利用質譜儀定序偵測細菌抗藥性

Antibiotic Resistance Detection Using Mass Spectrometry Profiling

F04066028　詹子毅
指導教授　莊坤達 教授

Problem Statement

In clinical practice, **antibiotic resistance** is a serious issue that often arises when treating infectious diseases and healthcare-associated infections (HAI).

Therefore, **speeding up the detection process with machine learning methods** is of great interest to clinical doctors and scientists.

Current method
MALDI-TOF
+
Vitek2 blood culture
(1~3 days)

Proposed method
MALDI-TOF
+
machine learning
(< 1 hr)

Visualizing the Dataset

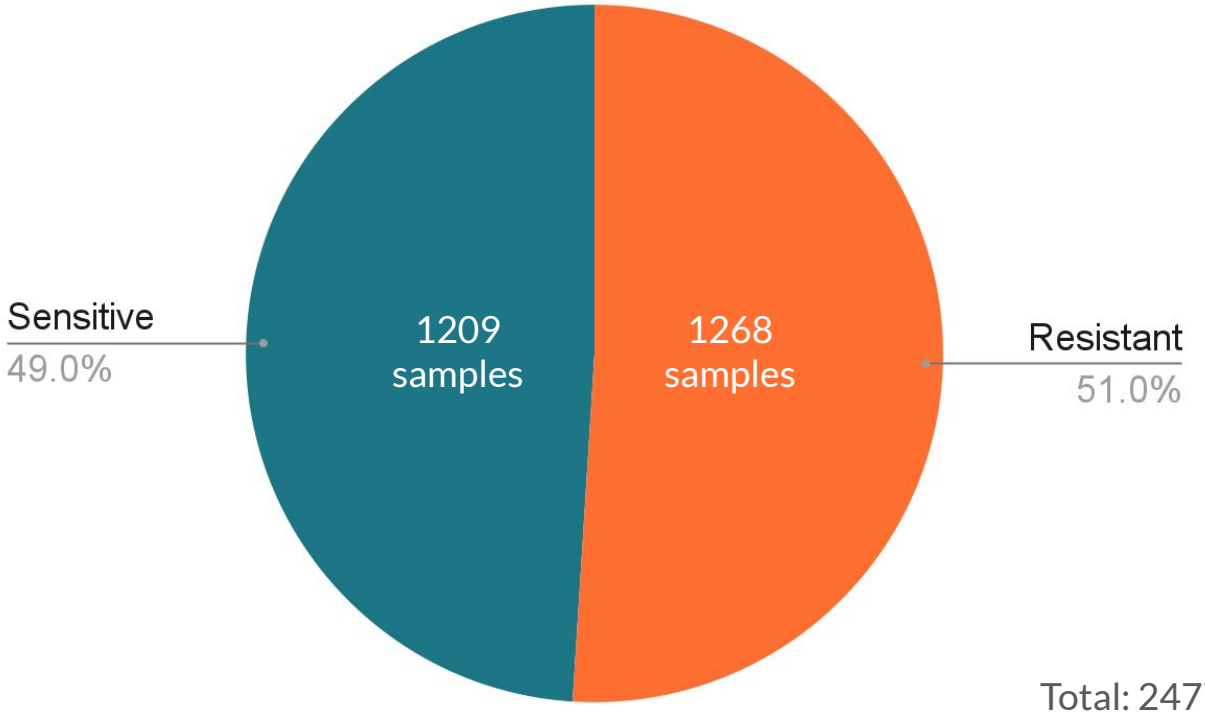
Source: Taipei Veterans General Hospital

