Cairo University

MASS SPECTROMETRY IMAGING IN DETECTING TUMOR HETEROGENEITY

By:
Ibrahim Elsayed
Donia Abd Elslam
Renad Taher
Mariem Ahmed
Mustafa Yehia

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Systems and Biomedical Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

MASS SPECTROMETRY IMAGING IN DETECTING TUMOR HETEROGENEITY

By:
Ibrahim Elsayed
Donia Abd Elslam
Renad Taher
Mariem Ahmed
Mustafa Yehia

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Systems and Biomedical Engineering

Under the Supervision of

Prof. Dr. Walid Abd Elmoula

.....

Research interests: Biomedical Imaging, Bioinformatics, Deep Learning, AI in Spatial-Omics
Faculty of Engineering, Harvard University

Prof. Dr. Ahmed Morsy

.....

Head of SBME Department
Faculty of Engineering, Some University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

Acknowledgments

First, we want to say “Thank You” for Dr/ Walid Abd Elmoula for his time, assistance, and supervision of us along the past *10* months and for offering a suitable materials, resources, and opportunities reached to him for us. We also want to say the same thing for Dr/ Ahmed Morsy; Both showed honest and appreciated work.

Second, we are so proud reaching this moment and with our love and appreciation, we want to thank each doctor/TA/person in Cairo University - Faculty of Engineering (CUFE) who helped us achieve this. It would be difficult without you. So, thank you is not enough, but totally respected.

Table of Contents

Acknowledgments ..... **iii**
List of Tables ..... **v**
List of Figures ..... **vi**
Abstract ..... **vii**
Chapter1: Introduction **1**
Chapter2: Cancer Overview and Further **2**
2.1 Motivation **2**
2.2 Diagnosis Techniques **2**
2.3 Prognosis Techniques **2**
2.3.1 Survival Analysis **2**
Chapter3: Mass Spectrometry Imaging (MSI) **4**
3.1 Motivation **4**
3.2 Matrix-assisted Laser Desorption Ionization (MALDI) **4**
3.2.1 Sample and Tissue Preparation **4**
3.2.2 Matrix Application **5**
3.2.3 Applications **5**
Chapter4: Machine Learning Fundamentals **6**
4.1 Motivation **6**
4.2 Dimensionality Reduction **6**
4.2.1 Curse of Dimensionality **6**
4.2.2 Reducing Dimensions **6**
4.2.2 Linear Dimensionality Reduction Technique (PCA) **6**
4.2.3 Non-Linear Dimensionality Reduction Technique (t-SNE) **7**
4.3 Clustering **8**
4.3.1 K-means Clustering **9**
4.4 Microarray Analysis Technique **9**
4.4.1 Significance Analysis of Microarrays (SAM) **10**
4.5 Cross Validation (CV) **10**
4.5.2 K-fold CV **11**
4.5.1 Leave-Out-Patient-Out (LOPO) **11**
Chapter5: Applications (Gastric/Breast Cancers) **12**
5.1 Data Exploratory **12**
5.2 MSI Data **12**
5.3 Dimensionality Reduction **12**
5.3.1 PCA for Both Gastric/Breast Data **12**
5.3.2 t-SNE for Both Gastric/Breast Data **13**
5.4 Clustering **13**
5.5 Survival Analysis **14**
5.6 SAM **14**
Chapter6: Results (Gastric/Breast Cancers) **15**
6.1 MSI Data **15**
6.2 Dimensionality Reduction (t-SNE) **15**
6.3 Clustering (K-means Clustering) **16**
6.4 Survival Analysis (Kaplan-Meier Curves) **16**
6.5 SAM **17**
Chapter7: Literature Review **18**
Discussion **19**
Conclusions and Future Work **20**
References **21**
Appendix A **22**

List of Tables

Table 1: Some Formats of Response Variables in SAM 10
Table 2: Contingency Table of Breast Data applied on different K Values 16
Table 3: Findings of the reviewed sources 18
Table A.1: Clinical Outcome of Gastric Data (Brief) 22
Table A.2: Clinical Outcome of Breast Data (Brief) 22

List of Figures

Figure 1: Gantt Chart 1
Figure 2: Cancer Formation (Left) & WHO 2020 Cancer Statistics (Right) 2
Figure 3: MALDI 4
Figure 4: PCA Space 6
Figure 5: t-SNE Space 9
Figure 6: K-means Clustering (Left) & K-means Initial (Right) 9
Figure 7: Assigning Datapoints to Closest Centroid (Left) & Moving K-means Centroids (Right) 9
Figure 8: Steps for Gastric Data (Up) and Breast Data (Bottom) Analysis 12
Figure 9: Gastric Data (HE Image) 15
Figure 10: Breast Data (HE Image) 15
Figure 11: Gastric Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right) 15
Figure 12: Breast Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right) 15
Figure 13: Gastric Data K-means Spatial Image (Left) & Count Plot of Patients in each Cluster (Right) 16
Figure 14: Breast Data K-means Count Plot with Metastasis Colored ($\boldsymbol{k = 6}$), $\boldsymbol{k = 7}$ & $\boldsymbol{k = 8}$ 16
Figure 15: Gastric Data K-M Curve between Clusters (Left) & Significance of 1 vs 3 (Right) 16
Figure 16: SAM applied on Gastric Data 17

Abstract

Tumor subpopulations have molecular phenotypes that drive tumor progression and determine disease outcome which is essential for a more personalized therapy. Mass spectrometry imaging has proven its ability to identify diagnostic and prognostic biomarkers. In this research, we seek to determine tumor subpopulations that affect patient outcomes and the statistically associated subpopulations with poor survival and tumor metastasis. Here we introduce spatially mapped t-distributed stochastic neighbor embedding (t-SNE), a nonlinear visualization of the data that can better resolve the biomolecular intratumor heterogeneity. The outcomes will allow us to uncover subpopulations statistically associated with patient survival in primary tumors of gastric cancer and with metastasis in primary tumors of breast cancer.

Keywords— Mass spectrometry imaging, t-SNE, intratumor heterogeneity, metastasis

Chapter1: Introduction

Mass Spectrometry Imaging (MSI) is a technology that provides the spatial distribution of hundreds of biomolecules directly from tissue simultaneously that leads to minimal loss of histological information. The same tissue section can be histologically assessed and registered to MSI Dataset. Tumor Cells can be extracted from the often highly heterogeneous tissues encountered in patient tumors. This high cellular specificity is behind the increasing popularity of MSI in cancer research and its proven ability to identify diagnostic and prognostic biomarkers. There is growing awareness that MSI also can be used to annotate tissues based on the local mass spectrometry profiles and thereby differentiate tissues/regions that are not histologically distinct.

A hierarchical cluster analysis of MSI Data revealed a patchwork of molecularly distinct regions, which were postulated to reflect the tumor's clonal evolution. It was recently demonstrated that such an approach, using multivariate analysis of the MSI data to identify regions with distinct mass spectral signatures and then linking these molecularly distinct regions to patient outcome, enables the identification of tumor subpopulations that are statistically associated with poor survival and tumor metastasis. All methods used to date for revealing intratumor heterogeneity have been linear dimensionality-reduction techniques, but this linearity constraint focuses the results on the global characteristics of the data space at the expense of finer details. Accordingly, linear methods might not be sensitive to the subtle changes expected to demarcate the clonal progression of tumors, in which the molecular differences between nearly sequential subpopulations may be minor.

Non-Linear multivariate methods can preserve both *local* detail and *global* data structure in the low dimension representation by emphasizing similarities between data points such as t-distributed stochastic neighbor embedding (t-SNE) which has rapidly established itself as a method of choice for summarizing high dimensionality datasets owing to its ability to overcome the "*Crowding Problem*" in which some of the higher-dimensional data similarities cannot be faithfully represented in a single map. t-SNE has been applied to high-dimensionality imaging data and has been shown to outperform other dimensionality-reduction techniques in several life-science applications. t-SNE demonstrated its superiority over linear multivariate methods for demarcating regions of tissues with different mass spectral signatures.

Significance: MSI can uncover molecular intratumor heterogeneity. The challenge has been to identify those tumor subpopulations that drive patient outcomes within the highly complex datasets (hyperdimensional data, intratumor heterogeneity, and patient variation). Here we report an automatic, unbiased pipeline to nonlinearly map the hyperdimensional data into a 3D space, and identify molecularly distinct, clinically relevant tumor subpopulations. We demonstrate this pipeline's ability to uncover subpopulations statistically associated with patient survival in primary tumors of gastric cancer and with metastasis in primary tumors of breast cancer.

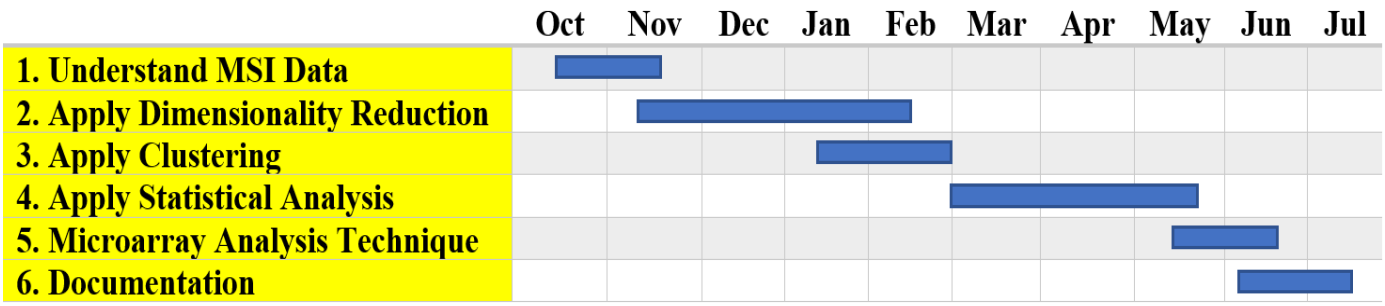


Figure 1: Gantt Chart

Chapter2: Cancer Overview and Further

2.1 Motivation

Cancer is considered as one of the most important topics from the 20th century till now and needs a lot of attention. According to World Health Organization (WHO) 2020 cancer statistics (Figure 2), it reflects our interest in both *breast* and *gastric* cancers. So, we are going to proceed with exploring causes of cancers, ways to discover/treat it and finally moving forward to a new approach called MSI to discover a lot of wonderful things about the field of cancer. Cancer is basically an abnormal growth of cells (Figure 2) with time that causes problems with Activity of Daily Living (ADL) for persons and sometimes complete disabilities or death. The earlier we discover it, the more chances to save someone's life.

