High Accuracy Cancer Prediction Using Machine Learning Inspired Mass Spectrum Analysis

Peter Liu
6/21/2018

Abstract

Early detection and treatment of cancer is essential to increase survival rate and life quality of patients with cancer. However, cancer was mostly discovered at late stage, because people tend not to do the cancer check up regularly due to the high expense. Alternatively, mass spectroscopy offers an easy and affordable approach to analyze people's routine check up samples in short time, and my interest is to apply machine learning (ML) algorithm to analyze mass spectra to diagnose cancer. Herein, ML models (e.g., Random Forest, SVM, KNN and ensemble model) were applied for prediction of ovarian cancer and prostate cancer, with high accuracy ranging from 93% to 100%. One of the fingerprint chemicals determining ovarian cancer was also successfully identified from over 9000 chemical candidates, which is confirmed by literature report. This work not only offers early cancer detection toolbox, but also is a powerful tool for professionals to discover new fingerprint molecules for cancer research and drug development, which will greatly improve R&D efficiency and save cost.

Problem statement

Cancer prediction is one of the hottest topics in healthcare industry. Both healthcare institutes and patients would be benefited from early cancer determination. Early detection and treatment of cancer is essential to increase survival rate and life quality of patients with cancer. However, cancer was mostly discovered at late stage, because people tend not to do the cancer check up regularly due to the high expense. In fact, **cancer is nearly always diagnosed by an expert who has looked at cell or tissue samples under a microscope**,[1] which is not only time consuming but also cannot avoid human errors. Moreover, both patients and doctors will not suspect cancer until symptoms appeared, *e.g.*,

vomiting, dazzy, which sometimes only showed up on late stage of cancer. It is thus preferable to predict cancer using regular health check up samples (e.g., blood samples), to ‘redflag’ cancer, even before the symptoms appeared.

One way is to use high-sensitive analytical instruments to analyze routine check up samples and predict cancer. Mass spectroscopy offers an affordable and fast solution to collect chemical information from saliva or other excretes, which has been widely applied in pharmaceutical companies for drug screening and test. In general, mass spectroscopy differentiates chemicals by their weight or mass, and it is of high sensitivity even with low concentration of chemicals.[2] Moreover, the mass spectroscopy analysis typically uses tiny amount of samples (milligrams), takes seconds to minutes to finish and could be easily coupled with robotic sample preparation techniques, which is an ideal approach for high throughput chemical screening and testing.

My work herein will rely on deep analysis of mass spectra collected from healthy people (control group) and people with cancer(cancer group) to see if I can find the difference in mass spectra from these two groups and thus predict cancer. In addition, I would like to predict fingerprint chemicals that can determine cancer. Again, this will greatly inspire the pharmaceutical research and development to reveal the cancer formation mechanisms, and develop more efficient drugs to prevent or cure cancer.

One of the simplest methods to compare two spectra is to directly compare them side by side. Typical mass spectra is shown as below. Apparently, side by side comparison is not a good way in this case, and it is difficult to find patterns (which chemicals are determinant for cancer diagnosis), even for experienced professionals.

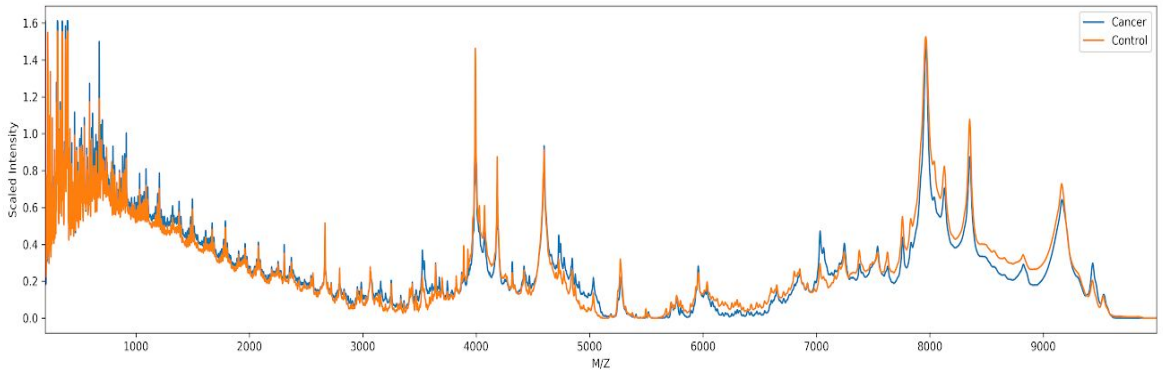


Figure 1. Overlay of mass spectrum from cancer and control group.

