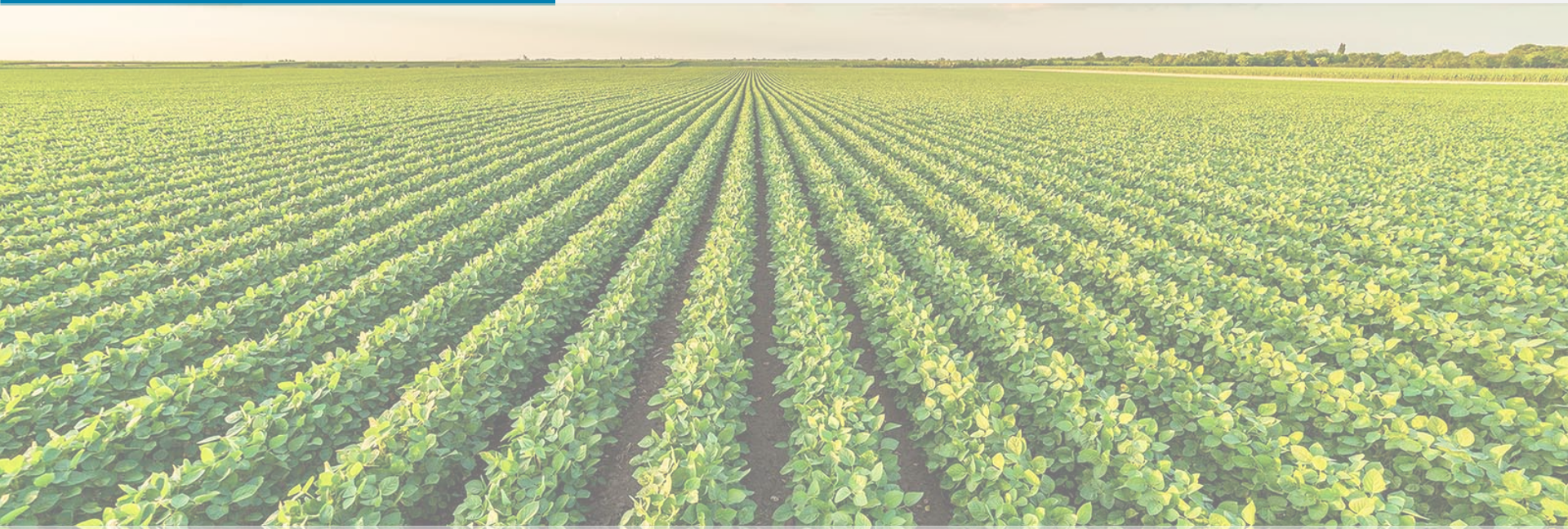
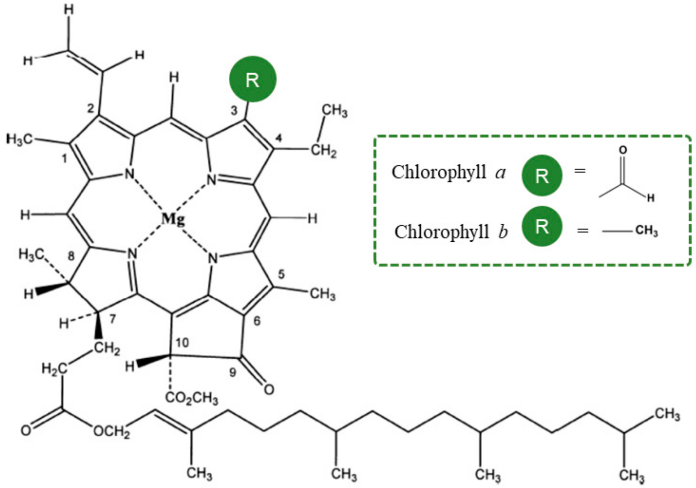


Exam: M. Streicher, March 25, 2025, Potsdam

Intelligent Data Analysis - Nitrogen (N) prediction

Introduction

- Nitrogen (N) part of chlorophyll (CHL) molecule
- Important fertilizer



<u>N fertilizer</u>

- Crucial for increasing crop yields
- Strong connected to environmental pollution

<u>Estimation of N:</u>

- Expensive, destructive and time consuming laboratory work

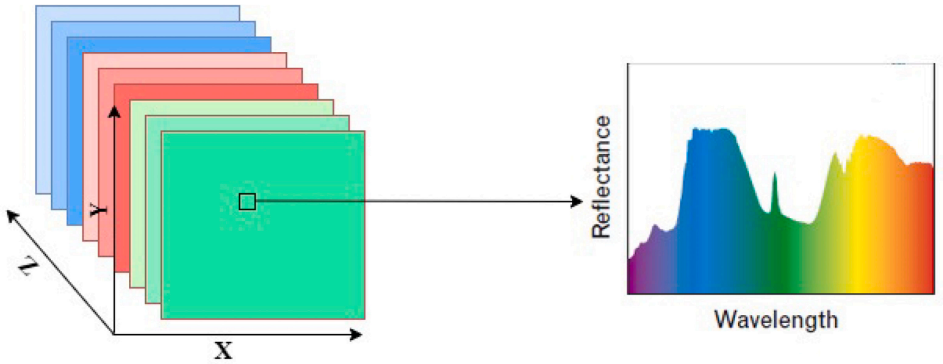
Introduction

New method should be:

