Single cell transcriptome deconvolution of breast cancer tissues reveals the cell fraction changes associated with cancer stages\*

|  |  |  |  |
| --- | --- | --- | --- |
| Rahul Bhansali Stony Brook University rbhansali@cs.stonybrook.edu | | Neha Mane Stony Brook University nmane@cs.stonybrook.edu | Dr. Daifeng Wang Stony Brook University daifeng.wang@stonybrookmedicine.edu.org |
|  |

ABSTRACT

In this study, we classify the single-cell data into clusters of cell-types. We then apply single-cell deconvolution on the TCGA-BRCA dataset to obtain the cell-type proportions of each sample. We use the clinical data of the TCGA-BRCA patients and machine learning to create classifiers that can predict the breast cancer stage of a patient with approximately 65% accuracy given the cell-type proportions of a breast cancer patient. We also compare the differentially expressed cell-type proportions between healthy and breast cancer patients. Finally, we analyze differentially expressed cell-types among patients having different survival rates.

1 INTRODUCTION

[Single cell sequencing](https://en.wikipedia.org/wiki/Single_cell_sequencing) in the general sense examines the sequence information from individual cells with optimized next generation sequencing (NGS) technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment. While, [single cell rna sequencing](https://en.wikipedia.org/wiki/Single_cell_sequencing#Single-cell_RNA_sequencing_(scRNA-seq)) (scRNA-seq) provides the expression profile of individual cells. Gene expression profiles help us understand the activity of thousands of genes at the same time, which makes it possible for us to get a wholesome idea of cellular function and how certain cells and its environment react to infections/diseases/particular treatment and ultimately helps us distinguish healthy and diseased states. RNA-seq has had a lot of applications to genomic medicine, several studies have used RNA-seq approach to identify differentially expressed transcripts between cancerous and normal tissue samples. [Estimating immune cell content](https://www.nature.com/articles/s41467-017-02289-3) to predict patient’s response to immunotherapy, studying [single cell expression profiles of melanoma](http://science.sciencemag.org/content/352/6282/189) cells before and after drug treatment showed that cell heterogeneity reflected two different states of expressions and also reflected drug resistance.

Deconvolution is a mathematical process used to extract constituent elements from a mixture of multiple signals. In genomic and biomedical research, deconvolution is widely applied to retrieve cell-type or tissue specific gene expression profiles from heterogeneous tissue samples. The heterogeneity of tissue samples and cells makes it difficult to identify the precise gene expression levels for each cell type. Relative changes in cell fractions, combined with variations such as disease state, complicate identifying true expressions. Changes in tissue composition are often indicative of disease progression or drug response, which makes the study of these cell fractions all the more important. A lot of [deconvolution algorithms](https://ieeexplore.ieee.org/document/7676285/) have been studied and proposed, most of which assume a linear model, in which the expression signal of the mixture is a weighted sum of the expression for its constitutive cell types. The reason a lot of research focuses on computational methods of deconvolution is due to the fact that in contrast to computational methods, experimental cell separation techniques like Fluorescence activated cell sorting (FACS), Laser capture microdissection (LCM) not only require a lot of time, effort and expense, they may also result in insufficient RNA abundance for further quantification of gene expression.

The results of this computational approach of deconvolution to estimate the cell proportions have been validated against experimental methods and the results are in [good agreement](https://www.nature.com/articles/s41467-017-02289-3#Fig6). While deconvolution has a lot of application to the biomedical field, our focus here is to use the knowledge of relative cell fractions to estimate the cancer stages/sub-types, which can then be further used to study the immunotherapy specific to each stage of cancer. Understanding the variations in relative cell-fractions in the early stages of cancer can help us direct specific therapy and understand the response to the therapy by changes in the tissue environment using single cell rna sequencing.

