CHAPTER-3

Data Collection and Pre-processing

3.1 Introduction

Every day, millions of face images are uploaded to social media, with the majority undergoing retouching using photo editing tools. In the digital age, virtually everyone possesses a smartphone, and photo editing software is readily accessible and user-friendly, even for those with little technical expertise. There is currently a lack of a comprehensive database for categorizing retouching and evaluating model robustness because there has been little study done on the effects of digitally altered or edited facial photographs. Standard face datasets are essential, but they are incredibly hard to find in real-world applications. Regarding sample availability for both actual and retouched faces, data size, and the editing methods used, the datasets used in this analysis vary greatly. The data collection and preprocessing chapter is a critical component of any research project, analysis, or machine learning endeavor. It lays the foundation for the quality and reliability of the data used, which in turn affects the validity and robustness of the conclusions and models built upon it. This chapter contains the two distinct datasets used for the research purpose and the preprocessing employed over the face images before training and evaluating the model.

3.2 **Dataset 1**

A substantial face image database is compiled from the ND-IIITD Retouched Faces database[28]. The original images are sourced from the Notre Dame - Set B database[29], and corresponding retouched images are generated automatically using the Portrait Pro Max tool. For each subject, seven distinct images are captured under various conditions, encompassing different backgrounds, lighting conditions, poses, and gestures. All seven samples are subjected to retouching using a single photo editing tool, applying different retouching measures. These retouching procedures change the texture of the skin, the dimensions of the eyes, nose, lips, and the whole face as well as the prominence of the smile the shape of the lips, and the color of the eyes. The dataset comprises, as summarize in Table 3.1, both male and female subjects, with varying degrees of retouching based on gender. Hence, total human faces are 325 including both gender, where 2600 faces are real images of all objects and corresponding fake(retouched) face images are 2275. The samples are categorized into different sets corresponding to the extent of retouching applied. Thus, Set 1 includes samples with minimal retouching, while Set 7 represents the highest level of retouching applied to facial images. The heat map, shown in fig 3.1, as obtained the knowledge from reference[9], visually illustrates the variations in retouching across all these sets.

FIGURE 3.1: Showcase the difference between real and fake samples for analysing the percentage of retouching applied. 1st row contains real face images of one subject, 2nd row contains fake images of same object, 3rd row shows the intensity difference between real and fake images

TABLE 3.1: Description of Datasets used for Retouching Detection

Label	Dataset	Category	Real	Retouched	Total
Dataset 1	ND-IIITD Retouched Faces[28]	Male(211),Female(114)	2600	2275	4875
Dataset 2	MDRF Dataset[30]	Male(100),Female(100)	400	800	1200

3.3 **Dataset 2**

The MDRF (Multi Demographic Retouched Faces) dataset contains the genuine and retouched samples of Caucasian, Indian and Chinese subjects[10]. The research utilizes sample images of both genuine and retouched faces belonging to the Caucasian race only. This dataset encompasses facial images of 100 male and 100 female subjects, totalling 200 individuals, as shown in table 3.1. Within this dataset, there are 1,200 images, consisting of 400 authentic images and 800 artificially enhanced images, retouched using the Portrait Pro and BeautyPlus photo editing tools. Dataset 2, which focuses on the Caucasian demography, shares the same ethnicity as Dataset 1. However, it diverges from Dataset 1 in several aspects. Notably, Dataset 2 exhibits an imbalanced distribution of real and fake samples, akin to Dataset 1, and employs two widely used photo editing tools (as per the rating obtained in reference[17]) for image manipulation. Unlike Dataset 1, the gender distribution in Dataset 2 is balanced, with equal representation of male and female subjects. The dataset's skewed distribution is utilized to evaluate the proposed model's generalization across diverse datasets.

FIGURE 3.2: Sample images of dataset 2. 1st column real images, 2nd column fake images doctored by Portrait Pro Max tool, 3rd column fake images doctored by Beautyplus photo editing tool

3.4 Data Pre-processing

Data pre-processing is a crucial step in preparing data for machine learning or computer vision tasks. It involves a series of operations and transformations to clean, format, and enhance the quality of the data. This involves following processing tasks.

