# 📘 MASTER PROJECT DOCUMENT

## Multimodal Skin Lesion Classification Using Image and Clinical Metadata

---

### এই document-টা কী?

তোমার প্রজেক্টের সবগুলো ফাইল — `PROJECT_PLAN.md`, `Project_Tracking.md`, `Dataset_Strategy.md`, `AI_Assistant_Instructions.md`, `PROJECT_OWNERSHIP.md` — এই সবগুলো একসাথে মিলিয়ে বানানো **একটাই single reference।**

এখানে যা পাবে:

- প্রজেক্টের উদ্দেশ্য
- এখন পর্যন্ত কী কী হয়েছে (ownership নেওয়ার জন্য বিস্তারিত)
- কী বাকি আছে
- সামনের প্রতিটা ধাপে কী করতে হবে
- কোন কোন skill লাগবে

🔑 চিহ্নিত অংশগুলো **সবচেয়ে গুরুত্বপূর্ণ/টেকনিক্যাল** — এগুলোই supervisor বা viva-তে সবচেয়ে বেশি জিজ্ঞেস করা হবে।

🗓️ **সর্বশেষ আপডেট:** 2026-07-09

---

## 📑 সূচিপত্র


| | Section |
|---|---|
| 🎯 | Status Dashboard |
| A | প্রজেক্টের পরিচয় |
| B | Folder/File Ownership |
| C | কালানুক্রমিক ইতিহাস (Phase 1-5) |
| D | 🔑 Leakage Audit — সবচেয়ে গুরুত্বপূর্ণ আবিষ্কার |
| E | বর্তমান Dataset অবস্থা |
| F | এখনো যা বাকি |

| | Section |
|---|---|
| G | সামনের Roadmap (Phase 6-10) |
| H | প্রয়োজনীয় Skill ও জ্ঞান |
| I | নিজে চালিয়ে দেখার Command |
| J | Supervisor/Viva Talking Points |


---

---

<a id="status-dashboard"></a>

# 🎯 Status Dashboard

## ১৫ সেকেন্ডে পুরো প্রজেক্টের ছবি

| | | | |
|---|---|---|---|
| ✅ | Phase 1 | Planning | সম্পূর্ণ |
| 🟡 | Phase 2 | Literature Review | চলমান (৩টা paper) |
| ✅ | Phase 3 | Dataset Collection | সম্পূর্ণ (৪টা dataset) |
| ✅ | Phase 4 | Dataset Preparation | সম্পূর্ণ |
| ✅ | Phase 5 | EDA | সম্পূর্ণ |
| 🟡 | Phase 6 | Baseline Model | কোড শেষ, training বাকি |
| ☐ | Phase 7 | Multimodal Fusion | শুরু হয়নি |
| ☐ | Phase 8 | Experiments & Evaluation | শুরু হয়নি |
| ☐ | Phase 9 | Thesis Writing | শুরু হয়নি |
| ☐ | Phase 10 | Paper Submission | শুরু হয়নি |

<br>

| ❓ প্রশ্ন | ✅ উত্তর |
|---|---|
| Dataset কি "ready"? | **হ্যাঁ** — সম্পূর্ণ clean, leakage-free, documented |
| Model কি "trained"? | **না** — কোড আছে, কিন্তু একটাও training run সম্পূর্ণ হয়নি |
| এখনকার blocker কী? | GPU নাই → Kaggle-এ shift করার প্রস্তুতি চলছে |
| সবচেয়ে বড় technical অর্জন | 🔑 ২২টা leakage/shortcut column খুঁজে বাদ দেওয়া |
| পরের কাজ কী | Kaggle-এ dataset mount → ৬টা baseline training run |

