Dataset Preparation and Preprocessing

1. Introduction

The objective of Milestone 1 is to prepare a clean, labelled, balanced, and preprocessed dataset for training a model that detects facial age-related skin features such as wrinkles, dark spots, puffy eyes, and clear skin.
This milestone ensures high-quality data for efficient model training in Milestone 2.

2. Dataset Setup

2.1 Folder Structure

```
project_root/
├── dataset/
├── cleaneddataset/
├── splittedataset/
│   ├── train/
│   └── validation/
```

dataset – Main directory containing all project image data, including the cleaned dataset and the train/validation split.

cleaneddataset – Contains cleaned and preprocessed images after removing low-quality files, duplicates, and incorrectly labelled examples. This acts as the master dataset before splitting.

splittedset – Contains the dataset split into train/ and validation/ folders for model training and performance evaluation. Generated from the cleaned dataset.

notebooks – Contains Jupyter notebooks for data analysis, visualization, preprocessing, and model experimentation.

2.2 Class Definitions

Images were categorized into the following four classes:

- Clear Skin
- Wrinkles
- Dark Spots
- Puffy Eyes

Ambiguous or low-quality images were removed to maintain dataset purity.

3. Dataset Inspection and Class Distribution

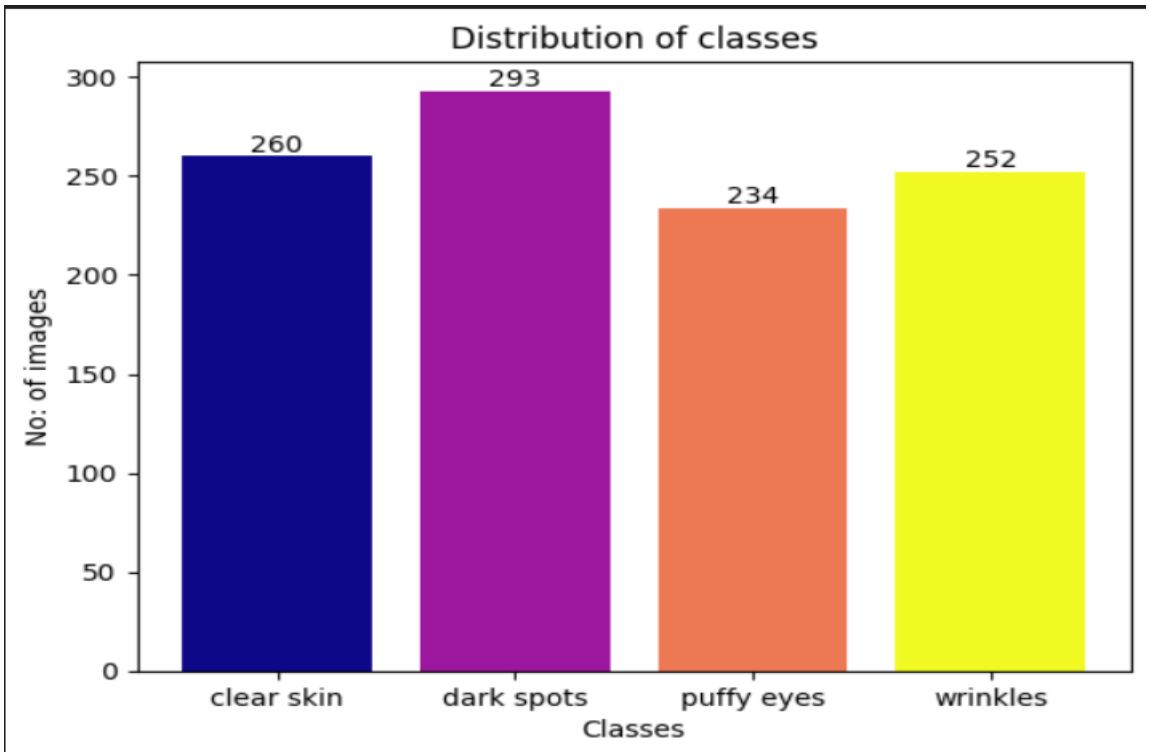
A short script was used to verify:

- Images are correctly labelled
- Samples display properly

```
Import os

for c in classes:

    clfolder = os.path.join(datasetpath,c)

    imagename = os.listdir(clfolder)[0]

    imgagepath = os.path.join(clfolder,imagename)

    img = cv2.imread(imgagepath)
```

3.1 Class Distribution Plot

To check dataset balance, the number of images per class was counted and plotted.



4. Train–Validation Split

The dataset was divided using an 80:20 split, where 80% of the images were used for training and 20% for validation. The training set contains four classes: clear skin, wrinkles, dark spots, and puffy eyes.
The validation set contains the same four classes and is used to evaluate the model on unseen data.

This splitting strategy ensures that the model is tested on images it has not encountered during training, improving its ability to generalize.

5. Image Preprocessing

5.1 Resize and Normalize

All images were resized to **224 × 224 pixels** for model compatibility. Pixel values were normalized by dividing each value by **255**, ensuring all inputs fall within the 0–1 range.

5.2 One-Hot Encoding

Labels were automatically converted into one hot encoding using class_mode="categorical",

Allowing the model to handle multi class classification.

6. Data Augmentation Pipeline

Augmentation was applied only to training images to increase diversity while preserving class features. These transformations simulate natural changes that occur in facial photos without altering the underlying age-related features.

Transformations Used

- **Rotation (±20°)** – Simulates slight head tilts, helping the model remain accurate even when faces are not perfectly aligned.
- **Zoom (up to 20%)** – Mimics variations in camera distance, ensuring the model can detect age features such as wrinkles or dark spots even when the face appears closer or farther away.
- **Horizontal/Vertical Shifts (10%)** – Represents small changes in face positioning within the frame, preventing the model from becoming sensitive to fixed face alignment.
- **Horizontal Flip** – Helps the model generalize to left/right orientation of facial features, since aging signs appear symmetrically.
- **Brightness Variation** – Accounts for different lighting conditions, ensuring age-related features are recognized in bright or dim environments.
- **Fill mode: Nearest** – Used to fill in newly created pixels during transformations while preserving important facial details.

These augmentations improve robustness by teaching the model to focus on **age-indicative patterns** rather than on fixed image positions, lighting, or orientation.

```
traingenerator = ImageDataGenerator(
    rescale=1./255,
    rotation_range=20,
    zoom_range=0.2,
    width_shift_range=0.1,
    height_shift_range=0.1,
    horizontal_flip=True,
    brightness_range=(0.9, 1.1),
    fill_mode="nearest")
```

7. Visualization of Output

7.1 Original Images

These images represent some sample images from the dataset before augmentation.

7.2 Augmented Images

These images demonstrate how augmentation transforms some sample images from the dataset.

8. Conclusion

Milestone 1 successfully produced a clean, structured, and enhanced dataset ready for model development. All required inspection, preprocessing, and augmentation tasks were completed. This forms the foundation for next module.