Duration: 2018 ~ 2019

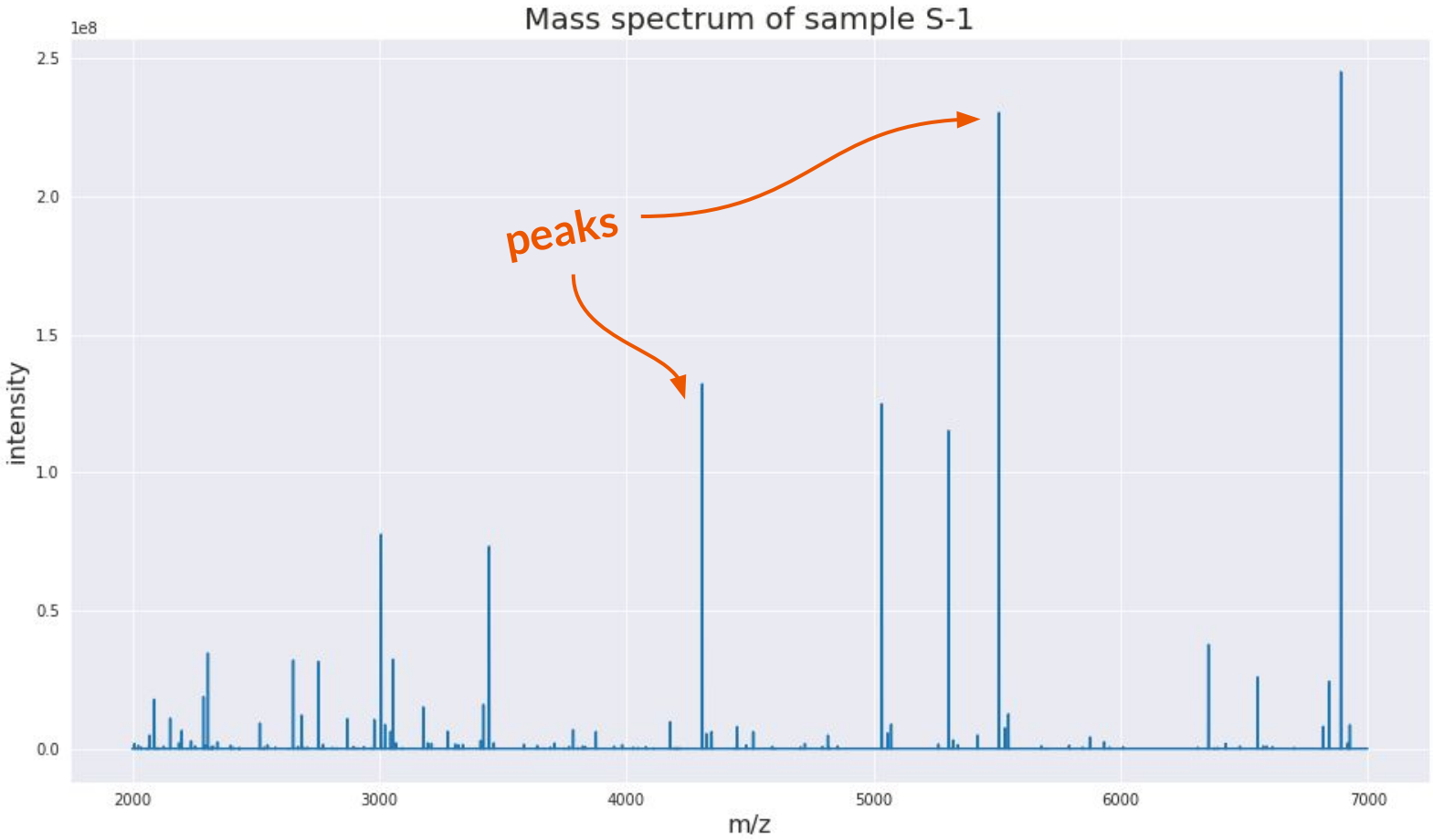
Bateria species: Staphylococcus aureus（金黃色葡萄球菌）

