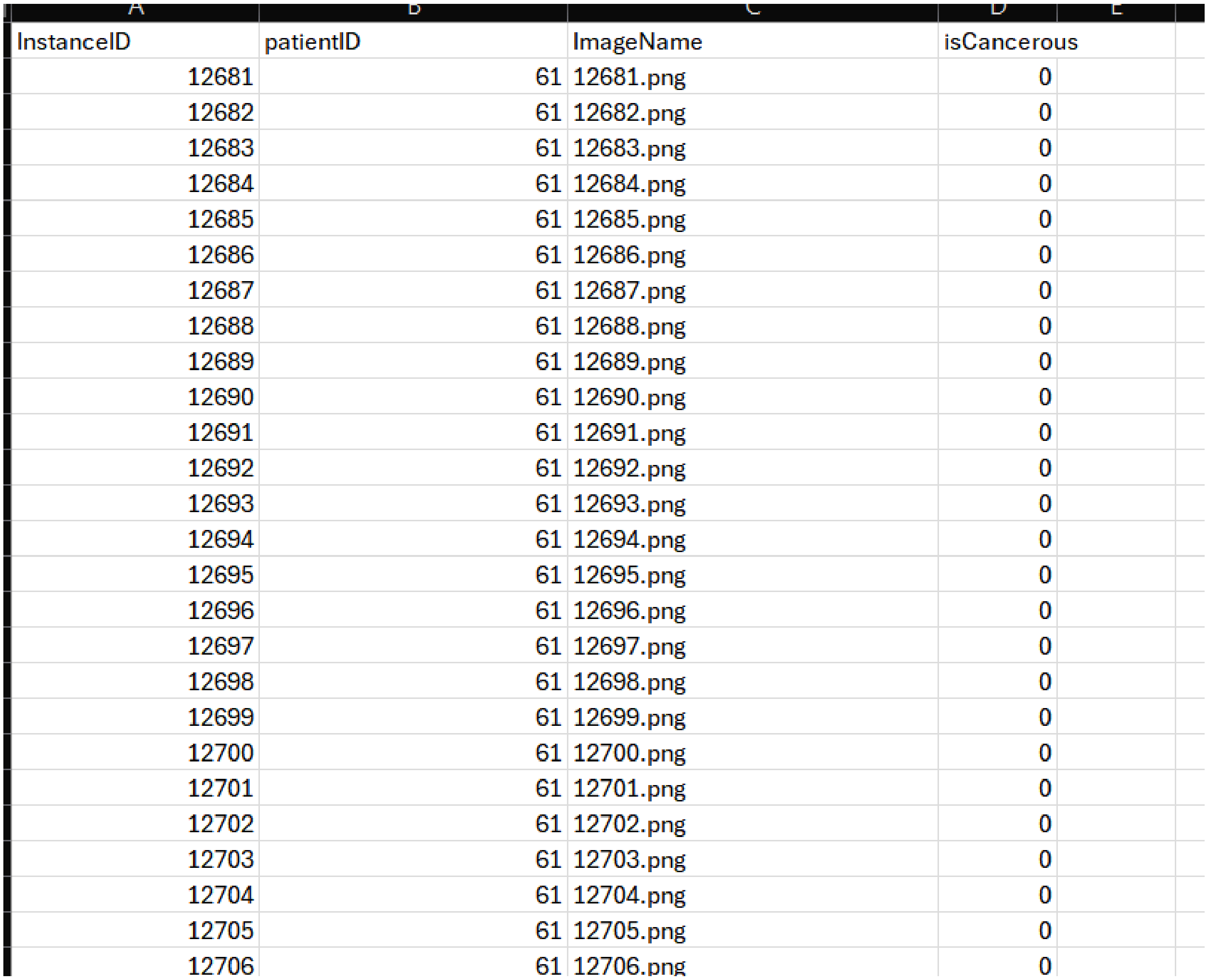
