Identification of Human Lymphocytes Using Optical Diffraction Tomography and Machine Learning

FAJAR WIMAR RAMADHAN

Outline

1. Introduction: Why lymphocytes and their classification are important
2. Method: 3D QPI & Machine Learning
3. Result and Discussion
4. Summary

Introduction

Introduction :

Understanding the role and Identify Lymphocytes are Important!

- The lymphocyte population are tightly regulated to defend the host.
- Disturbances in lymphocyte are related to various diseases such as cancers, autoimmune diseases and virus infections.
- If we can know the composition of the spesific lymphocytes cells in our body, we can diagnose some disease.

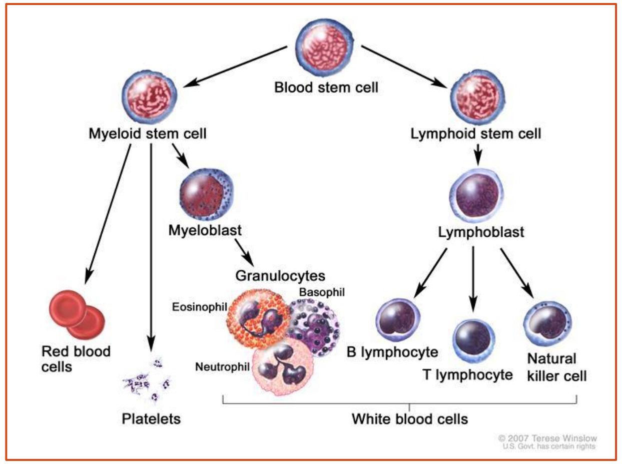


Fig. Blood cell type.

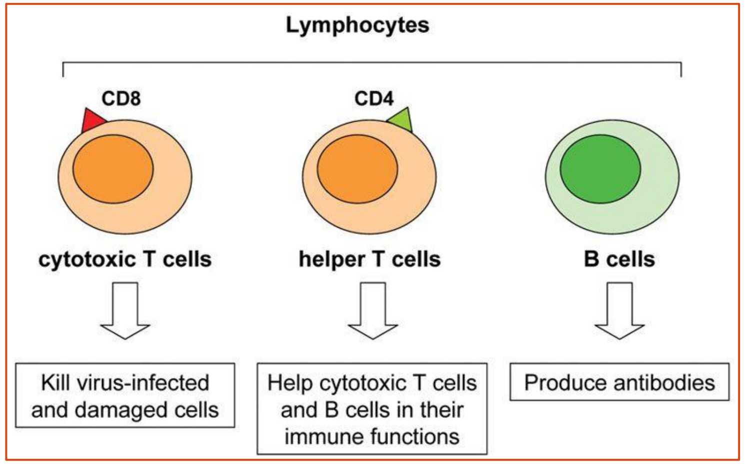


Fig. Lymphocytes cell type.

Alizadeh, A. A. *et al. Nature* **403,** 503 (2000).
Granular, T. L., Leukemia, L. & Disorders, R. ncologist. 247–258 (2004).
de Visser, K. E., Eichten, A. & Coussens, L. M. *Nat. Rev. Cancer* **6,** 24 (2006).
Ueda, H. *et al. Nature* **423,** 506 (2003).
von Boehmer, H. & Melchers, F. *Nat. Immunol.* **11,** 14 (2009).
Saez-Cirion, A. *et al. Proc. Natl. Acad. Sci.* **104,** 6776–6781 (2007).

Method

Method : Our Approach

Quantitative Phase Imaging :
Optical Diffraction Tomography

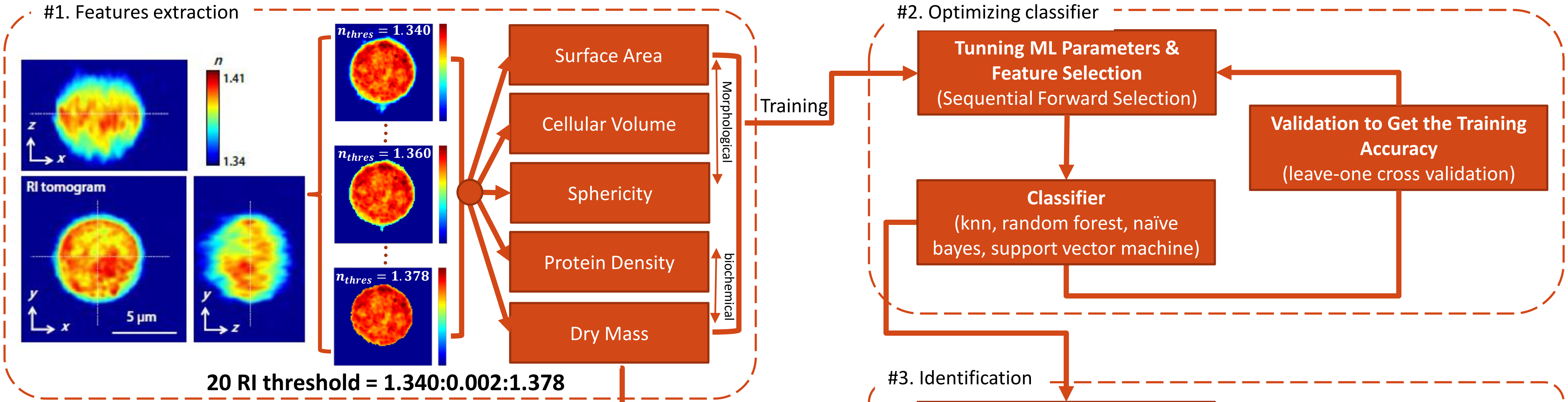
3D refractive index tomogram

Machine learning

Label-free Identification of WBCs.
We aim to classify CD4, naïve CD8, and memory CD8 cells.

Fig. Concept of our research goal.

Method : What I actually do ?
Label-free Lymphocyte Identification



Protein density

$n = n_0 + \alpha C$

α is the refractive index increment = 0.2mL/g.

Sphericity

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V_p)^{\frac{2}{3}}}{A_p} = 1\ round\ shape.$$

Fig. Label-free lymphocyte identification.

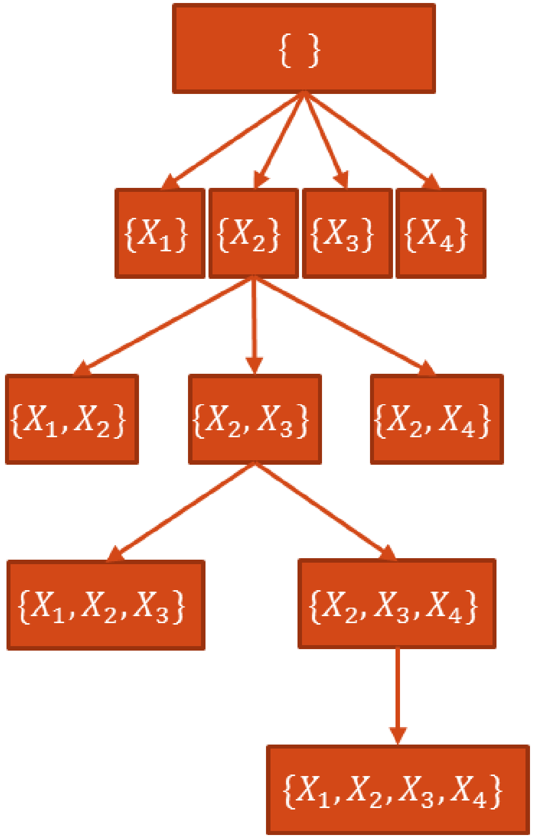
J. Yoon* & Y.J. Jo* et al. *Scientific Reports* (2017)