Data Wrangling

Instead, I will use machine learning techniques to find the difference between cancer and control group and use those differences to predict cancer. The data I am investigating are public available from National Cancer Institute.[3] Herein, I focused on two types for cancer: ovarian cancer and prostate cancer, which belongs to female and male cancer respectively. For ovarian cancer, I adopted two datasets, one is prepared by robotics, the other is prepared by hand following standard protocol. For prostate cancer, all samples were prepared by hand following standard protocol. Each dataset has cancer and control (healthy) groups (Figure 2).

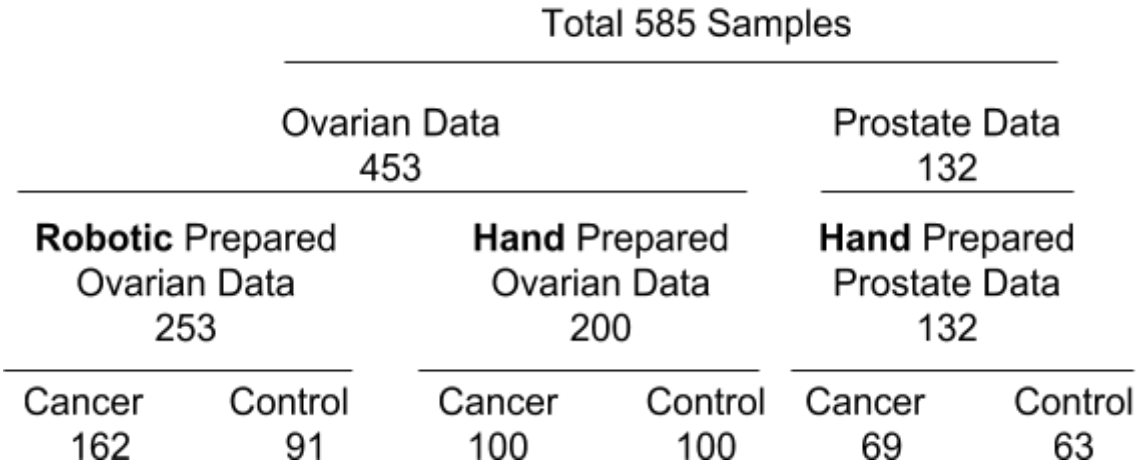
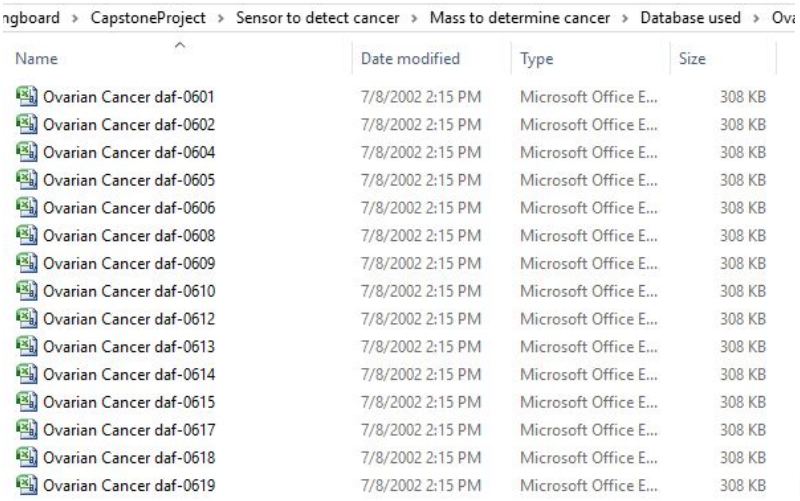


Figure 2. Mass Spectra Data (585 in total) Used for Analysis. Robotic prepared ovarian cancer dataset: 91 controls and 162 ovarian cancers; Hand prepared ovarian cancer dataset: 100 controls and 100 ovarian cancers; Prostate cancer dataset: 63 controls and 69 prostate cancers.

