COMP 6721 - FALL 2023
APPLIED ARTIFICIAL INTELLIGENCE

Project Assignment, Part I

Data Collection, Cleaning, Labeling & Preliminary Analysis

Submitted to:
Dr. René Witte

Submission by:
Group NS_19

Team Members:
Ankita Shinde : 40230690 (Evaluation Specialist)
Antas Jain : 40233532 (Data Specialist)
Muskan Gupta : 40230236 (Training Specialist)

Project Manager/TA:
Naghmeh Shafiee Roudbari

Github Link:
https://github.com/AntasJain/COMP-6721-Project

Dataset

For the Project, we have used following Datasets:

1. Dataset 1 - Challenges in Representation Learning: Facial Expression Recognition Challenge (ICML Face Data csv)

   Provenance Information:

   - Dataset Source - Kaggle Challenge
   - Images Used in the dataset are from unknown sources.
   - License - Unknown
   - URL - https://www.kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge/data

   Overview:

   - The dataset comprises a CSV with 35887 *grayscale* images of faces, each measuring *48x48 pixels*. The faces have been automatically aligned to ensure centralization and consistent spatial occupation in every image.
   - Each face is classified according to the emotion conveyed in the facial expression, assigning it to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).
   - Out of the images with these 7 emotions, we have extracted the required emotions - *Angry and Neutral*, using category number
   - Here is the Distribution of Images according to category:

Category	0	1	2	3	4	5	6
Count	4953	547	5121	8989	6077	4002	6198

   - The dataset has diverse faces with emotions, and mostly black and white backgrounds.

   Justification:

   - As per Project requirement, we needed 4 Classes, out of which three classes were required, and we chose Angry as the optional emotion.

- Out of this, Category 0 and 6 *viz* Angry(4953) and Neutral(6198) were extracted into a dataframe.
- Challenge: Using images in Array format throughout the project.

2. Dataset 2 - Student Engagement Dataset

   Provenance Information:

   - Author: Joyeeta Dey (Owner)
   - Source: Kaggle Datasets
   - License: Unknown
   - URL: https://www.kaggle.com/datasets/joyee19/studentengagement

   Overview:

   - The dataset comprises **2120** *jpg RGB* images of human faces, in landscape orientation.
   - Images are not necessarily face-centered.
   - Images are divided into two main categories:
     - Engaged - 1076 images belonging to 3 different subclasses: confused - 369 images, engaged - 347 images, frustrated - 360 images
     - Not Engaged - 1044 images belonging to 3 different subclasses: Looking away - 423 images, bored - 358 images, drowsy - 263 images
   - The dataset is less diverse than the previous dataset.
   - Total of 347 images labeled engaged and combined bored data with selected 802 Images were used for a refined dataframe.

   Justification:

   - For the project, Engaged and Bored/Tired were two classes of emotions that were necessary, this dataset provides an ample amount of images to use as a dataset for said categories.
   - Challenges: Removing noisy data, and making the images compatible with the images in the previous dataset for overall processing.

Data Cleaning

The following techniques were employed during the data cleaning process:

1. **Conversion of Images in Student Engagement Dataset (Dataset-2):**
    a. Images in the dataset were of different sizes and in RGB format, these images were cropped into face-centered copies.
    b. Images were then converted into 48x48 pixel grayscale images, to make them similar to Dataset-1 for making uniformity in the dataset.

2. **Removing useless images from Datasets:**
    a. The Dataset-1 consisted of 7 categories of emotion, out of which the required categories “Angry” and “neutral” were carefully extracted.
    b. Dataset-2 consisted of Sub-categories under Engaged Category, out of those, only Engaged Subcategories were selected, hence 347 images were selected for the combined dataset, removing subcategories: Frustrated and Confused.
    c. In Dataset-2 Length of bored data : 347, drowsy data : 32, looking away data : 423, all the three sub-categories were merged and to form total size of images under Bored/Tired Category to total of 802.

3. **Data Sampling:**
    The total count of Images per category was selected to be 1000
    a. Sampling: Data Under Category “Angry” and “neutral” which were obtained from Dataset-1 was in very large quantity, thus a total of 1000 Images from both categories were randomly selected and sampled for the refined/combined dataset.
    b. Over-Sampling & Data Augmentation: Data Under Category “Bored” and “Angry” was in less quantity, hence data augmentation was used to make the count equal to 1000 for both categories.
    For Augmentation, Random Images were selected and following filters were applied to them:
        - Random Cropping
        - Random Horizontal Flip
        - Random Color Jitters

    By Applying Data Augmentation, following benefits were achieved:
    - Dataset Size Increment, and handling class imbalance.
    - Augmenting the data with random transformations (e.g., flips, rotations, and color variations) makes the model more robust to variations

4. **Creating a new dataset:**
   A new dataset with data combination of Augmented data and originally sampled data was created for further processing. This data was shuffled to mix the classes for better processing, and would be useful for further Analysis.

