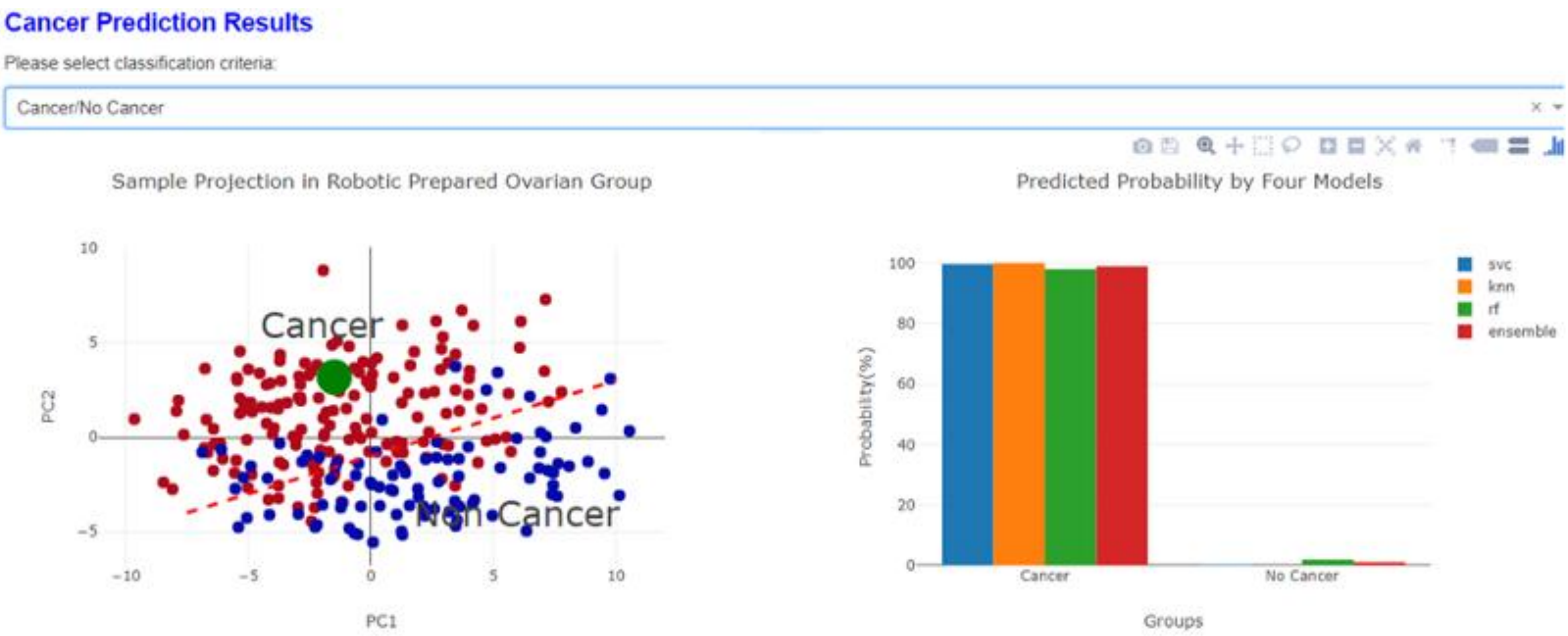


Cancer Prediction Results

Please select classification criteria:

Cancer/No Cancer

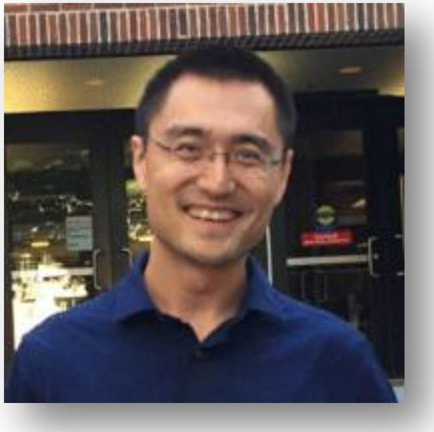
Early Cancer Detection using Data Science Tools

Peter Liu

Dallas AI

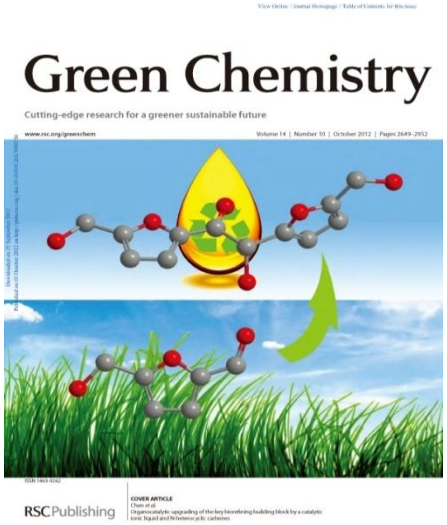
2018.7.26

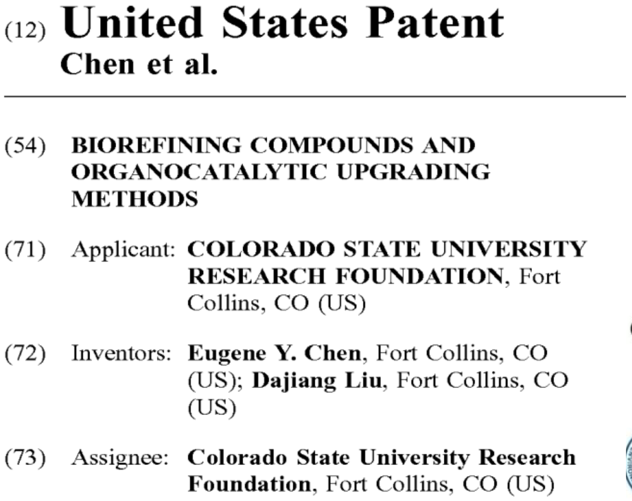
About me



Dr. Peter LIU
Energy Industry

Ph.D., Colorado State University
Scientist, Sandia National Labs


View Online / Journal Homepage / Table of Contents for this issue
Green Chemistry
Cutting-edge research for a greener sustainable future
www.rsc.org/greenchem
Volume 14 | Number 10 | October 2012 | Pages 2649–2952
ISSN 1463-9262
RSC Publishing
COVER ARTICLE
Chen et al.
Organocatalytic upgrading of the key biorefining building block by a catalytic ionic liquid and N-heterocyclic carbenes



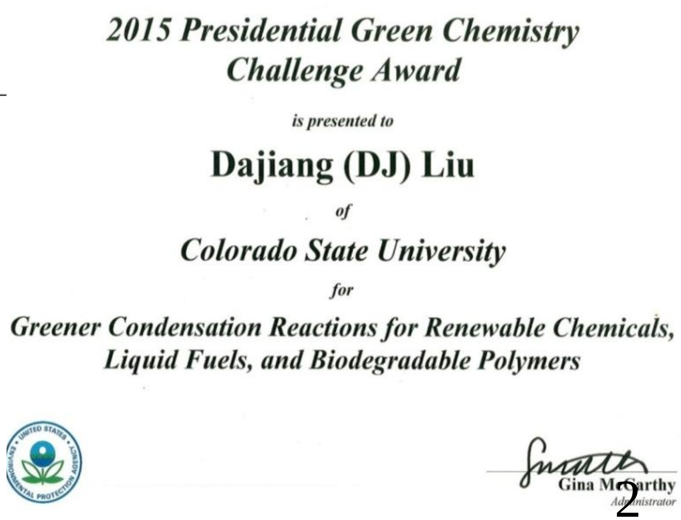
(12) **United States Patent**
Chen et al.

(54) **BIOREFINING COMPOUNDS AND ORGANOCATALYTIC UPGRADING METHODS**

(71) Applicant: **COLORADO STATE UNIVERSITY RESEARCH FOUNDATION**, Fort Collins, CO (US)

(72) Inventors: **Eugene Y. Chen**, Fort Collins, CO (US); **Dajiang Liu**, Fort Collins, CO (US)

(73) Assignee: **Colorado State University Research Foundation**, Fort Collins, CO (US)



2015 Presidential Green Chemistry Challenge Award

is presented to

Dajiang (DJ) Liu

of

Colorado State University

for

Greener Condensation Reactions for Renewable Chemicals, Liquid Fuels, and Biodegradable Polymers

Gina McCarthy
Administrator

Sections

- **Early cancer detection**

  "*How Data Science Enables Early Cancer Diagnosis*" on Medium

  - ML application to 'red-flag' suspected cancer patients

- **Machine learning product development**

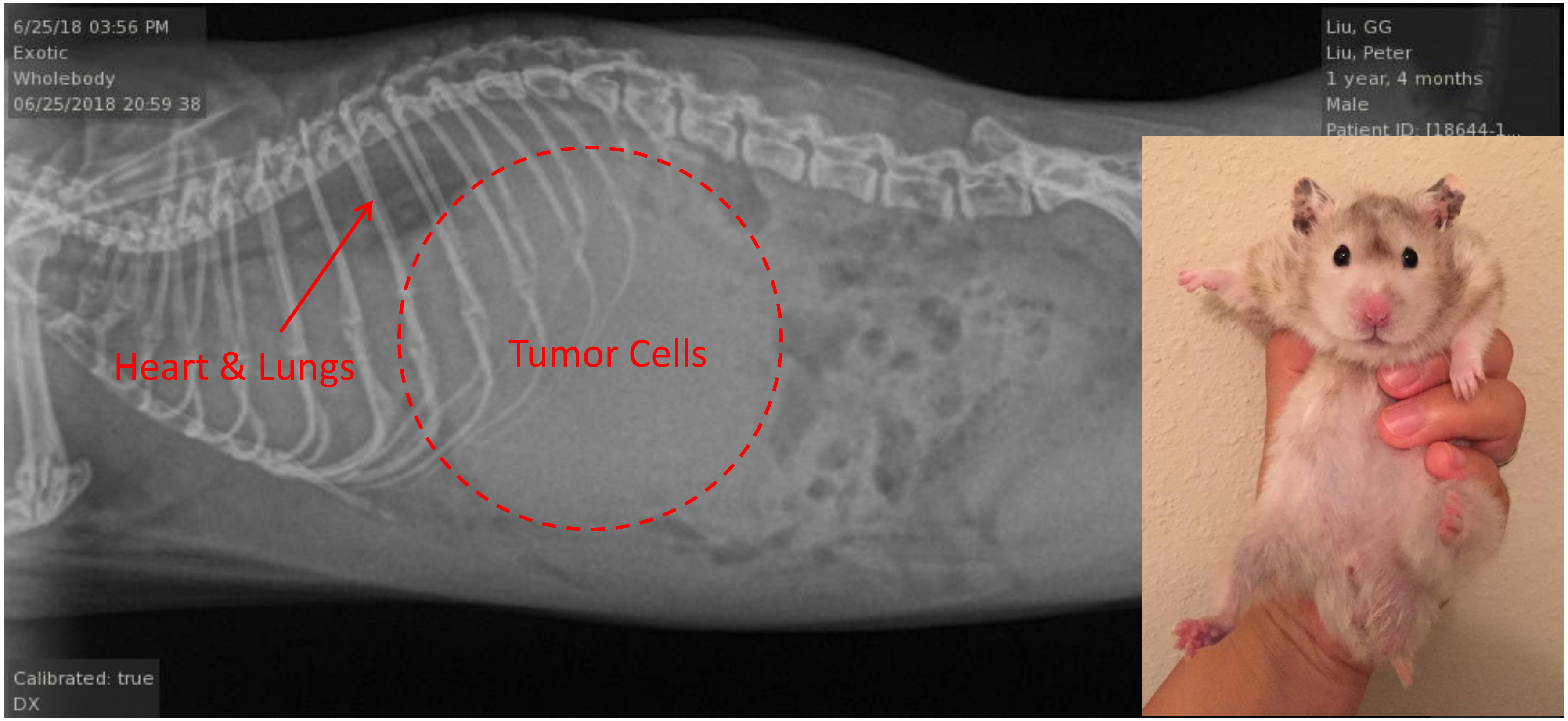
  - App development using Dash

- **Other interesting topic**

  "*Scan-and-Bingo Approach for Product Authentication*" on Medium

  - ML application to authenticate our food products

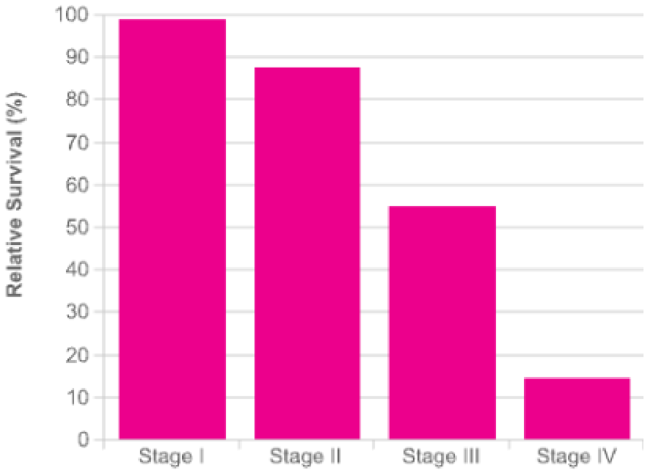
My Pet Hamster Passed Away Due to Cancer