There are two aspects to the challenges faced during this study. One involved the biological challenge of estimating the cell-type proportions based on the signature gene matrix and the mixture (sample) matrix. The other one is the computational problem faced due to a paucity of the number of breast cancer patients that hinders proper learning of the machine learning breast cancer stage classifier. Compounding to this problem is the fact that the dataset is unbalanced and has many Stage-2 patients compared to other cancer stages. For overcoming the first challenge, we used CIBERSORT which is probably the best performing cell-type proportion estimator and used its output for further calculations. For the computational challenge, we balanced the dataset using the well-known standard SMOTE and ADASYN sampling. The problem of increasing the number of samples for better model training is still left as an open challenge and as a future work.

We begin the study by reducing the gene expression profiles of 515 single cells (each containing ~58k gene expression data)  using t-SNE dimensionality reduction technique to reduce it to two dimensions and feed this as an input to the DBSCAN clustering algorithm that clusters the cells into 9 clusters (cell-types). We use a set of 546 marker genes obtained from the [nature article](https://www.nature.com/articles/ncomms15081), and contain marker genes for ER+, HER2+, TNBC tumor sets and immune gene sets and extract only those gene-rows from the raw TCGA-BRCA dataset. Finally, we obtain the refined signature matrix and mixture matrix. We then apply the well-known single-cell deconvolution algorithm CIBERSORT to obtain the relative cell fractions of each cell-type for all the 1094 breast cancer patients. We also apply the same technique on -- healthy individuals to obtain the cell fractions of each cell-type for them.

The clinical data for each of the patients are then used together with the obtained cell-type fractions for training various classifiers that can predict the breast cancer stage of an individual given the cell fractions of each cell-type. Training data preprocessing is done via SMOTE and ADASYN sampling to balance out the training dataset due to presence of biases in the initial training data. We then experiment by creating 8 different classifier models and obtain the best accuracy (~65%) while using a Random Forest Classifier after 5-fold cross validation.

We perform further analysis to find out the differentially expressed cell-type between healthy individuals and breast cancer patients through the means of box-plot of the cell-type proportions. We observe that cell-type 6 has low fraction in healthy-individuals whereas it is comparatively high in breast cancer patients. Furthermore, cell-type 2 has high fraction in healthy individuals but it has a low fraction in breast cancer patients.

The final comparison we perform is to figure out the differentially expressed cell-type fractions based on the survival rate of the cancer patients. We firstly classify the survival rate of the cancer patients into three categories: a) low: if the “days to death” lies from the minimum observed value up to the first quartile value; b) medium: if the “days to death” lies from the first quartile value up to the third quartile value and c) high: if the “days to death” lies beyond the third quartile value. Once this classification is performed, we use heat-maps to observe that for a low survival rate, the  2nd, 3rd, 4th, 5th and 8th cell-type have negligibly fractions and 6th and 7th cell-type have low fractions. For a medium survival rate, the 2nd, 4th and 8th cell-type have increase cell fractions. For a high survival rate, the proportion of the 2nd cluster type is very high and for the 6th cluster type is very low. This is similar to as seen for the cell-type fractions of healthy patients. Besides this we also used Kaplan-Meier (KM) plots to obtain median survival time for each of the breast cancer stages.

2 MATERIALS AND METHODS

2.1 Single-cell Dataset

The single-cell RNA sequencing (RNA-seq) data was obtained from the NCBI Gene Expression Omnibus database under the accession code [GSE75688](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688) with reference to the article published by [Woosung Chung et al.](https://www.nature.com/articles/ncomms15081) in nature

The dataset’s GEO entry contains raw TPM values of the single cell RNA-seq data of primary breast cancer taken from 11 patients which provides gene expression profiles for **57912 genes in 515 cells**. The 11 patients have distinct subtypes of ER+, ER+ and HER2+, HER2+ and TNBC and lymph node metastasis.

