CONCORDIA UNIVERSITY

COMP 6721 Applied AI Project by NS_01

Professor: *Dr. René Witte*

Student name:

Dara Rahmat Samii (XXXXXX)
Numan Salim Shaikh (40266934)
Shahab Amrollahibioki (40292670)

Project Repository of AI - Github

1. DATA SET

The data utilized for the first phase of our project was primarily collected from two sources: Google Images and a Kaggle dataset [1].

The dataset, denoted as "fer2013," was initially prepared by Pierre-Luc Carrier and Aaron Courville as part of their ongoing research project. We were fortunate to have access to a preliminary version of their dataset, generously provided by them, which served as a valuable resource for our project. This dataset is publicly available and has been integral to our research.

To gather images from Google Images, we employed the “Image Downloader” software. This tool allowed us to systematically retrieve images from Google, providing us with a diverse and extensive set of visual data. These images were essential for expanding our dataset and ensuring its relevance to our project objectives.

This blend of data sources, including the Kaggle dataset and the images obtained from Google, formed the foundation of our dataset and supported our efforts in the first phase of the project. It enabled us to curate a comprehensive collection of facial expressions, essential for our subsequent tasks of data cleaning, labeling, and model development. The samples of each class are shown in Fig. 1 below.

Figure 1: Samples of each class of Data set

Class	No. of images	Size
Angry	4462	6.73 MB
Disgust	492	740 KB
Fear	4593	6.97 MB
Happy	8110	12.1 MB
Sad	5483	8.08 MB
Surprise	3856	5.53MB
Neutral	5572	8.25MB

Table 1: The classes of the dataset

In addition to the "fer2013" dataset, a new dataset labeled "fer2013new" was created. This dataset extends the "fer2013" dataset with additional attributes, including "NF" (Not Face) scores and a new categorical label "new_emotions." The "fer2013new" dataset consists of 35,459 samples with 12 attributes: emotion scores, pixels, usage, image name, and the newly added attributes. The "NF" score quantifies the presence of non-facial elements in an image, while "new_emotions" categorizes the student's emotional state into new classes.

1-1- Justification for Dataset Choices:

1. **"fer2013" Dataset:** The "fer2013" dataset was chosen because it contains a substantial number of facial images, making it suitable for training deep learning models. Although it has some limitations, such as potential noise in emotion labels, it provides a solid foundation for building an emotion recognition system.
2. **"fer2013new" Dataset:** The creation of the "fer2013new" dataset was motivated by the need to incorporate additional attributes, such as "NF" scores and new emotion labels. These attributes enable a more granular analysis of students' responses during academic lectures and enhance the dataset's suitability for the project's objectives.

1-2- Provenance Information:

The provenance of the "fer2013" dataset is well-documented. It was sourced from Kaggle's FER2013 Challenge, and the dataset is available for public use. The "fer2013new" dataset was generated by combining the original "fer2013" dataset with new attributes.

Dataset Name	Source	Licensing Type
Fer 2013	Kaggle FER2013 Challenge	Publicly available
Fer 2013 new	Derived from fer2013 with new attributes	Custom (derived)

Table 2. Dataset Provenance and Licensing Information

The licensing for both datasets comply with public data usage policies, and they are intended for research and educational purposes.

2- Data Cleaning Report

2-1- Data Cleaning Techniques:

Data cleaning is a critical phase in the dataset preparation process. It ensures that the dataset is standardized, contains high-quality images, and is ready for further analysis and model training.

1. **Resizing and Standardization:**

The "fer2013" dataset originally contained facial images with varying resolutions. To establish a consistent and standardized dataset, all images were resized to a uniform 48x48 pixel resolution. This resizing was performed using OpenCV, a popular Python imaging library. Standardization is essential to ensure that the Convolutional Neural Network (CNN) model can effectively process and analyze each image.

2. **Light Augmentation:**

Data augmentation techniques were applied to the images to introduce diversity and reduce overfitting. These techniques included slight rotations, brightness adjustments, and cropping. Augmentation helps the model generalize better by exposing it to different lighting conditions, facial orientations, and perspectives. For instance, by slightly rotating an image, the model can better recognize facial expressions from different angles, enhancing its overall accuracy.

3. **Handling Class Imbalance:**

The original "fer2013" dataset exhibited significant class imbalance. Certain emotion categories had far fewer samples than others. We chose to address this class imbalance at later stages in the project to avoid introducing bias during data cleaning. Techniques like oversampling and undersampling will be applied in subsequent phases to balance the dataset.

4. **Labeling Noise:**

Emotion labels in the "fer2013" dataset are not entirely error-free. Some images might be inaccurately labeled, which can adversely affect model training. To minimize the impact of labeling noise, we retained the original labels and introduced new attributes. The "NF" scores and "new_emotions" attributes provide additional information that can be used to refine the dataset and reduce the influence of noisy labels.

2-2- Challenges and Solutions: Data cleaning encountered several challenges:

1. **Class Imbalance:**

The primary challenge was the class imbalance. While we considered techniques such as oversampling or under-sampling, we decided to handle class imbalance in later phases of the project to ensure that data cleaning and preprocessing did not introduce unintended biases.

