Enhanced Early Cancer Detection via Topological Data Analysis of Blood Biomarkers: A Non-Invasive Approach for Accurate, Multi-Cancer Screening

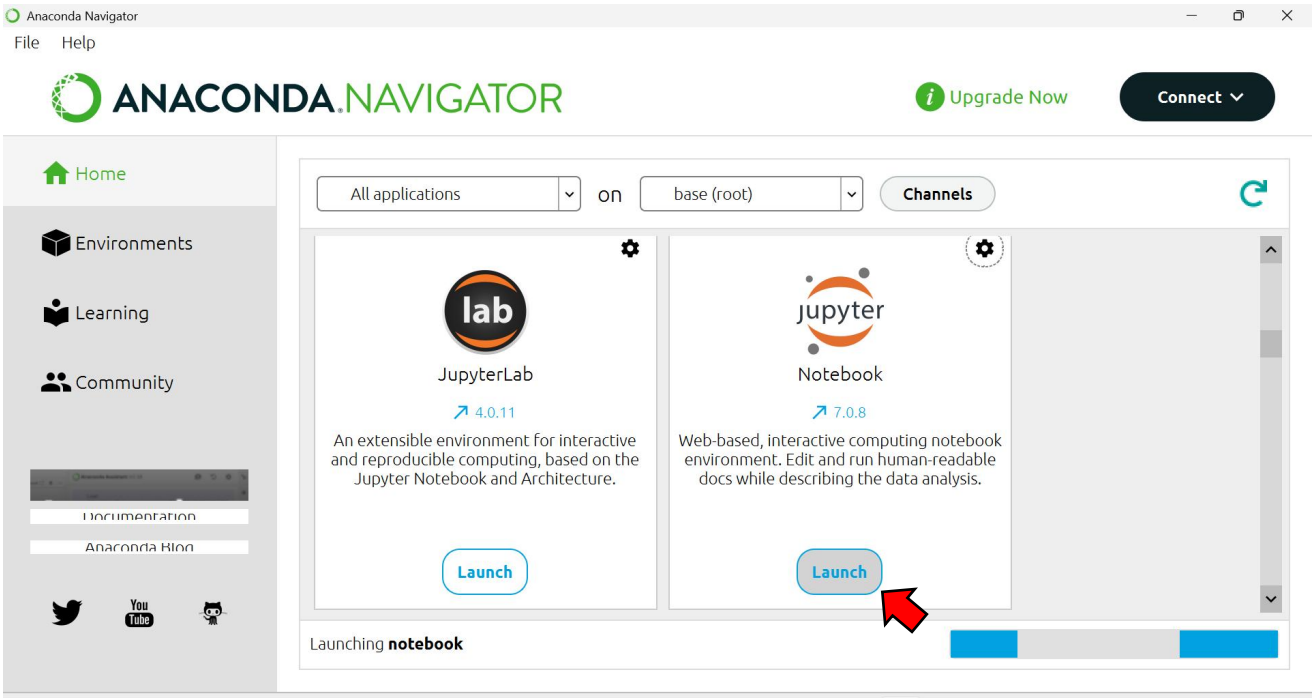
Sudarshan Gogoi[a], Bikash Thakuri[b], Amit Chakraborty[c,*]

[a,,c] Department of Mathematics, Sikkim University, Gangtok-737102, Sikkim, India
[b] School of Basic Sciences, SRM University, Sikkim

Tutorial on Biomarker Selection, Cancer Detection, and Tissue Localization Process

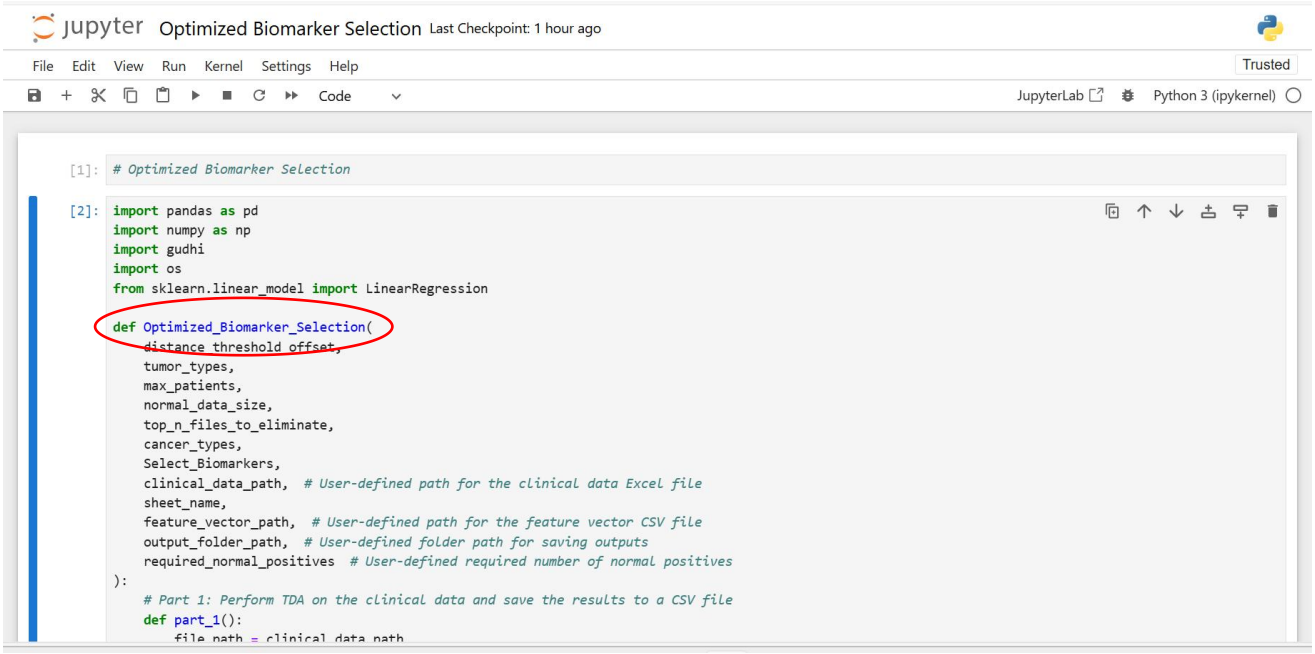
1. Open the link: https://github.com/Sudarshan-Gogoi/TDA-Detect-Cancer/tree/main .

2. Click on Codes and files
3. Download Files:
- "Optimized Biomarker Selection.ipynb"
- "TDAcancer_detect.ipynb"
- "Clinical cancer data.xlsx" (Example File)
4. Open Anaconda Navigator.
5. Launch Jupyter Notebook (Version 7.0.8):
- Click on Launch under Jupyter Notebook.



Biomarker Selection Process

6. Open Jupyter Notebook and load "Optimized Biomarker Selection.ipynb". (If the user wants to use the default example biomarker set for analysis, then directly move to the step. 28)



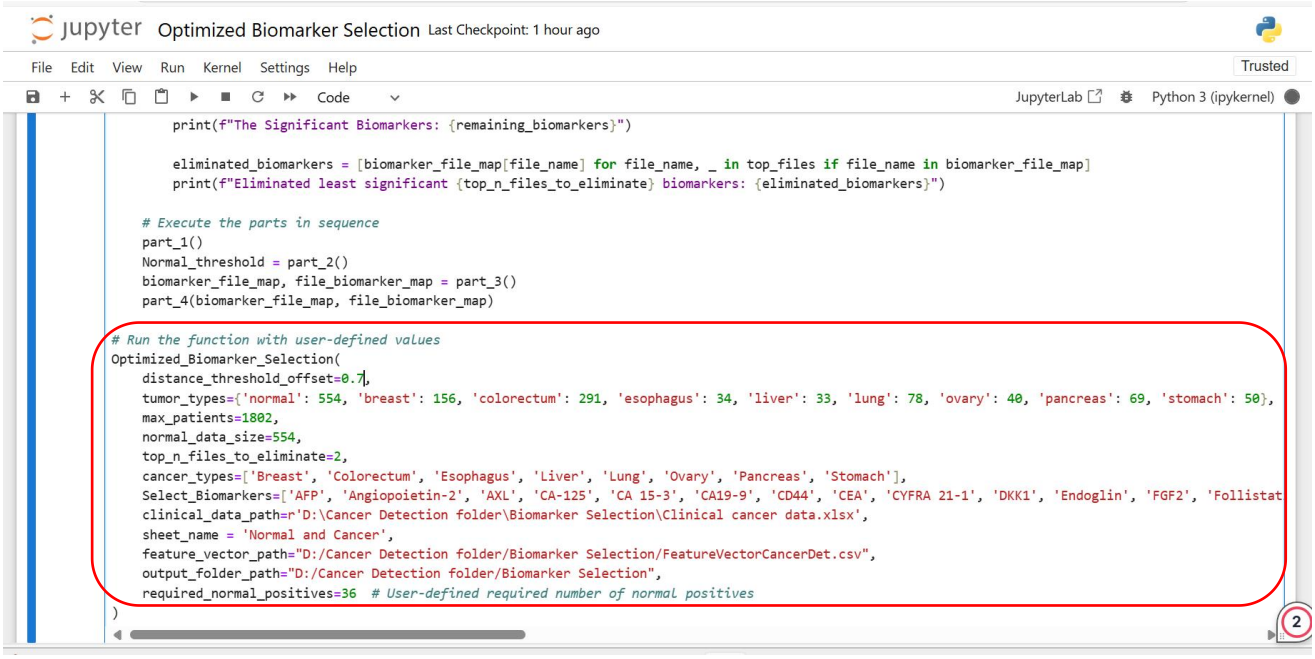
7. Enter the required inputs to the function ***Optimized_Biomarker_Selection***. The input descriptions are given below:

- **distance_threshold_offset:** Enter a value between 0 and 2.
- **tumor_types:** List the tumor types present in the dataset along with the number of patients to use as training samples for each tumor.
- **max_patients:** Specify the total number of patients in the dataset.
- **normal_data_size:** Provide the number of healthy individuals to use as training samples.
- **top_n_files_to_eliminate:** Mention the number of least significant biomarkers to eliminate from the set of biomarkers present in the dataset.
- **cancer_types:** List the cancer types present in the dataset.
- **Select_Biomarkers:** List the biomarkers present in the dataset.
- ***clinical_data_path:** Path to the file containing the Excel data file **"Clinical cancer data.xlsx"**.
- **sheet_name:** Input the patient data sheet **"Normal and Cancer"** from the provided Excel data file **"Clinical cancer data.xlsx"**.
- **feature_vector_path:** Path to store the feature vector file.
- **output_folder_path:** Path to store the intermediate result files.
- **required_normal_positives:** Minimum number of cancer false positives that can be considered to make specificity above 95%.