2.2 Bulk Dataset

The bulk dataset was obtained from The Cancer Genome Atlas Project from the Genome Data Commons Data Portal. Under the [TCGA Breast Invasive Carcinoma Cancer Project](https://portal.gdc.cancer.gov/projects/TCGA-BRCA) we obtained the transcriptome profiling data for 1094 breast cancer patients. This data contained the gene expression profiles of **1094 patients across 21000 genes** in FPKM format

2.3 Clinical Dataset

The clinical dataset was also obtained from the [TCGA Breast Invasive Carcinoma Cancer Project](https://portal.gdc.cancer.gov/projects/TCGA-BRCA) for the same **1094 patients with 108 clinical observations**. Our focus was mainly on the survival rate and HER2 attributes.

2.4 Data Preprocessing

The gene expression values were used on the transcripts per million (TPM) as provided for the single cell dataset. The bulk dataset gene expression values were converted from Fragments Per Kilo base of transcript per Million mapped reads(FPKM) to TPM to ensure that both datasets were on the same scale using the conversion formula

TPMi = FPKMi x 106 / Sum(FPKM)

Furthermore, the expression values in both the datasets were transformed to:

y = log2(TPM + 1)

2.5 Generate Signature Matrix (genes v/s cell-type)

The signature matrix will be the matrix of reference gene expressions that we will use to deconvolve the mixture. This Signature Genes File contains a group of marker genes that collectively define a unique gene expression for each cell type component.

The GSE75688 dataset contains data for about 58K genes, deconvolution is performed only on the set of marker genes to obtain the relative cell fractions according to cell-type. Hence, a list of marker genes were obtained from the [nature article](https://www.nature.com/articles/ncomms15081), for ER+, HER2+, TNBC tumor sets and immune gene sets as well. The gene expressions of the marker genes were then extracted from the GSE75688 dataset of single cell rna-sequenced data by mapping the ‘gene\_name’ of the two files.

With the marker genes, the dataset now contains gene expression of **515 cells across 546 marker genes.** In order to get the classify the cells, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the high dimensional dataset into a 2D map. t-SNE lets us visualize clusters that could be formed from the dataset in a 2D space. According to the algorithm, if two points are closely similar in the original space, they appear close together in the 2D map and therefore we can see similar points forming distinct clusters, it also makes sure that if two points are different in the original space then they are free to be placed anywhere not close to each other. Thus, the algorithm preserves the local structure of the high-dimensional data in the 2D map. This 2D map data was then feed as an input to the DBSCAN clustering algorithm that clusters the cells into 9 clusters (cell-types). Once the clusters were observed, the signature matrix was constructed by averaging the gene expression of the 515 cells across the 9 cell-types observed from the clusters.

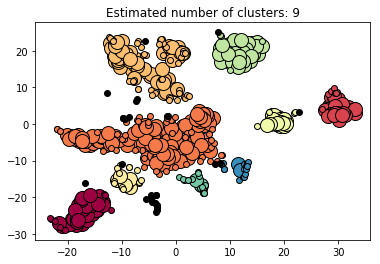


Figure 1: **Clusters of similar cells observed after applying t-SNE on the single-cell dataset**

2.6 Generate Mixture Matrix (genes v/s samples)

After preprocessing the bulk gene expression dataset, we have the gene expression for 1094 samples across 21000 genes, but in order to deconvolve this bulk dataset we require the gene expression to be across the same set of marker genes that we have obtained for the signature matrix. Therefore, after preprocessing the ‘gene\_name’  attribute of the bulk dataset, we join it with the marker genes file and obtain the bulk dataset for the marker genes. We now have our mixture file that contains the gene expression for 1094 samples across 546 genes.

2.7 Deconvolution using CIBERSORT

The Signature Matrix and Mixture Matrix are fed to [CIBERSORT](https://cibersort.stanford.edu/). The parameters used were 100 permutations. CIBERSORT performs the deconvolution on this rna-seq data and gives the relative cell-type fractions of each sample. The CIBERSORT method assumes that the gene expression profile of a bulk mixture can be obtained by solving a linear equation system computationally as it can be explained by the weighted sum of the cell type-specific profiles of which it is composed.