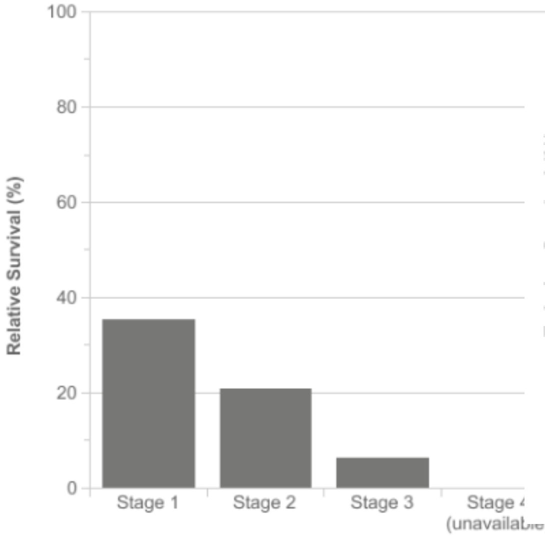
Top 3 Cancers in United States

Five Years' Survival Rates at Different Stages

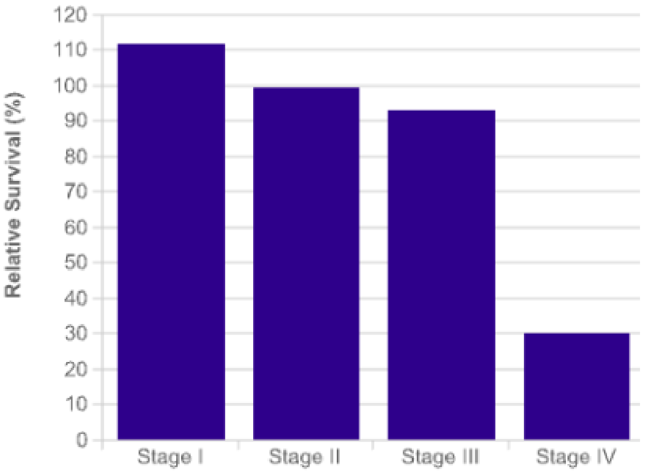
Breast Cancer



Lung Cancer



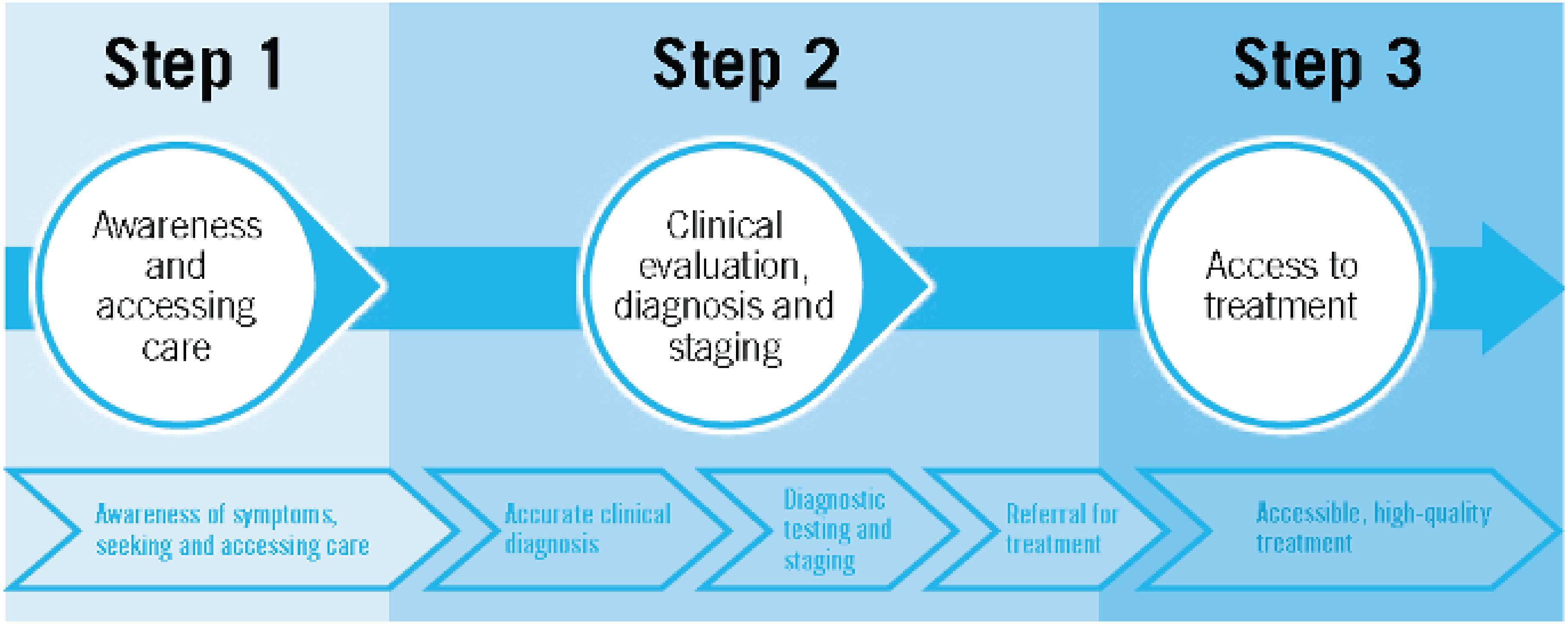
Prostate Cancer



Early Detection/Treatment = Higher Survival Rate

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

Typical Procedures for Cancer Diagnosis



Step 1

Awareness and accessing care

Step 2

Clinical evaluation, diagnosis and staging

Step 3

Access to treatment

Awareness of symptoms, seeking and accessing care

Accurate clinical diagnosis

Diagnostic testing and staging

Referral for treatment

Accessible, high-quality treatment

Problems:

- Some cancer does not have symptoms until late stage
- Difficult to link symptoms with cancer

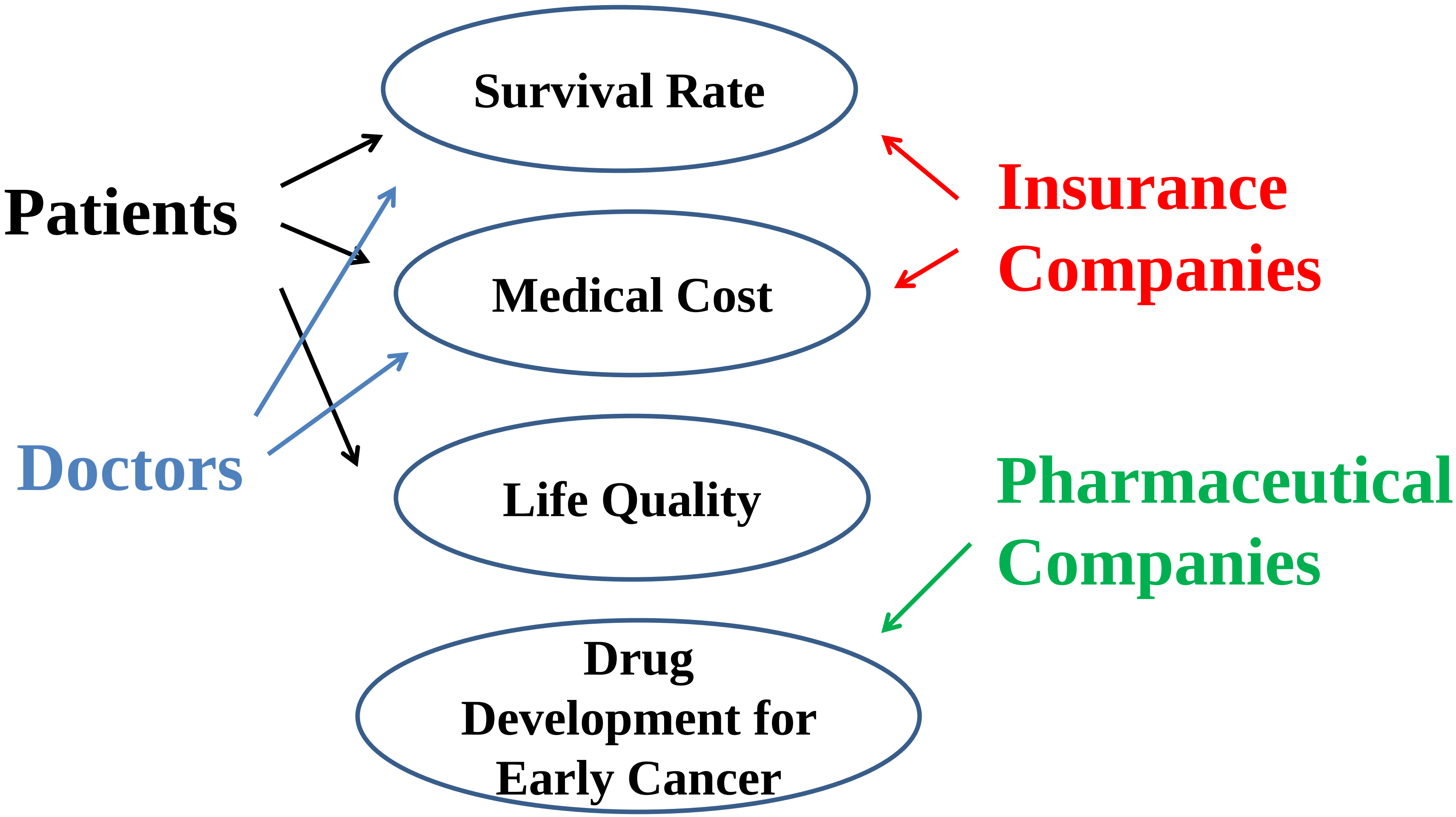
National Cancer Institute
https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis

Early Detection of Cancer

Goal

- "Red Flag" the suspected cancer samples using routine checkup samples(*e.g.*, blood, serum)
- App development to aid cancer screening

Who Cares?


Patients

Doctors

Survival Rate

Medical Cost

Life Quality

Drug Development for Early Cancer

Insurance Companies

Pharmaceutical Companies

Approach

- **Shotgun Method**

  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

Shotgun Method

- **Mass Spectrometry**

- Collect mass information of all chemicals

- **Low Sample Loading**

- Milligram (1/1000 grams) samples

- **High Sensitivity**

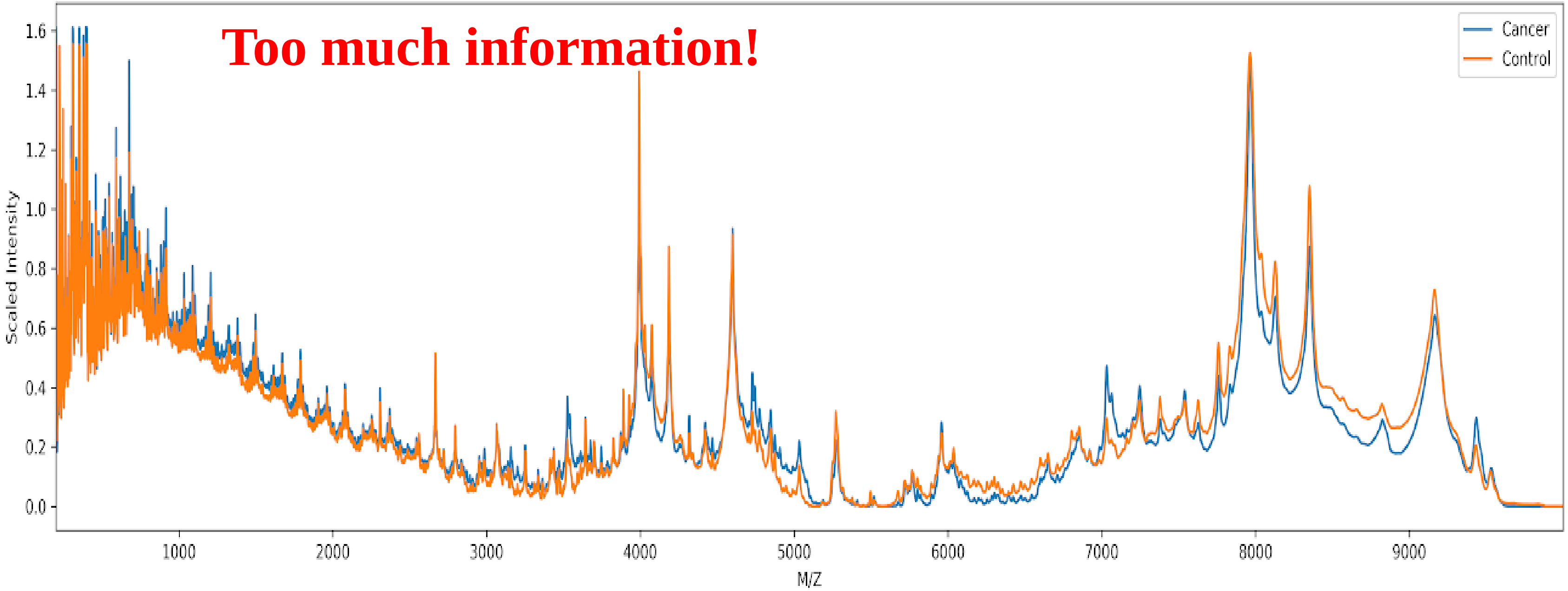
- Detect trace amount of chemicals at parts per billion(ppb) level

- **High Throughput Screening**

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

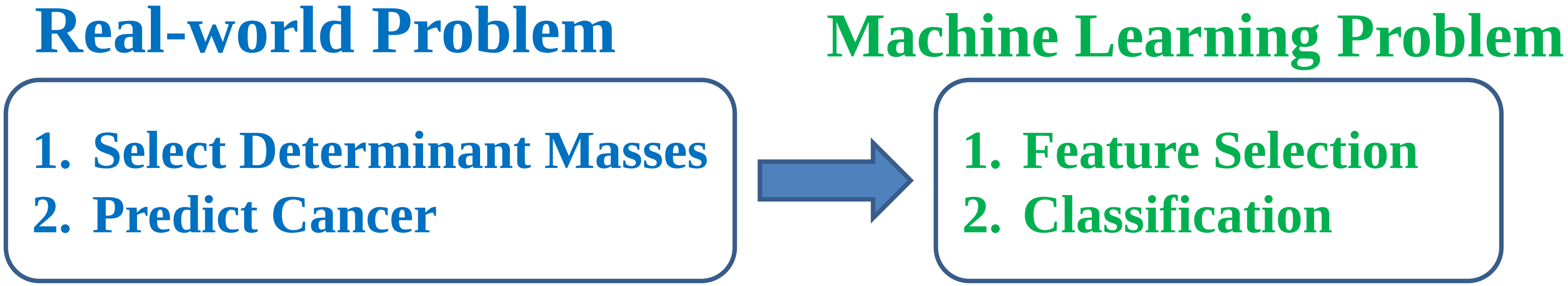
Problems with Shotgun Method

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



Too much information!

