Genetic Algorithm based Semi-supervised Convolutional Neural Network for Real-time Monitoring of Escherichia Coli Fermentation of Recombinant Protein Production Using a Raman Sensor

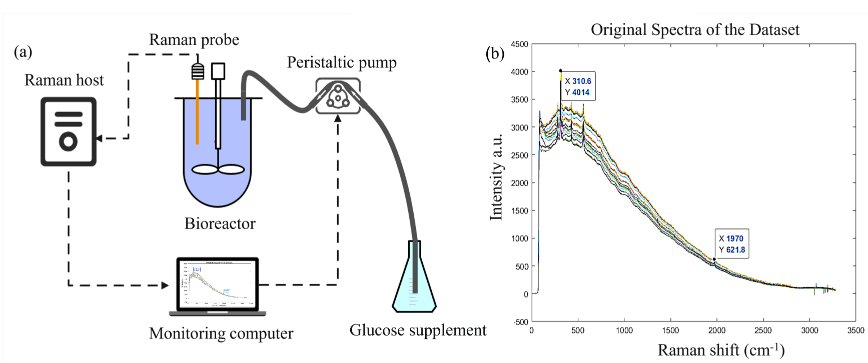
Zhenguo WEN[1], Yuan Liu[1], Xiaotian Zhou[1], Teng Wang[1], An Luo[1], Zhaojun Jia[1], Xingquan Pan[1], Weiqi Cai[1], Mengge Sun[1], Xuezhong Wang[1], and Guangzheng Zhou[1]

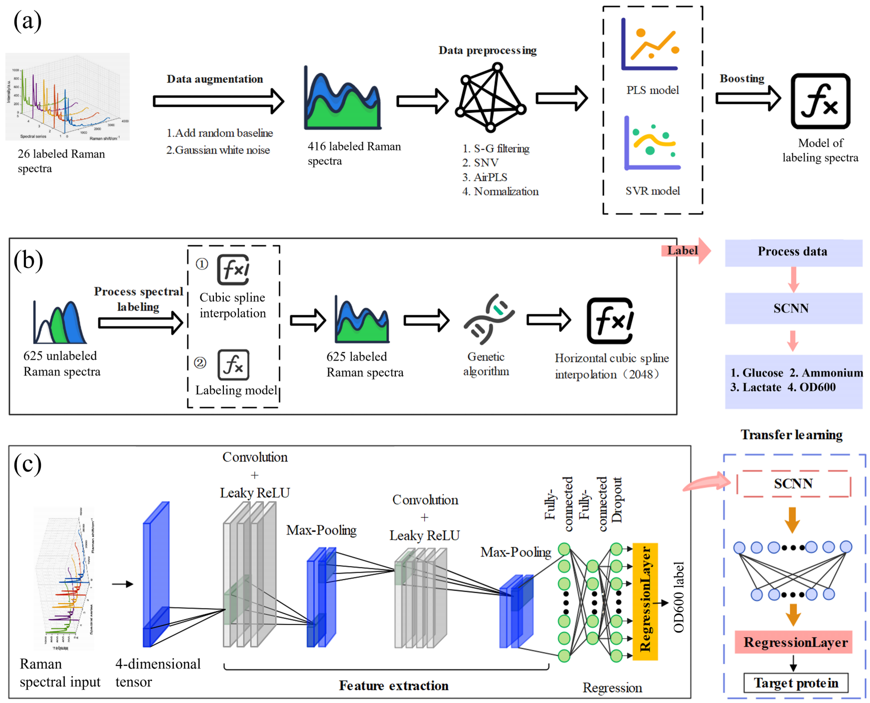
[1]Beijing Institute of Petrochemical Technology