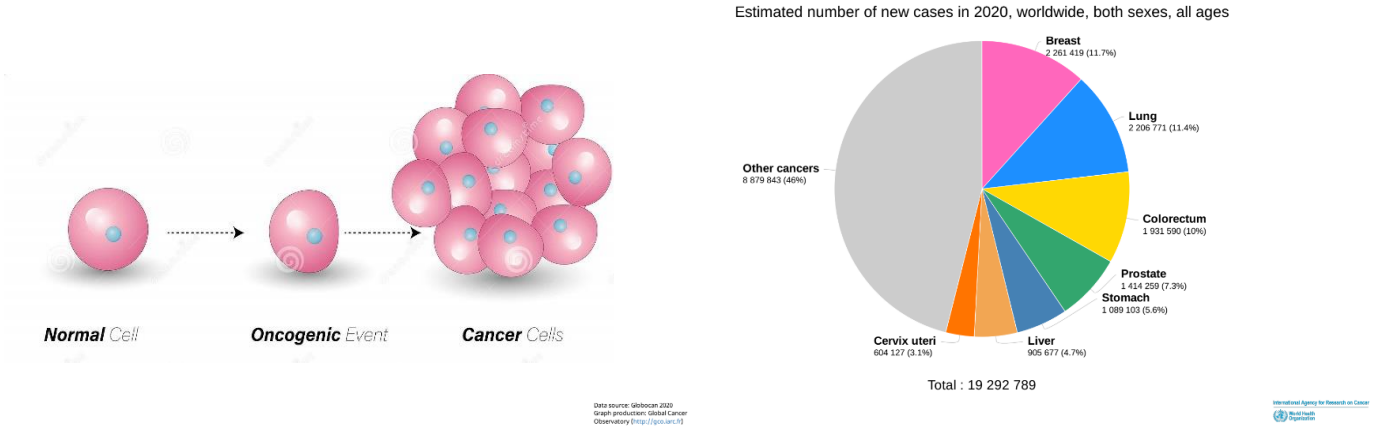


Figure 2: Cancer Formation (Left) & WHO 2020 Cancer Statistics (Right)

2.2 Diagnosis Techniques

It is known to many of us that cancer is now available to be treated (Diagnosis) using many techniques such as (MRI, Nuclear Medicine, CT, etc.) all of those became old-fashioned and the challenge now is to discover new ones; as the cancer itself are developing with time and be more cruel than previous. On the body level, it is better to use these techniques for sure, but when it comes to (Molecular Level) they have no benefit and would be better not to use them. For the molecular level, there is no doubt that MSI is the master of all techniques in the current time. It shows many things that the microscopic level itself could not determine (We will see that upon our thesis) and that helps us in different aspects in the cancer field.

2.3 Prognosis Techniques

Prognosis is simply a follow-up approach, but in an expertized way. For cancerous patients, it is very important to periodically checkup the patient's status. One of techniques used to assure this approach is "Survival Analysis" and we will procced with it in more details in the following section.

2.3.1 Survival Analysis

It is known as (time to event analysis) that describes some methods for analyzing the length of time till a well-defined end point of interest happens (The point of interest in cancer is death indeed). Sometimes the actual survival times for some patients is unknown as not all patients experience the event by the end of the observation period. This phenomenon, that is called *censoring*, must be put into consideration in the analysis to allow for valid inferences. Moreover, survival times are usually skewed, limiting the usefulness of analysis methods that assume a normal data distribution [1]. For patients who survive until the end of the study period, or who are lost to follow-up before the end of the observation period, full survival times are unknown. Instead all that is known is that the survival time is greater than the observation time. This unique feature of survival data is referred to as right censoring, which is described in more detail below.

The results would be biased by equaling the observed survival time (follow-up time) of the patients with the unobserved total survival time or by ignoring the censored patients in the analysis. Censoring makes the analysis of survival experiments a bit complicated [2]. Even if there was no censoring in the data set,

survival times usually have a heavily skewed distribution, limiting the usefulness of statistical tests that assume a normal data distribution.

A) Kaplan-Meier (K-M) Curve Method

It was developed to deal with incomplete observations, so it handles the cases where patient drops out from the study (censored). K-M is the easiest way to compute survival over certain interval. survival estimates are a familiar way to deal with time to event concepts, As K-M shows the probability of surviving over a small interval(s). [3]

$$S = \frac{No.of\ living\ patients\ at\ start - No.of\ died\ patients}{No.of\ living\ patients\ at\ start}$$

K-M curve is a step wise function where the estimated survival drops vertically whenever the event occurs. [7] In our project time to event is time until the cancer patient dies. In our study, each patient has three variables time to event, his status at the end of the study (censored or dead), and cluster as we said previously, each patient in clinical data is assigned to cluster using k-means.

1. Censoring

It means that we cannot determine the survival time that happens if the patient drops out of the study or the study came to an end before the event of interest occurs. [6] Censoring has three types right, left, and interval censoring. Left censoring occurs when the event of interest occurs before enrollment in study. Right censoring occurs when the event of interest occurs after the interval of the study. Interval censoring occurs when the event of interest occurs within interval and not observed in an exact time. In our study, we deal with only right censoring where we do not have definite information about patients' status at the end of the study.

In survival analysis we make *3* assumptions, Firstly, we assume censored patients have the same prospects as patients who continued the follow up. Secondly, we assume that patients who were recruited at the beginning and during the study have the same survival probabilities. Thirdly, we assume that we know the specific time the occurred at. [4]

2. Log Rank Test

It tests the null hypothesis that states that there is no significant difference between the cluster survival curves at any point in time [5], thus it compares the entire curves not a survival probability at a time. [7] We calculate the test statistic (X^2) using the equation,

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Where O_1, O_2 are the total number of the observe deaths in each cluster, and E_1, E_2 the expected number of events in each cluster. The calculated statistic is used then to calculate p-value which will help us in determining the significant difference between the clusters.

3. P-value

It is the probability value that helps you to accept or reject your null hypothesis. P-value is evidence against the null hypothesis, the smaller the p-value the stronger we bent to reject the null hypothesis, so the greater the statistical significant difference. A p-value less than 0.05 is statistically significant difference and we reject the null hypothesis, while p-value greater than 0.05 is a strong evidence to accept the null hypothesis. [8]

B) Fisher's Exact Test Method

It is a statistical significance test used to determine if there is a significant relationship between two categorical variables and it is also used in the analysis of contingency tables (Explained Below) [11][12]. We chose Fisher's exact test to apply it on our data as the sample is small [13]. After making our contingency tables, we applied fisher's test on them to get the required p-value.

C) Contingency Tables Method

It is very useful to condense many observations into smaller ones to make it easier to maintain tables and show the relation between the variables, so this helps in finding interactions between them [9]. It shows the distribution of a variable in the rows and another in its columns so this is considered a way of summarizing categorical variables [9]. (Special type of frequency distribution tables), where *2* variables are shown simultaneously [10].

Chapter3: Mass Spectrometry Imaging (MSI)

3.1 Motivation

Imagine yourself with a group of people at the same neighborhood, but all of you want to cross a crowded street. So, it is obvious that you will wait till a passage is opened for you all or you will face a problem known by "Crowded Data". Refer to all of you as a "Cell" and the spreading of you in the street as a "Tumor". The only highlighted case is that all of you can pass with the same percentage, but the most proper evidence is that you will face each other causing a ruin in the system (Refer to it as your Body) and that we want to focus on.

MSI enables untargeted investigations into the *spatial distribution* of molecular species in a variety of samples. It can image thousands of molecules (e.g., Metabolites, Lipids, Peptides, Proteins, and Glycans) in a single experiment *without labeling*. [14] Tumor Heterogeneity is not with the same percentage in cancerous patients' body. That could be noticed by distribution of ions of proteins in a tissue sample and that is the power of MSI. MSI could be divided into *3* main categories: Secondary Ion Mass Spectrometry (SIMS - Lipids), Matrix-assisted Laser Desorption Ionization (MALDI - Proteins and Metabolites), and Desorption Electrospray Ionization (DESI - Both). We will illustrate MALDI technique in details next.

3.2 Matrix-assisted Laser Desorption Ionization (MALDI)

It is simply a category of MSI as mentioned previously that investigate the proteins in a tissue sample. Why are we interested in this technique specifically? The answer is we want to investigate the effect of inter/intratumor heterogeneity and determine the biomarkers (Proteins) responsible for that. So, MALDI is indeed the best analytical method for that. Its methodology is so simple, all what we need is the preparation of the sample and the rest could be done using MSI instrument (e.g., Mass Analyzer or MALDI-time-of-flight "TOF") (Not Available in Egypt Yet). MALDI is best for its ability to image wide range of molecular weights and molecular species. [14]

Comparing with other techniques, they all need minimal sample preparation (No matrix as in MALDI). That makes MALDI unique and best in protein and big samples analysis. In brief, to detect mass-to-charge (m/z) values from a sample we shall do some important steps which are: Sample Preparation, Fixation Wash, Ion Excitation (Laser and Electric Field), and Detector for it (Figure 3). The fact that is here, we could get a data contain the spectra of all MSI Data of one spot and convert to any format (e.g., ".mat", ".py", etc.) to start deal with.

3.2.1 Sample and Tissue Preparation

Sample preparation is a critical step in MSI. We first take thin tissue slices mounted on conductive microscope slides and apply a suitable MALDI matrix to the tissue (Manually or Automatically). Secondly, microscope slide is inserted into a MALDI Mass Spectrometer (MS). MS records *spatial distribution* of molecular species (e.g., Peptides, Proteins, etc.). Suitable image processing software can be used to import data from MS to allow visualization and comparison with the optical image of the sample.