Real-world Problem to ML Problem



Real-world Problem

1. Select Determinant Masses
2. Predict Cancer

Machine Learning Problem

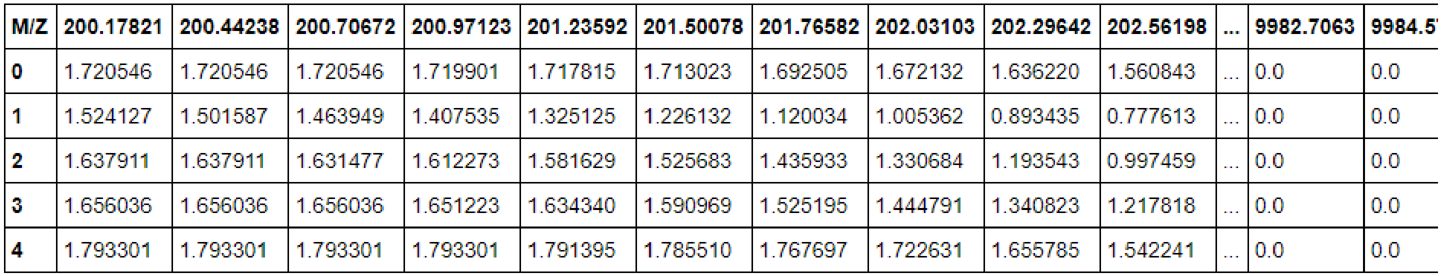
1. Feature Selection
2. Classification

Determinant Masses Selection vs. Important Features Selection

Cancer/No Cancer

vs.

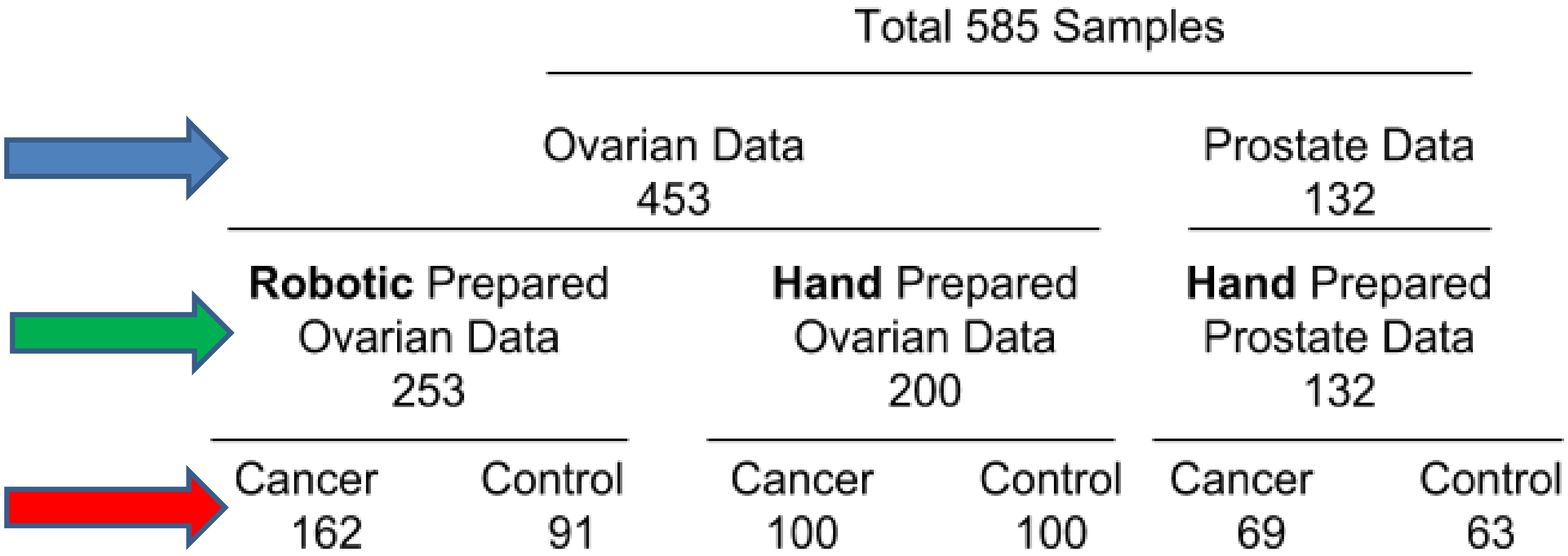
1/-1



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



Total 585 Samples

Ovarian Data 453

Prostate Data 132

Robotic Prepared Ovarian Data 253

Hand Prepared Ovarian Data 200

Hand Prepared Prostate Data 132

Cancer 162

Control 91

Cancer 100

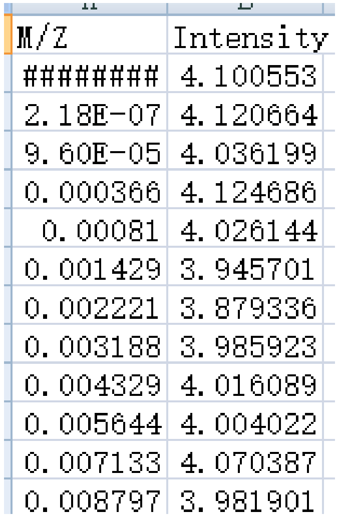
Control 100

Cancer 69

Control 63

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

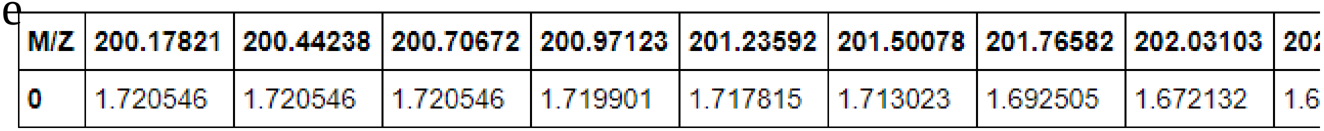
Mass cutoff

Scaling

Transpose

“Wide” single MS data

Mass values



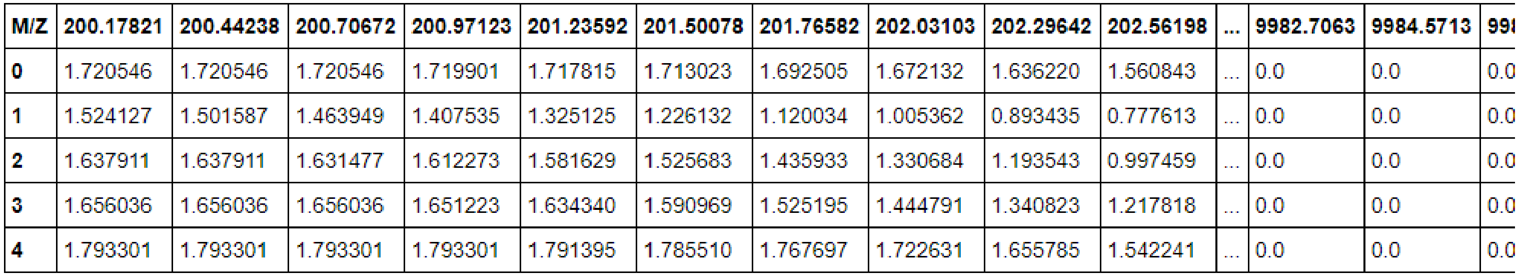
M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202
0	1.720546	1.720546	1.720546	1.719901	1.717815	1.713023	1.692505	1.672132	1.6

Concatenation by Row

Mass Spectra Data Matrix

Mass values

Samples

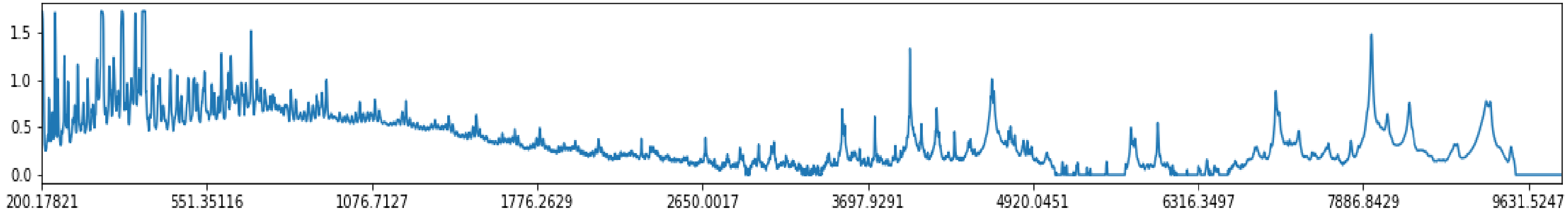


M/Z	200.17821	200.44238	200.70672	200.97123	201.23592	201.50078	201.76582	202.03103	202.29642	202.56198	...	9982.7063	9984.5713	99
0	1.720546	1.720546	1.720546	1.719901	1.717815	1.713023	1.692505	1.672132	1.636220	1.560843	...	0.0	0.0	0.0
1	1.524127	1.501587	1.463949	1.407535	1.325125	1.226132	1.120034	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

