Multi-stage Cancer Metastasis Classification and Prognosis

[*] Dataset obtained from Gene Expression Omnibus (GEO)

Komal Ashraf
Senior Software Engineer , NETSOL Technologies
MS Data Science ,University of Management and Technology (UMT)
Lahore, Pakistan

I. Introduction

Metastasis includes the spread of cancer cells from the essential tumor to encompassing tissues and to far off organs is the main cause of cancer dismalness and mortality. It is assessed that metastasis is answerable for about 90% of malignancy deaths. Breast cancer, as one of the leading causes of death, comprises several sub types with controversial and poor prognosis. Considering the metastasis based on classification for organizing of breast malignant, it is basic to analyze the malady at beginning phases. Considering the metastasis based grouping for organizing of bosom malignant growth, it is basic to analyze the malady at beginning phases.

Identification of genes that can be used to predict prognosis in patients with cancer is important in that it can lead to improved therapy, and can also promote our understanding of tumor progression on the molecular level [1].

Early detection of breast cancer and its correct stage determination are important for prognosis and rendering appropriate personalized clinical treatment to breast cancer patients.It has been reported in 2019 that the incidence and mortality of breast cancer worldwide are 24.2% and 15.0%, respectively, deserving more attention from healthcare systems and policy-makers [2].

Metastatic breast cancer (likewise called stage IV) is a malignancy that spread to another part of the body, most normally the liver, brain, bones or lungs. Malignant cells can split away from the first tumor in the breast and travel to different parts of the body through the circulation system or the lymphatic system, which is a huge system of nodes and vessels that attempts to evacuate microscopic organisms, infections, and cellular waste products.

In this paper, XGBoost, Support vector machine, K-nearest neighbours , Random Forest Classifier, Naive Bayes Classifier are used to perform classification. k-Best Feature Selection and K-Fold cross validation to evaluate our models. The best overall accuracy we achieved is 90% for SVM.

II. Methods

We obtained data from Gene Expression Omnibus (GEO) website using GEOparse library by accessing GSE14020 in python. Using GEOparse library, we became able to understand our data structure as it showed us the combined platform for both GPLs' so that we can define the controls containing GSM files of GPL96 and GPL570.

A. Data details

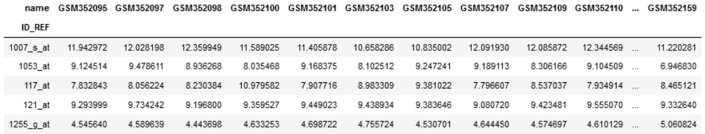
The data set we obtained from Gene Expression Omnibus (GEO) having identifier number GSE 14020 which contains 65 examples of various cancer metastases of Lungs, Brain ,Liver and Bone . Precisely, these samples are divided in two platforms (GPL96 and GPL570) having different dimensions. GPL96 have 36 samples and 22,283 genes. Whereas, GPL570 have 29 samples and 54,675 genes.

Data Table Description of GPL96 and GPL570:

- **ID:** Affymetrix Probe Set ID.
- **GB ACC:** GenBank Accession range.
- **SPOT ID:** Identifies controls.
- **Species Scientific Name:** The genus and species of the organism depicted by the probe set.
- **Annotation Date:** The date that the annotations for this probe array were last upgraded.
- **Sequence Type:** Consist of 2 types: "Exemplar sequence / Consensus sequence."
- **Sequence Source:** The database from where the sequence used to style this probe set was taken.
- **Target Description:** Detail description of the targets
- **Representative Public ID:** The accession range of a presentative sequence. To determine the database used it's referred to the "Sequence Source" field.
- **Gene Title:** Title of sequence described by the probe set.
- **Gene Symbol**: A sequence symbol, when one is obtained (from UniGene).
- **ENTREZ GENE ID:** Entrez Gene info UID
- **RefSeq Transcript ID:** References to multiple sequences in RefSeq. The sector contains the ID and outline for each entry, and there will be multiple entries per ProbeSet.