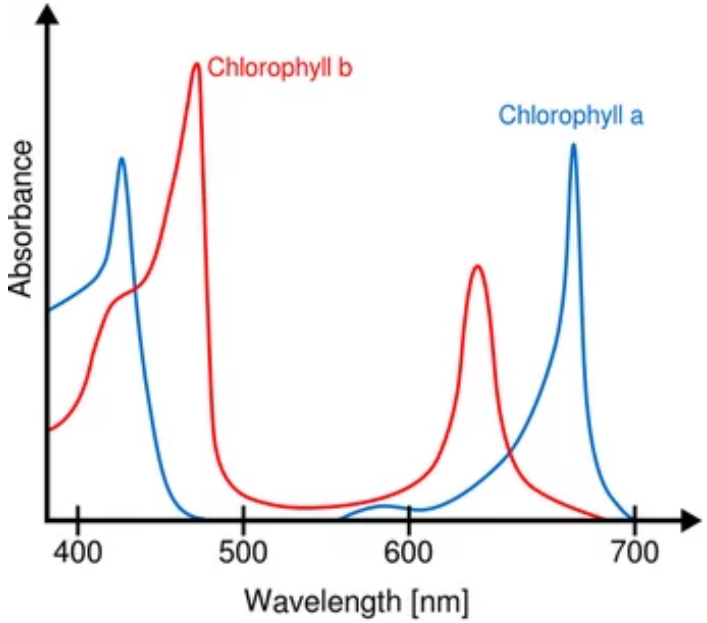
- Less expensive
- Applicable in large scale
- Non-destructive (*in vivo* usable)
- Not as time consuming as laboratory work

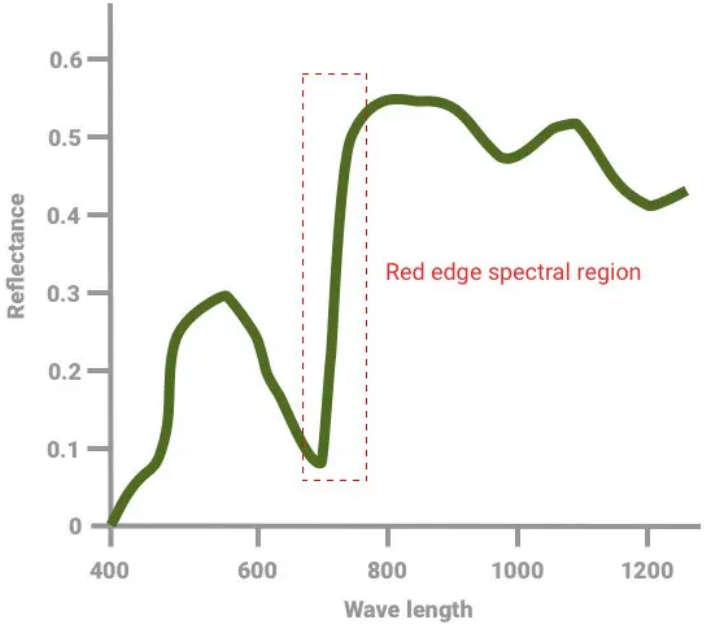
Solution:

Leaf level hyperspectral image data



Why are leaf-level spectral means so important for predicting N in plant leaf samples?





Data origin

RESEARCH PAPER

A leaf-level spectral library to support high-throughput plant phenotyping: predictive accuracy and model transfer

Nuwan K. Wijewardane[1], Huichun Zhang[2,6], Jinliang Yang[3,5], James C. Schnable[3,5], Daniel P. Schachtman[3,5] and Yufeng Ge[4,5,*]

- Several experiments conducted between 2018 and 2020
- University of Nebraska Lincoln

Domain	Number of samples
Soybean *(Glycine max)*	126
Camelina *(Camelina sativa)*	96
Maize *(Zea mays)*	1734
Sorghum *(Sorghum bicolor)*	949

Input data

- In total 2905 mean spectra from four domains
- Visible, near-infrared, and shortwave infrared regions (VIS-NIR-SWIR)
- Specifically: wavelengths from 350 up to 2500

Target data

- Plant samples: dried and homogenized
- Measuring: Dumas method with a LECO FP428 N analyzer (AOAC method 968.06)
- N content is expressed as a percentage of dry matter

Problem Setting

Input data D:

$$D = \left\{ \left(x_i, y_i \right) \right\}_{i=1}^{2905},$$

where $x_i = [x_{i1}, x_{i2}, \ldots, x_{i2151}]$ is the mean reflectance per wavelength of plant leaf sample i and $y_i \in \mathbb{R}$ the continuous label that represents the N content.

Model:

$f_\theta : \mathbb{R}^{2151} \rightarrow \mathbb{R}$, where θ is the model parameter vector.

Problem Setting

Task:

Regression problem

Type of learning:

Supervised Learning

Goal:

To train a model that reliable predict the N content of our leaf level hyperspectral image data.

Data analysis

Input data D:

$$D = \left\{ \left(x_i, y_i \right) \right\}_{i=1}^{2905}$$

Missing data:

No missing data found.