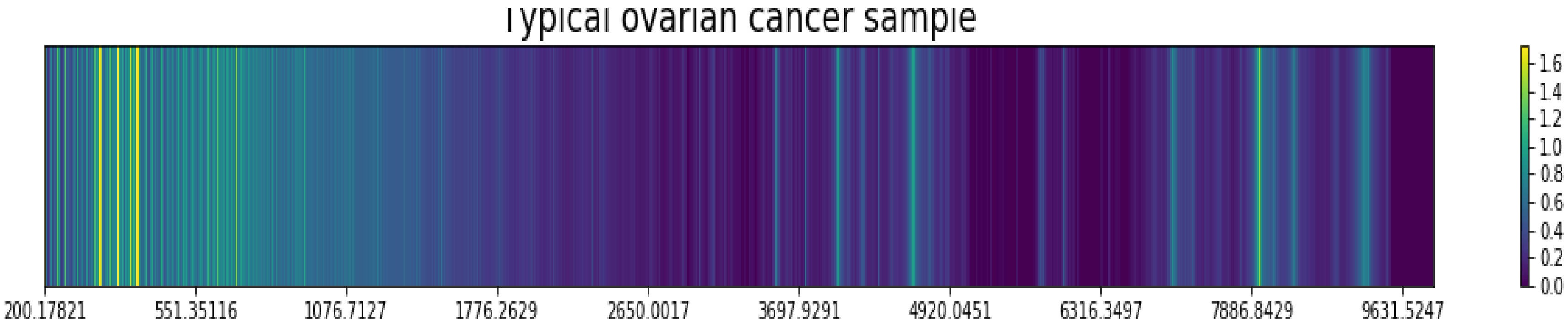
Exploratory Data Analysis

Heatmap is more preferable than plot view

Plot view of data



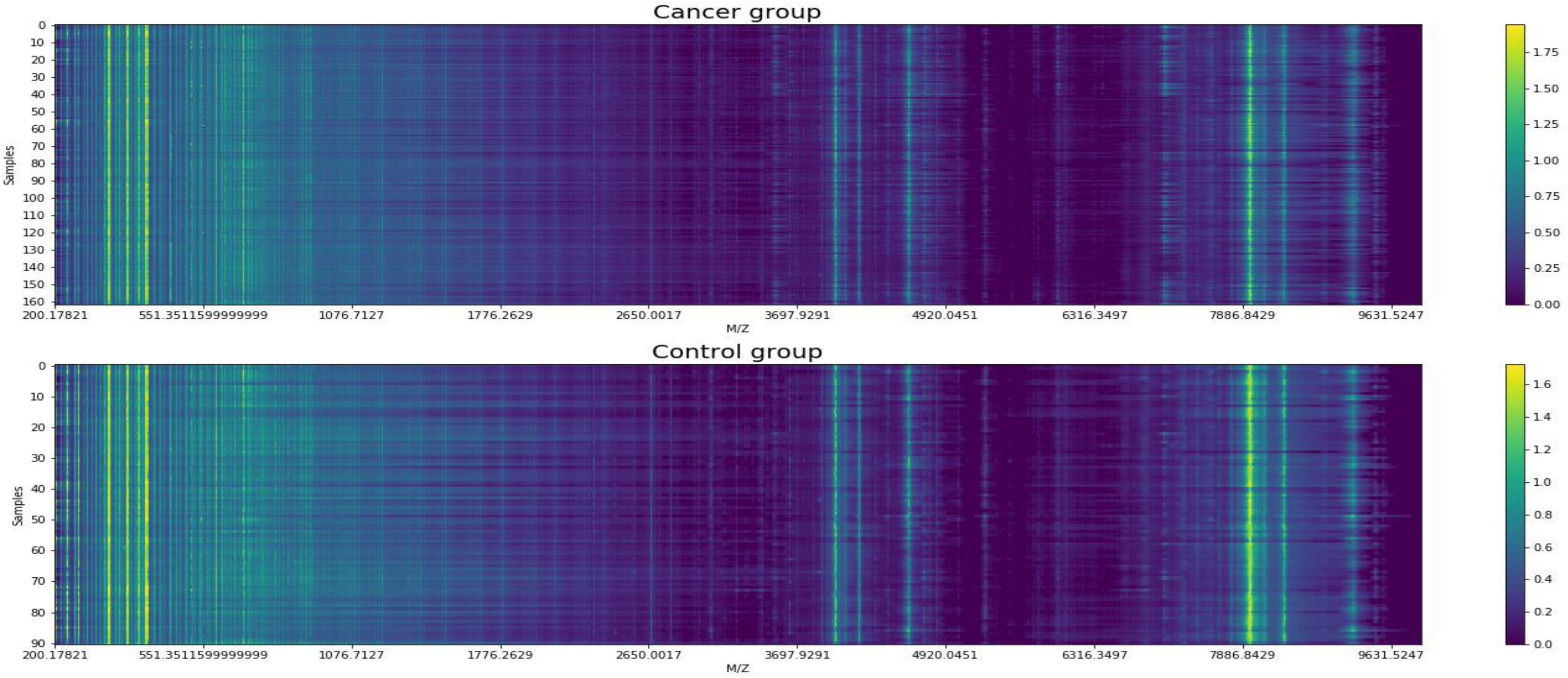
1D Heatmap



Exploratory Data Analysis

Heatmap of robotic prepared ovarian datasets

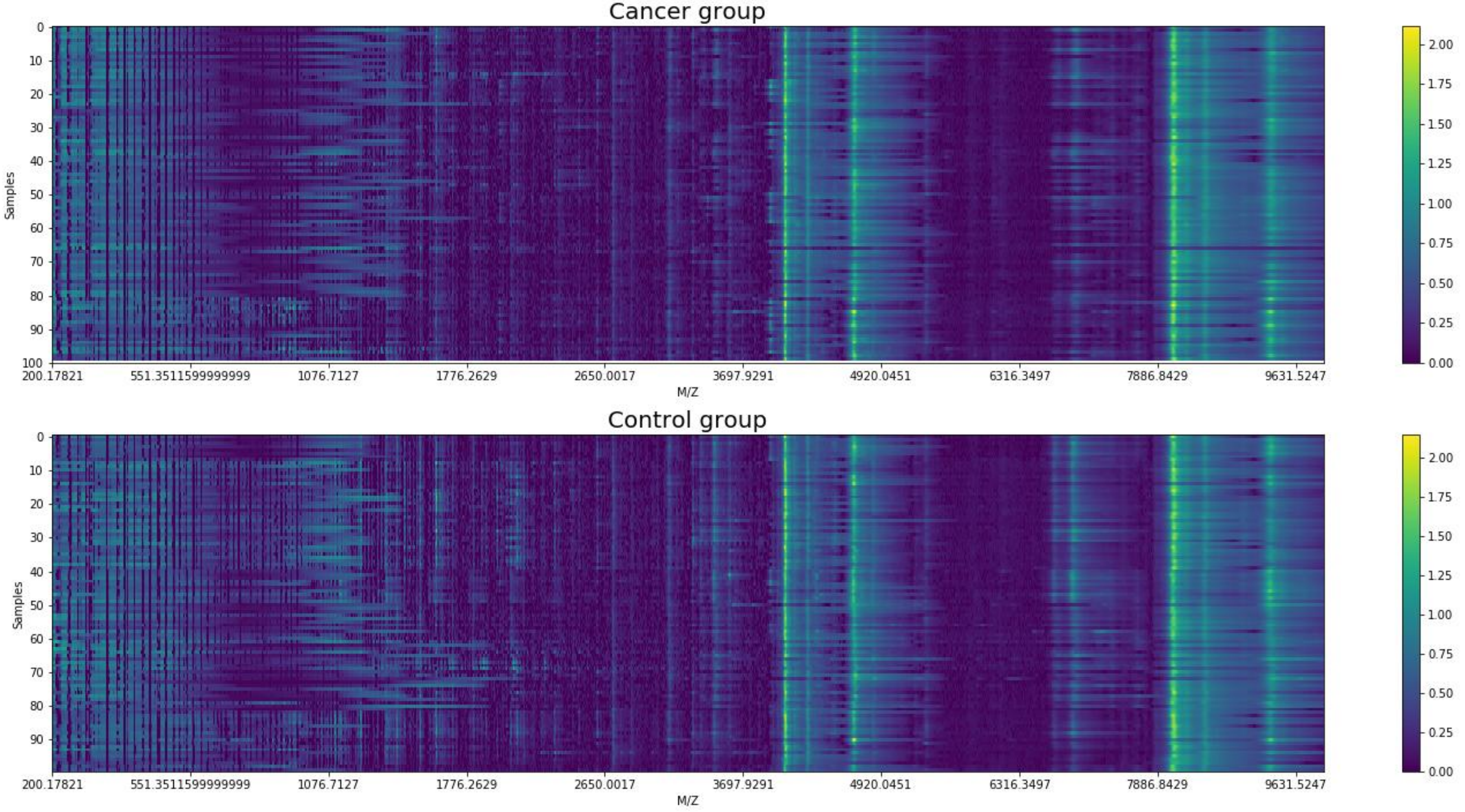
- Difficult to tell the difference between cancer and control group



Exploratory Data Analysis

Heatmap of hand prepared ovarian datasets

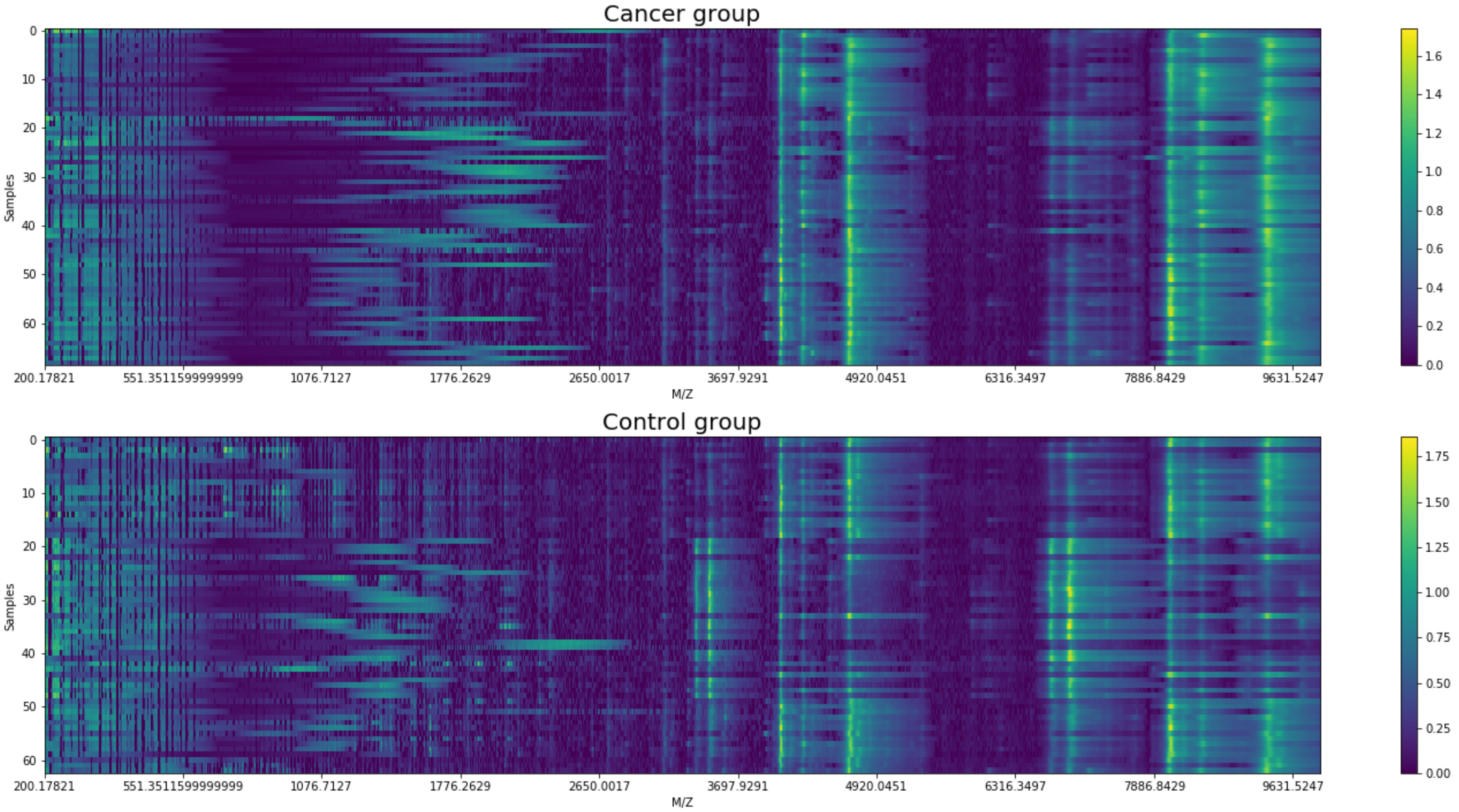
- Difficult to tell the difference between cancer and control group



Exploratory Data Analysis

Heatmap of hand prepared prostate samples

- Difficult to tell the difference between cancer and control group



Cancer group

Control group

Data Visualization based on PCA

Easy to tell the difference!

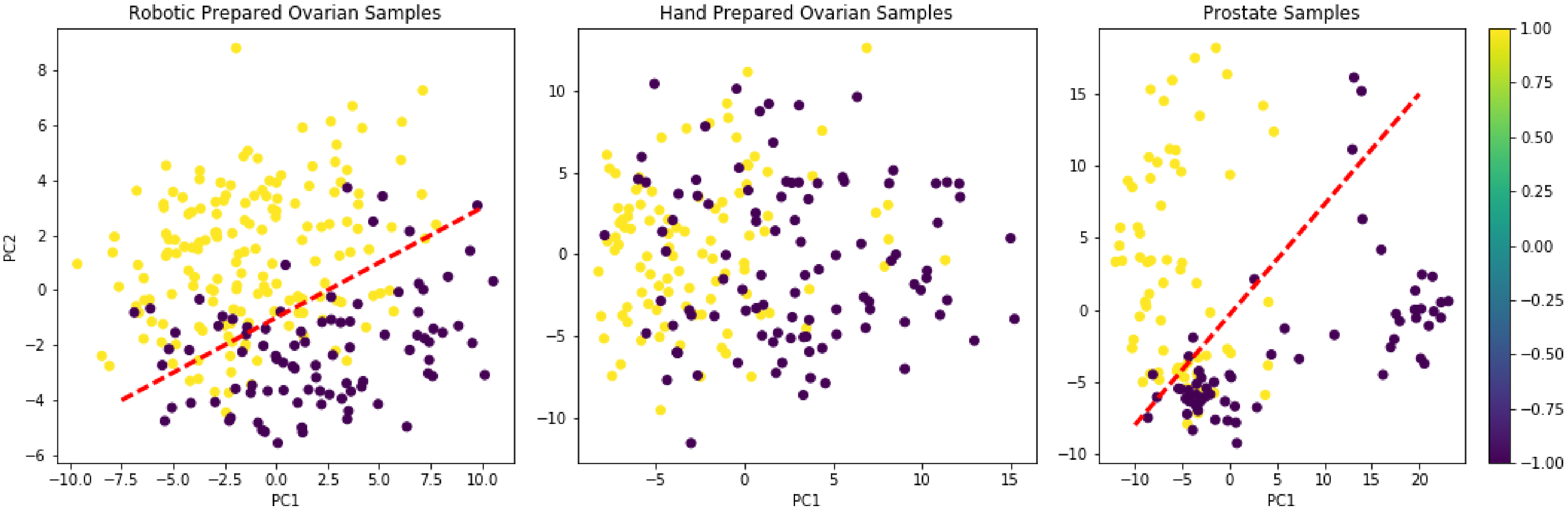


Figure. Comparison of cancer and non-cancer group in three datasets. **Purple** plots represent **non-cancer** group, while **yellow** plots represent **cancer** group

Feature Selection by Random Forest

Important Features = Fingerprint Mass

Less than 1% features were needed

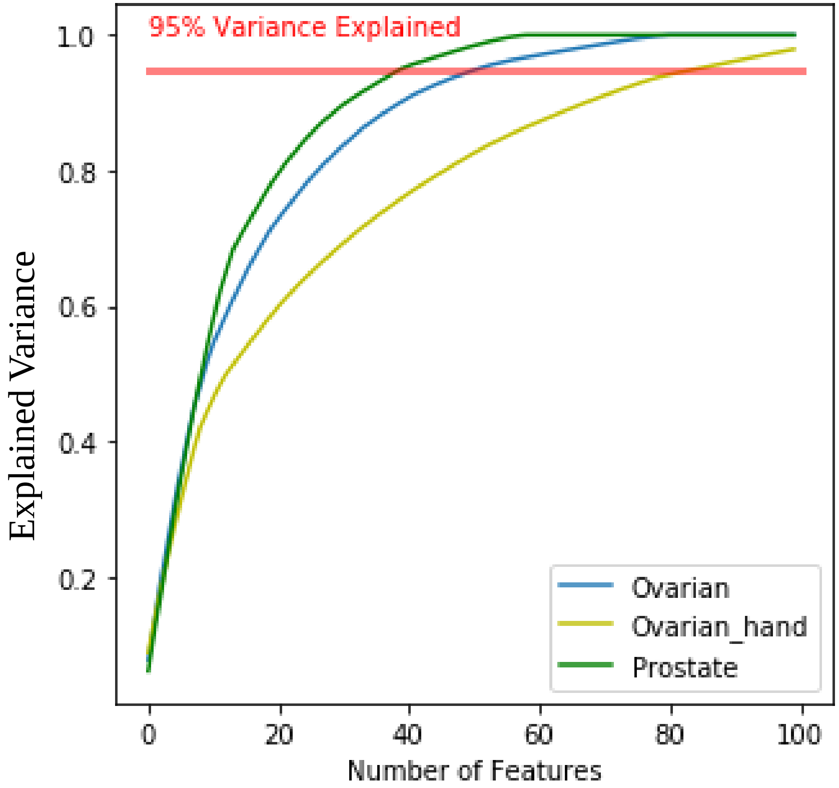


Figure. Explained Variance vs. Number of Features rendered by **Random Forest.** Decision Tree is a natural way of feature selection

Feature Selection

Feature Selection is not only for Modeling!

- Fingerprint masses inspire drug development
- One key molecule (molecular weight 472) to determine ovarian cancer is in our list of fingerprint masses

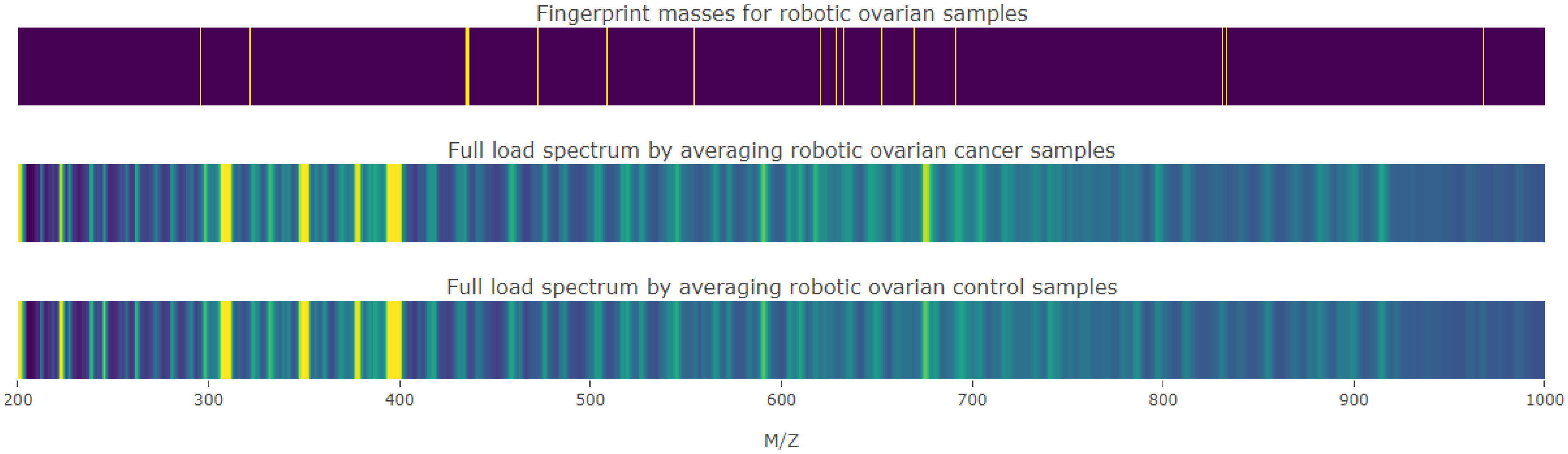


Figure. Selected fingerprint masses for robotic prepared ovarian samples

Reference
https://academic.oup.com/ajcp/article/134/6/903/1760577

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	0.92
	AUC	0.93	0.91	0.94	0.92
	F1-Score	0.94	0.93	0.96	0.93
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 detection of cancer?

