High Accuracy Cancer Prediction Using Machine Learning Inspired Mass Spectrum Analysis

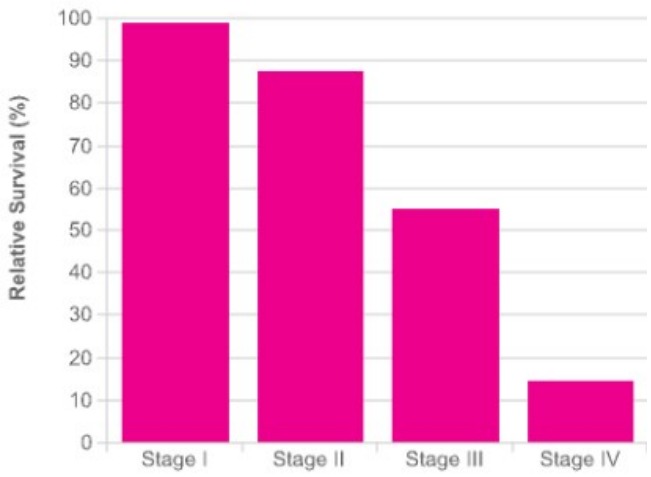
Peter Liu

2018.6

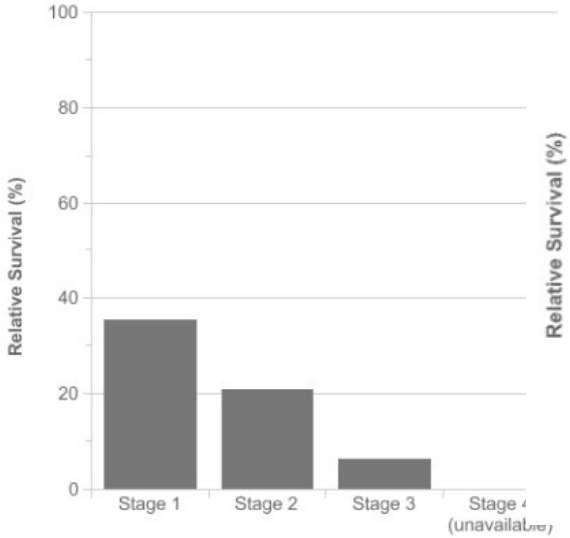
Top 3 Cancers in United States

Five Years' Survival Rates at Different Stages

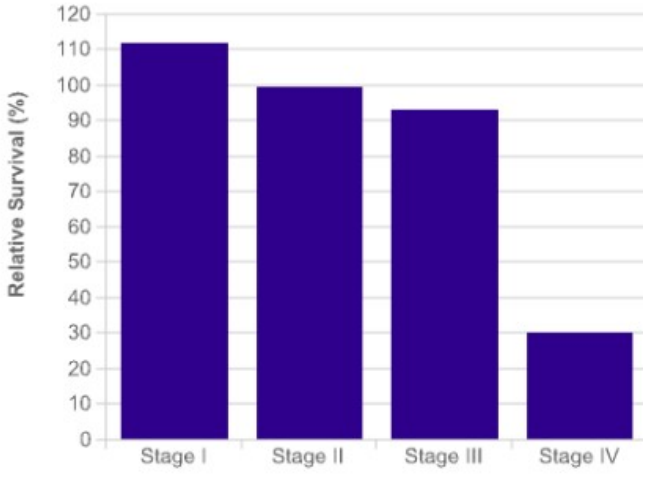
Breast Cancer



Lung Cancer



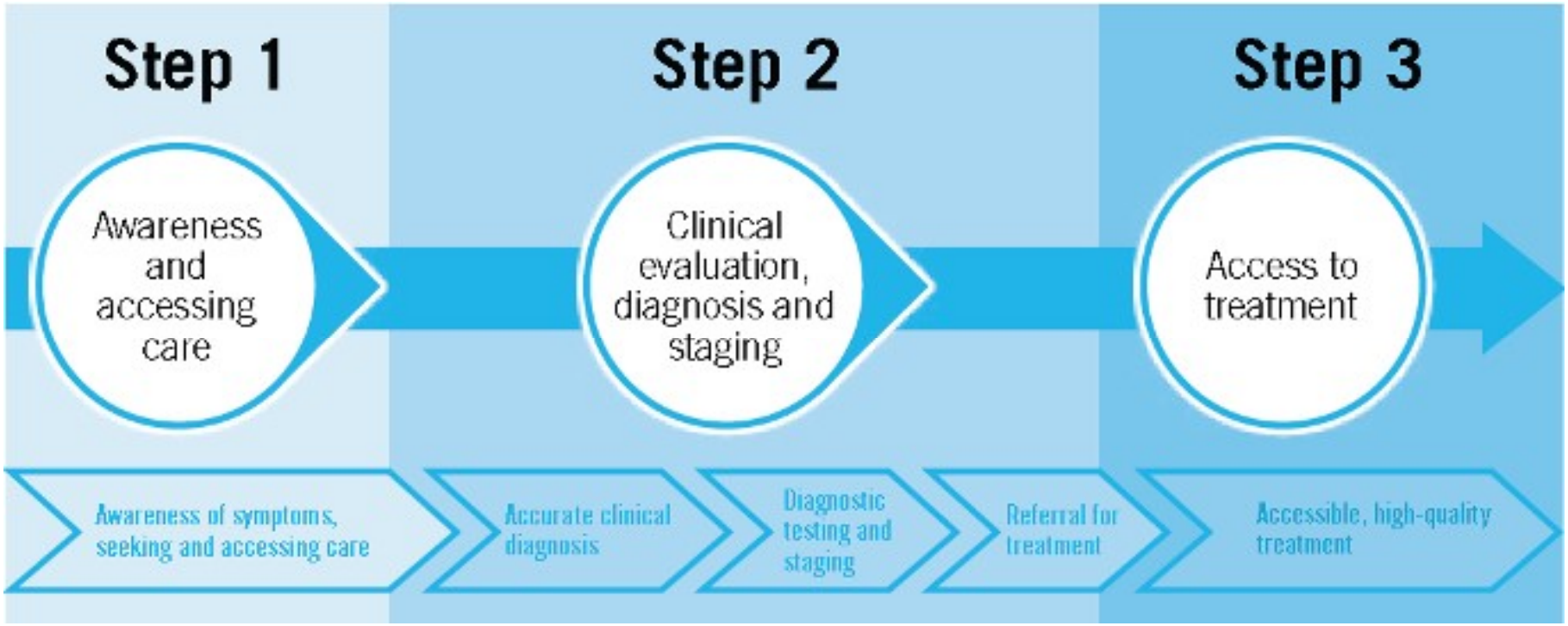
Prostate Cancer



- Survival rate decreases exponentially when cancer is diagnosed at later stage
- **Early Detection is important to increase SURVIVAL RATE!**

Cancer Statistic Source from: http://www.cancerresearchuk.org

Cancer Diagnosis



Problems:

- Some cancer does not have symptoms
- Not noticeable until late stage
- Routine check up specifically for cancer is not affordable