Original data includes 585 csv files (Figure 3). Each csv file represents the mass spectrum collected from one patient. This is a common way to store data for analysis instruments

a)



ngboard › CapstoneProject › Sensor to detect cancer › Mass to determine cancer › Database used › Ova

Name	Date modified	Type	Size
Ovarian Cancer daf-0601	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0602	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0604	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0605	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0606	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0608	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0609	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0610	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0612	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0613	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0614	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0615	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0617	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0618	7/8/2002 2:15 PM	Microsoft Office E...	308 KB
Ovarian Cancer daf-0619	7/8/2002 2:15 PM	Microsoft Office E...	308 KB

b)

	A	B
1	M/Z	Intensity
2	########	4.100553
3	2.18E-07	4.120664
4	9.60E-05	4.036199
5	0.000366	4.124686
6	0.00081	4.026144
7	0.001429	3.945701
8	0.002221	3.879336
9	0.003188	3.985923
10	0.004329	4.016089
11	0.005644	4.004022
12	0.007133	4.070387
13	0.008797	3.981901
14	0.010634	4.096531
15	0.012646	4.166918
16	0.014832	4.295626
17	0.017193	4.219206
18	0.019727	3.919558
19	0.022436	3.819005
20	0.025319	3.957768

Figure 3. a) A typical dataset was group of several csv files; b) Single csv file records M/Z and Peak Intensity

M/Z represents the mass of molecules (assuming charge = +1), intensity represents the signal intensity obtained by the instruments. As you can see, mass spectrum is with high accuracy, it can differentiate molecules with tiny difference in molecular weight. The drawback of high sensitivity is that it sometimes comes with low selectivity (percentage of TN over sum of TN and FP), so it needs professionals to identify useful information.

The data wrangling in this case is to transform data from long form to wide form, then combine different data frames into several groups. In the final data frame, each row represents one sample, each column represents one M/Z (Figure 4).

a) Single sample:

Step 5. Transform the dataframe to a 'wider' form

```
In [9]: # Pivot the dataframe, making the dataframe a 'wider' form
        df_transform= df.pivot_table(columns='M/Z')
        df_transform = df_transform.reset_index().iloc[:,1:]
```

```
In [10]: # A wide format of M/Z ratio
         df_transform
```

Out[10]:

M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202.29642	202.56198	...	9982.7063	9984.5713	99
0	1.720546	1.720546	1.720546	1.719901	1.717815	1.713023	1.692505	1.672132	1.63622	1.560843	...	0.0	0.0	0.0

1 rows × 9200 columns

b) Samples after concat:

```
In [88]: cancer_group.head()
```

Out[88]:

M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202.29642	202.56198	...	9982.7063	9984.5713	99
0	1.720546	1.720546	1.720546	1.719901	1.717815	1.713023	1.692505	1.672132	1.636220	1.560843	...	0.0	0.0	0.0
1	1.524127	1.501587	1.463949	1.407535	1.325125	1.226132	1.120034	1.005362	0.893435	0.777613	...	0.0	0.0	0.0
2	1.637911	1.637911	1.631477	1.612273	1.581629	1.525683	1.435933	1.330684	1.193543	0.997459	...	0.0	0.0	0.0
3	1.656036	1.656036	1.656036	1.651223	1.634340	1.590969	1.525195	1.444791	1.340823	1.217818	...	0.0	0.0	0.0
4	1.793301	1.793301	1.793301	1.793301	1.791395	1.785510	1.767697	1.722631	1.655785	1.542241	...	0.0	0.0	0.0