Method : Sequential Forward Selection (SFS)

For example we have n = 4 features.

n : number of total features

k : number of selected features



Step 0 (k=0)
Max acc. ?? → $\{Null\}$

Step 1 (k=1)
Max acc. ?? → $\{\boldsymbol{X_2}\}$

Step 2 (k=2)
Max acc. ?? → $\{\boldsymbol{X_2}, X_3\}$

Step 3 (k=3)
Max acc. ?? → $\{\boldsymbol{X_2}, \boldsymbol{X_3}, X_4\}$

Step 4 (k=4)
Max acc. ?? → $\{X_1, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}\}$

Step 5
Max acc. ?? → $\{\boldsymbol{X_2}, X_3\}$

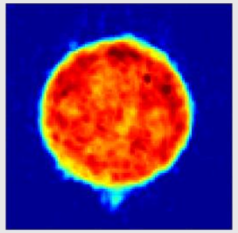
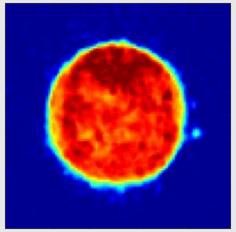
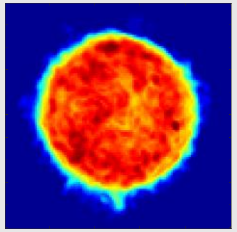
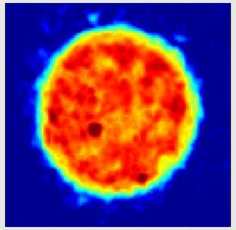
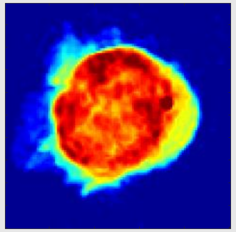
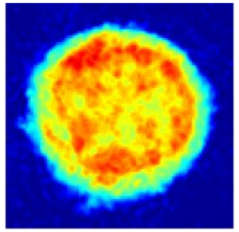
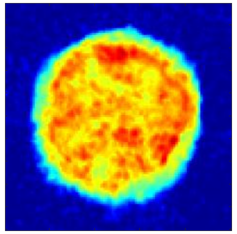
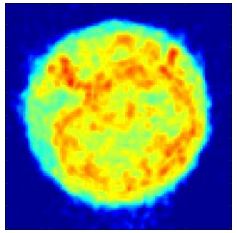
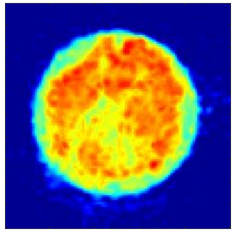
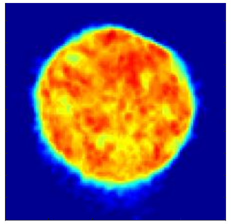
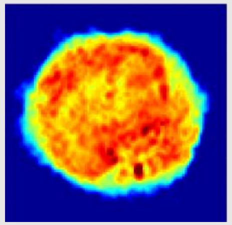
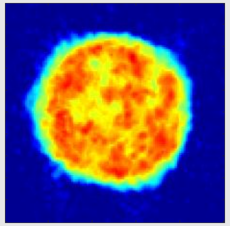
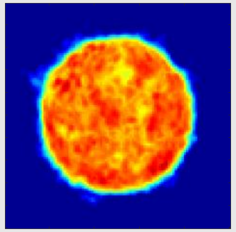
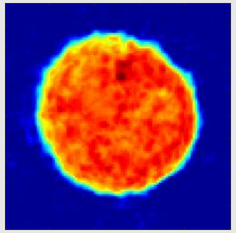
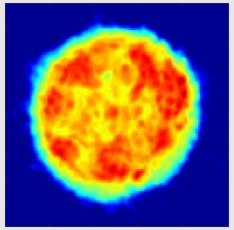
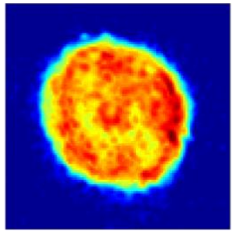
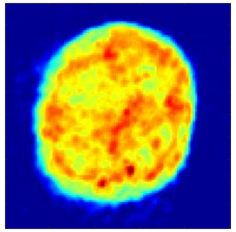
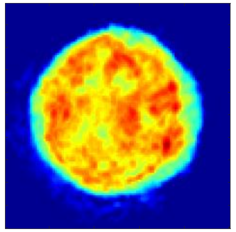
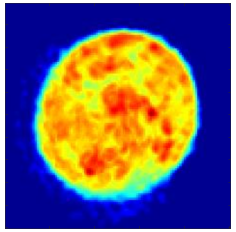
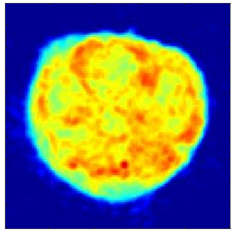
Fig. Sequential forward selection example algorithm.

Tab. Comparison number of all combination vs SFS combination.

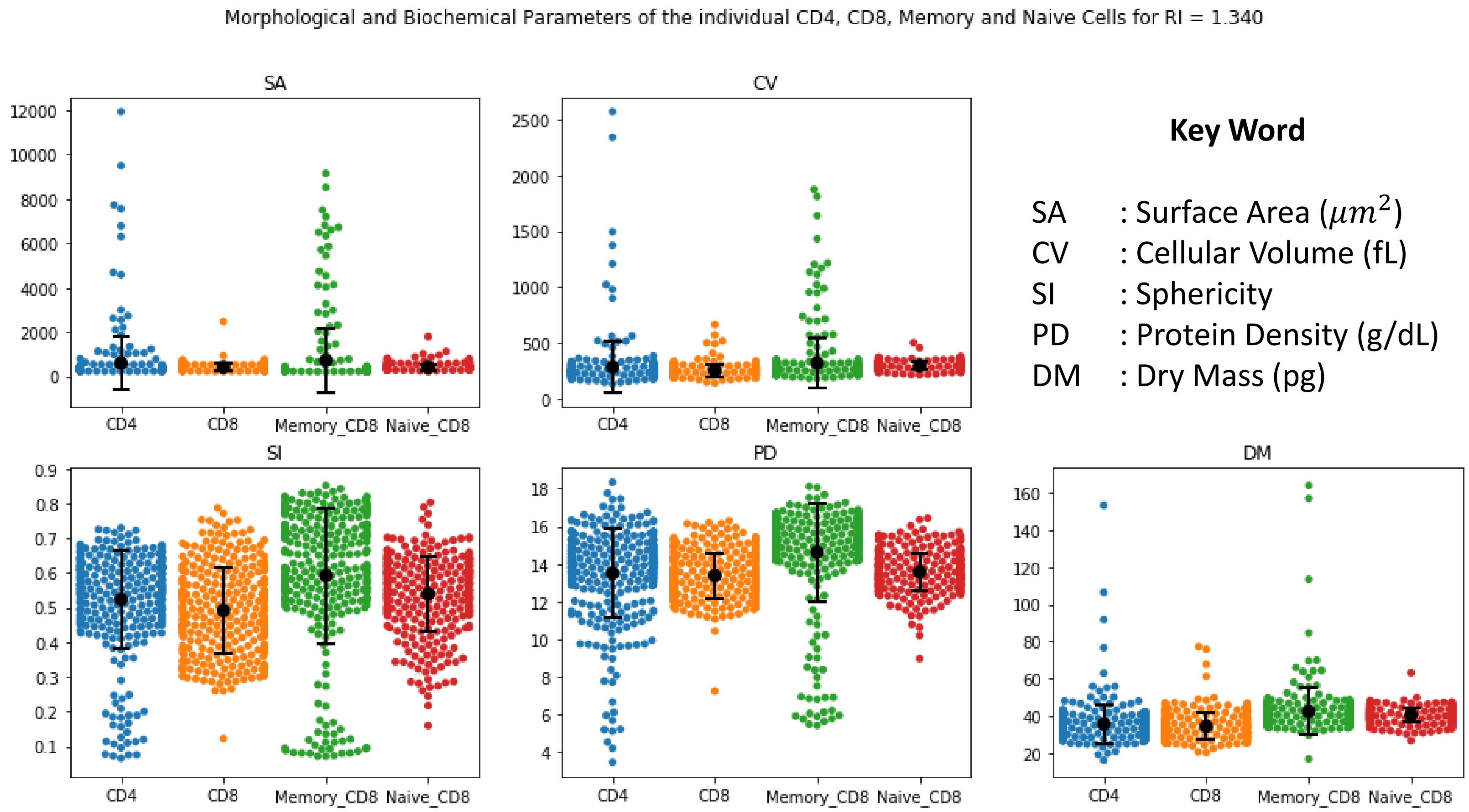
n	Number of all combination	Number of SFS Combination
4	16	11
100	1.2676506 × 10^30 (times = ??)	5,501 (times~3 hours)
n	2^n	$1 + \frac{n(n+1)}{2}$

