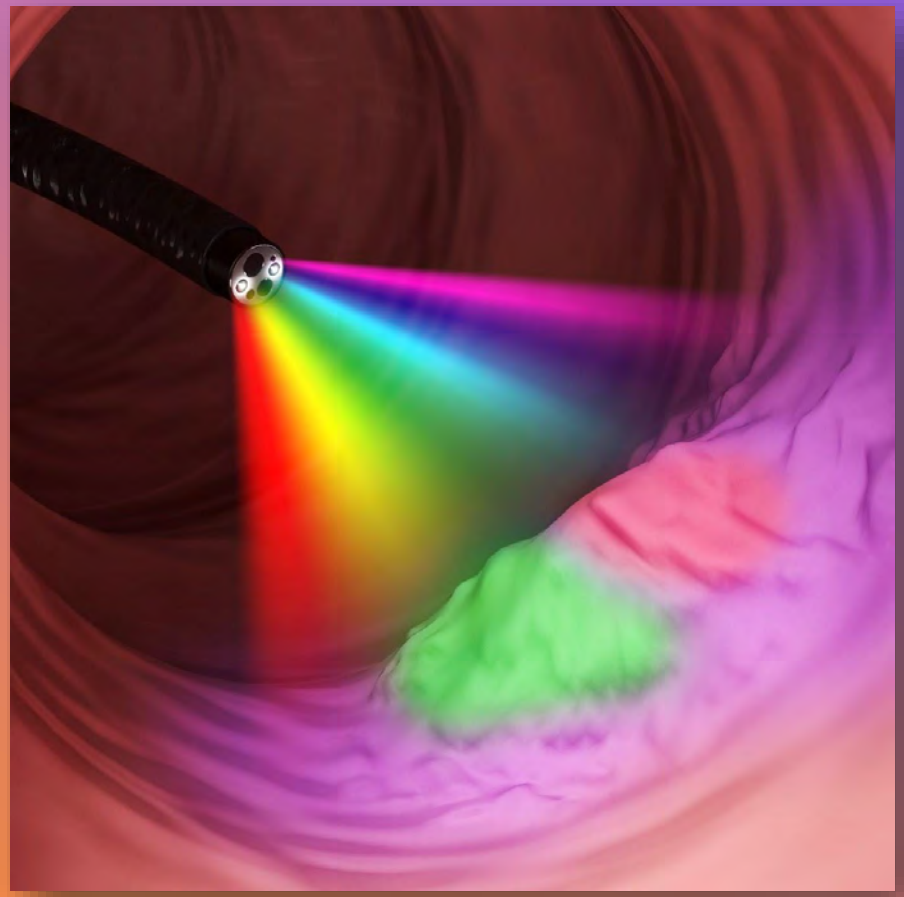
SUPERVISED MACHINE LEARNING IN HYPERSPECTRAL IMAGING APPLIED TO THE DETECTION OF DISEASE IN OESOPHAGEAL LESIONS