Tissue samples must be preserved quickly to reduce *molecular degradation*. First freeze the sample by wrapping the sample then submerging it in a cryogenic solution. Once frozen, the samples can be stored below $-80°C$ for up to a year. When ready to be analyzed, the tissue is embedded in a gelatin media. Finally,

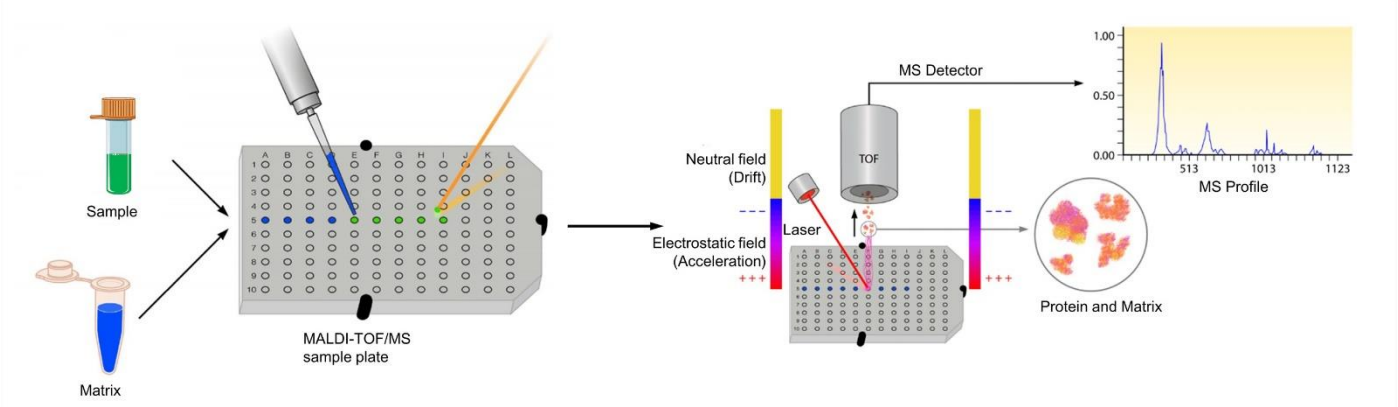


Figure 3: MALDI

we place the tissue sections sample on the surface of a conductive slide that is of the same temperature, and then slowly warmed from below. The section can also be adhered to the surface of a warm slide by slowly lowering the slide over the cold sample until the sample sticks to the surface. The sample can then be stained to easily target areas of interest and pretreated with washing to remove species that suppress molecules of interest. Washing with varying grades of ethanol removes lipids in tissues that have a high lipid concentration with little delocalization and maintains the integrity of the peptide spatial arrangement within the sample.

3.2.2 Matrix Application

Matrix must absorb at the laser wavelength and ionize the analyte. Matrix selection and solvent system relies heavily upon the analyte class desired in imaging. The analyte must be soluble in the solvent to mix and recrystallize the matrix. The matrix must have a homogeneous coating to increase sensitivity, intensity, and shot-to-shot reproducibility. Minimal solvent is used when applying the matrix to avoid delocalization.

One technique is spraying. The matrix is sprayed, as very small droplets, onto the surface of the sample, allowed to dry, and re-coated until there is enough matrix to analyze the sample. The size of the crystals depend on the solvent system used.

Sublimation can also be used to make uniform matrix coatings with very small crystals. The matrix is placed in a sublimation chamber with the mounted tissue sample inverted above it. Heat is applied to the matrix, causing it to sublime and condense onto the surface of the sample. Controlling the heating time controls the thickness of the matrix on the sample and the size of the crystals formed. Automated spotters are also used by regularly spacing droplets throughout the tissue sample. The image resolution relies on the spacing of the droplets.

Images are constructed by plotting ion intensity versus relative position of the data from the sample. Spatial resolution highly impacts the molecular information gained from analysis.

3.2.3 Applications

MALDI involves the visualization of spatial distribution of *proteins*, *peptides*, *lipids*, and *other small molecules* within thin slices of tissue. The application of this technique to biological studies has increased significantly since its introduction. MALDI is providing major contributions to the understanding of diseases, improving diagnostics, and drug delivery. Significant studies are of the eye, cancer research, drug distribution, and neuroscience. Also, it can differentiate between drugs and metabolites and provide histological information in cancer research, which makes it a promising tool for finding new protein biomarkers. However, this can be challenging because of ion suppression, poor ionization, and low molecular weight matrix fragmentation effects. To combat this, chemical derivatization is used to improve detection.

Source (Wikipedia): MALDI Imaging

Chapter4: Machine Learning Fundamentals

4.1 Motivation

Machine Learning (ML) is a key concept in data science which facilitates our life in many aspects. In our thesis, we used a lot of ML algorithms and concepts to reach to our results according to a specific approach and well-defined strategy. ML algorithms are essential and required for maintaining and exploring any type of data and tune its parameters to reveal some hidden features in it.

We used a lot of ML algorithms and concepts here in our research to reveal some hidden features in the world of cancer specially for our main two type "gastric cancer" and "breast cancer". Algorithms applied faced some problems and sometimes we had to prefer one technique on another just as in our case study

4.2 Dimensionality Reduction

4.2.1 Curse of Dimensionality

This problem appears or occurs when analyzing and organizing the data in high-dimensional space and does not occur in low-dimensional space. You can face this problem in different domains such as numerical analysis, machine learning like in our work, data mining and databases. Commonly these problems occur when the dimension of data increases, the volume of the data increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. As a goal for that is to obtain the reliable and statistically sound result; Make the amount of data grows exponentially with increasing of the dimensionality. But it is not enough, you need to organize and search data that relies on detecting areas with similar properties. In high dimensional data, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.

4.2.2 Reducing Dimensions

Before we talk more deeper in dimensionality reduction, you should know the importance of this part is and imagine the effect of it. Let us talk about the dimensionality reduction with more details. Dimensionality reduction is a concept depends on the feature selection. That system leads to select some of relevant features from the variables and predictors to construct your model. But we face the problem known as "short, fat data problem" which means the number of features "proteins" larger than the number of samples, which exists in Gastric Data and Breast Data. The main concept is the features are redundant or irrelevant and can thus be removed without incurring much loss of information.

When we talk about dimensionality reduction, we should note two notions; There is feature depend on another feature that is strongly correlated. From that, we should know the difference between feature selection and feature extraction. Feature selection is a technique to select important features or a subset of features, but in feature extraction generates new features from existed features. Then there are two categories of dimensionality reduction: linear dimensionality reduction and non-linear dimensionality reduction. Examples of linear dimensionality reduction are Principal Component Analysis (PCA), Isomap, locally linear Embedding, etc. And non-linear dimensionality reduction examples are t-distributed Stochastic Neighborhood Embedding (t-SNE), UMP, etc. So, we will talk with more details about PCA and t-SNE to show what the importance of these models.

4.2.2 Linear Dimensionality Reduction Technique (PCA)

PCA is a technique in modern data analysis like in diverse the fields from neuroscience to computer graphics. That is a simple, non-parametric method, which is used to extract information from confusing datasets. We can use it to find a linear projection of our dimensional data, in such a way that the variance of the projected data is maximized. And so, here is an example of a typical map that you would get from PCA. From that, we conclude that the PCA provides a roadmap to reduce complex datasets. PCA is mainly concerned with preserving large pairwise distances in the map (Figure 4).

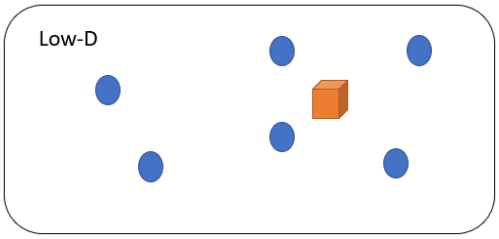


Figure 4: PCA Space

Particularly, the question; is PCA minimizing the right objective function? And if you think about PCA, it is mainly concerned with preserving large pairwise distances in the map. It is a way to maximize variance, or the same as to minimize a squared error between distances in the original data and distances in the map. And because you are looking at a squared error, you are mainly concerned with preserving distances that are very large. [15]

Our goal input data us characterized by user noise, this noise is the essence of a sample and hence is a good noise. This noise is basically variance of data and hence represents information. The goal of PCA is to reduce the dimensionality of data, while retaining as much of this variance as possible or retaining as much information simple enough. Let us have a dataset X, which contains p numerical variables. This data has a n-dimensional vectors,

$$X = [x_1 \ x_2 \ x_3 \ \dots \ x_n]$$

Using a linear transformation on input (X), we should be able to transform it into some (X_a) in the reduced feature space with maximum variance,

$$X_a = \sum_{j=1}^{p} a_j x_j$$

Where (a) is *Constant Vector*, $a = [a_1, a_2, \dots, a_p]$. We can detect the variance of each linear combination by,

$$var(X_a) = a^T S a$$

Where (S) is the simplest *Covariance Matrix* of the dataset, then, we need to identify the linear combination with maximum variance is equivalent to obtaining a p-dimensional vector (a) with maximizing $a^T S a$ in the quadratic form. Now that we have an optimization problem with equality constraints, we can solve using the method of LaGrange multipliers. So, to maximize

$$a^T S a - \lambda(a^T a - 1)$$

Where (λ) is *LaGrange Multiplier*. Then we must differentiate with respect to the vector (a) and equal this result to zero,

$$Sa - \lambda a = 0 \leftrightarrow Sa = \lambda a$$

Where (a) is *Eigen Vectors* and (λ) is *Eigen Values* of the matrix (S). then we need to calculate the variance,

$$var(X_a) = a^T S a = \lambda a^T a = \lambda$$

From that, we conclude the total variance of the dataset in the linear condition is equivalent to the eigen vector (λ) with n-dimension. So, we need to calculate the retained variance that is [15],

$$\left(Retained\ Variance = \sum_{i=1}^{N} \lambda_i \right) and \left(\%_{Info} = \frac{\sum_{i=1}^{D} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \right)$$

4.2.3 Non-Linear Dimensionality Reduction Technique (t-SNE)

So as a motivation, all that PCA is doing is it wants to make sure that stuff that is dis-similar, that ends up for apart like the zeros and ones. But there is one important question; Is that really what we want in visualization? Particularly, are those large pairwise instances in the data, are those things that are reliable? The answer is no, because if you think about data in terms of nonlinear manifold, to switch role here, then you will see that in this case the Euclidean distance between two points on this manifold would not reflect very well their similarity. Because the distance between these two points is similar, whereas if you would consider the entire structure of the data. (*Swiss-role Problem*).

t-SNE is an algorithm in ML used to visualize high-dimensional data that depends on stochastic neighbor embedding. For example, we can use it to visualize *30* features, where the features are intensity vectors that represents images. It can be used to represent word-count vectors or documents have thousands of dimensions. It is like the locally linear technique, but it solves the problem of collapse all points onto a single point. And solves the problem of visualize the real high-dimensional data with preserving the local and global structure of data in a single map.