Figure 2: Samples of inaccurate emotion of Data set

2. **Handling "Not Face" Images:**

Some images in the dataset were not of faces but of other objects. These non-facial images can confuse the model. To address this, we identified these images using "NF" scores and made decisions to exclude or re-label them during data cleaning.

Figure 3: Samples of non-facial of Data set

3- Labeling Report

3-1- Labeling Methodology:

Labeling the dataset is a pivotal phase in preparing it for the subsequent machine learning model development. This report comprehensively elucidates the labeling process, encompassing class mapping, the introduction of new attributes, and methods for resolving ambiguities.

3-2- Class Mapping:

One of the first tasks undertaken was mapping the existing numerical labels from the "fer2013" dataset to corresponding human emotions. This mapping served multiple purposes, including enhancing the interpretability of the dataset and aligning labels with their underlying emotions. The mapping scheme adopted was as follows:

0: Angry
1: Disgust
2: Fear
3: Happy
4: Sad
5: Surprise
6: Neutral

This mapping, while straightforward, was instrumental in making the dataset more user-friendly and facilitating the understanding of the emotions depicted in the images.

3-3- Handling Ambiguities:

The real-world portrayal of emotions is often intricate, with various expressions intertwined in a single image. In such scenarios, the labeling process was confronted with complexities related to assigning a single, unambiguous label. To mitigate this, a predominant emotion approach was employed, where the most dominant emotion within an image was designated as the label. This approach sought to present the dataset in a clear and intuitive manner, acknowledging the primary emotion conveyed by each image.

3-4- Challenges and Solutions:

While the labeling process advanced smoothly, several challenges demanded attention and resolution:

1. **Mixed Emotions:**

Certain images presented combinations of emotions, resulting in ambiguity. Resolving this required a focus on the most prominent emotion, aligning with the labeling strategy. While this strategy simplifies the labeling process, it is imperative to acknowledge that some images may not fully encapsulate the complexity of human emotions.

2. **Labeling Noise:**

The original dataset contained instances of labeling inaccuracies. To address this, new attributes were introduced, notably the "NF" (Not Face) scores and the "new_emotions" attribute. These additions contribute to a more nuanced analysis by

providing supplementary information regarding image quality and emotions.

3. Dataset Merging and Challenges:

The consolidation of the "fer2013" dataset with an additional dataset, "fer2013new," necessitated meticulous consideration of data consistency and coherence in labeling. Merging datasets requires stringent label matching, especially when amalgamating data from disparate sources. This process had to account for potential disparities in data quality between the two datasets.

4- Dataset Visualization:

4-1- Class Distribution:

Visualizing the distribution of classes within a dataset is an essential step in understanding the dataset's balance and can reveal if any class is overrepresented or underrepresented. In the context of our project, the dataset comprises images of guys in various emotional states during activities. The visualization aims to shed light on the distribution of these emotional states, providing valuable insights for subsequent phases of the project.

4-2- Matplotlib Visualization:

For class distribution visualization, we utilized the Matplotlib library, a powerful tool for creating a wide range of data visualizations. The code, which is an integral part of our project, enabled the creation of a bar graph representing the number of images in each class. This bar graph reveals the count of images for each emotional state, which allows for a comprehensive understanding of the dataset's composition.

The dataset comprises images categorized into distinct emotional states, as follows:

- Angry
- Disgust
- Fear
- Happy
- Sad
- Surprise
- Neutral

Let's delve into the Class Distribution analysis based on our dataset:

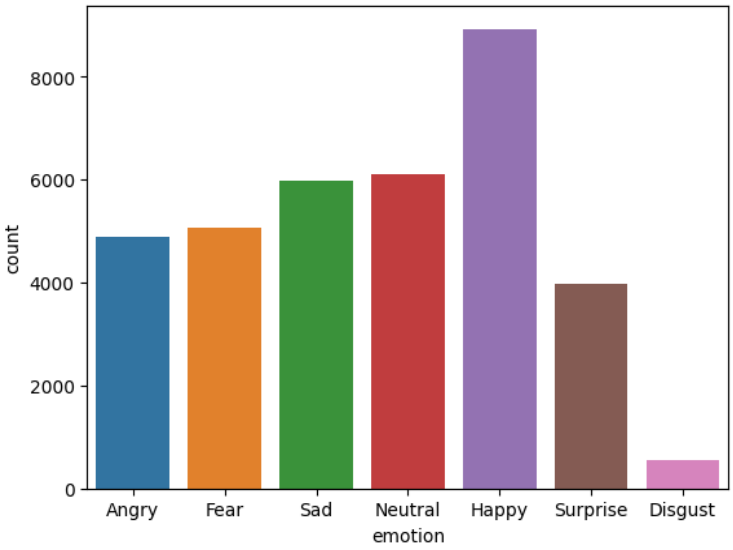

Figure 4: the Class Distribution analysis based on our dataset.