Early Detection of Cancer

Goal

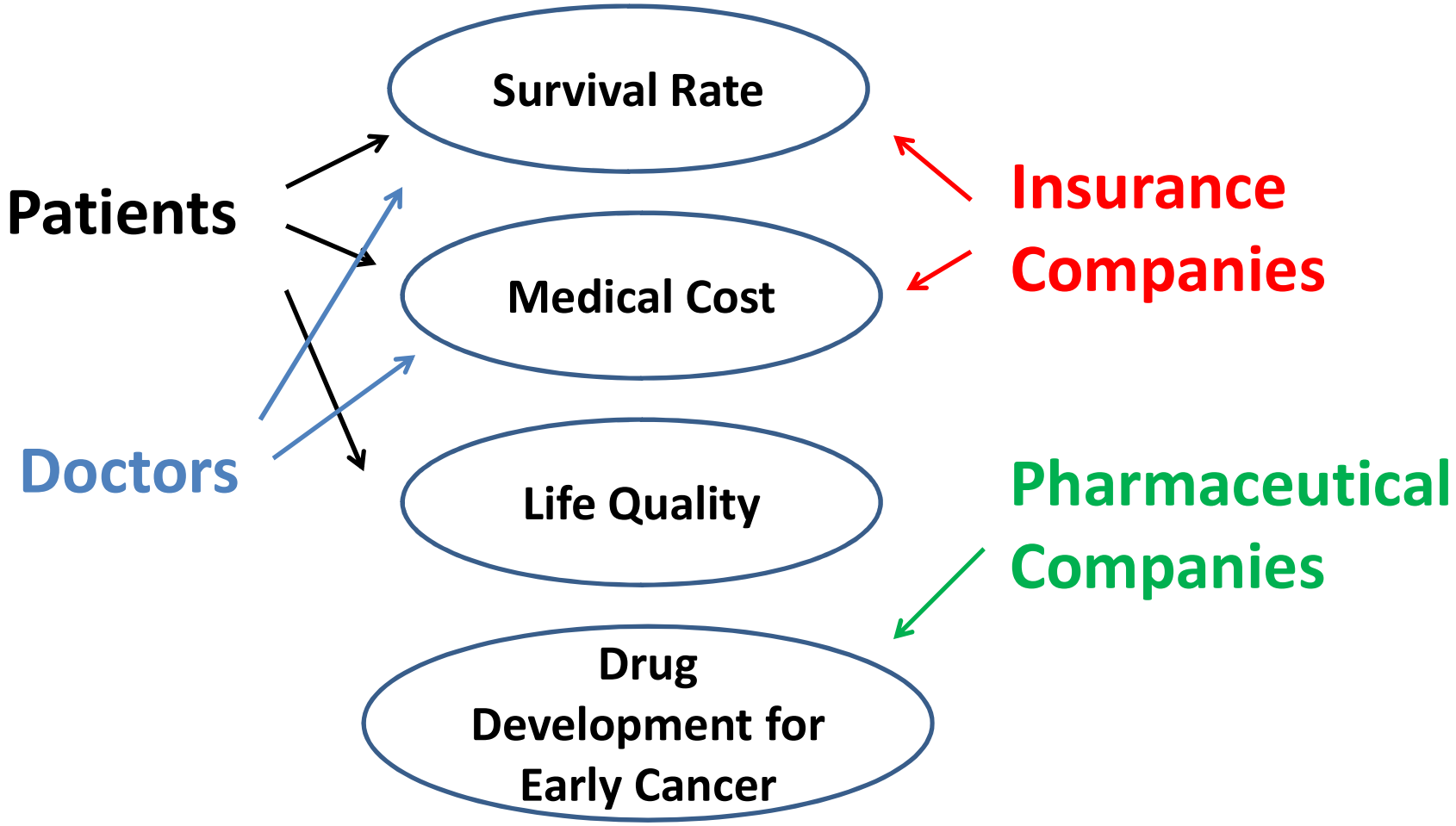
- SHOTGUN METHODS

  Analyze routine check up samples, *e.g.* blood samples, and **collect as much information as possible** for cancer detection

- 'RED FLAG' SUSPECTS

  Identify samples that have high probability of cancer and recommend for further testing

Who Cares?


Survival Rate
Medical Cost
Life Quality
Drug Development for Early Cancer
Patients
Doctors
Insurance Companies
Pharmaceutical Companies

Shotgun Method

- **Mass Spectroscopy**

- Collect mass information of all chemicals

- **Low Sample Loading**

- Milligram (1/1000 grams) samples

- **High Sensitivity**

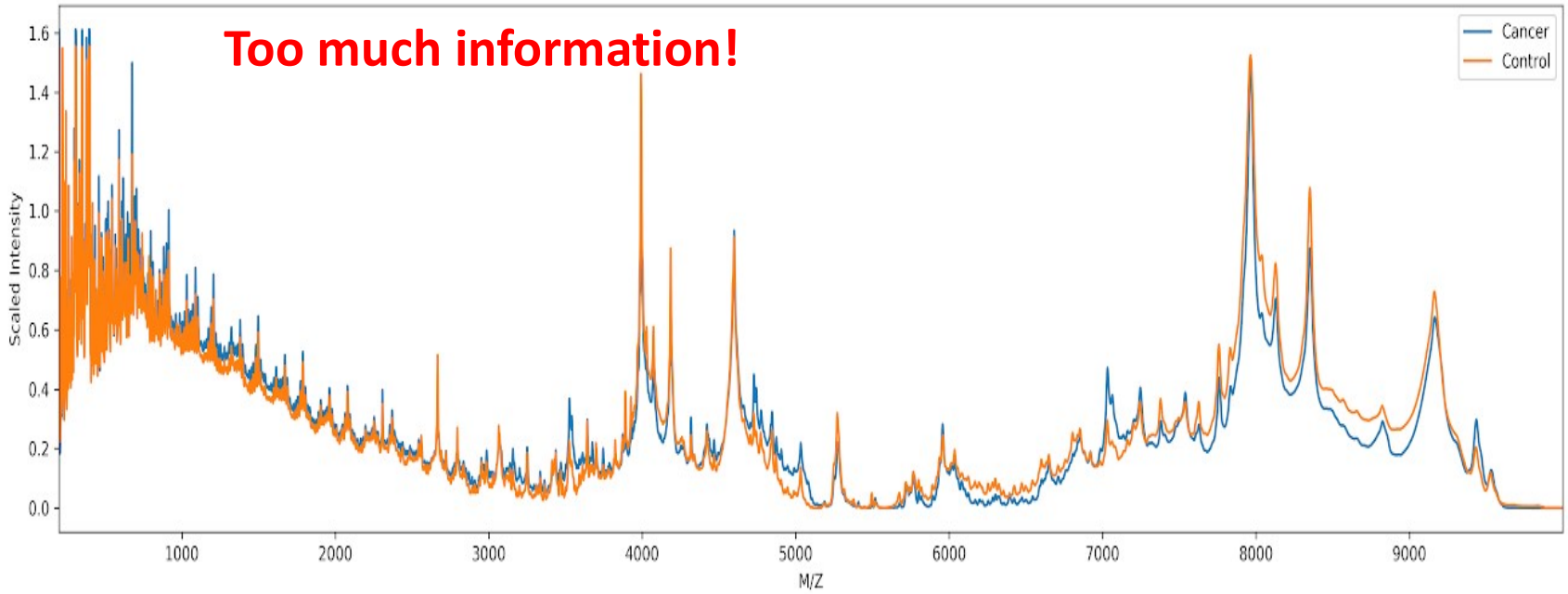
- Detect trace amount of chemicals at part per billion(ppb) level

- **High Throughput Screening**

- Easily coupled with robotic sample preparation process and results obtained within minutes