How this technique works? Consider here we are given a bunch of high dimensional objects, $\{x_1, x_2, \dots, x_N\}$. And we want to get what structure, sort of the underlying structure of data is. We want to know their certain clusters in there. And what is the more local structure of this data manifold that is formed by the body high-dimensional inputs? So, we need to answer these questions to solve previous problems in other techniques [16]. First, we need to measure the pairwise similarities between high-dimensional objects. In high-dimensional space, we are going to measure similarities between points (Figure 5).

$$P_{j|i} = \frac{\exp\left(-\frac{\left||x_i - x_j\right|\right|^2}{2\sigma_i^2}\right)}{\sum_{L\neq k}\exp\left(-\frac{\left||x_k - x_L||\right|^2}{2\sigma_i^2}\right)}$$

We will do that in a way that look at local similarities. So, it is similarities to nearby points. Let us assume the red square is a point (x_i). So, this is a point in the high dimensional space. And what we will do is to center a gaussian over this point. Then we are going to measure or compute the conditional distribution. Note that we do not normalize over all pairs of points but only over pairs of points that involve basically point (x_i). Second, we set the bandwidth (σ_i) such the conditional has a fixed perplexity. Because it allows us to set a different bandwidth for each point. And the way we are setting the bandwidth is basically in such a way that the conditional distribution has a fixed perplexity. So, you can think of this as basically scaling the bandwidth of the gaussian in such a way that a fixed number of points fall in mode of this gaussian. And we symmetrize the conditionals,

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2N}$$

What we do, it is sort of a hack, where we say the joint probabilities, so the distribution over pairs of points is just going to be the symmetrized version of the conditionals. That is given us the final similarities in the high dimensional space [16]. Third, we need to measure the pairwise similarities between low-dimensional map points,

$$q_{ij} = \frac{\left(1 + \left||y_i - y_j\right|\right|^2\right)^{-1}}{\sum_k \sum_{l\neq k}\left(1 + \left||y_k - y_l|\right|^2\right)^{-1}}$$

So, what are we going to do is now we are going to look at the low dimensional space? So, this is a two- or three-dimensional space. And we are going to layout points in that map. So, the red square in figure 2 in the low dimensional space, it will be called (y_i). And we are going to center some kernel over this point. Then we measure the density of all other points (y_j) under that distribution. And again, we are going to renormalize by dividing over all pairs of points. And what this gives this is a probability (q_{ij}), which is basically measures the similarity of two points in the low dimensional map. Then we want to do is what we want these probabilities (q_{ij}) to reflect the similarities (p_{ij}) which we compute in high dimensional space [16].

If the $q_{ij}(s)$ are physical identical to the $p_{ij}(s)$, then apparently the structure of the map is very similar to the structure of the data in the original high dimensional space. To measure the similarity between the low dimensional and high dimensional space we use the Kullback-Leibler (K-L) divergence,

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_{j\neq i} P_{ij}\log\frac{p_{ij}}{q_{ij}}$$

That is sort of the standard measure of natural divergence, natural distance measure, between probability distributions. And it takes the form to sum over all pairs of points. We want to layout this point in the low dimensional space in such a way that q_{ij} values are similar as possible to the p_{ij} values. And to do that, we are basically doing gradient descent in this K-L divergence, which boils down to just moving to points around in such a way that this K-L divergence becomes small.

$$\frac{\partial C}{\partial y_i} = 2\sum_j \left(p_{(j|i)} - \mathrm{q}_{(\mathrm{j}|\mathrm{i})} + \mathrm{p}_{(\mathrm{i}|\mathrm{j})} - \mathrm{q}_{(\mathrm{i}|\mathrm{j})}\right)\left(\mathrm{y_i} - \mathrm{y_j}\right)$$

Well, this is basically having to do with the asymmetry of the K-L divergence. If you think about us having two similar high dimensional points, the two high dimensional points that are close together, these points will have a large p_{ij} value.

4.3 Clustering

So, why clustering? As a motivation too, clustering is an important technique for "Unsupervised Data" and our data is a simple example for that, but that is not the only reason that makes us turns our mind to this technique. The answer is for the following; As the colors in the t-SNE image are continuous and close colors

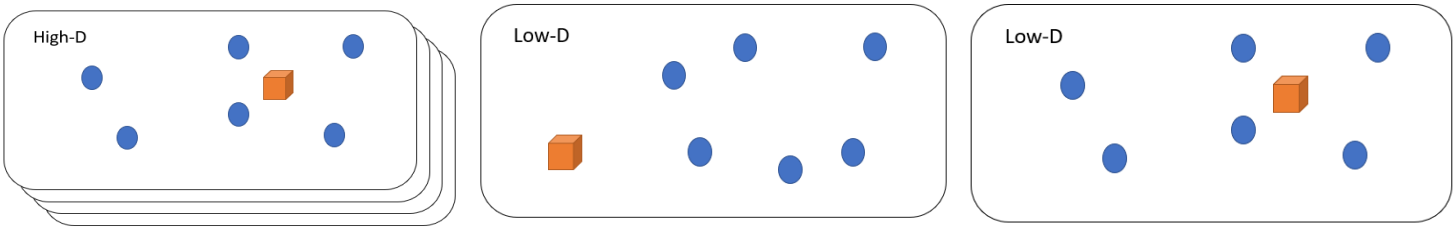


Figure 5: t-SNE Space

represents close datapoints in the t-SNE space we used clustering to color the datapoints according to the cluster it belongs to as each cluster represents a discrete color so, we can determine the clinical outcome of each cluster.

4.3.1 K-means Clustering

K-means clustering algorithm is an unsupervised learning algorithm which splits unlabeled datasets into number of clusters based on their properties such that datapoints with similar properties will be in the same cluster. K random centroids were initiated first in the data, where k is the number of clusters chosen. let us choose $k = 2$ then we will have *2* initial centroids (Figure 6). Then assigning every datapoint to its nearest centroid by calculating the Euclidean Distance (ED). Then moving every centroid to the average of the datapoints assigned to it by calculating the center of gravity (Figure 7) and repeating these steps until the position of the centroids is not changing [17].

$$ED = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

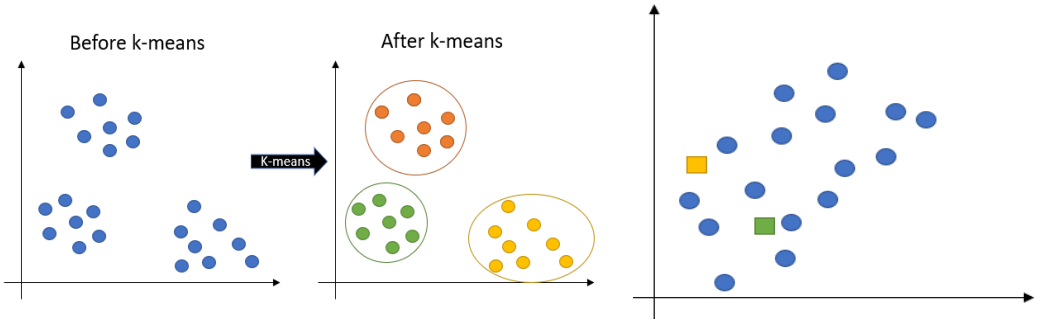


Figure 6: K-means Clustering (Left) & K-means Initial (Right)

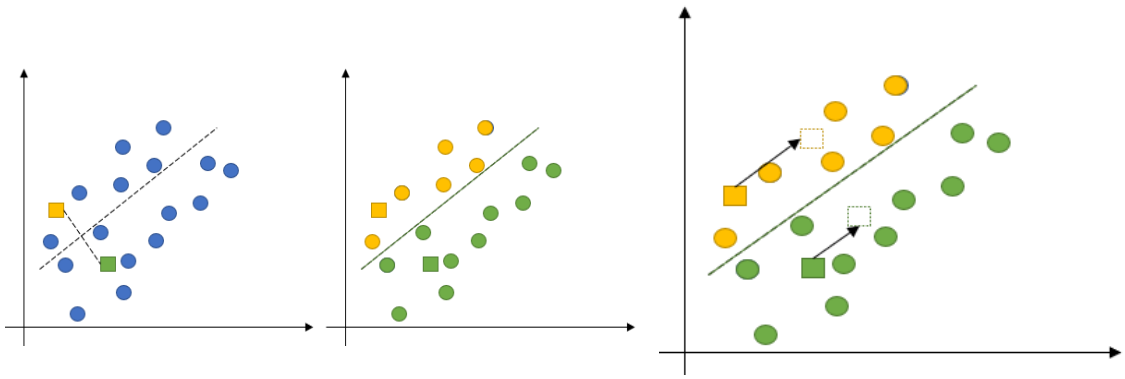


Figure 7: Assigning Datapoints to Closest Centroid (Left) & Moving K-means Centroids (Right)

4.4 Microarray Analysis Technique

Microarray analysis is the final step in reading and processing data produced by a microarray chip (as in our case "MALDI Data"). Process done we do not care about them for now, all we want to focus on is the significant behind this technique. We can determine a lot of impressive things from data using this technique (For example: significant DNA responsible for a specific type of cancer, newborn pedigree, etc.). Here, we will use it for determining the significant protein (Biomarker). - See next section (SAM Technique) -

4.4.1 Significance Analysis of Microarrays (SAM)

It is a statistical technique for finding significant genes in a set of microarray experiments. The input is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment (Table 4.4.1). It computes a statistic (d_i) for each gene (i), measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cutoff for significance is determined by a tuning parameter *delta*, chosen by the user based on False Positive Rate (FPR) [18]. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

Notes: [18]

- **Quantitative:** Real-valued (e.g., Blood Pressure)
- **Two Class (Unpaired):** *2* sets of measurements, in which the experiment units are all different in *2* groups. (e.g., Control and treatment groups with samples from different patients).
- **Multiclass:** More than two groups, each containing different experimental units. (Generalization Case)
- **Survival Data**: Time until an event (e.g., Death, Relapse), Possibly Censored.