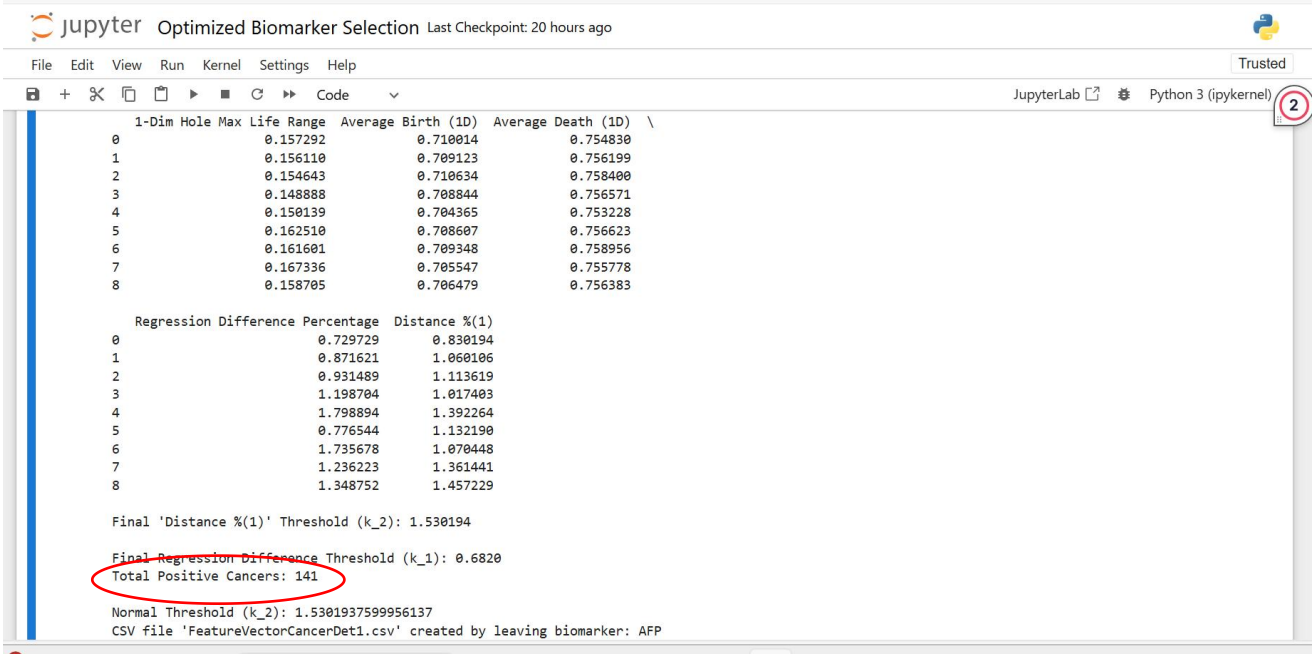
***Note:** For the example file "Clinical cancer data.xlsx", the user must provide **distance_threshold_offset**, **top_n_files_to_eliminate**, **clinical_data_path**, **feature_vector_path**, and **output_folder_path**. The other inputs are set as default. For **distance_threshold_offset** and **top_n_files_to_eliminate**, the user can use values from **Table 1** (given below) as default. The inputs

are case-sensitive, so refer to the example inputs and ensure that your inputs match the given case exactly.

8. Run the function ***Optimize_Biomarker_Selection***.

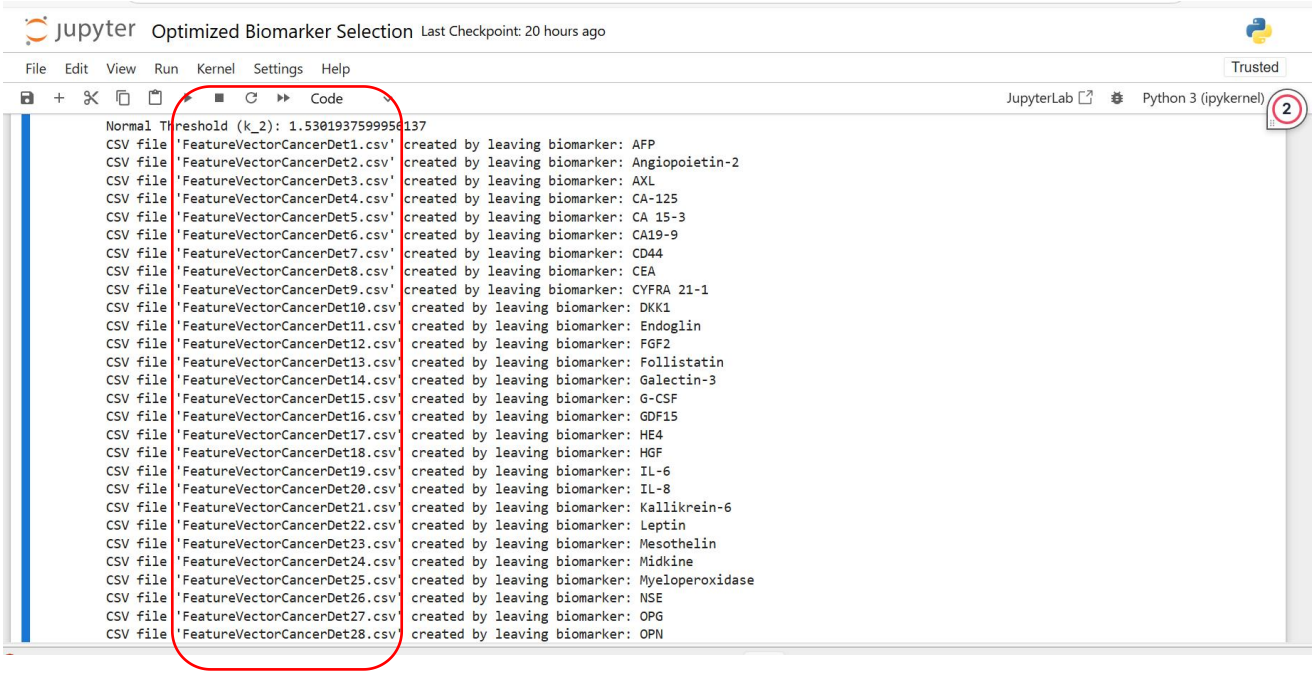


9. Observe the **Total Positive Cancers** number.

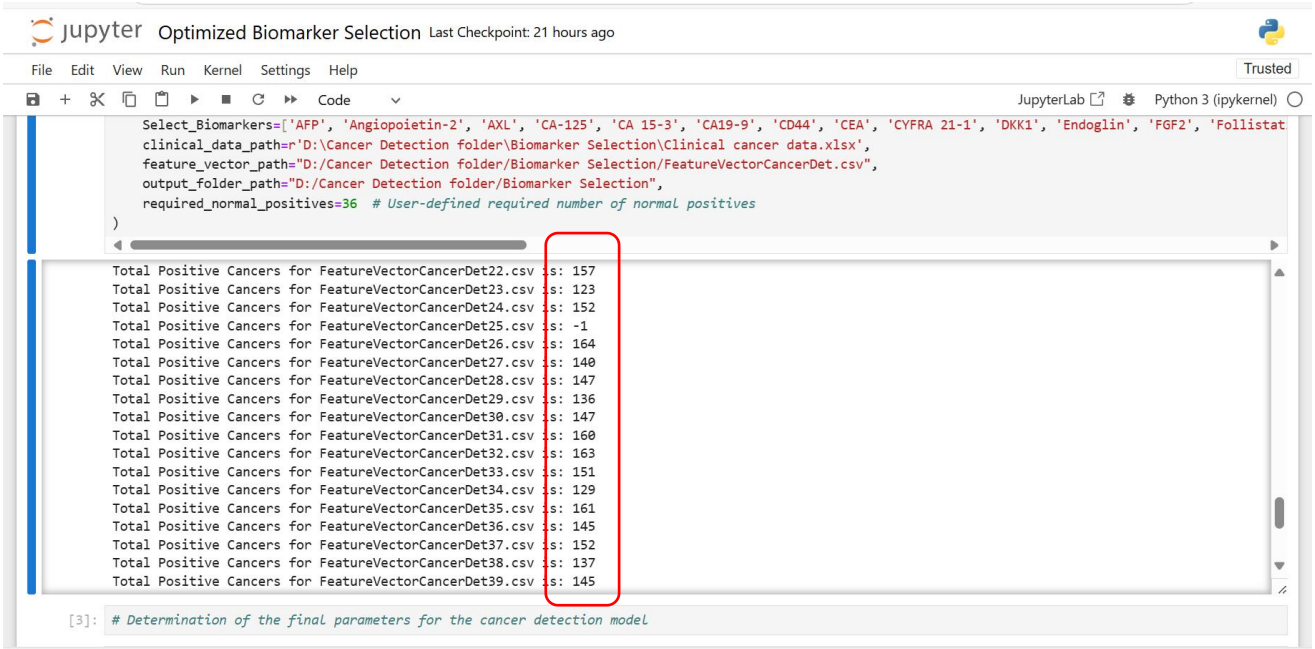


10. If the Total Positive Cancers number is not satisfactory, proceed to the next step.
11. A number of CSV files (e.g., FeatureVectorCancerDet1, FeatureVectorCancerDet2, etc.) will be generated and saved in the given "**output_folder_path**" for required analysis.