2.8 Classifiers for Breast Cancer Stage Prediction

The next goal was to come up with a good machine learning model that can predict the stage of breast cancer based on the cell-type fractions for each patient. The cell-type fraction results obtained previously from CIBERSORT was used for training as well as k-fold cross validation. We categorized the 12 sub-stages of cancer into three types (Stage I, IA and IB as 1, Stage II, IIA, IIB as 2 and Stage III, IIIA, IIIB, IIIC, IV and X as 3), so that the problem could be modelled as a multi-class classification problem.

As the training data classes were imbalanced, we applied two different [sampling methods](https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family) to try and balance the classes for better accuracy of the model. The two methods used were as follows:

* 1. **SMOTE Sampling** - Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature.
  2. **ADASYN Sampling** - ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data.  The algorithm uses Euclidean distance for KNN Algorithm.

Overall, we applied 8 types of machine learning models (of the Scikit-learn library) to check which model gives the best accuracy in terms of cancer stage prediction:

1. Multi-Class Logistic Regression.
2. Random Forest Classification.
3. Linear SVC.
4. Linear SVM.
5. kNN Classifier.
6. Naive Bayes Classifier.
7. Convolutional Neural Network (CNN).
8. Decision Tree.

We applied 5-fold cross validation as the metric for model prediction accuracy calculation and to prevent overfitting.

**Table 1:** Comparison of different breast cancer stage prediction classifiers

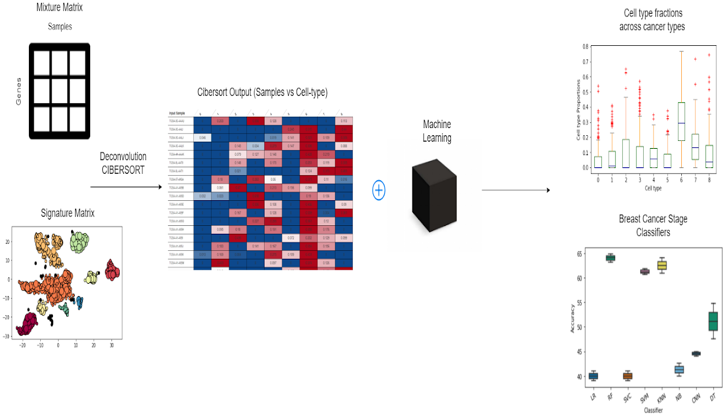
|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Model** | **Accuracy (with 5-fold cross validation on ADASYN sampling)** | **Accuracy (with 5-fold cross validation on SMOTE sampling)** |
| **1** | Multi-Class Logistic Regression | 39.26%  (multinomial =39.91%) | 40.44%  (multinomial = 40.44%) |
| **2** | Random Forest | 63.18% | 64.91% |
| **3** | Linear SVC | 39.97% | 40.65% |
| **4** | Linear SVM (C=10, gamma =10) | 61.82% | 60.63% |
| **5** | kNN Classifier (n\_neighbors = 4) | 60.90% | 64.10% |
| **6** | Naive Bayes Classifier | 39.97% | 42.58% |
| **7** | CNN | 44.52% | 44.66% |
| **8** | Decision Tree | 47.89% | 54.76% |

2.9 Survival Analysis

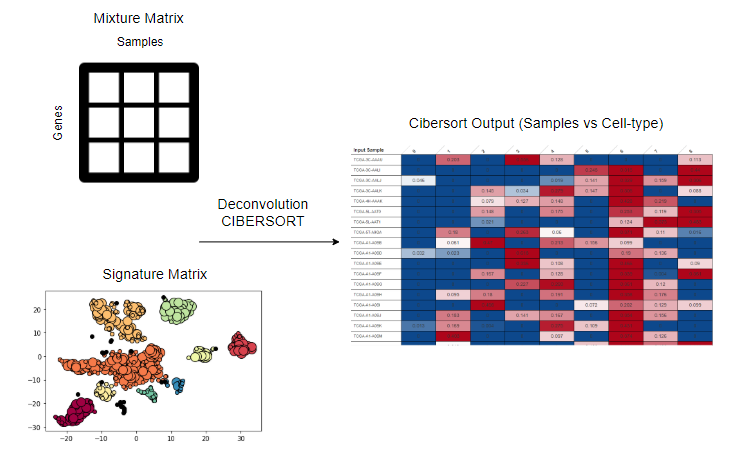
Using the clinical data available for the breast cancer patients, we estimated the survival plots using Kaplan-Meier (KM) estimate using the Python lifelines library. The KM fit API for survival analysis requires to be fed two arrays: one for the duration the subject was observed for (we took the DAY\_TO\_DEATH attribute of the clinical data with Nan values removed) and the second for the event observed: 1 if death was observed and 0 if death was not observed. We drew two types of KM plots. The first one considered all the stages of breast cancer together and the second one considering each stage of breast cancer differently.