- **Gene Ontology Biological Process:** Every annotation contain three parts: ”Accession Number / Description / Evidence”. The framework compares directly on to the GO_ID. The verification can be ”direct” or ”extended”.
- **Gene Ontology Cellular Component:** Every annotation contain three parts: ”Accession Number / Description / Evidence”. The framework compares directly on to the GO_ID. The verification can be ”direct” or ”extended”.
- **Gene Ontology Molecular Function:** Every annotation contain three parts: ”Accession Number / Description / Evidence”. The framework compares directly on to the GO_ID. The verification can be ”direct” or ”extended”.

After combining the data of GPL96 and GPL570, our dataset contains 65 GSM files columns. In order to draw meaning full results we obtain table with each GSM as column, ID_REF as index and VALUE in each cell using pivot_samples method from GSE object and this is the data we further proceeded with.



name	GSM352095	GSM352097	GSM352098	GSM352100	GSM352101	GSM352103	GSM352105	GSM352107	GSM352109	GSM352110	...	GSM352159
ID_REF												
1007_s_at	11.942972	12.028198	12.359949	11.589025	11.405878	10.658286	10.835002	12.091930	12.085872	12.344569	...	11.220281
1053_at	9.124514	9.478611	8.936268	8.035468	9.168375	8.102512	9.247241	9.189113	8.306166	9.104509	...	6.946830
117_at	7.832843	8.056224	8.230384	10.979582	7.907716	8.983309	9.381022	7.796607	8.537037	7.934914	...	8.465121
121_at	9.293999	9.734242	9.196800	9.359527	9.449023	9.438934	9.383646	9.080720	9.423481	9.555070	...	9.332640
1255_g_at	4.545640	4.589639	4.443698	4.633253	4.698722	4.755724	4.530701	4.644450	4.574697	4.610129	...	5.060824

Fig. 1.

We took the transpose of the data where all the rows containing the ID_REF became column and all the columns containing GSE files became rows. Then we add a new column in our data frame named as **Metastasis** which contains all the relevant the type of metastasis i.e. lungs, liver, brain and bone.This additional column **Metastasis** that we added in our data frame is our ”Dependent Variable” .

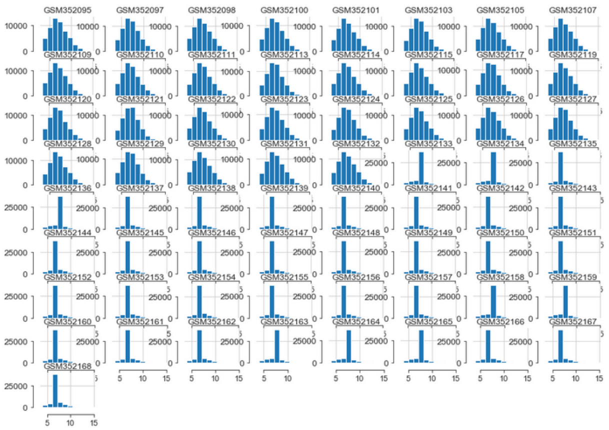


Fig. 2. Graphical Representation, GSE 14020 Selected Samples

B. Normalization

Real data-sets are regularly messy and come with a great diversity of range in their values: negative numbers, huge range differences within columns, etc. same is the case with our data set .That’s why ’Data Normalization’ is a key initial step. Normalizing the data implies setting it up in a way that is simpler for the system to process.

