Determining Cancerous Regions from Spatial Transcriptomics Data of Breast Cancer Tissue

Max Land
Princeton University
mland@princeton.edu

Abstract

Previous methods of determining normal cells vs cancer cells based on gene expression data have relied on analyzing a subset of tissue specific genes (i.e. BRCA2 mutation in breast tissue). However, the diversity of gene expression in cancer cells presents difficulties for this approach. In this paper, we present two novel classification methods are not dependent on the presence of specific genes. Instead, these methods consider the entire transcriptome as a whole and use PCA and percent dropout to reduce our data to one dimensional feature vectors. Next, we apply a simple thresholding classification method to these feature vectors to distinguish normal cells from healthy cells. We find that this novel clustering approach according to the first principal component of PCA and percent dropout have high performance and present a promising new approach in identifying tumor regions in cancer tissue.

1 Introduction

Due to the development of single-cell sequencing technologies [10], it is well known that multi-cellular organisms can generate and maintain homogeneous populations of healthy cells in individual tissues [4]. Cancer, however, can disrupt homogeneity and cause large changes to gene expression of infected cells. This is because cancer often involves mutations in proto-oncogenes, tumor suppressor genes, and other regulatory regions of the genome. These mutations alter the transcription rate and genetic diversity of cancer cells. This is reflected in single-cell sequencing studies of breast [7], kidney [12], bladder[5], and colon tumors [3]. Thus, given the changes in gene expression caused by cancer, an apparent question arises: Can we distinguish healthy cells vs cancer cells based on gene expression data?

The answer, as it turns out, is not so simple. Current methods involve constructing gene expression profiles of cancer cells to identify specific genes prevalent in cancer cells [2]. Researchers then look for the expression of these genes in cells to identify which cells are cancerous. However, the problem with this approach is that cancer is actually known for its heterogeneity at both the inter- and intra-tumor levels [1, 2]. Within a tumor, different spatial regions have been found to contain different cancer cell clones [1]. In addition, tumor cells are also infiltrated with stromal, immune and other cell types. These cells form the basis of tumor heterogeneity [11, 6]. As a result, characterization of tumor cells from tissue-wide gene expression data is an inherently difficult challenge. For example, CIBERSORT, a method for characterizing cell composition of complex tissues from their gene expression profiles, performs considerably worse on samples containing solid tumors [8]. Additionally, to analyze a different types of cancer tissue, researchers would need to construct separate gene expression profiles.

In this paper, we propose two novel methods to distinguish between normal cells and cancer cells without requiring prior characterization and establishment of gene expression profiles of various types of cancer cells. Instead, we focus on the transcriptome as a whole. Intuitively, we know there is a difference in gene expression data between normal and cancer cells. Our two models use PCA and percent dropout respectively to capture this difference in a one-dimensional feature

vector. Next, we apply a simple thresholding classification to classify our samples. To analyze the performance of our models, we utilize spatial transcriptomics data. Spatial transcriptomics (ST) is a new technology that allows researchers to measure gene activity in a tissue sample and map the location of the activity [9]. The spatial pattern of gene expression within a tissue is obtained by sequencing mRNA from a grid a spots, which refer to as ST spots, each containing a small number of cells.

In the case of cancer tissue, the nature of spatial transcriptomics allow us to view the location of our ST spots in the original tissue sample. We can visually see which ST spots come from cancerous vs non-cancerous regions of the tissue by plotting the spots on a cancer visualization image of the tissue. This creates a unique gene expression data set of cancer tissue ST spots with the true label known for each ST spot: cancerous or non-cancerous. This, in turn, allows us to perform exploratory cluster analysis to investigate novel methods to determine cancerous from non-cancerous tissue regions using the spatial transcriptomics data. In particular, we explored models based on PCA and percent dropout as mentioned previously in this paper.

2 Methods

2.1 Data/Data Pre-processing

We worked with publicly available breast cancer data set published by the Spatial Research Lab [9]. The data set consists of four layers of breast cancer tissue sliced from the same sample. For each layer, we have:

1. Spatial Transcriptomics Data: Gene expression matrix, where gene are columns and ST spots (identified by coordinates) are rows. For all datasets, spots outside the tissue were removed.
2. Visualization of Cancerous Regions: Hematoxylin & eosin stained brightfield image of the tissue layer. In these images, the dark purple tissue corresponds to the cancerous regions and light purple tissue corresponds to the healthy, non-cancerous regions.

Next, we generated our set of true labels l_{true} for each layer. To do so, we plotted each ST spot on our stained brightfield image. Then, we manually labeled the ST spot as cancerous or non-cancerous depending on the location. We report a visualization of l_{true} below:

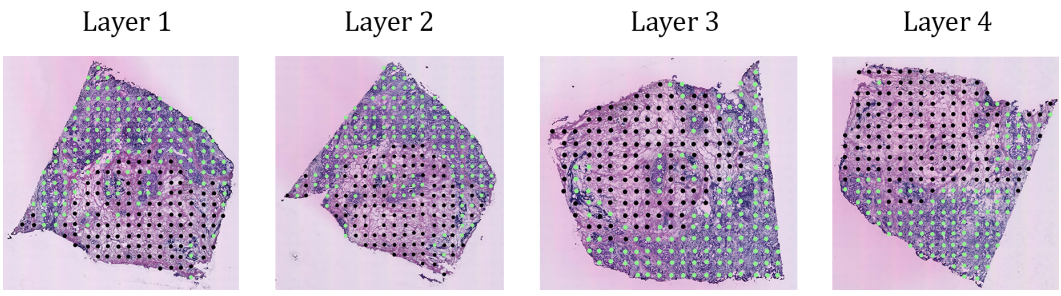

Figure 1: Visualization of our manually clustered ST spots.