Table 1: Some Formats of Response Variables in SAM

Response Type	Coding
Quantitative	Real Number (e.g., 27.4)
Two Class (Unpaired)	Integer (e.g., 1)
Multiclass	Integer (e.g., 1)
Survival Data	(Time, Status) Pair (e.g., (50, 1)) **Status:** (1): Died, (0): Censored

SAM Algorithm

SAM calculates a *test statistic* for relative difference in gene expression based on permutation analysis of expression data and calculates a false discovery rate. The principal calculations of the program are illustrated below:

$$d_i = \frac{r_i}{s_i + s_0} : i = 1,2, \dots, p$$

$$r_i = \frac{\sum_j y_j \left(x_{ij} - \sum_j \frac{x_{ij}}{n}\right)}{\sum_j \left(y_j - \hat{y}_j\right)^2}$$

$$s_i = \frac{\sqrt{\frac{\sum_j \left(x_{ij} - \hat{x}_{ij}\right)^2}{n-2}}}{\sqrt{\sum_j \left(y_j - \hat{y}_i\right)^2}}$$

The s_0 constant is chosen to minimize the coefficient of variation of d_i. r_i is equal to the expression levels (x) for gene (i) under (y) experimental conditions.

$$FDR = \frac{Median\ (90^{th}\ Percentile)\ of\ No.of\ Falsely\ called\ genes}{No.of\ genes\ called\ significant}$$

Source (Wikipedia): Microarray Analysis Techniques

4.5 Cross Validation (CV)

It is a highly ranked technique among other ML algorithms, it is considered as a beast in the world of ML, and we will have a quick review about it in our research but will never use it due to the shortage of time and resources. It is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. The most common types of CV and their related data resampling methods. [19] CV is like the repeated random subsampling method, but the sampling is done in such a way that no two test sets overlap.

4.5.2 K-fold CV

In K-fold CV, the available learning set is partitioned into (k) disjoint subsets of approximately equal size. Here, "fold" refers to the number of resulting subsets. This partitioning is performed by randomly sampling cases from the learning set without replacement. The model is trained using $(k-1)$ subsets, which, together, represent the training set. Then, the model is applied to the remaining subset, which is denoted as the validation set, and the performance is measured. This procedure is repeated until each of the (k) subsets has served as validation set. The average of the (k) performance measurements on the (k) validation sets is the cross-validated performance. (e.g., for $k = 10$, i.e., 10-fold CV). In the first fold, the first subset serves as validation set $D_{val,1}$ and the remaining nine subsets serve as training set $D_{train,1}$. In the second fold, the second subset is the validation set and the remaining subsets are the training set, and so on.

The cross-validated accuracy, for example, is the average of all ten accuracies achieved on the validation sets. More generally, let ˆf−k denote the model that was trained on all but the (k^{th}) subset of the learning set. The value $\hat{y}_i = \hat{f}_{-k}(x_i)$ is the predicted or estimated value for the real class label, (y_i), of case (x_i), which is an element of the (k^{th}) subset. The cross-validated estimate of the prediction error, $\hat{\epsilon}_{cv}$, is then given as.

$$\hat{\epsilon}_{cv} = \frac{1}{n}\sum_{i=1}^{n} ב\left(y_i, \hat{f}_{-k}(x_i)\right)$$

CV often involves stratified random sampling, which means that the sampling is performed in such a way that the class proportions in the individual subsets reflect the proportions in the learning set.

4.5.1 Leave-Out-Patient-Out (LOPO)

It is a CV technique which depends on hide only one sample from the dataset as a test-set and the remaining become training-set, then test the model on this state. It is useful when dealing with small dataset and its specialization is mainly in cancer research and bioinformatics.

Chapter5: Applications (Gastric/Breast Cancers)

5.1 Data Exploratory

We begin this part by defining some notations used throughout the thesis; we worked on two types of gastric cancer data 'Gastric Data', and breast cancer data 'Breast Data'. These two types could be divided from one category, which is 'MSI Data' as described previously. Our MSI data could be considered as pixel arrays that are high-dimensional data. There are many properties of this type of data that could be defined or explained by the fact 'units that are partially redundant'. The redundancy in high-dimensional data means that there are parameters or features that can characterize different units are dependent on each other. From that, we must understand all units in data that require taking the redundancy into account. Gastric Data is a pixel array, or a group of gastric samples taken from *63* patients who had gastric cancer. and Breast Data is a group of breast samples taken from *32* patients, who are in differently risky stages.

For both data, we applied different methods (MSI, ML Concepts, and Prognosis Analysis) that conclude us to our result. Exploring Gastric Data was the first thing to be done to understand chemistry behind it, reaching to SAM technique. Now, let us discuss each step of it then have a look at our results and interrupt it in a scientific way. As done in Gastric Data, we do the same in Breast Data except for the pre-final step we did a little bit change (Figure 8).

5.2 MSI Data

Gastric/Breast Data contain many variables such as 'HE_Image' which is a histological image for each sample, 'clinical data' for each patient that contained information like 'Sample ID', 'Survival time', 'Survival Status' and 'pN', etc. Malignant gastric cancer is a disease, which contains cells form in the lining of the stomach. Anything that increases the risk that are measured by the risk factor. And resulted from the growth out of control about cells. Breast cancer could be transferred from one place to another (Metastasis).

5.3 Dimensionality Reduction

5.3.1 PCA for Both Gastric/Breast Data

We apply PCA on MSI Data by projecting each data point 'pixel' onto the first few principal components to obtain lower-dimensional data while preserving as much of the variation of data as possible. Maximization of the variance of the projected data is the first principal component. The transformation of the MSI into three-dimensional representation maps the data vectors from an original space of p variables to the new space of p variables which are uncorrelated over the dataset. However, not all the principal components need to be kept. PCA is like other linear dimensionality reduction techniques focus on the results on the global characteristics of the data space. PCA can have the effect of concentrating much of the signal into the first few principal components, which can usefully be captured by dimensionality reduction; while the later principal components may be dominated by noise, and so disposed of without great loss.

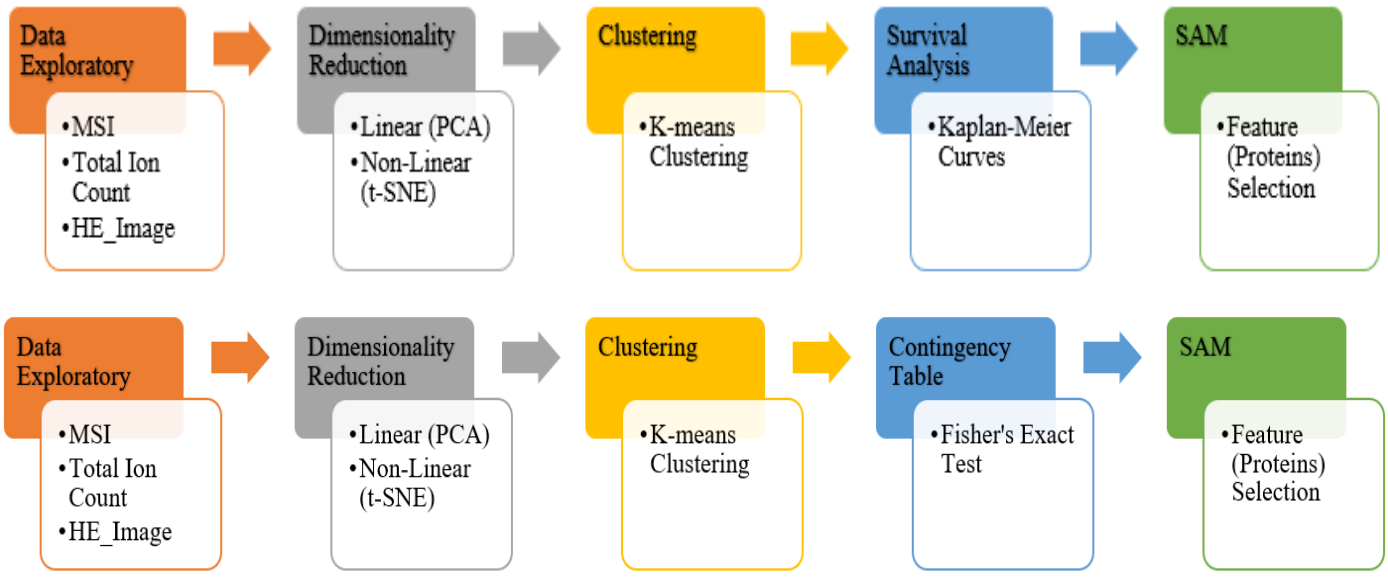


Figure 8: Steps for Gastric Data (Up) and Breast Data (Bottom) Analysis

This low-dimensional map was then used to enable visual exploration of the inter/intratumor heterogeneity and as input of the clustering algorithm. PCA focus on appropriately modeling large pairwise distances between protein expression profiles. The focus on modeling large pairwise distances came at the price of substantial errors in modeling small pairwise distances. However, it was exactly this local similarity structure that is essential in clustering and visual exploration: the goal of clustering is to find groups of nearby data points and, similarly, the goal of visual exploration is generally to determine which parts of the data are like a reference data point.