Problems with Shotgun Method

- Difficult to compare unless you already know which peak is the determinant



Real-world Problem to ML Problem

Real-world Problem

1. Select Determinant Masses
2. Predict Cancer

Machine Learning Problem

1. Feature Engineering
2. Classification

Determinant Masses Selection vs. Important Features Selection

Cancer/No Cancer

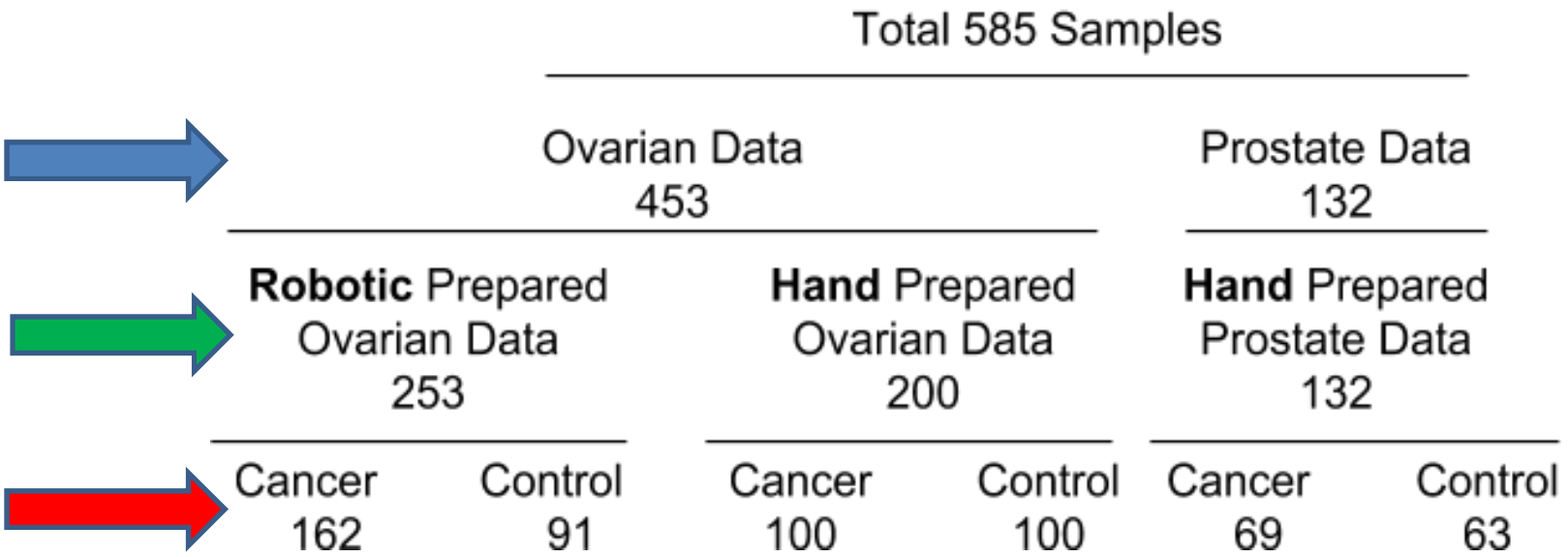
vs.

1/-1

M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202.29642	202.56198	...	9982.7063	9984.5
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
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
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
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
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

Data Source

- Sample mass spectra collected from National Cancer Institute (NCI)
- **Two cancers**
- **Three groups**
- **Six subgroups**



Data Wrangling

‘Long’ single MS data

M/Z	Intensity
########	4.100553
2.18E-07	4.120664
9.60E-05	4.036199
0.000366	4.124686
0.00081	4.026144
0.001429	3.945701
0.002221	3.879336
0.003188	3.985923
0.004329	4.016089
0.005644	4.004022
0.007133	4.070387
0.008797	3.981901

Mass cutoff
Scaling
Transpose

‘Wide’ single MS data

Mass values

M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202
0	1.720546	1.720546	1.720546	1.719901	1.717815	1.713023	1.692505	1.672132	1.6

Concatenation by Row

Mass Spectra Data Matrix

Mass values

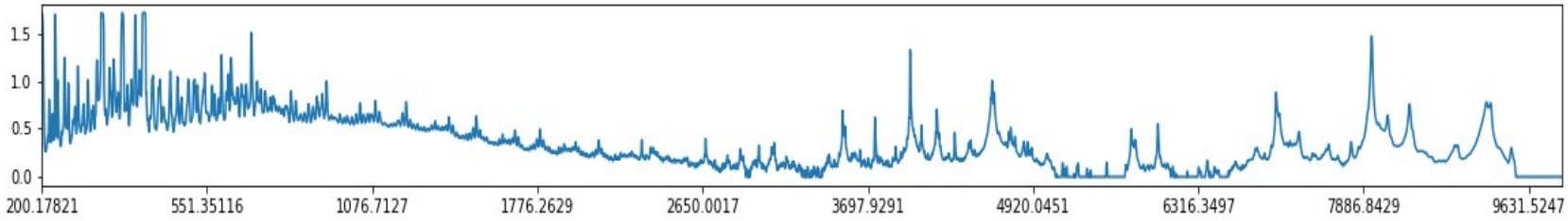
Samples

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	998
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

