How to use EBST?

Start: Run the MATLAB software. Then in the home menu, select the set path. In the popup window, select the option "add with subfolders" and select the unzipped EBST folder and save. Then run the **mainRUN.m** file.

Note: You can run standalone software without need for MATLAB from the compiled folder. Then, run the EBST.exe file.

1. First Page Descriptions

When you run the software, the application's home screen will appear as follows. The different sections of the home page of the software are numbered according to the Figure 1 which we will explain in the following sections.

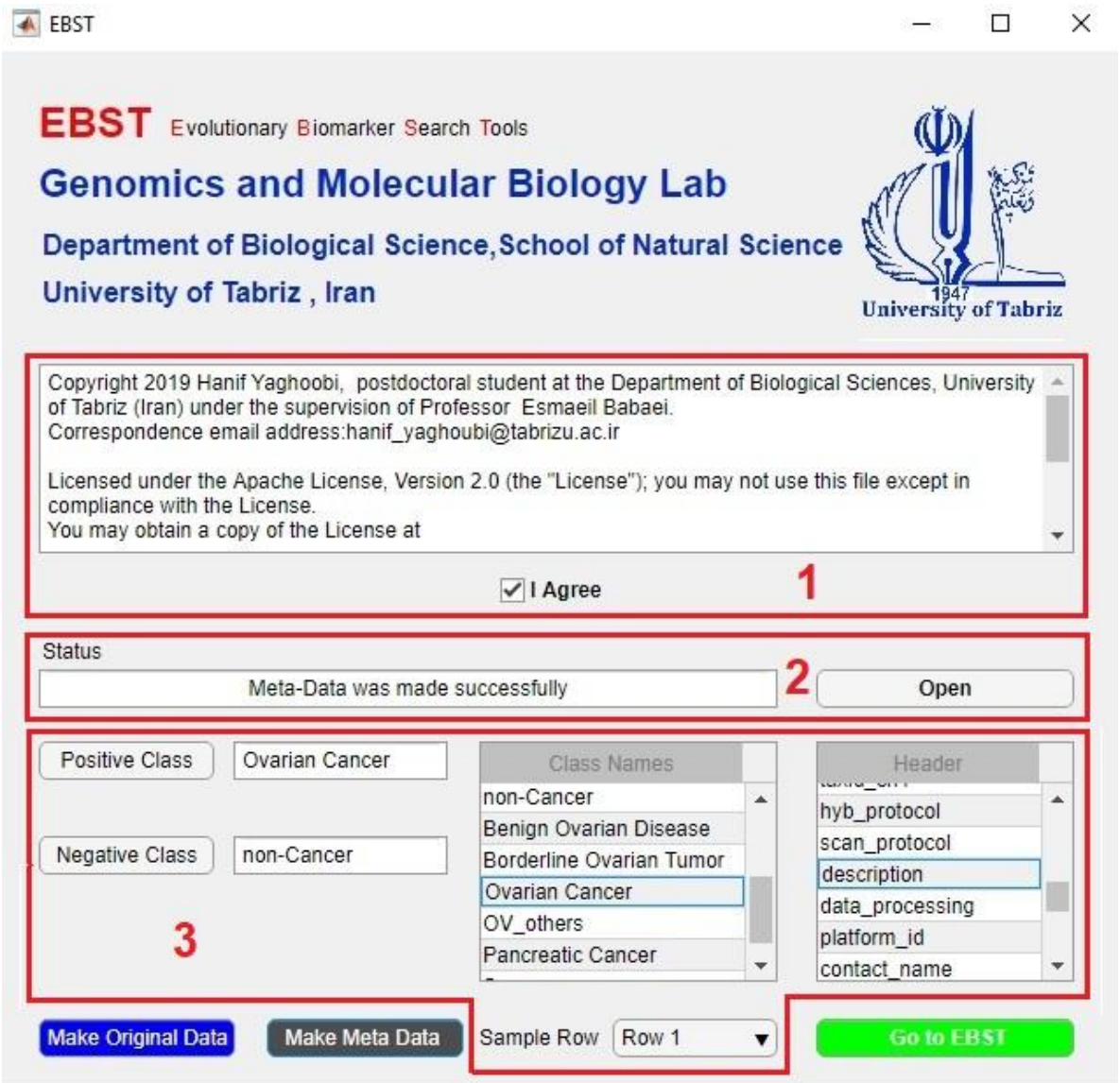


Figure 1: First Page Environment

1. This section is about Software Licensing. You should read this section and check the "**I Agree**" option if you agree.