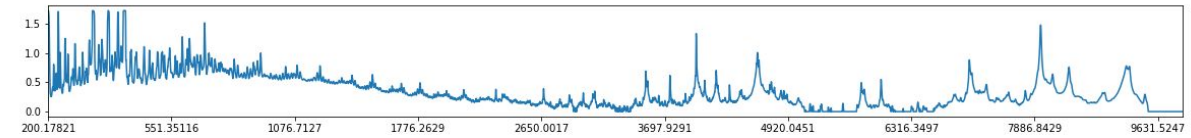
5 rows × 9200 columns

Figure 4. Comparison of data before and after combination

Exploratory Data Analysis

The typical way of viewing spectrum is shown in Figure 5a, where x-axis represent M/Z value, while y represents the scaled peak intensity. However, this visualization tends to be messy if comparing tons of spectra. In contrast, 1D heatmap offers a better way to view spectrum (Figure 5b), especially when it is necessary to compare different samples. In this case, X-axis represents M/Z value, the brightness of lines represent the intensity of peaks. The brighter the line is, the higher intensity of the peak is.

a) Typical view of mass spectra



b) 1D Heatmap

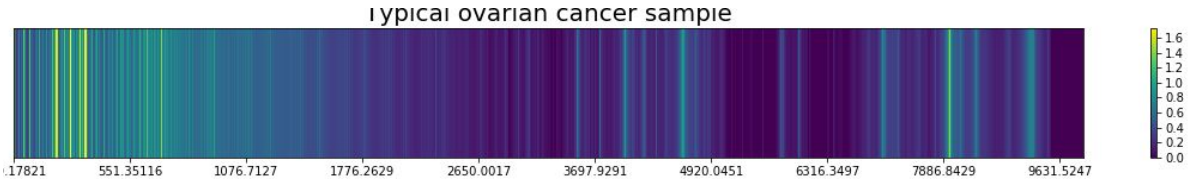
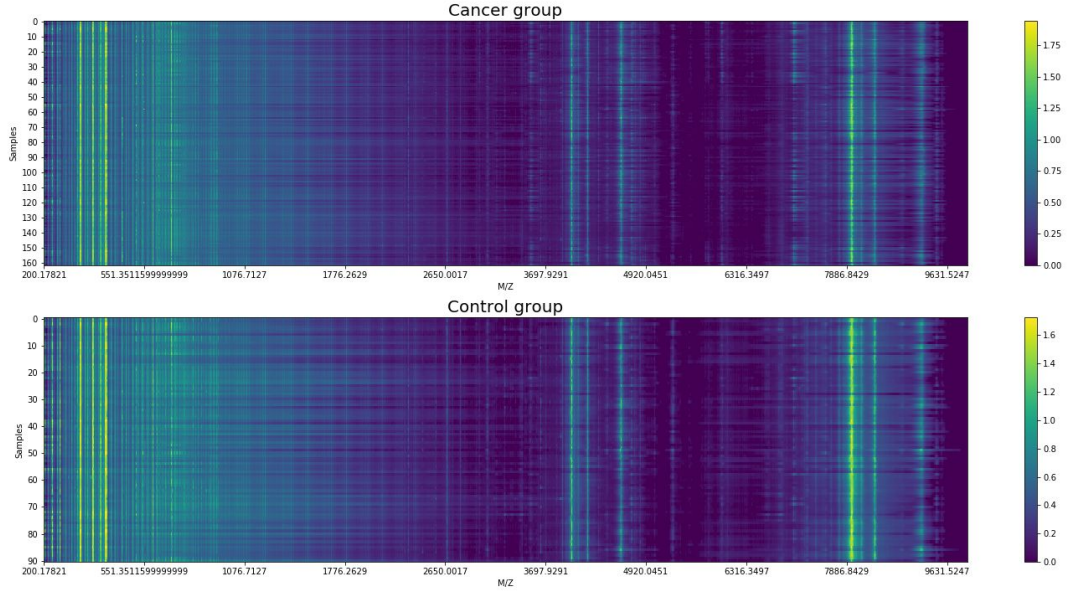