1. **Data Normalization/Standardization:** Scaling the features to have a similar scale can help improve the performance of many machine learning algorithms. Normalization scales the features to a specific range (usually between 0 and 1), while standardization makes the features have a mean of 0 and a standard deviation of 1.
2. **Data Augmentation:** In computer vision, data augmentation techniques are used to artificially expand the dataset by applying transformations like scaling, rotation, flipping, and cropping to the images[31]. This helps the model generalize better and reduces overfitting[32]. This research uses horizontal scaling and rotation by 0.2 to hypothetically increase the data size for analysis.
    - **Horizontal Scaling:** Horizontal scaling in image data preprocessing involves resizing images by changing their width or height while maintaining the same aspect ratio. This operation is commonly used to ensure that all images in a dataset have a consistent size, which is necessary for training deep learning models.
    - **Rotation:** Rotation is another important image preprocessing technique, which involves rotating an image by a certain angle. It's used to enhance dataset diversity and improve model robustness. Here's some detail on rotation

3. **Data Splitting:** The train-test split ratio, often known as the "train-test ratio" or "split ratio," is a crucial idea in machine learning and data analysis. A training set and a testing (or validation) set are the two subsets into which your dataset should be divided. Split the data into training, validation, and test sets to evaluate model performance and prevent overfitting[33]. Common train-test split ratios are 80%-20%, 70%-30%, 60%-40% and 50%-50% for classification task. The first number indicates the percentage allocated to the training set, while the second number represents the percentage assigned to the testing set. For instance, in an 80%-20% split, 80% of the data is used for training, and 20% is used for testing. Here, 20% is further divided into 10% for validation and 10% for test dataset, as shown in figure

3.3. The choice of the split ratio depends on various factors, including the size of the dataset, the nature of the problem, and the goals of the analysis. Smaller datasets may require larger testing sets to ensure an adequate evaluation, while larger datasets may allow for smaller testing sets. It's important to strike a balance when choosing the train-test split ratio. Too much data in the training set might lead to overfitting, while too little training data might lead to under fitting.

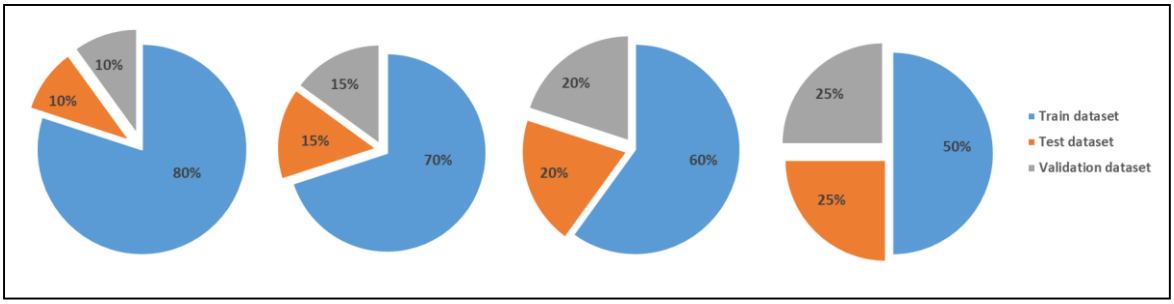

FIGURE 3.3: Train-Test Splitting defined for Facial Retouching Classification task

The right split ratio should allow the model to generalize well and provide a meaningful evaluation of its performance on unseen data[34]. Hence, this research endeavor aimed to determine the optimal split ratio for the proposed model, one that would enable the model to generalize effectively and deliver a substantial evaluation of its performance on unseen data. This investigation sought to identify the split ratio that yields the highest accuracy for both datasets. The table 3.2 and 3.3 displays facial images of Dataset 1and Dataset2 respectively categorized by various split ratios, along with the respective counts of real and fake images in each split. In each set, including the training, validation, and test datasets across all split ratios, an equal number of real and retouched (fake) samples are selected. The testing set specifically comprises samples from non-overlapping subjects, encompassing both male and female individuals. In other words, all these sets consist of authentic and retouched images from distinct and unique subjects.

TABLE 3.2: Train-Test Splitting for Dataset 1 for classification of Facial Retouching

Training-Testing Split Ratio	**Train**	**Validation**	**Test**
	No. of Images (Real + Retouched)	No. of Images (Real + Retouched)	No. of Images (Real + Retouched)
80%-20%	3624	462	460
70%-30%	3175	686	684
60%-40%	2713	910	920
50%-50%	2267	1133	1146

TABLE 3.3: Train-Test Splitting for Dataset 2 for classification of Facial Retouching

Training-Testing Split Ratio	Train	Validation	Test
	No. of Images (Real + Retouched)	No. of Images (Real + Retouched)	No. of Images (Real + Retouched)
80%-20%	960	120	120
70%-30%	840	180	180
60%-40%	720	240	240
50%-50%	600	300	300

3.5 **Summary**

After reviewing the previously published literature for the classification problem, special attention is paid to the selection of facial datasets. this chapter lays the foundation for subsequent analyses or machine learning tasks by ensuring that the data is standard, well-structured, and representative of the problem at hand. It creates a solid framework for understanding the characteristics of the data and prepares it for further investigation and model creation. In this chapter, the datasets' structure and properties are defined, and the key ideas of preprocessing and augmentation are introduced before being used to training.