Presenter: Yibo Wang

Supervisor: Steve Mead

VISIONLAB



AGENDA

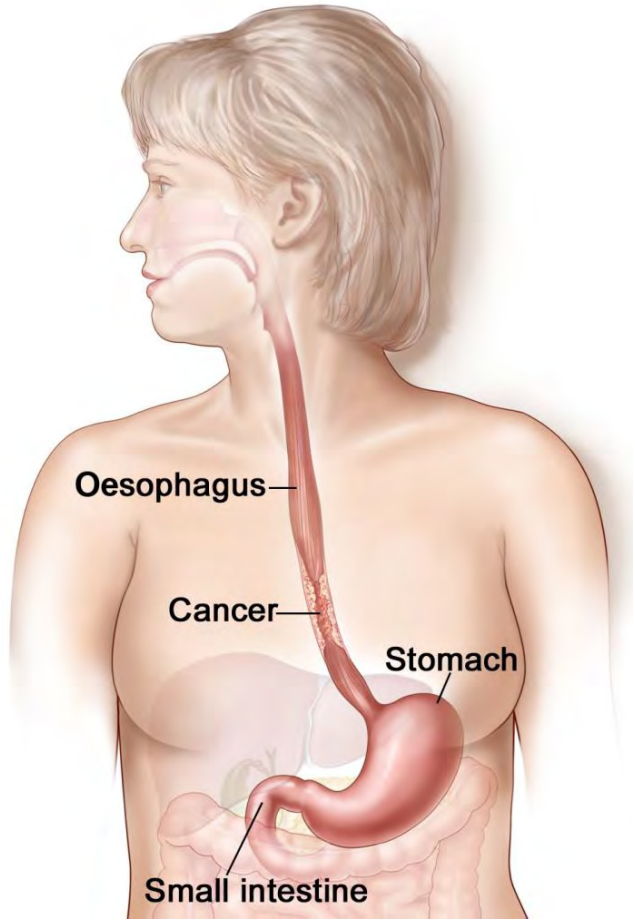
Section 1: Background

Section 2: Preprocessing Spectral Data

Section 3: Unsupervised Learning

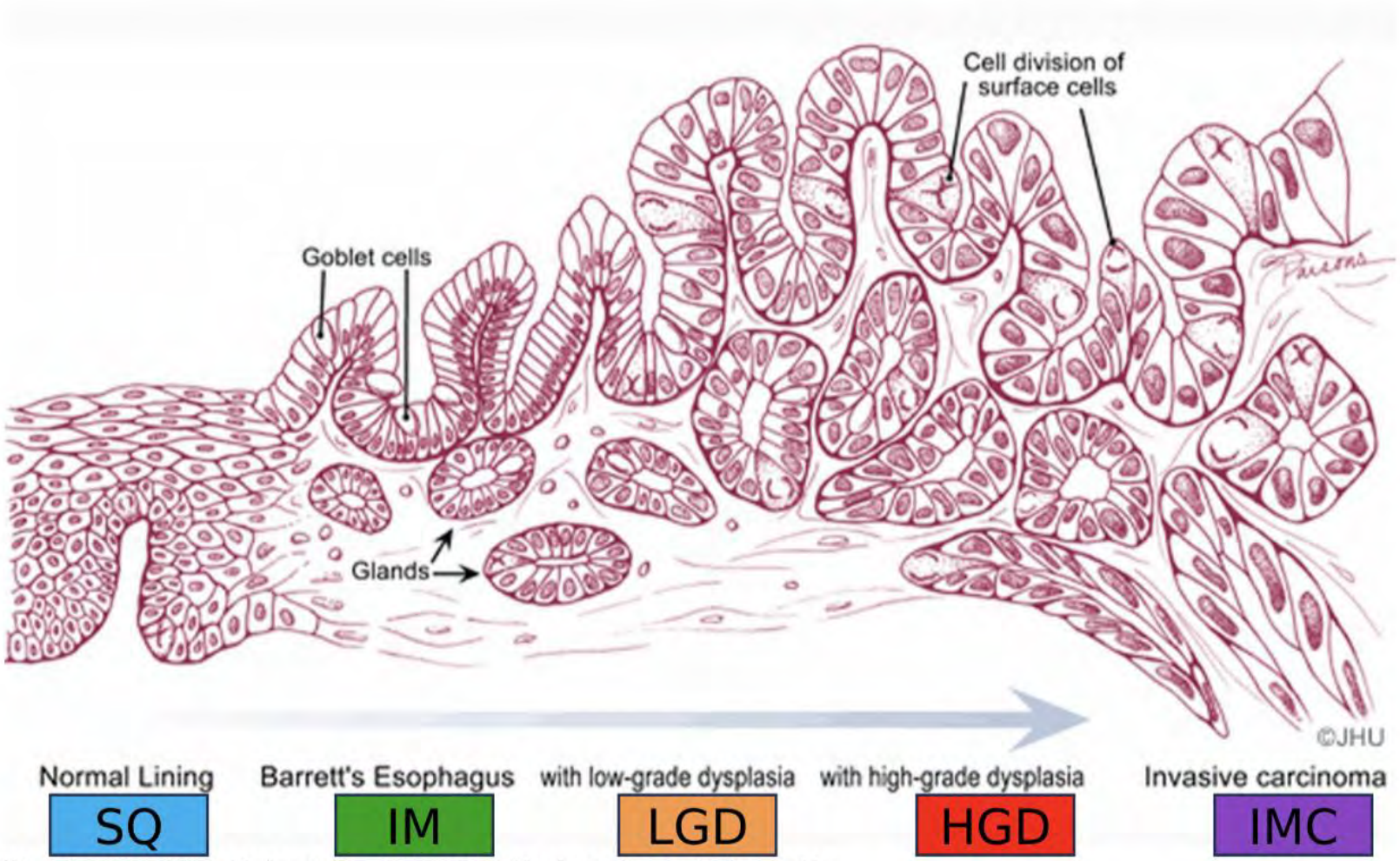
Section 4: Supervised Learning

Background: Importance of Early Detection



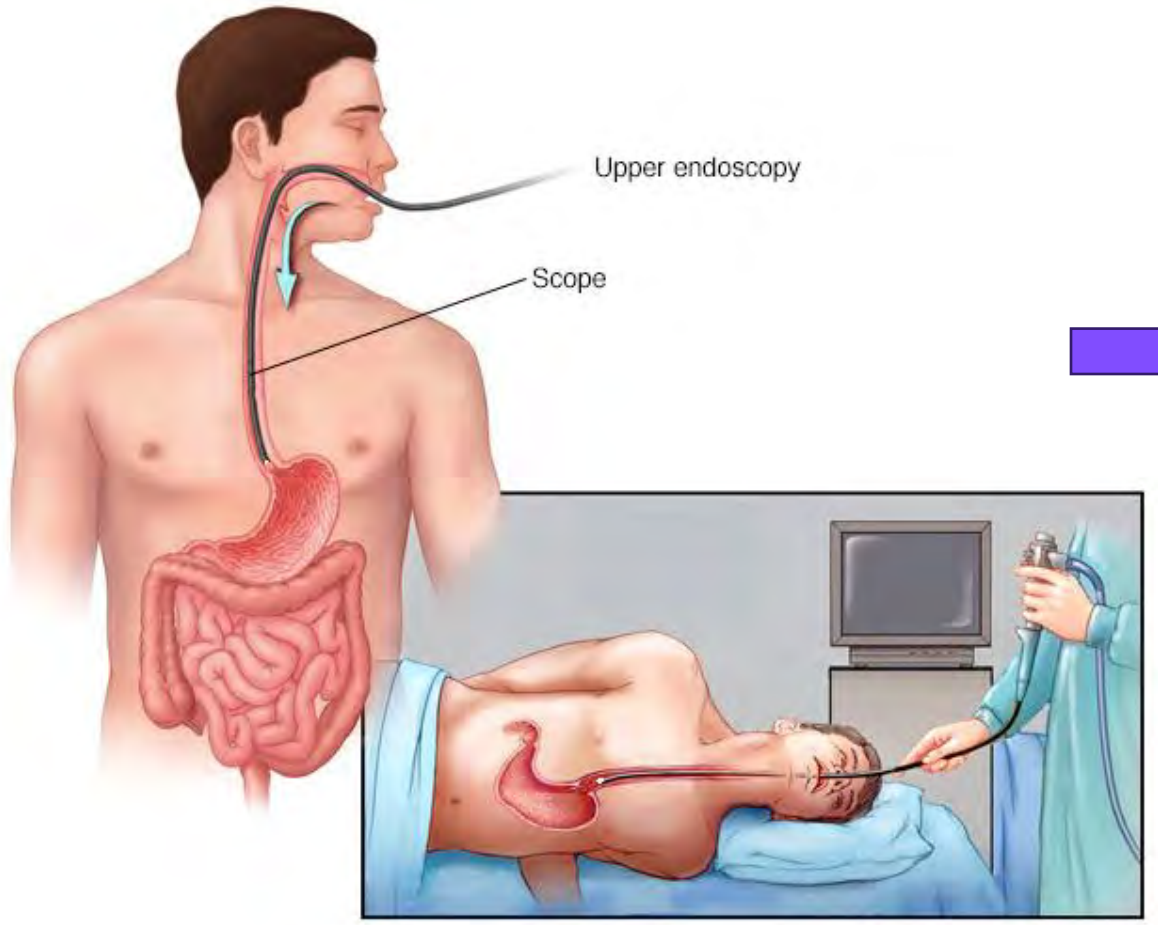
- 9,200 people per year in UK diagnosed with oesophageal cancer, while 70% are at a late stage
- 85% of people diagnosed with the earliest stage survive their cancer for 1 year or more.
- If oesophageal cancer is found early, surgical removal may be possible.

Background: Why Spectral Data is Useful



Credit: John Hopkins University, https://pathology.jhu.edu/barretts-esophagus/dysplasia

Background: How Data are Collected (1)



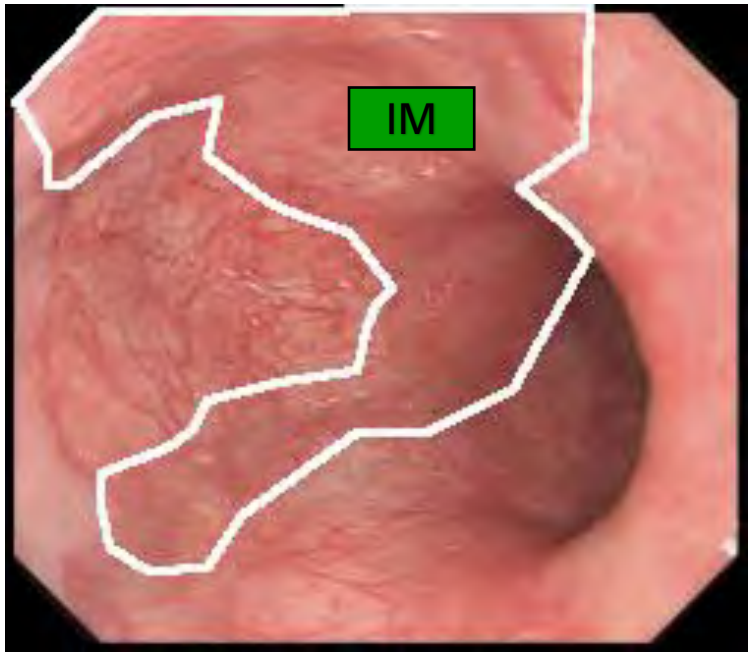
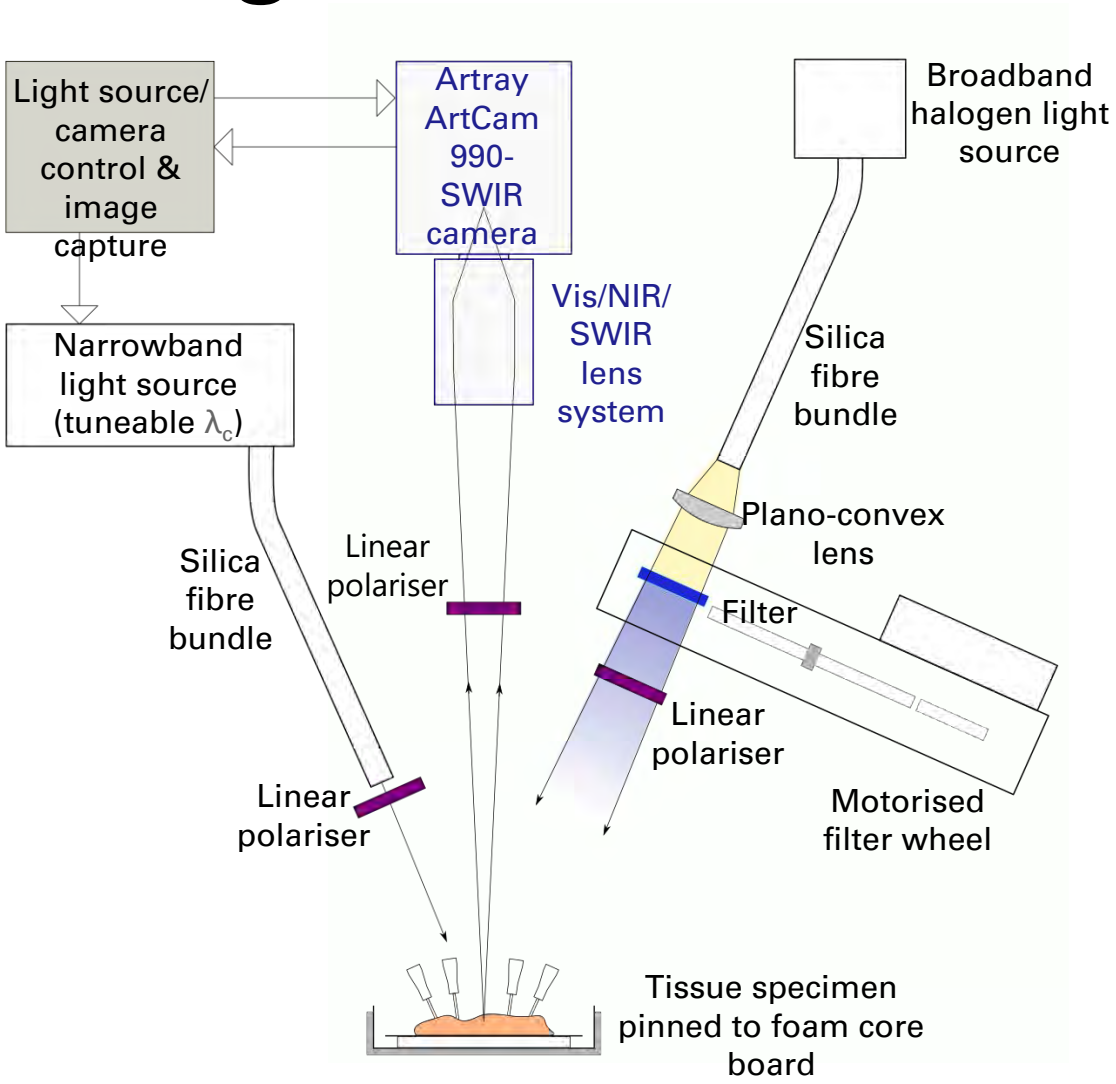
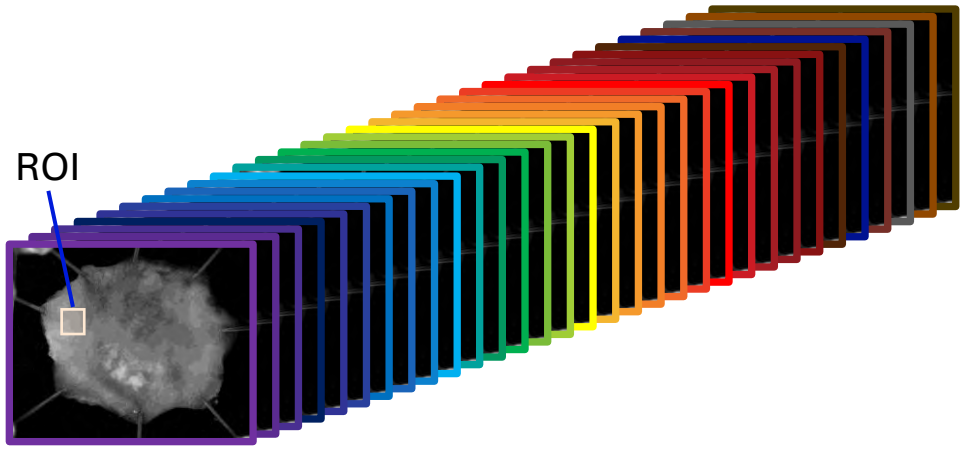