Biomarker Selection Process

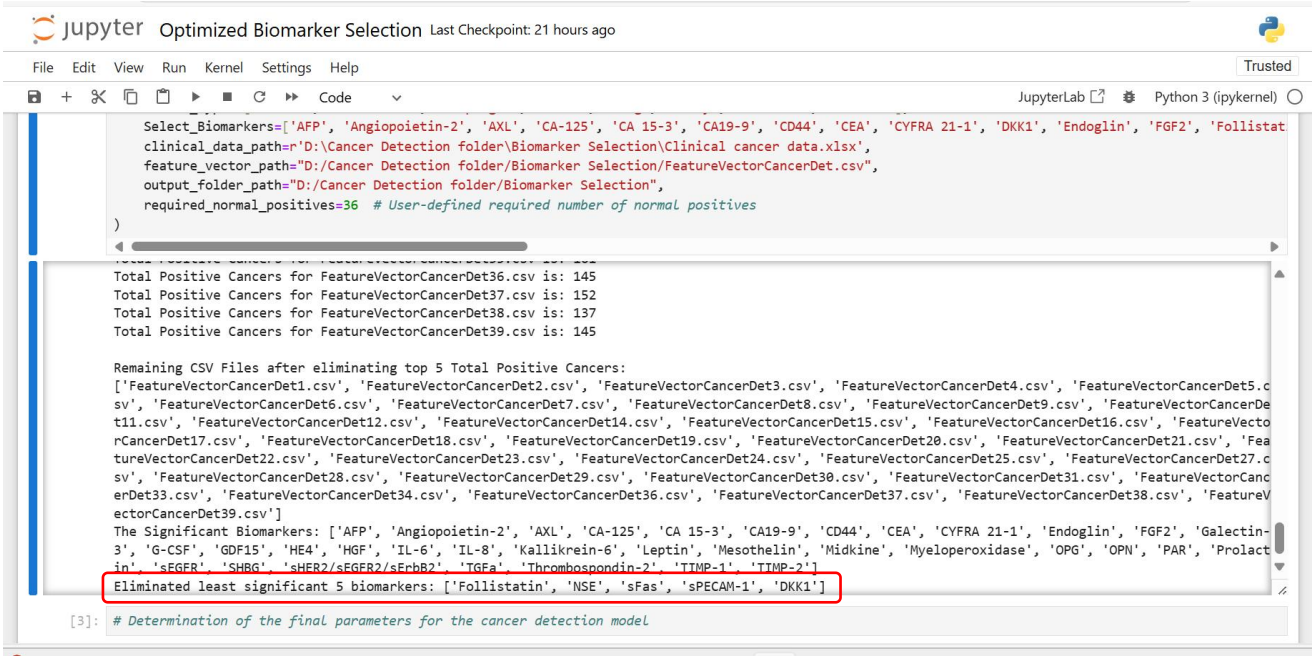


12. The Total Positive Cancers number for each CSV file will be shown. -1 indicates that no cancer positives were obtained for the given threshold.

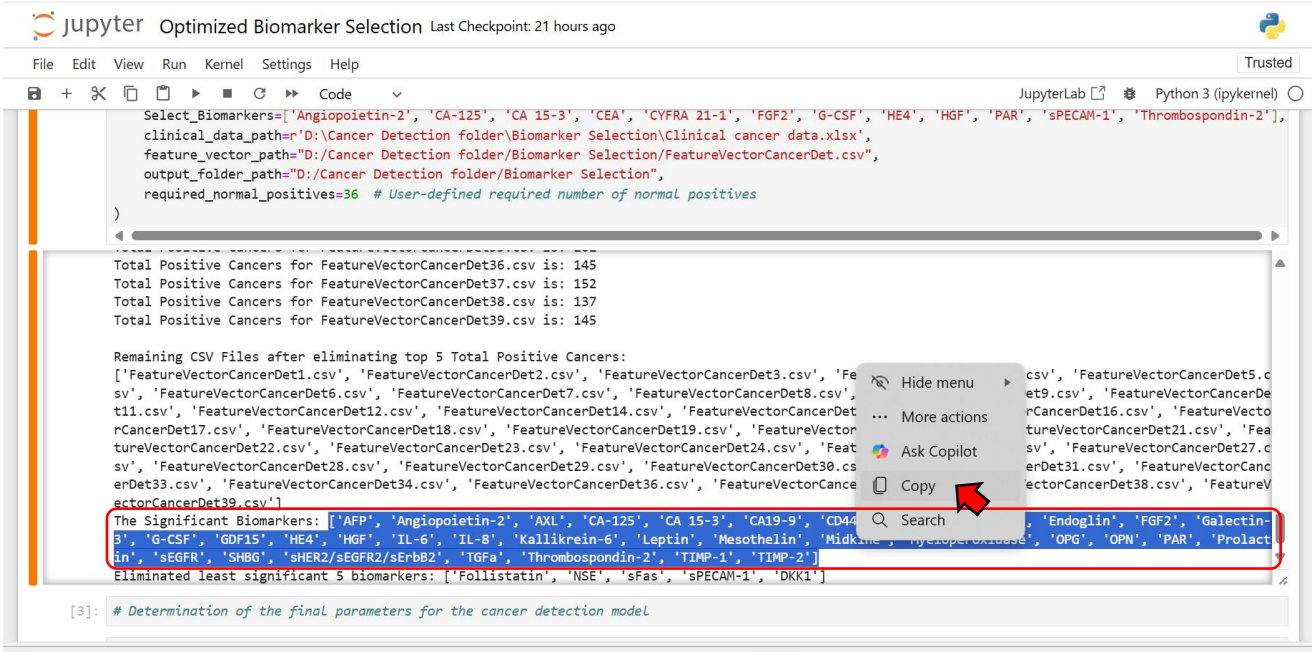


13. The least significant biomarkers will be eliminated based on the number of **top_n_files_to_eliminate** selected by the user.
14. **top_n_files_to_eliminate** will include those biomarkers related to the CSV files (e.g., FeatureVectorCancerDet1, FeatureVectorCancerDet2, etc.) with the top n Total Positive Cancers numbers.

Biomarker Selection Process

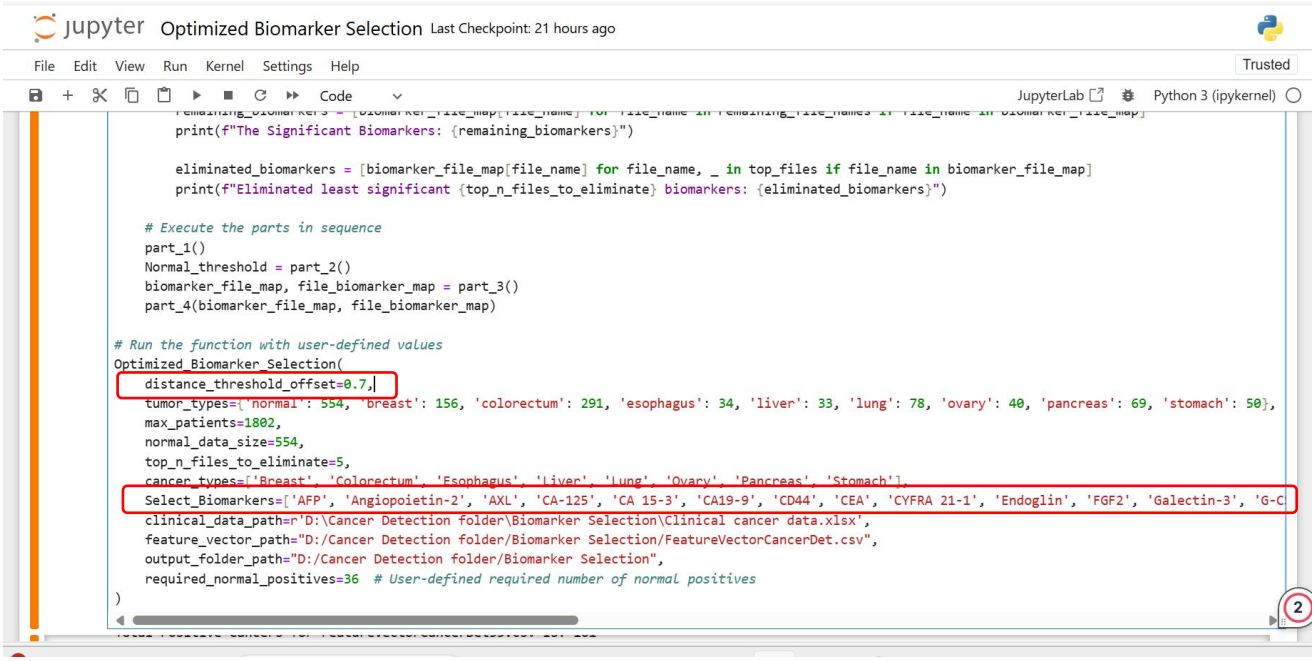


15. Verify the least significant biomarkers with existing cancer biomarker literature. If any biomarker is already validated as a major biomarker in cancer detection for the given "**tumor_types**," change the "**distance_threshold_offset**" parameter in Step 7 with another value and repeat the process.
16. If no major biomarkers are in the least significant biomarkers list, copy the Significant Biomarkers list.

