Machine Learning and Mass Spectrometry Imaging: Application in Tumor Heterogeneity

Hendwan Abozide, Youssef Saleh, Rokaia Mohammed

Cairo University, Faculty of Engineering, Department of Healthcare Engineering and Management

Under the Supervision of Dr. Walid Abdelmoula, Research Scientist at Brigham and Women's Hospital / Harvard Medical School

***Abstract* — Identification of tumor subpopulations that detriment the disease prognosis and the patient outcomes is essential for personalized therapy. Studies showed that Mass spectrometry imaging (MSI) could visualize intra-tumor heterogeneity and helped identify diagnostic and prognostic biomarkers. We studied the methods introduced by Abdelmoula et al. and used it as reference for our work [8]. Our goal was to study the intra-tumor heterogeneity of gastric and breast cancer data. We implemented our project using Python instead of using MATLAB that was used by our main reference. We used machine learning techniques and statistical methods to identify the significant tumor subpopulations that affect the patient status. Dimensionality reduction was performed using t-distributed stochastic neighbor embedding (t-SNE) and K-means clustering was applied to discretize the data. Statistical analysis was done on the discretized data to deduce the significant clusters, hazardous cluster in gastric cancer data and metastatic cluster in breast cancer data. Significance Analysis of Microarrays (SAM) was used to reveal the statistically significant protein ions with FDR < 0.001 associated with these significant clusters. We built a predictive model using the Support Vector Machine (SVM) to predict the survival status of gastric cancer patients and the metastatic status of breast cancer patients.**

I. Introduction

Mass spectrometry-based Imaging (MSI) was introduced as a technique that analyzes the complex-protein molecules in the tissue sample[1], [2]. The most common technique used in the clinical applications of MSI is Matrix-Assisted Laser Desorption/Ionization (MALDI) that can reveal the early onset of cancer biomarkers. Biomarkers are unique features in the tissue that can help in identifying the severity of the cancer tissue [3]. Histology is the study of microscopic anatomy of cells and tissues and is used to identify the severity of cancer tumors as it is considered the gold standard for almost all diagnosis of cancer diseases [4]. However, histological images can only show cancer biomarkers in later stages when compared to MSI.

The MALDI-MSI technique can be simplified as using a laser to irradiate a region of interest (pixel-by-pixel), which is coated by a special chemical matrix in the tissue sample to facilitate the ionization of molecular compounds [1]. It is capable of identifying proteins that could contribute to spatial variations in the tissue such that these variations could be linked to certain disease phenotypes and would allow for the possibility of personalized-patient therapy. The produced ions from MALDI are accelerated towards the detector and are sorted in their respective mass to charge ratio (m/z) values [1], [2]. A spectrum of m/z values from each pixel is produced, and the m/z values are considered the physical characterization of the released ions [5].

Cancerous tumor tissues contain within them different subpopulations which are distinct from one another [6]. This distinction causes different tissue phenotypes as well as influencing the cancer evolution. This is defined as tumor heterogeneity; it consists of intra-tumor and inter-tumor heterogeneity, where the tumor subpopulation differing between different samples is inter-tumor heterogeneity and different tumor subpopulations existing in the same sample is intra-tumor heterogeneity [6]. MALDI-MSI is used to visualize the intra-tumor heterogeneity that is linked to distinct tissue phenotype which helps in tracking tumor progression which could ultimately help in determining the patient status outcome [7].