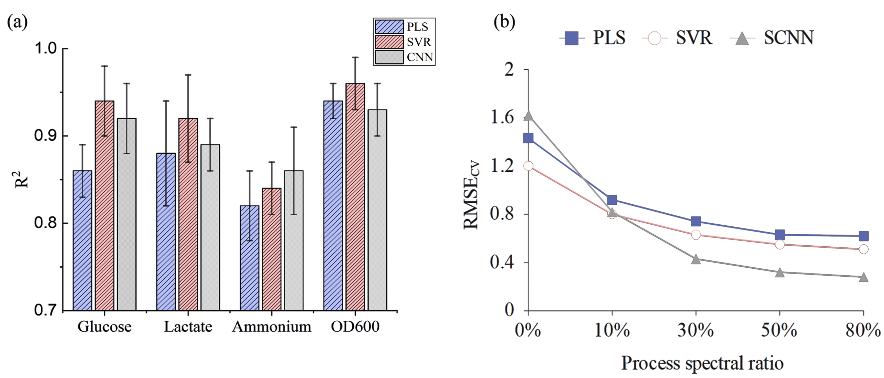
January 4, 2024

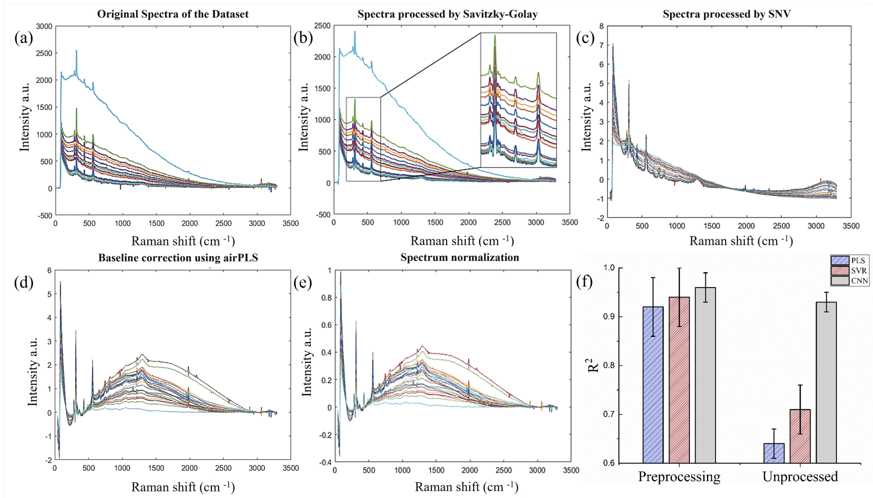
Abstract

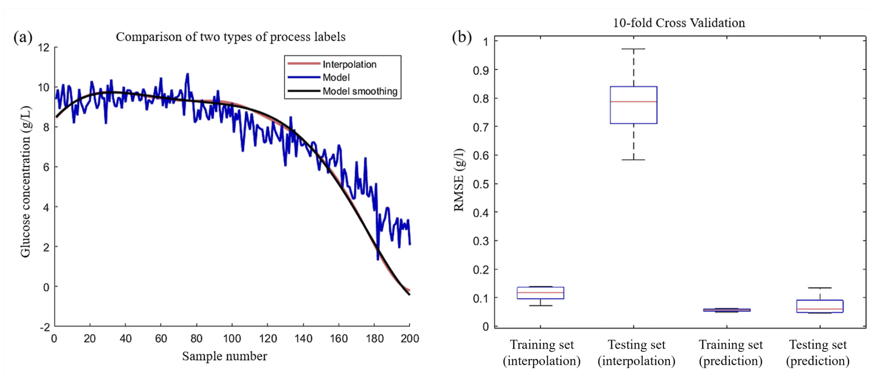
As a non-destructive sensing technique, Raman spectroscopy is often combined with regression models for real-time detection of key components in microbial cultivation processes. However, achieving accurate model predictions often requires a large amount of offline measurement data for training, which is both time-consuming and labor-intensive. In order to overcome the limitations of traditional models that rely on large datasets and complex spectral preprocessing, in addition to the difficulty of training models with limited samples, we have explored a genetic algorithm-based semi-supervised convolutional neural network (GA-SCNN). GA-SCNN integrates unsupervised process spectral labeling, feature extraction, regression prediction, and transfer learning. Using only an extremely small number of offline samples of the target protein, this framework can accurately predict protein concentration, which represents a significant challenge for other models. The effectiveness of the framework has been validated in a system of Escherichia coli expressing recombinant ProA5M protein. By utilizing the labeling technique of this framework, the available dataset for glucose, lactate, ammonium ions, and optical density at 600 nm (OD600) has been expanded from 52 samples to 1302 samples. Furthermore, by introducing a small component of offline detection data for recombinant proteins into the OD600 model through transfer learning, a model for target protein detection has been retrained, providing a new direction for the development of associated models. Comparative analysis with traditional algorithms demonstrates that the GA-SCNN framework exhibits good adaptability when there is no complex spectral preprocessing. Cross-validation results confirm the robustness and high accuracy of the framework, with the predicted values of the model highly consistent with the offline measurement results.