2. In this section, you can import GEO Series Data that you downloaded from the GEO Dataset database at https://www.ncbi.nlm.nih.gov/gds. Each Dataset with an access number is available as GSExxx or other forms on this website. This Series Data is usually in .txt format and usually contains the main data plus header data that contains information such as Description of Samples, Organism, Platform id, Geo accession number of samples, Row names, Column names and more. Use the **"Open"** button to select the downloaded file from the saved path. Wait a little until the file extraction process is finished. If the operation is successful, the address of the file is displayed in the Status bar.
3. If the file loading operation is performed correctly, in the **Header** table the titles of the various information of dataset are displayed. By default, the contents of the header information associated with the "Description" appear in the **Class Names** tab. At this point, you should look for a header that contains information about the class labels. Usually this information is in “Characteristics_ch1” or “Description”, and so on. Then you have to use the **Class Names** table to select a case as a positive class and select an item as a negative class, and in each case click the appropriate button. The labels of the positive and negative classes will appear in front of the buttons.

   Note: The label information for classes may not be in line 1 of the selected item from header. Then, with the Sample Row option, see the next rows.
4. Now you can press the **Make Original Data** button. If you want to use meta-analysis, you will need to perform steps 1 through 3 for another data, and then click the **Make Meta Data** button. The purpose of the meta analysis is to examine the generalizability of the results of the original data on another. This means that the biomarkers selected by the main data are also generalized in other similar data. Finally, click **Go to EBST** to enter the main application environment.

2. Main Page Descriptions

By pressing the Go to EBST button you enter the main environment as shown in Figure 2. The different parts of this environment are numbered and we will describe them in numerical order, respectively.

1. If you have not been able to enter your desired data from the first page, this section will be useful to you. This section is for entering the Data matrix, Label matrix and Gene names. You must separate the data matrix from the original database downloaded from the NCBI site. In the common form of data, Rows represent genes (attributes) and columns represent samples. This matrix should be in CSV format. For example, you can import the **Data.csv** file from the **results & data** folder. The label matrix is composed of 0 and 1 elements, such as 0 for normal samples and 1 for patient samples. For example, you can enter the **Lable.csv** file from the **results & data** folder. In the Gene Name field, you can enter the names of the genes as *.**csv** file. For example, you can import the **idn.csv** file from the **results & data** folder. You can also import all the necessary data and analysis on them into the Mat File Load section if you have previously saved them in *.mat format. This is explained in Section 7.

Note: When you enter data on the first page, the names of the genes are read from the series matrix file, which may be the actual name of the genes or probe id. Therefore, it may be necessary to re-enter the gene names file here.

2. This section deals with pre-processing the original data and meta data. The processes performed here are:

a) **Correct Imbalanced Data**: By checking this option, the number of samples in the two classes will be equal. This is done by random sampling of the more member class.

