**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 11 November 2024 |
| Team ID | 739761 |
| Project Title | Pixelprose - crafting visual stories with intelligent image captioning |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Dataset | Presence of duplicate images | Moderate | Identify and remove duplicates to ensure each image is unique and to prevent model bias. |
| Dataset | Imbalanced classes (some diseases may have fewer samples) | High | Apply data augmentation on underrepresented classes or oversample them to balance the dataset. |
| Dataset | Low-resolution or blurry images | |  | | --- | | High |  |  | | --- | |  | | Use image enhancement techniques or remove extremely low-quality images if they hinder model accuracy. |
| Dataset | Variability in lighting conditions | Moderate | |  | | --- | | Normalize lighting by applying histogram equalization or similar techniques to standardize image input. |  |  | | --- | |  | |
| Dataset | Inconsistent image dimensions | Low | Resize all images to a standard size (e.g., 224x224 pixels) suitable for the chosen CNN architecture. |