---

---

<a id="part-a"></a>

# Part A — প্রজেক্টের পরিচয়

### 🏷️ Research Title

**Multimodal Skin Disease Classification System Using Skin Images and Clinical Metadata**

<br>

## 🎯 Research Goal — ৪টা লক্ষ্য

1. একটা মানসম্পন্ন Master's thesis সম্পূর্ণ করা
2. বৈজ্ঞানিকভাবে অর্থপূর্ণ একটা multimodal AI system বানানো
3. একটা উপযুক্ত conference/journal-এ publish করা
4. Higher studies/scholarship-এর জন্য শক্তিশালী research profile তৈরি করা

<br>

## 🔑 Research Problem — আমরা কোন gap পূরণ করছি

বেশিরভাগ existing skin disease classification শুধু **ছবি-ভিত্তিক** (image-only)।

### এদের দুর্বলতা:

- Clinical তথ্য ব্যবহার না করা
- Patient metadata কম ব্যবহার করা
- বাইরের dataset-এ দুর্বল generalization
- কম explainability
- Dataset bias

### আমাদের সমাধান:

ছবি + clinical metadata + patient info + symptoms একসাথে ব্যবহার করে, **multi-class** (শুধু বাইনারি benign/malignant না) target করে, আর **cross-dataset generalization + leakage-audit rigor** যোগ করে — যেটা এখন পর্যন্ত reviewed কোনো paper একসাথে করেনি।

<br>

## 🧭 Working Philosophy (কাজের দর্শন)

- প্রতিটা সিদ্ধান্তে ব্যাখ্যা থাকা আবশ্যক (কেন, বৈজ্ঞানিক যুক্তি, সীমাবদ্ধতা, বিকল্প)
- Reproducibility বজায় রাখা
- Raw data কখনো পরিবর্তন না করা
- প্রতিটা পরিবর্তনের log রাখা
- শুধু accuracy-এর জন্য optimize না করা

<br>

### 📊 ৪টা Dataset-এর ভূমিকা

| Dataset | ভূমিকা | কী আছে |
|---|---|---|
| 🔑 **PAD-UFES-20** | **Primary/মূল dataset** | ছবি + rich clinical metadata + patient info + symptoms |
| **HAM10000** | Benchmark dataset | Dermoscopic ছবি + সীমিত metadata |
| **ISIC Archive 1** | External validation | কোনো metadata নাই, শুধু ছবি |
| **ISIC Archive 2** | External validation (২য়) | ছবি + rich metadata, বেশিরভাগ column leakage-risk বলে বাদ |

---

---

<a id="part-b"></a>

# Part B — Folder/File Ownership

## সব কোথায় আছে

```
Multimodal_Skin_Disease_Research/
│
├── README.md                          প্রজেক্টের প্রথম পাতা
├── requirements.txt                    প্রয়োজনীয় Python লাইব্রেরি
│
├── data/
│   ├── raw/              🔒 IMMUTABLE — কোনো script কখনো এখানে লিখে না
│   │   └── PAD_UFES20/, HAM10000/, ISIC_Archive_1/, ISIC_Archive_2/
│   │
│   ├── interim/          পরিষ্কার করার মাঝামাঝি ধাপের ফাইল
│   │
│   └── processed/        ✅ চূড়ান্ত, model-ready CSV
│       └── <প্রতিটা dataset>/
│             metadata_{train,val,test}.csv
│             label_mapping.csv
│             split_quality_report.csv
│             dataset_description.md
│             feature_whitelist.md
│
├── docs/                 🔑 সব সিদ্ধান্ত ও ব্যাখ্যা এখানে
│   ├── PROJECT_PLAN.md                 ক্যানোনিক্যাল প্ল্যান
│   ├── Project_Tracking.md             চলমান status + decision log
│   ├── Dataset_Strategy.md             dataset-এর ভূমিকা
│   ├── AI_Assistant_Instructions.md    মূল working directive
│   ├── Dataset_Preparation_Final_Report.md   cross-dataset verification
│   └── PROJECT_OWNERSHIP.md            file-by-file বিস্তারিত ব্যাখ্যা
│
├── src/
```

```
│   ├── data_audit/        Phase 4a — যাচাই কোড
│   ├── data_cleaning/     Phase 4b — পরিষ্কার, label, split কোড
│   ├── eda/                Phase 5 — বিশ্লেষণ কোড
│   ├── models/             Phase 6 — কোড লেখা শেষ, training বাকি
│   └── evaluation/         Phase 6 — evaluation কোড
│
└── notebooks/, reports/, logs/, papers/, _archive/
```
<br>