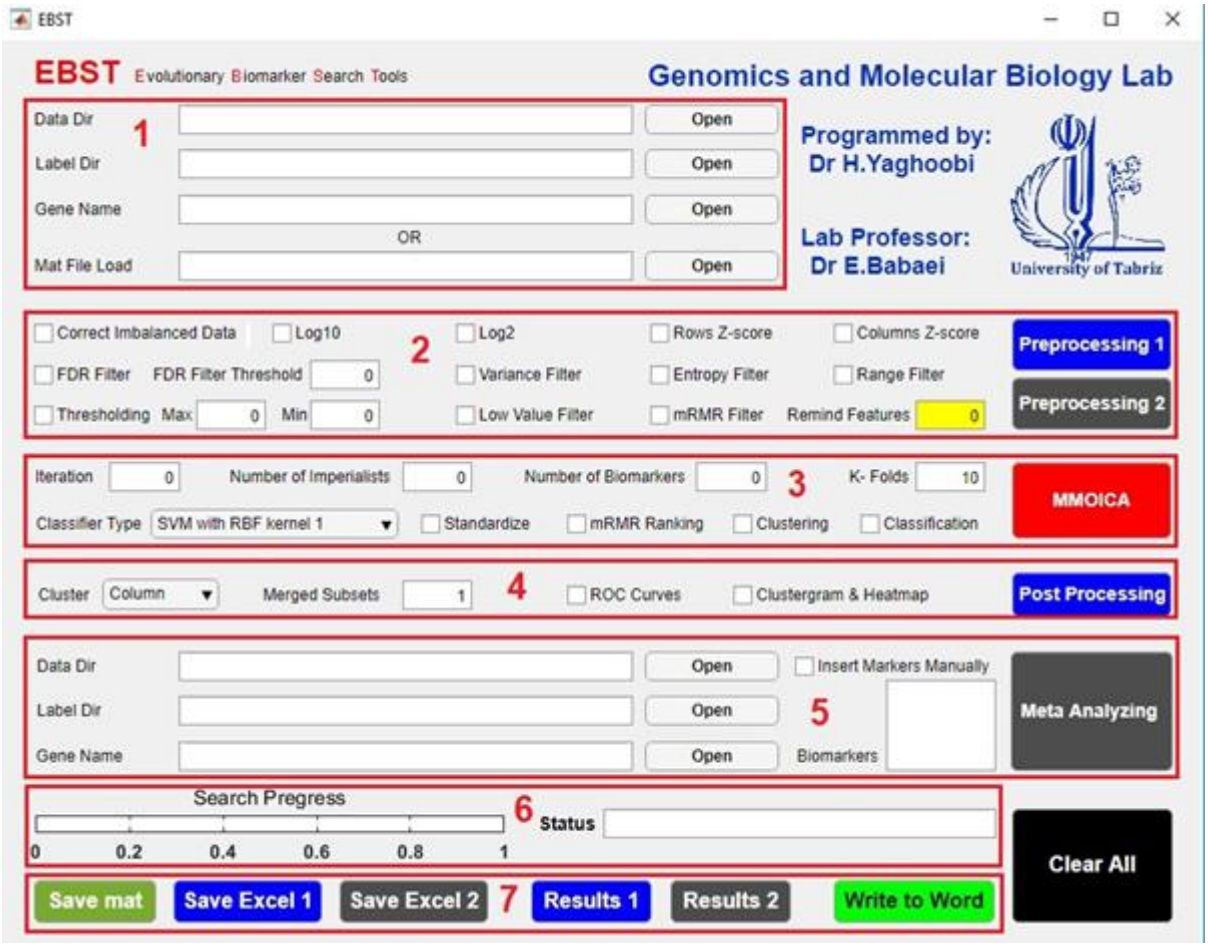


Figure 2: Main Page Environment

b) **Log10** and **Log2**: By selecting these options the logarithm data is extracted on a 10 or 2 basis.
c) **Rows Z-Score** and **Column Z-Score**: By selecting these options, the rows or columns of the data matrix are normalized by the Z-Score method.
d) **Thresholding**: This option sets the data between the maximum and the minimum thresholds so that data smaller than **Min** and larger than **Max** are written to **Min** and **Max** respectively.
e) **Variance, Entropy, Range** and **Low Value Filters**: These options filter the attributes based on their name. See the MATLAB Bioinformatics toolbox for further reading.
f) **FDR Filter** and **Threshold**: These options are for filtering features using the Fisher Discriminant Ratio (FDR) method. Selecting the FDR filter will remove features (here genes) that are less than the FDR threshold. Larger thresholds allow fewer features to remain.
g) **mRMR**: This option ranks the remaining features based on the Minimum Redundancy and Maximum Relevance (mRMR) criterion.

Note: The processes are executed in the alphabetical order mentioned above.

h) **Preprocessing1** Button: By pressing this button, the above processes are performed for the original data.
i) **Preprocessing2** Button: By pressing this button, the above processes are performed for the meta data.