A. Müller and S. Guido, *Introduction to Machine Learning with Python*. 2016.

Tab. Cross sectional cells.

Cell type	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
CD4					
CD8					
Memory CD8					
Naïve CD8					

Method :
The Example of Extracted Features



Key Word

SA : Surface Area (μm^2)
CV : Cellular Volume (fL)
SI : Sphericity
PD : Protein Density (g/dL)
DM : Dry Mass (pg)

Fig. Extracted features of individual CD4, CD8, Memory and Naïve CD8 cells for RI = 1.340.

Method :
K-Nearest Neighbors (k-NN) Algorithm

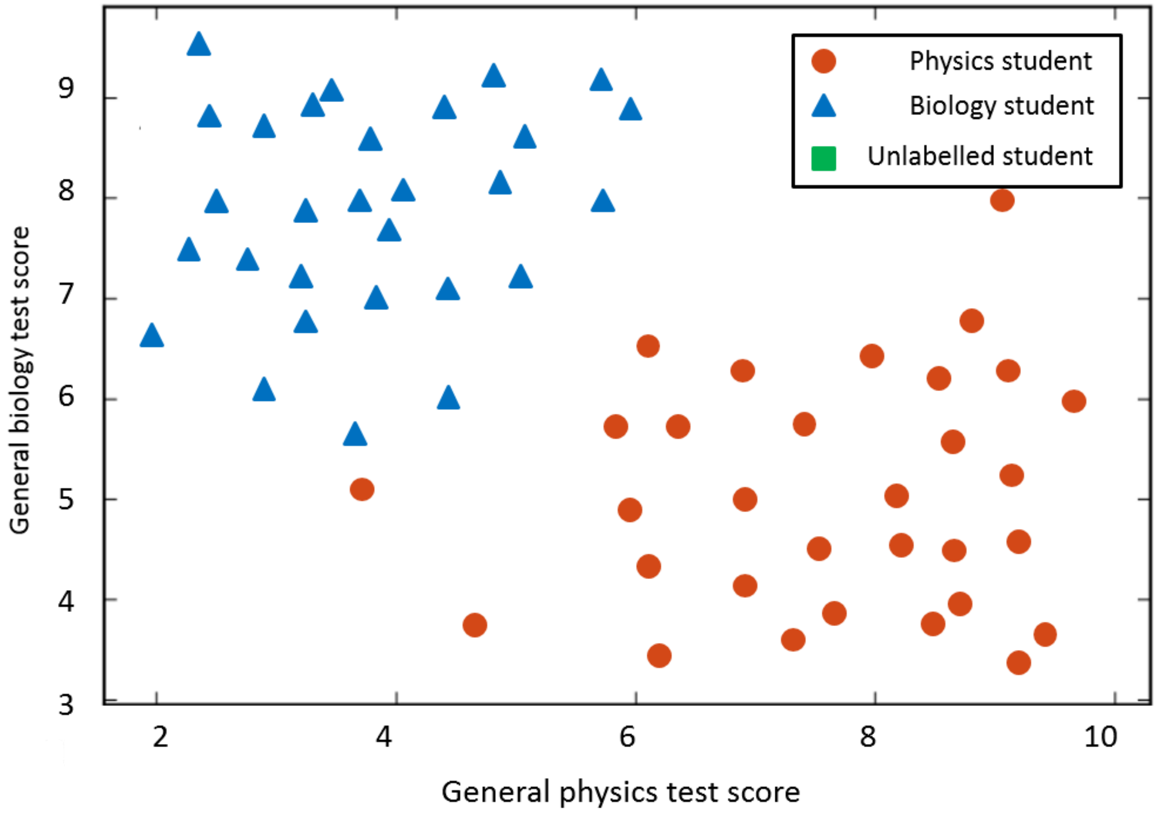


Fig. The example of knn method.

A. Müller and S. Guido, *Introduction to Machine Learning with Python*. 2016.

Method :
Random Forest Algorithm

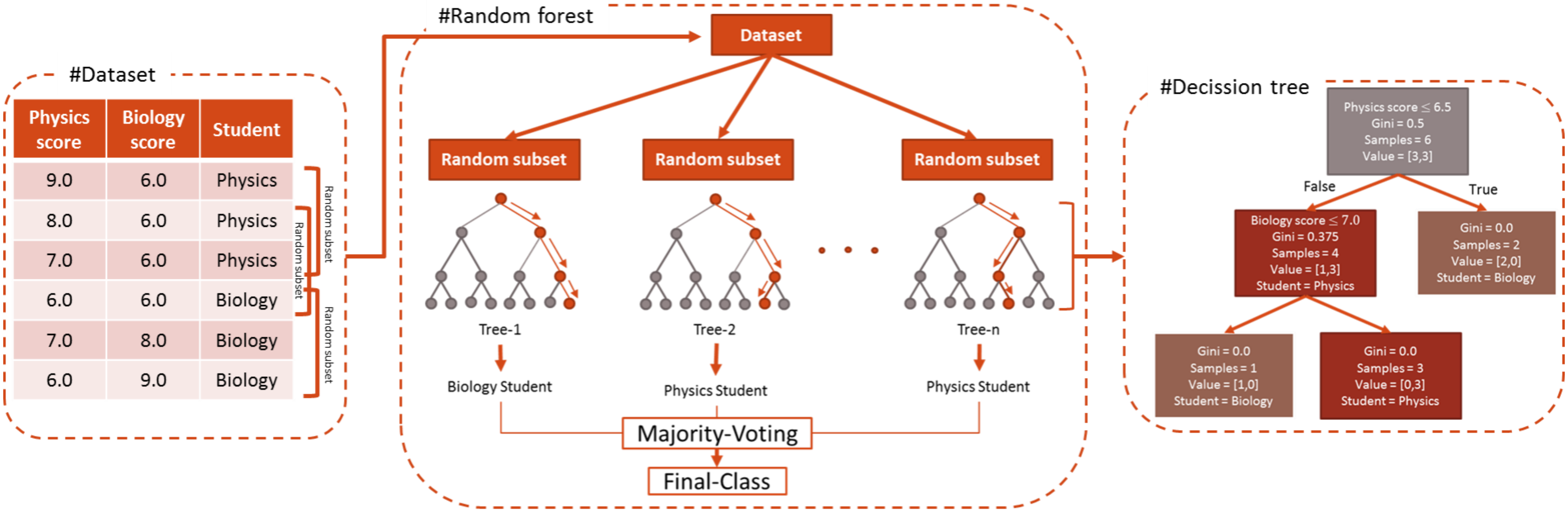


Fig. Random forest algorithm.