Standardisation:

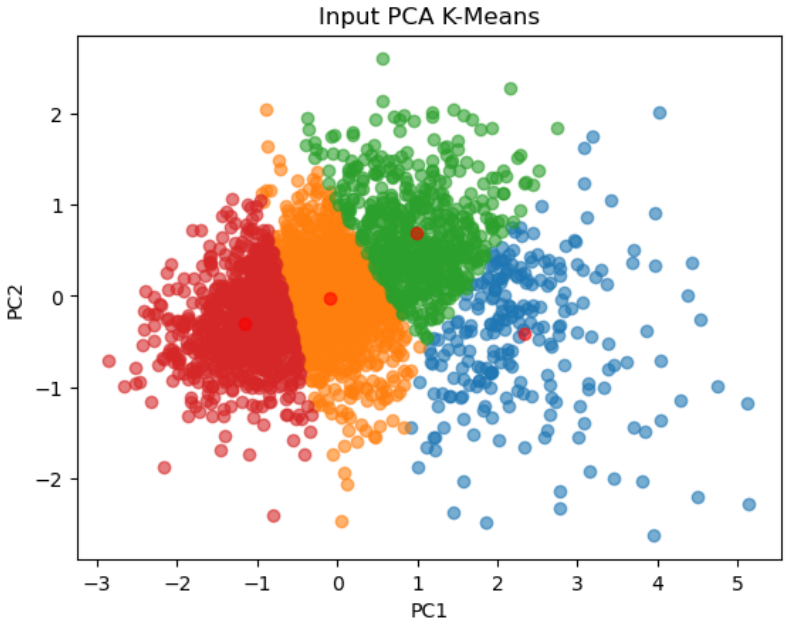
Z-score Standardization.

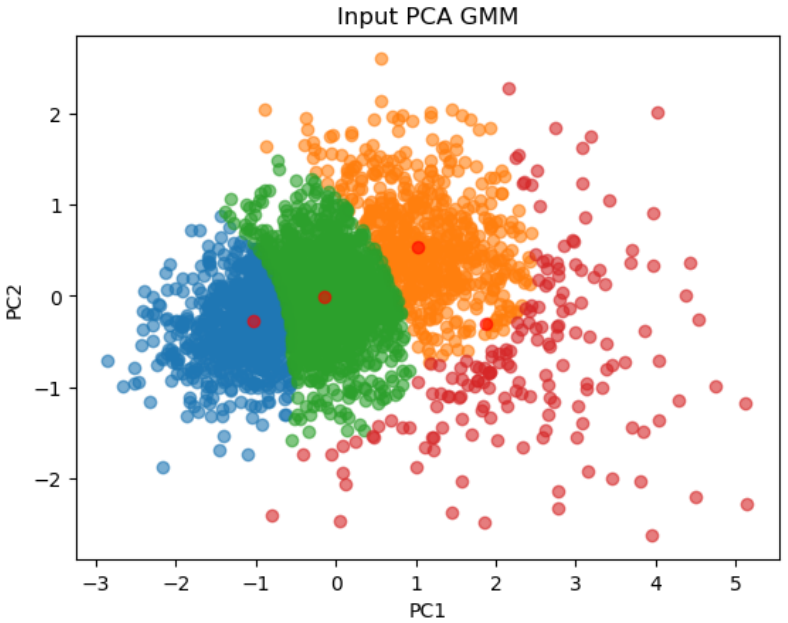
Data analysis

Clustering:

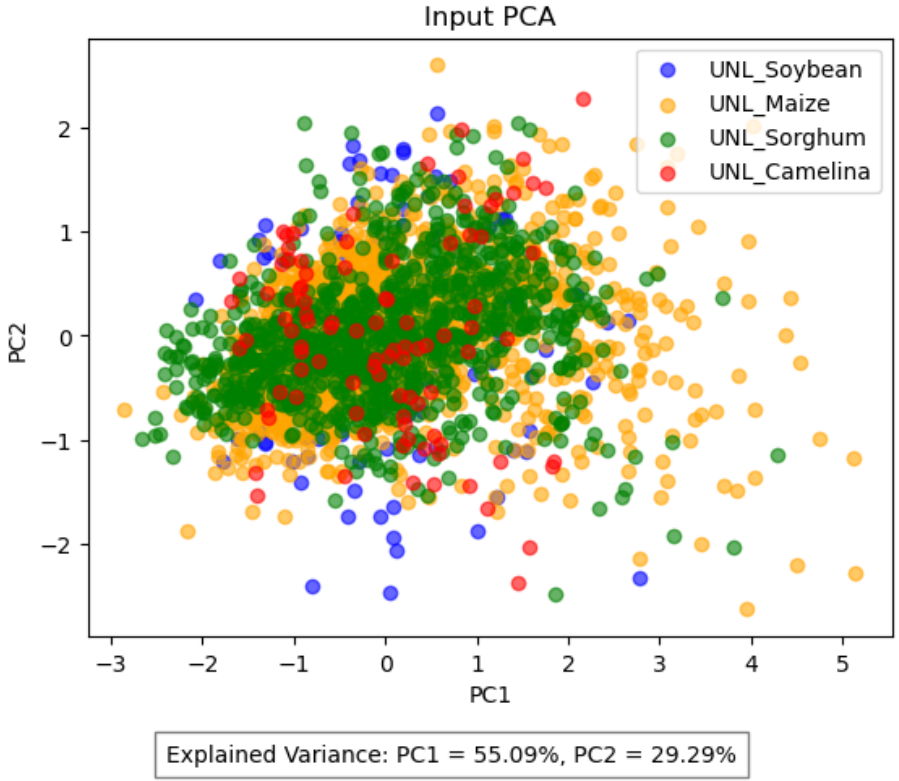
- Four different species (domains)
- Check for differences in input and target distributions
- Input data D: Clustering based on PC1 and PC2
- Target data y_i: 1D clustering
- Clustering methods: K-Means, GMM