- **Angry**: Approximately 4,000 images
- **Disgust**: Approximately 500 images
- **Fear**: Approximately 3,000 images
- **Happy**: Approximately 9,000 images
- **Sad**: Approximately 5,100 images
- **Surprise**: Approximately 3,900 images
- **Neutral**: Approximately 5,900 images
-

This analysis offers several insights:

1. **Class Imbalance:**

The dataset demonstrates a significant class imbalance, with emotions such as "Happy" and "Neutral" having a higher representation compared to emotions like "Disgust" and "Surprise." This imbalance can impact the performance of machine learning models and must be considered during model development.

2. **Data Quality:**

The dataset's class distribution also provides insights into data quality. Imbalanced classes could result from the real-world prevalence of emotions, but it could also indicate data collection biases.

3. **Model Training:**

The class distribution analysis is crucial for optimizing the training of machine learning models. Techniques such as class weighting or oversampling may be required to ensure the model can effectively learn from underrepresented classes.

4-3- Sample Images:

An integral aspect of dataset visualization is the presentation of a collection of sample images. These sample images provide a visual representation of the dataset's content and offer an opportunity to identify any anomalies or potential mislabelling. To ensure randomness and diversity, 25 images, five from each class, are selected for display. This approach allows for a representative snapshot of the dataset's composition and aids in understanding the diversity of emotional expressions captured.

The Matplotlib library has been employed to create a 5x5 grid displaying 25 sample images. Each image is randomly selected from different emotional classes to provide a balanced view of the dataset.

Below is a 5x5 grid displaying 25 sample images, with each row representing a different emotional class. The images are randomly chosen, and the grid showcases a range of emotional expressions:

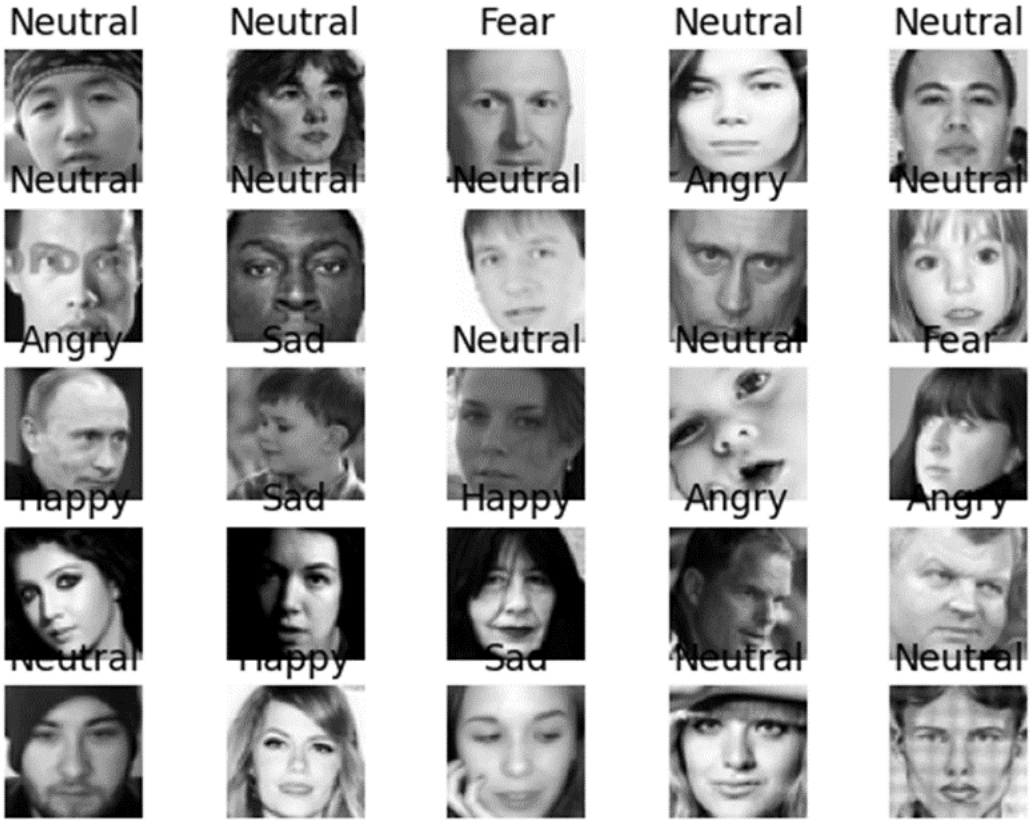

Figure 5: sample images dataset after cleaning

5- Conclusion:

In conclusion, Part 1 of the A.I.ducation Analytics project has successfully laid the groundwork for a transformative AI-driven educational feedback system. By meticulously curating, cleaning, and labeling a diverse dataset of students' facial expressions, we have ensured its quality and relevance to the project. The visualization and analysis of the dataset provided valuable insights, guiding us toward the development of a Deep Learning Convolutional Neural Network (CNN) for real-time emotion recognition. This dataset will play a pivotal role in the project's mission to enhance educational experiences through AI-driven insights and interventions.

REFERENCES

1- https://www.kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge/data