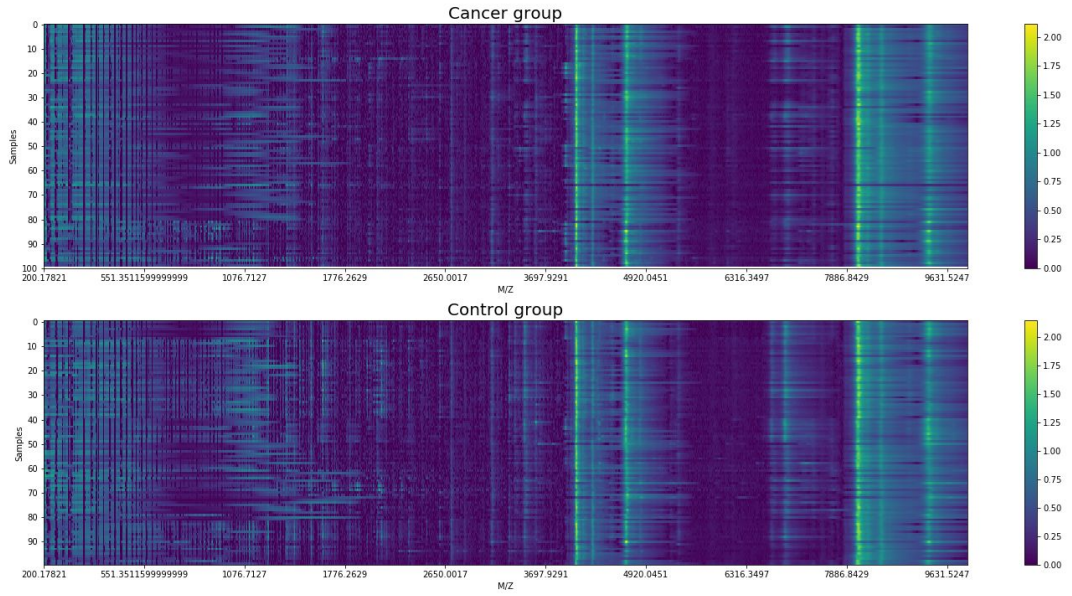
Figure 5. Visualization of Mass Spectrum

I can thus create heatmaps for three datasets by combining samples: robotic prepared ovarian datasets, hand prepared ovarian datasets, and hand prepared prostate datasets. It can easily be seen that robotic prepared samples showed better quality than hand prepared samples: less variance and more consistent, which is always good for data analysis and machine learning. I will jump into this topic again in later discussion.