Data analysis





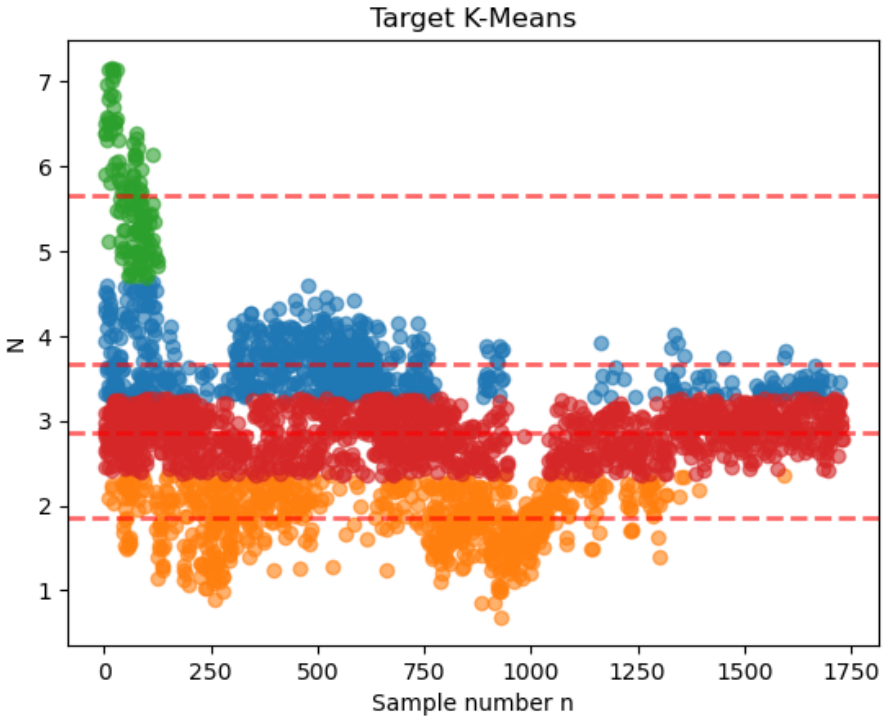
Data analysis

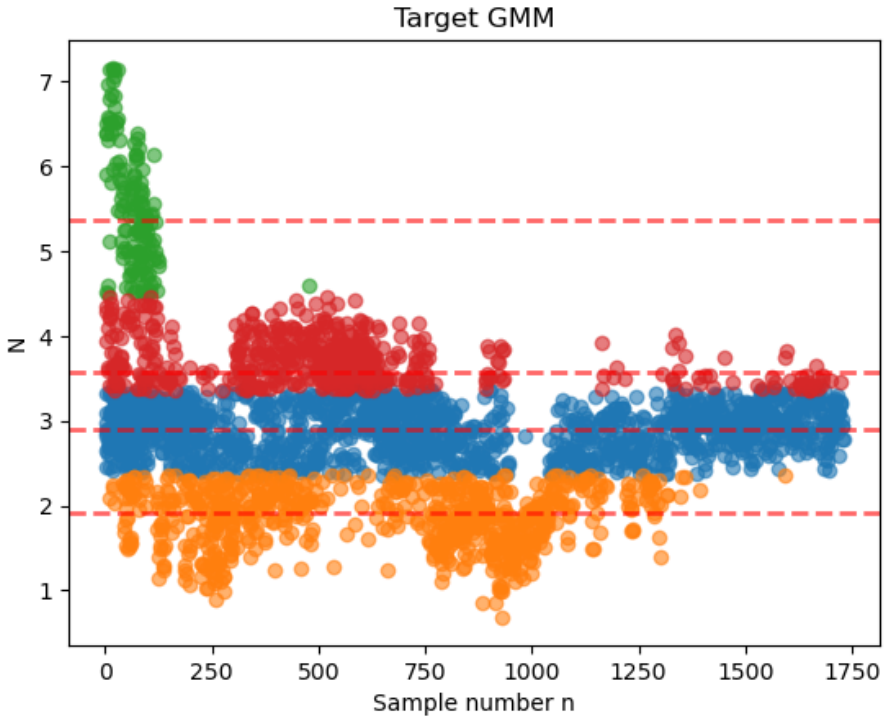


Result

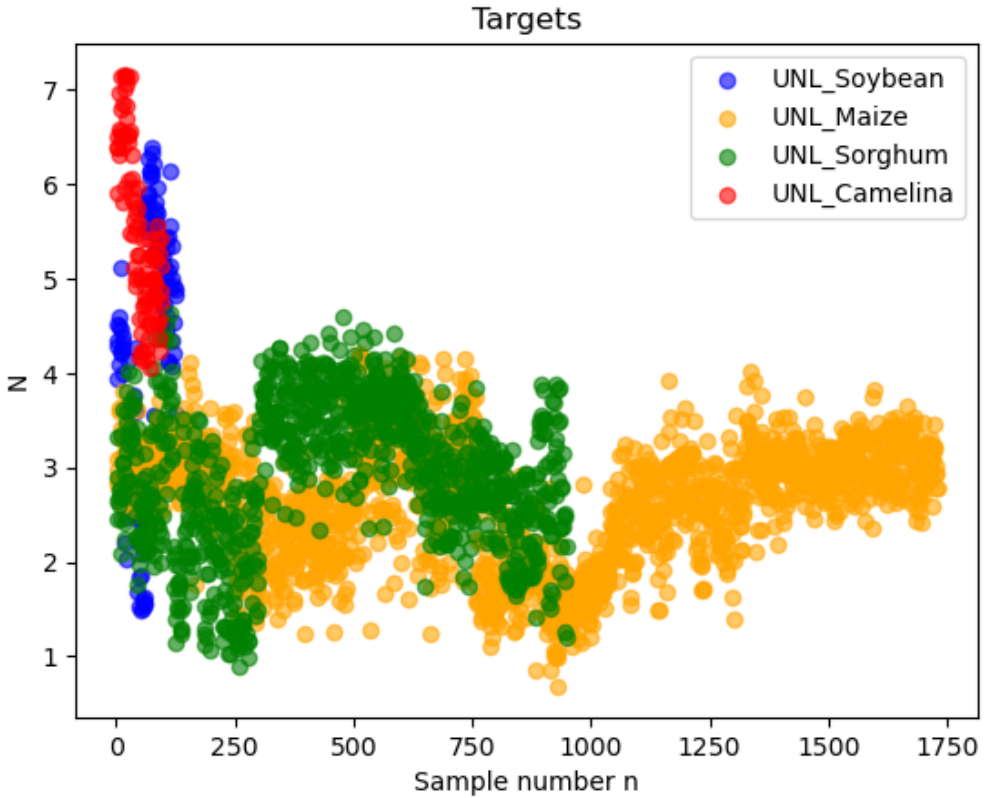
- Both cluster methods always provide a result.
- No distinguishable clusters in the linear principle component space for the input data.

Data analysis





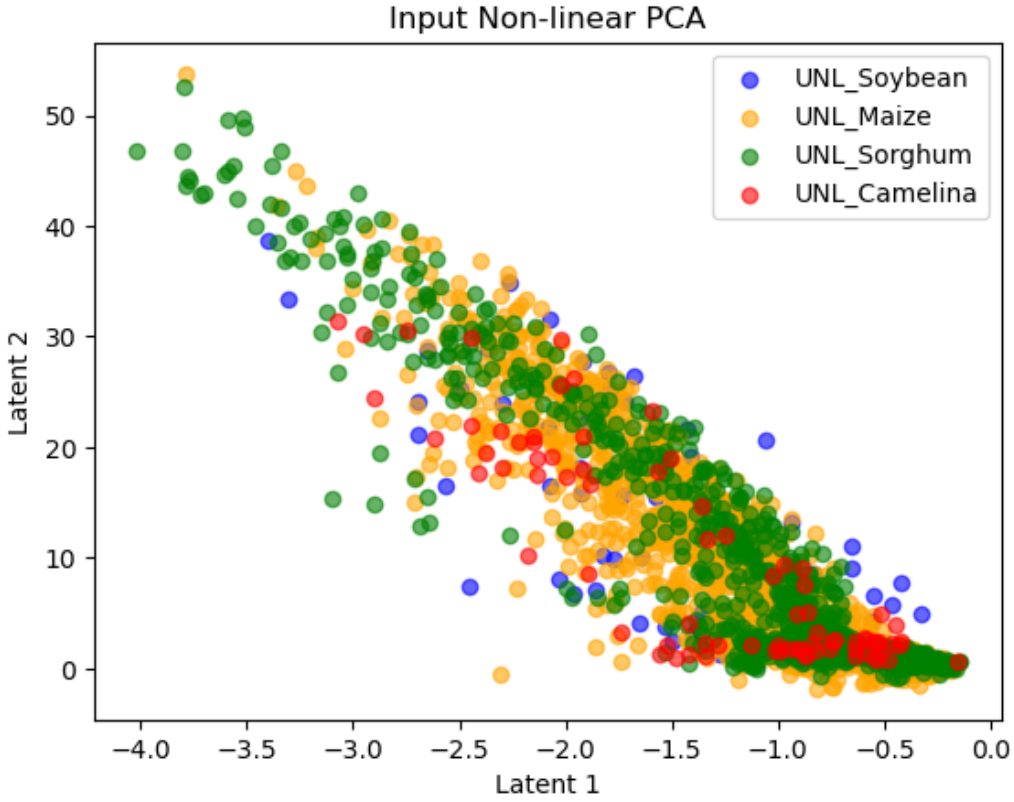
Data analysis



Result

- Both cluster methods always provide a result.
- Camelina cluster was found

Data analysis



Evaluation Protocol

$$A(\lambda \mid C) = E[max\{0, r^* - r'\}],$$
$$\textbf{where } r' \sim N[\mu(f(\lambda) \mid \{\lambda_c, r_c\}), \sigma^2(f(\lambda) \mid \{\lambda_c, r_c\})]$$

Data separation

Train, Validation and Test

Hyperparameter Search

- Hold-out testing with train and validation
- Weights & Biases
- Approach: Bayesian Optimization

(Gaussian Process, Matérn 5/2 kernel and Expected Improvement)