### 📄 কোন file কী কাজ করে

| File | কী বলে |
|---|---|
| 🔑 `PROJECT_PLAN.md` | সব confirmed সিদ্ধান্ত — এটাই "আইন" |
| 🔑 `Project_Tracking.md` | সবচেয়ে live/current ফাইল — "Session Handoff" দিয়ে শুরু |
| `Dataset_Strategy.md` | Dataset-এর role আর processing philosophy |
| `AI_Assistant_Instructions.md` | মূল কাজের নিয়মকানুন |
| `Dataset_Preparation_Final_Report.md` | Independent verification (overlap এখানে প্রথম ধরা পড়ে) |
| `PROJECT_OWNERSHIP.md` | প্রতিটা script/file-এর line-by-line ব্যাখ্যা |

---

---

<a id="part-c"></a>

# Part C — কালানুক্রমিকভাবে এখন পর্যন্ত কী হয়েছে

<br>

## ✅ Phase 1 — Planning

**সম্পূর্ণ · 2026-06-29**

- Thesis-এর scope ঠিক করা হয়েছে (multi-class, binary না)
- নিয়মকানুন লেখা হয়েছে (raw data read-only, missing value কখনো বানানো যাবে না, patient-wise split)
- ৩টা related paper review করে gap বের করা হয়েছে

<br>

# 🟡 Phase 2 — Literature Review

### চলমান

৩টা মূল paper পড়া হয়েছে:

- Mridha & Islam (2026)
- Suresh et al. (2026) — TG-CAVNet
- Watson et al. (2026)

🔑 **Watson et al.-এর সতর্কতা:** "diagnosis হওয়ার পরে তৈরি হওয়া field ব্যবহার করলে leakage হয়" — এটাই পরবর্তী পুরো leakage-audit কাজের মূল অনুপ্রেরণা।

<br>

# ✅ Phase 3 — Dataset Collection

### সম্পূর্ণ

৪টা dataset `data/raw/`-এ সংগ্রহ করা হয়েছে।

<br>

# ✅ Phase 4 — Dataset Preparation

**সম্পূর্ণ · 2026-07-07 → 2026-07-08**

### 4a. Audit (যাচাই) — প্রতিটা dataset আলাদা করে

- সব ছবি খোলে কিনা, নষ্ট কিনা যাচাই — **৪টাতেই ০টা corrupted ছবি**

🔑 **ISIC Archive 1-এ আসল data-quality সমস্যা ধরা পড়ে:** ১৫৫টা ছবি দুইটা ভিন্ন class-এ (label) একসাথে ফাইল করা ছিল (৭৮টা melanoma↔seborrheic keratosis, ৭৭টা actinic keratosis↔nevus) — আন্দাজে কোনটা সঠিক না ধরে, এই ছবিগুলো বাদ দেওয়া হয়েছে।

🔑 **PAD-UFES-20-এ ১৭৯ জন রোগীর একাধিক রোগ ধরা পড়ে** — তাই split করার সময় প্রতি রোগীর "dominant" (সবচেয়ে বেশি) রোগ ধরে stratify করা হয়েছে।

- HAM10000-এ কোনো patient identifier নাই — তাই `lesion_id` ধরে গ্রুপ করতে হয়েছে

- ISIC Archive 2-তে `patient_id` মাত্র ২% row-এ আছে — তাই `lesion_id` (fallback হিসেবে `image_id`) ব্যবহার করা হয়েছে