3. This is the core part of Evolutionary Biomarker Search Tool (EBST), which is the Modified Multi-Objective Imperialist Competitive Algorithm (MMOICA) settings and its implementation to search for biomarkers. The settings for this section are:

a) **Iteration**: Specifies the iteration of the search algorithm.
b) **Number of Imperialist**: Determines the number of imperialists. Read our original article for further reading.
c) **Number of Biomarker**: Determines the number of optimal biomarkers. Note that the number of biomarkers should be selected in proportion to the number of class samples. This is explained in our original article.
d) **K-Folds**: Specifies the number of cross validation folds in the classification if the search objective function is based on the classifier.
e) **Classification**: If this option is checked, the objective functions of the search algorithm will be based on the classifier. These objective functions include External cross validation error, Internal cross validation error, Sensitivity and Edge of classifier.
f) **Classifier Type**: In relation to the **Classification** option, this option selects the Classifier type.
g) **Standardize**: This option is also associated with Classification and by selecting this option the data will be standardized before classification.
h) **mRMR Ranking**: This option defines a new objective function based on the mRMR ranking. This is the Normalized Average Rank (*mRMR.N.A.R*) of the selected features described in the original article.
i) **Clustering**: This option selects the objective function of the cluster type. This clustering is of the type k-means with k = 2 and the criterion of similarity is the angle between two vectors.

Note: Any combination of the mentioned objective functions is possible. For example, six objective functions can be defined, four related to classification criteria, one related to mRMR rank and the other to clustering.

j) **MMOICA** Button: The search algorithm is executed by pressing this button. When the search is completed, the results will be displayed on a separate page. This page contains information about the data as well as the values of the target functions and selected attributes. An example of these results is shown in Figure 3.

4. This section deals with analyzing the results of the search algorithm. Here is the name of the biomarkers with cluster-gram & Heat-map and ROC curves that we have called post processing.

a) **Post Processing** Button: Pressing this button will display a results page, such as the one you see in Figure 4. The results page contains three table windows. The top left and bottom tables both show the results of classifying data by several classifiers with selected biomarkers. The upper left table shows the results of external and internal cross validation error and classification edge for each classifier. The bottom table shows the accuracy, sensitivity, specificity, positive predictive value (PPV) and area under the ROC curves (AUC) for each classifier. The table on the top right shows the selected biomarkers along with the signal-to-noise ratio (SNR), their FDR score, their mRMR rank, and the class correlation according to the negative / positive SNR. If you want to know whether a research or article on selected biomarkers has already been published, you can choose the biomarker you want and one of the classes in the class correlation column. Then select one of the NCBI or PMC options or both and browse directly to the selected databases.
b) **Merged Subsets**: In MMOICA, like other multi-objective optimization algorithms, all optimal pareto solutions can be the solution to the problem. However, our algorithm sorts the set of

solutions based on the least internal and external cross validation errors. This option specifies the number of sets of solutions that can be merged. Note that solution sets may have common features, in which case only one common feature is considered.

c) **Cluster-gram & Heat map**: Selecting this option adds the heat map and cluster gram of the selected biomarkers to the post processing results. An example of the result is shown in Figure 5.

d) **Cluster**: This option allows you to specify the type of clusters in the cluster gram so that if you want to cluster genes, the column option is selected and the row option is selected if you want to cluster the samples. The point is that it is possible to cluster both.

e) **ROC Curves**: Selecting this option adds ROC curves to the post processing results. These curves are plotted for all classifiers shown in the post processing results page. An example of these curves is shown in Figure 6.