2.10 Flowchart

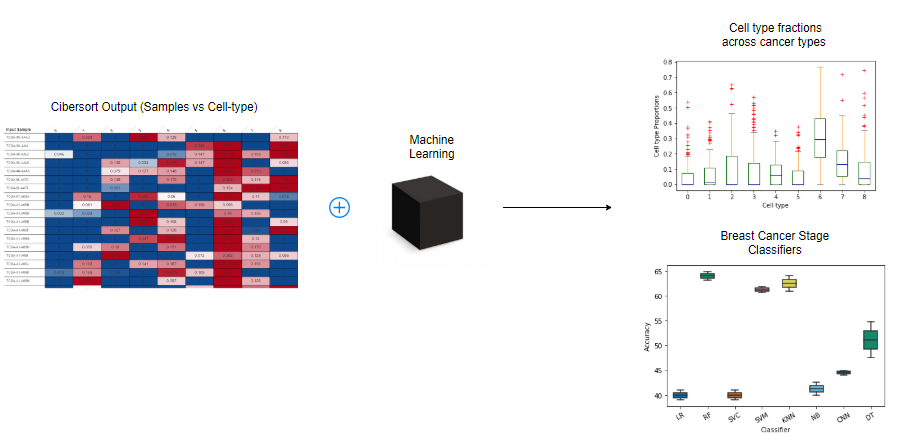
This section contains various flowcharts showing the pipeline of the process to firstly perform single-cell deconvolution and then create a breast cancer stage classifier based on the results of single-cell deconvolution.



**Figure 2.1: The complete flowchart of the process involved: from obtaining the cell-type fractions for each patient to creating a machine learning model for classification of breast cancer stage.**

****

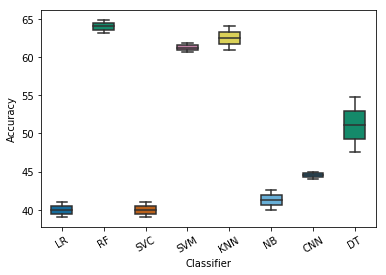
**Figure 2.2: A flowchart showing the steps involved to obtain the cell-type fractions of each breast cancer patient.**

****

**Figure 2.3: A flowchart depicting the elements involved in creating the breast cancer stage classifier.**

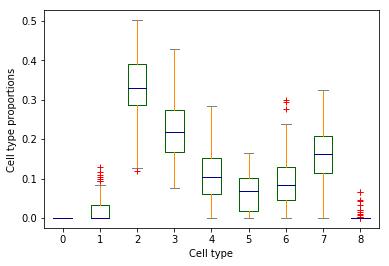
3 RESULTS

1. 8 types of classifiers were trained based on the BRCA dataset to obtain a model that could predict the stage of breast cancer for a patient.
2. Initial training on the raw dataset yielded low prediction accuracy scores due to the dataset being unbalanced (biased towards the stage II cancer stage).
3. Hence, the dataset was balanced through oversampling using two methods: SMOTE and ADASYN. This drastically improved the cancer stage prediction accuracy scores.
4. Overall, the top 3 models were: a) Random Forest (SMOTE sampling) with an accuracy of ~65%b) b) kNN classifier (SMOTE sampling) with an accuracy of ~64% c) Linear SVM with an accuracy of ~62%.