Note: This visual classification of tumor regions of the cell is reflective of the practice used by real-world oncologists. The light green represents a cancerous ST spot, and black represents a non-cancerous ST spot.

2.2 Method 1: PCA Classifier

We investigated the ability of the first principal component of PCA to differentiate between cancerous and non-cancerous regions of the tissue. To do so, we first preprocessed the spatial transcriptomics data using the following steps:

1. First, we filtered for genes with at least 15 nonzero entries. In others, genes with non-zero counts in fewer that 15 ST spots were filtered out.
2. The data is then library-size normalized, where counts for each gene in a spot are divided by the sum of counts for the ST spot. This normalization removes variance due to differences in size or number of cells in each ST spot.
3. Next, the data is log normalized with pseudocount 1 ($\log(X+1)$).
4. Finally, we Z-normalized each ST spots (by row).

Next, we applied PCA to generate our first principal component vector. For a cancer tissue layer with n ST spots sequenced, we get first principal component vector:

$$pc = (pc_1, \ldots, pc_n)$$

Then, we performed a thresholding classification to cluster the spots into cancerous vs non-cancerous. We wanted to get a vector of predictive labels for each spot:

$$l_{\text{pred}} = (l_{\text{pred}_1}, \ldots, l_{\text{pred}_n})$$

To do so, we clustered each spots' respective principal component vector pc against a threshold value t. For some t and $i = 1, \ldots, n$,

$$l_{\text{pred}} = \begin{cases} 1 & \text{if } pc_i \leq t \\ 0 & \text{if } pc_i > t \end{cases}$$

Note: Preliminary analysis showed that tumor cells tend to have lower pc values. So, we assigned our prediction labels l_{pred} to be consistent with l_{true} which assigned cancerous ST spots to 1 and non-cancerous ST spots to 0.

In summary, given a threshold t, we can generate a set of predictive labels l_{pred} based on PCA.

We will refer to the function as:

$$\text{pca\_classifier}(pc, t) = l_{\text{pred}}$$

2.3 Method 2: Percent Dropout Classifier

We investigated ability of using percent dropout to cluster cancerous vs non-cancerous spots. From the spatial transcriptomics data, we calculated the percent of genes with zero counts for each ST spot. For n ST spots, this resulted in a n-length percent dropout vector:

$$pd = (pd_1, \ldots, pd_n)$$

From here, we repeated the classification steps from our PCA Classifier. We performed a thresholding classification to cluster the spots into cancerous vs non-cancerous. The goal was to get a vector of predictive labels for each spot:

$$l_{\text{pred}} = (l_{\text{pred}_1}, \ldots, l_{\text{pred}_n})$$

To do so, we clustered each spots' respective percent dropout component against a threshold value t. For some t and $i = 1, \ldots, n$,

$$l_{\text{pred}_i} = \begin{cases} 1 & \text{if } pd_i \leq t \\ 0 & \text{if } pd_i > t \end{cases}$$

Thus, given a threshold t, we can generate a set of predictive labels l_{pred} based on percent dropout.

We will refer to the function as:

$$\text{pd\_classifier}(pd, t) = l_{\text{pred}}$$

2.4 Classifier Evaluation

We tested the effectiveness of our PCA classifier and percent dropout classifier on the breast cancer data set. We used two main metrics to evaluate performance:

1. Adjusted Rand Index (ARI): The Rand Index (RI) computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The raw RI score is then "adjusted for chance" by:

$$\text{ARI} = \frac{\text{RI - Expected\_RI}}{\text{max(RI) - Expected\_RI}}$$

   The ARI score is calculated using scikit-learn's: adjusted_rand_score().

2. Precision-Recall Curve: First, we define precision and recall:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \quad \text{and} \quad \text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

   where TP = # of True Positives, FP = # of False Positives, and FN = # of False Negatives.

   Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

2.5 Choosing a threshold

For each breast cancer tissue layer, we wanted to calculate ARI score and Precision-Recall score for each classifier. However, each classifier is function of a threshold t. Thus the question becomes, how do we pick a threshold t?

To address this, we calculated ARI scores and Precision-Recall scores for various thresholds t. To choose t, we took 200 different values of t evenly spaced between $[pc_{\text{min}}, pc_{\text{max}}]$ or $[pd_{\text{min}}, pd_{\text{max}}]$ (depending on the classifier) to generate 200 different predictions l_{pred}.