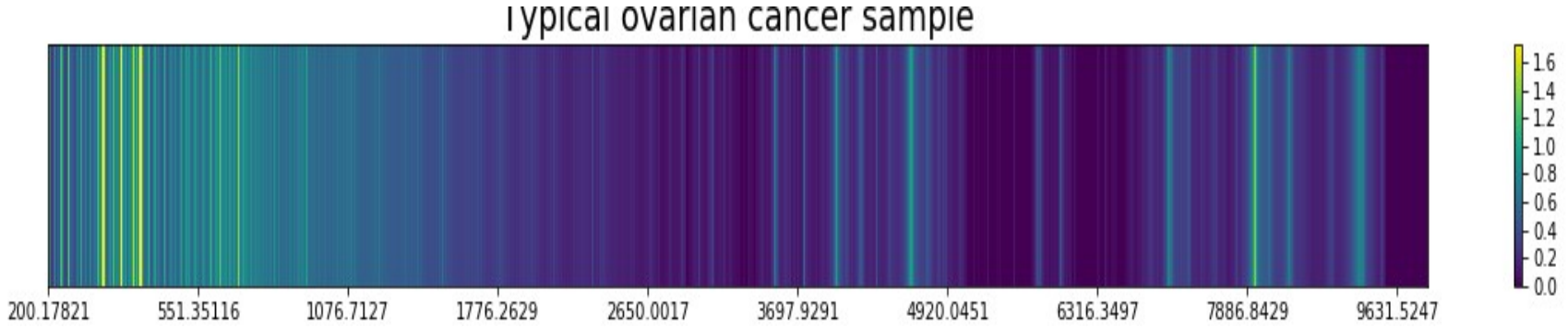
Exploratory Data Analysis

Heatmap is more preferable than plot view

Plot view of data



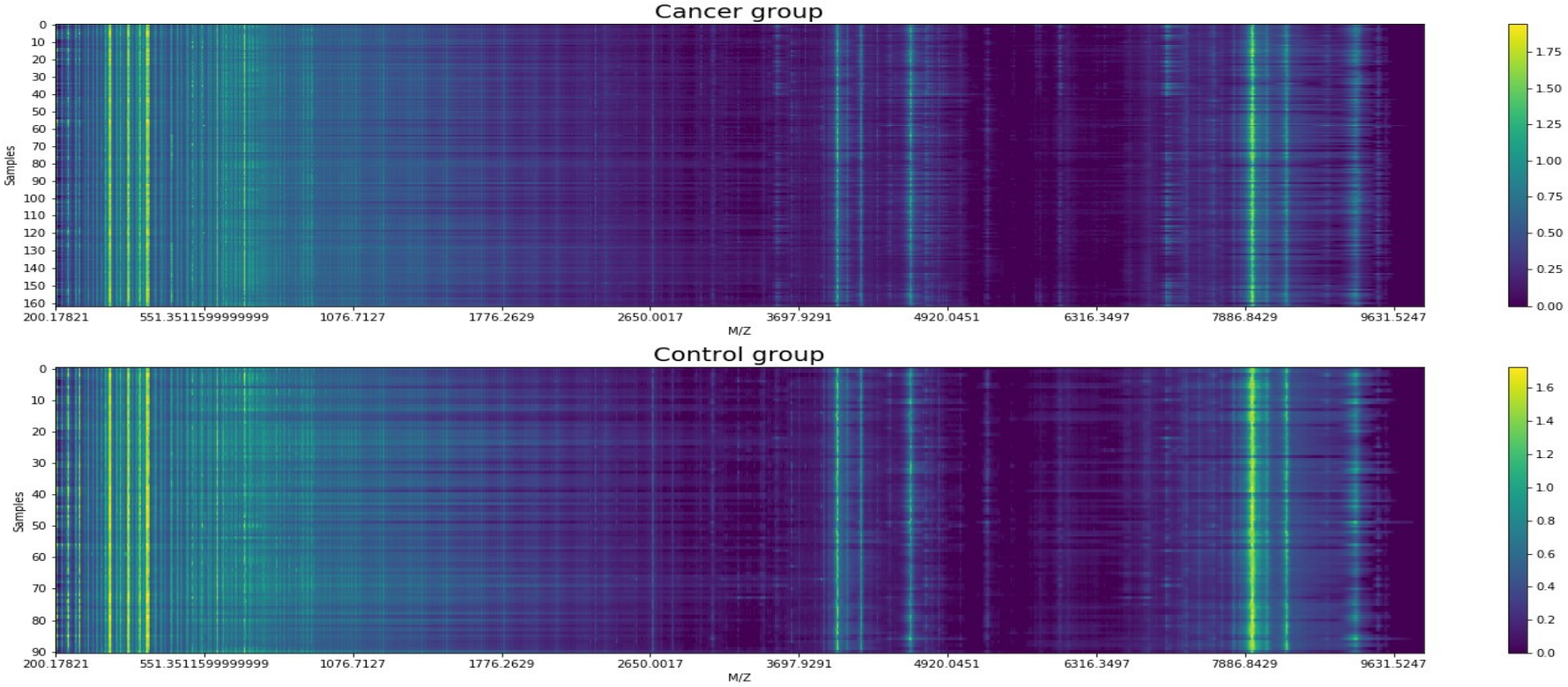
1D Heatmap



Exploratory Data Analysis

Heatmap of robotic prepared ovarian datasets

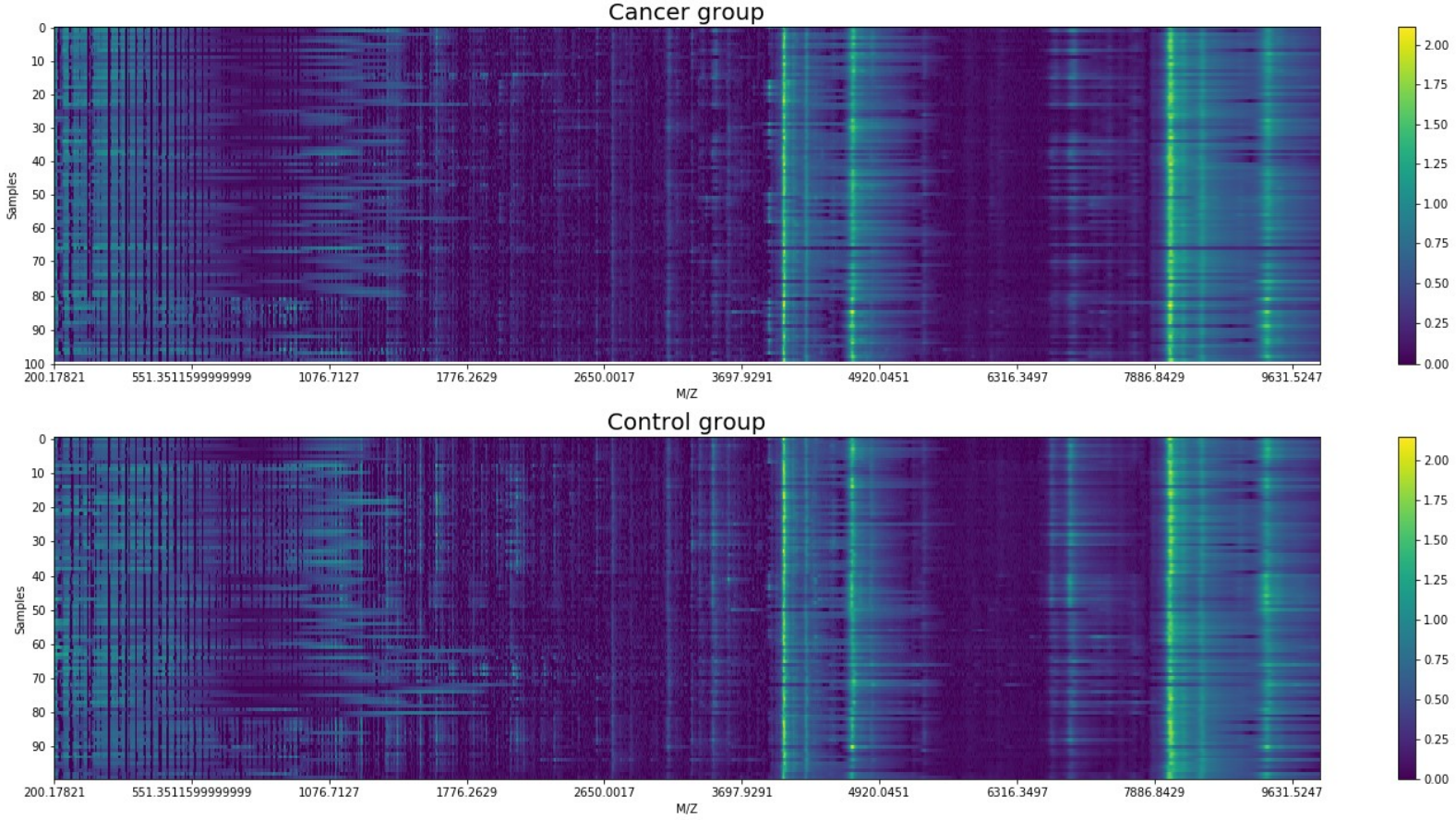
- Difficult to tell the difference between cancer and control group



