Development of an R Toolbox for Near-Infrared Spectroscopy Data Processing and Analysis of Plant Metabolic Phenotypes

Deggendorf Institute of Technology

MSc. Life Science Informatics

Methun George

Supervised by
PD Dr. habil. rer. nat. Steffen Neumann
Prof. Dr. Melanie Kappelmann-Fenzl

December 13, 2024

Contents

Chapter 1

Introduction

The advancement and widespread use of heigh-throughput experimental technologies in the field of plant biology have introduced significant challenges in managing and analysing the vast datasets effectively. Addressing these challenges require innovative methods that maximize the data utility while mimnimizing computational inefficiencies and resource consumption, ensuring robust insights into complex biological systems (https://www.frontiersin.org/research-topics/6856/machine-learning-in-plant-science/articles).

Chapter 2

Background

The background of this study include

2.1 Related Work

2.2 Near Infrared Spectroscopy (NIRS)

2.3 R Programming

2.4 Machine Learning

2.4.1 Partial Least Square Regression (PLSR)

2.4.2 Random Forest (RF)

2.4.3 Convolutional Neural Network (CNN)

2.5 Mass Spectrometry and Liquid Chromatography

Chapter 3

Implementation

3.1 Packages

Github, testing, actions

3.2 Contributions elsewhere

3.3 HPC runs

Chapter 4

Results and Discussion

4.1 Data charecterestics

histogram, spectra

4.2 Baseline Machine Learning Models Pablo

PLS, RF, CNN

4.2.1 Variable importance

4.3 Variations in Baseline systems

4.3.1 modifying the Test and Training split

4.3.2 input data length

4.4 Sues