For out dataset, we are using **Standard Scaler** normalization, assuming the data is normally distributed inside each feature and is scaled to such an extent that the distribution is now centred around 0, with a standard deviation of 1.

```
Text(0.5, 1.0, 'Normalized data')
```

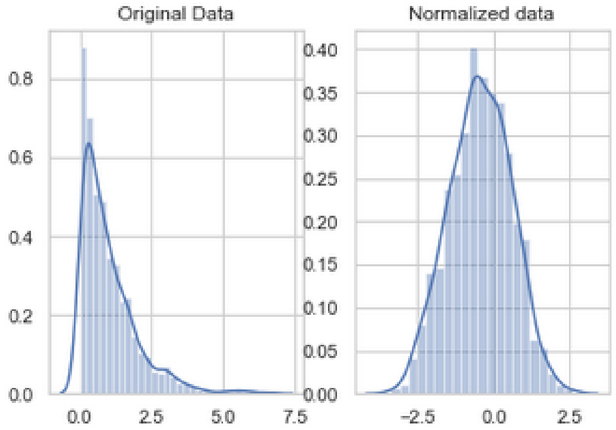


Fig. 3. Normalization using Standard Scaler

C. Sampling

The data set we are dealing with is too large/huge to be in any way dealt with. A workaround is to remove irregular samples from the dataset and work on it. In a data set, a training set is actualized to develop a model, while a test (or validation) set is to approve the model constructed. Data points in the training set are barred from the test (validation) set.

Our data set is partitioned into a training set and a test set.We essentially attempt to make a model to predict the test data. In this way, we utilize the training data to fit the model and testing data to test it. The models produced are to predict the outcomes unknown which is named as the test set.

For our dataset, using 70% of the data for training dataset and 30% for test set.

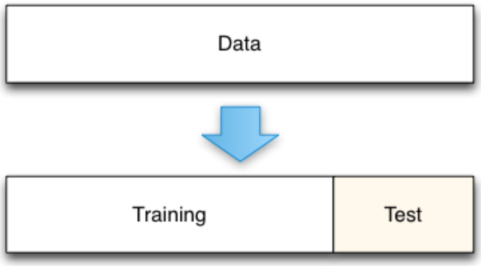


Fig. 4. Dividing dataset into training and test sets

D. Feature Selection/Extraction

The gene expression dataset contains 22,283 genes for lung, liver,brain and bone metastasis samples. The curse of dimensionality makes it difficult to classify the dataset in its current form. Thus, engaging in feature selection is essential to narrow down the number of genes to a handful at each node.K-Best and Chi-square are applied to select the best information gain of the selected genes, this step (Which is usually called filter feature selection) will drop down the number of genes to a couple of hundreds based on the correlation between each class and the gene expressions based on the default correlation threshold.

In this project for feature selection, we have used ”Univariate method”

- **Univariate Method:** In Univariate feature selection, the best features are chosen on the off chance that they are higher than a specific threshold. This threshold can be tuned in **SelectKBest** way.
- In the graph, these are the top 20 best features according to our dataset.

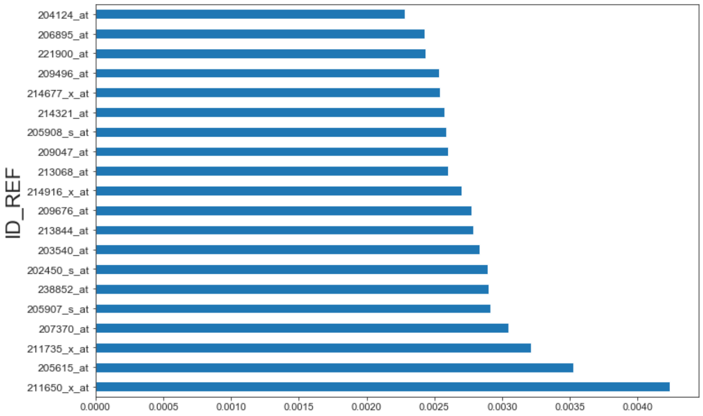


Fig. 5. Feature selection using univariate method (SelectKBest)