Final evaluation

- 5-fold cross validation
- Evaluation on validation and test set per fold

Models and Results

Mean Baseline, Neural Network, 1D CNN

Mean Baseline

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \vec{\mu} = [\mu, \ldots, \mu] \in \mathbb{R}^l, \text{where } l = \text{length test set}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| = \frac{1}{l}\sum_{i=1}^{l} |y_i - \vec{\mu}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{l}\sum_{i=1}^{l} (y_i - \vec{\mu}_i)^2$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \vec{\mu}_i)^2}{\sum_i (y_i - \mu)^2}$$

Metric	Value
MAE	0.71
MSE	0.93
R^2	0

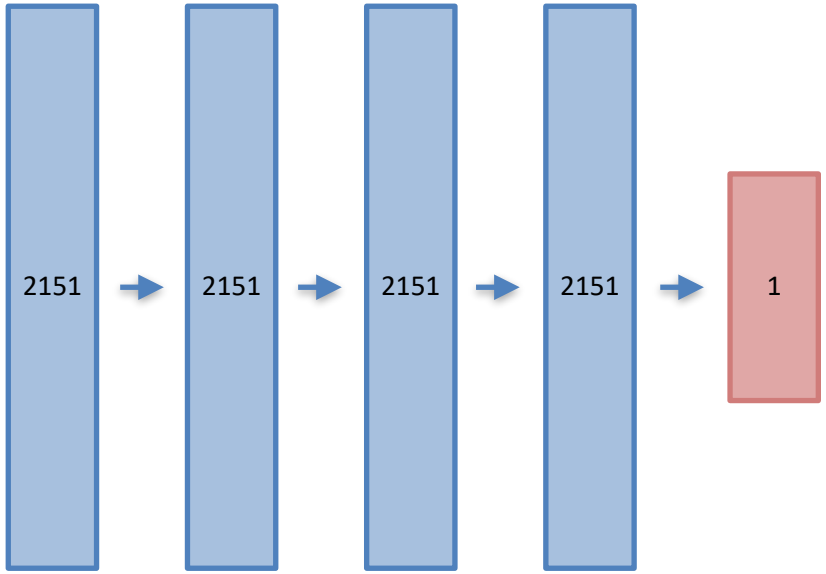
Neural Network

Regularized Empirical Risk Minimization:

$$argmin_\theta \left(\sum_i^n l((y_i - f_\theta(x_i))^2) + \lambda \Omega_2(\theta) \right)$$

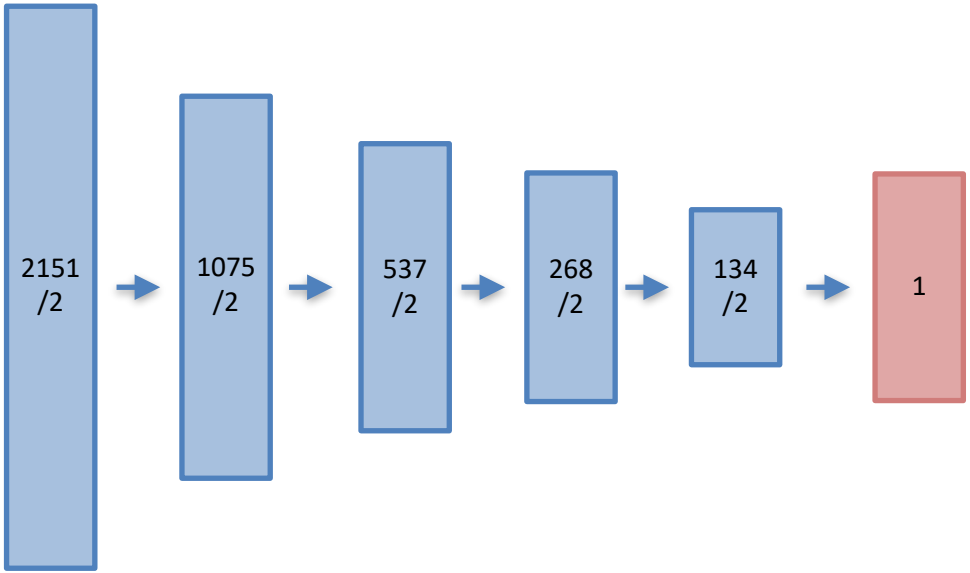
- Solve by using Gradient Descent Method
- Gradient is calculated by Back-propagation
- Gradient update step is done by Stochastic Gradient Descent (optim.SGD)