To clarify:

- Given principal component vector pc,

$$pc_{\text{max}} = \max\{pc_1, \ldots, pc_n\} \quad \text{and} \quad pc_{\text{min}} = \min\{pc_1, \ldots, pc_n\}$$

- Given percent dropout vector pd

$$pd_{\text{max}} = \max\{pd_1, \ldots, pd_n\} \quad \text{and} \quad pd_{\text{min}} = \min\{pd_1, \ldots, pd_n\}$$

Here is the algorithm in pseudo-code for PCA:

Algorithm : Calculate ARI score and Precision-Recall for 200 different threshold values

```
pc ← get_Principal_Component(SpatialTranscriptomicsData)
procedure EVALUATE_CLASSIFIER(pc, l_true)
    t ← linspace(min(pc), max(pc), 200)
    for i = 1 : 200 do
        l_pred ← pca_Classifier(pc, t_i)
        ari ← calculate_ARI(l_true, l_pred)
        precision ← calculate_Precision(l_true, l_pred)
        recall ← calculate_Recall(l_true, l_pred)
    end for
    return ari, precision, recall
end procedure
```

The algorithm for Percent Dropout follows the same design.

3 Results

3.1 Results of ARI Score: PCA Classifier

The publicly available breast cancer data set consists of four layers. For each layer, we tested the performance of our PCA classifier by calculating the ARI score as a function of a threshold t. We report the results below:

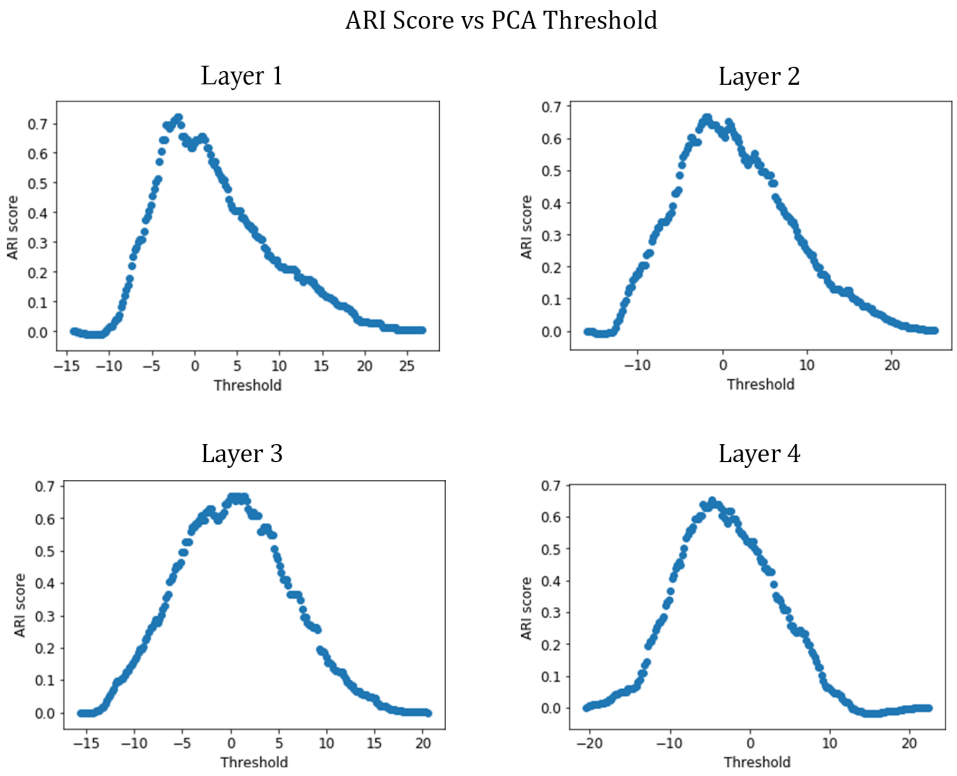

Figure 2: Performance Evaluation of the PCA Classifier: ARI Score

Next, we report the maximum ARI score and the threshold t associated with that score. Since we are classifying by PCA, we also report the proportion of total variance explained by the first principal component:

	Layer 1	**Layer 2**	**Layer 3**	**Layer 4**
Max ARI	0.722	0.666	0.668	0.653
Optimal t	-2.060	-1.932	1.478	-4.739
Prop of Total Variance Explained	0.012	0.015	0.013	0.015

As seen in Figure 2, the ARI score graphed as a function of t is a peaked graph. For optimal t, we see that the maximum ARI score ranges from $0.65-0.72$ across all four layers, despite our first principal components explaining only $1-2\%$ of the total variance of our data. This shows that clustering the ST spots according to the first principal component is an effective method to distinguish between cancerous and non-cancerous.

Next, we created a visualization of the best prediction by the PCA Classifier. For the optimal t that resulted in the maximum ARI score for each layer, we plotted the ST spots on the tissue image color-coded by its predicted label given by the PCA Classifier. As a reminder, light green corresponds to a cancerous ST spot, and black corresponds to a non-cancerous ST spot. We report the results below:

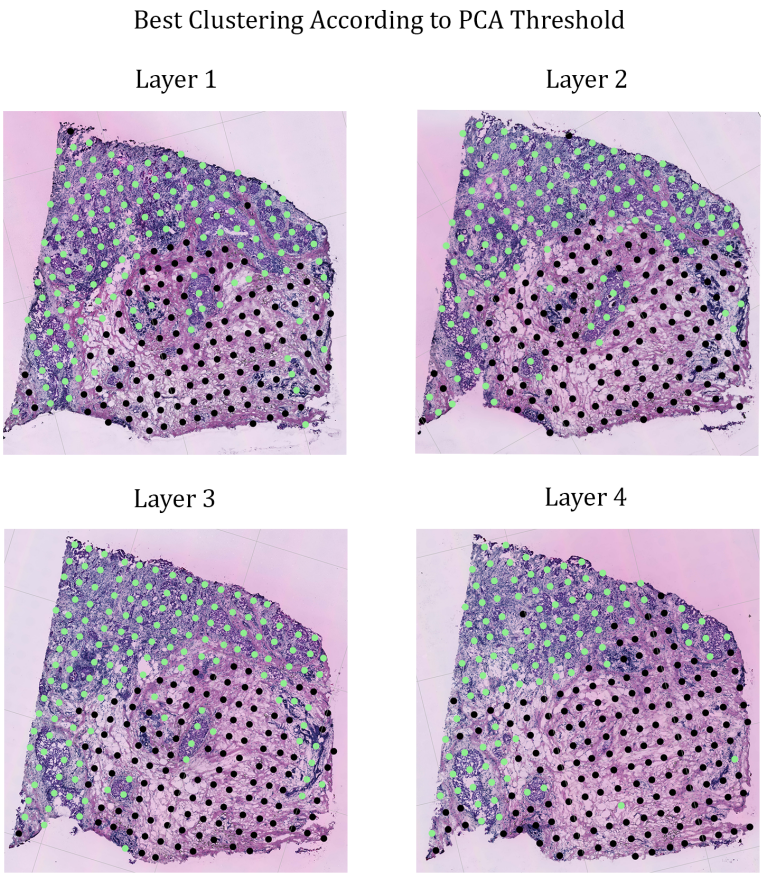

Figure 3: Each layer reflects the clustering corresponding to the highest ARI achieved in Figure 2.

The visualization of the predictions in Figure 3 show that the PCA classifier is effective at classifying the ”obvious” ST spots– those that are clearly in a cancerous region or non-cancerous region. Next, comparing the visualization of our PCA classifier (Figure 3) to the true labels (Figure 1), we see the misclassifications mainly occur at the ”no man’s land” where the cancerous and non-cancerous regions meet.

3.2 Results of ARI Score: Percent Dropout Classifier

Next, we repeated the same analysis for our Percent Dropout Classifier. For each layer, we tested the performance of our Percent Dropout classifier by calculating the ARI score as a function of a threshold t. We report the results below:

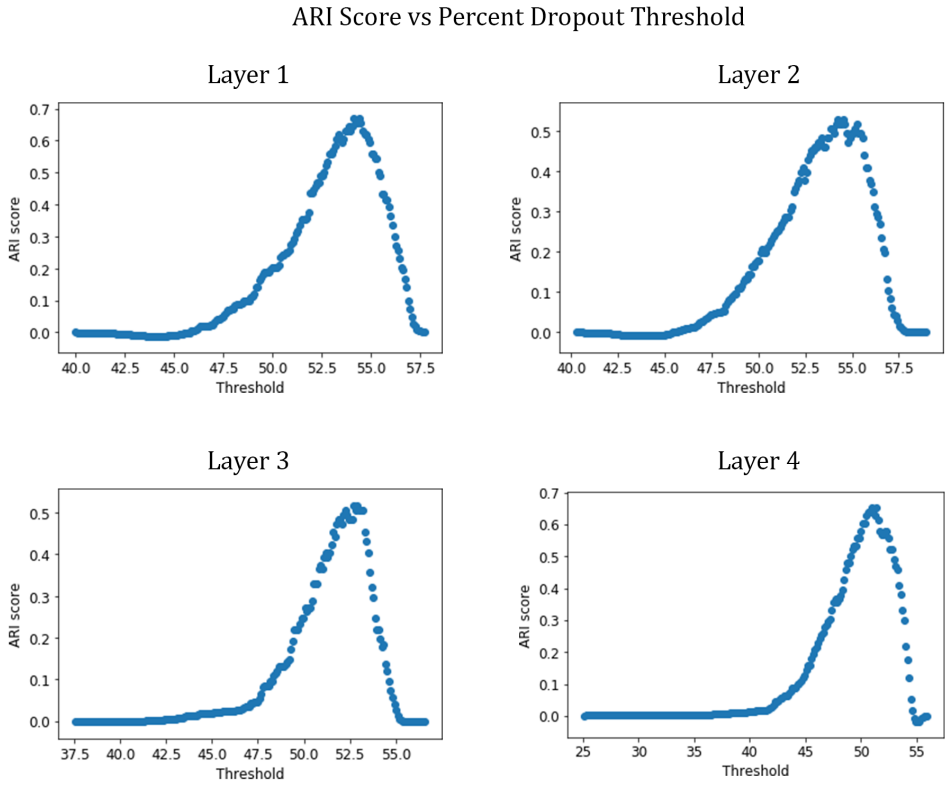

Figure 4: Performance Evaluation of the Percent Dropout Classifier: ARI Score

Next, we report the maximum ARI score and the threshold t associated with that score:

	Layer 1	**Layer 2**	**Layer 3**	**Layer 4**
Max ARI	0.669	0.530	0.516	0.653
Optimal t	54.128	54.240	52.916	51.410

For the Percent Dropout Classifier, we see that the ARI sore graphed as a function of t is also a peaked graph in Figure 4. Compared to the PCA Classifier, the maximum ARI score is slightly lower, but still ranges from $0.51 - 0.67$. This result suggest that the PCA classifier is more effective than the Percent Dropout classifier.

