Data Processing Requirements (DPR) Document

Retinal Image Disease Detection System

Table of Contents

1. Introduction

The Data Processing Requirements (DPR) document outlines the data processing pipeline and requirements for the retinal image disease detection project. It describes the steps involved in data ingestion, validation, transformation, and storage to ensure reliable and accurate predictions using the pretrained ResNet50 model and Kaggle dataset.

2. Data Ingestion

- The system should support data ingestion from the Kaggle dataset, including loading image files and associated metadata.
- The data ingestion process should include mechanisms for validating the data integrity, consistency, and format.
- Error handling and logging should be implemented to track any issues or anomalies during the data ingestion phase.

3. Data Validation

- The system performs data validation checks to ensure the correctness and quality of input data, including image files and labels.
- Common validation checks include verifying image file formats, checking for missing or corrupted images, and ensuring that labels correspond to images correctly.

4. Data Transformation

- Data transformation prepares the data for model training and prediction.
- This involves image preprocessing steps such as resizing, normalization, and augmentation to improve model performance and generalization.
- Data augmentation techniques may be employed to increase dataset diversity and mitigate overfitting.
- Ensure that image data is properly split into training, validation, and test sets.

5. Data Storage

- The system uses appropriate storage solutions for image data and model outputs, which could include local file storage or cloud-based solutions depending on project requirements.

6. Conclusion

The Data Processing Requirements (DPR) document outlines the key considerations and steps involved in data processing for the retinal image disease detection project. It covers data ingestion, validation, transformation, storage, and monitoring aspects. This document serves as a reference for the development team to ensure effective and reliable data processing within the system.