<br>

### 4b. Cleaning (পরিষ্কার)

- Column নাম একরকম করা (`gender`→`sex`, `region`→`anatomical_site`)

🔑 **লুকানো missing value ধরা:** PAD-UFES-20-এ `UNK` স্ট্রিং এবং HAM10000-এ `unknown` স্ট্রিং আসলে "তথ্য নাই" বোঝাচ্ছিল, কিন্তু সাধারণ `isna()` চেক এটা ধরতে পারেনি — আলাদা করে খুঁজে ঠিক করা হয়েছে।

- রোগের নাম একরকম করা (label mapping, প্রতিটার reason সহ)
- **Patient/lesion-wise split** (৭০/১৫/১৫), seed=42, প্রতিটাতে leakage শূন্য বলে verify করা হয়েছে

<br>

### 4c. 🔑 Cross-dataset overlap আবিষ্কার

**2026-07-08**

সব dataset আলাদাভাবে ঠিক হওয়ার পর, আরেকবার সব মিলিয়ে verify করার সময় একটা বড় জিনিস ধরা পড়ে:

| জোড়া | Overlap |
|---|---|
| HAM10000 ↔ ISIC Archive 2 | **৯৮.৬%** (প্রায় একই ছবি!) |
| ISIC Archive 1 ↔ ISIC Archive 2 | ৮১.৭% |
| HAM10000 ↔ ISIC Archive 1 | ৬৬.৫% |
| **PAD-UFES-20 ↔ যেকোনো অন্য dataset** | **0% (কোনো overlap নাই)** |

✅ **সমাধান (fix b, global re-split-এর বদলে):** প্রতিটা dataset-এর নিজস্ব split অক্ষত রাখা হয়েছে, কিন্তু ISIC Archive-কে HAM10000-trained model-এর "external validation" হিসেবে ব্যবহার করার সময় overlapping ছবিগুলো বাদ দিতে হবে — এর জন্য `external_validation_exclusions.csv` বানানো হয়েছে:

| Archive | মোট | বাদ | ব্যবহারযোগ্য থাকে |
|---|---|---|---|
| ISIC Archive 1 | ১,৩৬২টা | ৬৮৫টা | |
| ISIC Archive 2 | ৯,৮৭৩টা | ১৫,২০৩টা | |

<br>

### 4d. Feature Whitelist তৈরি

বিস্তারিত → [Part D দেখো]

<br>

### 4e. Documentation cleanup

**2026-07-08**

পুরনো/conflict-করা doc (`Research_Plan.md`, `Project_Cleanup_Review.md`) `_archive/`-এ সরানো হয়েছে (মুছে ফেলা হয়নি), `Dataset_Strategy.md` in-place সংশোধন করা হয়েছে।

<br>

## ✅ Phase 5 — EDA

**সম্পূর্ণ · 2026-07-09**

প্রতিটা dataset-এর জন্য বানানো হয়েছে:

- Class distribution
- বয়স/sex/anatomical site distribution
- Fitzpatrick distribution (PAD-UFES-20)
- Missing-value chart
- Sample image grid
- **Image dimension distribution** (নতুন সংযোজন)

🔑 এই শেষেরটাই পরে ISIC Archive 2-এর bimodal 600×450/1024×1024 cluster ধরতে সাহায্য করে, যেটা আবার `attribution` column-এর সাথে হুবহু মিলে যায় — cross-dataset overlap-এর আরেকটা independent প্রমাণ।

---

---

<a id="part-d"></a>

# Part D — 🔑 সবচেয়ে গুরুত্বপূর্ণ টেকনিক্যাল আবিষ্কার

# Leakage / Shortcut Audit

### এটাই এই প্রজেক্টের সবচেয়ে বড় টেকনিক্যাল contribution এই পর্যায়ে — supervisor/viva-তে সবচেয়ে বিস্তারিত বলার মতো অংশ।