Next, we created a visualization of the best prediction by the Percent Dropout Classifier. For the optimal t that resulted in the maximum ARI score for each layer, we plotted the ST spots on the tissue image color-coded by its predicted label given by the Percent Dropout Classifier. As a reminder, light green corresponds to a cancerous ST spot, and black corresponds to a non-cancerous ST spot. We report the results below:

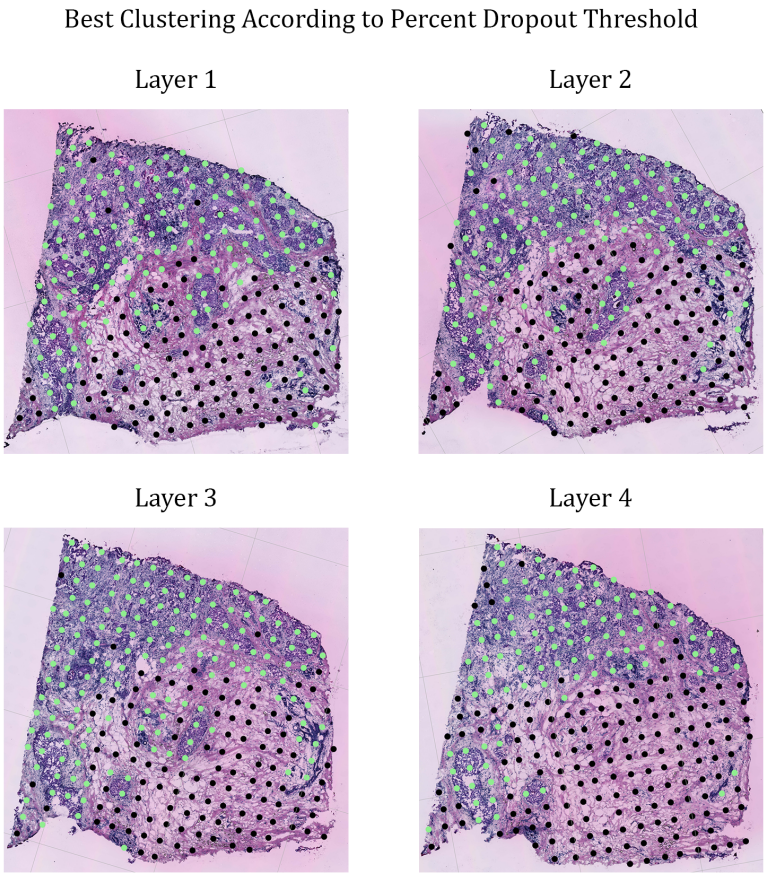

Figure 5: Each layer reflects the clustering corresponding to the highest ARI achieved in Figure 4.

Looking at the visualization of predictions in Figure 5, we see the predictions are slightly worse. For the most part, the Percent Dropout Classifier does classify the majority of ST spots correctly– the large regions of cancerous and non-cancerous are correct. However, compared to the PCA classifier in Figure 3, there are obvious misclassifications in the large cancerous regions. Again, this substantiates the claim from previous result that the PCA Classifier has better performance than the Percent Dropout Classifier.

3.3 Precision-Recall Curves

Another way to assess our classifiers is through precision-recall curves. As seen in the algorithm in section 2.5, for each threshold t, we calculate a precision and recall score. Because we calculate 200 values of t for each layer, this results in 200 data points of the form: $(\text{recall}, \text{precision})$

We then used those points to create a Precision-Recall curve for each layer, and we report the results below:

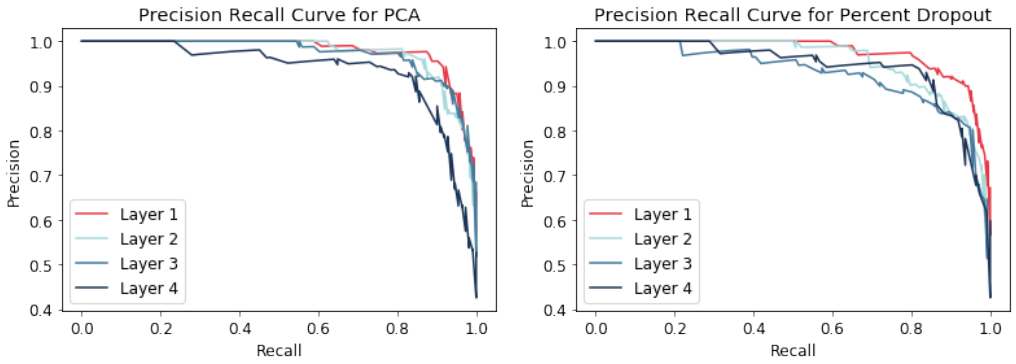

Figure 6: Precision-Recall Curves for PCA

Overall, the precision-recall curves for both the PCA Classifier and the Percent Dropout Classifier are curved appropriately. This indicates good performance.

Furthermore, we see that for for both classifiers, Layer 1 has the best performance. This is consistent with our ARI results seen in Figure 2 and 4. As a matter of fact, the order of performance of the four layers for each classifier reflects the order reported by our ARI results.