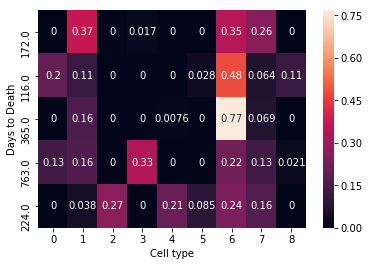
**Figure 3: A box-plot for comparison of the prediction accuracies of the classifiers**

1. Given below is a box-plot showing the proportions of each of the 9 cell-type clusters for HER2 positive patients. Here, we can see that the cluster type 6 is dominant for HER2 positive patients.
2. Given below is a box-plot showing the proportions for each of the 9 cell-type clusters for healthy patients. As we can see in the plot, the cell proportions for healthy patients are dominated by the 2nd cluster and has a lower proportion of the 6th cluster type. On the contrary, the 2nd cluster type was low and 6th cluster type most dominant for the HER2 positive patients.

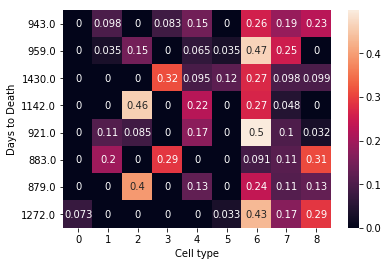


**Figure 4: Estimated cell type proportions of healthy humans without breast cancer**

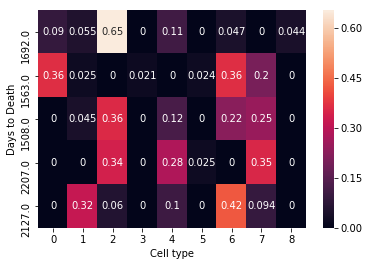
1. In each of the heat-maps given below we try to figure out the differentially expressed cell-type cluster for cancer patients with low, medium or high survival rates.
   1. In the case of patients with low survival rate (Figure 5), we see that there is a very low (**almost negligible**) cell-type proportion of the 2nd, 3rd, 4th, 5th and 8th cluster type. There are relatively low proportions of the 6th and 7th cluster type.
   2. In the case of patients with medium survival rate (Figure 6), we see that the cell proportions of the 2nd, 4th and 8th cluster type has increased compared to the concentrations with low survival rate. The 3rd and 5th cluster type proportions remain relatively unchanged.
   3. For patients with high survival rate (Figure 7), we see that the proportion of the 2nd cluster type is very high and for the 6th cluster type is very low. This is similar to as seen for the cell-type proportions of healthy patients (Figure 4).



**Figure 5: Heat map of cell-type fractions vs Days to Death for HER2 positive patients within the range of (116(min) - 793 days)**

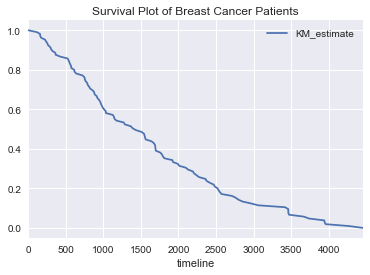


**Figure 6: Heat map of cell-type fractions vs Days to Death for HER2 positive patients within the range of (794-1489 days)**

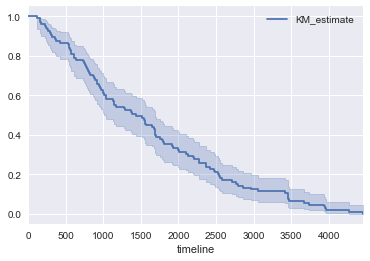


**Figure 7: Heat map of cell-type fractions vs Days to Death for HER2 positive patients within the range of (1489-2207(max) days)**

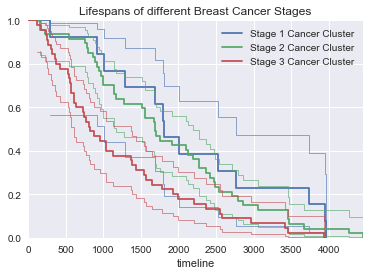
1. Here we see the various KM plots. We also obtained the median survival time for each stage of breast cancer. For Stage 1 cancer it was 1812 days. For Stage 2 cancer the median survival time is 1694 days and for Stage 3 it is 825 days.