5.3.2 t-SNE for Both Gastric/Breast Data

As concept explained previously, we now will talk about its effect of Non-Linear Dimensionality Reduction of these data and the visualization in the low dimensional space. We advocate the using of t-SNE for some reasons; it had rapidly established itself as a method of choice for summarizing high-dimensionality datasets owing to its ability to overcome the crowding problem, in which some of the higher-dimensional data similarities could not be visualized in a single map space. The important component of t-SNE method that we used is the Barnes-Hut-SNE implementation because it is faster in analysis of large data, visualizing the molecular intratumor heterogeneity and with complexity $\left(O\left(Nlog(N)\right)\right)$ that is faster than non-linear algorithms $\left(O(n^2)\right)$.

t-SNE technique was used to reduce gastric and breast cancer data in a single map representation (three-dimensional space). This was done by representing every pixel of data in high dimension to one point in the single map. This technique makes what we told previously in processing of the distances between neighboring data points using the non-linear embedding. This technique used a stochastic optimization algorithm which runed multiple times. So, to get the samples in a fully map, we had to run the t-SNE algorithm for two times; firstly, with default settings and fixed seed point of zero. Secondly used the reduced map of a first time as an initialization of the second t-SNE running. We did that because we had to achieve global convergence.

Then we applied the control check in result to know if there a systematic bias due to patient selection or not. We found that and according to it each point in space was colored related to their cancer subtype and measurement date. That indicated to not exist prominent batch effects visible in t-SNE map.

But in this part, there were some important notices like the significant of data, seed points and the effect of different operating systems. These notices we are going to talk about with more details. The problem in seed point is when we runed the algorithm, it chose random seed point which resulted different values of global convergence. This value leads effects in significantly of clusters and SAM analysis. And when we runed the algorithm in MATLAB you got different results compared to the result of python because the value of convergence in MATLAB was more suitable in significant of data. So, we used MATLAB to get this suitable dimensionality reduction. And tried to get the similar result about MATLAB.

5.4 Clustering

After applying t-SNE we applied k-means clustering on the t-SNE map with $random\_state = 0$ to start with same random datapoints as centroids every time to make the results reproducible. So, each pixel in a patient were assigned to a cluster then we assigned each patient to a cluster according to the major clusters in this patient in which if $\left(\frac{1}{k}\right) * 100\%$ of the pixels are assigned to a cluster then the patient is assigned to that cluster and some patients are assigned to more than a cluster. So, every cluster will have several patients assigned to it. We choose the number of clusters based on the k-means image with the highest correlation with the t-SNE image.

For Gastric Data: The highest correlation was at $k = 3$ as concluded from the edge correlation curve (Fig. 1D) [20], so we are having 3 clusters or 3 different stages of gastric cancer patients each with different clinical outcome that we will determine the significance between them and their survival time in the survival analysis section.

For Breast Data: The highest correlation was at $k = 8$ and the second and third K-means images with highest correlation with the t-SNE image were at $k = 7$ and $k = 6$ as shown in the edge correlation curve (Fig. 3B) [20]. So, we are having 8 clusters in which we will see their relevance to the metastatic status in the Fisher's exact test section.

5.5 Survival Analysis

We applied K-M curve on Gastric Data to notice the significant between clusters and determine which cluster has the low survival rate among the *3* clusters by set a significance level with a P-value below 0.05 as approved in statistics science (Null Hypothesis).

For Breast Data, we made contingency tables. Contingency table was applied to a breast cancer dataset of primary tumors from *32* patients, of whom *21* had lymph node metastasis $(pN = 1)$ and *11* were metastasis-free $(pN = 0)$. Firstly, we assembled our dataset into a table (the first column includes the sample ID, the second one states whether there's lymph node metastasis or not (pN) and the third one determines the cluster of each sample). Then, we made a contingency table of each k-means dataset. Both methods are done using R programming language.

5.6 SAM

For more investigation on both Gastric/Breast Data to determine the biomarkers (significant proteins) to find a method to treat cancer, we establish microarray analysis technique (Specially SAM technique) and applied it on both data after determining the significant clusters of data.

Chapter6: Results (Gastric/Breast Cancers)

6.1 MSI Data

Figure 9: Gastric Data (HE Image)

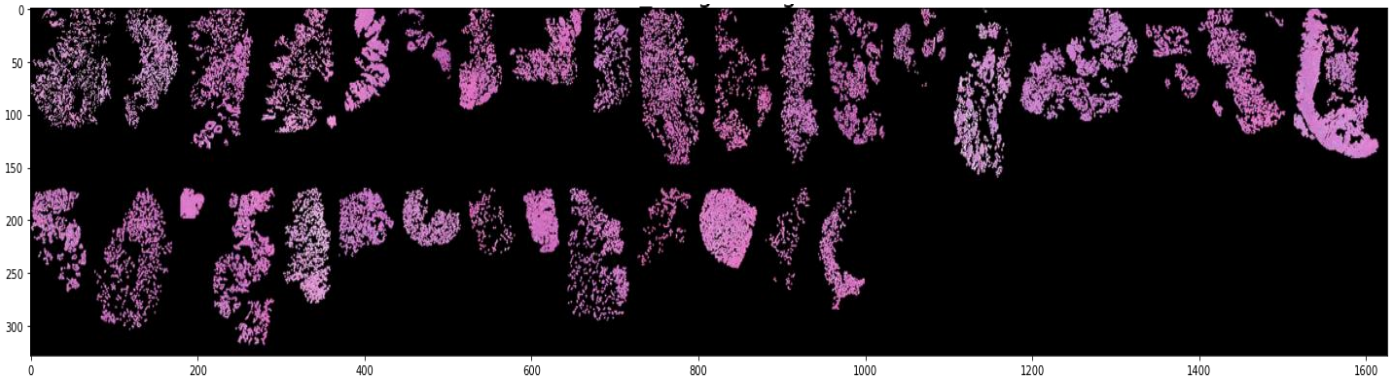


Figure 10: Breast Data (HE Image)

6.2 Dimensionality Reduction (t-SNE)

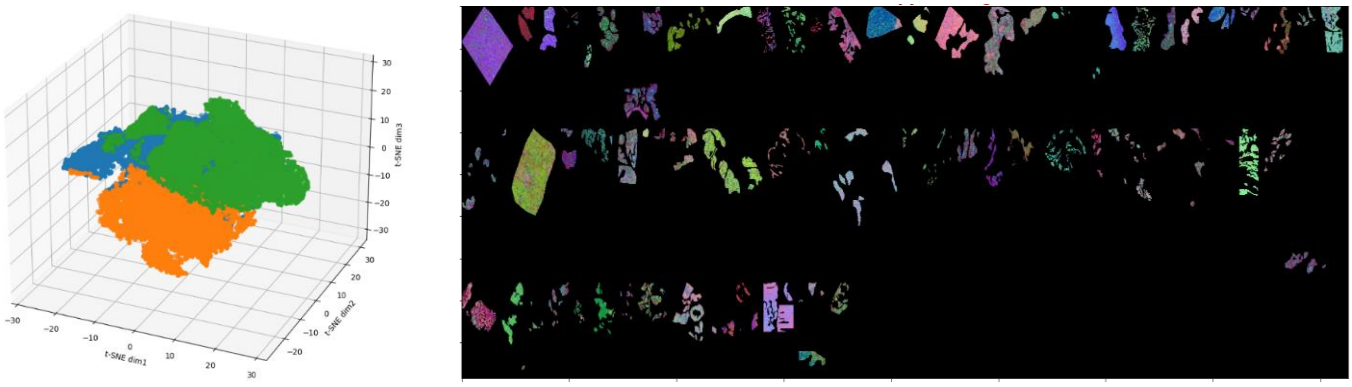


Figure 11: Gastric Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right)

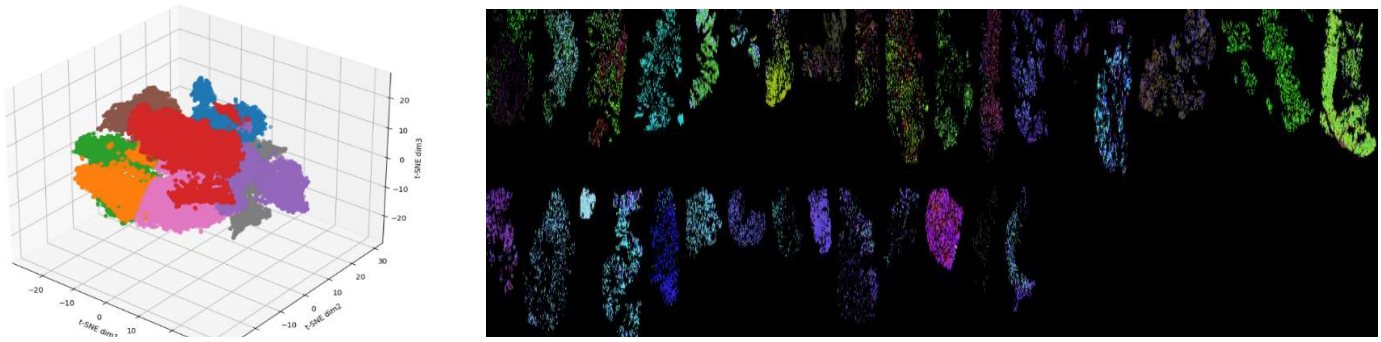


Figure 12: Breast Data t-SNE Scatter Space (Left) & t-SNE Spatial Image (Right)