It was stated by Abdelmoula et al. [8] that the intra-tumor heterogeneity is commonly visualized using dimensionality-reduction techniques applied on the MSI data. These techniques are divided into linear and non-linear methods; one of the most common linear methods is Principal Component Analysis (*PCA*), and one of the most common non-linear methods is t-distributed stochastic neighbor embedding (*t-SNE*). PCA aims to represent the variance and preserve the global details of the data [9], [10], while t-SNE focuses on representing the similarities between the data points as well as preserving both the global and local details of the data [8], [11]. Moreover, t-SNE can clearly represent the intra-tumor heterogeneity and show the molecular differences between tumor subpopulations that contribute to the cancer evolution [8]. There's also a present issue that faces any complex high-dimensional data which is the *Curse of Dimensionality* [12]. It is described as the increase of the complexity of the data due to the increase in the number of the dimensions (features), which could deteriorate the clustering and classification efficiency. The non-linear method (t-SNE) can overcome this issue as it can accurately represent the high dimensional data in a lower-dimensional space in 2D and 3D, when compared to linear method (PCA) [11], [13].

In our project, we applied the Barnes-Hut implementation of t-SNE for faster analysis on our biological datasets (gastric and breast cancer) [13] to visualize the intra-tumor heterogeneity. We then applied machine learning techniques and statistical approaches to link the t-SNE map with the clinical ground truth of the patient's status to create a machine-learning model that could predict the patient's

survival or metastatic status. We benchmarked our results based on our reference by Abdelmoula et al. [8] and implemented the project using Python. We have publicly released our implementation in GitHub. (Link: https://github.com/Youssef-Saleh/GP-MSI).

II. Methods & Results

The two MALDI-MSI datasets we used (gastric & breast) were taken from tissue sections of patients as described in our main reference [8]. The two datasets consisted of 63 samples of gastric cancer tissues and 32 samples of breast cancer tissues. The gastric cancer and breast cancer datasets have 82 and 62 protein ions (i.e., *m/z* ions), respectively. These number of proteins represent the number of dimensions at each pixel. To assess whether to use linear or non-linear dimensionality reduction method, we compared PCA and t-SNE in representing the complex multi-dimensional MSI data. PCA was applied on the MSI data as shown in Fig. 1A and 1C and the transformed data 's variance was not sufficient to represent the data accurately (see the percentage of the explained variance in Fig. 1).

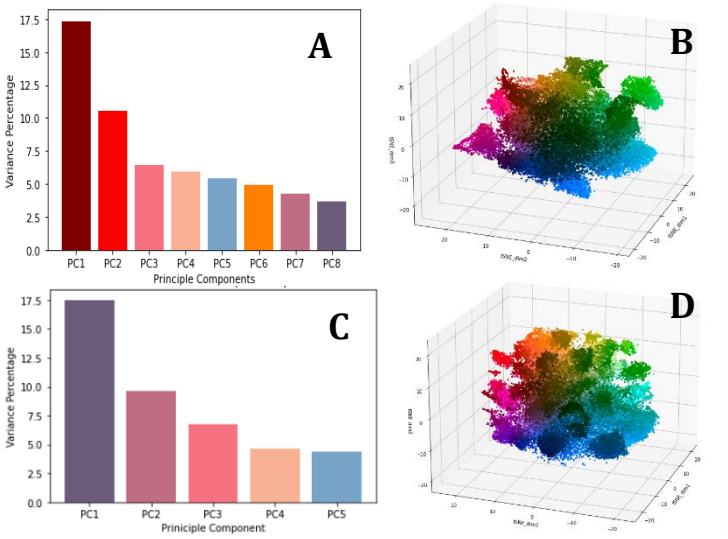

*Fig. 1. **PCA vs t-SNE:** (A) PCA applied on breast Cancer Data showing the 1st 8 principal components representing 58.53% of the original data. (B) t-SNE scatter map of the breast cancer represented in lower dimensional space. (C) PCA applied on gastric cancer data showing the 1st 5 principal components representing 42.53% of the original data. (D) t-SNE scatter map of the gastric cancer represented in lower dimensional space.*

The t-SNE maps shown in Fig. 1B and 1D represented the dataset in lower dimensional space and showed the inter-tumor and intra-tumor heterogeneity. The t-SNE map was colored using the LAB color space; we interpreted the t-SNE coordinates as the LAB color coordinates. This showed that there are unique distributions in the data. These distributions were shown in a continuous domain, and it was therefore difficult to identify the tumor subpopulations that contributed to the intra-tumor heterogeneity. Therefore, we need to transform it into a discrete space through data clustering.