a) Heatmap of robotic prepared ovarian datasets



b) Heatmap of hand prepared ovarian datasets



c) Heatmap of hand prepared prostate datasets

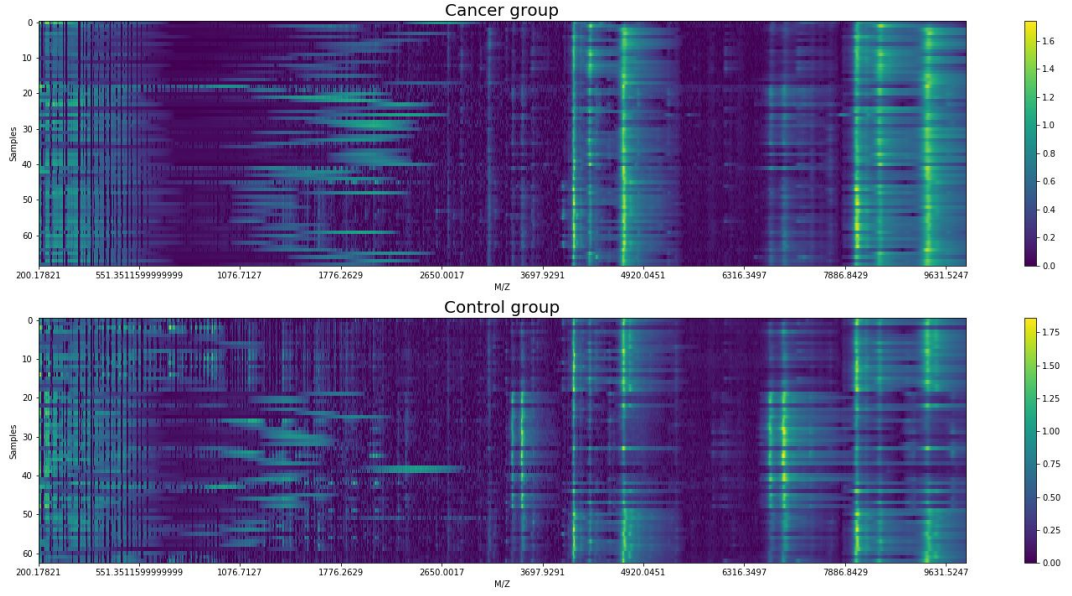


Figure 6. Overlay of heatmap of mass spectra of three different datasets

The heatmap is still too complicated for us to know the difference between control and cancer groups. So the next question is: can we visualize how easily our data could be separated into cancer and non-cancer group? It would be great to get a general idea of where should we focus on before we do any further deep analysis. In this case,

exploratory data analysis (EDA) is necessary to evaluate the dataset (e.g., complexity and data quality).

Rightnow we are facing a dataset with significantly larger features (different masses) than samples (number of mass spectra). This is common for all spectra data, where it is relatively difficult to collect large amount of samples through experiments, but it was fairly easy to obtain tons of features or data points through spectroscopy analysis. Principal Component Analysis provides us with great tools to project high dimensional data into two-dimensional space, where we can easily see and know our data.[4] Herein, I would like to visualize the data distribution using the first two principal components (Figure 7).We can see that for robotic prepared ovarian samples and prostate samples, cancer and non-cancer samples could be reasonably separated, while for hand prepared ovarian samples, cancer and non-cancer samples are largely overlapped and cannot be separated using only the first two principal components (more difficult to predict cancer/non cancer compared with predict robotic prepared ovarian datasets and prostate datasets).

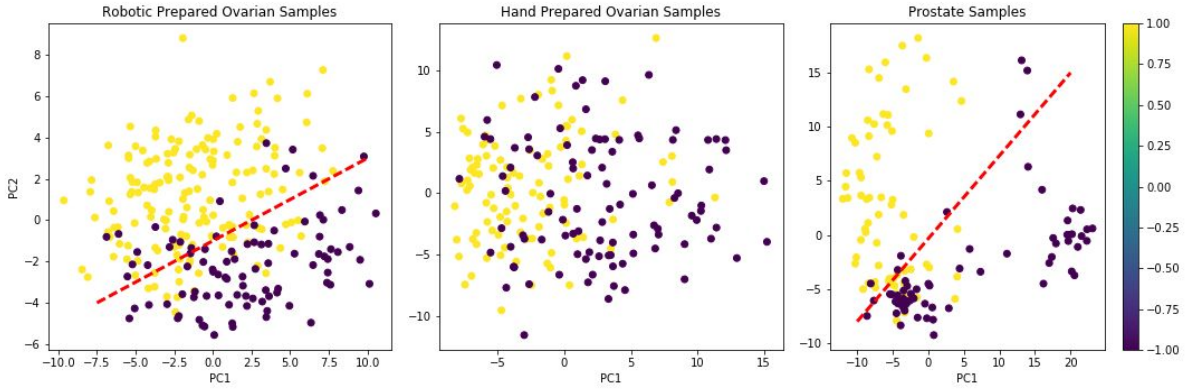


Figure 7. Comparison of cancer and non-cancer group in three datasets. Purple plots represent non-cancer group, while yellow plots represent cancer group. From left to right: robotic prepared ovarian samples, hand prepared ovarian samples, and prostate sample

In-depth analysis using machine learning

We know that our spectra data is high-dimensional data. In fact, high dimension data not only bring the curse of high dimensionality, but also bring correlated and noise features, which might cause our model to overfit the data or difficult to converge. So we need to select important features before we apply machine learning algorithm.

Decision Tree is a natural way of feature selection. Tree split is based on maximum gain of gini impurity, so tree always splits towards more important features. Random forest algorithm is an ensemble method using tree bagging and random feature selection for each split. Herein, I used Random Forest to select the most important features. I set the threshold at 95%, meaning I expected the most important features could explain more than 95% variance of dataset.

