Title: Image Data Preprocessing and Splitting for Melanoma Detection Tasks

Description:

This Python script is designed to preprocess and split image data for two melanoma detection tasks (Task 1 and Task 2). It reads image data and corresponding ground truth labels from CSV files, resizes the images, and organizes them into separate directories based on the task and label type (melanoma, nevus, seborrheic keratosis).

Requirements:

1. Python 3.x

2. pandas

3. NumPy

4. Pillow (PIL)

Usage:

1. Ensure that the required Python libraries are installed.

2. Modify the paths for the data and ground truth CSV files according to the specific dataset location.

3. Execute the script in a Python environment.

Functions:

1. `resize\_images\_in\_folder(input\_folder, output\_folder, new\_size=(224, 224))`:

Description: Resizes images in the input folder and saves them in the output folder with the specified new size.

Parameters:

- `input\_folder` (str): Path to the folder containing the input images.

- `output\_folder` (str): Path to the folder where resized images will be saved.

- `new\_size` (tuple): Tuple representing the new size in the format (width, height). Default is (224, 224).

2. `task1\_spliter(df, data\_dir, final\_dir)`:

Description: Splits data for Task 1 (Melanoma vs. Nevus and Seborrheic Keratosis) into separate folders based on labels.

Parameters:

- `df` (Pandas DataFrame): DataFrame containing the ground truth information for Task 1.

- `data\_dir` (str): Path to the directory containing the original images.

- `final\_dir` (str): Path to the directory where the Task 1 data will be organized.

3. `task2\_spliter(df, data\_dir, final\_dir)`:

Description: Splits data for Task 2 (Seborrheic Keratosis vs. Nevus and Melanoma) into separate folders based on labels.

Parameters:

- `df` (Pandas DataFrame): DataFrame containing the ground truth information for Task 2.

- `data\_dir` (str): Path to the directory containing the original images.

- `final\_dir` (str): Path to the directory where the Task 2 data will be organized.

4. `main()`:

Description: Main function to execute the preprocessing and data splitting tasks. It initializes the paths for train, test, and validation datasets, resizes images, and organizes them into separate directories for each task type.

Execution Steps:

1. The script reads the ground truth information from CSV files using Pandas DataFrames.

2. The `resize\_images\_in\_folder()` function is applied to resize images in the train, test, and validation directories.

3. The `task1\_spliter()` and `task2\_spliter()` functions are used to split data for Task 1 and Task 2, respectively, into separate folders based on the labels.

4. The script creates new directories for each task type (melanoma, nevus, seborrheic keratosis) in the train, test, and validation directories.

5. It copies the images into the corresponding task-specific directories.

6. The script merges the training and validation data for cross-validation purposes into a new directory named "train\_merged".

Note: The script assumes that the image files are stored in the specified paths and follow the ".jpg" or ".jpeg" format. Any other file format will be skipped during the resizing and copying process.

For Data augmentation, the images are generated and saved to disk. Images generated for each task is 6 times their original size as suggested by Gayathri.