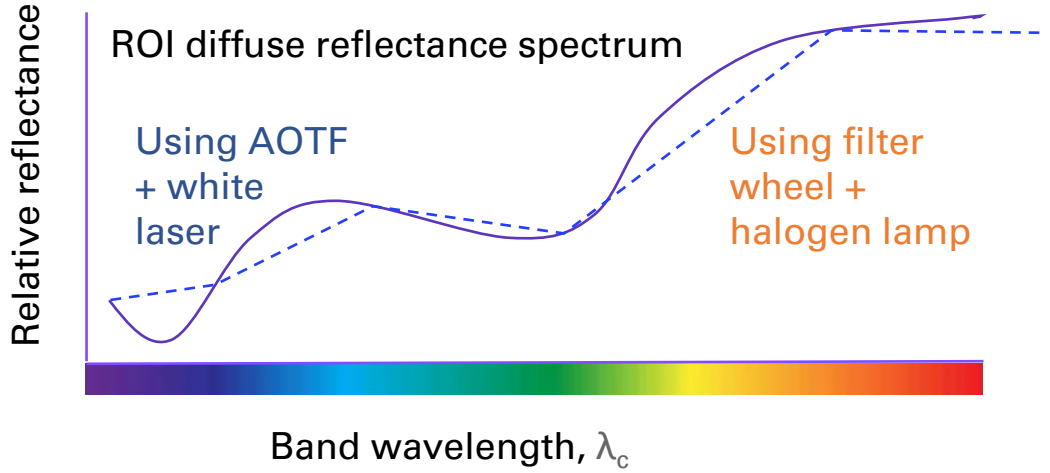
Image of oesophageal lesion

Label is given by histopathologist

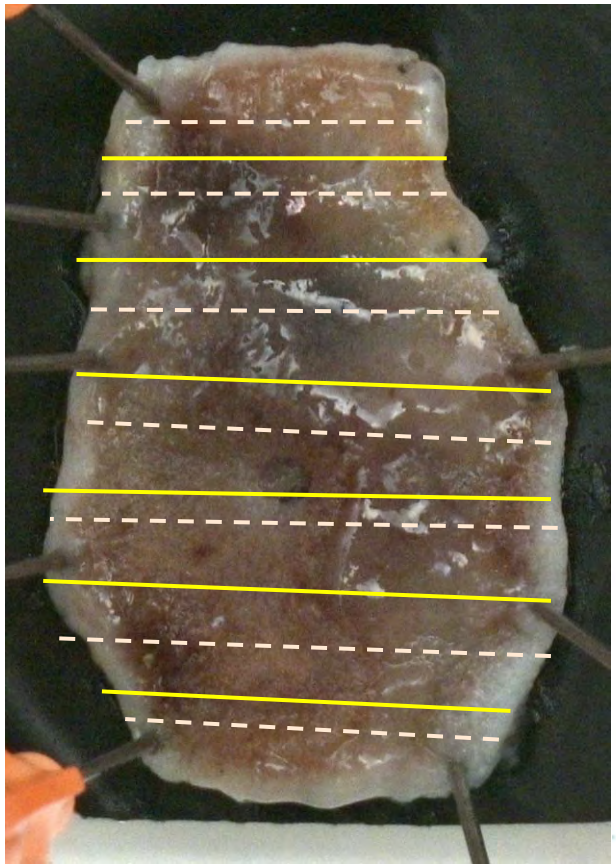
Background: How Data are Collected (2)



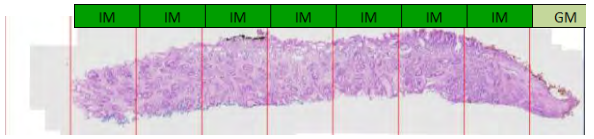




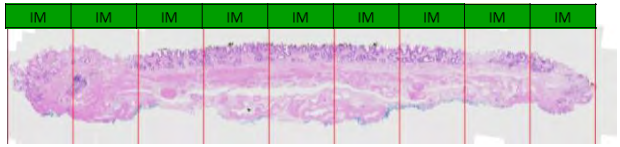
Background: How Label is Given



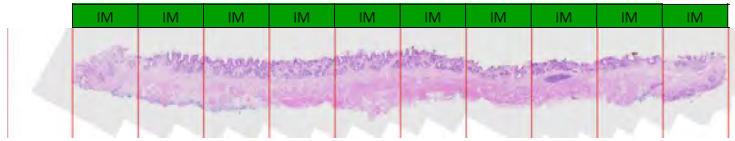
PS23-23645
A1-1/L3

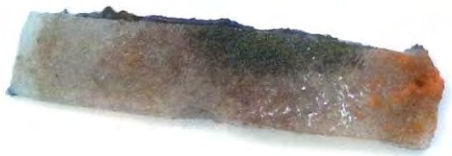


PS23-23645
A2-1/L3

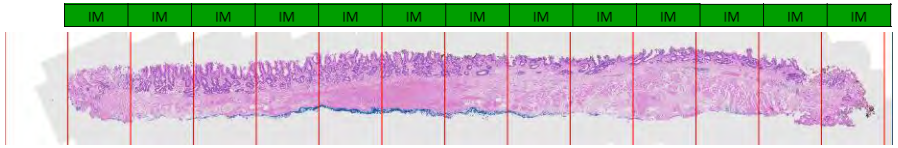


PS23-23645
A3-1/L3

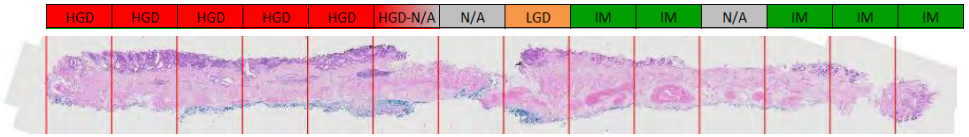




PS23-23645
A4-1/L3



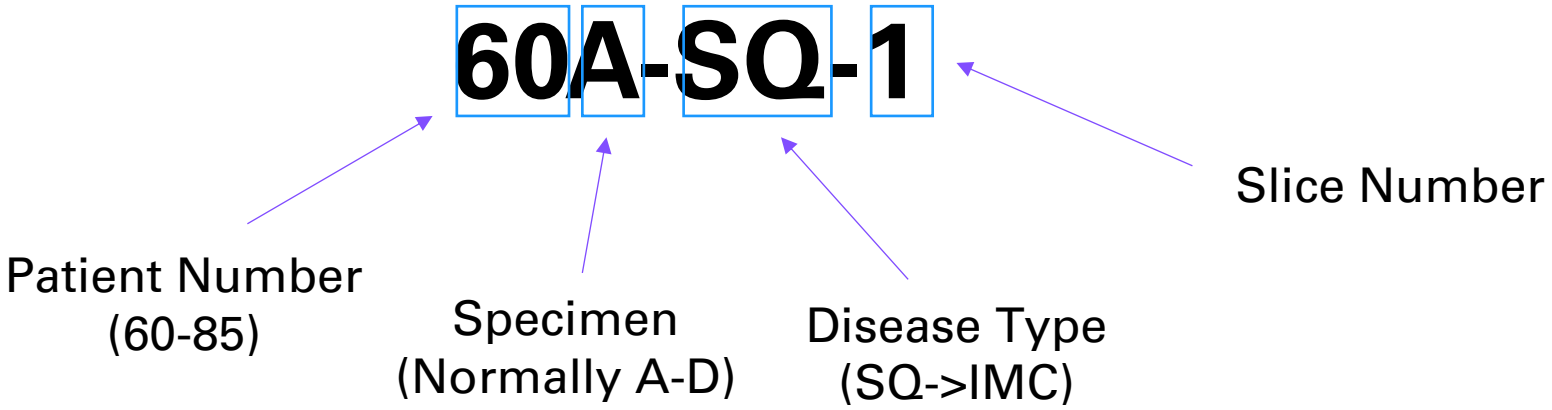
PS23-23645
A5-1/L3



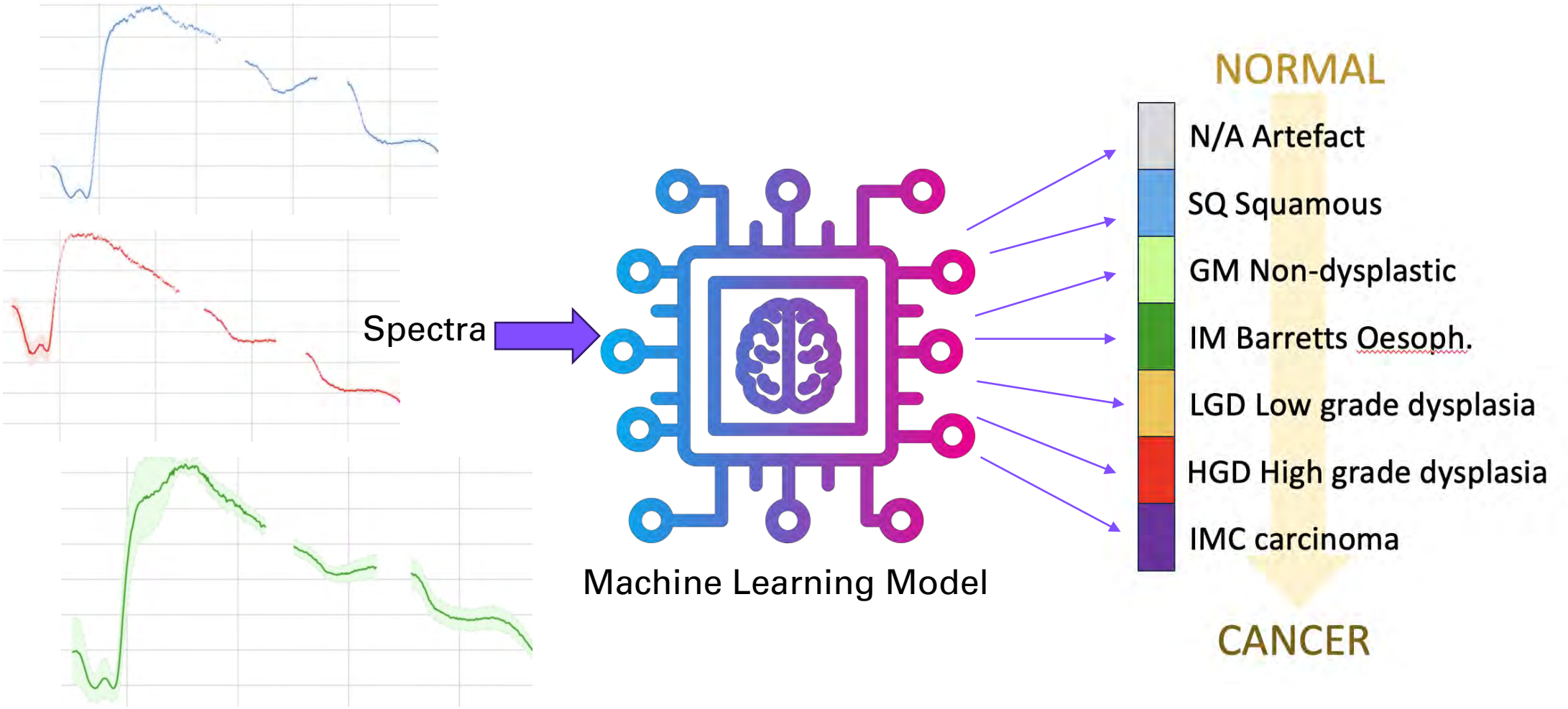
Structure of Spectral Data

Label	500nm	502nm	504nm	…	700nm	702nm	…	1398nm	1400nm
60A-SQ-1	N/A	0.06575	0.06607		0.19877	0.19625		0.01022	0.00958
……									

Note: Data of 850-900nm and 1052-1114nm are labelled ‘N/A’ due to restriction of equipment.