For better visualization of tumor subpopulations, discretization of the t-SNE maps was applied using K-means clustering method. The t-SNE map of gastric cancer was clustered from K=3 to K=8 and the t-SNE map of breast cancer was clustered from K=3 to K=10. The optimum K clusters was deduced by our main reference [8], which is K=3 in gastric cancer and K=8 in breast Cancer. However, when we implemented the statistical analysis on the gastric data, it showed that K=4 had more significant statistical results.

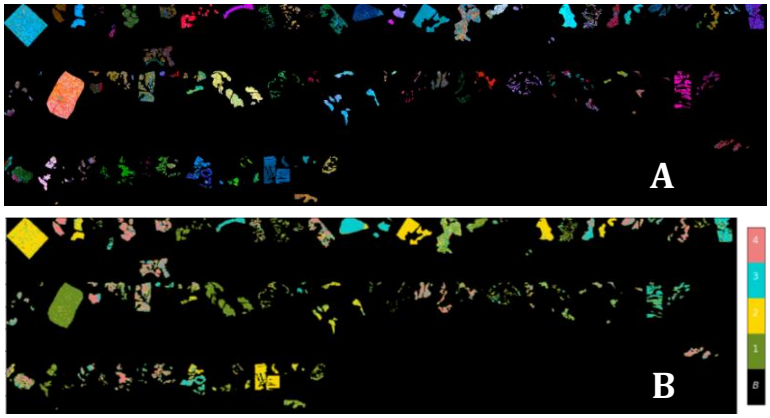

*Fig. 2. **Colored images showing all gastric cancer samples:** (A) shows the t-SNE colored image of the 63 gastric cancer samples. (B) shows K=4 clustering of the 63 gastric cancer tissues. Each color label in the color bar represents a cluster.*

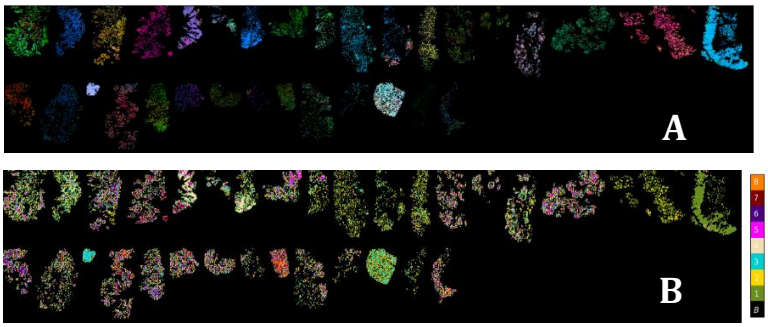

*Fig. 3. **Colored images showing all breast cancer samples:** (A) t-SNE colored image of the 32 breast cancer samples. (B) K=8 clustering of the 32 breast cancer tissues. Each color label in the color bar represents a cluster.*

It was demonstrated in Fig. 2B and 3B that the t-SNE clustered images showed that the discretization was able to represent different clusters (tumor subpopulation) more visibly within the samples. K-means clustering was able to highlight the boundaries between different clusters, as illustrated in Fig. 4A and 5A, unlike the t-SNE maps in Fig. 1B and 1D where the samples' distributions were intertwined.

Next, we were interested in determining the significant clusters in both the breast and gastric samples. The significant clusters for each sample were defined as clusters that contributed with more than $^1/_K$ of the total sample pixels, such that K is the number of clusters. In the results shown in this paper, the gastric cancer significant threshold was $^1/_4$ and in breast cancer significant threshold was $^1/_8$. After linking the patient samples to their significant clusters,

we were also interested in linking the patient samples' significant clusters to the patient's clinical ground truth. The clinical ground truth in gastric samples is clinical survival time in months and censored survival status for unknown or confirmed dead, while it is the metastasis status in breast samples.