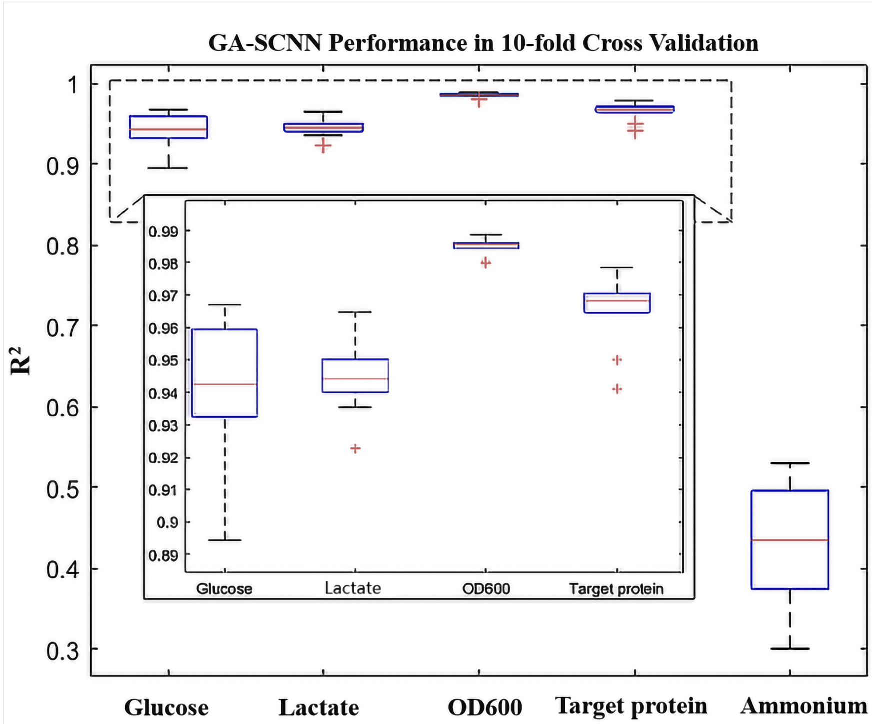


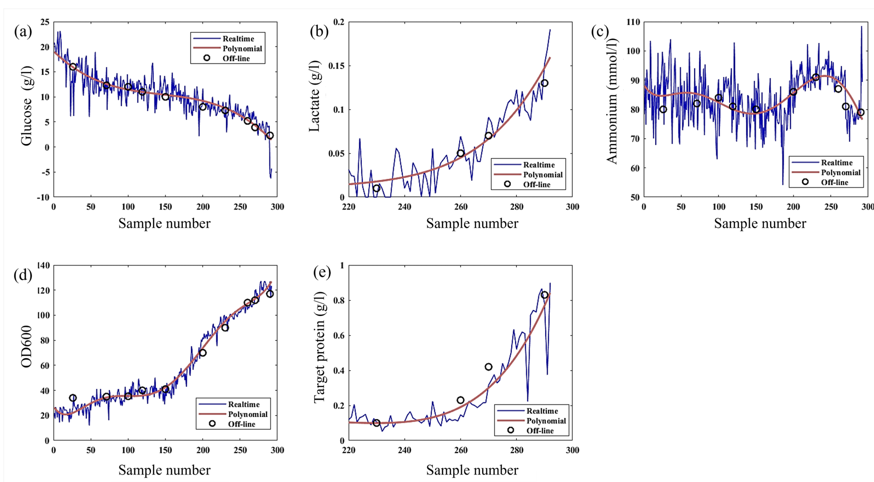

(a)
26 labeled Raman spectra
Data augmentation
1.Add random baseline
2.Gaussian white noise
416 labeled Raman spectra
Data preprocessing
1. S-G filtering
2. SNV
3. AirPLS
4. Normalization
PLS model
SVR model
Boosting
Model of labeling spectra
(b)
625 unlabeled Raman spectra
Process spectral labeling
① Cubic spline interpolation
② Labeling model
625 labeled Raman spectra
Genetic algorithm
Horizontal cubic spline interpolation（2048）
Label
Process data
SCNN
1. Glucose 2. Ammonium
3. Lactate 4. OD600
Transfer learning
(c)
Raman spectral input
4-dimensional tensor
Convolution + Leaky ReLU
Max-Pooling
Convolution + Leaky ReLU
Max-Pooling
Fully-connected
Fully-connected
Dropout
RegressionLayer
OD600 label
Feature extraction
Regression
SCNN
RegressionLayer
Target protein


(a)
PLS
SVR
CNN
R2
1.0
0.9
0.8
0.7
Glucose
Lactate
Ammonium
OD600
(b)
PLS
SVR
SCNN
RMSECV
2
1.6
1.2
0.8
0.4
0
0%
10%
30%
50%
80%
Process spectral ratio


(a)
Original Spectra of the Dataset
(b)
Spectra processed by Savitzky-Golay
(c)
Spectra processed by SNV
(d)
Baseline correction using airPLS
(e)
Spectrum normalization
(f)
Intensity a.u.
Raman shift (cm -1)
PLS
SVR
CNN
R2
Preprocessing
Unprocessed


(a)
Comparison of two types of process labels
Interpolation
Model
Model smoothing
Glucose concentration (g/L)
Sample number
(b)
10-fold Cross Validation
RMSE (g/l)
Training set (interpolation)
Testing set (interpolation)
Training set (prediction)
Testing set (prediction)


GA-SCNN Performance in 10-fold Cross Validation
R^2
Glucose
Lactate
OD600
Target protein
Ammonium


(a)
(b)
(c)
(d)
(e)
Realtime
Polynomial
Off-line
Glucose (g/l)
Lactate (g/l)
Ammonium (mmol/l)
OD600
Target protein (g/l)
Sample number