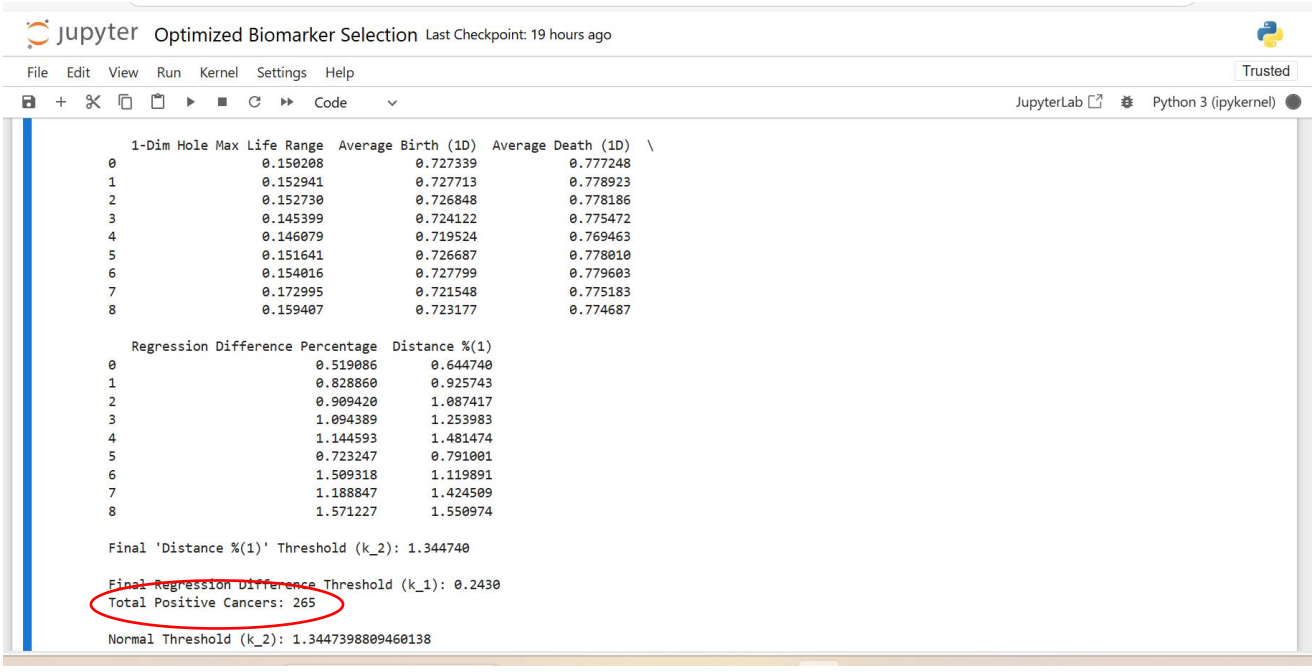


Biomarker Selection Process

17. Paste the copied biomarkers list in "**Select_Biomarkers**".
18. Enter a "**distance_threshold_offset**" value between 0 and 2.



19. Run the function and observe the Total Positive Cancers number again (Iteration. 2).



20. If the Total Positive Cancers number is not satisfactory, then continue from Step 15 to 19 again.
21. Repeat this process until you get a satisfactory number of Total Positive Cancers.

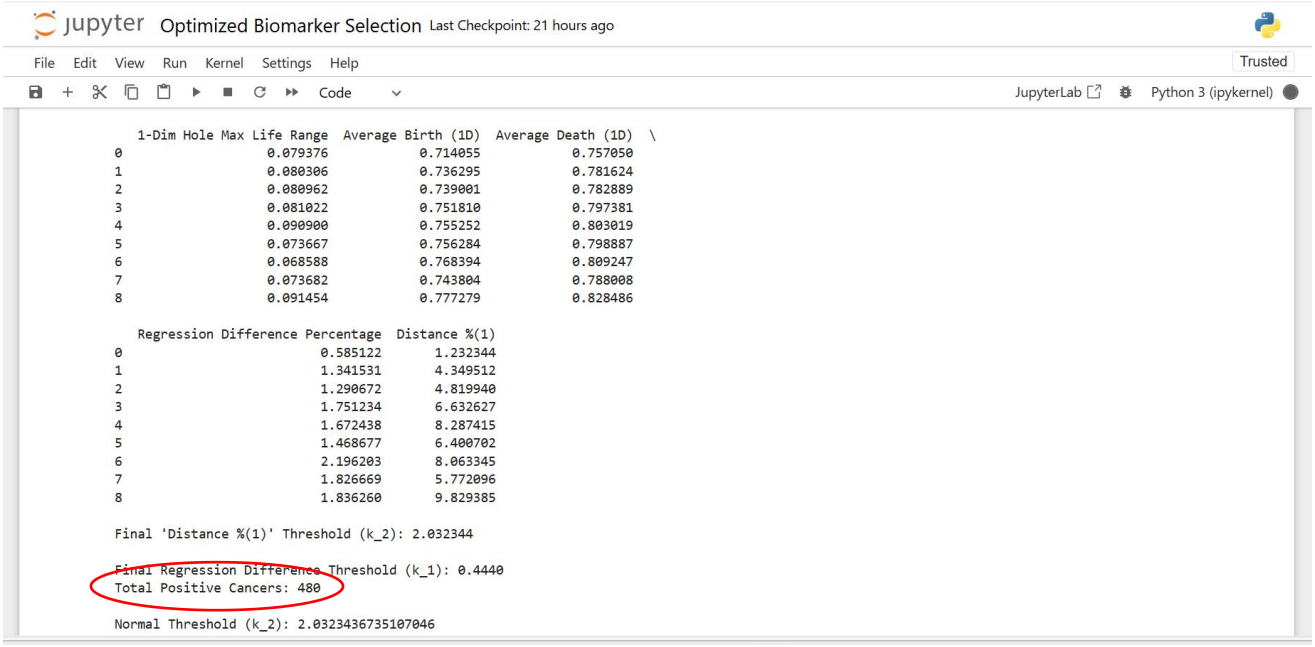
22. If a satisfactory number of Total Positive Cancers is obtained for a certain biomarker set, and further elimination of biomarkers does not improve the Total Positive Cancers number, that biomarker set can be considered the final significant biomarker set.

Table 1 (Example)

Iteration No.	Regression Difference Threshold (k_1)	Total Cancer Positives	Distance %(1) Threshold (k_2) (Distance %(1) + distance_threshold_offset)		Biomarkers eliminated	top_n_files_to_ eliminate
			Distance %(1)	distance_threshold _offset		
1	0.682	142	0.830194	0.7	'Follistatin', 'NSE', 'sFas', 'DKK1', 'sEGFR'	5
2	0.243	265	0.64474	0.7	'TIMP-2', 'Galectin-3', 'TGFa', 'TIMP-1', 'SHBG'	5
3	0.454	332	0.59581	0.4	'AXL', 'Leptin', 'Midkine', 'Kallikrein-6', 'IL-8'	5
4	0.248	317	0.710615	0.8	'OPG', 'Prolactin', 'AFP', 'Mesothelin', 'Endoglin'	5
5	0.723	353	0.900462	0.6	'Myeloperoxidase', 'CD44', 'OPN', 'GDF15', 'sHER2/sEGFR2/sErbB2'	5
6	0.662	448	1.053634	0.8	'IL-6', 'CA19-9'	2
7	-	480	-	-	-	-

23. Use the default parameters For **distance_threshold_offset** and **top_n_files_to_eliminate**, from **Table 1** in the respective iterations to get the example significant biomarkers set: ‘Angiopoietin-2’, ‘CA-125’, ‘CA 15-3’, ‘CEA’, ‘CYFRA 21-1’, ‘FGF2’, ‘G-CSF’, ‘HE4’, ‘HGF’, ‘PAR’, ‘sPECAM-1’, ‘Thrombospondin-2’.

