NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Developing AI model for cell status classification using Raman acquired data

Amulya Mathur

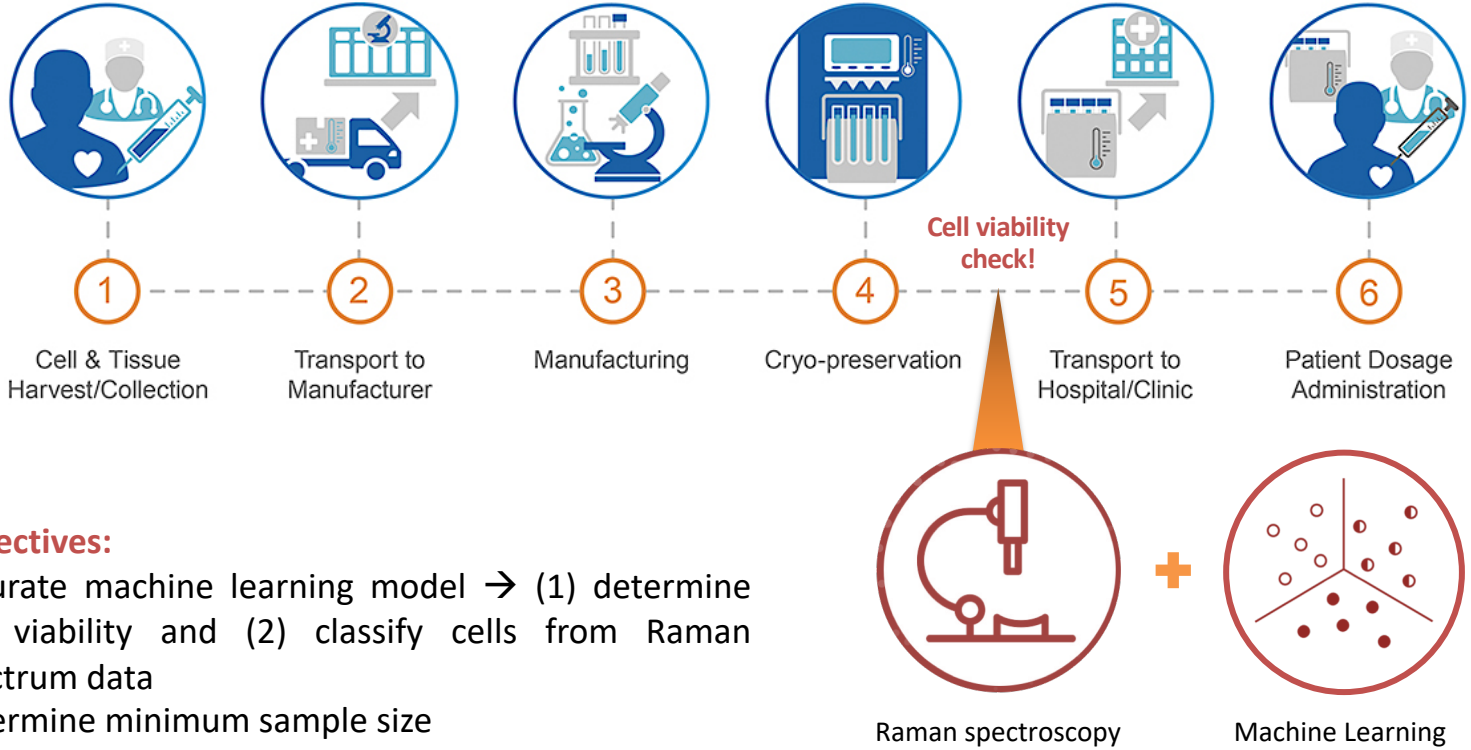
School of Chemistry, Chemical Engineering & Biotechnology

8th May 2023

2022-23 BG4801 FYP

Background

Cell therapy:

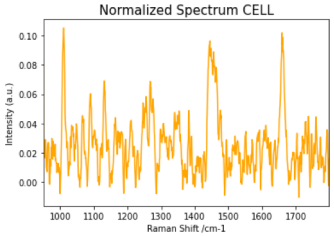


FYP Objectives:

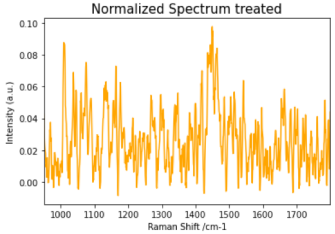
- Accurate machine learning model → (1) determine cell viability and (2) classify cells from Raman spectrum data
- Determine minimum sample size