Exploratory Data Analysis

Heatmap of hand prepared ovarian datasets

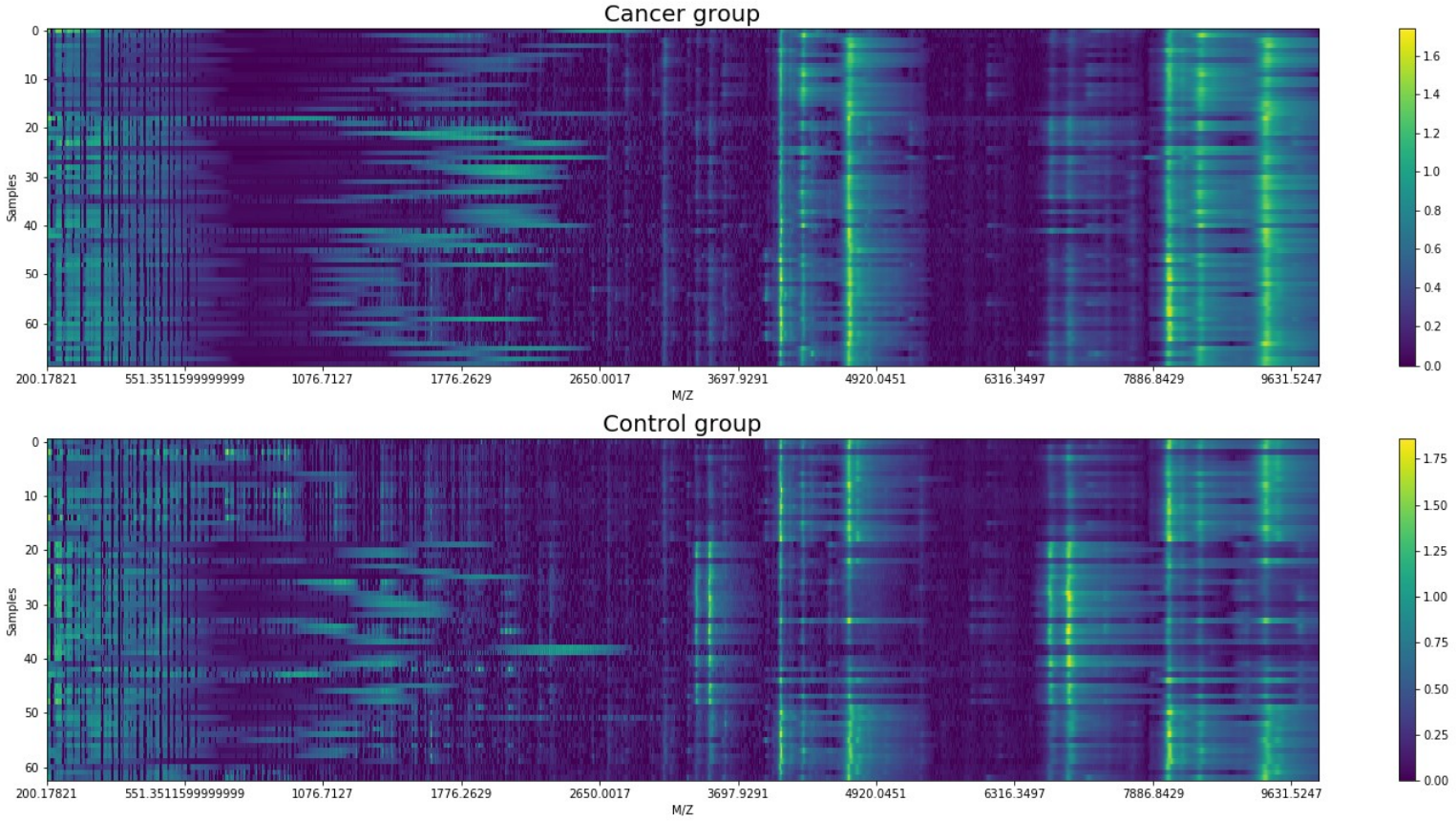
- Difficult to tell the difference between cancer and control group



Exploratory Data Analysis

Heatmap of hand prepared prostate samples

- Difficult to tell the difference between cancer and control group



Data Visualization in 2D

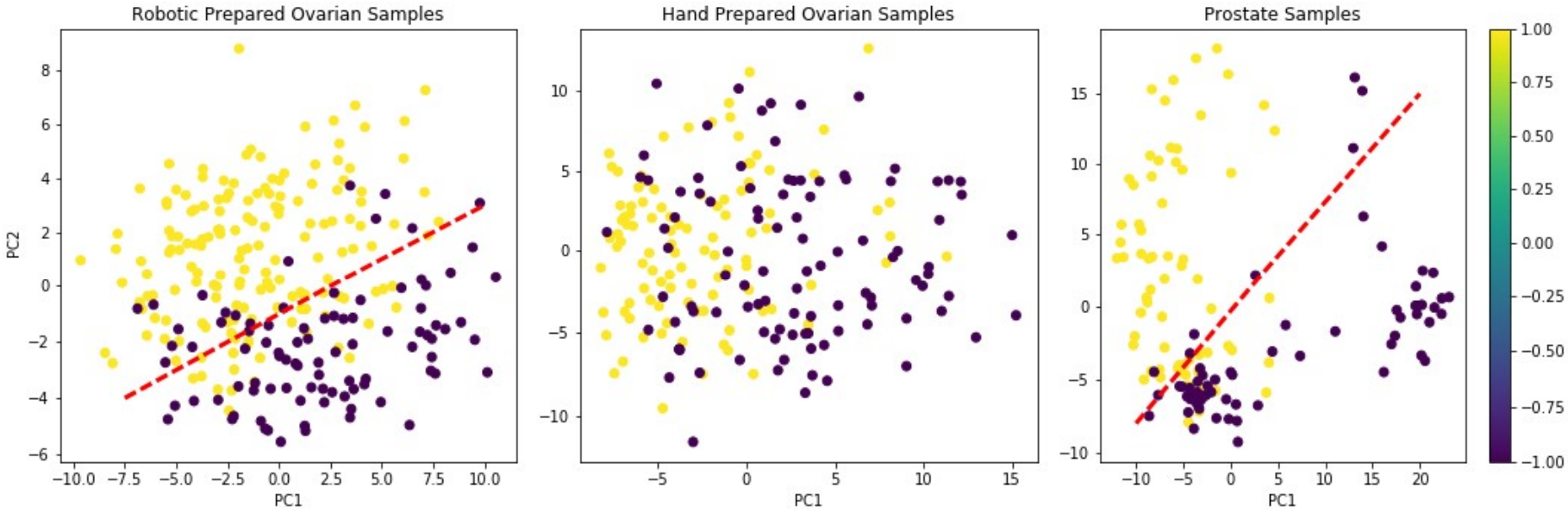


Figure. Comparison of cancer and non-cancer group in three datasets. **Purple plots represent non-cancer group, while yellow plots represent cancer group. From left to right: robotic prepared ovarian samples, hand prepared ovarian samples, and prostate sample**

Feature Selection

Important Features == Fingerprint Mass

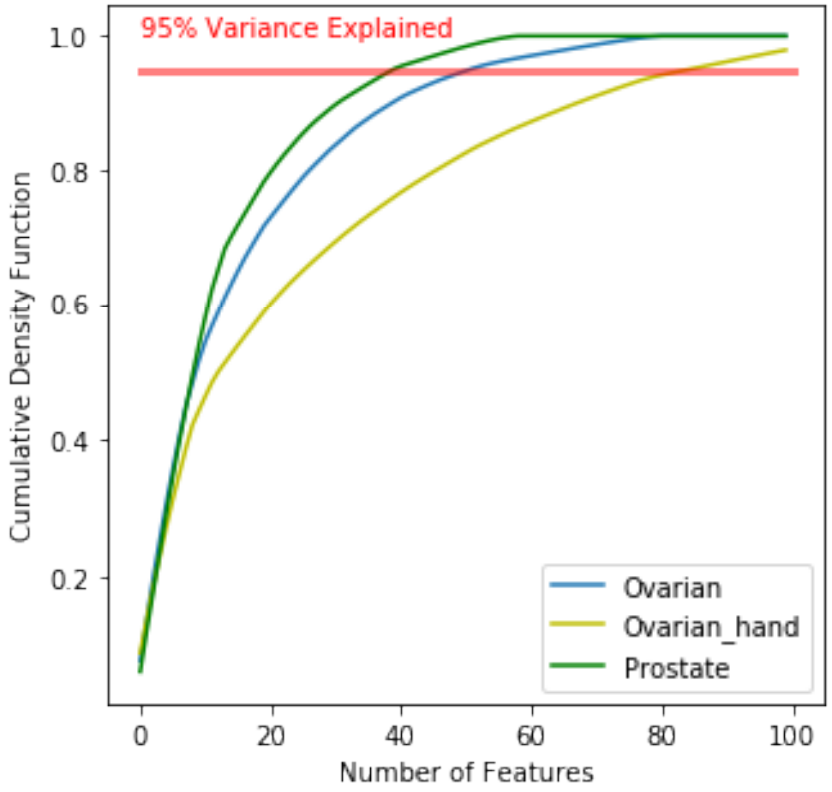


Figure. Explained Variance vs. Number of Features rendered by **Random Forest.** Decision Tree is a natural way of feature selection.