5. **Calculating Pixel Density and Filtering out low Pixel Density Images:**
   - Calculating Pixel Density: The 'pixel_density' function computes the ratio of non-zero pixels to total pixels for each image.
   - Filtering Low Pixel Density Images: The next step involves filtering out images with low pixel density. A threshold is set, and images below this threshold are identified based on quantiles of the pixel density distribution.
   - Enhancing Dataset Quality: This filtering process contributes to dataset quality improvement by eliminating images with very few non-zero pixels. Such images may not offer sufficient information for effective model training. Adjusting the threshold provides flexibility in defining low pixel density based on task-specific requirements.

6. **Enhancing Images with Low Pixel Intensity:**
   - Calculating Image Intensity: Image intensity, representing the average pixel value, is calculated for each image using a function named ‘calculate_intensity’. This function calculates the mean pixel value.
   - Threshold-based Filtering: A threshold for low intensity is established using intensity distribution quantiles. Images falling below this threshold are identified for further processing with enhancement methods.
   - Enhancement Techniques: Images undergo enhancement by adjusting brightness and contrast. The optimal intermediate value is chosen through thorough analysis of images across a range of values. The following illustrates an example of image enhancements.

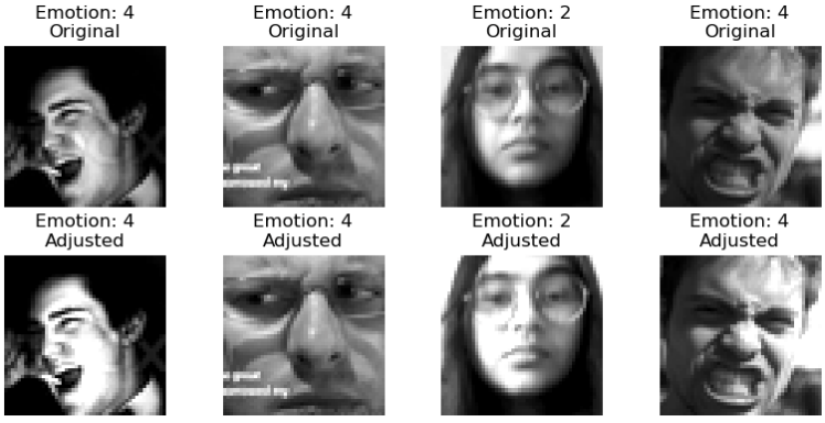

Fig: (Col 1) Low Intensity Image (Col 2) Enhanced Images

7. **Filtering and Removing High Intensity Images:**
   A threshold for high intensity was defined based on the 99th percentile of the intensity distribution. Images with intensity values exceeding this threshold were identified, as they carried little to no information, they were removed from the dataframe, and a new clean dataframe was created that can be used in next phases. This helped remove noise images such as text images and non-human images from the dataset as well.
   Example of High Intensity Images removed from dataframes:

Fig: Removed High intensity Images

Labeling

Obtaining labels from pre-existing datasets, combining datasets, and mapping emotion classes were some of the tasks involved in labeling the dataset. The decision-making process, platforms/tools utilized, ambiguities encountered, and approaches are all described in depth below.

Sources of Labels:

- **Dataset-1 (ICML)** - The main source of labeled emotion data was the Dataset-1. There were anger and neutral emotion classes in this dataset. In order to use this dataset, it was first necessary to examine the 'emotion' column, where the values 0 and 6 denoted angry and neutral emotions, respectively. These labels were changed to a more traditional format, with 1 standing for neutral feelings and 4 for angry feelings.
- **Dataset-2** - By processing image data from particular folders, additional emotion classifications were produced, including engaged, bored, tired, and looking away. After converting the image files into image arrays, distinct data frames were made for every type of emotion.

Class Mapping and Data Merging:
To create a comprehensive dataset, the dataframes from the ICML face data and the additional emotion classes were merged into a single dataframe. The labels were mapped as follows:

Neutral (Dataset-1 class-6) → 1
Angry (Dataset-1 class-0) → 4
Engaged → 2
Bored → 3
Drowsy → 3
Looking Away → 3

The decision to group Bored, drowsy and looking away into a single class (3) was made to balance class sizes and simplify the emotion classification task. The merged dataset was then shuffled to ensure randomness in the data order.

Ambiguities and Decision-Making:
Variations in image quality and conflicting label representations were found as ambiguities. Particularly, it was difficult to distinguish distinct emotion boundaries due to differences in lighting, background, and facial expressions.

Achieving a fair representation of emotion classes and taking the viability of training a model on the available data into account guided decision-making. Due to the closeness in facial expressions and the small amount of samples in each class, it was decided to combine bored, drowsy and looking away into a single class.

