**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| **Date** | 14 June 2025 |
| **Team ID** | SWTID1749876754 |
| **Project Title** | SynapseScan- AI Driven Classification of Ovarian Cancer Variants |
| **Maximum Marks** | 6 Marks |

**Data Exploration and Preprocessing Report**

Dataset variables will be statistically analyzed to identify patterns and outliers in ovarian cancer medical imaging data, with Python employed for preprocessing tasks like image normalization, data augmentation, and feature engineering. Data cleaning will address missing values and image quality issues, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for accurate cancer variant classification and predictions.

|  |  |
| --- | --- |
| **Section** | **Description** |
| **Data Overview** | Medical imaging dataset containing histopathological images of ovarian cancer variants  Both training set and test set were divided into 5 subfolders: CC, EC, HGSC, LGSC, MC, which are the 5 main types of ovarian cancer.  Each subfolder in training set had between 3000-12000 images.  Each subfolder in test set had about 800 images. |

**Data Preprocessing Code Screenshots**

|  |  |
| --- | --- |
| **Preprocessing Step** | **Implementation** |
| **Loading Data** |  |
| **Handling Missing Data** |  |
| **Data Transformation** |  |
| **Feature Engineering** |  |
| **Save Processed Data** |  |