- Lower false negatives

Model 1

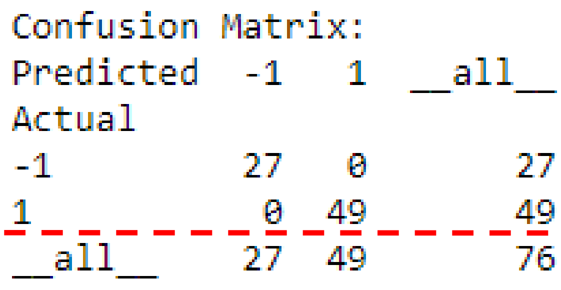
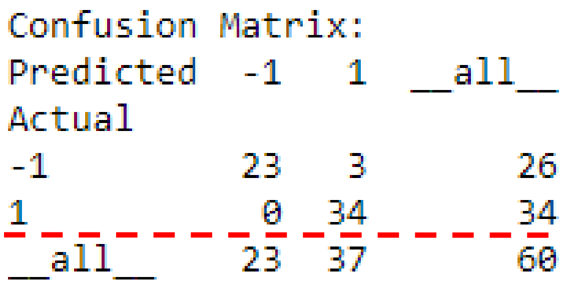
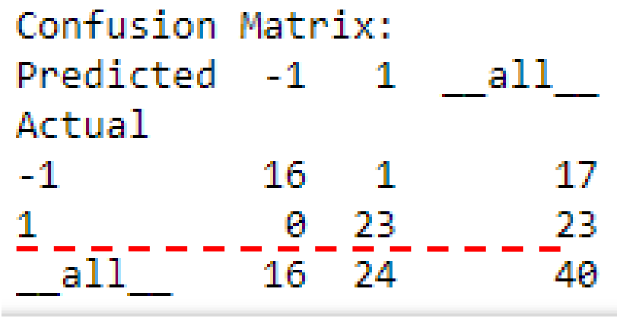
Predicted Actual	-1 (No Cancer)	1 (Cancer)	Total
-1 (No Cancer)	23	3 (FP)	26
1 (Cancer)	0 (FN)	34	34
Total	23	37	60

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

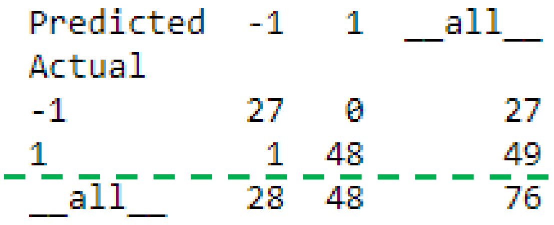
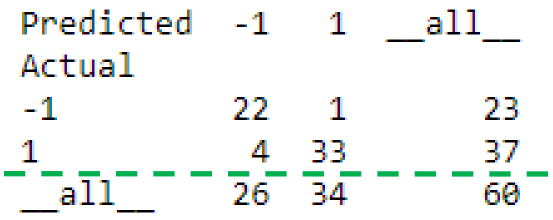
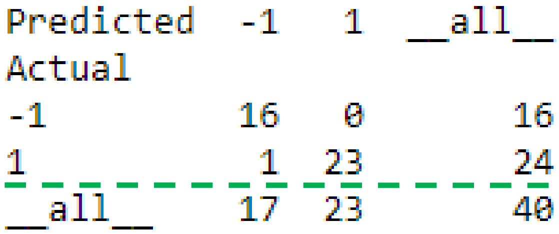
Robotic prepared Ovarian Samples	Hand prepared Ovarian Samples	Prostate Samples
		

```
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
```

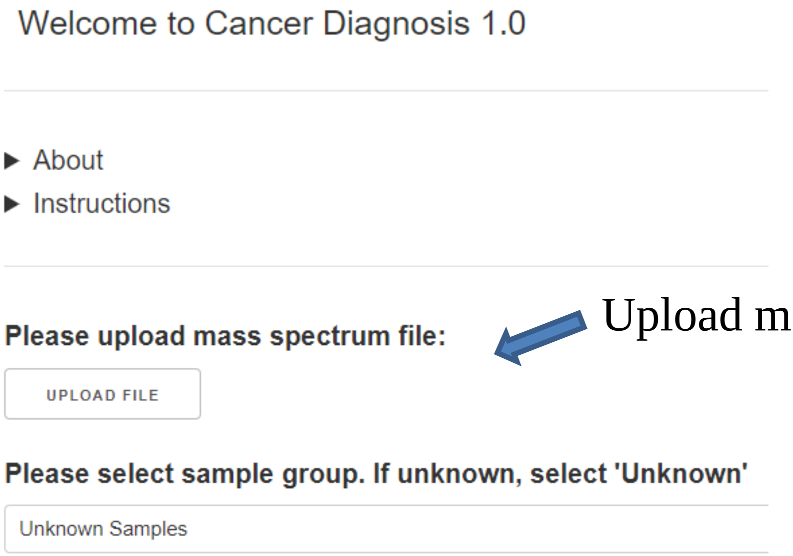
- SVM is our best model
- 0% of FN rate (fail to detect cancer) by SVM

Product Development

Cancer Diagnosis 1.0

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

Upload file



Welcome to Cancer Diagnosis 1.0

▶ About

▶ Instructions

Please upload mass spectrum file:

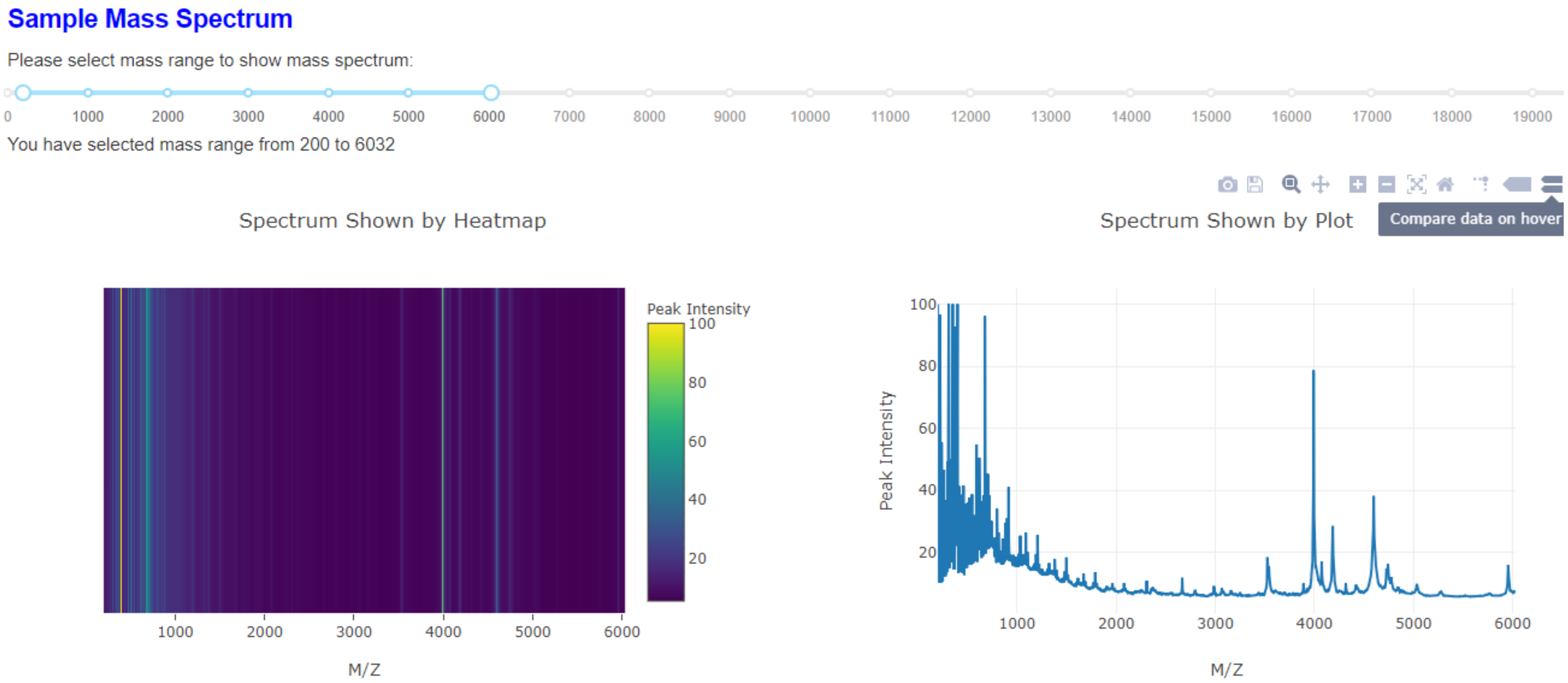
Please select sample group. If unknown, select 'Unknown'

Unknown Samples

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



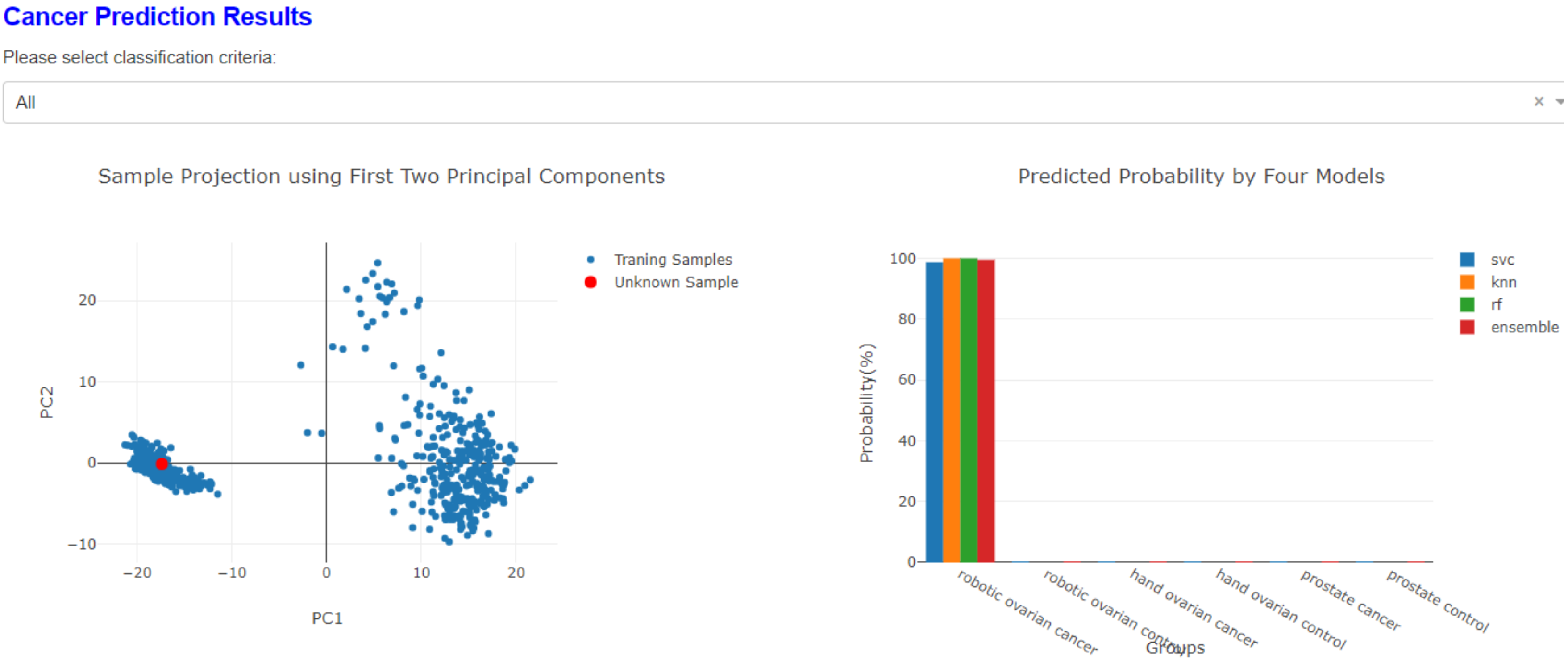
Sample Mass Spectrum

Please select mass range to show mass spectrum:

You have selected mass range from 200 to 6032

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



Cancer Prediction Results

Please select classification criteria:

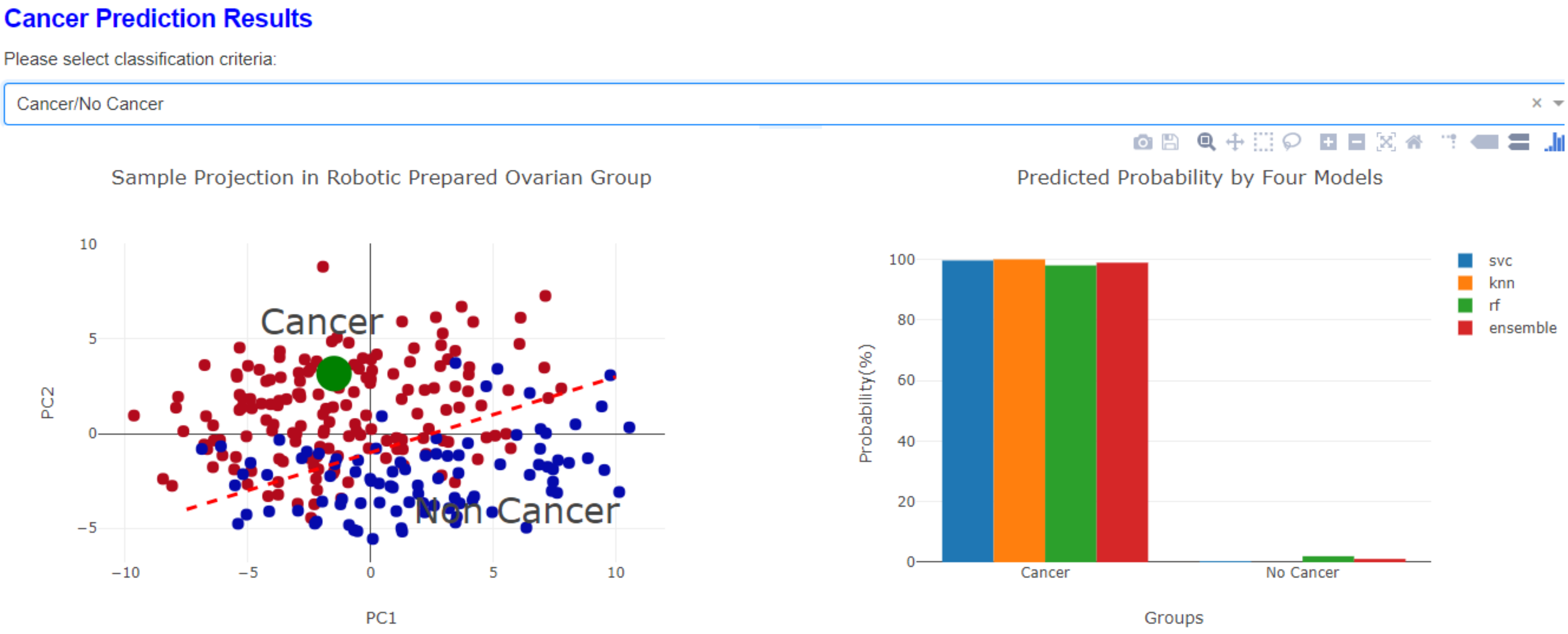
All

Sample Projection using First Two Principal Components

Predicted Probability by Four Models

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



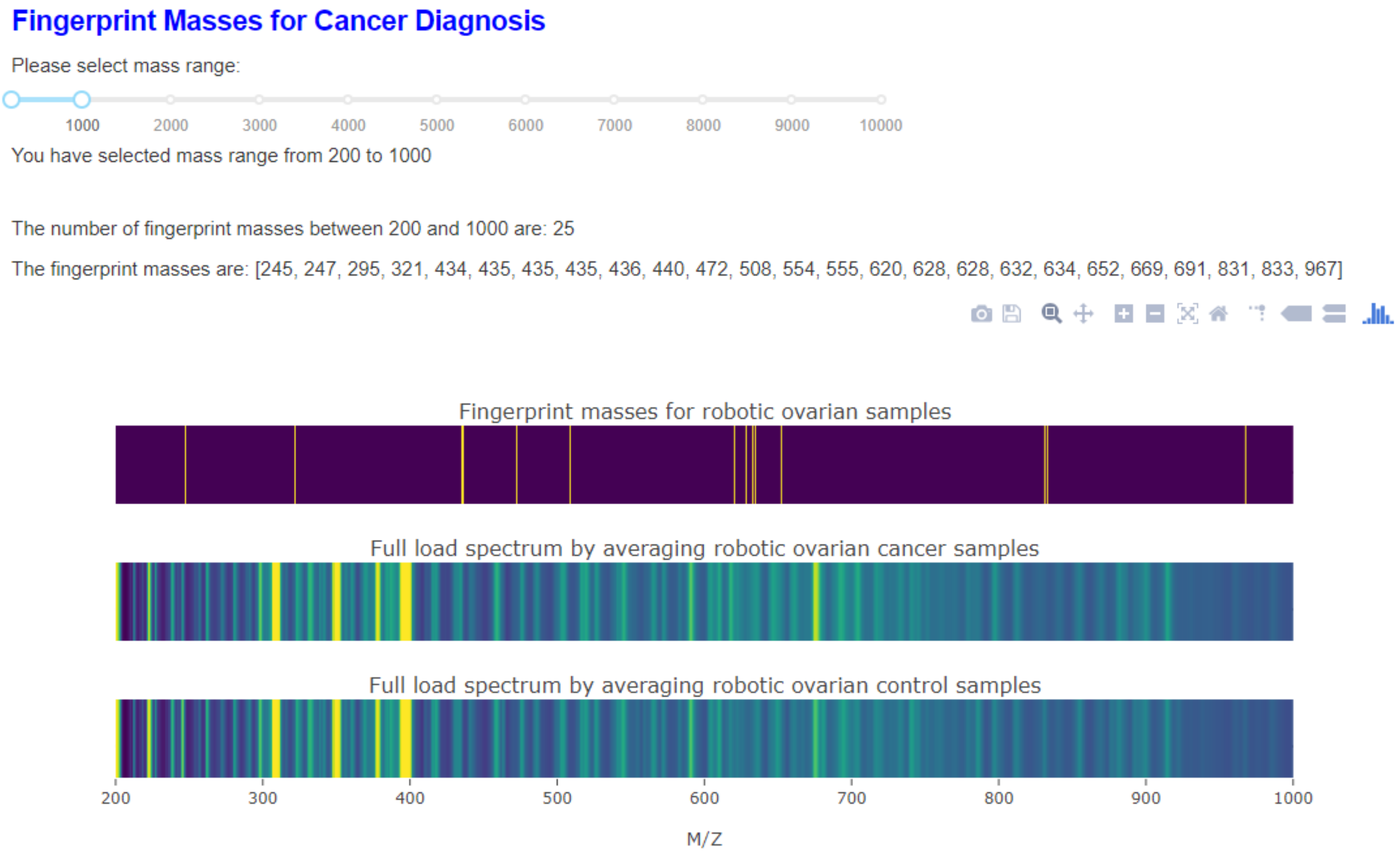
Cancer Prediction Results

Please select classification criteria:

Cancer/No Cancer

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



Fingerprint Masses for Cancer Diagnosis

Please select mass range:

1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

You have selected mass range from 200 to 1000

The number of fingerprint masses between 200 and 1000 are: 25

The fingerprint masses are: [245, 247, 295, 321, 434, 435, 435, 435, 436, 440, 472, 508, 554, 555, 620, 628, 628, 632, 634, 652, 669, 691, 831, 833, 967]

Fingerprint masses for robotic ovarian samples
Full load spectrum by averaging robotic ovarian cancer samples
Full load spectrum by averaging robotic ovarian control samples
200
300
400
500
600
700
800
900
1000
M/Z

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

√ "Red Flag" the suspected cancer samples using routine checkup samples(*e.g.*, blood, serum)

√ App development to aid cancer screening

× Increase the number of training/testing samples

× Add more cancers

Look into the future

Cancer diagnosis as easy as blood sugar test!

Product Authentication

“***Scan-and-Bingo Approach for Product Authentication***” on Medium
- ML application to authentic your food and cosmetic products

Goal

- Develop a more reliable and efficient way to “**Red Flag**” those suspected counterfeits in the first place, e.g., cosmetic products, food, fuels

All data are publicly available through Quadram Institute

Approach

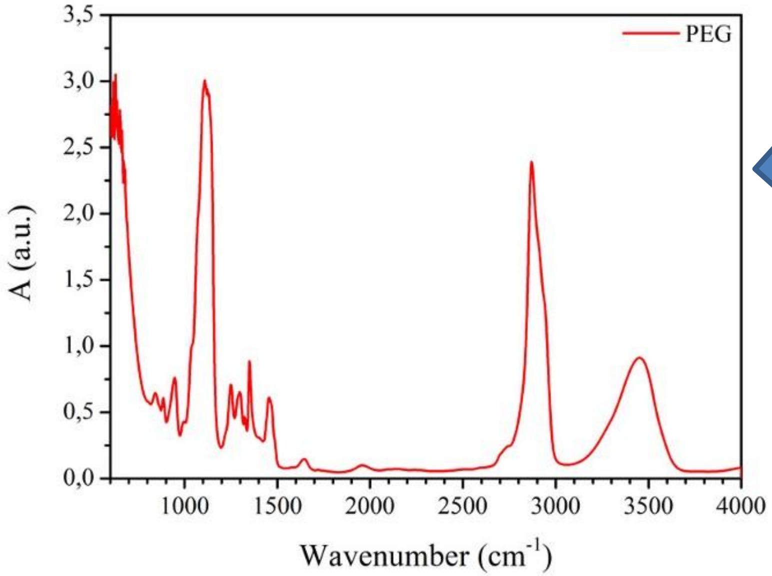
- Shotgun Method

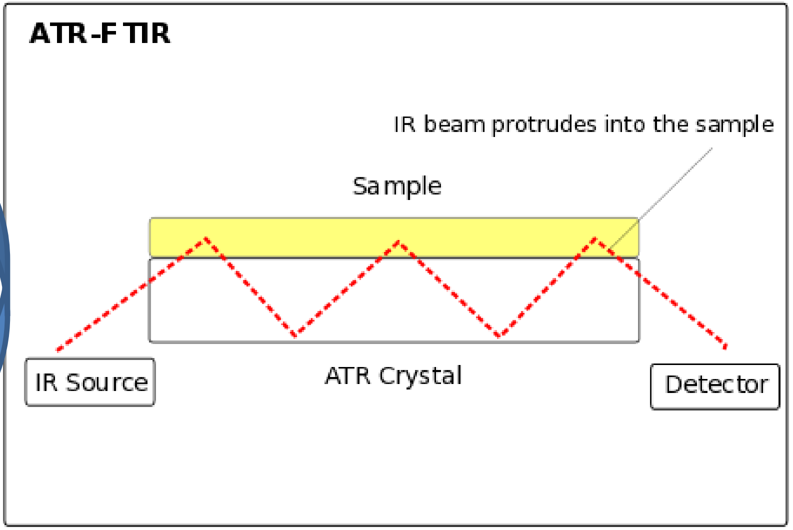
  Scan sample using hand-held FTIR, and **collect chemical information**

- “Red Flag” Suspects

  Identify whether the product is counterfeit

Fourier Transform Infrared Spectrometry (FTIR)