Next, comparing the two precision-recall curves against each other, we see that the precision-recall curve for PCA is curved more towards $(1, 1)$. This substantiates our analysis that the PCA Classifier has better performance than the Percent Dropout Classifier.

3.4 Distribution of Components

Both the PCA and Percent Dropout classifiers return a prediction of cancerous or non-cancerous for each ST spot dependent on a threshold t. For optimal t, we see that the our PCA and Percent Classifier perform well. However, in real applications, we don't know the true labels, thus we cannot experimentally determine the optimal threshold. Thus, this begs the following question: Is there a way to select a good threshold t?

To gain some insight into choosing a threshold t, we plot the distributions of our PCA components. Specifically, we split the PCA components into their underlying cancerous and non-cancerous distributions based on the true labels. We also plot the experimentally calculated optimal threshold t for reference. As a reminder:

	Layer 1	**Layer 2**	**Layer 3**	**Layer 4**
Optimal t	-2.060	-1.932	1.478	-4.739

Next, for each layer, we investigate the distribution of $(pc_1, \ldots, pc_n)$ for our principal component vector pc. We report our distribution plots:

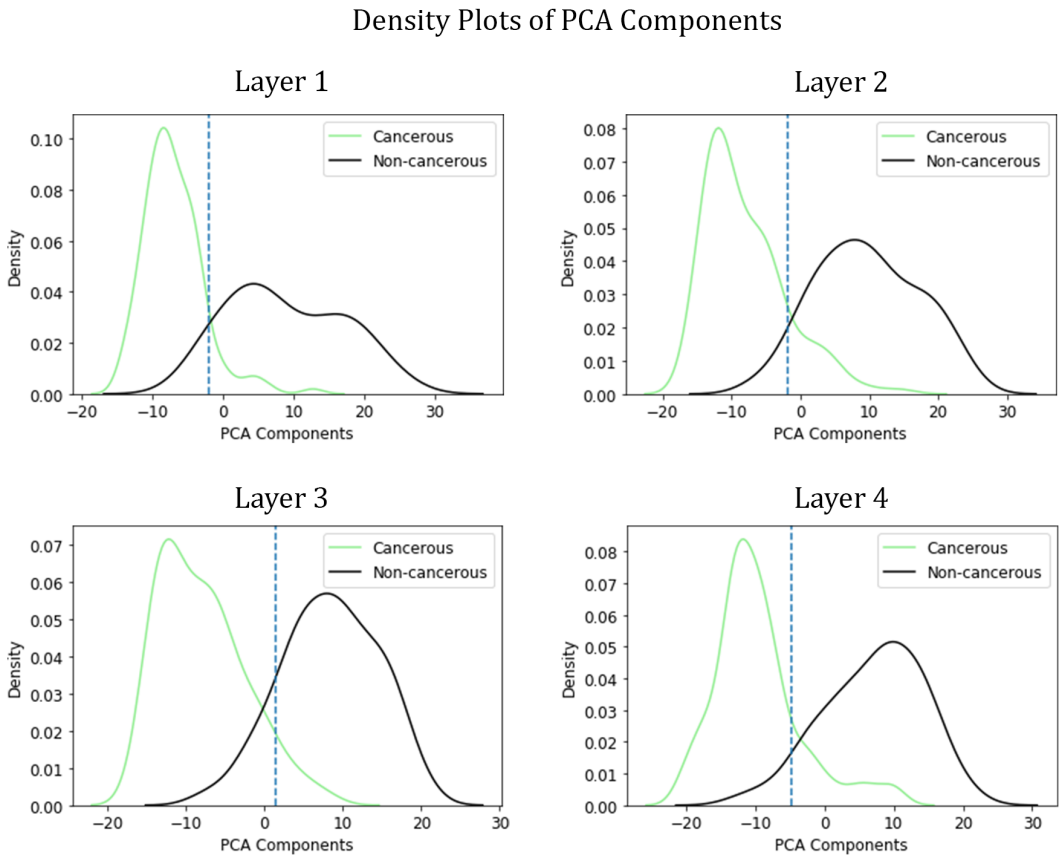

Figure 7: Distribution of PCA components for cancerous and non-cancerous ST spots.

From Figure 7, we immediately notice that the cancerous and non-cancerous ST spots are drawn from two separate underlying distributions. This further substantiates that the classifying by first principal component can reasonably distinguish between cancerous and non-cancerous ST spots. In light of this, it appears picking the optimal t becomes as simple as picking the overlap between the two tails of the respective distributions. This is confirmed by our experimental results– the optimal ARI occurs at the overlap of these two distributions, as indicated by the dotted blue line.

Again, however, it is important to recognize that it is not this straightforward in practice. In real applications, we do not know the true set of cancerous and non-cancerous spots. Thus, we cannot split the pc values into the two underlying distributions. Instead, we will only have the combined overall distribution. Thus, it might make more sense to view our problem in a more general sense:

> "Given that we know our vector of values pc are drawn from two underlying distributions, can we accurately classify each individual data points into the respective distribution it is drawn from?"