Plot the cumulative explained variance of features

It is noted that within 9200 features(M/Z), using only 40 features(0.43% of total features) can explain more than 95% variance for prostate samples, 52 features (0.58% of total features) will explain more than 95% variance for robotic prepared ovarian samples and 86 features (0.93% of total features) is needed for hand prepared ovarian samples (Figure 8). Feature selection will significantly reduce the noisy and redundant features.

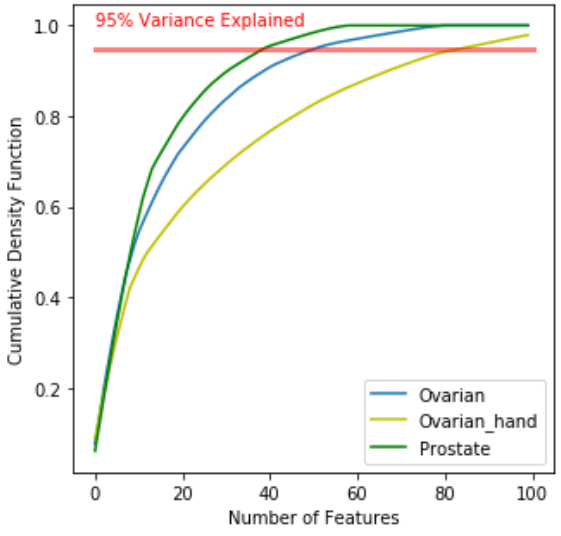


Figure 8. Explained Variance vs. Number of Features

Comparison of model performance of SVM, Random Forest, KNN and Ensemble method

Herein, I applied supported vector machine (SVM), random forest (RF), and k nearest neighbors (KNN) and ensemble method by voting for cancer prediction using selected features. It is noticed that for the prediction of both ovarian and prostate cancers, all ML models performs well: For robotic prepared ovarian data, Random Forest and SVM can achieve 100% accuracy; For hand prepared ovarian data, SVM and Ensemble method perform similar well and achieved 95% accuracy; For prostate data, SVM, Random Forest and Ensemble method can achieve up to 98% accuracy. What I mean here by accuracy is that given 98% accuracy, the prediction has 98% chance to be correct. However, we should not be too confident on our models, because we will definitely need much larger data to optimize our models and test the model performance. Our models also have to be flexible, meaning it can deal with situations where more noise than usual appears in the mass spectrum (e.g., the impurities in the samples during sample preparation and instruments errors).

Herein, I added ensemble models to know whether it will improve the performance of prediction. It is shown that for robotic prepared ovarian cancer prediction. In all the predictions, ensemble model performs similarly as single models in terms of prediction accuracy, but the single model, especially SVM, has a 0% of false negative rate, which is what we were targeting for in cancer screening. For cancer screening, we need to have lower false negative rate, because our model needs to be sensitive enough to ‘red flag’ suspected cancer patients for further diagnosis.

Table 1. Comparison of different models on cancer prediction

Datasets	Measure	Models			
		KNN	Random Forest	SVM	Ensemble by Voting
Ovarian Robotic	Accuracy	0.99	1.00	1.00	0.99

	AUC	0.99	1.00	1.00	0.99
	F1-Score	0.99	1.00	1.00	0.99
Ovarian Hand	Accuracy	0.93	0.92	0.95	0.95
	AUC	0.93	0.91	0.94	0.96
	F1-Score	0.94	0.93	0.96	0.96
Prostate	Accuracy	0.95	0.98	0.98	0.98
	AUC	0.95	0.97	0.97	0.98
	F1-Score	0.96	0.98	0.98	0.98

Fingerprint molecules that determine ovarian cancer and prostate cancer

From feature selection rendered by random forest, we can easily obtained the important features by sorting the variance explained from highest to lowest. The important features correspond to the molecular weights of fingerprint molecules that determine ovarian and prostate cancer.