Drug: Oxacillin

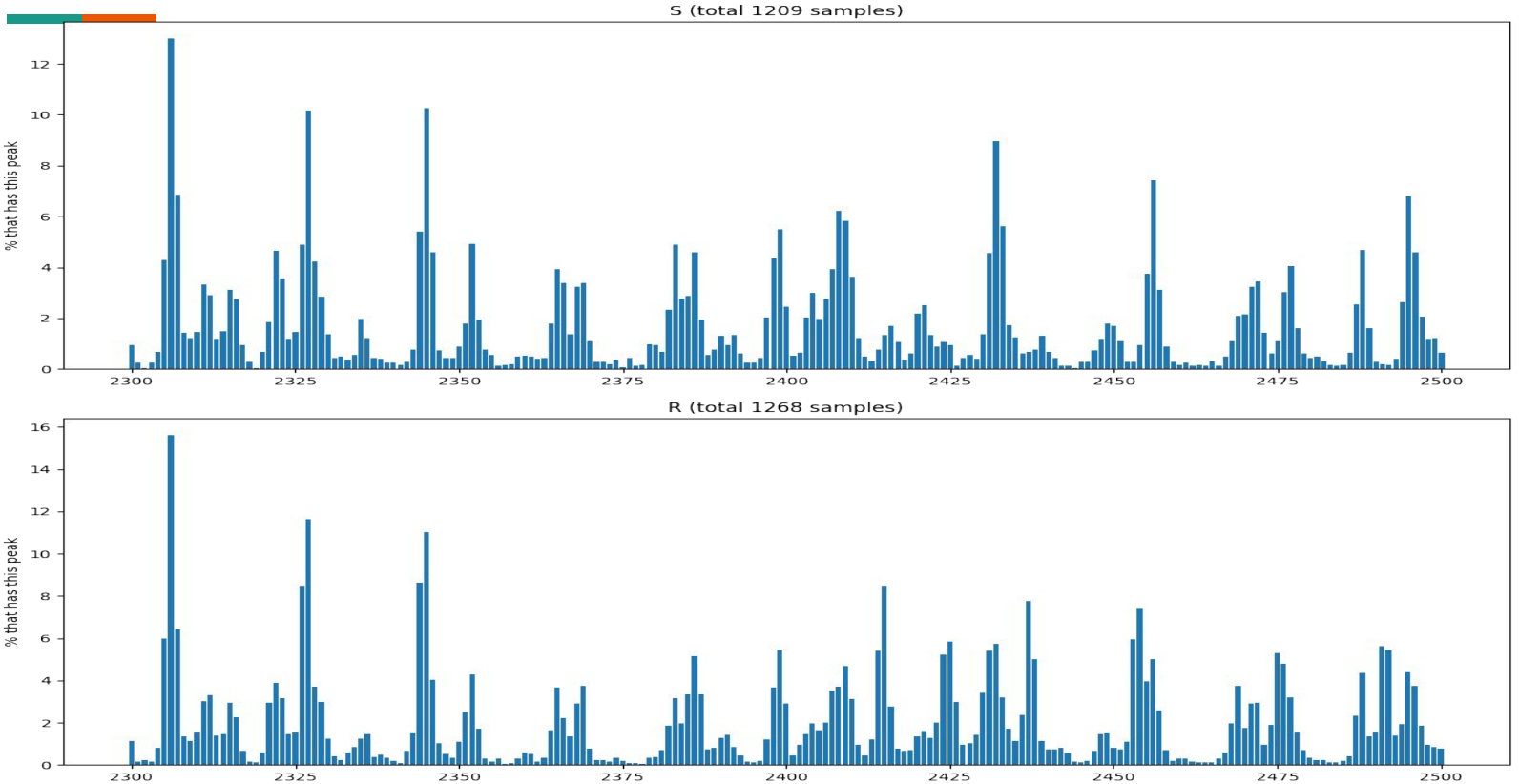
Class Distribution



Total: 2477 samples


Mass spectrum of sample S-1
1e8
2.5
2.0
1.5
1.0
0.5
0.0
intensity
peaks
2000
3000
4000
5000
6000
7000
m/z