A. Survival Analysis of Gastric Cancer

To differentiate between the clusters of the t-SNE clustered gastric cancer image. We applied the following statistical methods; Kaplan-Meier, log rank test and Cox regression. Kaplan-Meier is used to calculate the survival probability of the patient to a certain point of time [14]. To identify significant differences between the clusters, we used log rank test. It calculates a probability value (p-value) that represents significant difference between the clusters if the p-value is less than 0.05 (95% confidence interval) [15]. Cox regression is then used to calculate the Cox Hazard ratio for each cluster, which represents the probability of death occurring given the cluster's survival till a certain point of time [14].

Kaplan-Meier curves are shown in Fig. 4B and 4D, in which statistical difference is shown in Fig. 4D with p-value = 0.02 (p-value < 0.05) highlighting that clusters 2 and 3 are significantly different. Furthermore, Fig. 4C highlights that cluster 2 and 3 have high and low Cox hazard ratio values respectively, showing further confirmation towards them being significantly different.

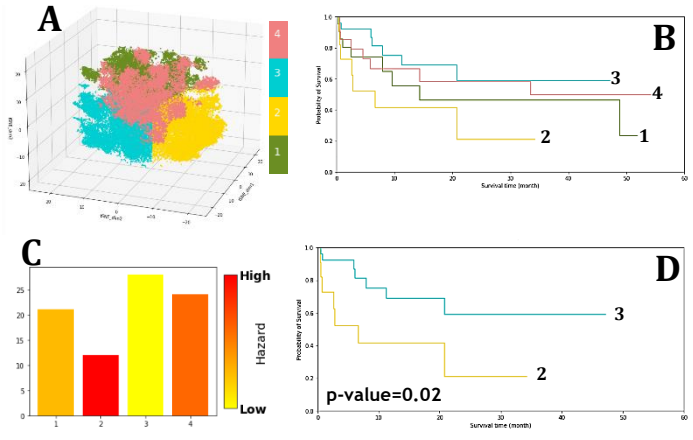

*Fig. 4. **Survival Analysis of Gastric Cancer**: (A) t-SNE clustered image using our optimum K-value = 4 of the MSI gastric cancer. (B) Kaplan-Meier curves showing the survival distribution for each cluster. (C) Bar plot colored using Cox hazard ratio showing the number of patients in each cluster. (D) Significant difference between clusters 2 and 3 is observed; they are represented as completely distinct regions in the t-SNE clustered image in (A).*

B. Metastasis Analysis of Breast Cancer

The breast cancer dataset consists of 32 patients; 11 of which are non-metastatic (pN = 0) and 21 that are metastatic (pN = 1). A fully metastatic cluster was observed in Fig. 5B when illustrating the distribution of metastatic and non-metastatic patients for each cluster using a histogram. This fully metastatic cluster (cluster 8) is highlighted in the t-SNE clustered image in Fig. 5C and can be taken as the significant cluster to compare with against all other clusters.