J. Brownlee, *Master machine learning algorithms*. 2016.

Method : Naïve Bayes Algorithm

#Dataset

Physics score (PS)	Biology score (BS)	Student (S)
9.0	5.0	Physics (P)
8.0	6.0	Physics (P)
7.0	6.0	Physics (P)
6.0	6.0	Biology (B)
7.0	8.0	Biology (B)
6.0	9.0	Biology (B)

#Prediction

Physics score (PS)	Biology score (BS)	Student (S)
8.5	6.5	??

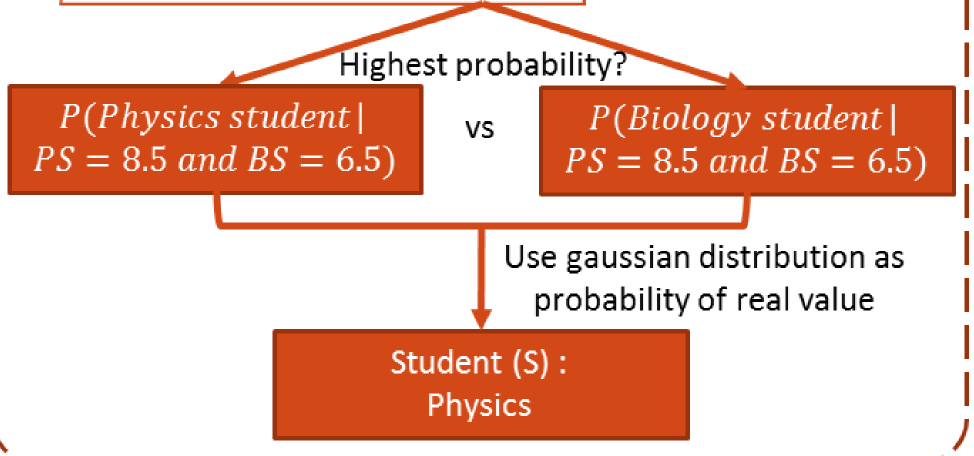


Fig. Naïve bayes algotirhm.

J. Brownlee, *Master machine learning algorithms*. 2016.

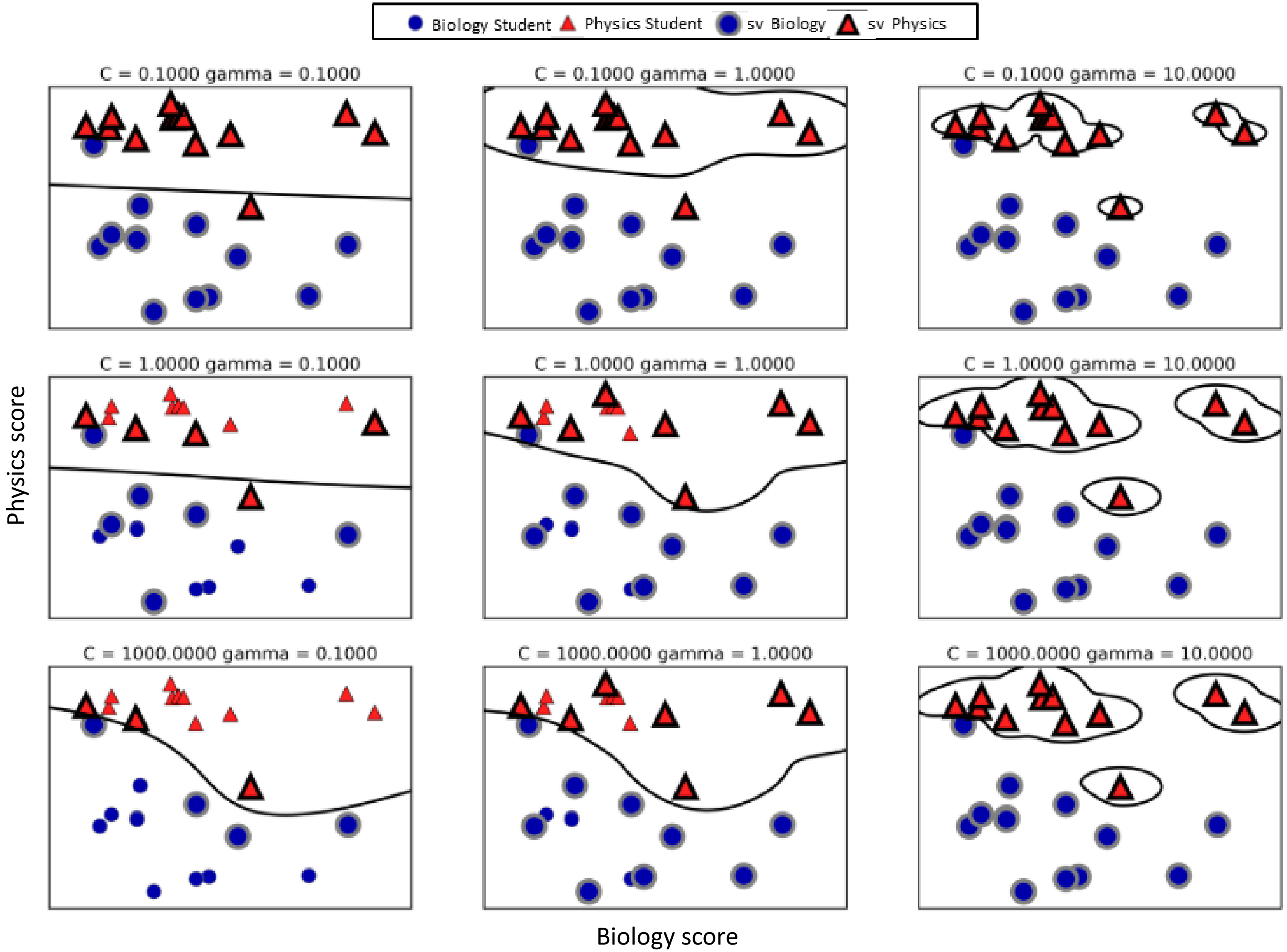


Fig. Support vector machine algorithm.

Method : Support Vector Machine

<u>Radial Kernel</u>

- The larger **gamma** parameters, the smaller data radius of influence.
- The larger the **C** value, the more sensitive the algorithm to the data.