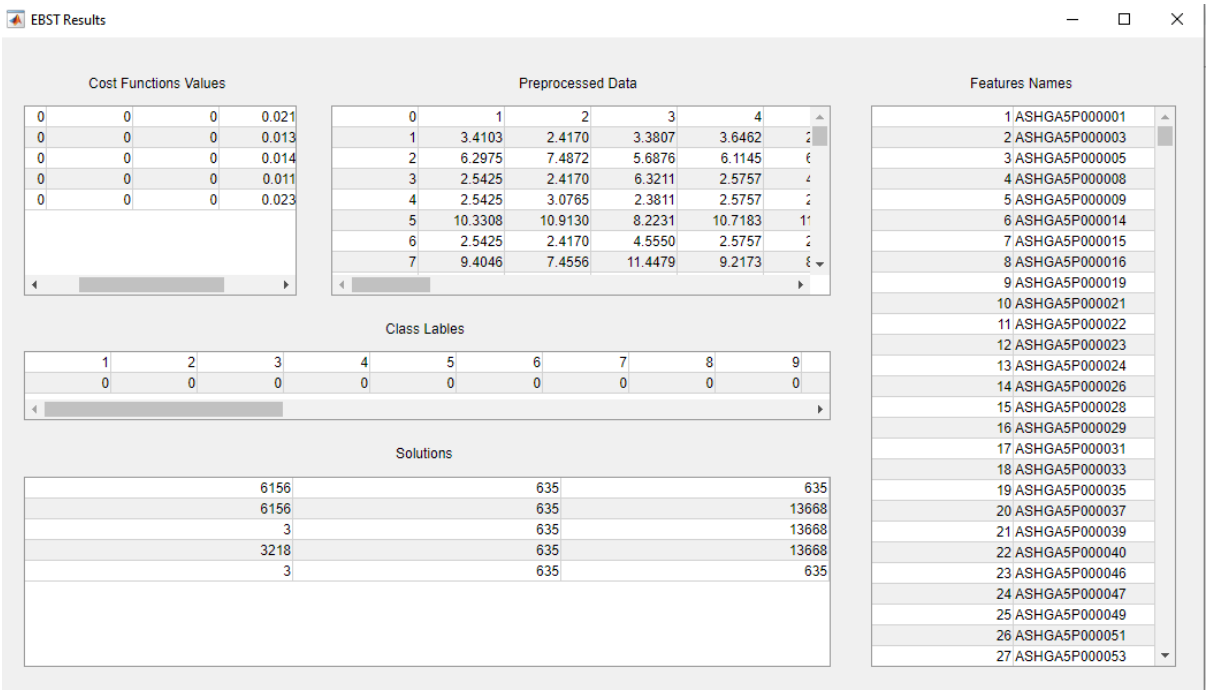
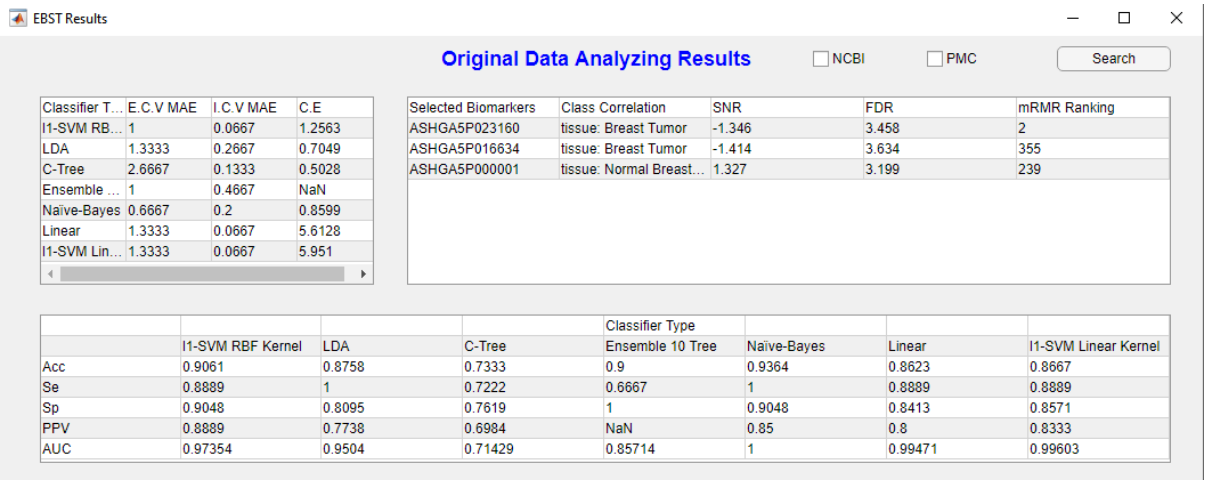


Figure 3: Example of MMOICA results page

5. This section examines the generalizability of selected features from the original data to other data previously entered into the software as meta data. The purpose of meta-analysis is to investigate whether the biomarkers selected from the original data can be biomarkers in the meta data as well. So here we have results similar to the post processing section for meta data.