Machine learning approach 1: Cell Viability

1. Collect Raman spectrums and preprocess data (6 different processing techniques)
[2200 untreated hMSCs]



4. Collect Raman spectrums and preprocess data (6 different processing techniques)
[200 PG-treated hMSCs]



2000 untreated hMSCs

200 untreated hMSCs

2. Train model

3. Validate model

Train & Validate

One Class SVM Model

2 different models (next slide)

Test

200 PG-treated hMSCs

5. Test model

6. Classify

Non-viable (Anomaly)

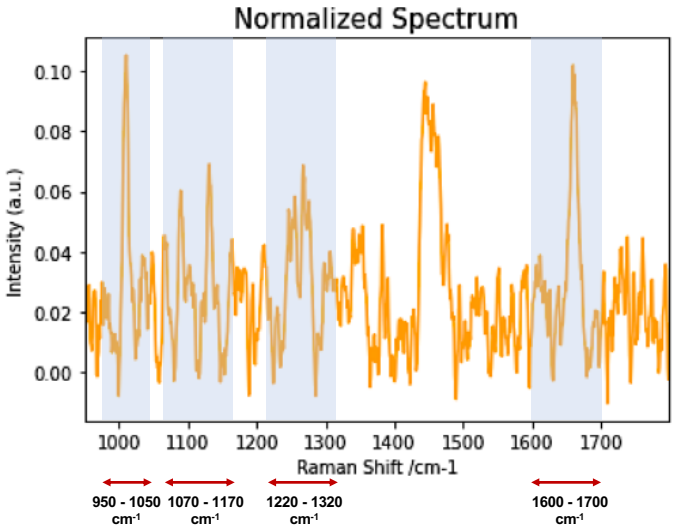
Viable (Normal)

Determine Cell viability

$$= \frac{\text{Viable}}{\text{Viable + Non-viable}} \times 100\ \%$$

Machine learning approach 1: Cell Viability

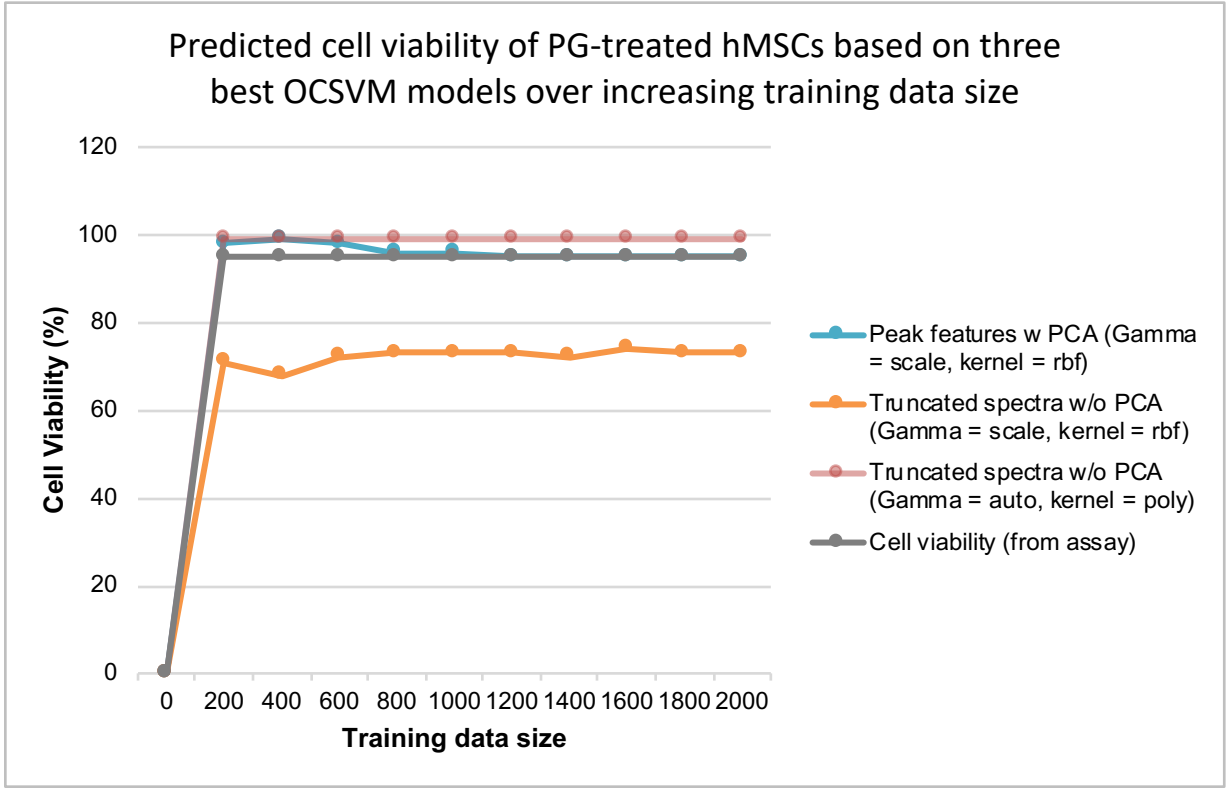
- **6 different Raman spectrum processing methods:**
  - **Full spectrum (950 – 1800 cm^{-1})**
  - **Full spectrum + PCA**
  - **Truncated spectrum:** 4 characteristic peak wavelengths (1010 cm^{-1}, 1096 cm^{-1}, 1257 cm^{-1}, and 1656 cm^{-1})
  - **Truncated spectrum + PCA**
  - **Four peaks data extraction:** Raman shift, peak height, and FWHM of the same 4 characteristic peak wavelengths
  - **Four peaks data + PCA**
- Including one raw spectrum