<br>

## সমস্যাটা কী? (সহজ ভাষায়)

Data-তে এমন কিছু column থাকতে পারে যেগুলো রোগের নামের সাথে **প্রায় সরাসরি মিলে যায়** — model তখন ছবি না দেখেই শুধু সেই column পড়ে "উত্তর" বলে দিতে পারে।

⚠️ এটা ধরা না পড়লে result fake ভাবে ভালো দেখাবে, কিন্তু বাস্তবে model কিছুই শেখেনি।

<br>

## যা যা ধরা পড়েছে

*(প্রতিটা real সংখ্যা দিয়ে verified, অনুমান না)*

| Dataset | Column | সমস্যা | প্রমাণ (সংখ্যা) |
|---|---|---|---|
| PAD-UFES-20 | `biopsed` | ১০০% malignant case-এ biopsy হয়েছিল, ব্যতিক্রম নাই | **Phi = 0.80**, chi²=1474.5, n=2,298 |
| PAD-UFES-20 | `diagnostic_code` | Label যেখান থেকে বানানো, ১:১ মিল | সরাসরি label-এর উৎস |
| HAM10000 | `diagnosis_confirm_type` | সব malignant case histopathology দিয়ে confirm | **Phi = 0.41**, chi²=1700.67, n=10,015 |
| HAM10000 | `diagnostic_code` | Label-এর উৎস, ১:১ মিল | ৭/৭ code unambiguous |
| ISIC Archive 1 | `class_label` | Label-এর সরাসরি rename | ৯/৯ class unambiguous |
| ISIC Archive 2 | `diagnosis_confirm_type` | Malignant case কখনো "serial imaging"-এ confirm হয়নি | **Phi = 0.36**, chi²=3171.74, n=25,076 |
| ISIC Archive 2 | `diagnosis_1/2/3/4/5` | পুরো diagnostic hierarchy, label এখান থেকেই বানানো | `diagnosis_3` = ১:১ label |
| ISIC Archive 2 | `concomitant_biopsy` | `diagnosis_confirm_type`-এর সাথে হুবহু (duplicate সংকেত) | একই Phi = 0.36 |

| Dataset | Column | সমস্যা | প্রমাণ (সংখ্যা) |
|---|---|---|---|
| ISIC Archive 2 | 🔑 `melanocytic` | **নিখুঁত বিভাজন** — Melanoma/Nevus সব সময় True, বাকি ৭ class সব সময় False | **০ ব্যতিক্রম**, ১৭,৩৯৫+৭,৬৮১ row |
| ISIC Archive 2 | `attribution`, `copyright_license`, `image_type` | Dataset-এর উৎস (হাসপাতাল) চিনিয়ে দেয়, ক্লিনিক্যাল না | `attribution`-এর একটা value = ঠিক HAM10000 overlap সংখ্যা (৯,৮৭৩) |
| ISIC Archive 2 | 🔑 ৬টা sparse field | Missingness নিজেই source (হাসপাতাল) চিনিয়ে দেয় | ৫টা field কখনোই Hospital Clínic/ViDIR row-এ populated না |

<br>

## 📌 সারসংক্ষেপে

**১৬টা column** (leakage / label-source / non-clinical-administrative বিভাগে) + **আলাদা ৬টা sparse field** (source-institution leak প্রমাণিত, ২০২৬-০৭-০৯) = **মোট ২২টা column**, ৪টা dataset জুড়ে খুঁজে বের করে বাদ দেওয়া হয়েছে — প্রতিটার জন্য statistical প্রমাণ সহ, শুধু অনুমান না।