Architecture I - Simple



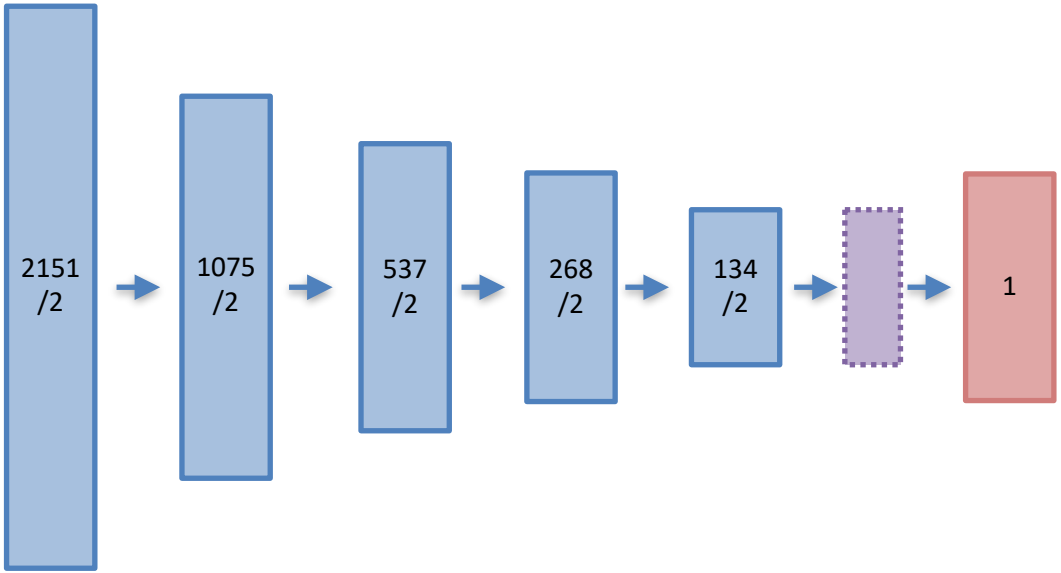
	Hyperparameter
Layers	4
Learning rate	0.001
Weight decay	0.001
	Result
MSE	0.25
R^2	0.68
Parameter	8,605

Architecture II - Bottleneck



	Hyperparameter
Layers	5
Learning rate	0.001
Weight decay	0.0001
	Result
MSE	0.24
R^2	0.68
Parameter	2,081

Architecture II - Bottleneck + Dropout



	Hyperparameter
Layers	5
Learning rate	0.001
Weight decay	0.0001
	Result
MSE	0.23
R^2	0.69
Parameter	2,081

1D CNN

Regularized Empirical Risk Minimization:

$$argmin_\theta \left(\sum_i^n l((y_i - f_\theta(x_i))^2) + \lambda \Omega_2(\theta) \right)$$

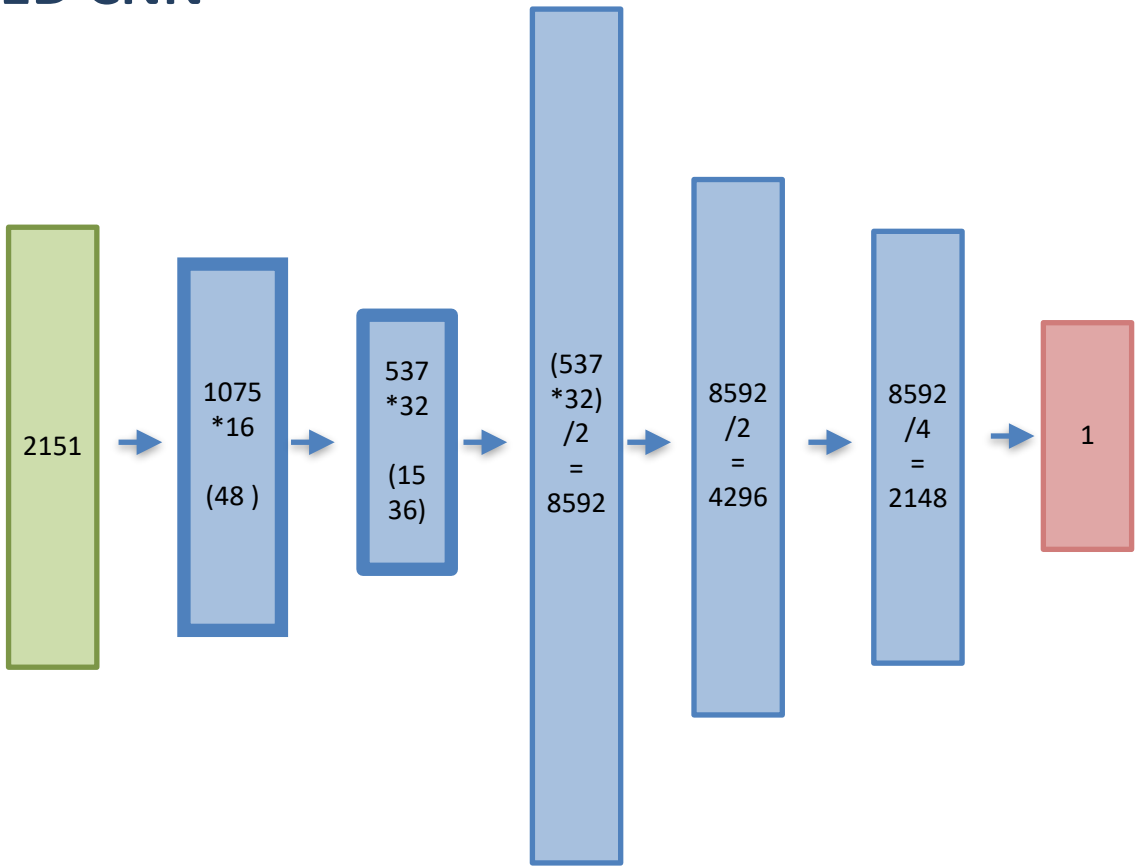
RESEARCH Open Access

Check for updates

A hyperspectral deep learning attention model for predicting lettuce chlorophyll content