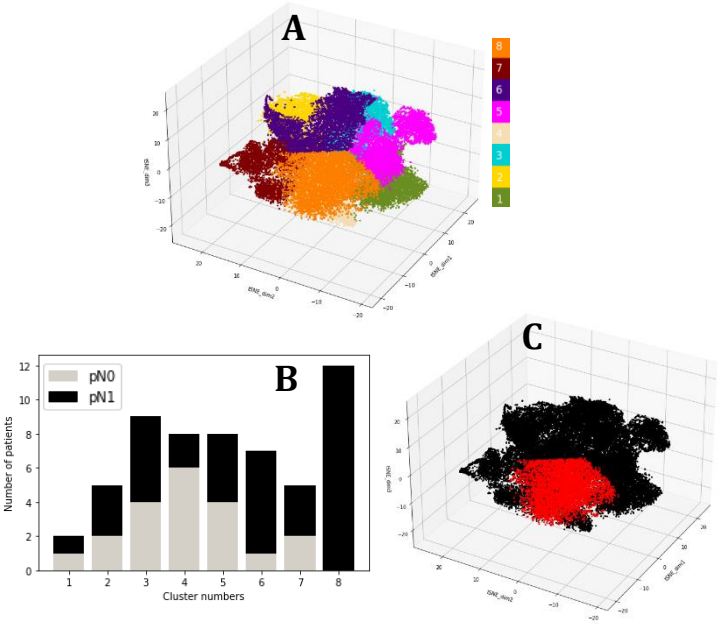

*Fig. 5. **Metastasis Analysis of Breast Cancer**: (A) t-SNE clustered image using the optimum K-value = 8 of the MSI breast cancer. (B) Distribution of the metastatic (black) and nonmetastatic (grey) patients represented in a histogram and the fully metastatic cluster 8 can be observed. (C) The fully metastatic cluster, highlighted in red, is shown in the t-SNE map.*

C. Significance Analysis of Microarrays & Machine Learning

We investigate the possibility of building a predictive pixel classifier able to predict the survival & metastasis status of newly introduced MSI gastric and breast cancer data respectively. To train the classifier with protein ions (features) able to differentiate between all clusters, Significance Analysis of Microarrays (SAM) is used. It applies a t-test at the individual gene or protein level to determine if the expression for that gene or protein is significant [16]. We used it to determine the significant protein ions, with False Discovery Rate (FDR) < 0.001, contributing to the hazardous cluster in gastric cancer and fully metastatic cluster in breast cancer.

The pixel classifier will then mainly use the Support Vector Machine (SVM) along with k-nearest-neighbor (KNN) models and then be trained on the patients' MSI data and cross-validated using the main reference's method called Leave One Patient Out (LOPO) [8]. The LOPO method dictates that we set one patient aside and run the entire processes mentioned in this paper (t-SNE, clustering, clinical linking, SAM, SVM/KNN classification) and then test the classifier on the left-out patient's MSI data. This would allow our classifier to avoid over-fitting and information leakage when testing the performance of the trained classifier.

LOPO method is then repeated for all the patients to produce unbiased results from the classifier. The absence of ground truth for patient samples to compare with our pixel classifier results led us to follow our main reference's method [8] of applying thresholds for detecting the good and poor subpopulations. This would let us classify each patient sample as good or poor outcome. In gastric cancer, the thresholds for the prediction of poor survival were set as *t1 = 10%* and *t2= 50%* [8]. If the pixels contributing to poor survival were less than *10%*, then the patient's survivability was high, and if it is higher than *50%*, then the patient's survivability was low. Otherwise, if it was in between *t1* and *t2* then it was considered moderate survivability. In breast cancer, the thresholds for the prediction of metastasis were set using Youden's index [8], such that *t1 = 2%* and *t2=100%*. If the pixels contributing to the patient's metastasis were less than *2%*, then the patient was classified as non-metastatic. Otherwise, the patient was classified as metastatic. Both gastric and breast cancer thresholds were taken from our main reference.