<br>

## ফলাফল: প্রতিটা Dataset-এর জন্য একটা `feature_whitelist.md`

| Dataset | মোট column | ✅ Model-এ ব্যবহারযোগ্য |
|---|---|---|
| PAD-UFES-20 | ২৯ | **২১টা** |
| HAM10000 | ১০ | **৩টা** (age, sex, anatomical_site) |
| ISIC Archive 1 | ৬ | **০টা** (শুধু ছবি) |
| ISIC Archive 2 | ৩০ | **৪টা active** (age_approx, sex, anatom_site_1, anatom_site_general) |

---

---

<a id="part-e"></a>

# Part E — বর্তমান Dataset অবস্থা

## এক নজরে

| Dataset | Train | Val | Test | Class | Imbalance |
|---|---|---|---|---|---|
| PAD-UFES-20 | 1,606 | 338 | 354 | 6 | ~16:1 |
| HAM10000 | 7,004 | 1,501 | 1,510 | 7 | ~58:1 |
| ISIC Archive 1 | 1,655 | 292 | 100 | 9 | **~393:1** ⚠️ |
| ISIC Archive 2 | 17,535 | 3,769 | 3,772 | 9 | ~63:1 |

⚠️ ISIC Archive 1-এ ~393:1 imbalance — মাত্র ১টা sample একটা class-এ। এটাই সবচেয়ে বেশি imbalance (`PROJECT_PLAN.md`-এ ভুলবশত "239:1" লেখা ছিল, এখন সংশোধিত)।

<br>

✅ **Dataset preparation সম্পূর্ণভাবে "ready":**

- ছবি ও metadata clean, একসাথে যুক্ত (`image_path` দিয়ে, copy না করে)
- Leakage-free split (patient/lesion-wise, cross-dataset overlap handled)
- কোন column নিরাপদ তা নির্ধারিত (feature whitelist)
- সব সিদ্ধান্ত document করা

⚠️ **"Ready" মানে "trained" না** — এখনো কোনো model আসলে পুরো data দিয়ে train হয়নি, শুধু code লেখা ও ছোট sample-এ test করা হয়েছে।

---

---

<a id="part-f"></a>

# Part F — এখনো যা বাকি

## Open Items (সৎভাবে বলা)

<br>

### 📝 Documented সীমাবদ্ধতা

*(এগুলো bug না, জেনেবুঝে রাখা সিদ্ধান্ত/সীমা)*

- Cross-dataset overlap **কমানো হয়েছে, পুরোপুরি মুছে ফেলা হয়নি** — exclusion list সবসময় সক্রিয়ভাবে apply করতে হবে, automatic না
- HAM10000 ও ISIC Archive 2-তে সত্যিকারের patient-level control নাই (শুধু lesion/group-level)
- ISIC Archive 1 শুধু image-only, Test set খুবই ছোট ও imbalanced
- ISIC Archive 2-এর ৩টা column (`anatom_site_2`, `anatom_site_special`, `dermoscopic_type`) whitelist-এ আছে কিন্তু নিজস্ব missingness/attribution check এখনো বাকি
- ৩টা image-এ (`ISIC_0028619`, `ISIC_0011126`, `ISIC_0011118`) দুই archive-এর মধ্যে label-এ মতভেদ আছে — inherited, শুধু flag করা

<br>

### ☐ এখনো শুরু হয়নি

*(correctly deferred)*

- Literature review চলমান (আরও paper যোগ হবে)
- Fairness/bias analysis (Fitzpatrick, বয়স ভিত্তিক) — Phase 8-এ হবে

🔑 **Model training — এখনো একটাও সম্পূর্ণ run হয়নি।** Phase 6 Stage 1-এর কোড লেখা শেষ, GPU-না-থাকা confirm হয়েছে, কিন্তু কোনো checkpoint/metrics file তৈরি হয়নি — Kaggle-এ shift করার প্রস্তুতি চলছে (বিস্তারিত Part G)।