A. Müller and S. Guido, *Introduction to Machine Learning with Python*. 2016.

Result and Discussion : Data Distribution

Result and Discussion : Data Distribution of CD4 and CD8

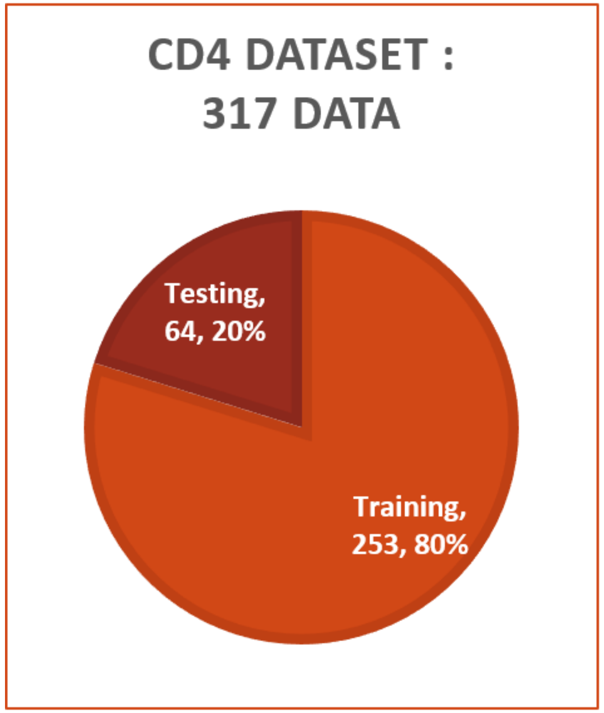


Fig. CD4 dataset

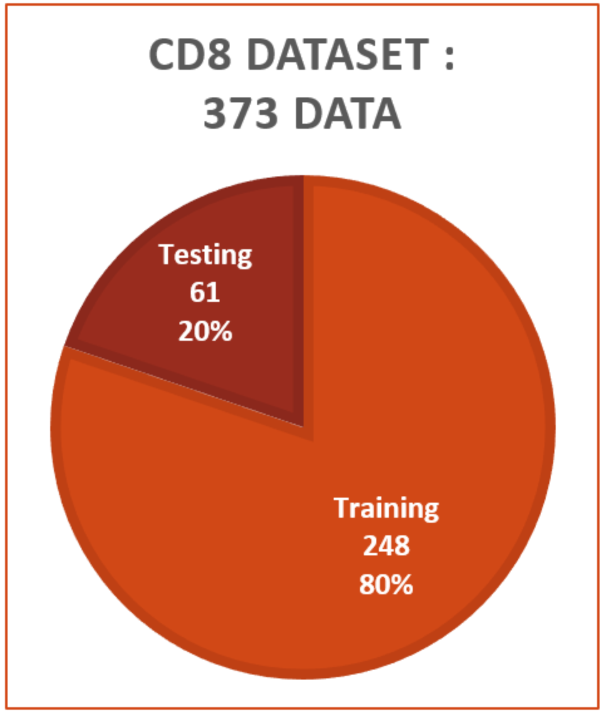


Fig. CD8 dataset

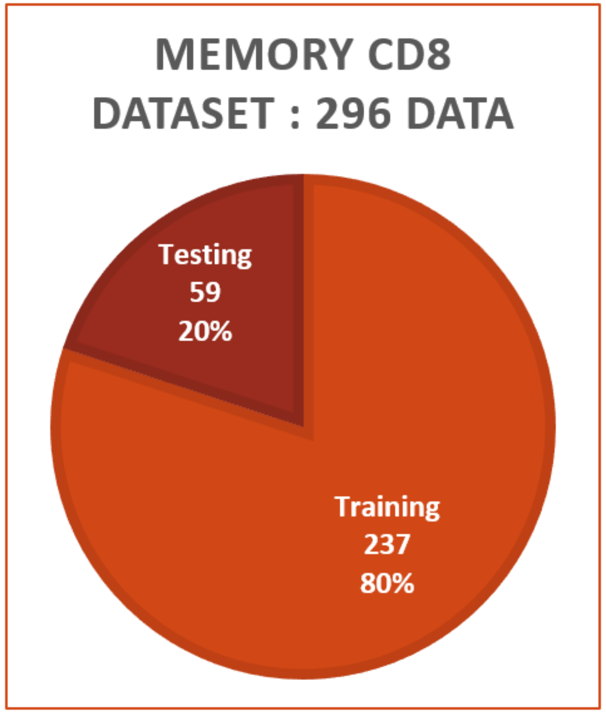


Fig. Memory CD8 dataset

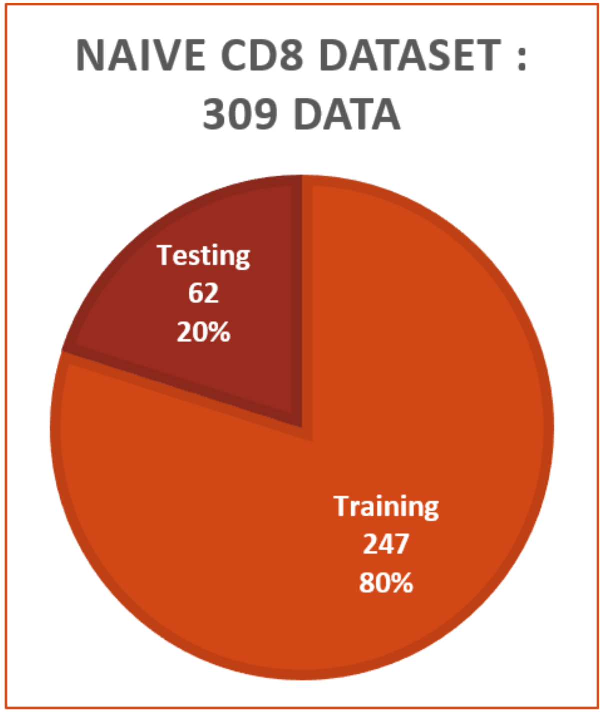


Fig. Naïve CD8 dataset

Result and Discussion : KNN Classifier

Result and Discussion (KNN): Selecting k Value

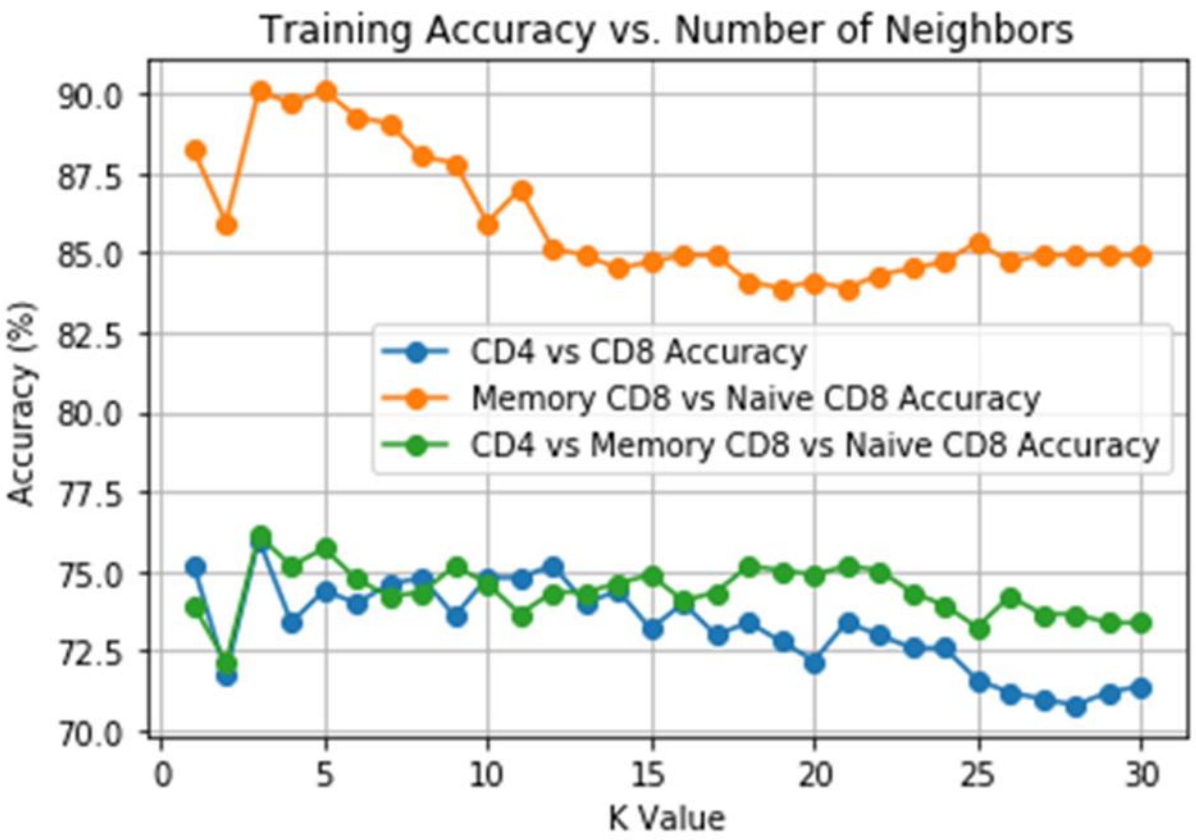


Fig. Accuracy vs. number of neighbors.

