Introduction to Machine Learning (67577): Hackathon 2025
Challenge 2: Detecting Breast Cancer Attributes 👨‍⚕️

June 11, 2025

1 Hackathon Challenge 2: Predicting Attributes of Breast Cancer

Breast cancer is one of the most pervasive malignant diseases among women worldwide and in Israel. If treated early and appropriately, the chance of remission is high.[1] In this task, you're given *anonymized* data of breast cancer patients treated in Israel over the last years. Your task is to predict certain medical characteristics of the disease for each patient based on their available data. This can help save doctors' time, validate and "double-check" their decisions, and alleviate the cost of expensive tests.

Important Note:
You're given actual medical data provided with consent by the medical center after thorough anonymization for research purposes. Please treat it respectfully and don't distribute it beyond the course participants. The findings made by Hackathon projects have the potential to contribute to ongoing research, and we'll be happy to follow up with groups that have made interesting findings with regard to the specific tasks and the data in general.

Challenges:

- The data requires a **normalization and preprocessing**. For example, some of the categorical values were inputted in free text, other values are missing.
- **Multilabel problem** - While in class we dicussed scenarios where there is exactly one correct label, the categorical value in Part 1 can have between 0-3 correct values.

Dataset:
The training dataset contains 65,798 entries (across train and test), each with 34 features, as detailed below in Section 1.7. It is supplied in the following files:

- **train.feats.csv**: 49,351 records in a CSV file, where each row contains a single instance (single patient visit) with 34 features. The first row describes feature names.
- **test.feats.csv**: 16,447 records in a CSV file, each containing a single instance (single patient visit) with 34 features. The first row describes feature names.

1.1 Getting the data and supporting scripts

Follow instructions in https://github.com/gabrielStanovsky/iml-hack-oncology to download the data, scripts, and a few baselines we implemented to demonstrate how to work with the data.

[1] https://www.cancer.org.il/template/publications.aspx?maincat=12

1.2 Part 1: Predicting Metastases

Given each visit characteristics, predict metastases sites (multi-label, multi-class categories).[2]
You are given the corresponding labels for the train set:

- train.labels.0.csv: 49,351 records, each line corresponds to the patient visit described in the same line index in train.feats.csv.

You must submit code to train a model and make predictions, along with its prediction for test features. It should be in the same format as the train labels but with 16,447 entries, corresponding to predictions for the test features.

1.2.1 Submit

- model and code.
- **predictions.csv**: 16,447 entries each corresponding to *test.feats.csv*, following the same format as *train.labels.0.csv*.

1.2.2 Evaluation

Will be done using the following:

1. Micro average F1 score
2. Macro average F1 score

You are given the evaluation script with which you can test your output format validity: *evaluate_part_0.py* (run the file without parameters to get help).
Note: We suggest that before submitting, you run your model on the train set and compare it with the given gold predictions to verify your output format and calculate the empirical loss.

1.3 Part 2: Predicting Tumor Size

Given each visit characteristics, predict the diagnosed tumor size (in mm).
You are given the corresponding labels for the train set:

- train.labels.1.csv: 49,351 records, each corresponding with the features described in Section 1.7.

You must submit code to train a model and make predictions, along with its prediction for test features. It should be in the same format as the train labels but with 16447 entries corresponding to predictions for the test features.

1.3.1 Submit

- Your code and model.
- predictions.csv: 16,447 entries each corresponding to *test.feats.csv*, following the same format as *train.labels.1.csv*.

1.3.2 Evaluation

Will be done using mean squared error.
You are given the evaluation script to test your output format validity: *evaluate_part_1.py* (run the file without parameters to get help).
Note: We suggest that you run your model on the train set and compare it with the given gold predictions to verify your output format and calculate the empirical loss before submitting.

[2] https://en.wikipedia.org/wiki/Metastasis

1.4 Part 3: Unsupervised Data Analysis

Teach us something interesting about the data!

We've discussed in class various techniques for unsupervised analysis (e.g., clustering, dimensionality reduction, principal component analysis). In this task, you are required to perform any of these methods to identify interesting trends in the data. For example (but not limited to):

1. What does the principal component patient look like?
2. Can you identify recurring patterns?
3. Can you provide interesting clustering of the data (k-means or spectral)?