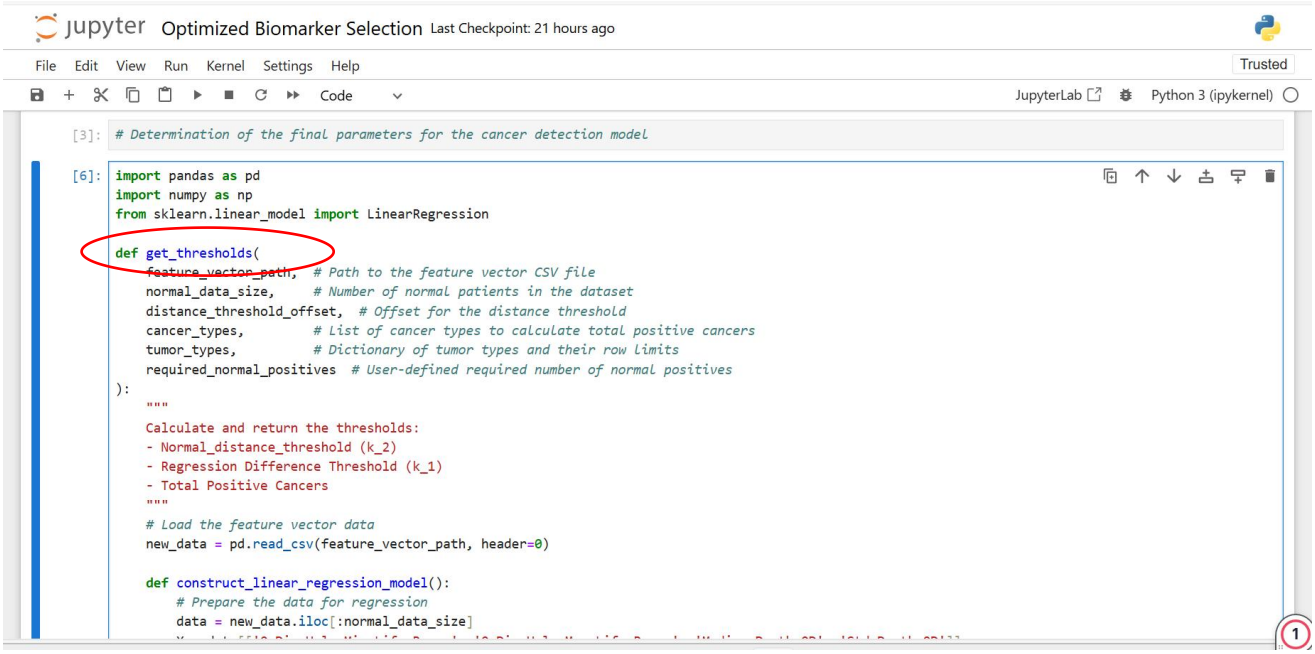
Biomarker Selection Process



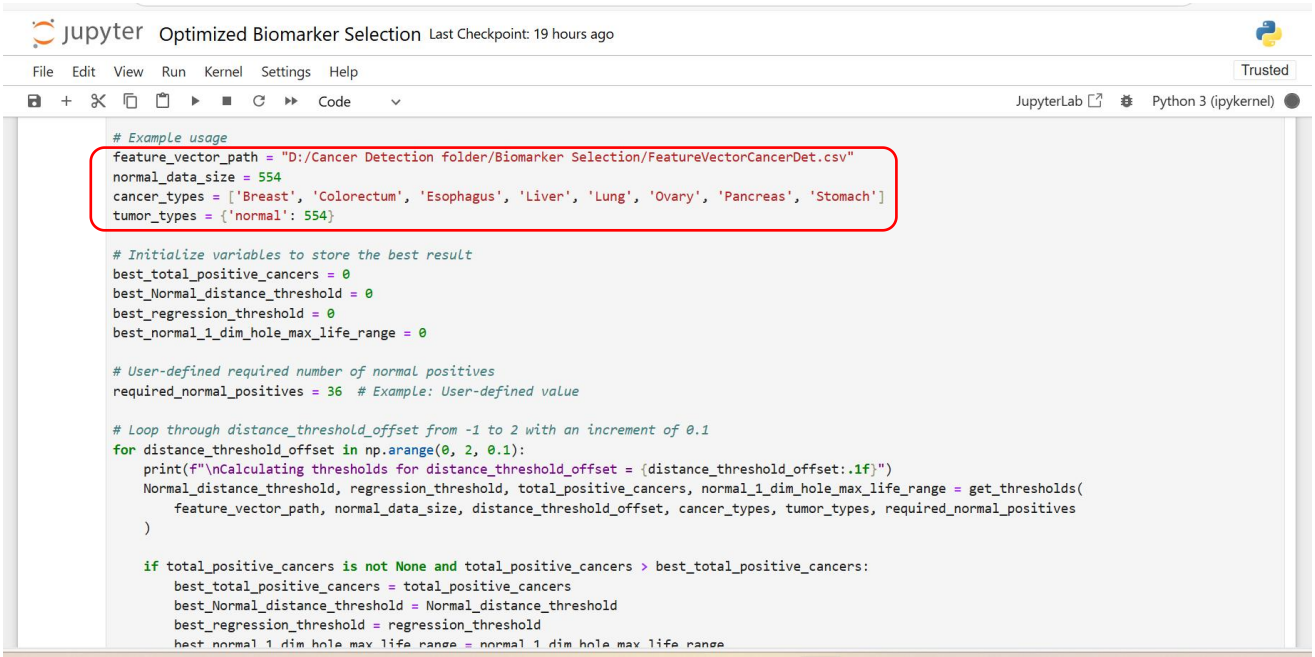
24. **Final Significant Biomarkers set:** ‘Angiopoietin-2’, ‘CA-125’, ‘CA 15-3’, ‘CEA’, ‘CYFRA 21-1’, ‘FGF2’, ‘G-CSF’, ‘HE4’, ‘HGF’, ‘PAR’, ‘sPECAM-1’, ‘Thrombospondin-2’.

Threshold Determination

25. Enter the same inputs for feature_vector_path, Selection/FeatureVectorCancerDet.csv", normal_data_size, cancer_types and tumor_types in the ***get_thresholds*** function.