---

---

<a id="part-g"></a>

# Part G — সামনের Roadmap

## বিস্তারিত (Phase 6-10)

<br>

## ☐ Phase 6 — Baseline Model Development

### এখন চলছে

**Stage 1 সিদ্ধান্ত (approved):**

| সিদ্ধান্ত | মান |
|---|---|
| Image branch | **EfficientNet-B0** (কম overfitting risk, ছোট dataset-এর জন্য ভালো) |
| Metadata branch | Simple MLP |
| Loss/Metric | Class-weighted loss, macro-F1 প্রধান (accuracy না) |
| Seeds | ৩টা প্রতি branch (মোট ৬টা run), mean±std রিপোর্ট |
| Image size | 224×224 resize+pad, load-time-এ |
| Test split | একদম শেষে একবারই ব্যবহার |

**অবস্থা** *(২০২৬-০৭-০৯, `Project_Tracking.md`-এ log করা)*:

Stage 1-এর কোড (image + metadata branch) **লেখা শেষ**, কিন্তু **এখনো কোনো training run সম্পূর্ণ হয়নি** — কোনো checkpoint বা metrics file ডিস্কে তৈরি হয়নি। GPU নাই তা নিশ্চিত করা হয়েছে (`torch.cuda.is_available() == False`), তাই Kaggle-এ shift করার সিদ্ধান্ত হয়েছে এবং কোডে (`config.py`) প্রতিফলিত করা হয়েছে, কিন্তু Kaggle dataset upload/config এখনো বাকি।

🔑 এটাকে **"verified end-to-end" বলা যাবে না** যতক্ষণ না একটা আসল training/evaluation run checkpoint, metrics CSV, এবং evaluation report তৈরি করে।

<br>

## ☐ Phase 7 — Multimodal Fusion

Late fusion (ছবি+metadata feature জোড়া লাগানো) → তারপর cross-attention fusion (মূল contribution, লিটারেচার gap পূরণ)।

<br>

## ☐ Phase 8 — Experiments & Evaluation

- 🔑 **PAD-UFES-20 ↔ HAM10000 cross-dataset generalization** — headline result (এই জোড়ায় কোনো overlap নাই, সবচেয়ে পরিষ্কার তুলনা)
- HAM10000→ISIC external validation (exclusion filter সহ)
- Fitzpatrick skin-tone fairness analysis
- Bootstrap significance testing

<br>

## ☐ Phase 9-10 — Thesis Writing → Publication

arXiv preprint → workshop (ISIC @ MICCAI/CVPR) বা journal (Diagnostics, IEEE Access, Computers in Biology and Medicine)

---

---

<a id="part-h"></a>

# Part H — প্রয়োজনীয় Skill ও জ্ঞান

<br>

## 🧠 এখনই বুঝতে হবে

*(concept, code লেখা না লাগলেও)*

- Train/Val/Test split, data leakage
- Class imbalance কেন শুধু accuracy বিভ্রান্তিকর (macro-F1 কেন লাগে)
- Overfitting কী

## 🐍 Basic Python

- pandas দিয়ে CSV পড়া/বোঝা (`df.head()`, `df['col'].value_counts()`)

## 🤖 Deep Learning concepts

*(নাম চেনা, কাজ বোঝা যথেষ্ট)*

- CNN, Pretrained model/Transfer learning, Loss function, epoch, batch

## 🛠️ Tools

*(হাতে-কলমে শিখতে হবে)*

- Git/GitHub (backup, version রাখা)
- Kaggle Notebooks (GPU দিয়ে training)
- Basic command line (`python -m src.models.train --branch image`)

<br>

## 🔑 সবচেয়ে গুরুত্বপূর্ণ (technical না হলেও critical)

**প্রতিটা সিদ্ধান্তের "কেন" ব্যাখ্যা করতে পারা** — biopsed leakage কেন বাদ দেওয়া হলো, patient-wise split কেন জরুরি, cross-dataset overlap কেন সমস্যা — এগুলো নিজের ভাষায় বলতে পারা viva-তে সবচেয়ে বেশি কাজে লাগবে।