Classifier T...	E.C.V MAE	I.C.V MAE	C.E
I1-SVM RB...	1	0.0667	1.2563
LDA	1.3333	0.2667	0.7049
C-Tree	2.6667	0.1333	0.5028
Ensemble ...	1	0.4667	NaN
Naïve-Bayes	0.6667	0.2	0.8599
Linear	1.3333	0.0667	5.6128
I1-SVM Lin...	1.3333	0.0667	5.951

Selected Biomarkers	Class Correlation	SNR	FDR	mRMR Ranking
ASHGA5P023160	tissue: Breast Tumor	-1.346	3.458	2
ASHGA5P016634	tissue: Breast Tumor	-1.414	3.634	355
ASHGA5P000001	tissue: Normal Breast...	1.327	3.199	239

				Classifier Type			
	I1-SVM RBF Kernel	LDA	C-Tree	Ensemble 10 Tree	Naïve-Bayes	Linear	I1-SVM Linear Kernel
Acc	0.9061	0.8758	0.7333	0.9	0.9364	0.8623	0.8667
Se	0.8889	1	0.7222	0.6667	1	0.8889	0.8889
Sp	0.9048	0.8095	0.7619	1	0.9048	0.8413	0.8571
PPV	0.8889	0.7738	0.6984	NaN	0.85	0.8	0.8333
AUC	0.97354	0.9504	0.71429	0.85714	1	0.99471	0.99603