We set the mass window from 200~1000, which include small molecules that could be further separated, purified and characterized. Interestingly, **one of the literature reported key molecules (molecular weight 472) to determine ovarian cancer is in our important mass list for ovarian prediction**. In another word, **we developed a tool to select the possible fingerprint molecules for cancer diagnosis, which is of great beneficial for new discovery of metabolisms and cancer-causing molecules**. In this case, instead of focusing on all 9300 possible molecules, researchers could just focus on 52 molecules for ovarian cancer prediction, or 40 molecules for prostate cancer prediction, which will greatly improve R&D efficiency and save cost.

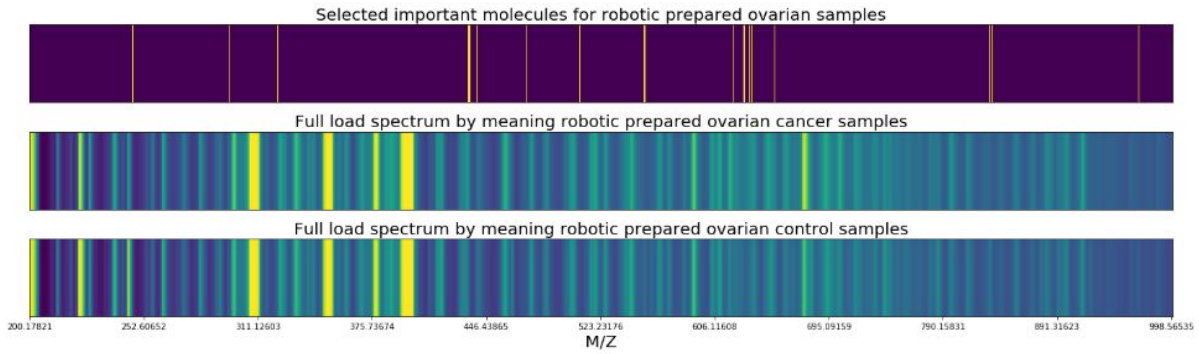


Figure 9. Selected import molecules for robotic prepared ovarian samples

If samples were accidentally mixed up, can we tell which group it belongs to?

There are often cases when people mixed up samples, especially dealing with large amount of samples. Herein, we offered a solution of how to use machine learning tools to assign the unknown samples to the group. In this dataset, we have 6 individual group, indicated by Figure 2. We have to decide which group the sample belongs to using multi classification. Comparing three models (e.g., SVM, Random Forest, and KNN), we concluded SVM performs best in this multi classification, with up to 93% accuracy. It was further proved that our model could separate samples according to sex (up to 97% accuracy), and robotic prepared and hand prepared (up to 100% accuracy).

Conclusion

For the prediction of ovarian and prostate cancers, SVM were selected as the best model in terms of accuracy (95-100%), and 0% false negative rate. Best model for multi classification (e.g, totally six groups) was selected to be SVM, achieving up to 93% of accuracy. We also successfully identified one of the fingerprint molecules determining ovarian cancer, which is confirmed by literature report. In another word, we developed a tool to select potential fingerprint molecules for cancer diagnosis. This work not only offers early cancer detection product, but also will be a powerful tool for professionals to discover new fingerprint molecules for cancer diagnosis. Specifically, instead of focusing on all

9300 possible molecules, researchers could just focus on 52 molecules for ovarian cancer prediction, or 40 molecules for prostate cancer prediction, which will greatly improve R&D efficiency and save cost.

Reference

1. How cancer is diagnosed?
https://www.cancer.org/treatment/understanding-your-diagnosis/tests/testing-biopsy-and-cytology-specimens-for-cancer/how-is-cancer-diagnosed.html

2. General knowledge on mass spectrometry was referenced here: https://en.wikipedia.org/wiki/Mass_spectrometry

3. Original mass spectra are downloaded from
https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp
Data Also available through
https://github.com/liudj2008/CapstoneProject1

4. For Principal Component Analysis, please refer to
https://en.wikipedia.org/wiki/Principal_component_analysis

5. Reference paper is available through
https://academic.oup.com/ajcp/article/134/6/903/1760577
The authors used other detection methods (e.g., HPLC-HRMS) to collect mass spectrum data from other ovarian cancer patients. Interestingly, their results closely agree with our discovery using public data