Challenges:

- Label Consistency: Making sure that different datasets and emotion classes are labeled consistently.

In order to overcome obstacles and produce a balanced representation of emotion classes, the labeling process comprised a combination of modifying already-existing labels, combining datasets, and making deft choices. The final dataset is an extensive compilation with the goal of capturing a wide variety of facial expressions that correlate to various emotional states.

Data Visualization

1. Class Distribution:
    a. Before Data Sampling and Augmentation, Raw Data from merging 2 datasets:

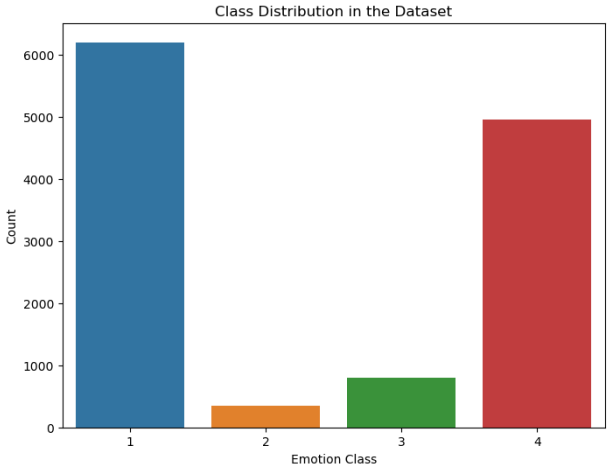

Fig. Raw Data Class Distribution

    b. After Sampling and Data Augmentation:

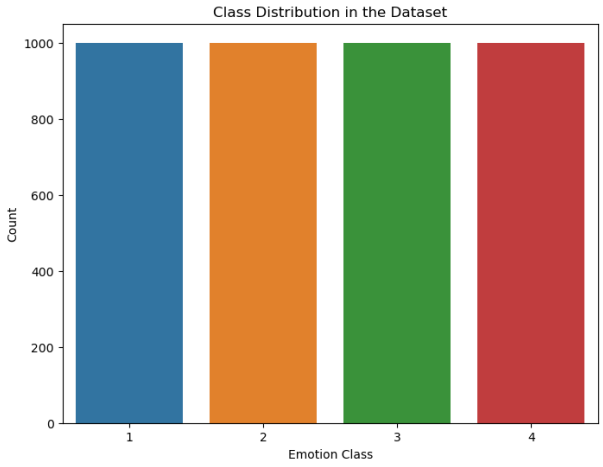

Fig. Class Distribution after Augmentation and Sampling

These Bar Diagrams represent the number of images in each category before(a) and after sampling and augmentation (b).

Where: Class 1 - Neutral, Class 2 - Engaged, Class 3 - Bored/Tired and Class 4 - Angry.

2. Sample Images

   A collection of 25 random images was presented in a 5 × 5 grid. The images were chosen randomly from each class during each code execution.

Fig. 5x5 grid of random emotions

3. Pixel Intensity Distribution

   For the same set of random images, a histogram was plotted to illustrate the distribution of pixel intensities. This histogram provides insights into variations in lighting conditions among images.

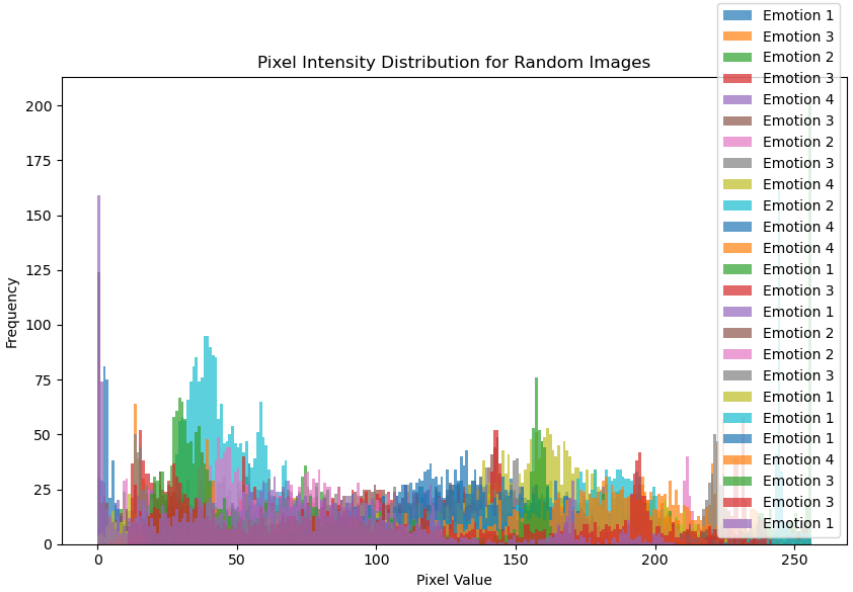

Fig. Pixel Intensity Distribution for above 25 images

4. Exploratory Analysis:
   By Filtering Low Density Images and High intensity Images, there were significant removal of images, filtering out images that cannot be used for expression analysis.