Selected k = 3

- **Case A : CD4 vs CD8**
  Accuracy = 76.00%.
- **Case B : Memory CD8 vs Naïve CD8**
  Accuracy = 90.08%
- **Case C : CD4 vs Memory CD8 vs Naive CD8**
  Accuracy = 76.13%

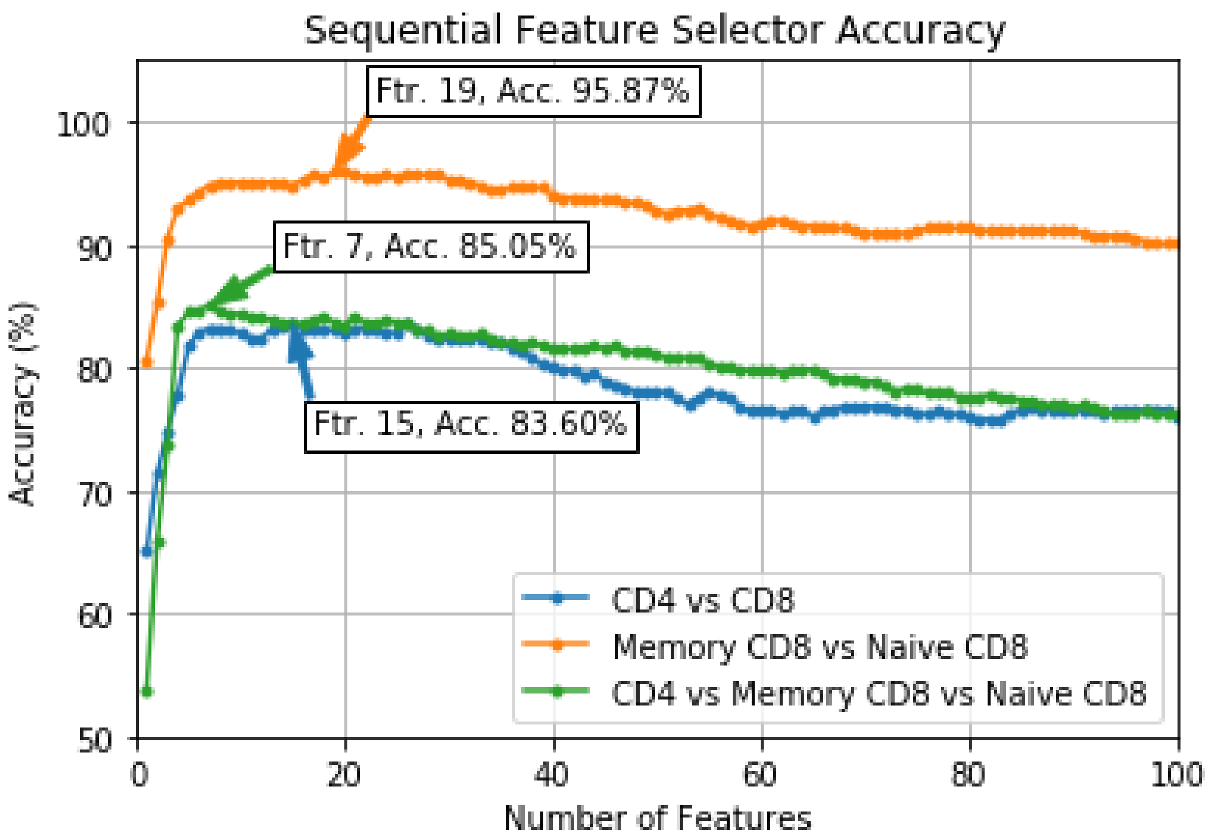


Fig. Feature selection in KNN.

Result and Discussion (KNN): Selected Features

- **Case A : CD4 vs CD8**
  Training Acc. = **83.60%,** best combination :
  ('**SA** 1.340', 'SA 1.342', 'SA 1.344', 'SA 1.346', '**SI** 1.342', 'SI 1.344', 'SI 1.370', 'SI 1.372', 'SI 1.374', '**PD** 1.346', 'PD 1.354', 'PD 1.360', 'PD 1.362', 'PD 1.364', 'PD 1.378')
- **Case B : Memory CD8 vs Naïve CD8**
  Training Acc. = **95.87%**, best combination :
  ('**SA** 1.340', 'SA 1.342', 'SA 1.344', 'SA 1.346', 'SA 1.348', '**CV** 1.342', '**SI** 1.340', 'SI 1.342', 'SI 1.344', 'SI 1.348', 'SI 1.362', 'SI 1.364', 'SI 1.374', 'SI 1.376', '**PD** 1.340', 'PD 1.364', 'PD 1.372', 'PD 1.374', 'PD 1.376')
- **Case C : CD4 vs Memory CD8 vs Naive CD8**
  Training Acc. = **85.05%**, best combination :
  ('**CV** 1.376', '**SI** 1.340', 'SI 1.374', '**PD** 1.340', 'PD 1.376', 'PD 1.378', '**DM** 1.368')

Keyword

SA	: Surface Area (μm^2)	SI	: Sphericity	DM	: Dry Mass (pg)
CV	: Cellular Volume (fL)	PD	: Protein Density (g/dL)	X.XXX	: RI value

Note :
This selected features will be used for another Machine Learning

Result and Discussion (KNN) : Training and Testing for CD4 vs Memory CD8 vs Naïve CD8

Training

Accuracy = 85.05%		Output cell type (number of cell)			Sensitifity (%)
		CD4 cell	Memory CD4 cell	Naïve CD8 cell	
Input cell type	CD4 cell	198	33	14	80.82
	Memory CD48 cell	25	197	15	83.12
	Naïve CD8 cell	13	9	225	91.09
Specificity (%)		83.90	82.43	88.58	85.05

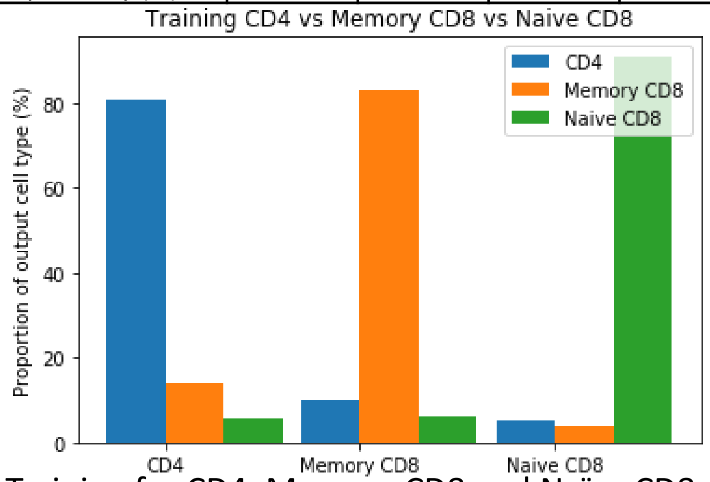


Fig. Training for CD4, Memory CD8 and Naïve CD8 cells.