As mentioned above, this data is part of ongoing research, so any finding here would be valuable, and we'd be happy to follow up with the teams that make the most insightful finding. Be creative!

Note:

Make sure to do this part using just *train.feats.csv*. This task is unrelated to the other two.

1.4.1 Submit

A short report outlining your findings, along with any supporting code. This can be achieved via a Python notebook, streamlit, or other suitable form.

1.4.2 Evaluation

As this is an open-ended task, we'll manually examine the results and test their soundness, validity, and observations.

1.5 Questions to domain experts

Join the dedicated google group to read and post questions to domain experts.[3] Importantly, these experts are volunteers in this domain (without knowledge in ML), so keep your questions concise, polite, and about the domain. They will only answer a subset of the questions, don't get blocked while waiting for an answer.

1.6 Submission Instructions

Submit a zip file which follows the format in Figure 1.

- `README.txt` — contains a file list and a brief description of each file.
- `USERS.txt` — contains the team members' logins and IDs. Use one line per student, in the format: `login, ID`.
- All other files are specific to the subtasks, see above.

[3] Google group: https://groups.google.com/a/mail.huji.ac.il/g/iml-2025-cancer-hackathon

```
<student-id>.tar
├── USERS.txt
├── README.txt
├── subtaskI
│   ├── requirements.txt
│   ├── predictions.csv
│   └── <additional files you developed for the model to run>
├── subtaskII
│   ├── requirements.txt
│   ├── predictions.csv
│   └── <additional files you developed for the model to run>
└── subtaskIII
    └── project.pdf
```

Figure 1: *The structure of a valid submission.*

1.7 Feature description

Column Name	Description
Id-hushed_internalpatientid	Id of patients, we have 11623 different patients - 11623 values, Hash - long HEX
Form Name	Type of medical visit - 9 values, Hebrew words
Hospital	Hospital code - 4 values, float
User Name	User name of the doctor reporter - 154 values, <NUM>_Onco
Age	Age of patient - 11623 values, float
Basic stage	Carcinoma Basic stage - 4 Different stages (c, p, r, null)
Diagnosis date	Date of diagnosis - Datetime
Her2	Tumor marker test that determines the number of copies of the HER2 gene or the amount of HER2 protein in a cancer cell - various formats
Histological diagnosis	Histological diagnosis - 41 values, CONSTS (English caps)
Histopathological degree	Histopathological degree - 6 values - G1 to G4 + GX + Null
Ivi -Lymphovascular invasion	Whether the tumor invaded blood vessels or lymph nodes - 17 values, various formats.
KI67 protein	The rate of cell multiplication in the tumor - a number - various formats.
Lymphatic penetration	How much the Lympha was penetrated - 5 values, L<x> + null
M -metastases mark (TNM)	Amount of existence of metastases - 6 values, M<x>
N -lymph nodes mark (TNM)	Amount of lymph invasion - 21 values, N<x>
T -Tumor mark (TNM)	Size of tumor in the first exam - 22 values, T<Something>
Margin Type	Tumor margin type - 3 values, Hebrew consts
Nodes exam	How many Lymph nodes were examined - 42 values, float
Positive nodes	How many of Lymph nodes contained carcinoma metastases - 28 values, integers
Side	Breast side of tumor - 3, Hebrew const
Stage	Stage of cancer - 17 values, English const
Surgery date1	Date of first surgery - Datetime
Surgery date2	Date of second surgery - Datetime
Surgery date3	Date of third surgery - Datetime
Surgery name1	Name of first surgery - 23 values, CONSTS (English caps)
Surgery name2	Name of second surgery - 18 values, CONSTS (English caps)
Surgery name3	Name of third surgery - 6 values, CONSTS (English caps)

Surgery sum	Number of surgeries: 3 values (1, 2, 3)
Tumor depth	Depth of tumor - 6 values, float
Tumor width	Width of tumor - 31 values, float
er	Tumor marker test that determines the sensitivity to estrogen of the cancer cell - various formats
pr	Tumor marker test that determines the sensitivity to progesterone of the cancer cell - various formats
surgery before or after-Actual activity	Name of surgery before diagnosis - 10 values, Hebrew short string
surgery before or after-Activity date	Date of surgery before diagnosis - Datetime

Table 1: *Breast Cancer data column descriptions*