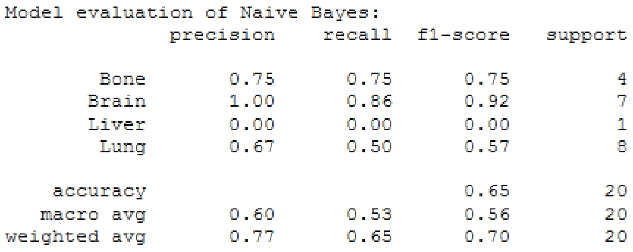
E. Classification

For Classification, we have applied different classification algorithms on our data set to check which model is more suitable and gives highest accuracy. For assessing, we have implemented ”confusion matrix, precision, recall and f1 score” which are the most generally utilized metrics.

After normalization and splitting our data set, we designed the models using XGBoost, Gaussian Naive Bayes classifier, KNN, Random Forest classifier.

For evaluating our designed models, we used **K fold cross validation** method in order check the consistence and remove the randomness in models. For our data set, we have utilized ”K-Fold CV” technique to split data into a K number of ares i.e. folds, where each fold is used as a testing set sooner or later at some point.

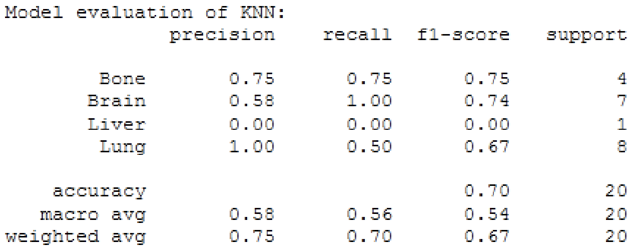
- **Gaussian Naive Bayes:** The outcome shows that our Naive bayes calculation has classified the 20 records with 65% accuracy.



```
Model evaluation of Naive Bayes:
              precision    recall  f1-score   support

        Bone       0.75      0.75      0.75         4
       Brain       1.00      0.86      0.92         7
       Liver       0.00      0.00      0.00         1
        Lung       0.67      0.50      0.57         8

    accuracy                           0.65        20
   macro avg       0.60      0.53      0.56        20
weighted avg       0.77      0.65      0.70        20
```

Fig. 6. Gaussian Naive Bayes

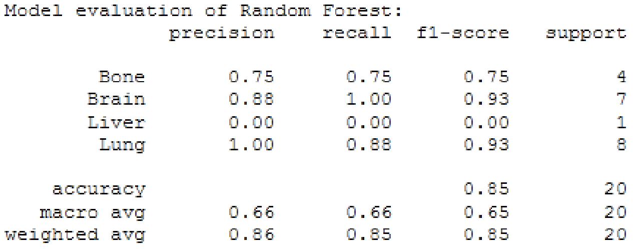
- **KNN - K-Nearest Neighbors:** K-Nearest Neighbors works by checking the distance from test set to the known estimations of training set. The outcome shows that our KNN calculation has classified the 20 records with 70% accuracy.



```
Model evaluation of KNN:
              precision    recall  f1-score   support

        Bone       0.75      0.75      0.75         4
       Brain       0.58      1.00      0.74         7
       Liver       0.00      0.00      0.00         1
        Lung       1.00      0.50      0.67         8

    accuracy                           0.70        20
   macro avg       0.58      0.56      0.54        20
weighted avg       0.75      0.70      0.67        20
```