Figure 4: Example of Post Processing results page

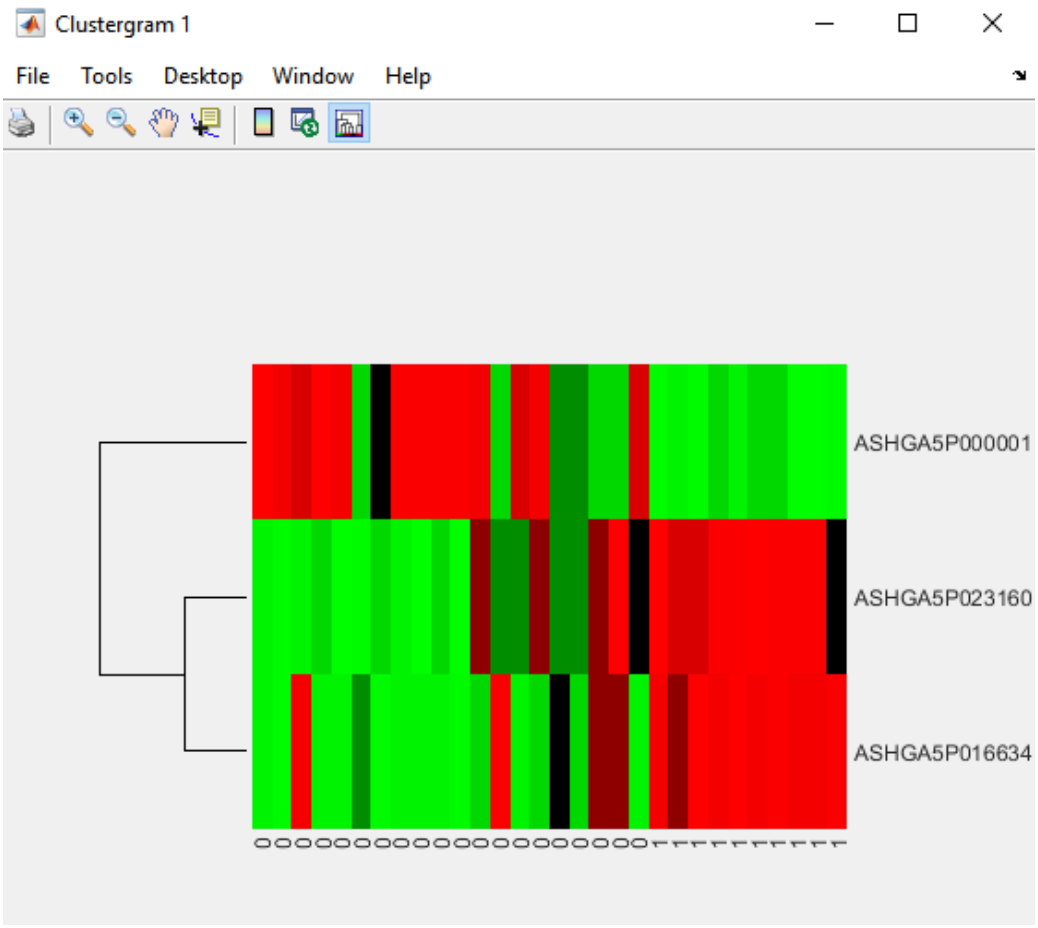


Figure 5: Example of Cluster-gram & Heat map

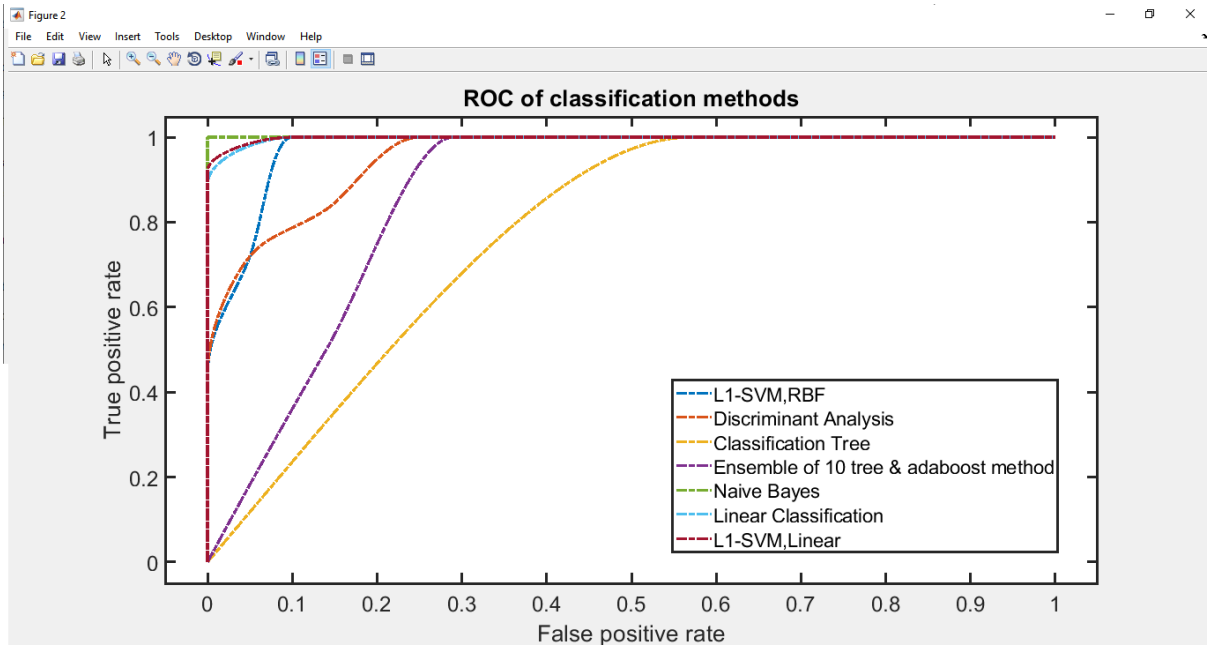
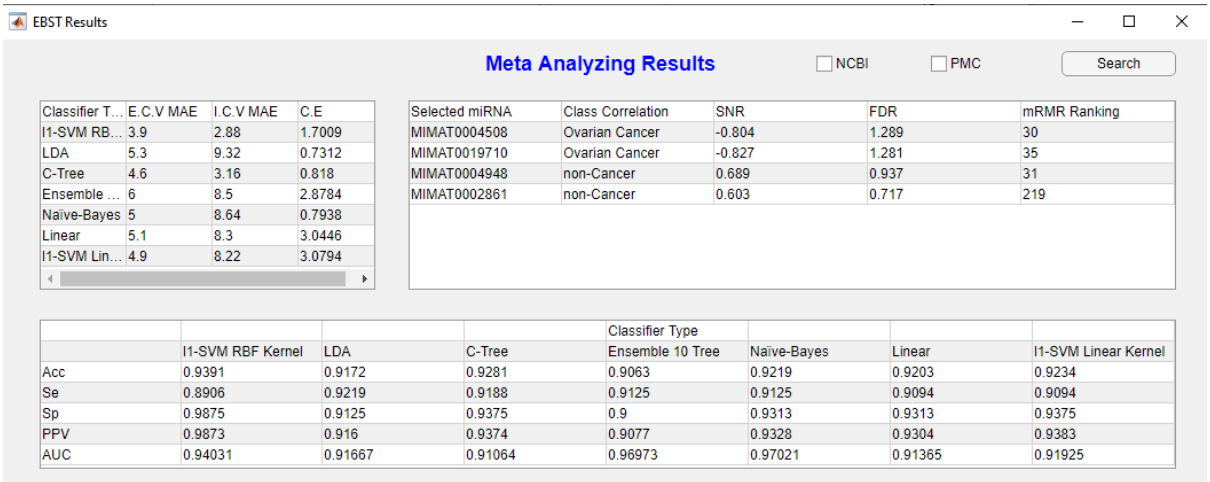