- **2 OCSVM models:**

Model type	Parameters used
OCSVM Model 1	**kernel = 'rbf', gamma = 'scale',** degree = 3, coef0 = 0, tol = 0.001, nu = 0.5
OCSVM Model 2	**kernel = 'poly', gamma = 'auto',** degree = 3, coef0 = 0, tol = 0.001, nu = 0.5

Results and observations 1: Cell Viability

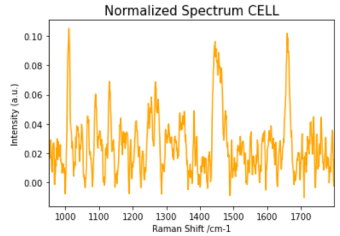


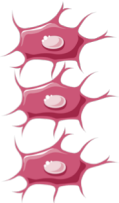
- ✓ **Best result:**
  - **Model:** OCSVM (Gamma = 'scale', kernel = 'rbf')
  - **Data processing technique:** 4 peaks data extraction + PCA

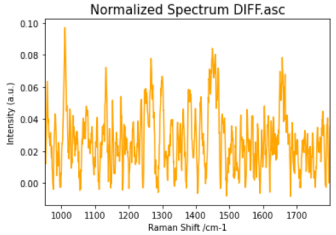
- ✓ Model most accurate with targeted peak extraction → precise information