Feature Selection

Fingerprint masses for robotic prepared ovarian samples

- The number of fingerprint masses between 200 and 1000 are 25
- **One key molecule (molecular weight 472) to determine ovarian cancer is in our important mass list for ovarian prediction**

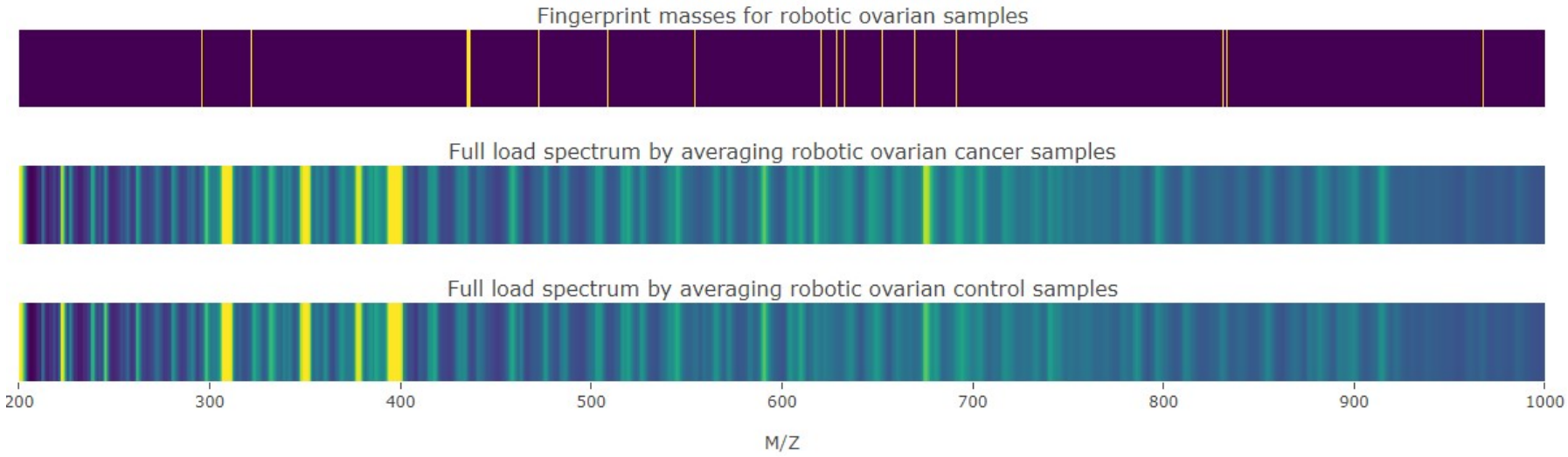


Figure. Selected fingerprint masses for robotic prepared ovarian samples

Models for Cancer Prediction

Table 1. Comparison of different models on cancer prediction

Datasets	Measure	Models after tuning parameters			
		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

FP and FN

Which model is better in early determination of cancer?

- We want lower rate of false negative (FN/fail to detect cancer)

Model 1

Predicted Actual	-1 (No Cancer)	1 (Cancer)	Total
-1 (No Cancer)	23	3 (FP)	26
1 (Cancer)	0 (FN)	34	34
Total	23	37	60

Model 2

Predicted Actual	-1 (No Cancer)	1 (Cancer)	Total
-1 (No Cancer)	23	0 (FP)	26
1 (Cancer)	3 (FN)	34	34
Total	23	37	60

SVM vs. Ensemble

SVM

Prostate Samples	Robotic prepared Ovarian Samples	Hand prepared Ovarian Samples
Confusion Matrix: Predicted -1 1 __all__ Actual -1 27 0 27 1 0 49 49 __all__ 27 49 76	Confusion Matrix: Predicted -1 1 __all__ Actual -1 23 3 26 1 0 34 34 __all__ 23 37 60	Confusion Matrix: Predicted -1 1 __all__ Actual -1 16 1 17 1 0 23 23 __all__ 16 24 40

Ensemble

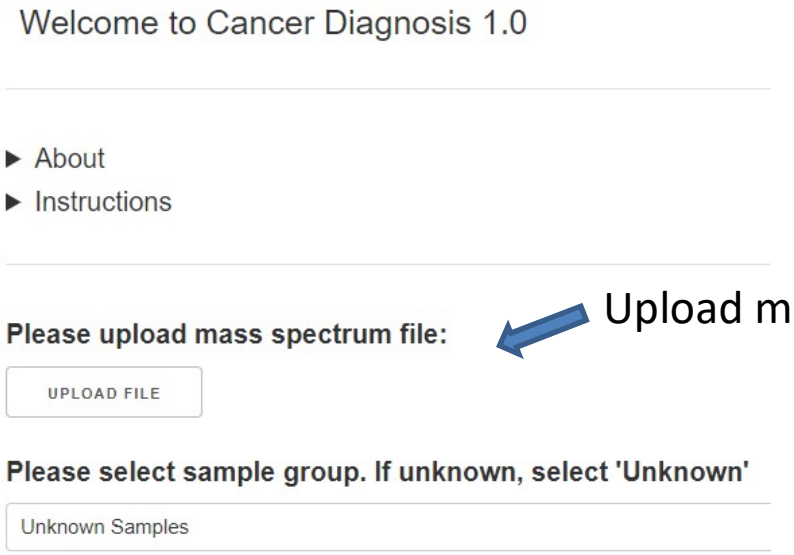
Predicted -1 1 __all__ Actual -1 27 0 27 1 1 48 49 __all__ 28 48 76	Predicted -1 1 __all__ Actual -1 22 1 23 1 4 33 37 __all__ 26 34 60	Predicted -1 1 __all__ Actual -1 16 0 16 1 1 23 24 __all__ 17 23 40

- SVM is our best model
- 0% of FN rate (fail to detect cancer) by SVM

Cancer Diagnosis 1.0

- **Web App** developed based on Dash
- Simply **upload spectrum file** and cancer diagnosis results will be shown