Machine learning applied to spectral data



Preprocessing

Moving Window (Outlier Removal)

Set window size to be 5.

The outlier is defined to be the datapoint which is outside median ± 8/10 median absolute deviation

Approx. 1% has been detected to be outliers.

Savitzky Golay Filter (Smoothing)

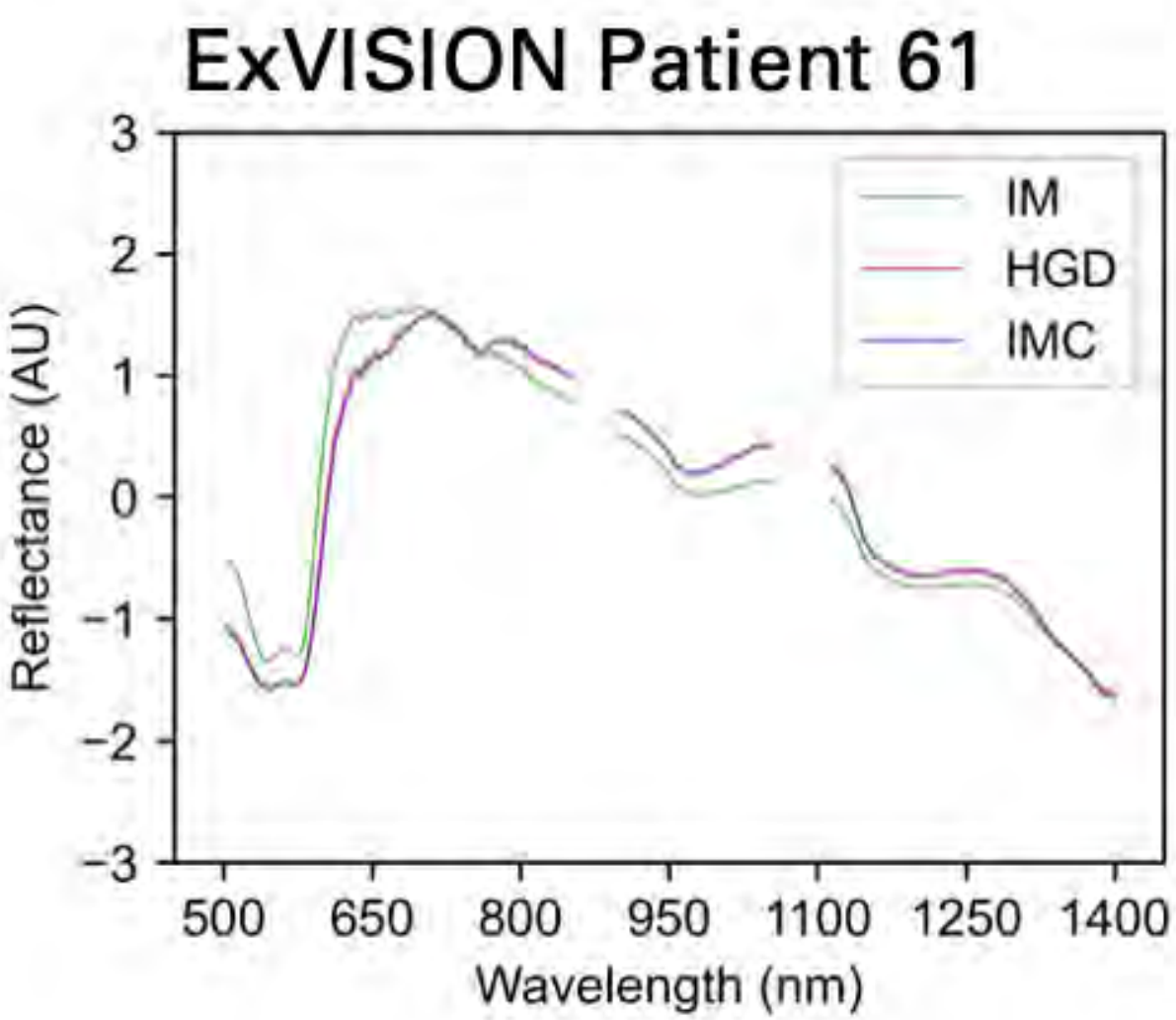
Set window size to be 11 (conventional setting) and polynomial order to be 7 (smaller order may cause over-smoothing).

Standard Normal Variate (Normalising)

One way of normalising spectral data. Proved to have the best performance from other research groups.

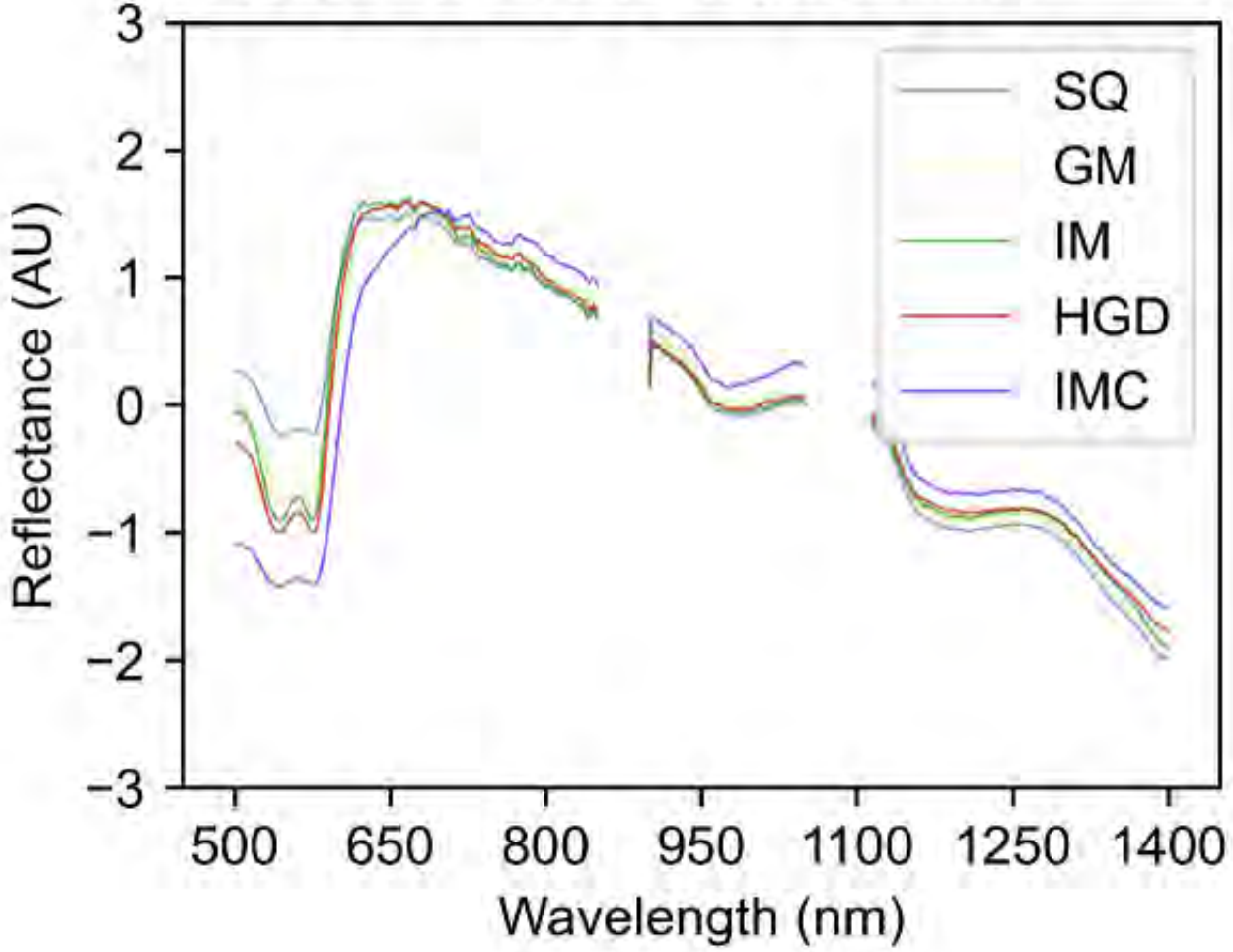
Plotting Mean Spectra

Plot the mean spectra of 18 patients, as well as the '± 1 standard deviation' version, to try to find some patterns of the difference between each stage.



