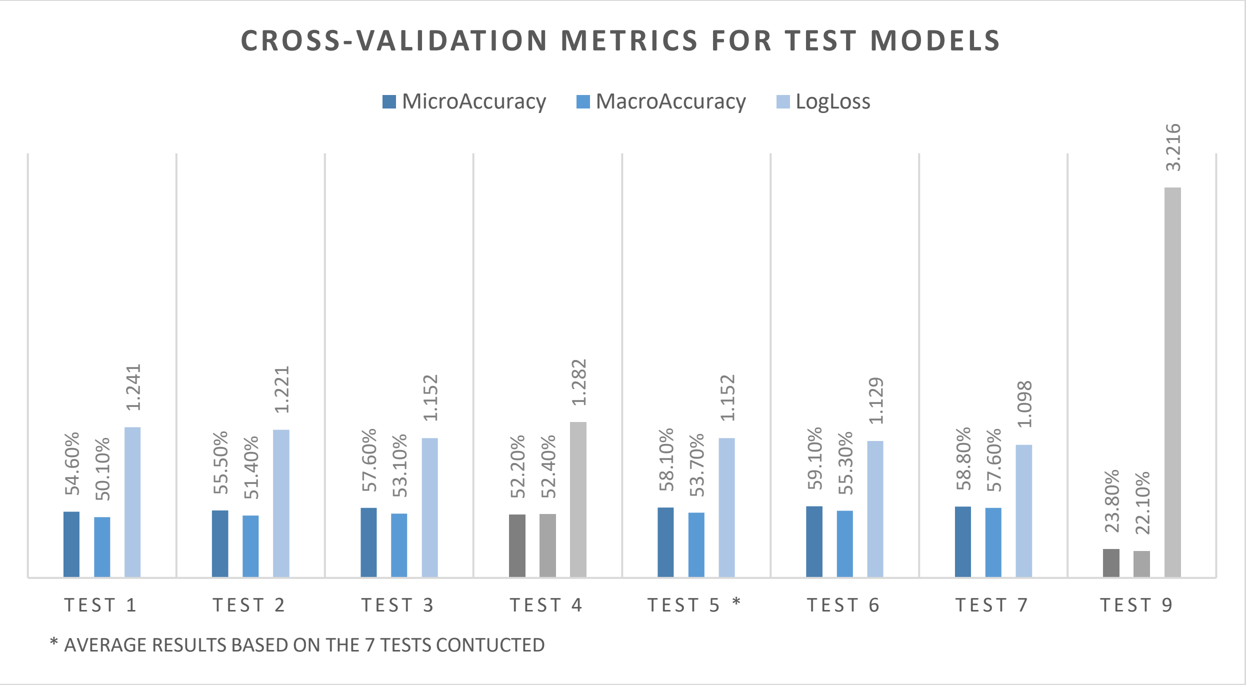


*Figure 10, cross-validation metrics for test models*

## Discussion

The first test shows the problem of underfitting of the model, which can be deduced based on the analysis of individual precision and recall of some categories, the algorithm had a problem with detecting 4 of 7 categories. Underfitting had also been confirmed by the test on validation data which showed the problem with categories: neutral, fear, disgust, and anger.

The results of the second test show that one of the reasons for the deterioration of the results may have been faulty face detection, resulting from poor quality of images used that the face detection algorithm cannot cope with. The incorrectly generated part of the data is not the ground truth, on which bases the algorithm learns. This data interpretation becomes worthless and misleads the rest of the model. Removing and fixing the faulty data, caused a noticeable improvement in recognition of some categories, to a greater or smaller extent, and overall improvement of the performance of the model. There was also a significant reduction in standard deviation for log loss, which means that the discrepancy between data had been reduced by this method.

The results of the third test show the biggest achievements in the improvement of the model caused by manually sorting the images to the best matching categories. This means that the wrongly assigned data also contributed to the wrong recognition of emotions by the algorithm. Such a change resulted in a more balanced recognition of all categories. A significant improvement is noticeable in recall and precision for the previously problematic for algorithm, neutral, fear and disgust categories, however it caused a slight decrease of recall and precision for categories which were dominant in previous tests. This means that a good classification of the data is very important, because an incorrect interpretation of individual photos by the researcher and inaccurate sorting may cause confusion of individual emotions by the algorithm, resulting in overfitting of the model.

The reduction of the total database, which was the goal of test 4, had cased worse results, significantly reducing the performance of this model, in which the neutral category was mostly interpreted as angry and again reducing the recognition and precision of this category to zero. The results are practically comparable to those from the first test, this method had neutralized virtually all improvements resulting from database preparation introduced in previous tests. Log Loss and accuracy returned to their original values. One of the reasons for such a poor performance could had been caused by removal of most of the valuable images from the dataset, leaving the less meaningful and of a poor-quality one in dataset. The second reason for the deterioration of results could had been caused by the insufficient amount of data to detect solid patterns, anomalies, and exceptions or distinctive. Therefore, for the sake of the model, the changes were abandoned, and tests were continued with the dataset in the previous test.