Fig. 7. KNN - K-Nearest Neighbors

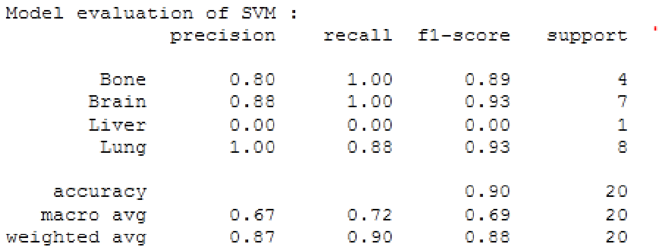
- **Random Forest:** The random forest algorithm integrate various algorithm of a similar sort i.e. multiple decision trees, bringing about a forest of trees.
  Using Random forest classification method for our 4 categories (lungs, liver, brain and bone) 85% of accuracy is achieved.



```
Model evaluation of Random Forest:
              precision    recall  f1-score   support

        Bone       0.75      0.75      0.75         4
       Brain       0.88      1.00      0.93         7
       Liver       0.00      0.00      0.00         1
        Lung       1.00      0.88      0.93         8

    accuracy                           0.85        20
   macro avg       0.66      0.66      0.65        20
weighted avg       0.86      0.85      0.85        20
```

Fig. 8. RF - Random Forest

- **SVM:** Support Vector Machines work by drawing a line between various clusters of data points to group them into classes. Focuses on one side of the line will be one class and focuses on the other side belong to another class.
  By applying SVM classifier method on our categories i.e. Brain, bone, liver and lung, the outcome shows 90% of accuracy.
- **XGBooster:** XGBoost is an open-source software library which provides a gradient boosting framework. It's a

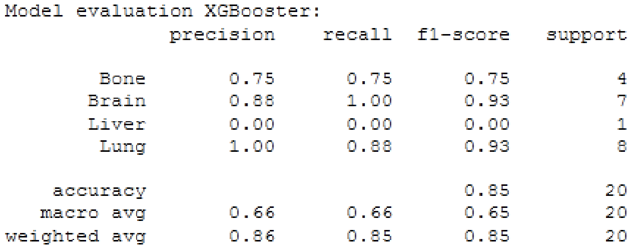


```
Model evaluation of SVM :
              precision    recall  f1-score   support

        Bone       0.80      1.00      0.89         4
       Brain       0.88      1.00      0.93         7
       Liver       0.00      0.00      0.00         1
        Lung       1.00      0.88      0.93         8

    accuracy                           0.90        20
   macro avg       0.67      0.72      0.69        20
weighted avg       0.87      0.90      0.88        20
```

Fig. 9. SVM - Support Vector Machine

usage of gradient boosted decision trees intended for speed and performance.

Using XGBooster method for our categories, the outcome shows 85% of accuracy.

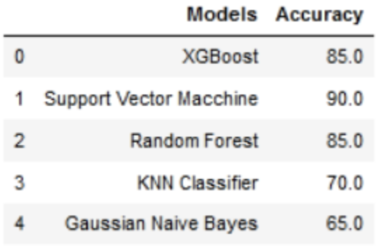


```
Model evaluation XGBooster:
              precision    recall  f1-score   support

        Bone       0.75      0.75      0.75         4
       Brain       0.88      1.00      0.93         7
       Liver       0.00      0.00      0.00         1
        Lung       1.00      0.88      0.93         8

    accuracy                           0.85        20
   macro avg       0.66      0.66      0.65        20
weighted avg       0.86      0.85      0.85        20
```

Fig. 10. XGBooster

F. Accuracy

Using different classification models, **SVM** model is the most accurate model that we tried (Fig. 11), having largest accuracy i.e 90% .



	Models	Accuracy
0	XGBoost	85.0
1	Support Vector Macchine	90.0
2	Random Forest	85.0
3	KNN Classifier	70.0
4	Gaussian Naive Bayes	65.0

Fig. 11. Accuracy results of different models

III. Results and Discussion

In this paper we have endeavored to explain, compare and evaluate the performance of various machine learning strategies that are being applied to multi stage cancer metastasis prediction and prognosis. Essentially we distinguished various patterns with respect to the types of machine learning strategies being used, the types of training data being incorporated, the sorts of endpoint expectations being made, the kinds of malignant diseases being examined and the general performance of these techniques in anticipating cancer growth or results.