---

---

<a id="part-i"></a>

# Part I — নিজে চালিয়ে দেখার Command

```
# ১. সেটআপ (একবারই)
python -m venv .venv
.venv\Scripts\Activate.ps1
pip install -r requirements.txt

# ২. Audit (যেকোনো order-এ চালানো যায়)
.venv\Scripts\python.exe -m src.data_audit.run_audit_pad_ufes20
.venv\Scripts\python.exe -m src.data_audit.run_audit_ham10000
.venv\Scripts\python.exe -m src.data_audit.run_audit_isic_archive_1
.venv\Scripts\python.exe -m src.data_audit.run_audit_isic_archive_2

# ৩. Cleaning (৪টা আগে শেষ হতে হবে, তারপর leakage-filter)
.venv\Scripts\python.exe -m src.data_cleaning.run_cleaning_pad_ufes20
.venv\Scripts\python.exe -m src.data_cleaning.run_cleaning_ham10000
.venv\Scripts\python.exe -m src.data_cleaning.run_cleaning_isic_archive_1
.venv\Scripts\python.exe -m src.data_cleaning.run_cleaning_isic_archive_2
.venv\Scripts\python.exe -m src.data_cleaning.cross_dataset_leakage_filter

# ৪. EDA
.venv\Scripts\python.exe -m src.eda.eda_pad_ufes20
.venv\Scripts\python.exe -m src.eda.eda_ham10000
.venv\Scripts\python.exe -m src.eda.eda_isic_archive_1
.venv\Scripts\python.exe -m src.eda.eda_isic_archive_2
.venv\Scripts\python.exe -m src.eda.eda_cross_dataset
```

### ✅ সফল হয়েছে কিনা যাচাই:

- `reports/<Dataset>/`-এ summary `.md` ফাইল তৈরি হয়েছে কিনা
- `corrupted_images.csv` খালি কিনা
- `split_quality_report.csv` error ছাড়া তৈরি হয়েছে কিনা

---

---

<a id="part-j"></a>

# Part J — Supervisor/Viva Talking Points

<br>

**১.** *"৪টা public dataset ব্যবহার করেছি, প্রতিটাতে rigorous audit করেছি — শুধু download করে ব্যবহার করিনি।"*

**২.** *"Patient-wise split ব্যবহার করেছি, image-wise না — এই field-এর একটা common ভুল এড়িয়েছি।"*

**৩.** 🏆 *"১৬টা leakage/label-source column + আলাদা ৬টা source-identity-leak sparse field (মোট ২২টা) খুঁজে বের করে বাদ দিয়েছি, প্রতিটা statistical প্রমাণ (phi coefficient) সহ — এটা প্রমাণ করে আমি critical ভাবে চিন্তা করেছি, শুধু model বানাইনি।"* → **সবচেয়ে শক্তিশালী পয়েন্ট**

**৪.** *"HAM10000 ও ISIC Archive আসলে একই source-এর ছবি — এই লুকানো overlap ধরেছি এবং documented ভাবে সমাধান করেছি।"*

**৫.** *"মূল contribution হবে cross-dataset generalization (PAD-UFES-20 ↔ HAM10000) — বেশিরভাগ paper এটা মাপে না।"*

---

---

## ⚠️ একটা জরুরি নোট

এই document-টা তোমার existing ফাইলগুলো (2026-07-09 পর্যন্ত পড়া অবস্থা) ভিত্তি করে বানানো। যদি এরপর কোনো নতুন কাজ (যেমন Kaggle training result, sparse-field check-এর বাকি ৩টা column) হয়ে থাকে, সেগুলো এখানে নাই — Claude Code-কে জিজ্ঞেস করে `Project_Tracking.md`-এর সবচেয়ে উপরের "Session Handoff" অংশ পড়ে latest status নিশ্চিত করে নাও।