Histogram of Peak Occurence



Designing a Solution

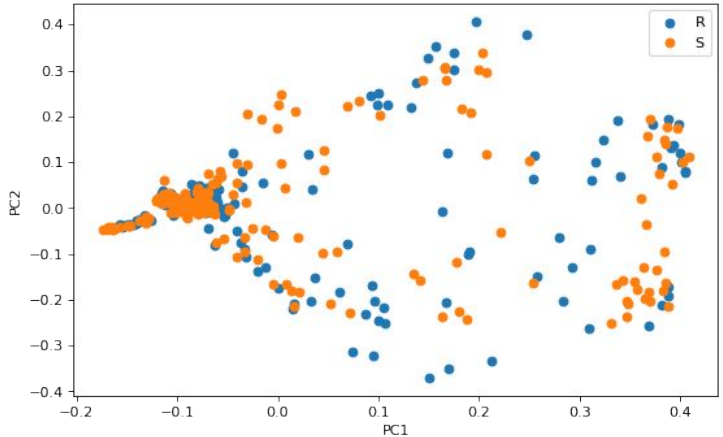
Model Prediction Baselines

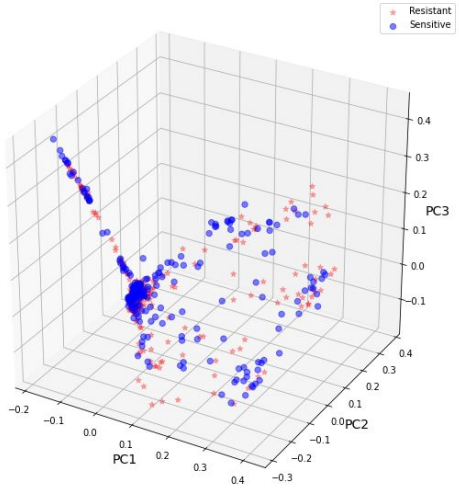
x_test = 20% = 496 samples

	Decision Tree	Random Forest	AdaBoost
parameters	depth = 10	trees = 100	trees = 1000
accuracy	0.639	0.677	**0.700**
R precision	0.838	0.728	0.700
R recall	0.366	0.591	0.724
R f1 score	0.510	0.652	0.712
S precision	0.582	0.641	0.700
S recall	0.926	0.769	0.674
S f1 score	0.715	0.699	0.686

Existing Problems

1. The variance of S / R samples is greater than the variance of bacteria subgroups
2. Subgroups are unlabeled and hard to detect
3. Samples developed from different batches have varying degrees of linear spectrum shift





PCA

Thought Process

- **Hypothesis:**

  Samples from the **same cluster** are likely to **share the same peaks**

  (e.g. samples A and B both have peak 2636)