Also, it important to keep in mind that the two distributions will most likely overlap as seen in Figure 7. Thus, our current method of simply classifying by an optimal t will inherently result in misclassifications (both false positives and false negatives). However, there are ways to improve prediction by taking advantage of the spatial location of our ST spots. Our results show that cancerous and non-cancerous cells are drawn from distinct distributions. The overlap, therefore, most likely represent ST spots that are in the process of becoming cancerous. In Figure 3, we see that most of the misclassifications are ST spots that lie in between cancerous and non-cancerous regions. Intuitively, the ST spots that lie closer to definitive cancerous ST spots are more likely to be cancerous and

vice versa. So, for ST spots that lie in that overlap region, future models can incorporate the spatial information to resolve misclassifications.

Next, we plot the distribution of our Percent Dropout components. Specifically, we split the Percent Dropout components into their underlying cancerous and non-cancerous distributions based on the true labels. We also plot the experimentally calculated optimal threshold t for reference. As a reminder:

	Layer 1	**Layer 2**	**Layer 3**	**Layer 4**
Optimal t	54.128	54.240	52.916	51.410

Next, for each layer, we investigate the distribution of $(pd_1, \ldots, pd_n)$ for our percent dropout component vector pd. We report our distribution plots:

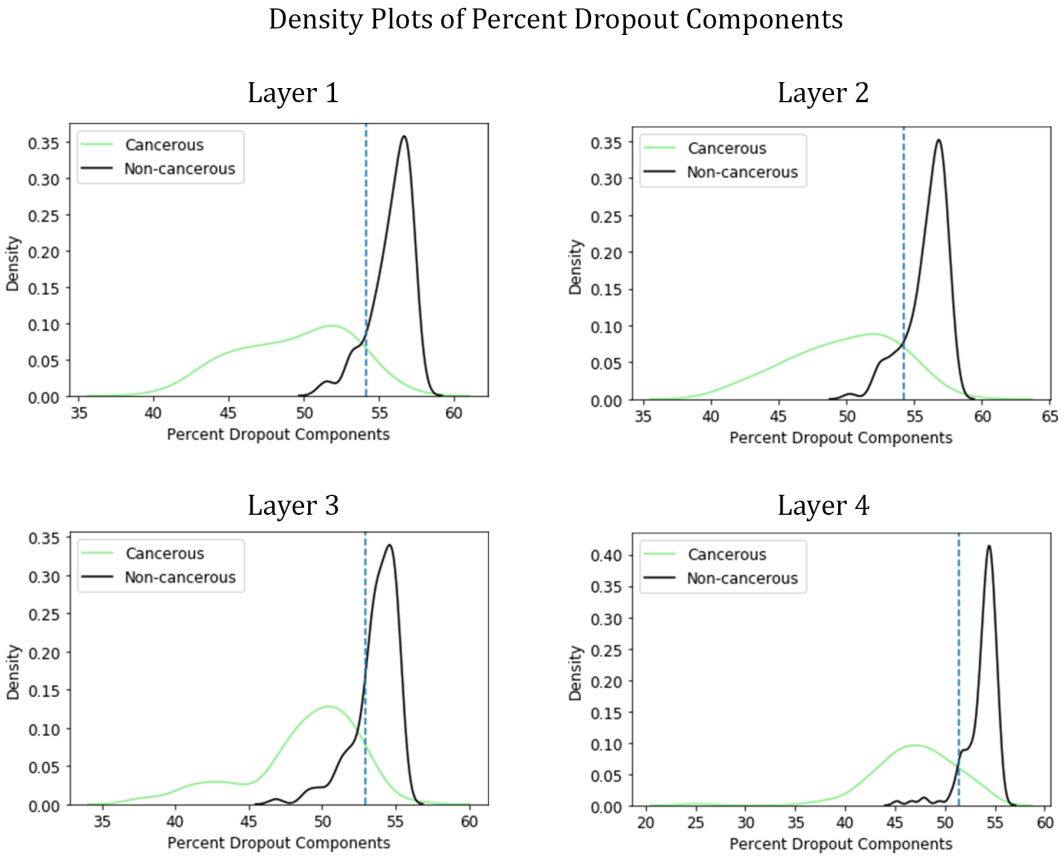

Figure 8: Distribution of Percent Dropout components for cancerous and non-cancerous ST spots.

Similarly to the PCA density plots, we see that cancerous and non-cancerous cells form two different underlying distributions in Figure 8. Most notably, the non-cancerous distribution is extremely peaked compared to the cancerous distribution.

The results for percent dropout are also more interpretable– cancerous ST spots tend to have a lower percent dropout. Percent dropout measures the percent of genes with zero expression. Thus, cancerous ST spots have a higher number of unique genes expressed. This makes sense, as cancerous cells arise from additional mutations in proto-oncogenes, tumor suppressor genes, and other regulatory regions of the genome. This biological understanding also explains the difference in kurtosis between the two distributions. Normal breast cancer cells drawn from the same tissue sample have homogeneous gene expressions [4]. This is reflected in the sharply peaked non-cancerous distribution. On the other hand, cancerous mutations are a result of random mutations. Different ST spots

are also at different stages of cancer (some spots have been infected longer than others). Thus, there is a higher chance for variance in the diversity of gene expression in cancerous ST spots. This is reflected in the distribution being less peaked and more distributed across percent dropout values. Furthermore, previous research has shown cancer has heterogeneity within tumors [1].