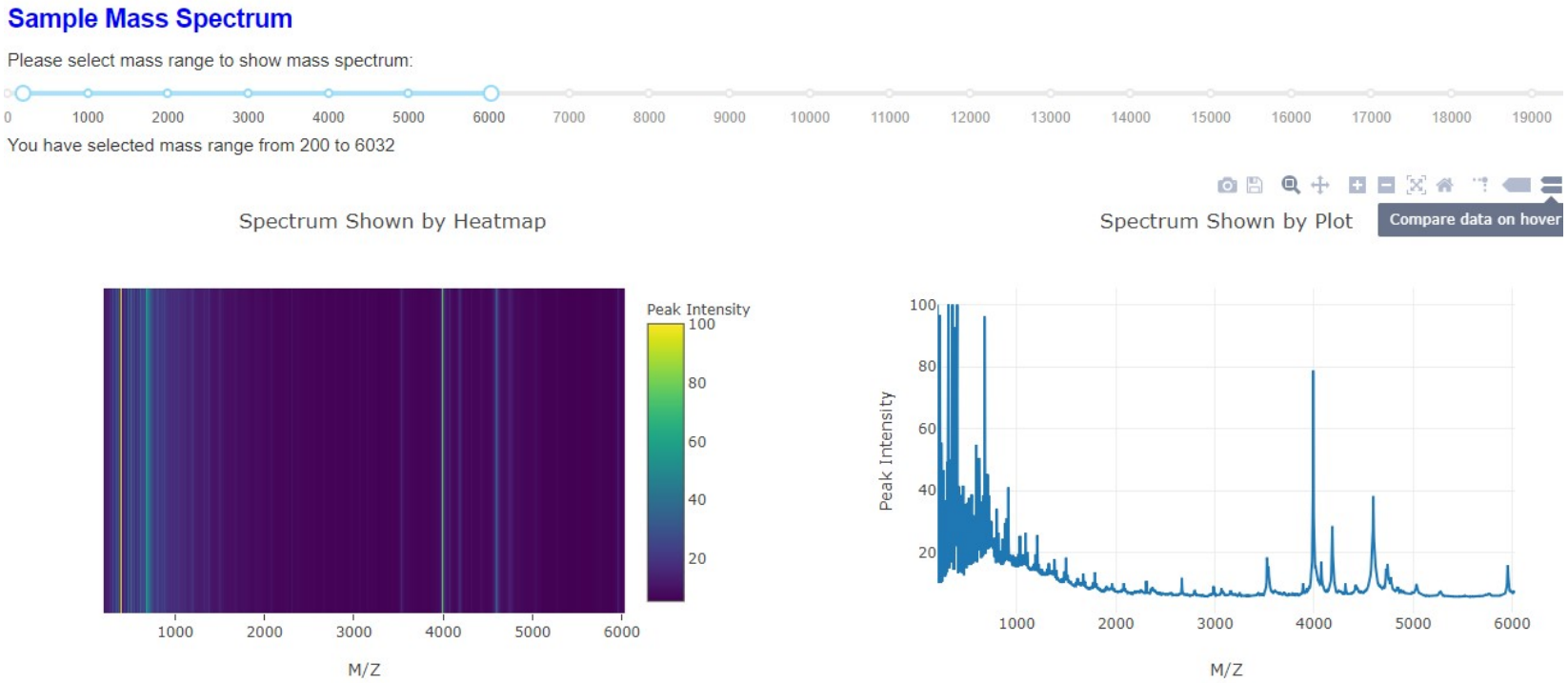
Upload file



Upload mass spectrum csv/excel from your own computer

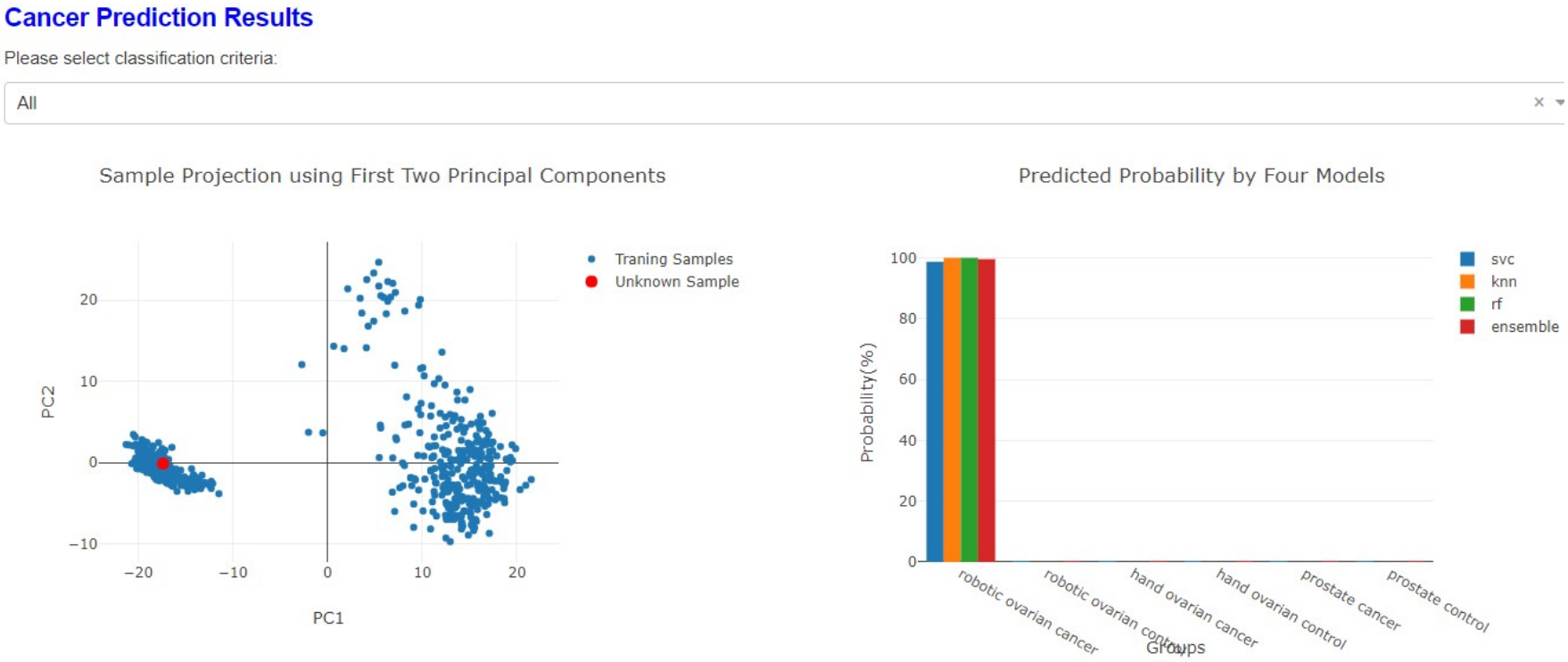
Mass Spectrum Preview

- Mass spectrum will be shown using heatmap and plot, and you can choose the mass range