SAM (Ref: IM)
HGD: 0.240
IMC: 0.260

Pearson
HGD: 0.971
IMC: 0.966



SAM (Ref: SQ)
GM: 0.178
IM: 0.204
HGD: 0.261
IMC: 0.535

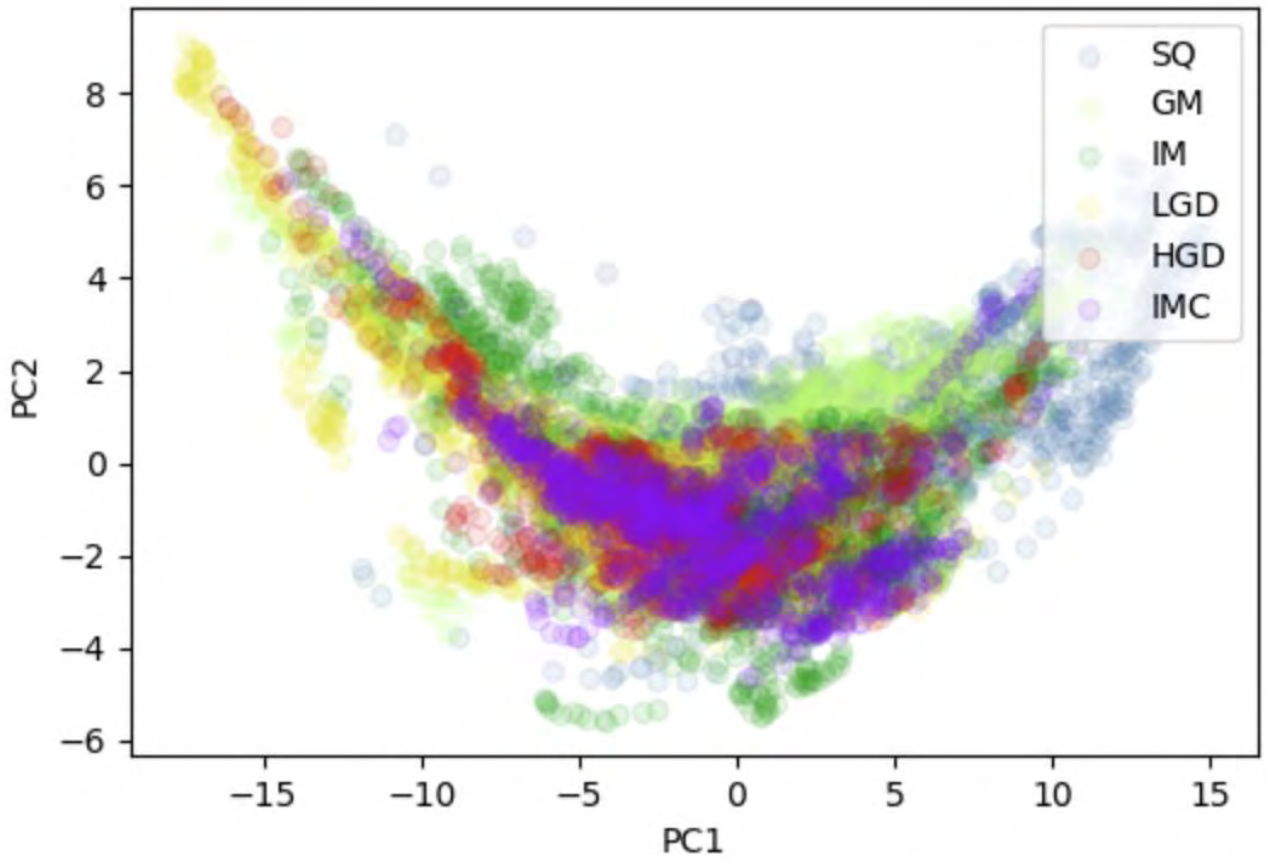
Pearson
GM: 0.984
IM: 0.979
HGD: 0.966
IMC: 0.860

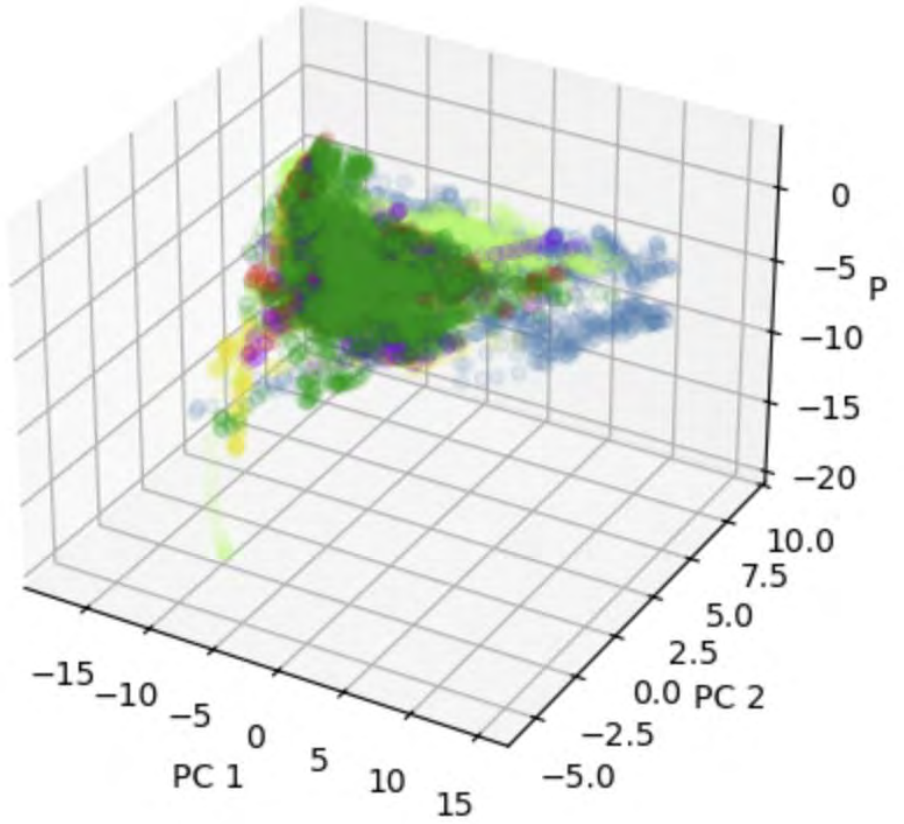
Machine Learning Part

Unsupervised Learning
(No Label Used)

Supervised Learning
(Label Used)