ATR-FTIR

IR beam protrudes into the sample

Sample

IR Source

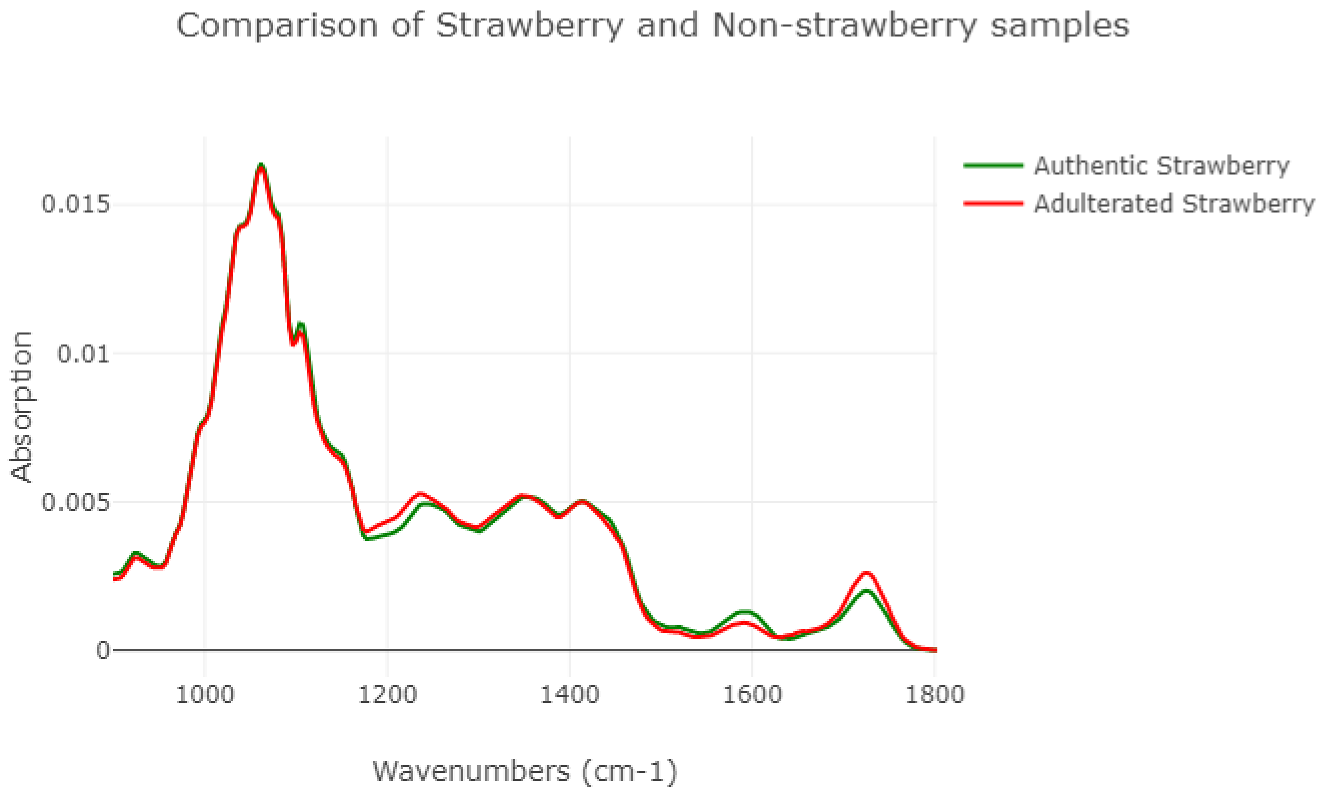
ATR Crystal

Detector

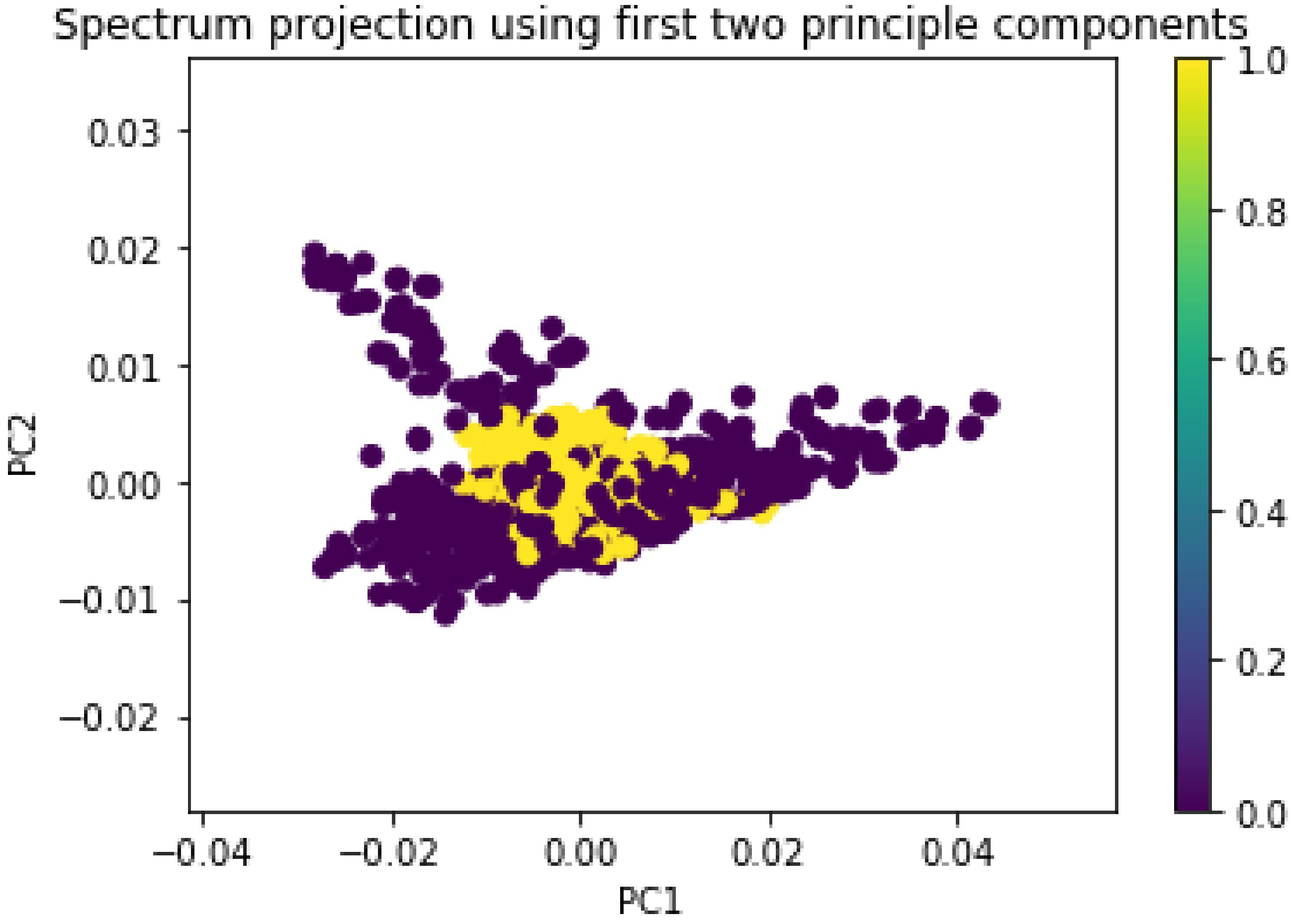
Authentic Or Not?

Strawberry Jam FTIR

- Total 983 samples: 351 authentic strawberry samples, 632 non-strawberry samples



Spectra Projection based on PCA

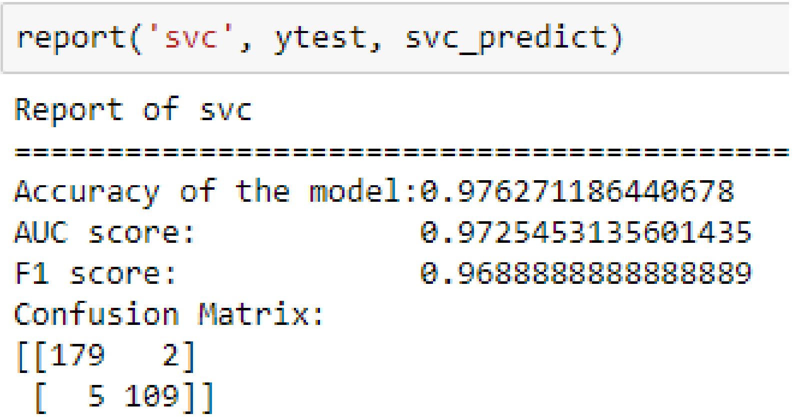


Authentic Strawberries

Counterfeits

Model Performance

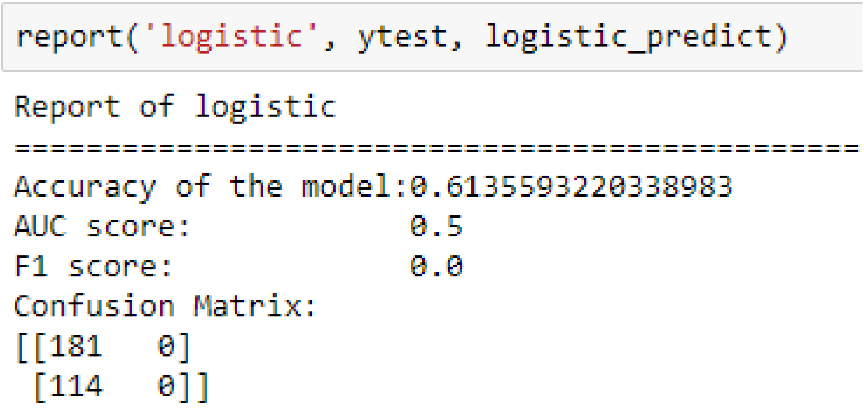
Supported Vector Classifier



```
report('svc', ytest, svc_predict)
```

```
Report of svc
==========================================
Accuracy of the model:0.976271186440678
AUC score:             0.9725453135601435
F1 score:              0.968888888888889
Confusion Matrix:
[[179   2]
 [  5 109]]
```

Logistic Regression (Problematic!)

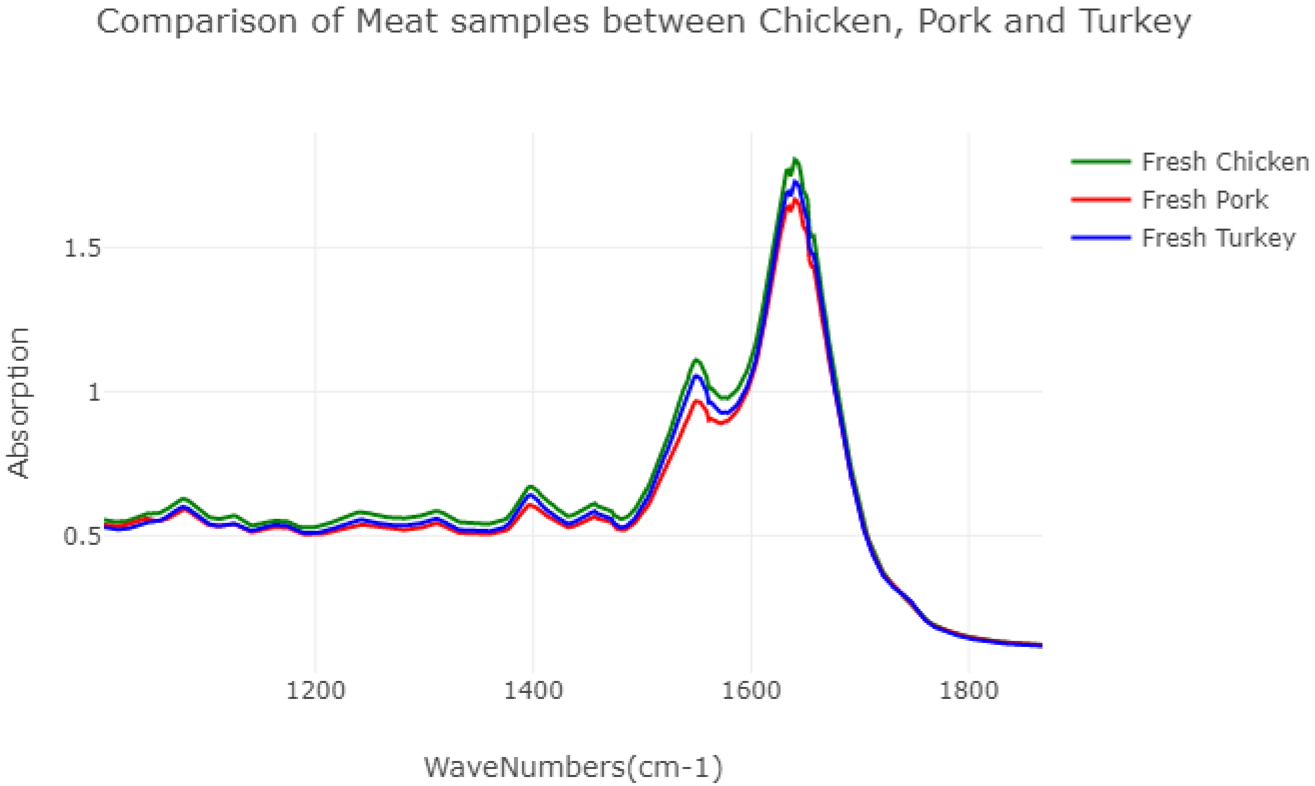


```
report('logistic', ytest, logistic_predict)
```

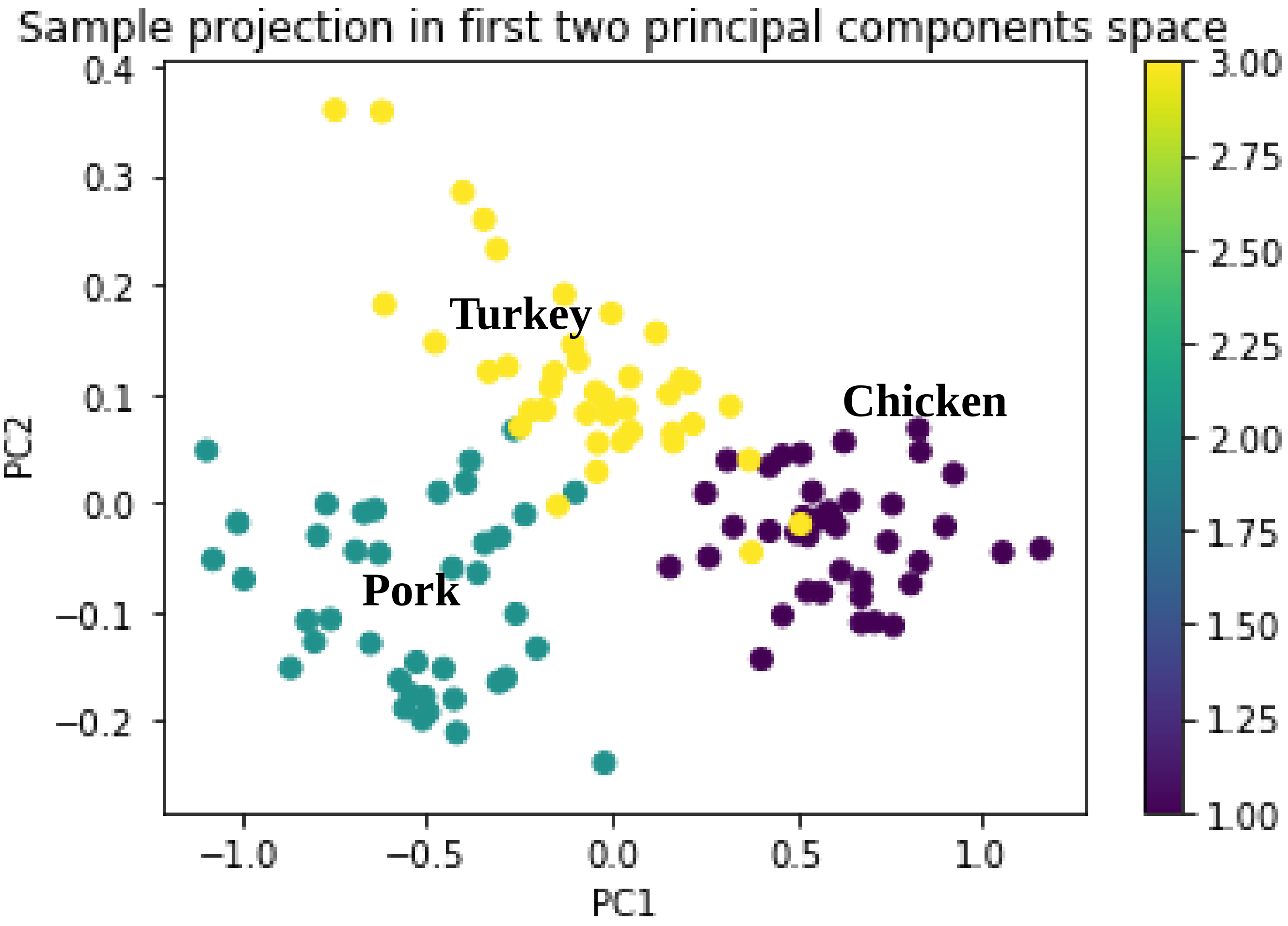
```
Report of logistic
===============================================
Accuracy of the model:0.6135593220338983
AUC score:            0.5
F1 score:             0.0
Confusion Matrix:
[[181   0]
 [114   0]]
```