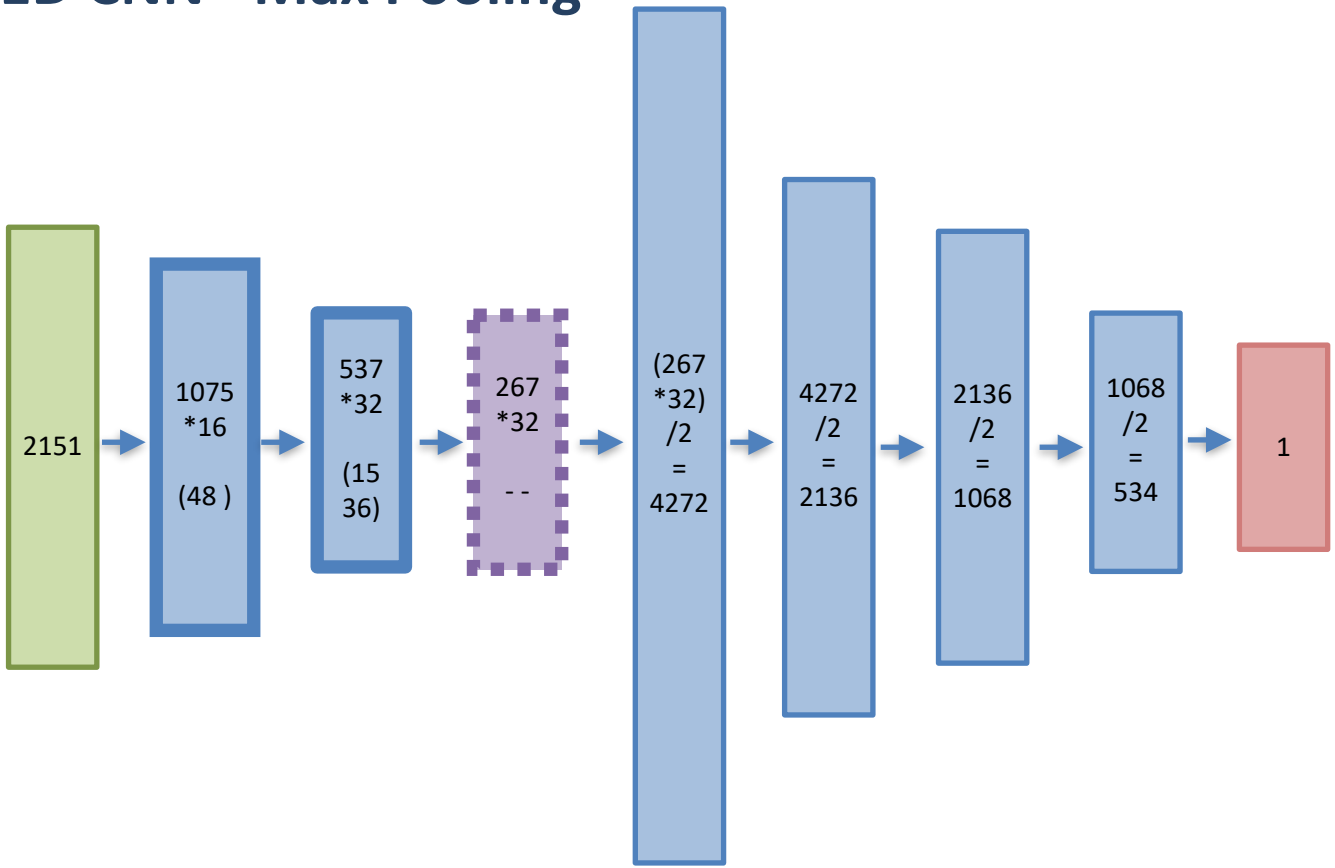
Ziran Ye[1], Xiangfeng Tan[1], Mengdi Dai[1], Xuting Chen[1], Yuanxiang Zhong[1], Yi Zhang[2], Yunjie Ruan[3,4] and Dedong Kong[1*]

1D CNN



	Hyperparameter
Layers	3
Learning rate	0.01
Weight decay	0.001
Kernel size	3
Padding	0
Stride	2
Out channel 1	16
Out channel 2	32
	Result
MSE	0.22
R^2	0.72
Parameter	16,621

1D CNN - Max Pooling



	Hyperparameter
Layers	4
Learning rate	0.001
Weight decay	0.001
Kernel size	3
Padding	0
Stride	2
Out channel 1	16
Out channel 2	32
	Result
MSE	
R^2	0.69
Parameter	9,595

Conclusion

Clustering, Mean Baseline, Neural Network, 1D CNN

Conclusion

- Bottleneck structure improved model performance
- Drop out improved model performance
- NN: Lowest R^2 of 0.68
- CNN: Highest R^2 of 0.72

Conclusion: None of the models could reached a reliable R^2 score > 0.9.

Sources

[1] Image cover slide: rbb, 2016, Retrieved from: https://www.rbb-online.de/content/dam/rbb/rbb/fernsehen/rbb_praxis_bilder/2016/05/25/COLOURBOX8440750.jpg.jpg/size=966x543.jpg

[2] Image question slide: Gekonnt wirken, Retrieved from: https://www.gekonnt-wirken.de/wp-content/uploads/2018/07/AdobeStock_496887170-scaled-82625_1080x675.jpeg

[3] Source Code: https://github.com/MarStreicher/IDA_Laser/tree/main/