Visualisations of PCA

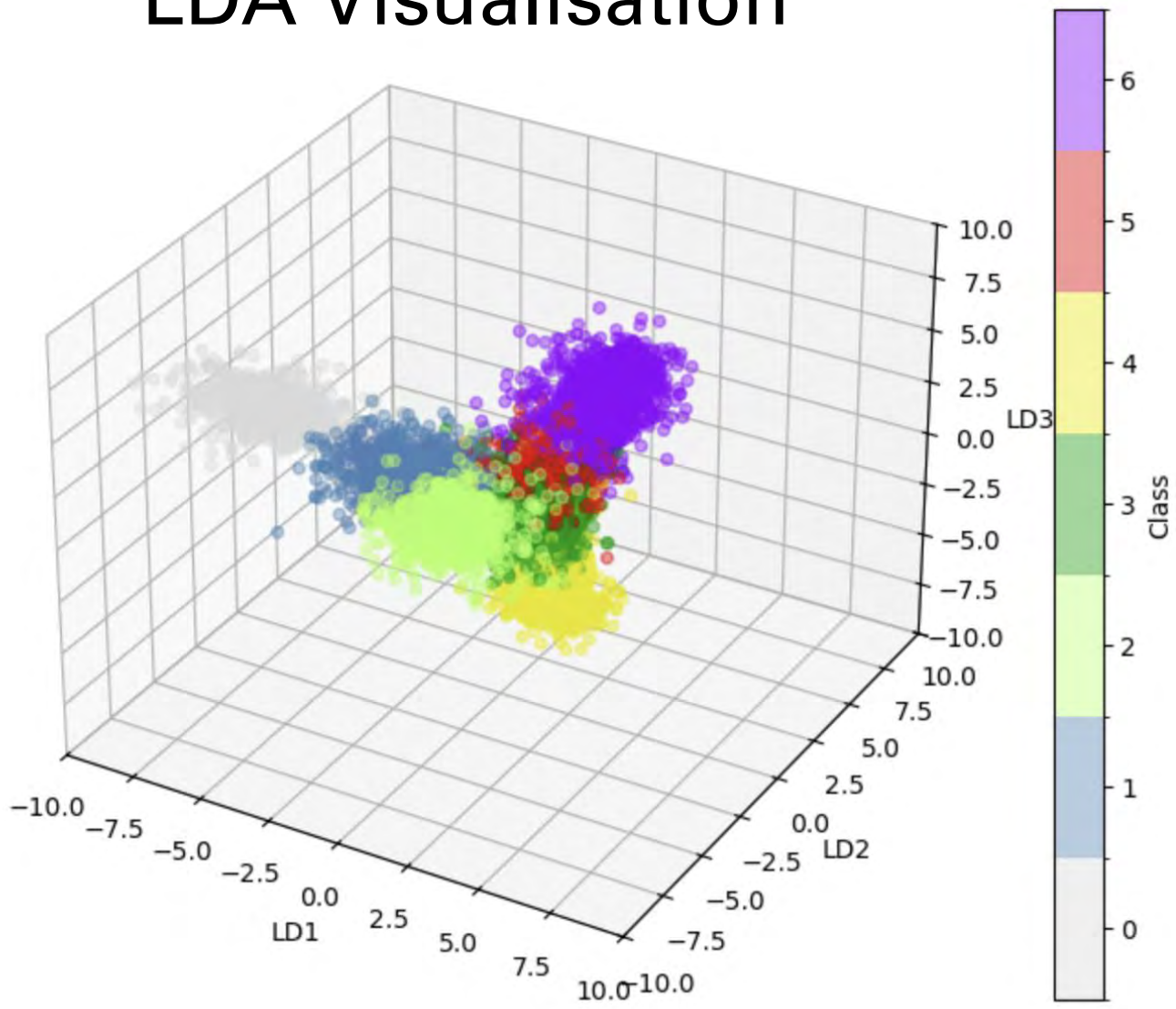


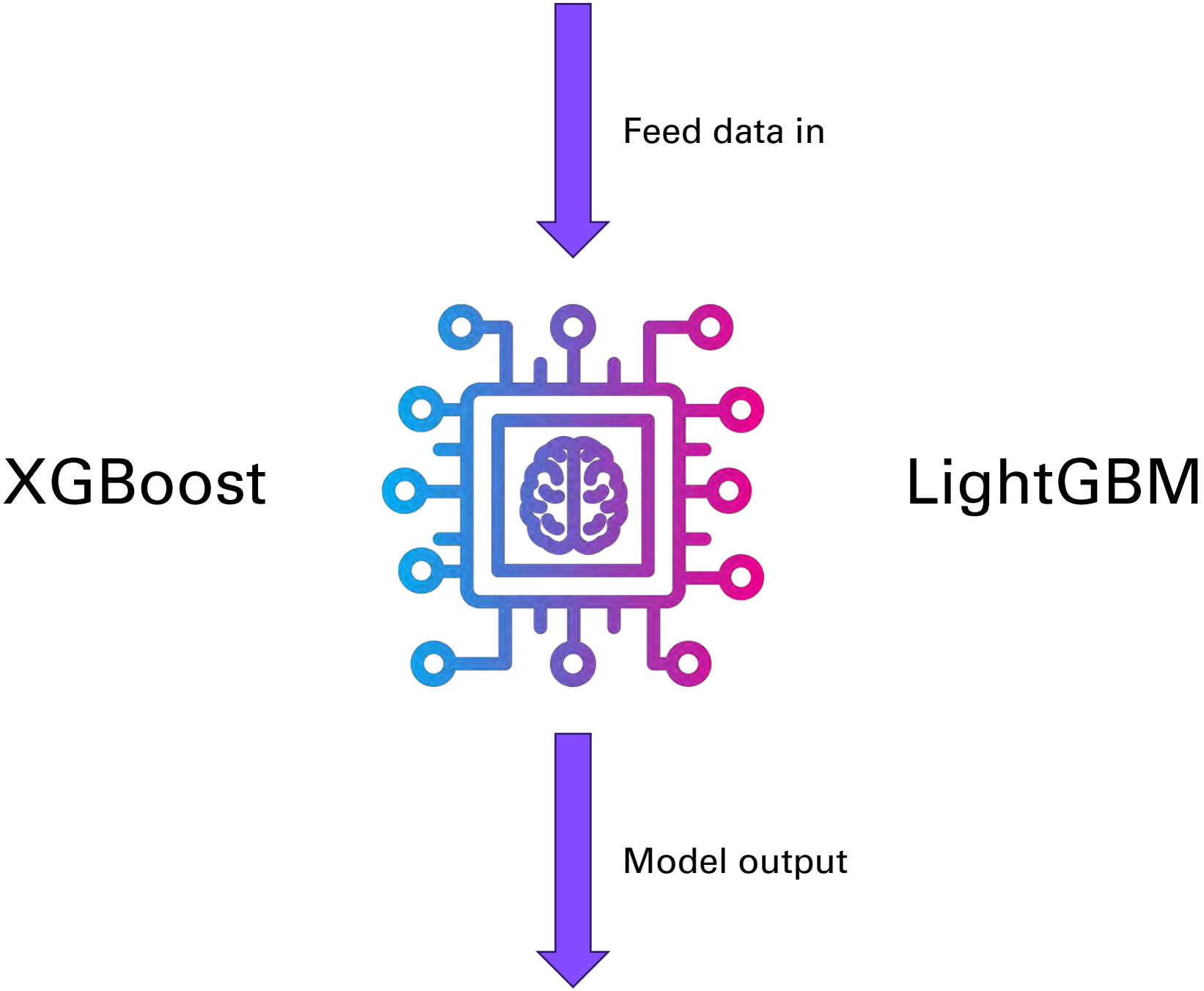


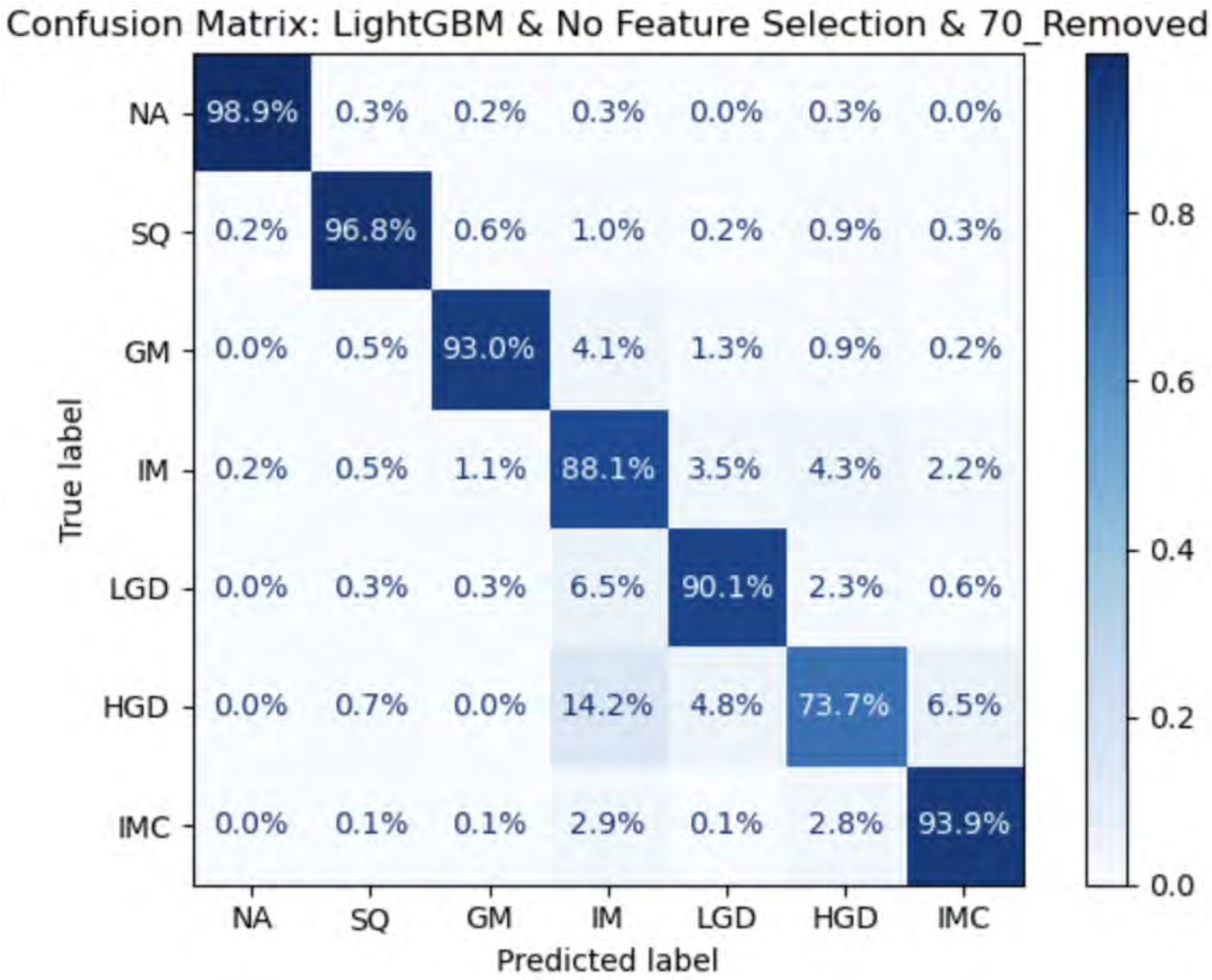
LDA Visualisation

LDA is similar as PCA, where it is also used to reduce the dimension of data.

After applying LDA, we obtain a transformation matrix.




Feed data in
XGBoost
LightGBM
Model output


Confusion Matrix: LightGBM & No Feature Selection & 70_Removed
NA
SQ
GM
IM
LGD
HGD
IMC
True label
98.9% 0.3% 0.2% 0.3% 0.0% 0.3% 0.0%
0.2% 96.8% 0.6% 1.0% 0.2% 0.9% 0.3%
0.0% 0.5% 93.0% 4.1% 1.3% 0.9% 0.2%
0.2% 0.5% 1.1% 88.1% 3.5% 4.3% 2.2%
0.0% 0.3% 0.3% 6.5% 90.1% 2.3% 0.6%
0.0% 0.7% 0.0% 14.2% 4.8% 73.7% 6.5%
0.0% 0.1% 0.1% 2.9% 0.1% 2.8% 93.9%
NA
SQ
GM
IM
LGD
HGD
IMC
Predicted label
0.8
0.6
0.4
0.2
0.0

It seems that we can stop at this point…

But something went wrong…

Where is the problem?

Supervised Learning Pipeline

Linear Discriminant Analysis (LDA)

Labels are used, so that we can distinguish data of different diseases by maximising the distance between each clusters

Feature Selection

Available methods include:

- Boruta (recommended by Imperial College)
- Spectral Band Selection

Training the Model

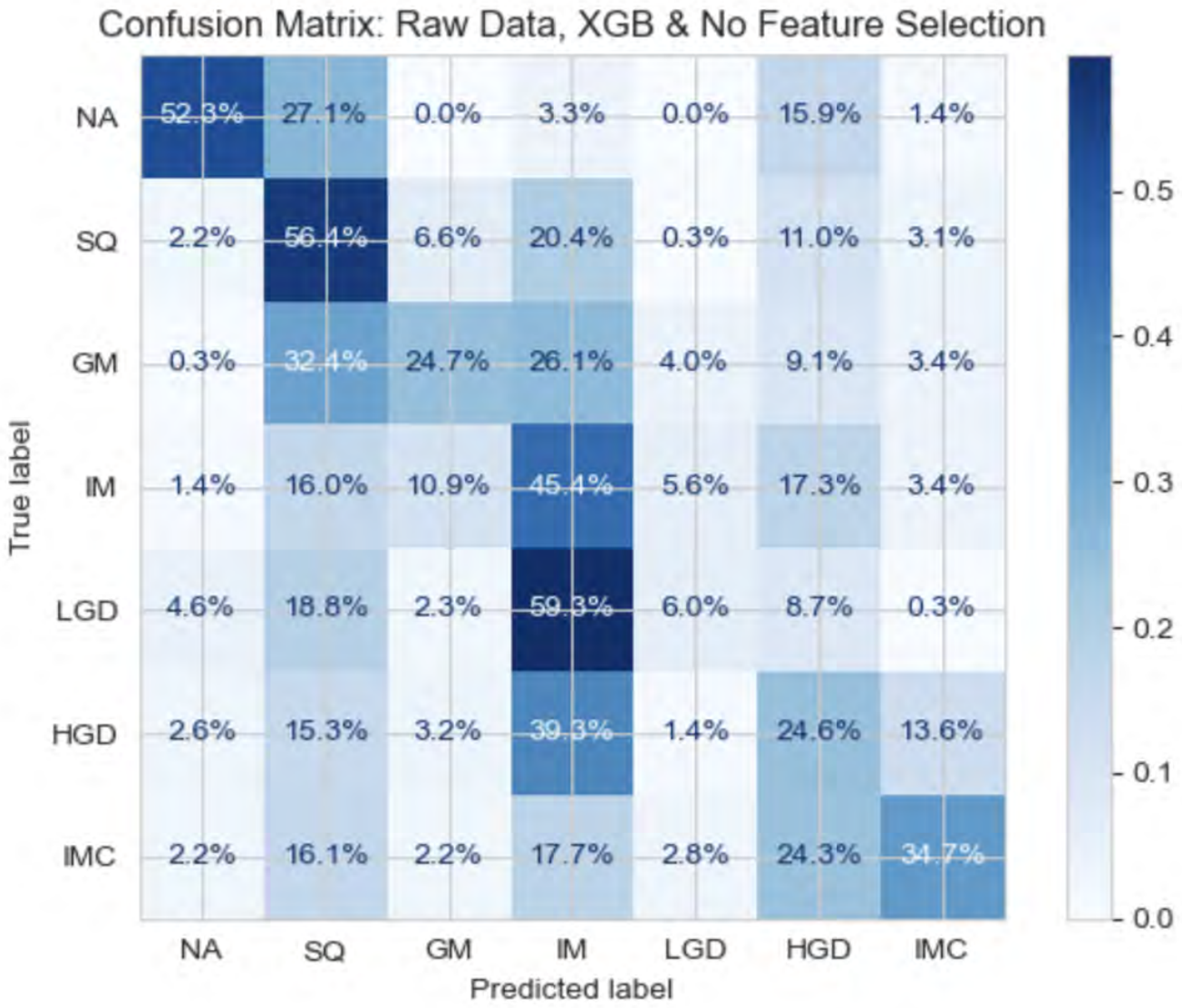
Available methods include:

- XGBoost
- LightGBM

Test

To avoid overfitting, cross validation is used to separate the training and test sets.
Confusion matrices and receiver operating characteristics are used to measure the performance of the model.