Figure 6: Example of ROC Curves

As can be seen, the data entry field is similar to that of part 1 and applies when data are not entered from page 1. Here we have:

a) **Insert Markers Manually**: If this option is selected, the biomarker names should be manually entered in the box below. Enter names below each other and use the Enter key to separate names. If this option is not selected, the biomarker names will be the ones obtained in the search algorithm and shown in the post processing section.

b) **Meta Analyzing** Button: Pressing this button will perform meta-analysis on meta data. Here's a results page similar to what we had in post processing, along with clustergram & Heatmap and ROC curves. An example of the results page is shown in Figure 7.



Classifier T...	E.C.V MAE	I.C.V MAE	C.E
l1-SVM RB...	3.9	2.88	1.7009
LDA	5.3	9.32	0.7312
C-Tree	4.6	3.16	0.818
Ensemble ...	6	8.5	2.8784
Naïve-Bayes	5	8.64	0.7938
Linear	5.1	8.3	3.0446
l1-SVM Lin...	4.9	8.22	3.0794

Selected miRNA	Class Correlation	SNR	FDR	mRMR Ranking
MIMAT0004508	Ovarian Cancer	-0.804	1.289	30
MIMAT0019710	Ovarian Cancer	-0.827	1.281	35
MIMAT0004948	non-Cancer	0.689	0.937	31
MIMAT0002861	non-Cancer	0.603	0.717	219

	Classifier Type						
	l1-SVM RBF Kernel	LDA	C-Tree	Ensemble 10 Tree	Naïve-Bayes	Linear	l1-SVM Linear Kernel
Acc	0.9391	0.9172	0.9281	0.9063	0.9219	0.9203	0.9234
Se	0.8906	0.9219	0.9188	0.9125	0.9125	0.9094	0.9094
Sp	0.9875	0.9125	0.9375	0.9	0.9313	0.9313	0.9375
PPV	0.9873	0.916	0.9374	0.9077	0.9328	0.9304	0.9383
AUC	0.94031	0.91667	0.91064	0.96973	0.97021	0.91365	0.91925

Figure 7: Example of Meta Analyzing results page

6. This section deals with the software status report, which consists of two parts:
   a) **Search Progress**: When the algorithm starts searching by pressing the MMOICA button, this section shows how many iterations of the algorithm have been performed.
   b) **Status**: This box shows what process is currently running.

7. This section is about displaying and saving the results, which we describe below:
   a) **Results1**: Pressing this button will display all the results of the analysis on the original data.
   b) **Results2**: Pressing this button will display all meta-analysis results again.
   c) **Save Excel 1**: Pressing this button will save all the results of the original data analysis as an excel file.
   d) **Save Excel 2**: Pressing this button will save all meta-analysis results as an excel file.
   e) **Save mat**: This button stores all the necessary data and analysis on it in *.mat format.
   f) **Write to Word**: This button saves all results including figures and tables in a *.doc file.