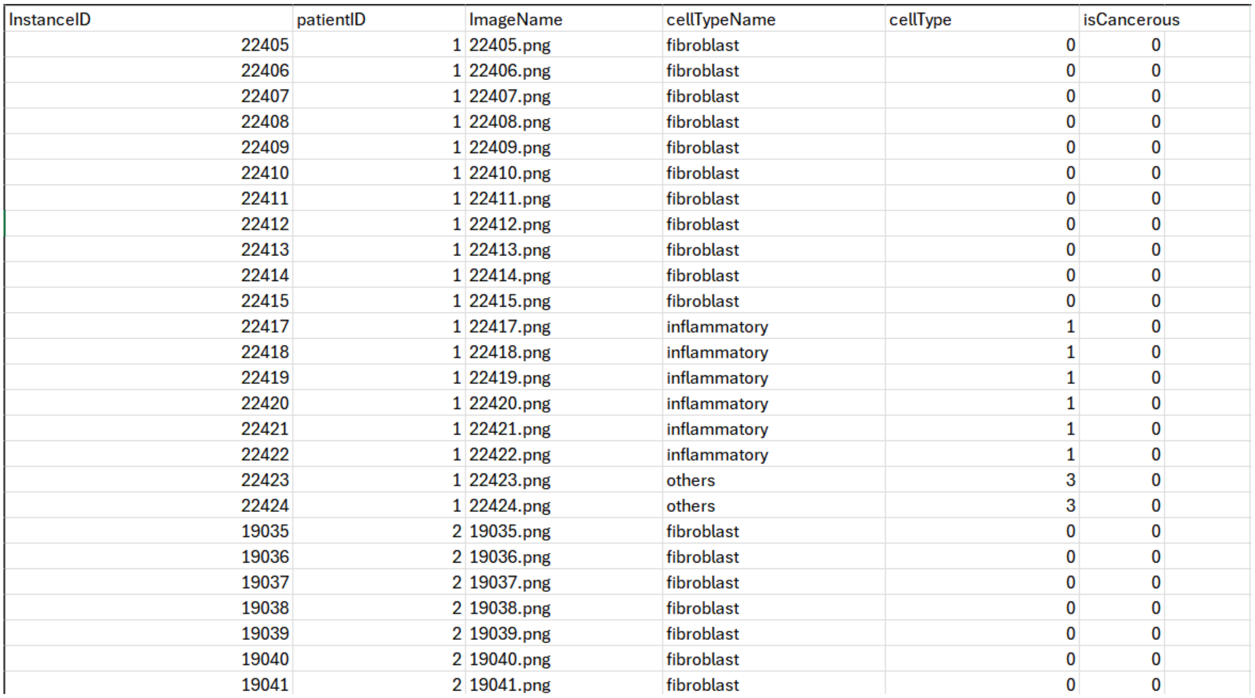
Assignment 2