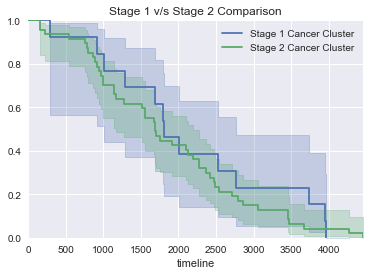
**Figure 8: KM survival plot of Breast Cancer Patients.**



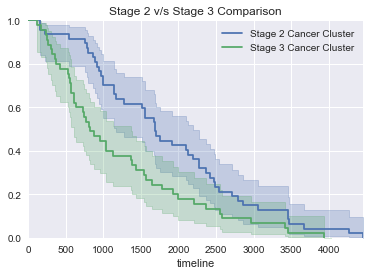
**Figure 9: KM survival plot of Breast Cancer Patients with 95% Confidence Intervals.**

****

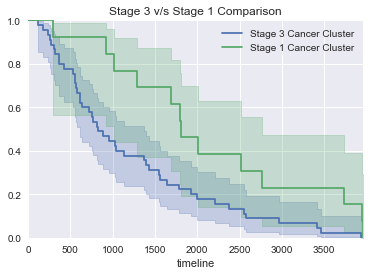
**Figure 10: KM survival plots for different stages of Breast Cancer Patients with 95% Confidence Intervals. The lines in bold show the survival line. The same colored lighter lines depict the 95% Confidence Interval line for that breast cancer stage.**



**Figure 11: KM survival plot of Stage 1 v/s Stage 2 Breast Cancer.**



**Figure 12: KM survival plot of Stage 2 v/s Stage 3 Breast Cancer.**



**Figure 13: KM survival plot of Stage 3 v/s Stage 1 Breast Cancer.**

**Table 2:** p-value and Test statistic values of the Log-Rank Test applied for comparison between various Breast Cancer Stages.

|  |  |  |  |
| --- | --- | --- | --- |
| Log-Rank Test (null\_distribution = chi squared, df=1,t\_0=-1) | | | |
| Comparison | alpha | p-value | Test statistic |
| Stage 1 v/s Stage 2 | 0.99 | 0.4391 | 0.5986 |
| Stage 2 v/s Stage 3 | 0.99 | 0.0048 | 7.9418 |
| Stage 3 v/s Stage 1 | 0.99 | 0.0097 | 6.6978 |

4 CONCLUSIONS

In summary, Using t-SNE on the single-cell data followed by DBSCAN clustering we were able to classify the single-cells into 9 different cell-type clusters. Using these cell-type clusters we were able to successfully apply the single-cell deconvolution technique CIBERSORT on the TCGA-BRCA patients dataset and obtain the relative cell-fraction of the cell-types for each patient. Based on the results obtained from CIBERSORT, we were also able to successfully implement a Random Forest model that could predict the breast cancer stage with around 65% accuracy. We also compared the cell-type proportions of breast cancer patients with healthy patients to find differentially expressed cell-types for the two categories using box-plots. We also analyzed differentially expressed cell-type proportions for patients having low, medium and high survival rates by means of a heat-map.

Future Work:

1. Experimenting with other single-cell deconvolution techniques (besides CIBERSORT) to obtain cell fraction of samples and then use these fractions as features to create new models and compare them to the models based on CIBERSORT generated cell fractions to see if it leads to better prediction accuracies.
2. The cancer stage prediction classifier should also be able to correct classify a healthy patient.
3. Improving the accuracy of the cancer stage prediction model through:

Hyper-parameter tuning of the best models.

Obtaining a larger training dataset on cancer patients.