In this research, we obtained data from Gene Expression Omnibus (GEO) database using GEOparse library where we accessed to GSM14020 file. This GSM14020 file contained two different platforms named as GPL6 and GPL570 which contains data of lungs, liver, brain and bone metastasis. Hence, after obtaining all the data, performing pre-processing and normalization, we analyzed that the data is quite large to be handled by the machine with space less then 32 GB and also for the machine learning models to perform efficiently. So we used k-best feature selection method to handle our data. The best thing about k best is that is used chi square test in it which selects the features which are highly dependent (higher Chi-Square value) on the response so that it can be selected for model training.

By validating the challenging dataset, the concepts of the data preprocessing methods, the feature selection techniques and the classification algorithms being used, the performance shows that our Machine Learning models are capable of learning higher level discriminating features and has the best accuracy in multi-class breast cancer classification. SVM in our dataset gave the highest accuracy of 90% where as Random forest and XGBoost gave 85% accuracy. However, KNN and Naive Bayes classifiers gave 70% and 65% accuracy respectively.

Improvements in experimental design along with improved biological validation would no doubt enhance the overall quality, generality and reproductibility of many machine-based classifiers. Overall, we believe that if the quality of studies continues to improve, it is likely that the use of machine learning classifier will become much more commonplace in many clinical and hospital settings.

IV. Conclusion

The use of a machine learning models for identifying gene biomarkers for different cancers survival is a significant step in determining the proper treatment for each patient and will potentially increase survival rates [3]. In this paper, we perform classification using different machine learning models in order to analyze the different genes and provide prognosis for lung, liver, brain and bone cancer metastasis. Hence the highest accuracy we achieved is of SVM where the viability of the model anticipated 90% of accuracy. This is because the fact that SVM produce accurate and robust classification results, even when input data are non-monotone and non-linearly separable. So they can help to evaluate more relevant information in a convenient way. Since they linearize data on an implicit basis by means of kernel transformation, the accuracy of results does not rely on the quality of human expertise judgement for the optimal choice of the linearization function of non-linear input dat [4]. On the other hand KNN and Naive Bayes gave less accuracy as these are non parametric algorithms. Hence, this study will be very helpful for Hospitals and different heath case department for prognosis and diagnosis of multistage cancer metastasis so that they can provide therapy, treatment and other clinical assessments.

REFERENCES

[1] Seyfried TN, Huysentruyt LC. On the origin of cancer metastasis. Crit Rev Oncog. 2013;18(1-2):43-73. doi:10.1615/critrevoncog.v18.i1-2.40

[2] Martin TA, Ye L, Sanders AJ, et al. Cancer Invasion and Metastasis: Molecular and Cellular Perspective. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013.

[3] Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L and Ngom A (2019) A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Front. Genet. 10:256. doi: 10.3389/fgene.2019.00256
Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. BioMed Central. 2017;50(33):1-23.

[4] International Agency for Research on Cancer. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018.

[5] Chambers AF, Groom AC, MacDonald IC. Dissemination and growth of cancer cells in metastatic sites. Nature Rev Cancer. 2002;2(8):563–72.

[6] Samundeeswari ES, Saranya PK. An artificial neural network model for prediction of survival time of breast cancer dataset. Int J of Research in Engineering and Applied Sciences. 2016;6(1): 161-168

[7] Welch DR. Defining a cancer metastasis. Amer Assoc Cancer Res. Education Book. 2006:111–5.

[8] Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. Science. 2011;331(6024):1,559–64.

[9] : Musa AA, Aliyu UM (2020) Application of Machine Learning Techniques in Predicting of Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto Northwestern Nigeria. J Data Mining Genomics Proteomics. 11:220. DOI: 10.35248/2153-0602.20.11.220.

[10] Duffy MJ, McGowan PM, Gallagher WM. Cancer invasion and metastasis: changing views. J Pathol. 2008;214(3):283–93.

[11] Seyfried TN. Cancer as a Metabolic Disease: On the Origin, Management, and Prevention of Cancer.

[12] Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. BioMed Central. 2017;50(33):1-23.