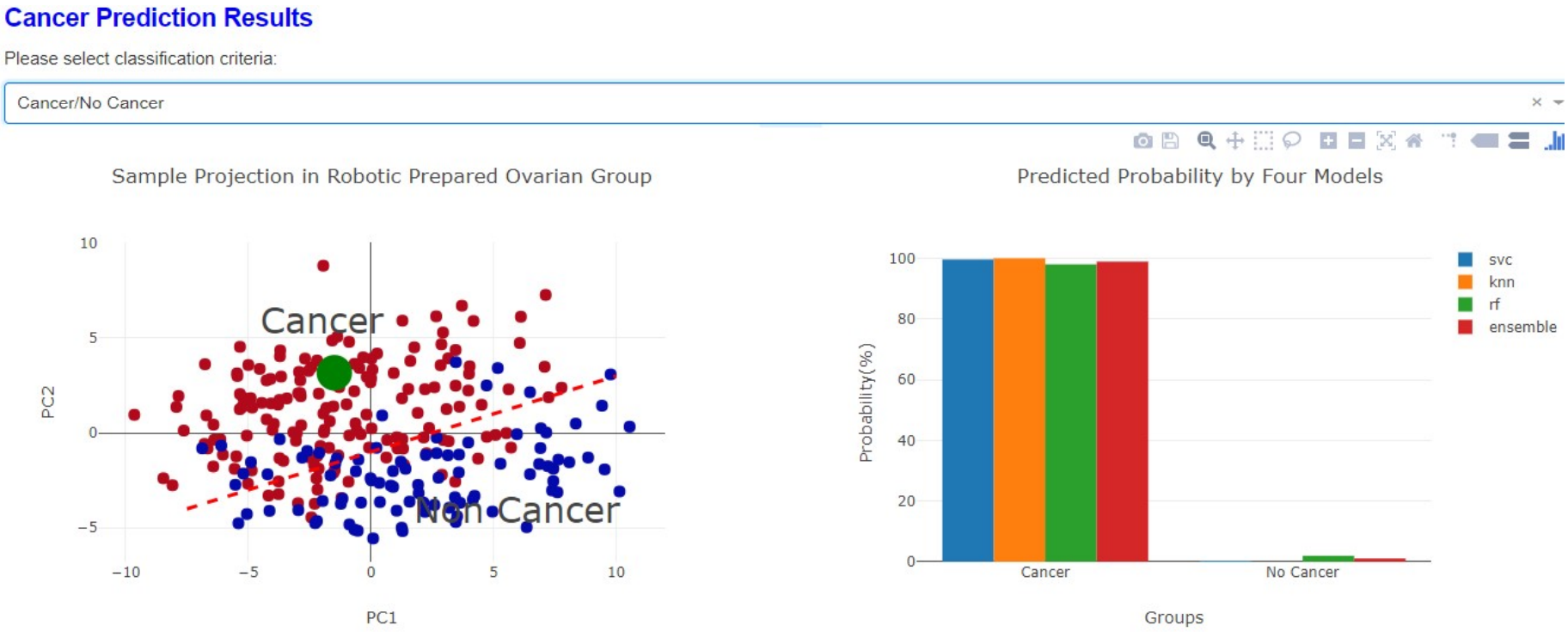
Classification of Unknown Sample

- It shows the visualization of new sample within training samples and predict the probability by four models. You can choose different classification criteria: all, sex or preparation



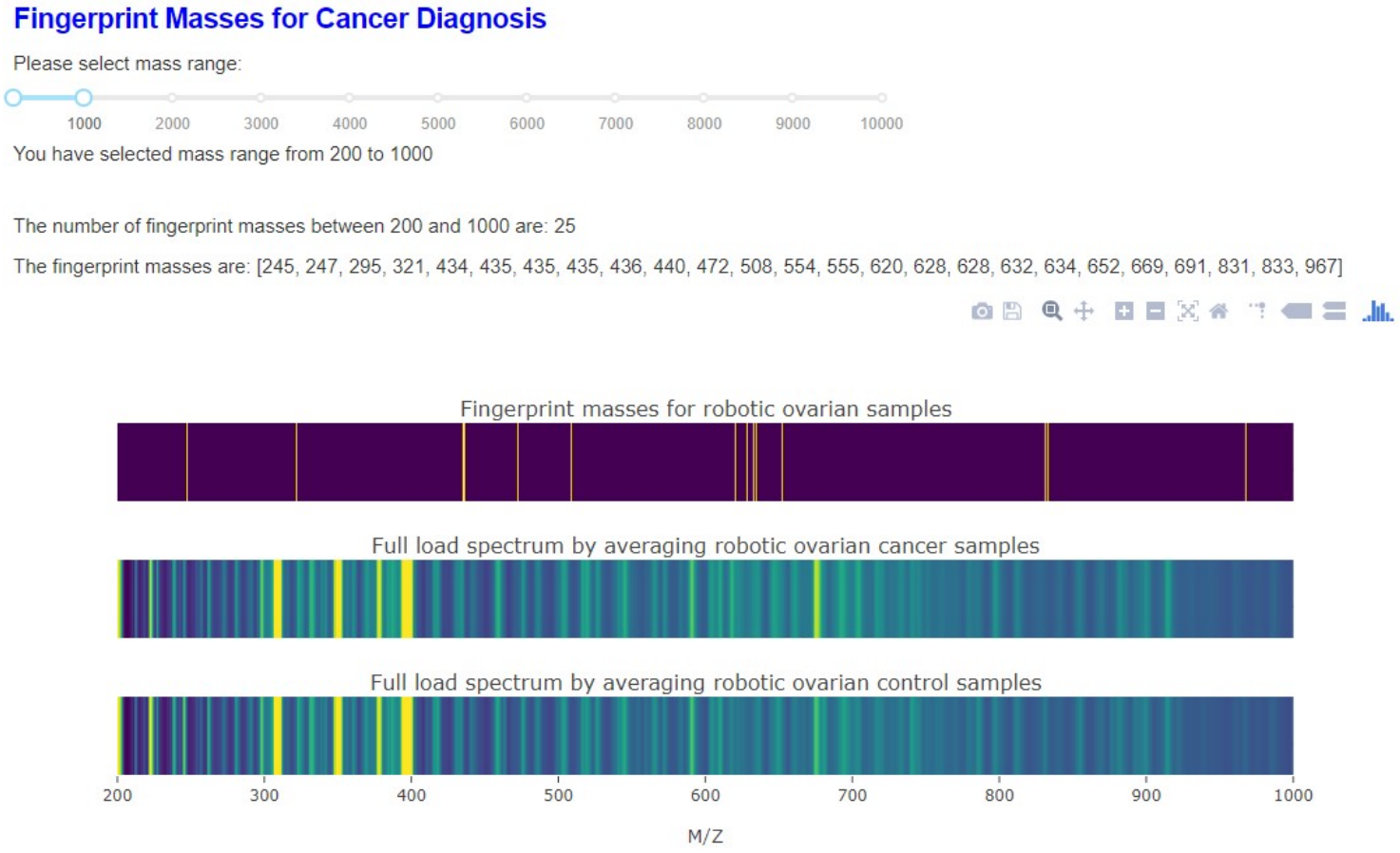
Prediction of Cancer/No Cancer in Specific Group

- If you choose specific group (Robotic prepared ovarian group herein), it shows the visualization of new sample within training samples in this group, and predict the probability of cancer/no cancer by four models



Fingerprint Masses in Specific Group

- It will also show the fingerprint masses within specific group (robotic prepared ovarian group herein), you can select the mass range to show interested fingerprint masses



Conclusion

- SVM were selected as the best model to predict ovarian and prostate cancers with high accuracy (95-100%), and 0% false negative rate, making it ideal to 'red flag' the suspected cancer samples
- One of the fingerprint molecules determining ovarian cancer was identified, which is confirmed by literature report
- A cancer diagnosis app was developed to offer quick cancer prediction results as well as lists of fingerprint molecules for cancer diagnosis

Recommendations

- Patients should ask for mass spectrum test during routine check up for cancer screening
- Doctors should recommend patients to do mass spectrum test during routine check up
- Insurance company should cover the mass spectrum test fee as preventative test to encourage people do routine cancer screening

Goal of Early Cancer Detection

√ SHOTGUN METHODS

Analyze routine check up samples, *e.g.* blood samples, and **collect as much information as possible** for cancer detection

√ 'RED FLAG' SUSPECTS

SVM were selected as the best model to predict ovarian and prostate cancers with high accuracy (95-100%), and 0% false negative rate

? Increase the number of training/testing samples

? Add more cancers

? To be added…