Testing

Accuracy = 83.06%		Output cell type (number of cell)			Sensitifity (%)
		CD4 cell	Memory CD4 cell	Naïve CD8 cell	
Input cell type	CD4 cell	51	9	2	82.26
	Memory CD48 cell	9	45	5	76.27
	Naïve CD8 cell	5	1	56	90.32
Specificity (%)		78.46	81.82	88.89	83.06

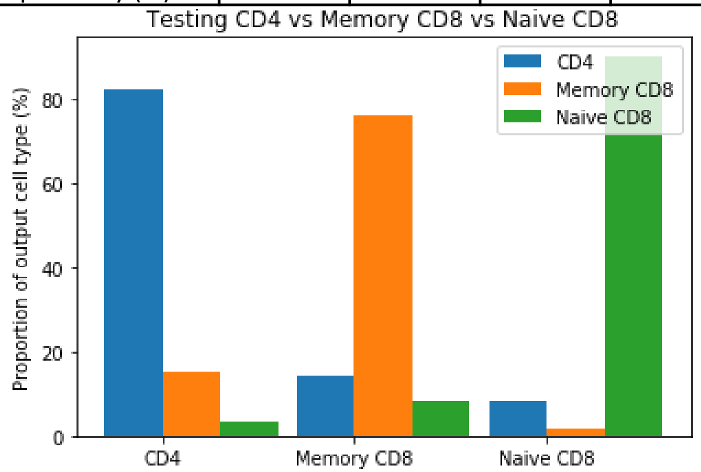


Fig. Training for CD4, Memory CD8 and Naïve CD8 cells.

Result and Discussion : Parameter tunning another machine learning

Result and Discussion : Random Forest, Selecting Number of Tree

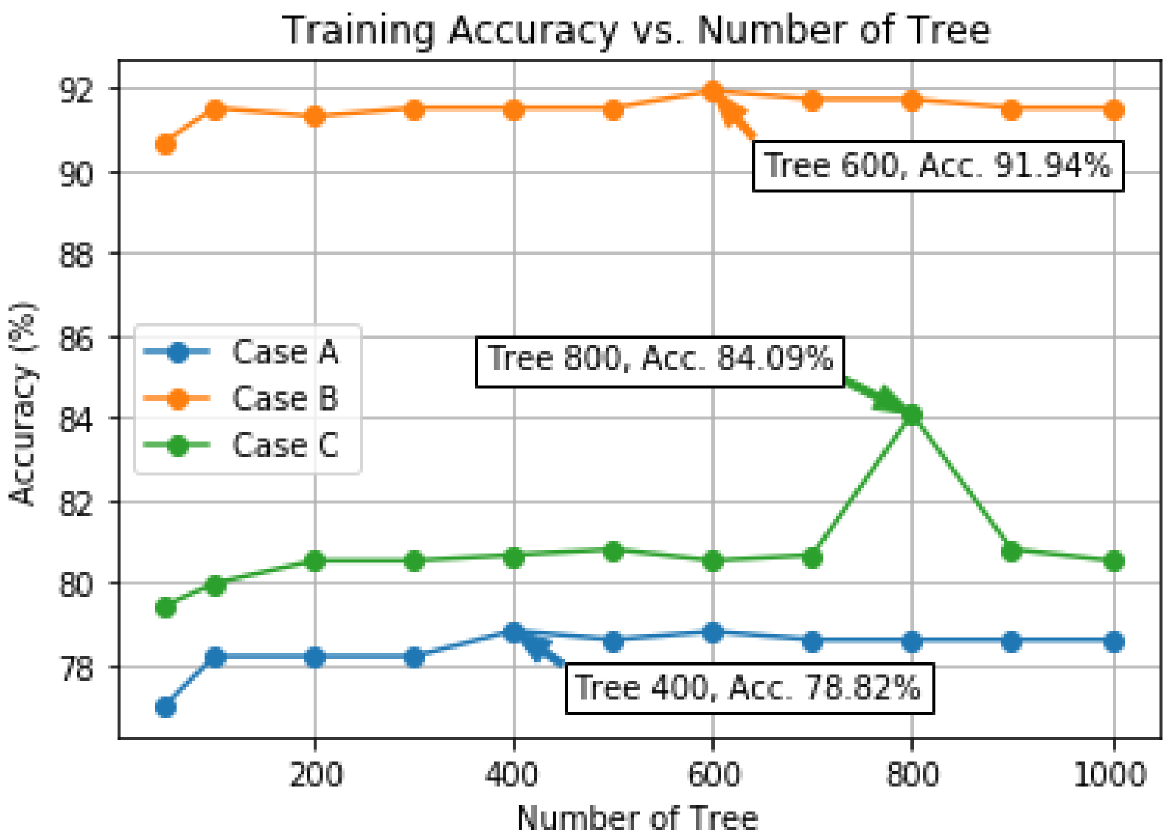


Fig. Selecting Number of Tree.

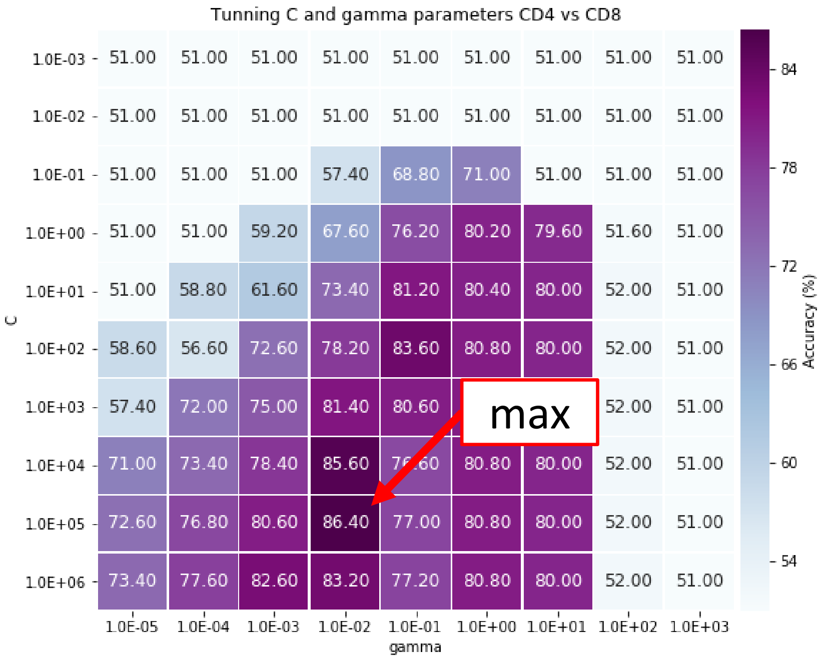
Selected Number of Tree

- **Case A : CD4 vs CD8**
  Number of Tree = 400
  Accuracy = 78.82%.
- **Case B : Memory CD8 vs Naïve CD8**
  Number of Tree = 600
  Accuracy = 91.74%
- **Case C : CD4 vs Memory CD8 vs Naive CD8**
  Number of Tree = 800
  Accuracy = 80.93%