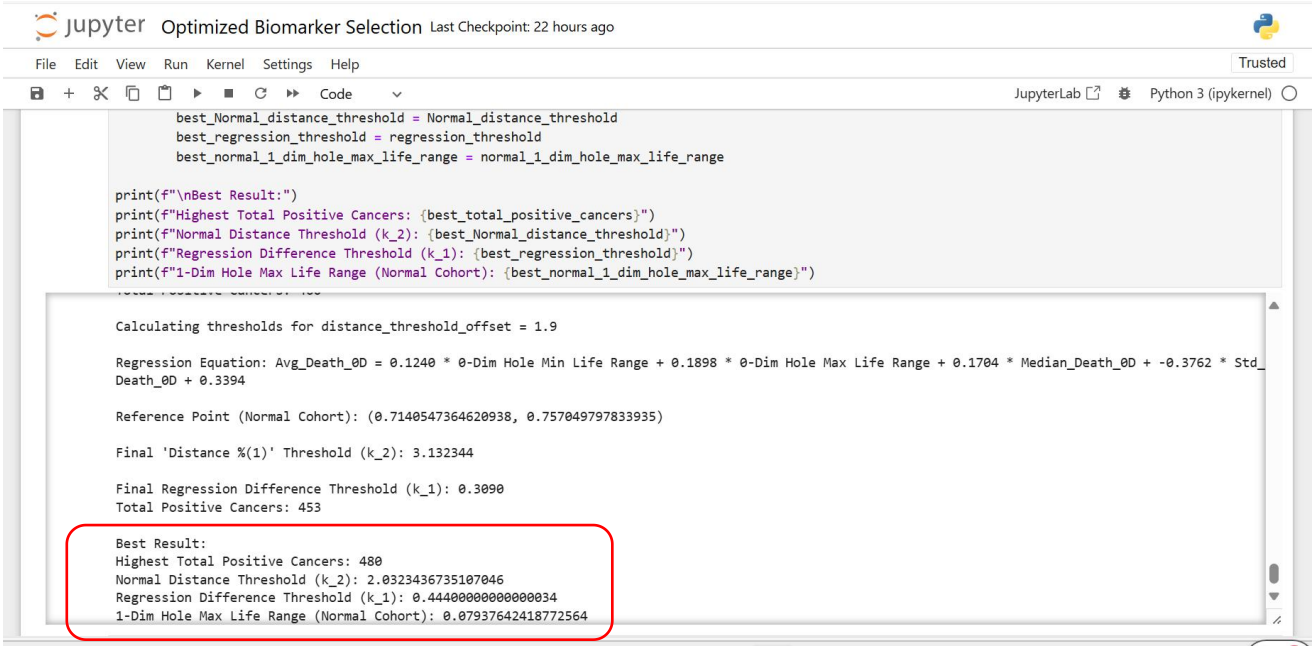


26. Run the function ***get_thresholds***.



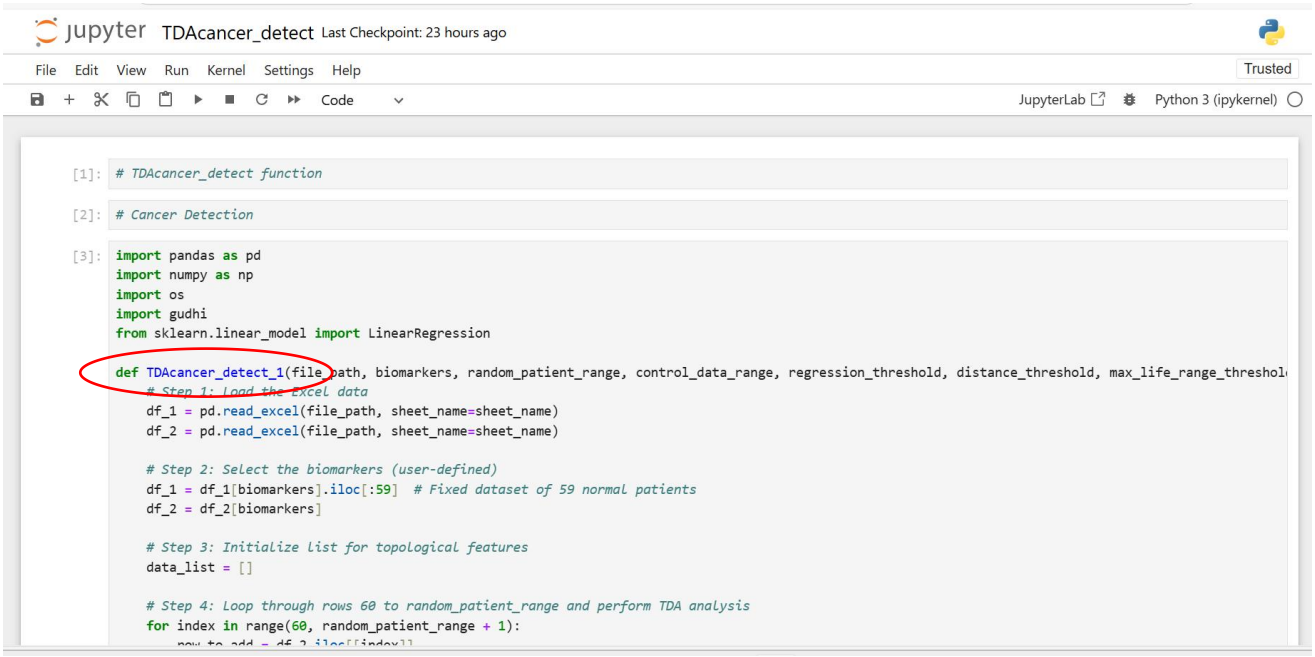
Threshold Determination

27. The thresholds to maximize cancer detection sensitivity for the selected biomarker set will be obtained.



Cancer Detection

28. Load *"TDAcancer_detect.ipynb"* in a Jupyter Notebook.



29. Enter the required inputs to the function ***TDAcancer_detect_1***. The inputs' descriptions are given below.

- **file_path:** Path to the file containing the Excel file **'Clinical cancer data.xlsx'**.
- **sheet_name:** Input the patients' data sheet 'Normal and Cancer' from the provided Excel Data File **'Clinical cancer data.xlsx'**.
- **biomarkers:** Select the biomarkers obtained from the ***Optimized_Biomarker_Selection*** algorithm (Step 24).
- **random_patient_range:** Specify the total number of patients in the dataset.
- **control_data_range:** Provide the number of healthy individuals to use as training samples.
- **tumor_types:** List the tumor types present in the dataset.

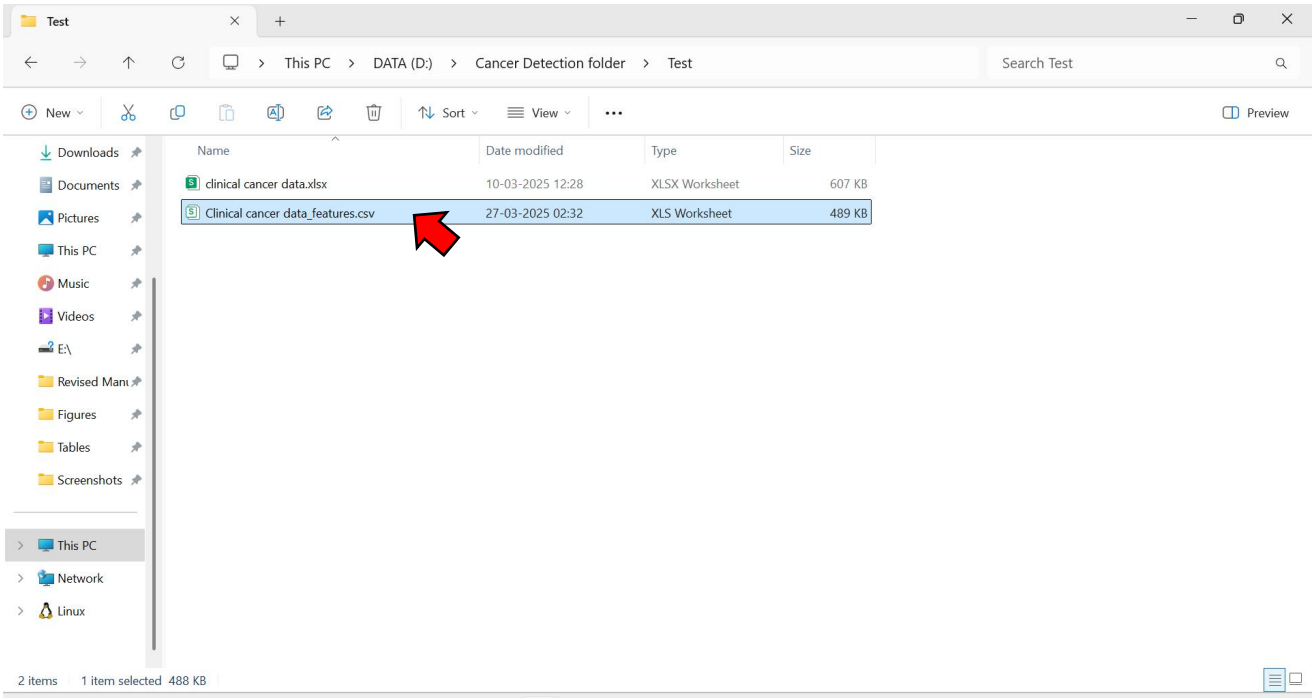
30. Enter the parameters **k_1** and **k_2** obtained from the ***get_thresholds*** algorithm.
31. After entering **k_1** and **k_2**, enter the parameters **k_3** and **k_4**. Select **k_3** and **k_4** such that both have the same distance from the value "**1-Dim Hole Max Life Range (Normal Cohort)**" obtained in the ***get_thresholds*** algorithm. Additionally, in selecting **k_3** and **k_4**, maintain the "Normal Positive" (False Cancer Positives) to ensure specificity is at least over 95%.

Note: For the example file **"Clinical cancer data.xlsx"**, the user must provide **file_path**. The other inputs are set as default. The inputs are case-sensitive, so refer to the example inputs and ensure that your inputs match the given case exactly.

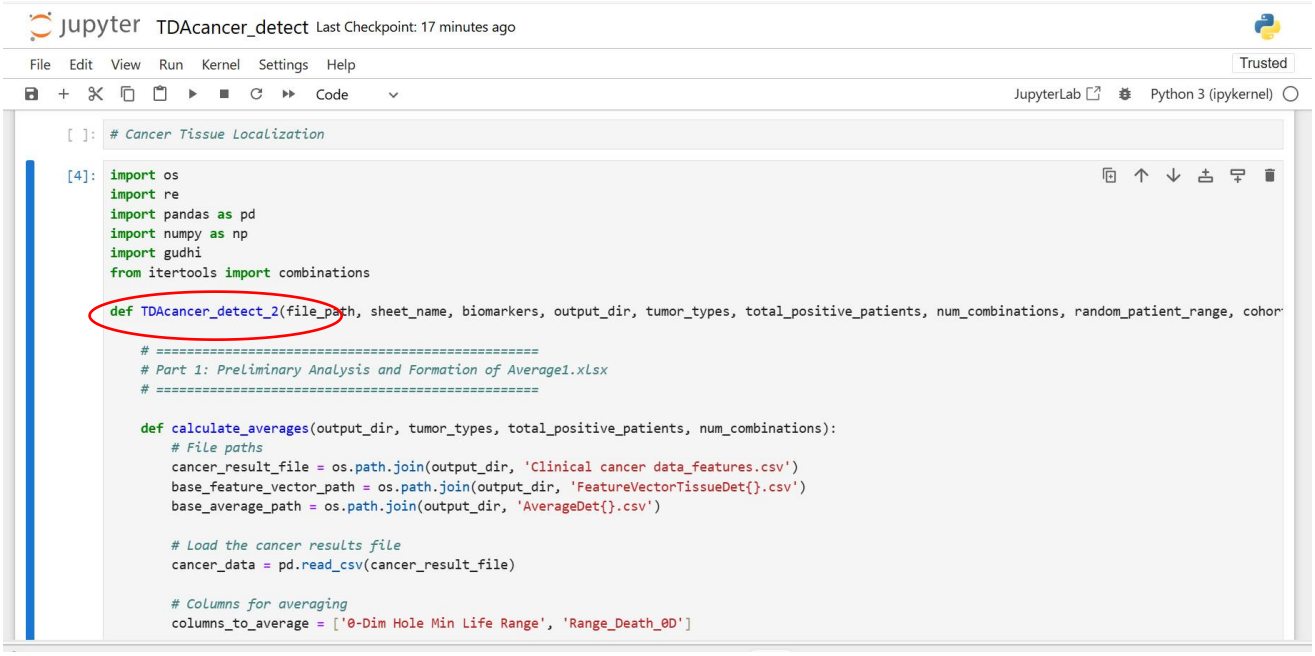
Cancer Detection

32. Run the function ***TDAcancer_detect_1***.

33. A CSV file **"Clinical cancer data_features.csv"** will be generated in the given "**file_path**." This file contains the cancer detection results (Positive or Negative) for each individual patient.