Based on the fifth test conducted, it can be concluded that data shuffling did not significantly affect the performance of the model and the test results were comparable. Nevertheless, using only one algorithm it cannot be state that doing so would not affect an algorithm based on other calculations. Despite the lack of visible impact on the model data shuffling is a good solution because its task is to reduce variance and to make sure the model remains general and less overfit. Training/validation sets stay representative of the overall distribution of the data.

Test 6 showed a general improvement in performance of the model, there had been a noticeable increase in recall of such categories as angry, disgust and fear, although at a small cost of correct recognition of surprise category, and significant decrease in the correct recognition of sadness emotions by the model. Despite this, the changes carried out can be considered beneficial as the overall model efficiency, precision and recall have increased for almost every category, coming closer to the matrix centre diagonal table**,** the exception is still category neutral, which is most often mistaken with the category of anger.

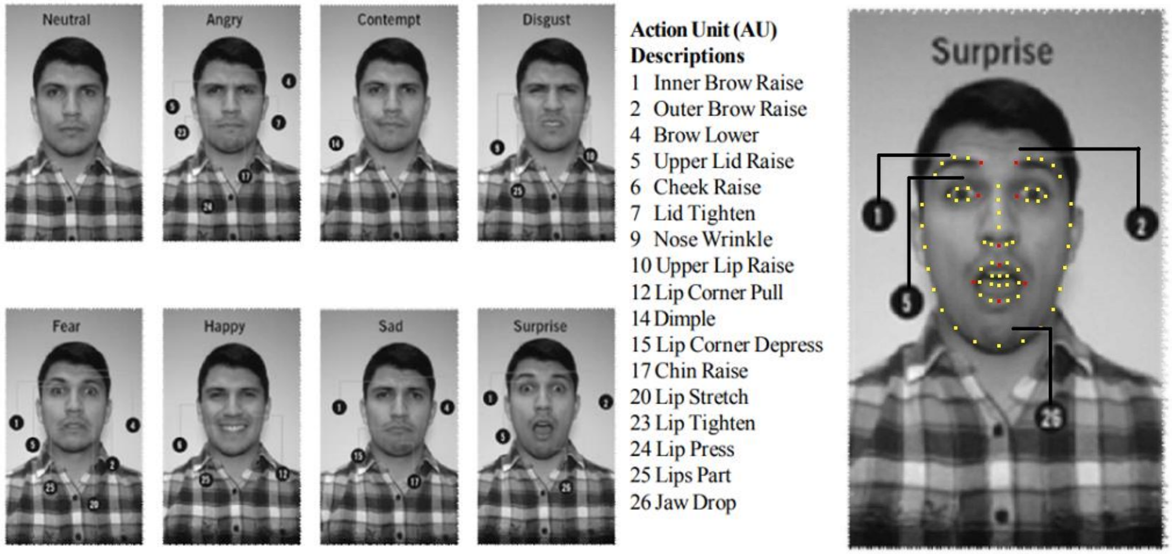


*Figure 11, example of feature extraction*

New features had improved performance, however, not as much as could it done by features based on other facial parts to classify emotions, it could be, for example, the detection of wrinkles on the face (Bayburt, 2019). Nevertheless, the “Dlib” framework does not allow the creation of such features due to its limitations, which are limited to detection of only the most important parts of the face. A consideration of using a different library can have been done to achieve this goal.

The best results so far were achieved in test 7, which included using all techniques in previous tests and appropriate sorting and reduction of datasets. The importance of the quality of the entered data is highly notable in this test, despite the smaller database size, much better results were achieved than in the first tests. Although, there had been an increase in standard deviation compering to previous achieved values, which could be improve with more data of the same high quality.

Based on the model test using validation data, the model is less underfitted than at the beginning of the project, but still requires a little of extra training to achieve the best balance between overfitting and underfitting. Small problems remained with recognition of categories such as neutral, disgust and fear. Similar conclusions can be drawn based on the confusion tables from previous model evaluation, which presents similar results.

Test 8 showed that the currently used algorithm is the best algorithm among all tested in the experiment. Nevertheless, decision tree binary classification algorithm had achieved comparable results and could have been use for further experimentation with this model.

The use of the pre-train algorithm did not work in this project, its results were significantly worse than the ones achieved by the algorithm used in this project. This may be due to the fact that the pre-train model was not designed to recognize emotions from photos, but only simple objects such as cups, cars etc.