6.3 Clustering (K-means Clustering)

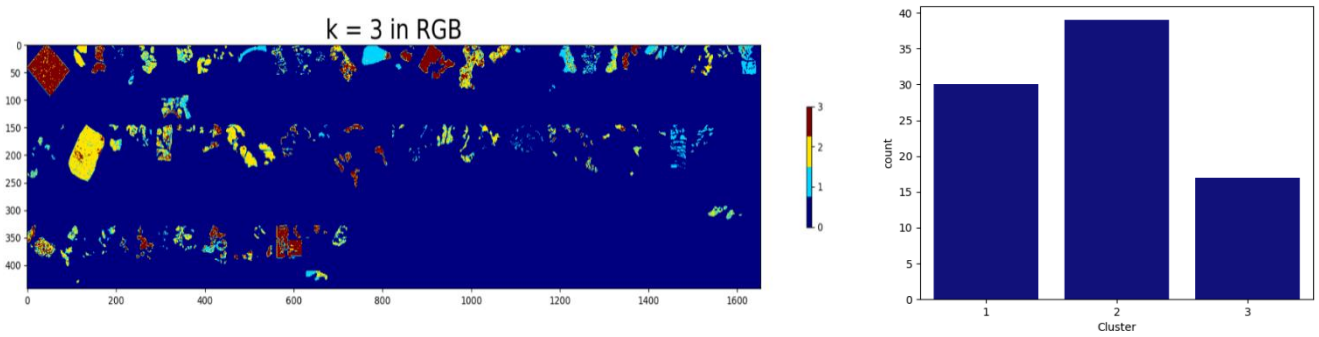


Figure 13: Gastric Data K-means Spatial Image (Left) & Count Plot of Patients in each Cluster (Right)

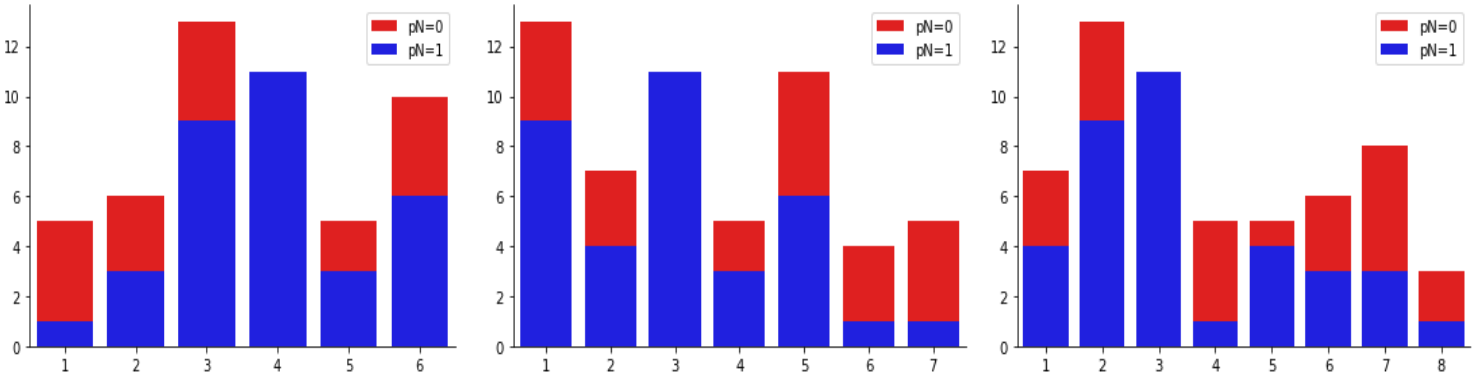


Figure 14: Breast Data K-means Count Plot with Metastasis Colored $(k = 6)$, $(k = 7)$ & $(k = 8)$

6.4 Survival Analysis (Kaplan-Meier Curves)

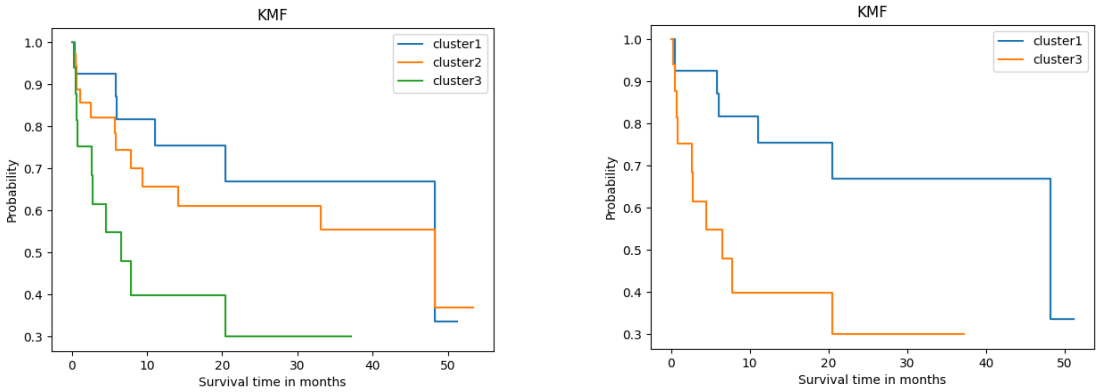


Figure 15: Gastric Data K-M Curve between Clusters (Left) & Significance of 1 vs 3 (Right)