Cancer Tissue Localization



34. Enter the required inputs to the function ***TDAcancer_detect_2***. The inputs' descriptions are given below.

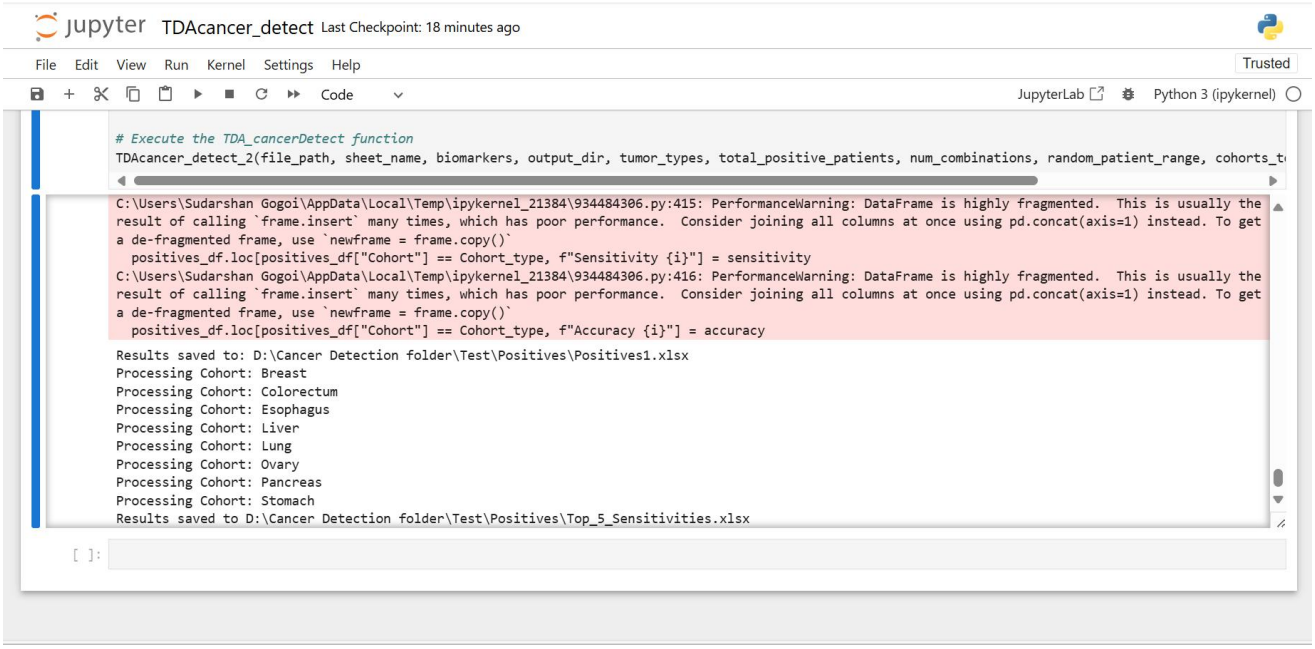
- **file_path:** Path to the file containing the Excel file **'Clinical cancer data.xlsx'**.
- **sheet_name:** Input the patients' data sheet 'Normal and Cancer' from the provided Excel Data File **'Clinical cancer data.xlsx'**.
- **biomarkers:** Select the biomarkers.
- **output_dir:** Path to store the result files.
- **tumor_types:** List the tumor types present in the dataset along with the number of patients to use as training samples. The sample numbers should be selected from **true negative** healthy individuals, and for cancer tumors, from **true positives** obtained from ***TDAcancer_detect_1***.
- **total_positive_patients:** List the cancer types present in the dataset along with the number of **cancer-positive patients** identified by ***TDAcancer_detect_1***.
- **cohorts_to_analyze:** List the cancer types present in the dataset.
- **num_combinations:** Provide the number of biomarker combinations user want to analyze.
- **random_patient_range:** Specify the total number of patients in the dataset.

Note:** For the example file **"Clinical cancer data.xlsx"**, the user must provide **file_path** and **output_dir**. The user should set the same input and output file paths for the functions ***TDAcancer_detect_1 and ***TDAcancer_detect_2*** because the files used in these functions are interconnected. The other inputs are set as default. The inputs are case-sensitive, so refer to the example inputs and ensure that your inputs match the given case exactly.

Cancer Tissue Localization

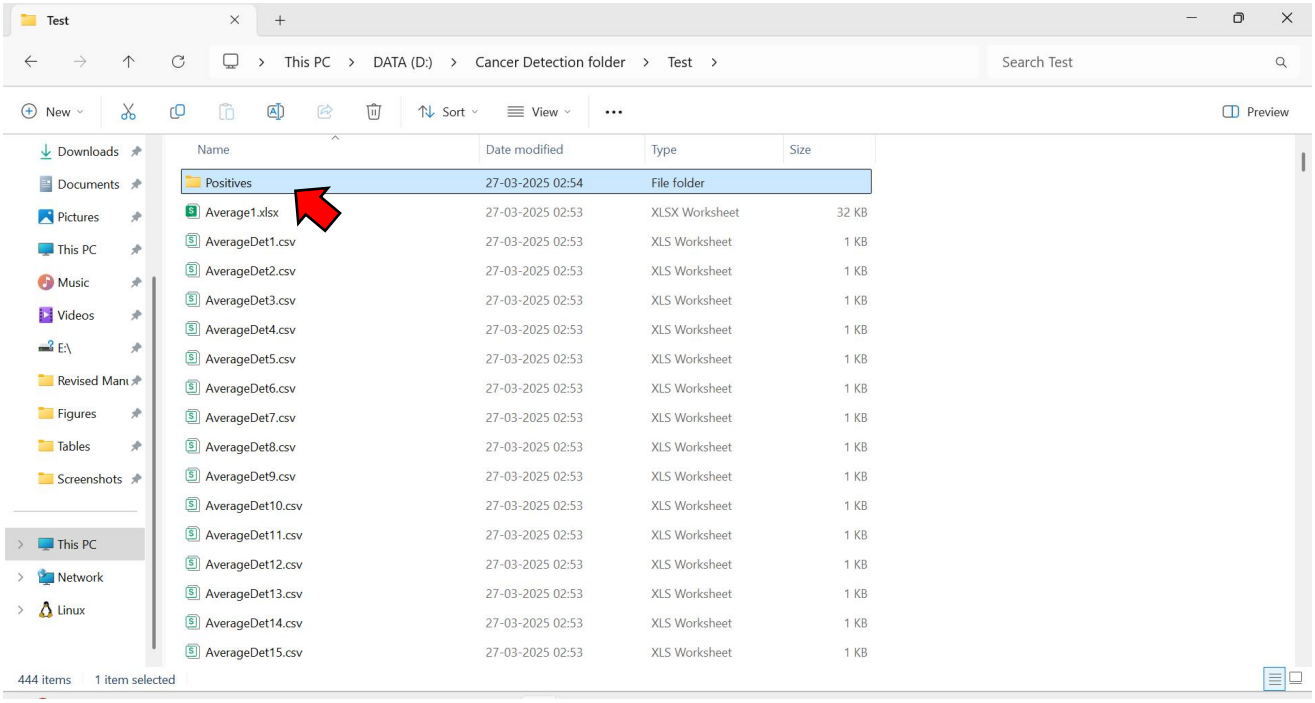
35. Run the function ***TDAcancer_detect_2.***