Problem of Data Leakage



The performance dropped down drastically…

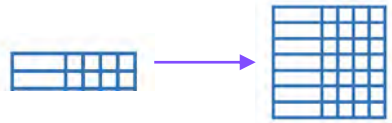
Conclusions:

- Spectral data might not be enough.
- Read papers from other research groups.

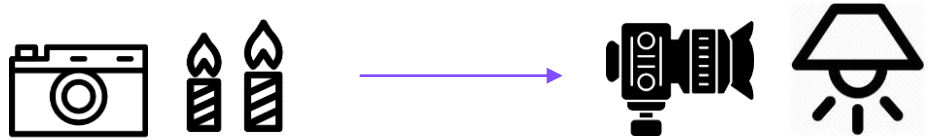
What to do next:

- Improve the preprocessing stage to reduce the differences in spectra between patients.
- Read papers from other research groups to find out ways of reduce batch effects.

- Increase the number of data



- Try improving the experiment (use camera and light source with less noises to catch more data).



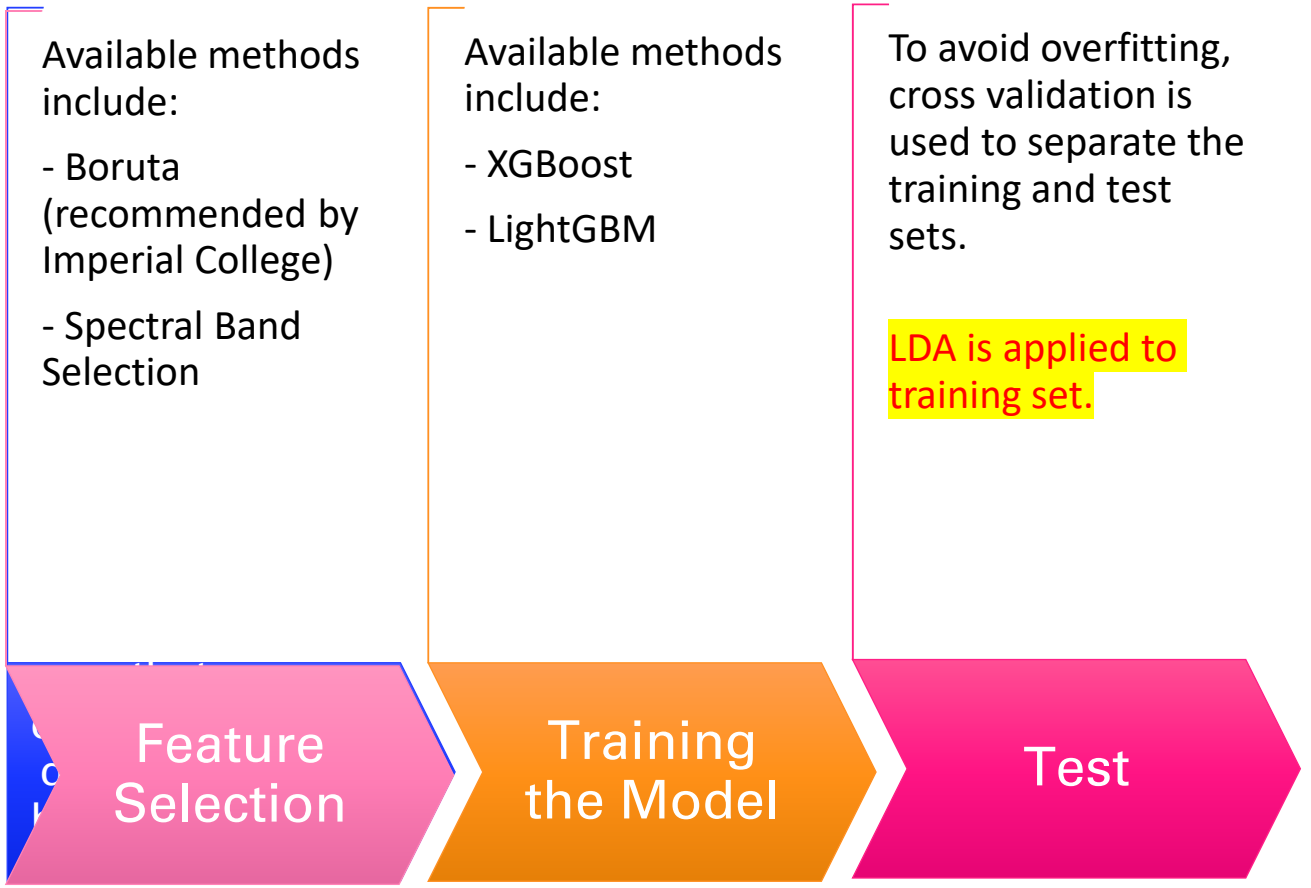
THANK YOU

Presenter name

Email address

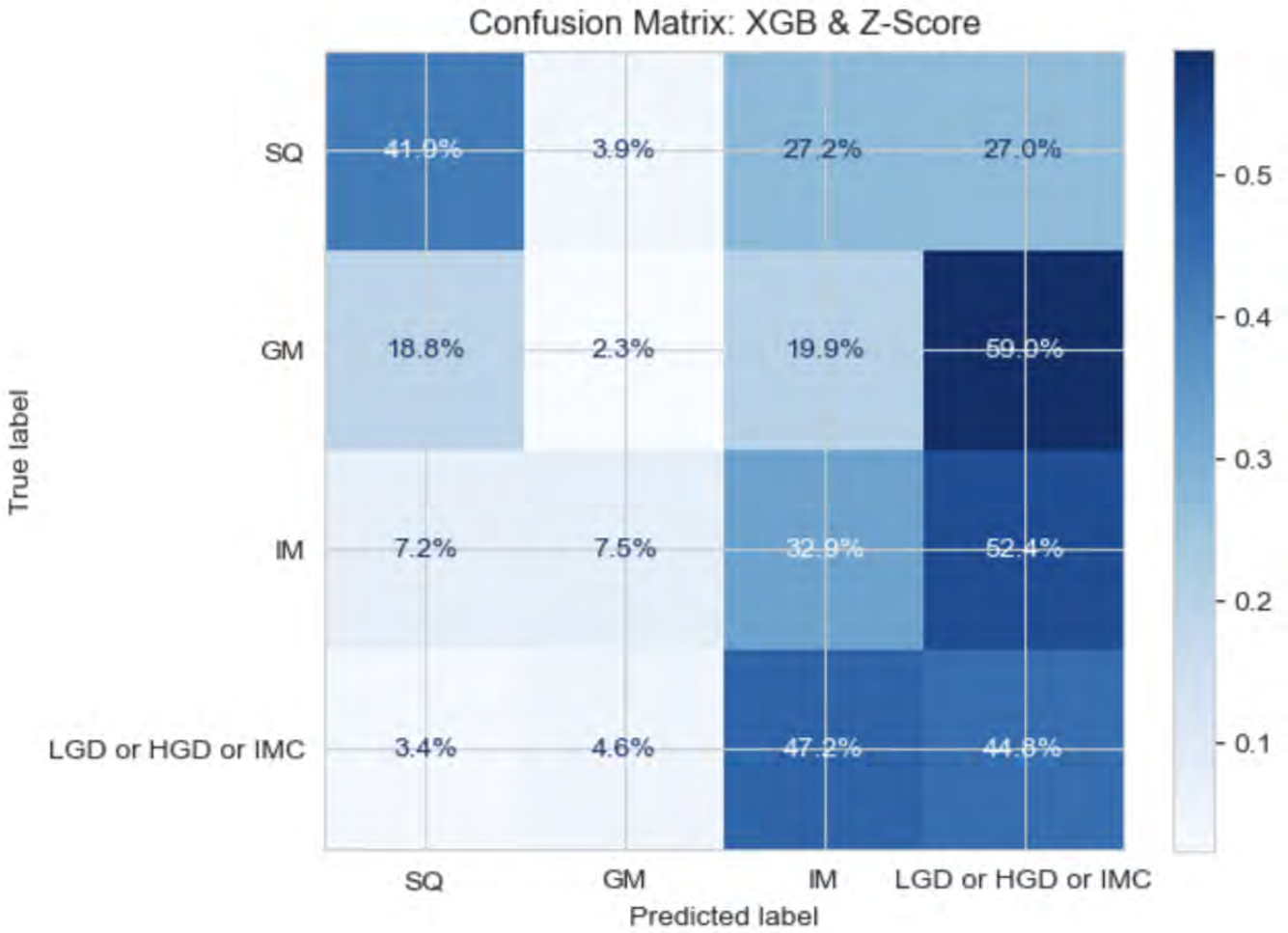
Website

Supervised Learning Pipeline (After Correction)



To improve, try

- Normalising the training set by Z-score
- Regrouping the data to reduce the difficulty of classification
- pycombat: a python library used to reduce the batch effects