III. Discussion

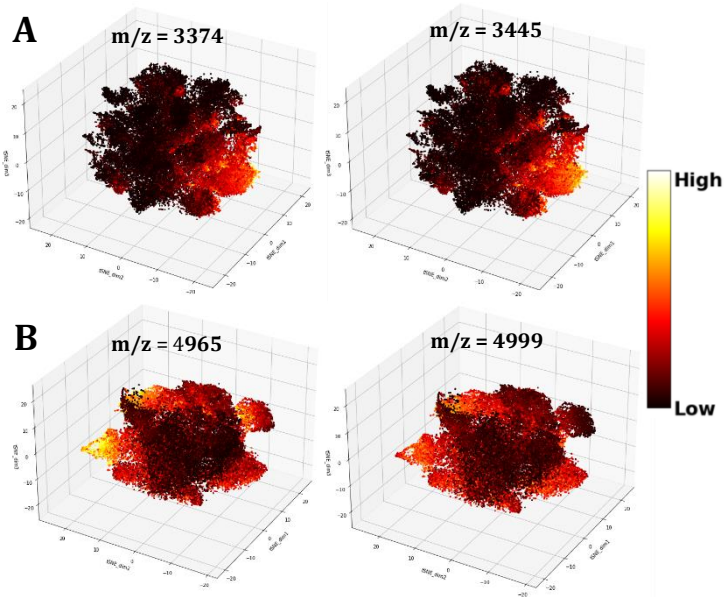

Fig. 6. ***Color-mapped specific protein ions in t-SNE space****: (A) Gastric cancer t-SNE map highlighting the most stable (100% detected in all LOPO runs) m/z protein ion intensities: 3374 and 3445. (B) Breast cancer t-SNE map highlighting the most stable m/z protein ion intensities: 4965, 4999. The color gradient shows the overexpression (high intensity) and underexpression (low intensity) of these m/z values.*

In each LOPO run, we were interested in identifying what the significant protein ions were for each patient sample and determining from them what the most recurring, present in *>80%* of samples, protein ions were. After completing the LOPO runs, the most stable ions in gastric cancer that contributed to the patient's survival were *m/z:* ***[3374, 3445, 3409, 3711, 3670, 3482, 3516, 13166]***. In breast cancer, the m/z values affecting metastasis status were *m/z:* ***[4999, 4965, 5067, 5171]***. Fig. 6 highlighted the most stable protein ions in gastric and breast cancer data; present in 100% of all LOPO samples. Fig. 6A showed that the overexpression of both gastric protein ions is present in the identified hazardous cluster (cluster 2) in Fig. 4A and Fig. 6B shows that the underexpression of both breast protein ions corresponds to the fully metastatic cluster highlighted in Fig. 5C. The resulted stable m/z values, present in >80% from LOPO runs, shows that we could minimize the selection of available protein ions that would be used for the classifier.

The classification results for the gastric cancer were categorized into 3 survival statuses: poor, medium, and high. Only *5* out of *23* dead patients were classified correctly with poor status, while *16* out of *40* unknown/alive patients were classified with high survival status. The remaining 42 patients were classified as moderate survivability. It could be concluded that the classifier wasn't able to accurately predict the poor patients due to the imbalanced ground truth in gastric cancer. In breast cancer, the overall classification accuracy was *78.125%* with *6* out of *11* correct non-metastatic predictions and *19* out of *21* correct metastatic predictions.

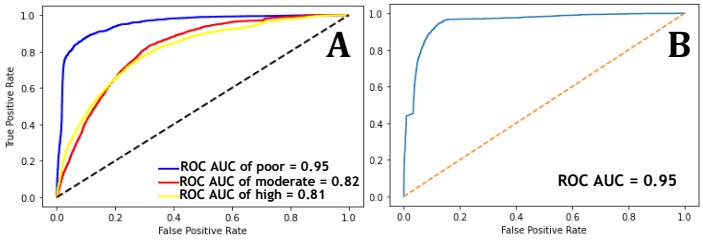

Fig. 7. ***Receiver Operating Characteristic Curves (ROC) for both datasets****: (A) ROC of gastric cancer patient no.23 showing ROC Area Under Curve (AUC) for each class with values 0.95, 0.82 and 0.81 for poor, moderate and high classes respectively (B) ROC of breast cancer patient no.21 showing ROC AUC for metastasis class of value 0.95.*

During each LOPO run, the classifier's performance was checked using ROC and measuring the classifier's Accuracy, Precision, F1-score, and Recall. The classifier performance is shown in Fig. 7 with one patient sample for each of gastric and cancer. These quality checks helped us determine if the classifier had no issues during its training before we test it on the left-out patient and recorded the final test result.