Fresh Meat FTIR

- 120 meat samples: 40 chicken samples, 40 pork samples and 40 turkey samples

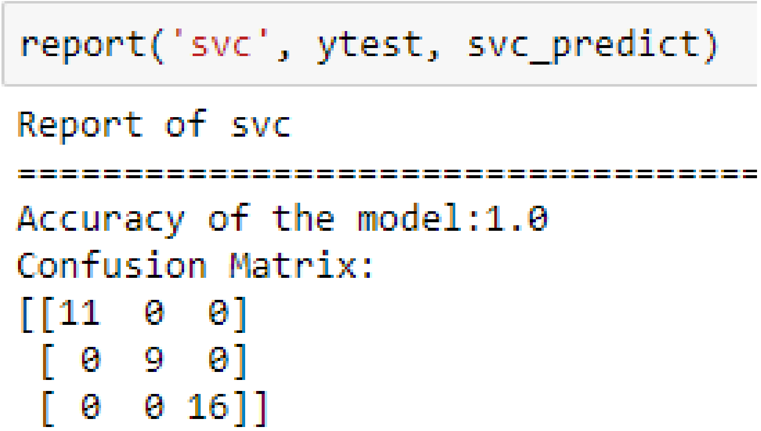


Spectra Projection based on PCA



Model Performance

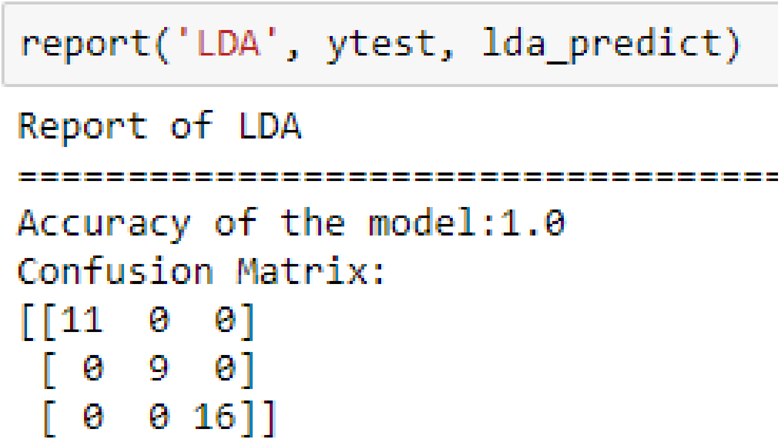
Supported Vector Classifier (SVC)



```
report('svc', ytest, svc_predict)
```

```
Report of svc
===================================
Accuracy of the model:1.0
Confusion Matrix:
[[11  0  0]
 [ 0  9  0]
 [ 0  0 16]]
```

Linear Discriminant Analysis (LDA)

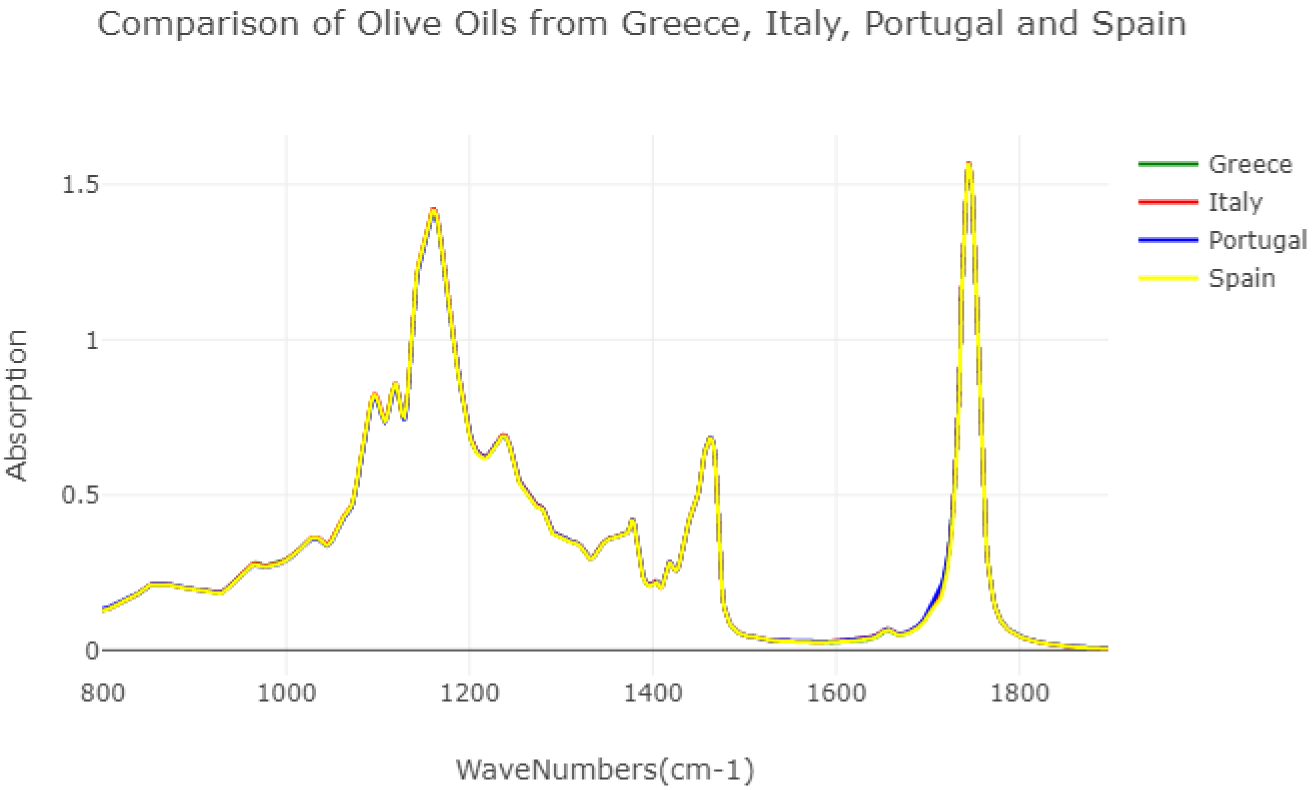


```
report('LDA', ytest, lda_predict)
```

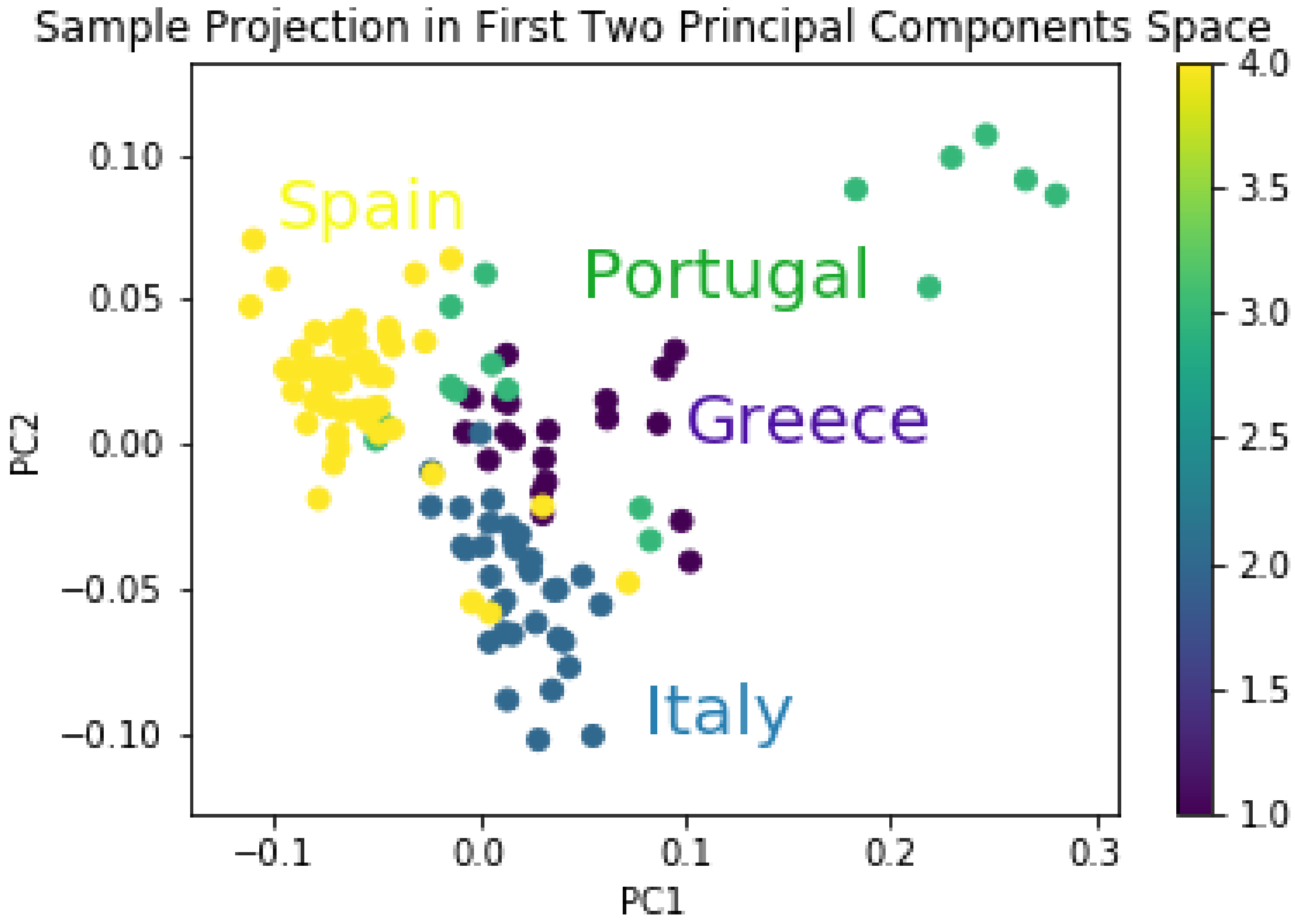
```
Report of LDA
===================================
Accuracy of the model:1.0
Confusion Matrix:
[[11  0  0]
 [ 0  9  0]
 [ 0  0 16]]
```

Origin of Olive Oil

- 120 olive oil samples: 20 from Greece, 34 from Italy, 16 from Portugal and 50 from Spain

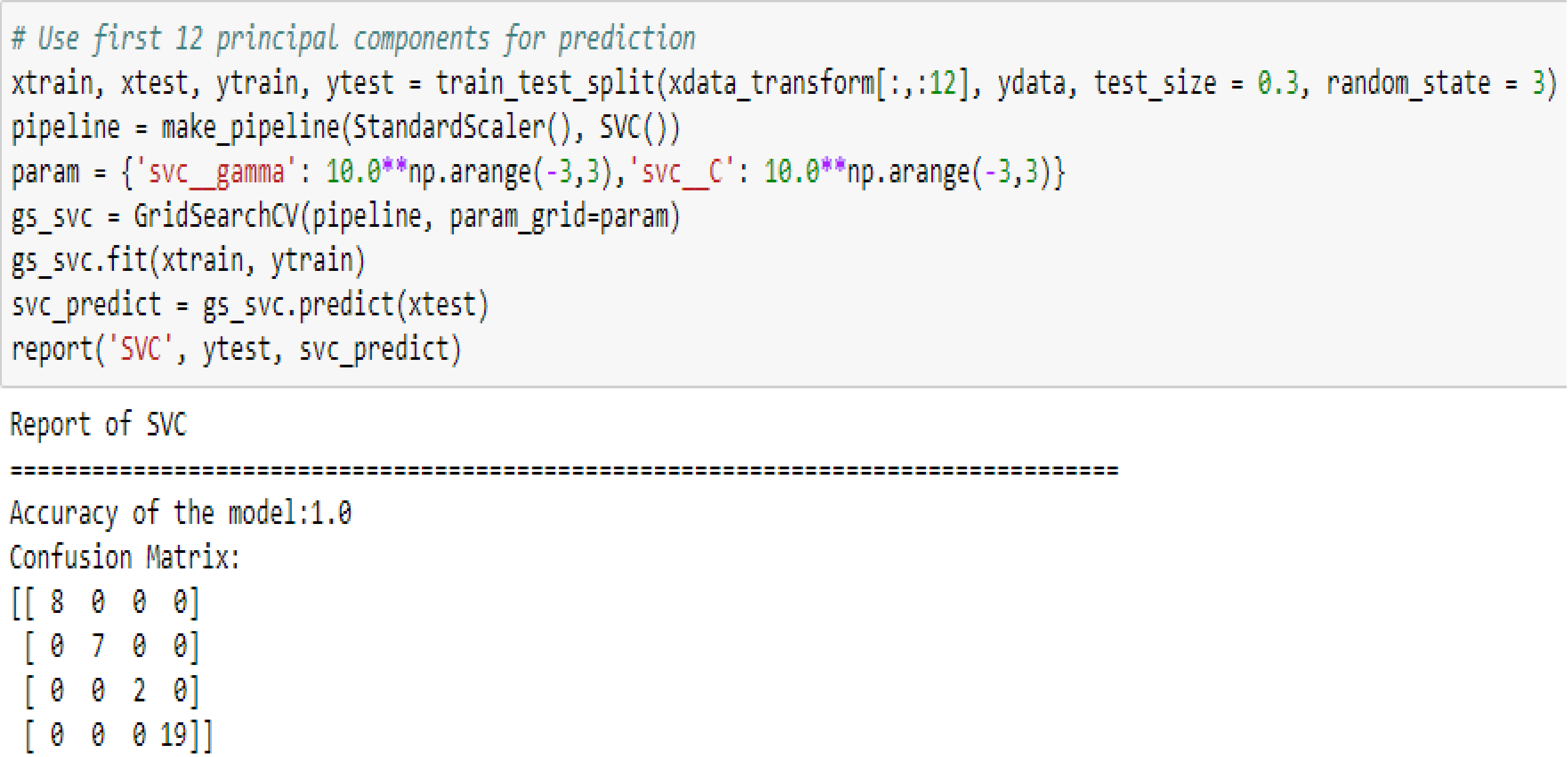


Spectra Projection based on PCA



Model Performance

Supported Vector Classifier (SVC)



```
# Use first 12 principal components for prediction
xtrain, xtest, ytrain, ytest = train_test_split(xdata_transform[:,:12], ydata, test_size = 0.3, random_state = 3)
pipeline = make_pipeline(StandardScaler(), SVC())
param = {'svc__gamma': 10.0**np.arange(-3,3),'svc__C': 10.0**np.arange(-3,3)}
gs_svc = GridSearchCV(pipeline, param_grid=param)
gs_svc.fit(xtrain, ytrain)
svc_predict = gs_svc.predict(xtest)
report('SVC', ytest, svc_predict)
```

```
Report of SVC
================================================================================
Accuracy of the model:1.0
Confusion Matrix:
[[ 8  0  0  0]
 [ 0  7  0  0]
 [ 0  0  2  0]
 [ 0  0  0 19]]
```

Conclusion

- A scan-and-bingo approach was proposed as proof of authentication: Scan sample by hand-held FTIR and analyze by machine learning tools
- A satisfactory high accuracy (>98%) was achieved, and this method also offered us with deep understandings on the key difference between samples.
- This toolkit is especially interesting for product distributors, such as Walmart, Target, and Wholefoods, for quick assessment of product quality, as well as product manufacturer for quick assessment of quality

Thank you!

Stay cool!