Table 2: Contingency Table of Breast Data applied on different K Values

K Values	P-value
6	0.004824
7	0.005738
8	0.003732

Alternative Hypothesis: two.sided

6.5 SAM

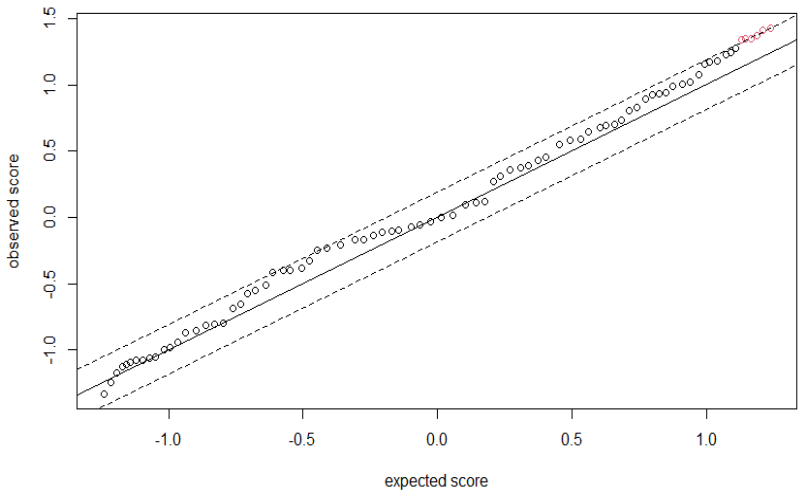


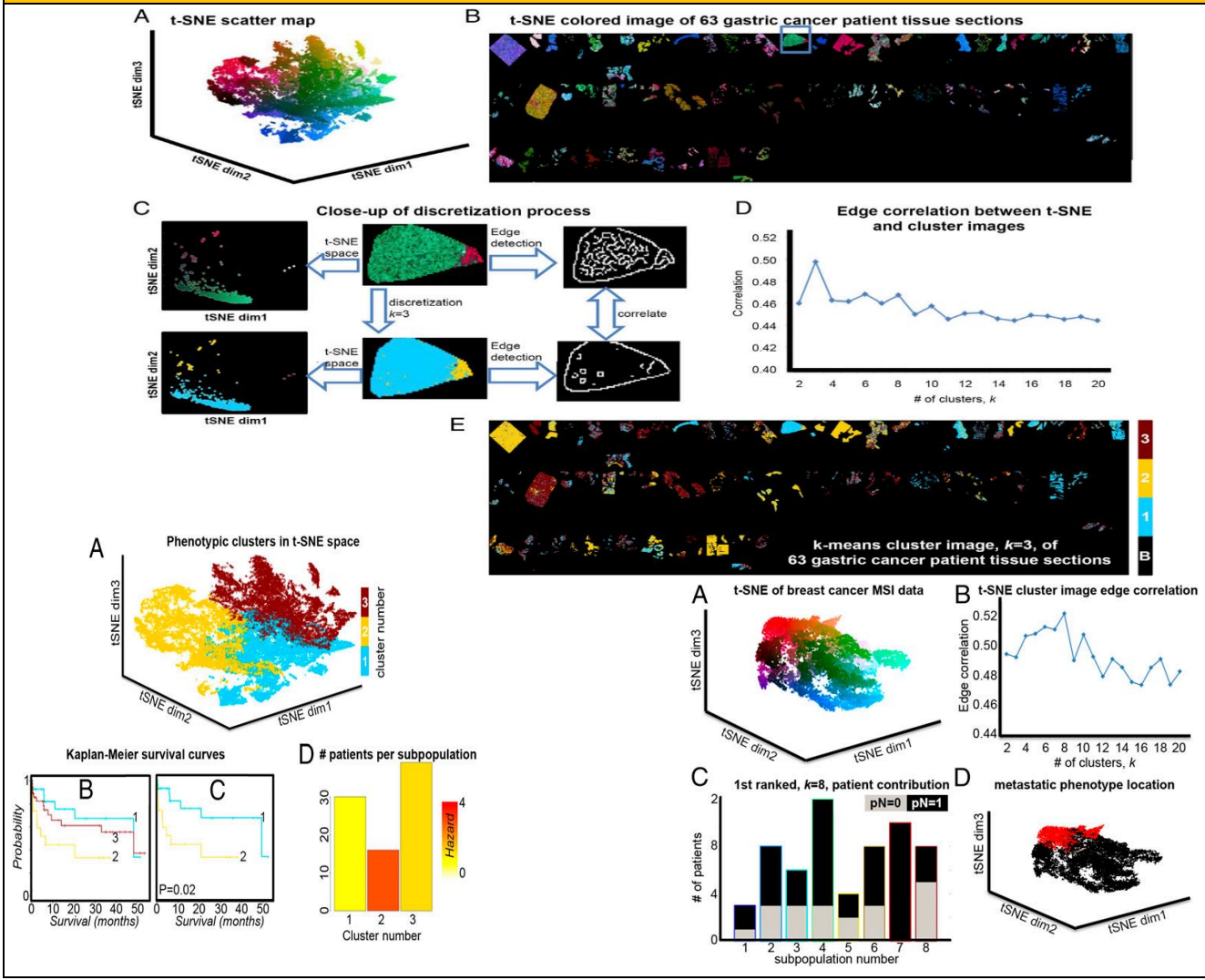
Figure 16: SAM applied on Gastric Data

Chapter7: Literature Review

Table 3: Findings of the reviewed sources

Author(s)	Year	Data Source	Method
Reference No. [20] **1.** Abdelmoula, Walid **2.** Balluff, Benjamin **3.** Englert, Sonja **4.** Dijkstra, Jouke **5.** Reinders, Marcel **6.** Walch, Axel **7.** Mcdonnell, Liam **8.** Lelieveldt, Boudewijn	2016	Link to Data	Tumor-specific signatures obtained by protein matrix-assisted laser desorption MSI analysis of primary tumors of gastric cancer $(n = 63)$ and breast cancer $(n = 32)$ were nonlinearly mapped to a 3D space using t-SNE. Using the perceptually linear LAB color map to color each pixel according to its position in the t-SNE space, a t-SNE colored image can be obtained that depicts regions characterized by similar mass spectral profiles with similar colors. To segment the image into a discrete number of clusters, bisecting k-means and edge correlation algorithms were applied. The resulting clusters, or tumor subpopulations, were then statistically compared with the patients' clinical data (survival for gastric cancer and lymph node metastasis for breast cancer) to identify the subpopulations statistically associated with patient phenotype. LOPO pixel-based and patient-based classifiers were built to cross-validate the identification of tumor subpopulations and patient outcomes.

Results



Discussion

MSI Data

In this research, the common industrial problem is cancer treatment. HE Image is a remarkable tissue section (Figure 9 & Figure 10). The recognition of the tumor subpopulations that affect the results of patients is important for better describing the changes in molecules, which chaperon tumor development and for enhancing the patient management. MSI has various properties that are suitable for this work, which is untargeted analysis. This analysis could analyze more molecular ions, which is applied to some tissue sections that are characterized as inexpensive and fast. We used MSI because it has an ability to detect tumor subpopulations in histologically identical regions of tumor tissue.

Dimensionality Reduction (t-SNE)

We have used non-linear dimensionality reduction based on t-SNE because it preserves the global and local similarity structure of the dataspace in the low dimensional representation. The superior representation of the t-SNE and gene expression data are shown previously. The non-linear nature of the t-SNE would also better prepare the data to recognize the mass spectrometry differences between tumor subpopulations. We can see the entire *63* patients, representing two types of tumor heterogeneity; inter/intratumor heterogeneity of MSI Data. And contribution of the molecular heterogeneity in Gastric Data due to intratumor heterogeneity and patient variability and it reveals clear structural separations based on molecular heterogeneity. To assess whether the structure revealed by t-SNE, which could be linked to clinical outcome and identify phenotypic tumor subpopulations. We used t-SNE technique because it preserves the local structure of data and does not need information to differentiate between the local differences due to various clones, various patient samples, or any measurement bias.

Then, we utilize the local image structure of the MSI data. After using t-SNE, we found some edges visualized in the sample that are natural boundaries of molecularly distinct subpopulations. To explore these edges, we converted the 3D t-SNE scatter map to LAB color space, then colored each pixel's data point in the scatter by using the Laboratory color space coordinates (Figure 11). As we knew previously from the Dataset section, Breast Data consists of *32* patients, *21* of them had lymph node metastasis ($pN = 1$) and *11* of them were metastasis free ($pN = 0$). Then, we applied t-SNE to focus analysis of tumor subpopulations. Data points had been colored based on their location in clusters. Pixels are represented in a single map space as we described previously. To explore these edges, we converted the 3D t-SNE scatter map, and this image is explored in LAB color space as in Gastric Data (Figure 12). From that image, we can see the spatial organization of similarities in local and global structure data space. It reflects intratumor heterogeneity and patient variation.

Clustering (K-means Clustering)

The *3* different clusters of the Gastric Data (Figure 13); blue color represents the background, and each color represents a cluster. We are having *32* patients in cluster 1, *13* patients in cluster 2 and *34* patients in cluster 3, so a total of *79* patients which is more than our number of patients and that shows us that some patients are assigned to more than one cluster. The different clusters of Breast Data in which each cluster have several *metastatic* and *non-metastatic* patients where $k = 8$ We found cluster *3* full of metastatic patients only (Figure 14).

Survival Analysis (Kaplan-Meier Curves)

The process involves calculations of probabilities of event occurrence at a certain point multiplied by the successive probabilities computed earlier in time. [4] K-M survival curves for the subpopulations represented by *3* clusters and by applying a log rank test between them; we found the greatest significant difference in survival between the subpopulations in clusters 1 and 3 with P-value of 0.02 (Figure 15). From the output we see that the p-value is less than the significance level of 5%. Like any other statistical test, if the p-value is less than the significance level, we can reject the null hypothesis [21]. In our context, rejecting the null hypothesis for Fisher's exact test of independence means that there is a significant relationship between the two categorical variables (lymph node metastasis and the assigned clusters). Therefore, knowing the value of one variable helps to predict the value of the other variables.

SAM

The significance level of SAM output could be found at the high end of SAM plot revealing the expected score along the line of interest (Figure 16).

Conclusions and Future Work

Intratumor heterogeneity is a key factor in tumor progression, affecting patient outcomes and treatment. Tumor subpopulations can be histologically indistinguishable but still have molecular phenotypes that drive tumor progression and determine disease outcome. The identification of these clinically relevant tumor subpopulations is of utmost importance for understanding cancer development and the management of cancer patients. Although localized genomic techniques have established branched evolution of tumors and single-cell transcriptional heterogeneity, the cost and throughput of these techniques are prohibitive for large scale multi-site sequencing of patient tissues. The automated identification of phenotypic tumor subpopulations reported here will allow better targeting of these powerful genomic methods to those subpopulations that are statistically associated with patient outcomes.

Future Work: We shall now establish a technique to distinguish between biomarkers we got from SAM analysis, and to do that we have first to classify them using any classification technique (e.g., KNN) and then we will have to differ our data (As explained previously, our data is relatively small, but significant) so, we will use a technique known as Leave-One-Patient-Out (LOPO) and apply the same procedures on data extracted. That will give us a strong evidence for our result and solve the problem of small data.

References

[1] Schober, Patrick, and Thomas R. Vetter. "Survival analysis and interpretation of time-to-event data: the tortoise and the hare." Anesthesia and analgesia 127.3 (2018): 792.

[2] Klein, John P., and Melvin L. Moeschberger. Survival analysis: techniques for censored and truncated data. Vol. 1230. New York: Springer, 2003.

[3] Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times. In Practical statistics for Medical research; pp. 365–93.

[4] Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research, 1(4), 274–278.

[5] Bland, J. M., & Altman, D. G. (2004). The log rank test. BMJ (Clinical Research Ed.), 328(7447), 1073.

[6] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery, 143(3), 331–336.

[7] Schober, P., & Vetter, T. R. (2021). Kaplan-Meier curves, log-rank tests, and cox regression for time-to-event data. Anesthesia and Analgesia, 132(4), 969–970.

[8] Beers, B. (2021, July 7). P-Value. Retrieved July 17, 2021, from Investopedia.com website: https://www.investopedia.com/terms/p/p-value.asp

[9] Masashi Sugiyama, in Introduction to Statistical Machine Learning, 2016

[10] Karl Pearson, F.R.S. (1904). Mathematical contributions to the theory of evolution. Dulau and Co.

[11] Weisstein, E. W. (n.d.). Fisher's exact test.

[12] Agresti, Alan (1992). "A Survey of Exact Inference for Contingency Tables". Statistical Science. 7 (1): 131–153.

[13] Kim, Hae-Young. "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test." Restorative dentistry & endodontics 42.2 (2017): 152-155.

[14] Buchberger, A. R., DeLaney, K., Johnson, J., & Li, L. (2018). Mass spectrometry imaging: A review of emerging advancements and future insights. Analytical Chemistry, 90(1), 240–265.

[15] J. Cadima, "Principal component analysis: a review and recent developments," THE ROYAL SOCIETY PUBLISHING, vol. 16, 2016.

[16] L. v. d. Maaten, "Visualizing non-metric similarities in multiple maps," Springerlink, vol. 23, 2011.

[17] Sinaga, Kristina & Yang, Miin-Shen. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2988796.

[18] Chu, G., Seo, M., Li, J., Narasimhan, B., Tibshirani, R., & Tusher, V. (n.d.). Users guide and technical document. Retrieved July 18, 2021, from Stanford.edu website: http://statweb.stanford.edu/~tibs/SAM/sam.pdf

[19] Berrar, D. (2019). Cross-Validation. In Encyclopedia of Bioinformatics and Computational Biology (pp. 542–545). Elsevier.

[20] Abdelmoula, Walid & Balluff, Benjamin & Englert, Sonja & Dijkstra, Jouke & Reinders, Marcel & Walch, Axel & Mcdonnell, Liam & Lelieveldt, Boudewijn. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proceedings of the National Academy of Sciences. 113. 10.1073/pnas.1510227113.

[21] Bower, Keith M. 2003. "When to Use Fisher's Exact Test." In American Society for Quality, Six Sigma Forum Magazine, 2:35–37. 4.

Appendix A

Link to Git-hub Repo: Mass Spectrometry Imaging in Detecting Tumor Heterogeneity
Link to Website: Main Website

Table A.1: Clinical Outcome of Gastric Data (Brief)

Sample_ID	T	N	M	Surv_time	Surv_status
1	2	0	'x'	1027	0
2	2	0	'x'	20	1
3	2	0	'x'	13	0
4	2	0	'x'	1350	0
5	2	0	'x'	1006	1

Table A.2: Clinical Outcome of Breast Data (Brief)

Sample_ID	pN
1	1
2	1
3	1
4	1
5	1