

# 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 15, 2024

# Contents

# Chapter 1

## Introduction

The advancement and widespread use of high-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 minimizing 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>).

The understanding of interplay between plant physiology and its hidden biochemical process is crucial for the improvement of basic plant science and addressing global challenges such as food security, crop resilience and combating climate change [1]. In recent years, advanced High-throughput analytical techniques such as Near-Infrared Spectroscopy (NIRS) and Liquid Chromatography-Mass Spectrometry (LC-MS) has instigated a paradigm shift in plant biology [2][3]. These High-throughput techniques are mostly used in areas like genomics, imaging and spectroscopy and is known for their ability to collect and analyse the data faster than traditional techniques[3].

# 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

# Chapter 5

## Reference

1. Pieruschka R, Schurr U. Plant Phenotyping: Past, Present, and Future. *Plant Phenomics*. 2019 Mar 26;2019:7507131. doi: 10.34133/2019/7507131. PMID: 33313536; PMCID: PMC7718630.
2. Pulok K. Mukherjee, Quality control and evaluation of herbal drugs, Evaluating natural products and traditional medicine. 2019, doi:10.1016/C2016-0-042328, ISBN:978-0-12-813374-3
3. Nizamani, M. M., Zhang, Q., Muhae-Ud-Din, G., Wang, Y. (2023). High-throughput sequencing in plant disease management: A comprehensive review of benefits, challenges, and future perspectives. *Phytopathology Research*, 5(44). <https://doi.org/10.1186/s42483-023-00215-7>
4. Lane, H. M., Murray, S. C. (2021). High throughput can produce better decisions than high accuracy when phenotyping plant populations. *Crop Science*, 61(3), 1473–1484. <https://doi.org/10.1002/csc2.20514>