- **Ensemble selection:**

  Train a model for each subset of samples that have a specific peak

Model Pipeline

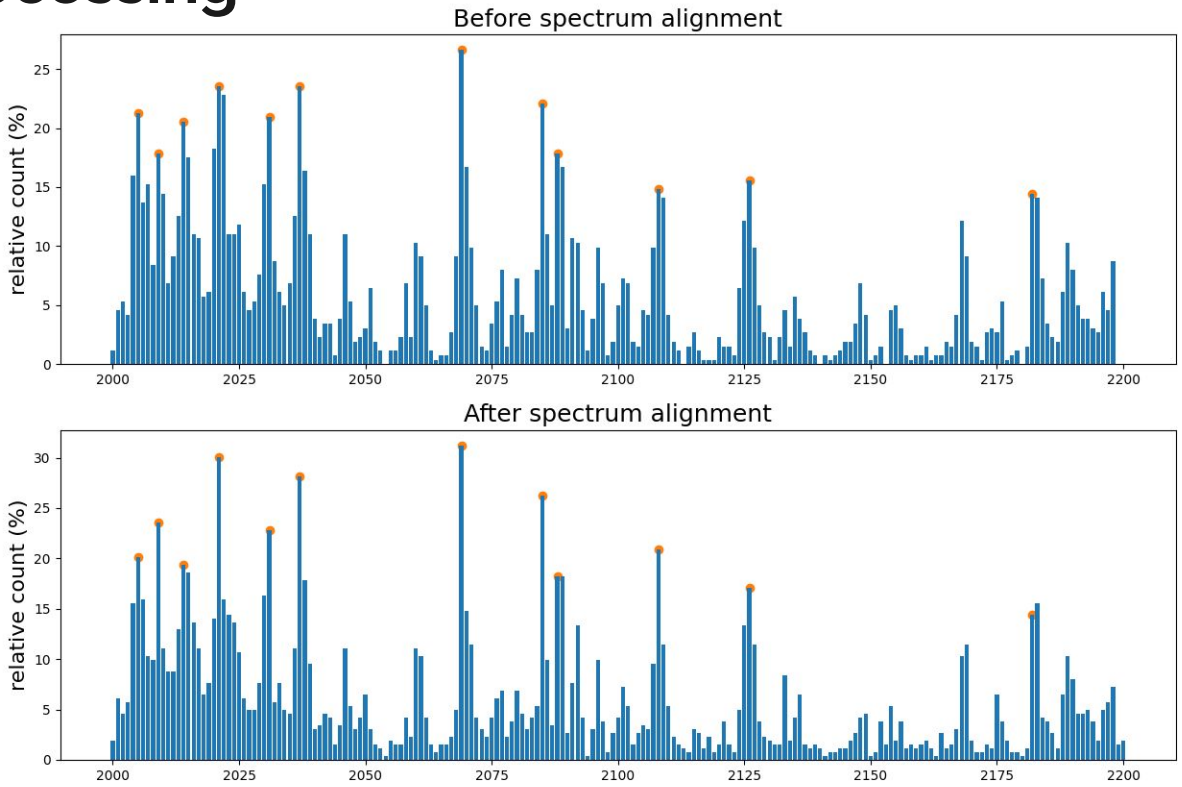
TRAINING

1. **ALIGN** each off-center spectrum
2. **NORMALIZE** using base intensity
3. **common_peaks** = **FIND** the top N peaks with the highest sample count in x_train until all samples have at least 1 peak from **common_peaks**
4. **FOR** each peak **pk** in **common_peaks**:
   clf = **RandomForest**()
   clf.fit(samples that have **pk**)
   scores = cross_validate(clf)
5. **SAVE** all clfs and **scores**

TESTING

1. **leader_peaks** = all **pk**s in **scores** where **“R precision” score > 0.85**
2. **FOR** each **pk** in x_test_sample:
   pred = clf[**pk**].predict(sample)
   IF **pk** in **leader_peaks** and **pred** == **Resistant**:
   ANS = **Resistant**
   ELSE:
   SAVE **pred** to a list
3. **ANS** = **VOTE** using **pred**s from the list, the higher vote wins

Preprocessing



Percentage of samples that have at least 1 leader peak: 72% → 75% after alignment

Decision Function

- Let leader_peaks = [**2636**, 2288, …] (leader_peaks ⟶ R precision > 0.85)
- Case 1:

  Test sample = [2001, 2010, **2636**]

  Predict labels = [S, S, **R**] ⟶ Final verdict: **R**!

- Case 2:

  Test sample = [2001, 2010, 3069]

  Predict labels = [S, S, R] ⟶ Final verdict: S

Results

Comparison with Baseline Models

x_test = 20% = 496 samples

	Decision Tree	Random Forest	AdaBoost	My Voting RFs
parameters	depth = 10	trees = 100	trees = 1000	
accuracy	0.639	0.677	0.700	**0.781**
R precision	0.838	0.728	0.700	**0.881**
R recall	0.366	0.591	0.724	0.660
R f1 score	0.510	0.652	0.712	0.754
S precision	0.582	0.641	0.700	0.720
S recall	0.926	0.769	0.674	0.908
S f1 score	0.715	0.699	0.686	0.803

Conclusions

1. **Test accuracy increased** from approx. 69% → **78%**
2. **High R precision (88%)** with 66% of R samples detected by the model (crucial for clinical practice)
3. Although R precision is high, **model is still S-biased** thus S recall > 0.9
4. Retention time (another parameter like m/z, intensity) was not present in this dataset, maybe its addition could solve point 3

Thank you for listening!