**Step by step guidelines to Run the AutoNSGACytoNet framework:**

1. **Collect data from** [**https://xenabrowser.net/datapages/**](https://xenabrowser.net/datapages/)**:**

**Collect GDC TCGA gene expression RNAseq TPM data for BRCA, LAML, LUAD, LUSC, COAD, SKCM, GBM and LIHC Pancancer Cohorts.**

**Already downloaded data can be found at:** [**Dataset**](https://drive.google.com/drive/folders/1_G0InAE3ZyQYqjS77BW0rwSxtElYzidY?usp=sharing)

1. **The** [**Dataset**](https://drive.google.com/drive/folders/1_G0InAE3ZyQYqjS77BW0rwSxtElYzidY?usp=sharing) **folder contains directory wise carcinogenic Data. Here Each directory is named after each cancer type and each directory contains .tsv format files for each cancer type. These initial tsv files need to be Transposed so that we can use the genes as feature and merged together to result into a final Dataset. Using the** [**merging**](https://drive.google.com/file/d/1Yy8-W23SQJDc_C8JQhyKhbn5U8aJvaXn/view?usp=drive_link) **script the final dataset can be achieved.**

**PS: Step 2 has one directory related dependency. Need to set the directory/path of** [**Dataset**](https://drive.google.com/drive/folders/1_G0InAE3ZyQYqjS77BW0rwSxtElYzidY?usp=sharing) **manually to run the code.**

1. **Step 3 is the preprocessing step. Data in** [**file**](https://drive.google.com/file/d/1glUzz2QOmWGvGp-pBvU1IOp39dXWBqwu/view?usp=sharing) **can be used or it can be generated using the** [**merging**](https://drive.google.com/file/d/1Yy8-W23SQJDc_C8JQhyKhbn5U8aJvaXn/view?usp=drive_link) **script. The notebook** [**Preprocessing Data**](https://drive.google.com/file/d/1Fd747Ooqt-zHtepRZeCsFxXas1Il-rzy/view?usp=sharing) **can be used to generate preprocessed CSV file. It also has director/path dependency. Need to set the path manually. The outcome is already stored as** [**Preprocessed\_8\_cancer\_genes.csv**](https://drive.google.com/file/d/1Ovg0b5nvPIES6SUPgzWe4QozruM38n13/view?usp=sharing) **in google drive.**
2. **Step 4 is the autoencoder based feature selection step. Here the input is as** [**Preprocessed\_8\_cancer\_genes.csv**](https://drive.google.com/file/d/1Ovg0b5nvPIES6SUPgzWe4QozruM38n13/view?usp=sharing) **file. After running the notebook** [**autoencoders-with-cv.ipynb**](https://colab.research.google.com/drive/1ngGEXxiaD62SclrGDxyzSgXov4KkFGK1?usp=sharing) **we will get out csv files** [**top\_0.5\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1yxWPcWSBHLQY3UbmYiyW6pYGf2YahVEO/view?usp=sharing)**,** [**top\_0.25\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1gGwXLi4Y1k0h1JXm4-QVUHwobOXioO5y/view?usp=sharing)**,** [**top\_1.0\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1whbaizfpgp4G_Dm3e5KTxl38cE3yDnG0/view?usp=sharing)**. Again, this notebook has directory/path dependency. Need to change accordingly.**
3. **Next step is the NSGA-2 Step. This step is implemented in** [**nsga2-with-rf-cv2.ipynb**](https://colab.research.google.com/drive/1N9eExHQGhmoPU9omu7xJbhlOgIHhen2_?usp=sharing) **notebook. It will take these files (**[**Preprocessed\_8\_cancer\_genes.csv**](https://drive.google.com/file/d/1Ovg0b5nvPIES6SUPgzWe4QozruM38n13/view?usp=sharing)**,** [**top\_0.5\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1yxWPcWSBHLQY3UbmYiyW6pYGf2YahVEO/view?usp=sharing)**,** [**top\_0.25\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1gGwXLi4Y1k0h1JXm4-QVUHwobOXioO5y/view?usp=sharing)**,** [**top\_1.0\_percent\_features\_cv.csv**](https://drive.google.com/file/d/1whbaizfpgp4G_Dm3e5KTxl38cE3yDnG0/view?usp=sharing)**) as input. Need to set the path manually here as well.**
4. **After running the notebook** [**nsga2-with-rf-cv2.ipynb**](https://colab.research.google.com/drive/1N9eExHQGhmoPU9omu7xJbhlOgIHhen2_?usp=sharing)**. We will get 3 output files** [**NSGA2\_77\_compression\_1.csv**](https://drive.google.com/file/d/1c6bkbqN6RaL3PPIpuKsvv64dIHGKpec1/view?usp=sharing)**,** [**NSGA2\_308\_compression\_3.csv**](https://drive.google.com/file/d/1i2ofm8YWFhzgU3e9AkYAVS4DeaPATPNs/view?usp=sharing)**,** [**NSGA2\_154\_compression\_2.csv**](https://drive.google.com/file/d/18e1QAcHpVtAp_vgz9h1arRAqoga2f3Rm/view?usp=sharing) **with gene subsets.**
5. **At step 7, we wil need the gene subset from** [**NSGA2\_308\_compression\_3.csv**](https://drive.google.com/file/d/1i2ofm8YWFhzgU3e9AkYAVS4DeaPATPNs/view?usp=sharing) **file. We will uplpad the gene set in** [**STRING database**](https://string-db.org/cgi/network?taskId=bbu2ZvRC7BLz&sessionId=bH4thSSRvOus) **to generate Protein-Protein-Interaction PPI network. The network is needed to be downloaded in .tsv format. Which is also available in google drive as** [**string\_interactions.tsv**](https://drive.google.com/file/d/1PrCtrxiP-19T3GIoDMH3MbKDcBmiNNYs/view?usp=sharing)**.**
6. **At step 8 we need to upload the** [**string\_interactions.tsv**](https://drive.google.com/file/d/1PrCtrxiP-19T3GIoDMH3MbKDcBmiNNYs/view?usp=sharing) **network in Cytoscape software with cytohubba dependency. It is an open source tool. At this step we will select 13 hub genes. These genes will be used for evaluation purpose.**
7. **This is the final and evaluation step. Here we need to run the** [**classification.ipynb**](https://colab.research.google.com/drive/1xgvRDGgkYmXV7rPQSJDXT1a3gG-x_njx?usp=sharing) **notebook. This notebook requires** [**Preprocessed\_8\_cancer\_genes.csv**](https://drive.google.com/file/d/1Ovg0b5nvPIES6SUPgzWe4QozruM38n13/view?usp=sharing) **Dataset. The gene subset acquired from Cytoscape is already hardcoded in the notebook. After running the notebook we will get desired evaluation metrics.**