Machine learning approach 2: Cell Classification

1. Collect Raman spectrums and preprocess data
[110 untreated hMSCs]







2. Collect Raman spectrums and preprocess data
[110 differentiated osteocytes]

- Raw
- Full spectrum
- Full spectrum + PCA

3. Split data

80 %
Training data

20 %
Testing data

Train & Test

1. KNN
2. NB
3. SVM
4. RF
5. XGB
6. CNN

4. Train each model to classify between 2 cell types

Classify

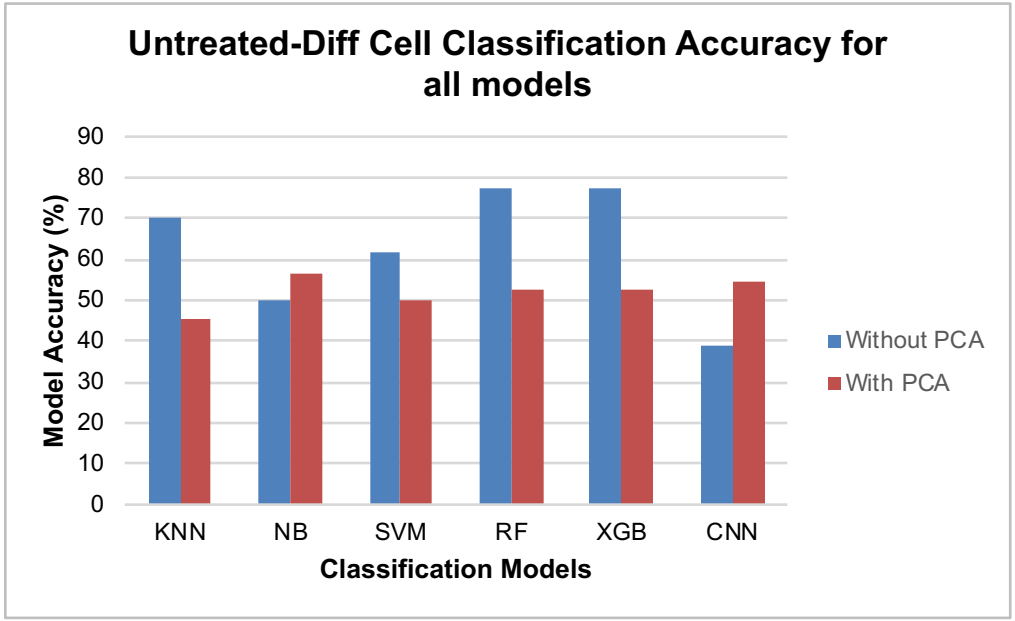
Untreated hMSCs

Differentiated osteocytes

Compare accuracies

5. Test all 6 models and select the most accurate one

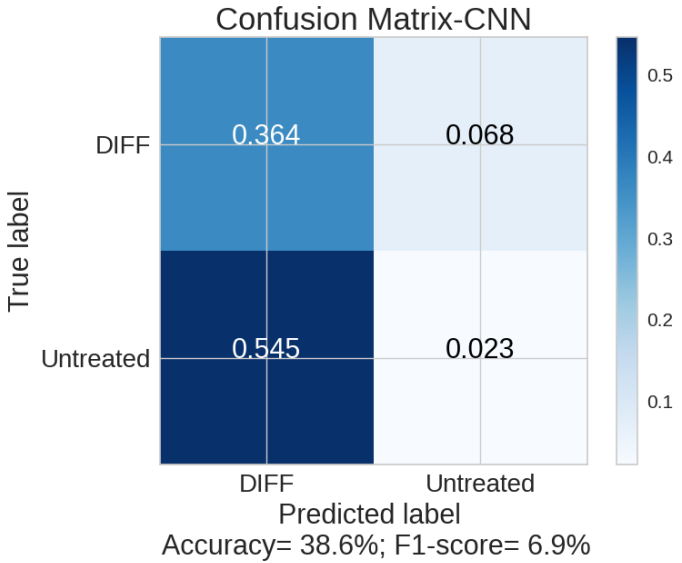
Results and observations 2: Cell Classification

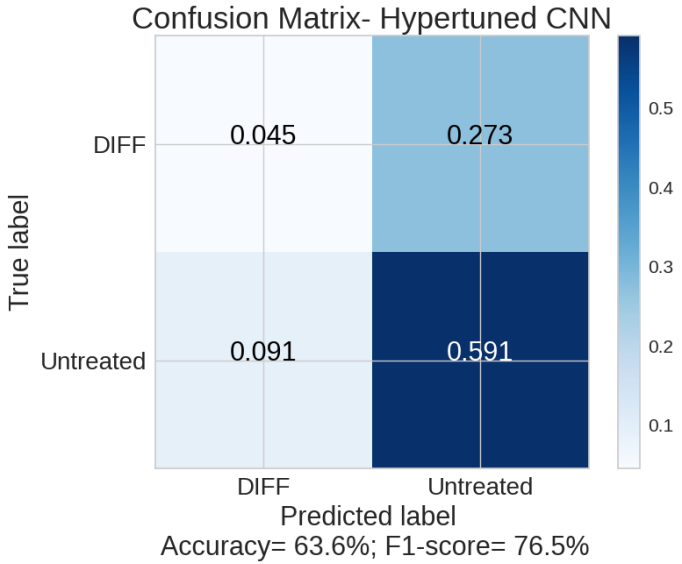


- ✓ **Best result:**
  - **Model:** Random Forest and XGBoost (77.3 %)
  - **Data processing technique:** Full spectrum + PCA
- ✓ Ensemble tree-based models → Bagging & Boosting methods

Results and observations 2: Cell Classification (CNN)

CNN model constructed using pre-processed data (w/o PCA)





Effect of Hyper-parameter tuning → higher classification accuracy

Errors, Conclusion & Future works

Sources of Error

- Non light-tight optical set-up → Noise effect
- Inconsistent data collection → data variation

Conclusion

- OCSVM, RF, and XGB best models
- Need complex models
- Need more training data

✓**Machine learning significant in Raman spectroscopy analysis**

Future Works

- Improvements:
  - Hyper-parameter tuning of CNN model
  - Data augmentation
  - Stratified k-fold sampling
  - Better Experimental set-up
- One-step Raman spectrum pre-processing: Autoencoders