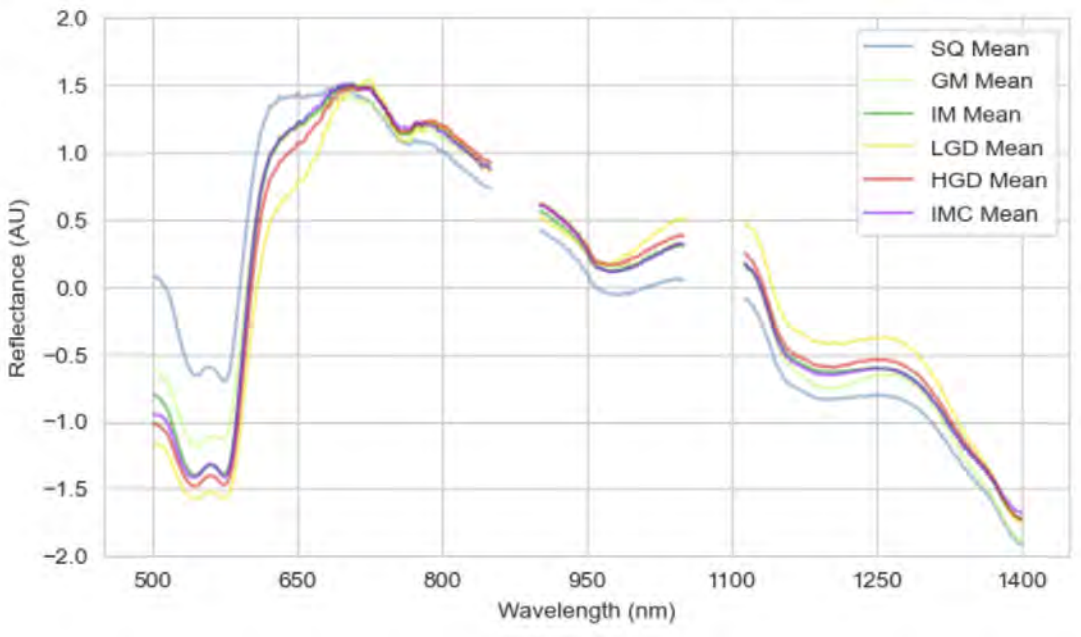


To improve:

- Change LDA dimension (6 instead of 3)
- Use preprocessed data (examine performance for each step)

  (e.g. adjusting parameters in smoothing filter, different normalizing methods)
- Add first derivative data
- Add filter wheel data (new source of data)
- Eliminate data from patient 70

Enlightened by →



Method 1: Spectral Angle Mapper (SAM)

We will calculate the angle between two spectral vectors:

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right)$$

Same as before, the smaller the Euclidean distance is, the more similar the two vectors are.

Method 2: Pearson Correlation

The Pearson correlation is:

$$r = \frac{\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})}{\sigma_x\sigma_y}$$

where $\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})$ is the covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$, and σ_x, σ_y are the corresponding standard deviations.

The range of r is $[-1, 1]$.

- If $r = 1/-1$, $\boldsymbol{x}$ and $\boldsymbol{y}$ are perfect positive/negative correlated.
- If $r = 0$, $\boldsymbol{x}$ and $\boldsymbol{y}$ are not correlated.

SAM Values

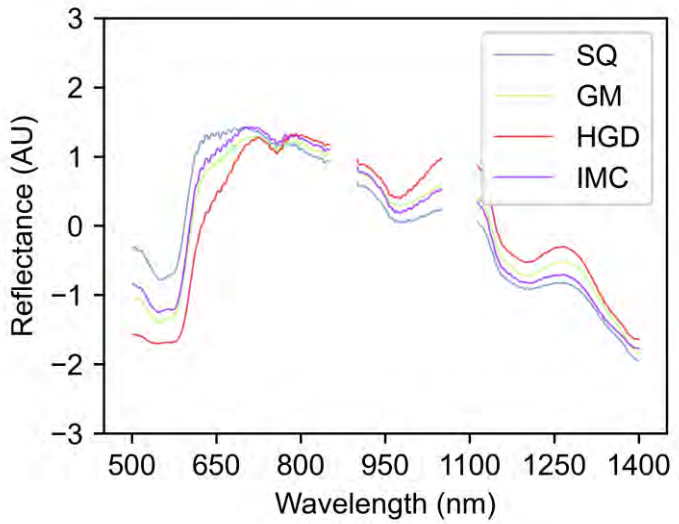
	60	61	62	66	67	70	71	72	73	74	75	76	79	80
SQ	0.119	NaN	0.541	0.275	0.448	0.154	NaN	0.37	0.064	NaN	0.157	NaN	0.186	0.38
GM	0.193	NaN	0.944	NaN	0.063	0.228	0.267	0.131	NaN	NaN	NaN	0.243	0.165	NaN
IM	NaN	0.155	NaN	0.219	NaN	0.331	0.301	0.059	0.231	0.06	NaN	0.104	0.082	0.124
LGD	NaN	NaN	NaN	NaN	0.292	NaN	NaN	NaN	0.345	0.146	NaN	0.121	NaN	NaN
HGD	0.362	0.045	NaN	0.059	0.329	0.339	NaN	NaN	NaN	0.038	0.102	0.095	0.062	NaN
IMC	0.138	0.11	NaN	0.063	NaN	0.062	0.053	NaN	NaN	0.174	0.092	0.147	NaN	NaN

PCC Values

	60	61	62	66	67	70	71	72	73	74	75	76	79	80
SQ	0.993	NaN	0.857	0.962	0.901	0.988	NaN	0.932	0.998	NaN	0.988	NaN	0.983	0.929
GM	0.982	NaN	0.587	NaN	0.998	0.974	0.964	0.991	NaN	NaN	NaN	0.971	0.986	NaN
IM	NaN	0.988	NaN	0.976	NaN	0.946	0.955	0.998	0.974	0.998	NaN	0.995	0.997	0.992
LGD	NaN	NaN	NaN	NaN	0.958	NaN	NaN	NaN	0.941	0.989	NaN	0.993	NaN	NaN
HGD	0.935	0.999	NaN	0.998	0.946	0.943	NaN	NaN	NaN	0.999	0.995	0.996	0.998	NaN
IMC	0.99	0.994	NaN	0.998	NaN	0.998	0.999	NaN	NaN	0.985	0.996	0.989	NaN	NaN

Note: Reference disease is picked to be with the lowest risk appeared in the patients' tissues

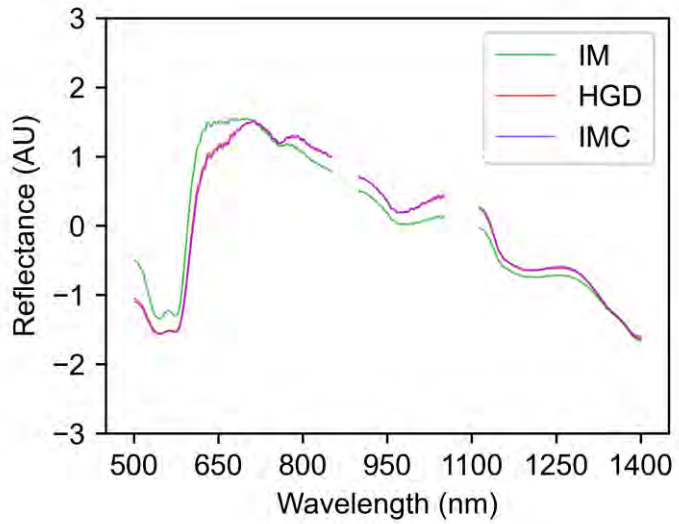
ExVISION Patient 60



SAM (Ref: SQ)
GM:0.357
HGD:0.648
IMC:0.255

Pearson
GM:0.937
HGD:0.797
IMC:0.968

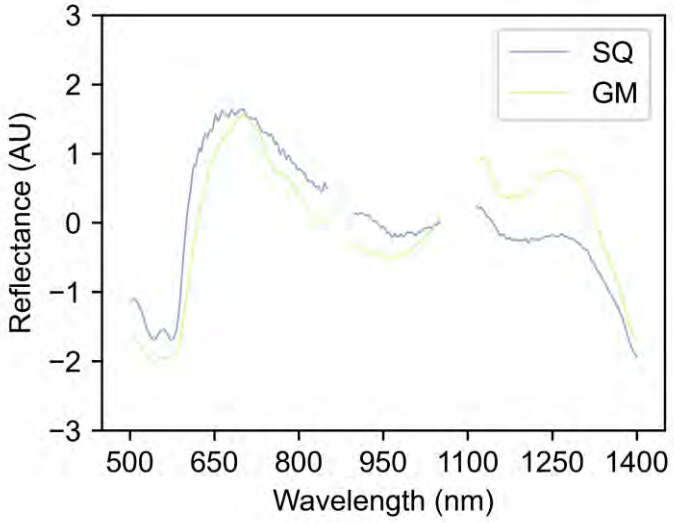
ExVISION Patient 61



SAM (Ref: IM)
HGD: 0.240
IMC: 0.260

Pearson
HGD: 0.971
IMC: 0.966

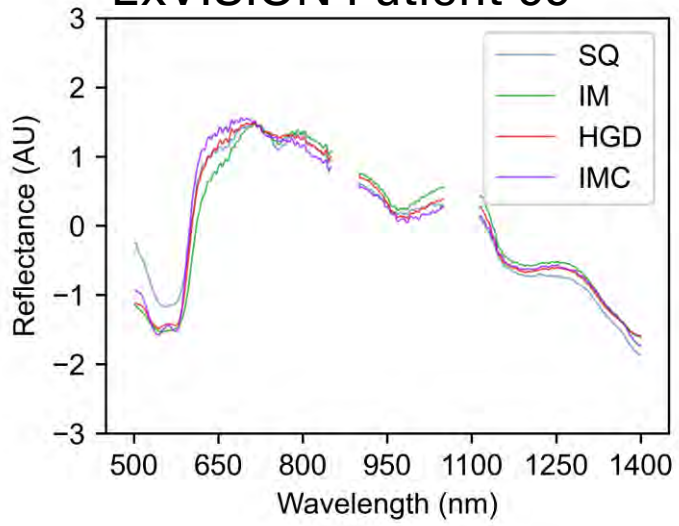
ExVISION Patient 62



SAM (Ref: SQ)
GM: 0.591

Pearson
GM: 0.830

ExVISION Patient 66



SAM (Ref: SQ)
IM: 0.271
HGD: 0.188
IMC: 0.201

Pearson
IM: 0.963
HGD: 0.982
IMC: 0.980

Supervised Learning Pipeline

Linear Discriminant Analysis (LDA)

Labels are used, so that we can distinguish data of different diseases by maximising the distance between each clusters.

Use processed low-dim data to move on.

Feature Selection

Available methods include:

- Boruta (recommended by Imperial College)
- Spectral Band Selection

Surprisingly, not helpful

Training the Model

Available methods include:

- XGBoost
- LightGBM

Test

To avoid overfitting, cross validation is used to separate the training and test sets.
Confusion matrices and receiver operating characteristics are used to measure the performance of the model.