## Conclusion

The project was based on the use of two datasets with a total of 1009 images that evolved during testing and created more or less complex dataset. The implementation of the framework for face recognition enabled the generation of numerical data for the needs of machine learning. By manipulating input data, experimenting with various other classification algorithms, or by creating a model using ResNet50 pre-train algorithms based on neural networks, it was possible to evaluate each model and the differences in its accuracy, errors, or precision.

Ultimately, during the research, a satisfactory increase of 5% in model accuracy was achieved, giving a result close to 60% and a low log loss close to 1. Various tests conducted, using different techniques (with the exception of test 4) showed a clearly visible upward trend in model performance. Based on the results, it can be concluded that each of the stages had more or less but still significant effect on the improvement of the performance of the model created by the algorithm.  Despite using a small number of images to learn the model, compared to other studies found during research, the underfitting problem was almost eliminated. Possibly increasing the dataset to several thousand with similar data quality, could had to achieved even better results.

The key limitation in this project was the use of a very limited framework for face detection based on which it is difficult to extract different meaningful measurements. Leaning only on the contours of the face and its few parts in comparison to what a human being can see seems to be insufficient. Adding the ability to detect wrinkles or details such as skin tone and shadow may be a valuable data for the algorithm, although such an excess of data can result in overfitting.

The test results could have been higher as there are more ways to improve models such as influencing the algorithm itself or using different hyperparameters that were not used during this project.

The use of the pre-train algorithm based on DNN + ResNet50 seemed very promising at first, but the results did not bring satisfactory results. This is probably because this algorithm had not been created in terms of face element detection but only object detection. Currently, the use of one of the classification algorithms and the appropriate preparation of data and parameters is a better solution as it allows to have a greater control during implementation of different techniques to improve performance.

Each of the techniques implemented had smaller or greater impact on model creation, that's why results should be analyse at every step and compared to previous ones. Such an information can be used to better understand and analyse the problem and decide which steps will contribute more to solving the problem.

# References

Aifanti, N. P. C. a. D. A., 2010. *The MUG facial expression database..* [Online]
Available at: https://mug.ee.auth.gr/fed/
[Accessed 10 05 2020].

Bayburt, E. D. &., 2019. *The Effect of Taboo Content on Incidental Vocabulary Acquisition in a Foreign Language: A Facial Expression Analysis Study Emrah Dolgunsöz Bayburt University, Turkey,* Turkey: s.n.

Clara R. Freeman, C. E. W., 2018. *Emotion Recognition Biases in Alcohol Use Disorder,* s.l.: s.n.

Developers, G., 2020. *Data Preparation and Feature Engineering for Machine Learning.* [Online]
Available at: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality
[Accessed 06 05 2020].

Gupta, S., 2018. Facial Emotion Detection using AI: Use-Cases. *Paralleldots.*

K.S.Gayathri, A. S. &. G. P. &. D., 2019. *Facial Emotion Recognition using Convolutional,* N/A: N/A.

Kanade, T. C. J. a. T. Y., 2000. *Cohn-Kanade AU-Coded Facial Expression Database.* [Online]
Available at: http://www.pitt.edu/~emotion/ck-spread.htm
[Accessed 10 05 2020].

Mahoor, A. M. &. D. C. &. M. H., 2016. *Going Deeper in Facial Expression Recognition using Deep Neural Networks,* Denver: University of Denver.

Microsoft, 2020. *How to choose an ML.NET algorithm.* [Online]
Available at: https://docs.microsoft.com/en-us/dotnet/machine-learning/how-to-choose-an-ml-net-algorithm
[Accessed 06 05 2020].

ODSC - Open Data Science, 2019. Properly Setting the Random Seed in ML Experiments. Not as Simple as You Might Imagine. *Medium.*

Pintelas, S. B. K. &. D. K. &. P. E., 2006. Data Preprocessing for Supervised Leaning. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME,* Issue 1.

Qiao, L. T. &. K. Z. &. K. W. &. X. Z. &. X. P. &. Y., 2017. *Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs,* s.l.: International Conference on Multimodal Interaction .

Ramirez, C. R. F. &. C. E. W. &. M. E. S. &. A. Z. &. V., 2015. *Emotion Recognition Biases in Alcohol Use Disorder,* s.l.: s.n.

Scikit-Learn & Keras, a. T., 2019. Chapter 1. The Machine Learning Landscape. In: *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow.* s.l.:Aurélien Géron, p. Chapter 1.

Tensorflow, S.-L. &. K. &., 2019. Chapter 2. End-to-End Machine Learning. In: *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow .* s.l.:Aurélien Géron, p. Chapter 2 .

Vishwanathan, A. S. a. S., 2008. *Introduction to Machine Learning,* Cambridge: Cambridge University Press.