4 Conclusion

In this paper, we proposed two novel methods to determine cancerous regions from spatial transcriptomics data of cancer tissue. Specifically, the PCA Classifier clusters based on the first principal component after applying PCA and the Percent Dropout Classifier clusters based on the proportion of zero expression genes. We demonstrated that both classifiers show promising results with good performance on breast cancer data. Out of the two, the PCA Classifier performs slightly better. Further analysis revealed that both classifiers effectively reduce the transcriptome to one dimensional feature vectors, and that the distribution of these feature vectors are not random, but rather a combination two underlying distributions (cancerous vs non-cancerous). This allows a thresholding classification technique to be effective at clustering accurately.

The results of this paper are an initial exploration of analyzing spatial transcriptomics data of cancer tissue. There several directions for future work. The first is incorporating spatial information. In general, tumors arise and proliferate rapidly– cancerous ST spots will often be found in large connected regions. In the visualizations of our classifier results (Figures 3, 5), we see there are obvious misclassficiations (i.e. a supposedly non-cancer ST spot surrounded by cancerous ST spots). The spatial information can be used ensure our classification results are more in line with our fundamental understanding of cancer. We can also stack the layers together to create a 3-D spatial network. Another important extension is to evaluate our classifiers on other types of spatial transcriptomics data. For example, there is available transcriptomics data on prostate cancer tissue [2]. Furthermore, it would be interesting to see what our PCA classifier would detect if there were no tumor regions to detect.

5 Code Availability

The methods in this paper are implemented in Python and are available at: `https://github.com/mrland99/spatial-research`

References

[1] Noemi Andor, Trevor Graham, Marnix Jansen, Li Xia, C. Aktipis, Claudia Petritsch, Hanlee Ji, and Carlo Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22, 11 2015.

[2] Emelie Berglund, Jonas Maaskola, Niklas Schultz, Stefanie Friedrich, Maja Marklund, Joseph Bergenstråhle, Firas Tarish, Anna Tanoglidi, Sanja Vickovic, Ludvig Larsson, Fredrik Salmén, Christoph Ogris, Karolina Wallenborg, Jens Lagergren, Patrik Ståhl, Erik Sonnhammer, Thomas Helleday, and Joakim Lundeberg. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications*, 9, 12 2018.

[3] Ellen Heitzer, Martina Auer, Christin Gasch, Martin Pichler, Peter Ulz, Eva Hoffmann, Sigurd Lax, Julie Waldispuehl-Geigl, Oliver Mauermann, Carolin Lackner, Gerald Höfler, Florian Eisner, Heinz Sill, Hellmut Samonigg, Klaus Pantel, Sabine Riethdorf, Thomas Bauernhofer, Jochen Geigl, and Michael Speicher. Complex tumor genomes inferred from single circulating tumor cells by array-cgh and next-generation sequencing. *Cancer research*, 73, 03 2013.

[4] Hai Le, Monika Looney, Benjamin Strauss, Michael Bloodgood, and Antony Jose. Tissue homogeneity requires inhibition of unequal gene silencing during development. *The Journal of Cell Biology*, 214:jcb.201601050, 07 2016.

[5] Yingrui Li, Xun Xu, Luting Song, Yong Hou, Zesong Li, Shirley Tsang, Fuqiang Li, Kate Im, Kui Wu, Hanjie Wu, Xiaofei Ye, Guibo Li, Linlin Wang, Bob Zhang, Jie Liang, Wei Xie, Renhua Wu, Hui Jiang, Xiao Liu, and Jun Wang. Single-cell sequencing analysis characterizes

common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience*, 1:12, 08 2012.

[6] Carlo Maley, C. Aktipis, Trevor Graham, Andrea Sottoriva, Amy Boddy, Michalina Janiszewska, Ariosto Silva, Marco Gerlinger, Yinyin Yuan, Kenneth Pienta, Karen Anderson, Robert Gatenby, Charles Swanton, David Posada, Chung-I Wu, Joshua Schiffman, E Hwang, Kornelia Polyak, Alexander Anderson, and Darryl Shibata. Classifying the evolutionary and ecological features of neoplasms. *Nature reviews. Cancer*, 17, 09 2017.

[7] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alexander Krasnitz, W. Mccombie, James Hicks, and Michael Wigler. Tumor evolution inferred by single cell sequencing. *Nature*, 472:90–4, 03 2011.

[8] Aaron Newman, Chih Liu, Michael Green, Andrew Gentles, Weiguo Feng, Yue Xu, Chuong Hoang, Maximilian Diehn, and Ash Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12, 03 2015.

[9] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[10] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian Tuch, Asim Siddiqui, Kaiqin Lao, and M Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6:377–82, 05 2009.

[11] Kenneth Valkenburg, Amber Groot, and Kenneth Pienta. Targeting the tumour stroma to improve cancer therapy. *Nature Reviews Clinical Oncology*, 15, 04 2018.

[12] Xun Xu, Yong Hou, Xuyang Yin, Bao Li, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, Weiming He, Li Zeng, Manjie Xing, Renhua Wu, Hui Jiang, Xiao Liu, Dandan Cao, Guangwu Guo, Xueda Hu, and Jun Wang. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148:886–95, 03 2012.