Extra Dataset



A	B	C	D	E
InstanceID	patientID	ImageName	isCancerous	
12681	61	12681.png	0	
12682	61	12682.png	0	
12683	61	12683.png	0	
12684	61	12684.png	0	
12685	61	12685.png	0	
12686	61	12686.png	0	
12687	61	12687.png	0	
12688	61	12688.png	0	
12689	61	12689.png	0	
12690	61	12690.png	0	
12691	61	12691.png	0	
12692	61	12692.png	0	
12693	61	12693.png	0	
12694	61	12694.png	0	
12695	61	12695.png	0	
12696	61	12696.png	0	
12697	61	12697.png	0	
12698	61	12698.png	0	
12699	61	12699.png	0	
12700	61	12700.png	0	
12701	61	12701.png	0	
12702	61	12702.png	0	
12703	61	12703.png	0	
12704	61	12704.png	0	
12705	61	12705.png	0	
12706	61	12706.png	0	

Data



InstanceID	patientID	ImageName	cellTypeName	cellType	isCancerous
22405	1	22405.png	fibroblast	0	0
22406	1	22406.png	fibroblast	0	0
22407	1	22407.png	fibroblast	0	0
22408	1	22408.png	fibroblast	0	0
22409	1	22409.png	fibroblast	0	0
22410	1	22410.png	fibroblast	0	0
22411	1	22411.png	fibroblast	0	0
22412	1	22412.png	fibroblast	0	0
22413	1	22413.png	fibroblast	0	0
22414	1	22414.png	fibroblast	0	0
22415	1	22415.png	fibroblast	0	0
22417	1	22417.png	inflammatory	1	0
22418	1	22418.png	inflammatory	1	0
22419	1	22419.png	inflammatory	1	0
22420	1	22420.png	inflammatory	1	0
22421	1	22421.png	inflammatory	1	0
22422	1	22422.png	inflammatory	1	0
22423	1	22423.png	others	3	0
22424	1	22424.png	others	3	0
19035	2	19035.png	fibroblast	0	0
19036	2	19036.png	fibroblast	0	0
19037	2	19037.png	fibroblast	0	0
19038	2	19038.png	fibroblast	0	0
19039	2	19039.png	fibroblast	0	0
19040	2	19040.png	fibroblast	0	0
19041	2	19041.png	fibroblast	0	0

✅ Final Clarified Plan (Based on Spec + Marking Guide)

🧪 Task Summary

You must solve:

1. **Binary classification:** `isCancerous` (0/1)
2. **Multiclass classification:** `cellTypeName` (4 types)

🧠 Datasets

- **mainData.csv** (60 patients): Full labels for both tasks.

- **extraData.csv** (39 patients): Only `isCancerous` labels. Missing `cellTypeName`.

🧬 Step-by-Step Approach

📍 Step 1: EDA and Preprocessing (6 marks)

- Analyze image sizes, distributions, pixel intensity.
- Identify class imbalance in `isCancerous` and `cellTypeName`.
- Normalize images, maybe augment (rotation, flip).
- Split by **patient ID**, not randomly (to avoid leakage).

✅ This addresses: EDA, class imbalance, leakage, data handling.

📍 Step 2: Supervised Baseline Models (12 marks)

Task 1 – `isCancerous` (All 99 patients):

- CNN (or classic ML if you're short on time).
- Use all data, because label exists in both files.

Task 2 – `cellTypeName` (Only 60 patients for now):

- Train model on `mainData.csv`.
- Justify model (e.g., CNN good for image patterns).
- Handle imbalance (class weights, oversampling, augmentation).

✅ This gives you supervised models and satisfies minimum requirement.

📍 Step 3: Semi-Supervised Learning (HD only)

> For HD, you must use extraData.csv to improve Task 2 (cellTypeName)

✅ Do This:

1. Use the Task 2 model (from 60 patients) to **predict** **`cellTypeName`** for 39 patients.

2. Keep **only high-confidence predictions** (e.g., softmax > 90%).
3. Add these pseudo-labeled examples to the training set.

Now, your `cellTypeName` model is trained on **60 + ~X pseudo-labeled patients**.

✅ This directly matches what your supervisor said and what the spec means for HD.

📍 Step 4: Final Unified Model (HD strategy)

> Combine both datasets (main + pseudo-labeled extra) to train a final model that can do both tasks.

You can do:

- **Two heads CNN**: One output for `isCancerous`, one for `cellTypeName`
- Or train two separate CNNs (simpler)

✅ Final model + pseudo-labeled data = HD

📍 Step 5: Independent Evaluation (10 marks)

- Compare your results to at least 2–3 papers using the **original CRCHistoPhenotypes dataset**
- Discuss strengths, weaknesses, and real-world impact.

✅ Do this to complete the 20% Independent Evaluation.

💯 How to Hit HD in the Marking Guide

Section	Action
Class Imbalance	Show bar plots of class counts, use class weights or augmentation
EDA	Visualize pixel intensities, sizes, sample images
Justification	Explain every preprocessing choice: why normalize, why CNN, etc.
Base Models	Use CNN or similar with good justification
Semi-Supervised	Pseudo-label `cellTypeName` on extraData.csv and combine

Model Eval	Plot accuracy, precision, recall, F1, confusion matrix
Independent Eval	Compare with 2–3 published results
Report	Add clean visuals, structured headings, describe every figure

---

Would you like a starting template for your report or notebook pipeline next?