IV. Conclusion

The applied machine learning techniques in this paper helped in analyzing the intratumor heterogeneity, which is a crucial factor in identifying the tumor evolution and subsequently the patient status. Tumor subpopulations affect the cancer evolution, and we were interested in identifying unique tumor subpopulations that influence the patient's outcome. Statistical analysis of the data helped in determining the tumor subpopulations with the most detrimental impact on the patient's status. Using the identified significant subpopulations, we built a classifier capable of predicting the metastasis status in breast cancer patients and survivability of gastric cancer patients. This classifier proves that we could identify early onset of cancer biomarkers, unlike histological images, and allows focused treatment for cancerous patients.

Acknowledgement

We would like to express our appreciation to Rokaia for her contributions and efforts with us during the first term of our graduation year on the project. She was a valued team member and unfortunately had to leave us in the second term.

References

[1] L. A. McDonnell and R. M. A. Heeren, "Imaging mass spectrometry," *Mass Spectrometry Reviews*, vol. 26, no. 4, pp. 606–643, May 2007, doi: 10.1002/MAS.20124.

[2] K. Schwamborn and R. M. Caprioli, "Molecular imaging by mass spectrometry — looking beyond classical histology," *Nature Reviews Cancer 2010 10:9*, vol. 10, no. 9, pp. 639–646, Aug. 2010, doi: 10.1038/nrc2917.

[3] L.-H. Gam, "Breast cancer and protein biomarkers," *World Journal of Experimental Medicine*, vol. 2, no. 5, p. 86, 2012, doi: 10.5493/WJEM.V2.I5.86.

[4] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Comput Methods Programs Biomed*, vol. 107, no. 3, p. 538, May 2012, doi: 10.1016/J.CMPB.2011.12.007.

[5] R. M. Caprioli, T. B. Farmer, and J. Gile, "Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS," *Anal Chem*, vol. 69, no. 23, pp. 4751–4760, May 1997, doi: 10.1021/AC970888I.

[6] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?," *Nature Reviews Cancer 2012 12:5*, vol. 12, no. 5, pp. 323–334, Apr. 2012, doi: 10.1038/nrc3261.

[7] B. Balluff *et al.*, "De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry," *Journal of Pathology*, vol. 235, no. 1, pp. 3–13, May 2015, doi: 10.1002/PATH.4436.

[8] W. M. Abdelmoula *et al.*, "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of Mass spectrometry imaging data," *Proc Natl Acad Sci U S A*, vol. 113, no. 43, pp. 12244–12249, May 2016, doi: 10.1073/PNAS.1510227113.

[9] M. Ringnér, "What is principal component analysis?," *Nature Biotechnology 2008 26:3*, vol. 26, no. 3, pp. 303–304, Mar. 2008, doi: 10.1038/nbt0308-303.

[10] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification," *Journal of Computer Science*, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/JCSSP.2018.1521.1530.

[11] L. van der Maaten and H. Geoffrey E, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[12] B. Christiansen, "Ensemble Averaging and the Curse of Dimensionality," *Journal of Climate*, vol. 31, no. 4, pp. 1587–1596, May 2018, doi: 10.1175/JCLI-D-17-0197.1.

[13] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *The journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[14] V. Bewick, L. Cheek, and J. Ball, "Statistics review 12: Survival analysis," *Critical Care*, vol. 8, no. 5, p. 389, Oct. 2004, doi: 10.1186/CC2955.

[15] J. M. Bland and D. G. Altman, "Statistics Notes: The logrank test," *BMJ : British Medical Journal*, vol. 328, no. 7447, p. 1073, May 2004, doi: 10.1136/BMJ.328.7447.1073.

[16] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001, doi: 10.1073/PNAS.091062498/ASSET/9AF693F7-706B-4AAC-A15C-23FB385DCD77/ASSETS/GRAPHIC/10624T2.JPEG.