Project Planning Phase

Planning Logic (Data Collection, Data Cleaning, and Exploratory Data Analysis)

Date	**15 February 2026**
Team ID	**LTVIP2026TMIDS81330**
Project Name	**Deep Learning Fundus Image Analysis for Early Detection of Diabetic Retinopathy**
Maximum Marks	**5 Marks**

1. Introduction

The Project Planning Phase focuses on preparing the dataset and defining the logical steps required before model development. In a medical AI system like Diabetic Retinopathy Detection, data quality directly impacts prediction accuracy. Therefore, structured planning was performed in three major stages:

1. Data Collection
2. Data Cleaning
3. Exploratory Data Analysis (EDA)

This phase ensures that the dataset is reliable, balanced, and medically meaningful before training the deep learning model.

2. Data Collection

2.1 Source of Dataset

The retinal fundus image dataset was collected from publicly available medical image repositories such as:

- Kaggle Diabetic Retinopathy Dataset
- Hospital retinal scan datasets
- Public ophthalmology research datasets

The dataset contains retinal fundus images categorized into five classes:

1. No Diabetic Retinopathy
2. Mild DR
3. Moderate DR
4. Severe DR
5. Proliferative DR

2.2 Dataset Structure

The dataset is organized into folder-based classification:

```
dataset/
├── train/
│   ├── No_DR/
│   ├── Mild/
│   ├── Moderate/
│   ├── Severe/
│   └── PDR/
├── test/
└── validation/
```

Each folder contains retinal images corresponding to specific DR stages.

2.3 Challenges During Data Collection

- Large variation in image brightness
- Different image resolutions
- Class imbalance (more Normal images than Severe cases)
- Noisy and blurred images

These challenges required further preprocessing.

3. Data Cleaning

Data cleaning is essential in medical image processing to remove irrelevant or corrupted samples.

3.1 Removal of Corrupted Images

- Checked unreadable images
- Removed blurred or blank images
- Removed duplicate samples

3.2 Image Resizing

All images were resized to:

299 × 299 pixels

This size matches the Xception model input requirement.

3.3 Normalization

Pixel values were normalized using:

preprocess_input() from Xception

This ensures:

- Faster convergence
- Stable training
- Better model performance

3.4 Data Augmentation

To avoid overfitting and improve generalization:

- Rotation
- Horizontal flipping
- Zooming
- Brightness adjustment

This increases dataset diversity without collecting new data.

4. Exploratory Data Analysis (EDA)

EDA helps understand dataset distribution and identify patterns.

4.1 Class Distribution Analysis

Observed imbalance:

- Normal class had highest samples
- Severe and PDR had fewer samples

Solution:

- Applied augmentation on minority classes
- Used balanced batch generation

4.2 Image Visualization

Random samples from each class were visualized to understand:

- Blood vessel abnormalities
- Cotton wool spots
- Microaneurysms
- Hemorrhages

4.3 Statistical Insights

- Mean pixel intensity distribution analyzed
- Histogram analysis of brightness levels
- Identified lighting variation patterns

5. Conclusion of Planning Logic

The planning logic ensured:

- Clean and structured dataset
- Balanced class distribution
- Model-ready formatted images
- Reduced noise and inconsistencies

This structured preparation directly improved final model accuracy.