36. The result files will be saved in the given **output_dir**.

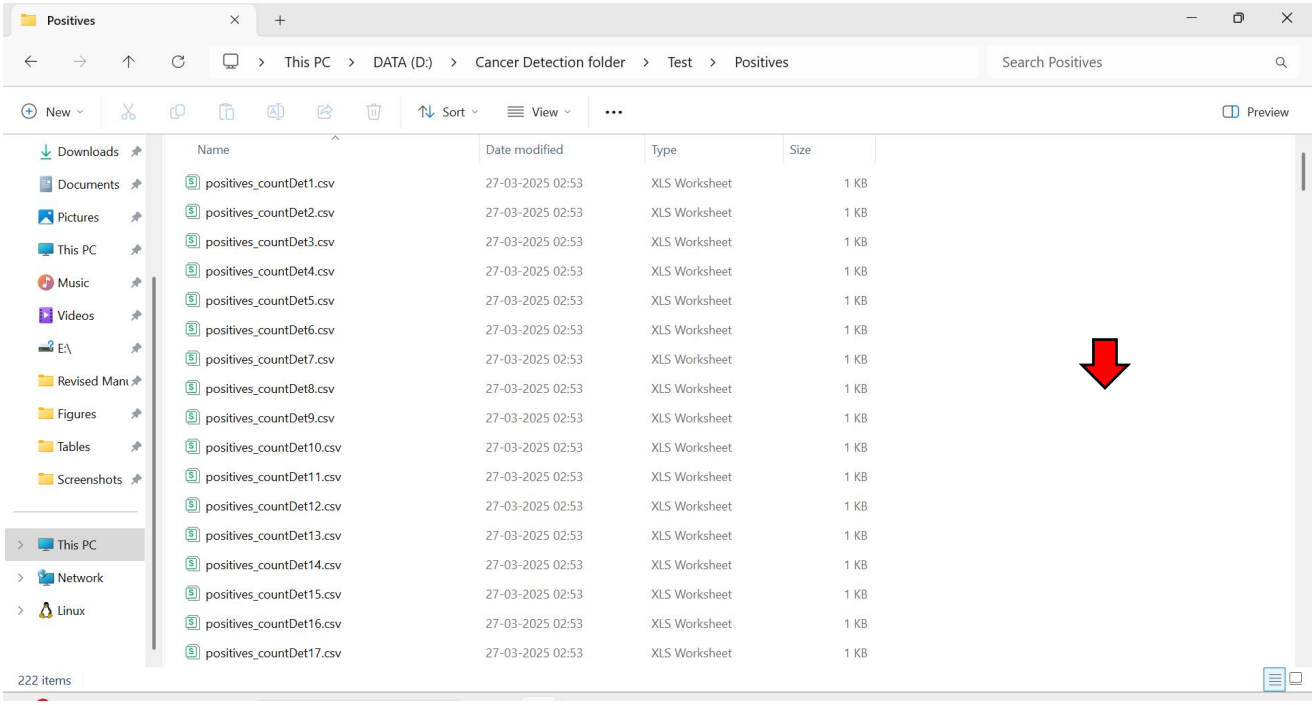


Cancer Tissue Localization

37. Open the **output_dir**.
38. Open the folder **“Positives”**

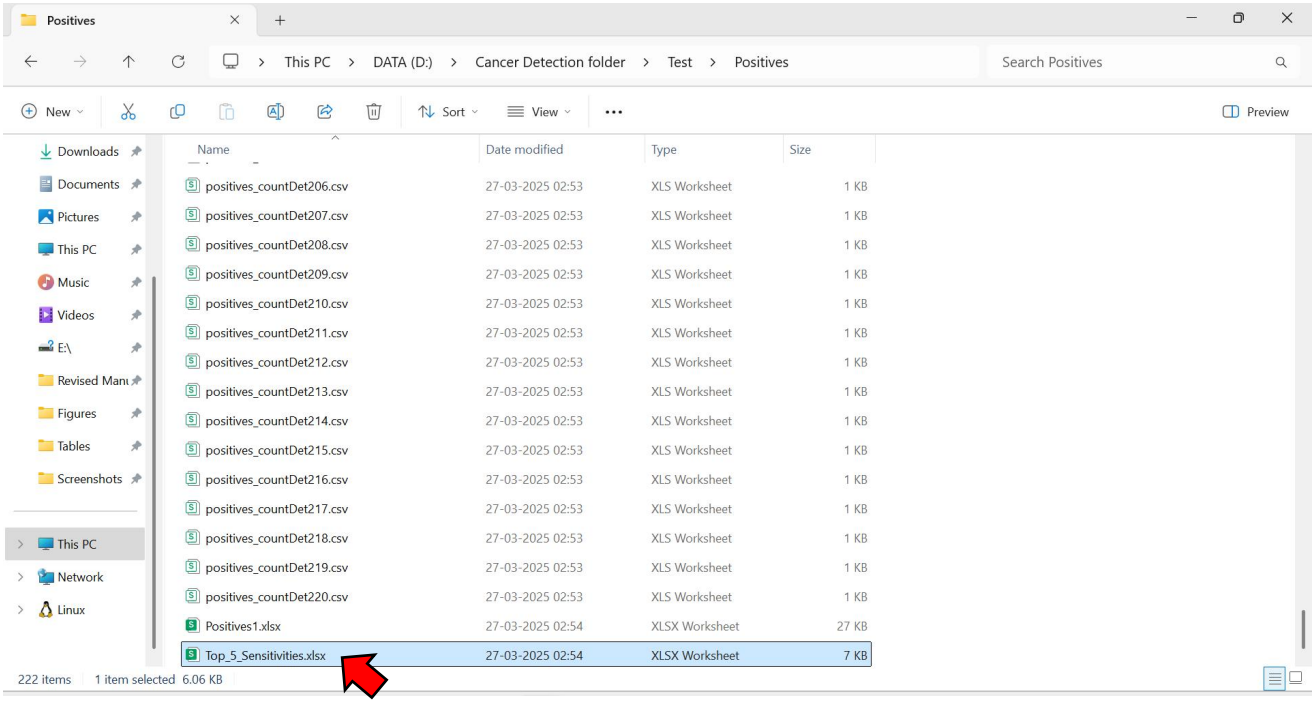


39. Scroll down to the last file.

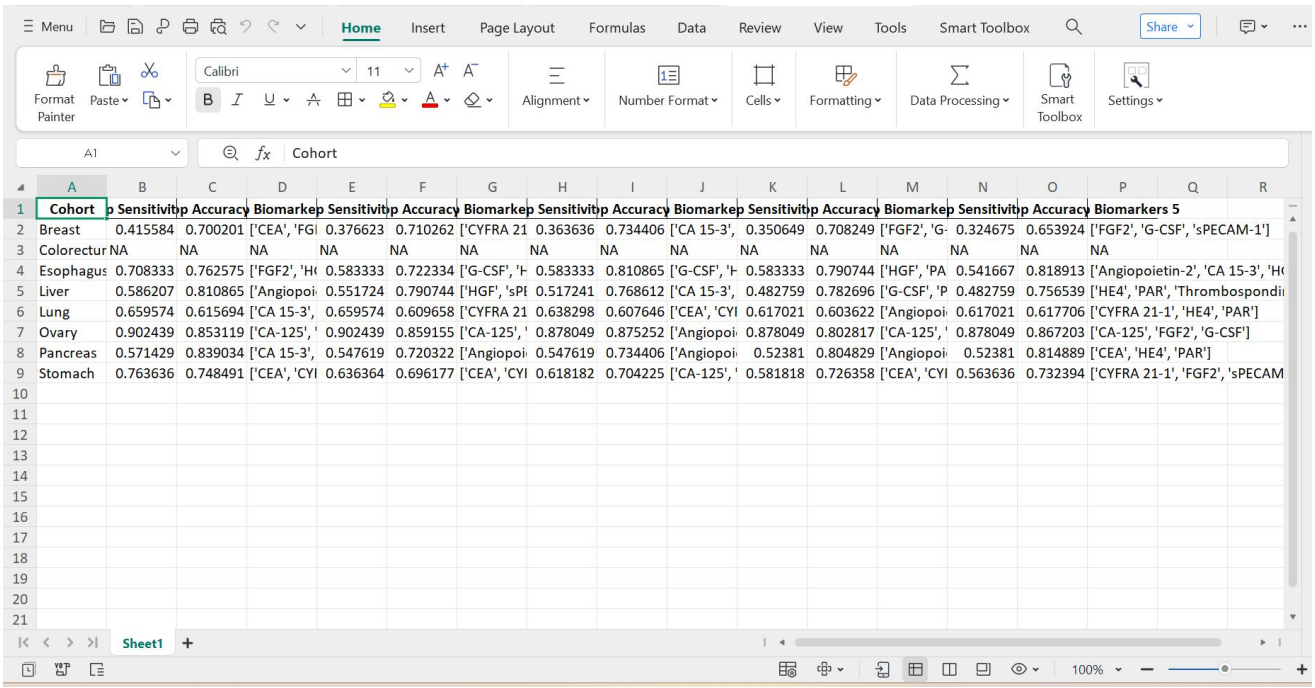


Cancer Tissue Localization

40. Open **Top_5_Sensitivities.xlsx** file



41. Tissue localization sensitivity and accuracy for different biomarkers will be obtained.



Cohort	p Sensitivity	p Accuracy	Biomarkers	p Sensitivity	p Accuracy	Biomarkers	p Sensitivity	p Accuracy	Biomarkers	p Sensitivity	p Accuracy	Biomarkers	p Sensitivity	p Accuracy	Biomarkers	p Sensitivity	p Accuracy	Biomarkers 5
Breast	0.415584	0.700201	['CEA', 'FGI	0.376623	0.710262	['CYFRA 21	0.363636	0.734406	['CA 15-3',	0.350649	0.708249	['FGF2', 'G-	0.324675	0.653924	['FGF2', 'G-CSF', 'sPECAM-1']			
Colorectur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA			
Esophagus	0.708333	0.762575	['FGF2', 'H(	0.583333	0.722334	['G-CSF', 'H	0.583333	0.810865	['G-CSF', 'H	0.583333	0.790744	['HGF', 'PA	0.541667	0.818913	['Angiopoietin-2', 'CA 15-3', 'H(			
Liver	0.586207	0.810865	['Angiopoi	0.551724	0.790744	['HGF', 'sPI	0.517241	0.768612	['CA 15-3',	0.482759	0.782696	['G-CSF', 'P	0.482759	0.756539	['HE4', 'PAR', 'Thrombospondi			
Lung	0.659574	0.615694	['CA 15-3',	0.659574	0.609658	['CYFRA 21	0.638298	0.607646	['CEA', 'CYI	0.617021	0.603622	['Angiopoi	0.617021	0.617706	['CYFRA 21-1', 'HE4', 'PAR']			
Ovary	0.902439	0.853119	['CA-125', '	0.902439	0.859155	['CA-125', '	0.878049	0.875252	['Angiopoi	0.878049	0.802817	['CA-125', '	0.878049	0.867203	['CA-125', 'FGF2', 'G-CSF']			
Pancreas	0.571429	0.839034	['CA 15-3',	0.547619	0.720322	['Angiopoi	0.547619	0.734406	['Angiopoi	0.52381	0.804829	['Angiopoi	0.52381	0.814889	['CEA', 'HE4', 'PAR']			
Stomach	0.763636	0.748491	['CEA', 'CYI	0.636364	0.696177	['CEA', 'CYI	0.618182	0.704225	['CA-125', '	0.581818	0.726358	['CEA', 'CYI	0.563636	0.732394	['CYFRA 21-1', 'FGF2', 'sPECAM			