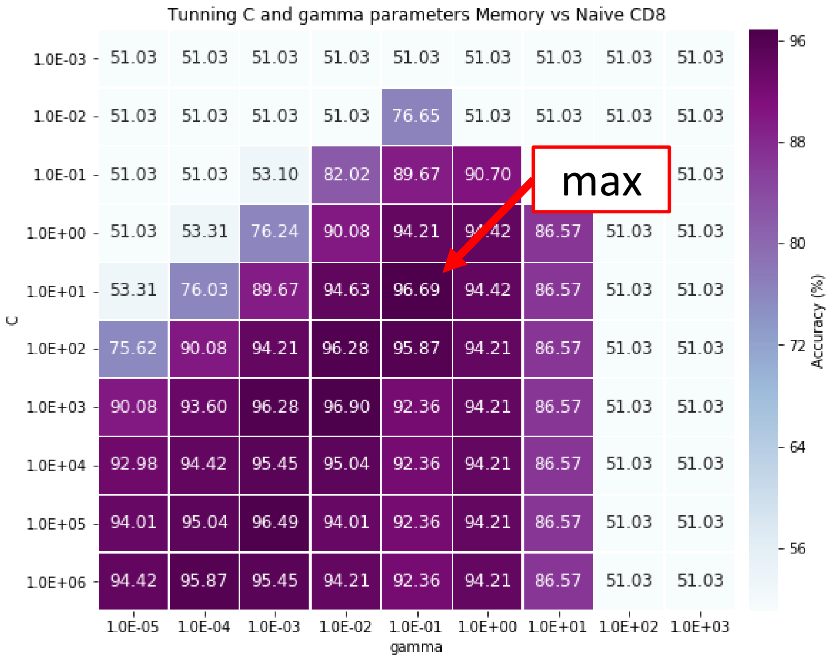
Result and Discussion : Naïve Bayes Classifier

Note:
Have no hyper-parameters

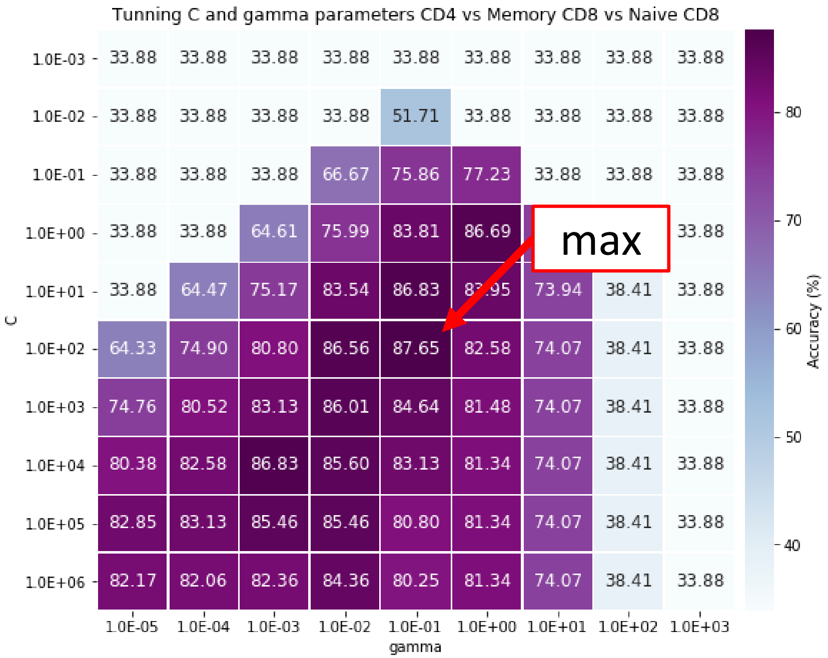
Result and Discussion : Support Vector Machine, Selecting C and γ



$C = 10^5, \gamma = 10^{-2}$



$C = 10^1, \gamma = 10^{-1}$



$C = 10^2, \gamma = 10^{-1}$

Fig. Hyper parameter in SVM.

Result and Discussion : Comparison of all classifier methods

Result and Discussion : Comparison of all classifier methods

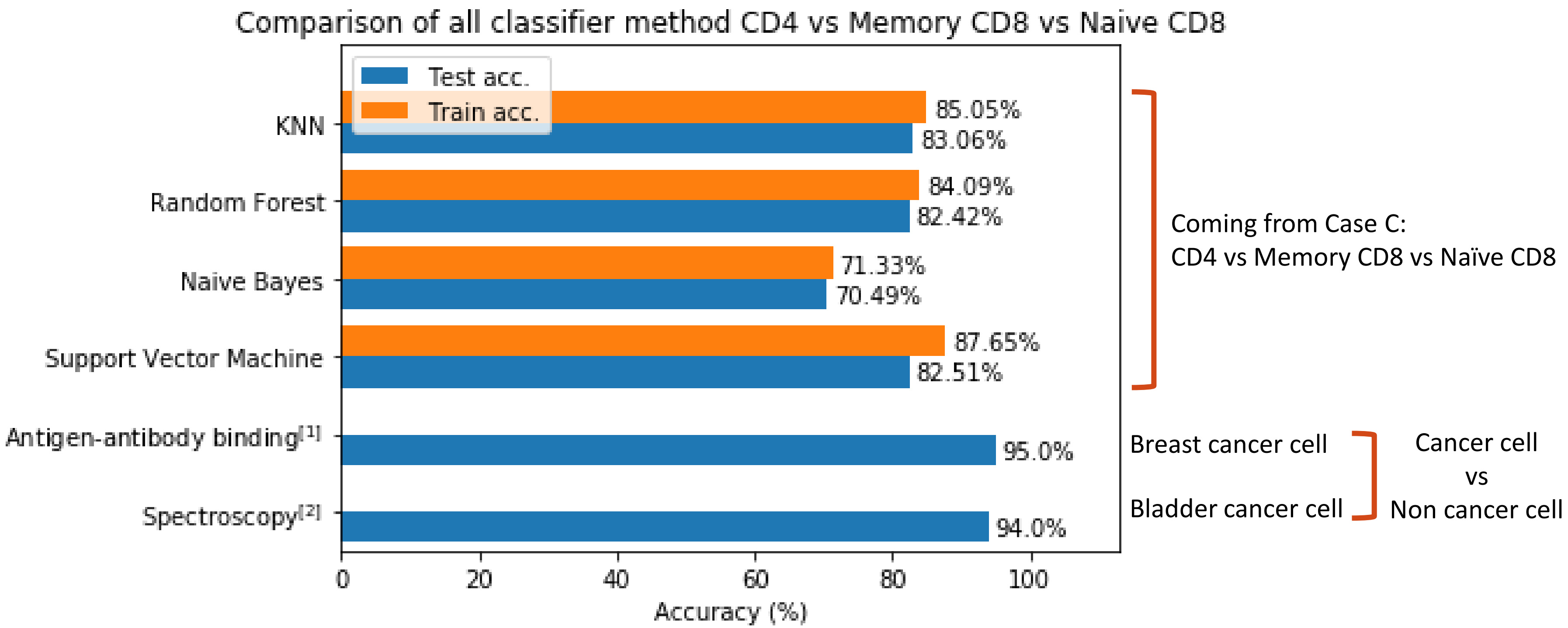


Fig. Comparison of all classifier.

[1] Ma, Z., Zhou, Y., Collins, D. J. & Ai, Y. *Lab Chip* **17,** 3176–3185 (2017).
[2] Draga, R. O. P. *et al. Anal. Chem.* **82,** 5993–5999 (2010).

Summary

Summary

Summary

- White blood cells are important because these cells defending the our body against harmful invaders in our immune system and here we want to classify three lymphocyte cell types (CD4, Naïve and Memory CD8).
- We can diagnose some disease if we can identify lymphocyte cells.
- ODT with machine learning enable identification of lymphocyte cell with testing accuracy up to 95.04% and comparable even more than conventional methods (depend on the testing cell).

Acknowledgements

KAIST

-

Biomedical Optics Laboratory

Prof. YongKeun Park

TA: Donghun Ryu

ITB

-

Department of Physics, Bandung Institute of Technology

Thank you