

Dataset Description Reference — NATURE Fundus Dataset

Overview

This dataset is a retinal fundus image dataset collected from ophthalmology clinical records. It contains colour fundus photographs paired with pixel-level segmentation masks for multiple retinal pathology and anatomy classes. The dataset is intended for supervised learning tasks such as multi-label semantic segmentation of retinal structures and lesions.

Dataset Statistics

Item	Count
Total fundus images	500
Total patients	230
Total clinical records	230
Images with at least one mask	280 (see Mask Availability)

Directory Structure

```

NATURE/
├── FUNDUS/                # Raw colour fundus photographs (500 × .JPG)
├── MASK/                  # Segmentation masks, one sub-folder per class
│   ├── A_Mask/           # Artery masks (100 images)
│   ├── V_Mask/           # Vein masks (100 images)
│   ├── DB_Mask/          # Dot and Blot Hemorrhage masks (46 images)
│   ├── FP_Mask/          # Fibrous Proliferation masks (10 images)
│   ├── FS_Mask/          # Flame Shaped Hemorrhage masks (14 images)
│   ├── HE_Mask/          # Hard Exudate masks (60 images)
│   ├── MA_Mask/          # Microaneurysm masks (60 images)
│   ├── PRP_Mask/         # Panretinal Photocoagulation masks (10 images)
│   └── SE_Mask/          # Soft Exudate masks (60 images)
└── MASTER_EXCELS/
    ├── METADATA.xlsx      # PRIMARY LABEL FILE – patient demographics &
    # comorbidities per record
    └── Mask_Availability.xlsx # Per-image mask presence flags

```

Imaging Data

Fundus Images (FUNDUS/)

- **Format:** JPEG (.JPG)
- **Naming convention:** `ophth_<IMAGE_ID>_<RECORD_ID>_<EYE>_FUNDUS.JPG`
 - `IMAGE_ID` — zero-padded sequential image number (e.g., 001)
 - `RECORD_ID` — zero-padded clinical record number (e.g., 001)
 - `EYE` — laterality: `L` (left eye) or `R` (right eye)
- **Example:** `ophth_001_001_R_FUNDUS.JPG`

Segmentation Masks (MASK/)

- **Format:** PNG (.png), binary or class-coloured masks
- **Naming convention:** `ophth_<IMAGE_ID>_<RECORD_ID>_<CLASS>.png`
 - `CLASS` — abbreviation matching the sub-folder (see table below)
- **Example:** `ophth_001_001_A.png` (artery mask for image 001)

Mask Classes

Sub-folder	Class Code	Description	# Masks
------------	------------	-------------	---------

A_Mask	A	Arteries — retinal arterial vasculature	100
V_Mask	V	Veins — retinal venous vasculature	100
DB_Mask	DB	Dot and Blot Hemorrhage — Intraretinal hemorrhage	46
FP_Mask	FP	Fibrous Proliferation — Abnormal retinal fibrous tissue growth	10
FS_Mask	FS	Flame Shaped Hemorrhage — Superficial hemorrhage	14
HE_Mask	HE	Hard Exudates — lipid deposits	60
MA_Mask	MA	Microaneurysms — small vascular bulges, early DR sign	60
PRP_Mask	PRP	Pan-Retinal Photocoagulation — laser treatment scars	10
SE_Mask	SE	Soft Exudates/cotton-wool spots — nerve fiber layer infarcts	60

Note: Not every image has a mask for every class. The `Mask_Availability.xlsx` file is the authoritative reference for which classes are annotated for each image.

Metadata Files (MASTER_EXCELS /)

METADATA.xlsx ← Primary Label File

This is the **central label file** for the dataset. Each row corresponds to one clinical record (patient visit) and provides the patient identifier, demographics, and systemic comorbidity labels. All supervised learning tasks should join other files to this file via `RECORD ID`.

Column	Type	Description
RECORD ID	Integer	Clinical record identifier — unique per row; links to <code>Mask_Availability.xlsx</code> and to the <code>ophth_*</code> image filenames
AGE	Float	Patient age in years at time of record
GENDER	Categorical	Male / Female
DIABETES MELLITUS (DM)	Categorical	Yes / No — diagnosis of diabetes mellitus
HYPERTENSION	Categorical	Yes / No / N/A — diagnosis of hypertension

CORONARY ARTERY DISEASE	Categorical	Yes / No / N/A — diagnosis of coronary artery disease
STROKE	Categorical	Yes / No / N/A — history of stroke

- **N/A values:** Present in HYPERTENSION , CORONARY ARTERY DISEASE , and STROKE for a subset of records where comorbidity data was not collected
- **Join key:** RECORD ID links to the RECORD_ID component of the oph_<IMAGE_ID>_<RECORD_ID> unique image ID

Mask_Availability.xlsx

Indicates which annotation classes are available for each unique image.

Column	Description
UNIQUE_ID	Unique image identifier (e.g., oph_001_001)
A	Artery mask present: Yes / No
V	Vein mask present: Yes / No
DB	Dot and Blot Hemorrhage mask present: Yes / No
FP	Fibrous Proliferation mask present: Yes / No
FS	Flame Shaped Hemorrhage mask present: Yes / No
HE	Hard exudate mask present: Yes / No
MA	Microaneurysm mask present: Yes / No
PRP	Pan-Retinal Photocoagulation mask present: Yes / No
SE	Soft Exudates mask present: Yes / No

Unique ID Convention

All images and masks share the common identifier format:

```
oph_<IMAGE_ID>_<RECORD_ID>
```

- **oph** — dataset prefix (ophthalmology)

- **IMAGE_ID** — three-digit zero-padded sequential index across the full image set
- **RECORD_ID** — three-digit zero-padded index linking to Patient RECORD ID Mapping

This ID is the join key across all EXCELS, fundus images, and mask files.

Notes for Data Description Writing

- The dataset is **multi-label**: a single image can have masks from multiple classes simultaneously.
- Mask counts per class are **unequal** — some pathologies (FP, PRP) are rare and have very few annotated examples.
- Artery (**A**) and Vein (**V**) masks are the most prevalent, each covering 100 images (20% of the full set).
- The patient-record relationship is **many-to-one**: one patient may contribute multiple records (and therefore multiple images), so image count > patient count.