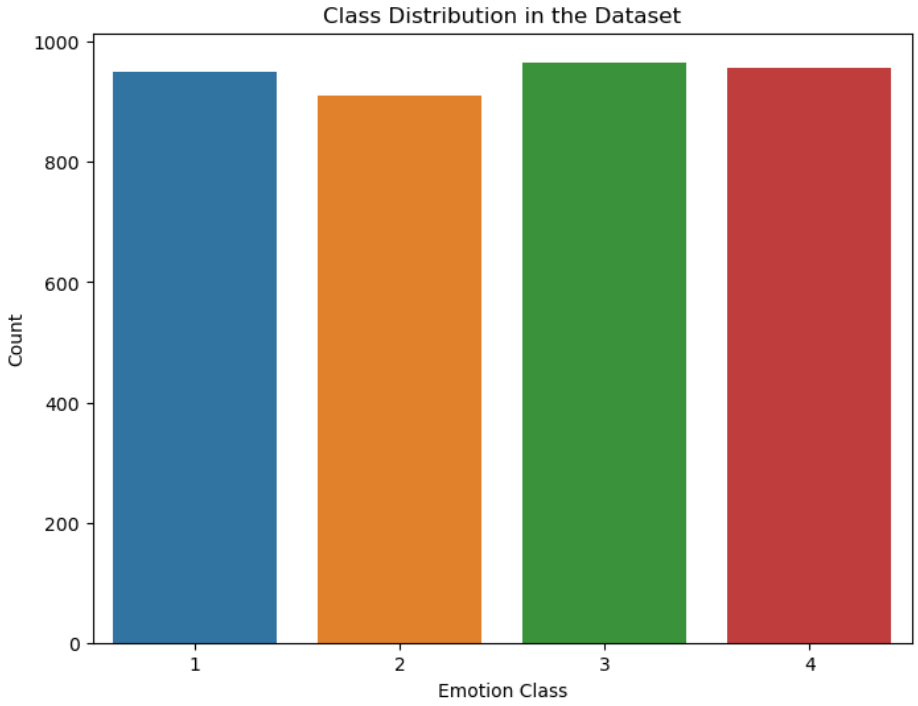

Fig. Class Distribution after Exploratory Data Analysis/Selective cleaning

References:

1] Tonichi Edeza , “ Introduction to Image Processing with Python ” , Towards Data Science, 2020 [Online], Available : https://towardsdatascience.com/introduction-to-image-processing-with-python-histogram-manipulation-for-beginners-8b107d4c4fef

2] Er. Shrawan, “Treatment of Imbalance Dataset for Human Emotion Classification”, Medium, 2023 [Online], Available : https://medium.com/@ershrawan014/treatment-of-imbalance-dataset-for-human-emotion-classification-863e4b9342d8

3]Manmeet Singh, “Image Data Augmentation for Facial Recognition”, Medium, 2020 [Online], Available: https://manmeet3.medium.com/face-data-augmentation-techniques-ace9e8ddb030

4]Pulkit Sharma, “ Build your First Multi-Label Image Classification Model in Python”, Analyticsvidhya , 2020 [Online], Available: https://www.analyticsvidhya.com/blog/2019/04/build-first-multi-label-image-classification-model-python/

5]M. T. H. Fuad et al., "Recent Advances in Deep Learning Techniques for Face Recognition," in IEEE Access, vol. 9, pp. 99112-99142, 2021, doi: 10.1109/ACCESS.2021.3096136., https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9478893

6]Debnath, T., Reza, M.M., Rahman, A. et al. Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. Sci Rep 12, 6991 (2022). https://doi.org/10.1038/s41598-022-11173-0,

7]Diah Anggraeni Pitaloka, Ajeng Wulandari, T. Basaruddin, Dewi Yanti Liliana, Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition, Procedia Computer Science, Volume 116, 2017, Pages 523-529, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.10.038.

8]Zhu D, Fu Y, Zhao X, Wang X, Yi H. Facial Emotion Recognition Using a Novel Fusion of Convolutional Neural Network and Local Binary Pattern in Crime Investigation. Comput Intell Neurosci. 2022 Sep 22;2022:2249417. doi: 10.1155/2022/2249417. PMID: 36188698; PMCID: PMC9522492. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9522492/

9]JManansala , “ Image Processing With Python: Digital Image Sampling and Quantization”, Medium, 2021 [Online], Available: https://medium.com/swlh/image-processing-with-python-digital-image-sampling-and-quantization-4d2c514e0f00