# Predictive modeling of Zinc concentration in *Arabidopsis halleri* leaves using Artificial Intelligence and plant photos

A thesis submitted as a fulfilment of requirements for Master's degree in Bioinformatics

By

**Dima SOULEMAN**

Supervised by

**Dr. Yasser KHADRA**

# Table of contents

1 **Introduction** ..... 7
1.1 Background on Zinc Contamination ..... 7
1.1.1 Importance of studying Zn contamination in the environment. ..... 7
1.1.2 Impact on ecosystems and human health. ..... 8
1.1.3 Zinc contamination in soils and its effects on plant health ..... 9
1.2 Role of Arabidopsis helleri ..... 12
1.2.1 Characteristics of *Arabidopsis halleri* as a hyperaccumulator and tolerant plant. . 12
1.2.2 *Arabidopsis halleri* as a bioindicator of Zn contamination. ..... 14
1.3 AI in Environmental Science ..... 15
1.3.1 AI approaches that used in environmental science ..... 15
1.3.2 Overview of AI applications in environmental monitoring. ..... 17
1.4 Logistic Regression ..... 17
1.4.1 Logistic Regression: The Art of Predicting Probabilities ..... 17
1.4.2 Logistic regression in the field of environmental science ..... 19
1.5 Machine Learning ML for predicting environmental metals pollutant ..... 21
2 **Literature study** ..... 23
2.1 Studies using ML to detect metals contamination and accumulation ..... 23
2.2 Studies using plant images and machine learning ML ..... 24
2.2.1 Plant disease recognition ..... 24
2.2.2 Plant classification ..... 25
3 **Methodology** ..... 28
3.1 Data Collection ..... 28
3.1.1 Description of data and images collection methods for *Arabidopsis halleri*. ..... 28
3.2 Image processing ..... 30
3.3 Data Preprocessing ..... 32
3.3.1 Data cleaning ..... 32
3.3.2 Data visualization: ..... 33
3.3.3 Statistical analysis: ..... 37

3.4 AI Model Development ..... 39
3.4.1 Data preparation: ..... 39
3.4.2 Model Training for Logistic Regression ..... 40
3.4.3 Metrics for evaluating model performance ..... 40
3.5 Zn content prediction in A.halleri leaves ..... 41
3.5.1 Data preparation ..... 42
3.5.2 Model development ..... 42
3.5.3 Evaluate model performance ..... 43
4 **Results** ..... 45
4.1 Statistical results for all features ..... 45
4.2 Model Performance for Zn contamination prediction ..... 45
4.2.1 Without images feature ..... 45
4.2.2 With images data ..... 48
4.1 Model Performance for Zn content prediction ..... 51
5 **Discussion** ..... 53
5.1 Interpretation of Statistical Results ..... 53
5.1 Interpretation of logistic regression Results ..... 53
5.2 Interpretation of multiple regression Results for Zn content prediction ..... 54
5.3 Advantages of Using ML ..... 54
5.4 Limitations of the Study ..... 55
6 **Conclusion** ..... 55
6.1 Summary of Findings ..... 55
6.2 Future Directions ..... 56
7 **References** ..... 57

## Figure and table list:

Figure 1: Soil heavy metal pollution from Pb/Zn smelting regions ……..5

Figure2: Impact of Zn contamination on human health (Angon et al., 2024)….7

Figure 3: Normal plant vs plants with signs of Zn toxicity…..8

Figure 4: Effects of heavy metals HMs on plants and crop production (Angon et al., 2024)….9

Figure 5: Arabidopsis helleri in natural habitat….10

Figure 6: Mechanisms of heavy metal tolerance and accumulation…..11

Figure 7: The main genes involved in the hyper-accumulation of Zn (Verbruggen et al., 2009)….12

Figure 8: Non-linear Relationship of logistic regression …….16

Figure 9: Hydroponic culture…...26

Figure 10: *Arabidopsis halleri* plants in polluted and non-polluted condition……28

Figure 11: Visualization of RGB Channels in Red, Green and Blue ……29

Figure 12: Samples count in polluted (1) and non-polluted (0) conditions…..30

Figure 13: Morphological Features distributions in the non-polluted (0) and polluted (1) conditions for all plants…...31

Figure 14: physiological Features distributions in the non-polluted (0) and polluted (1) conditions for all plants……32

Figure 15: Visualization of relationship between variables using Scatter Plots, for all samples in polluted and non-polluted condition…..33

Figure 16: Visualization of relationship between variables using Scatter Plots. In green, samples in polluted condition, in orange, samples in non-polluted condition…...34

Figure 17: Correlation rates between the features in the non-polluted and Zn-polluted conditions…..35

Figure 18: Correlation between morphological and physiological features, stars indicate the significance of this correlation as following: $p < 0.001$ = ***, $p < 0.01$ = **, $p < 0.05$ = *.…36

Figure 20: Confusion matrix with advanced classification metrics…...39

Figure 21: confusion matrix of the model's performance without images features…..42

Figure 23: confusion matrix of the model's performance with images features…….45

Table 1: Description of A.halleri Morphological and Physiological Features…27

Table 2: Accuracy, precision, recall, and F1 score for the first model, features without images

Table 3: Accuracy, precision, recall, and F1 score for the first model, with images feature

## Table of abbreviations

| | | | |
|---|---|---|---|
| **Zn** | Zinc | **GANs** | Generative Adversarial Networks |
| **P** | Polluted condition by Zn | **NLP** | Natural Language Processing |
| **NP** | Non-Polluted condition by Zn | **XAI** | Explainable AI |
| **ROS** | reactive oxygen species | **ML** | Machine learning |
| **HMs** | heavy metals | **BAFs** | transfer of HMs from soil to crops |
| **AI** | Artificial intelligence | **ENPs** | metallic Engineered Nanoparticles |
| **PCA** | Principal Component Analysis | **RCF** | root concentration factor |
| **CNNs** | Convolutional Neural Networks | **TF** | translocation factor |
| **RNNs** | Recurrent Neural Networks | **ANNs** | artificial neural networks |
| **BPNN** | back propagation neural network | **ConvNet** | off-the-shelf convolutional neural network |
| **LonR2** | root length | **massR** | root dry biomass |
| **moyLarF** | leaf width | **massF** | shoot dry biomass |
| **moyFluo2** | photosystem II yield | **green** | green channel values |
| **Zn_acc** | Zn content in the leaves | **p** | P_value |
| **ROC** | Receiver Operating Characteristic | **TN** | True negative |
| **TP** | True positive | **FN** | False negative |
| **FP** | False positive | | |

## Abstract

Zinc contamination resulting from anthropogenic activities poses a significant threat to human and plant health. Predicting such contamination is crucial to enable the implementation of appropriate treatment methods. *Arabidopsis halleri* is widely known for its remarkable mechanisms to tolerate and hyperaccumulate heavy metals, particularly cadmium and Zinc, in its tissues.

This study aimed firstly, to develop a logistic regression model to predict Zn contamination in soil using a combination of morphological and physiological features extracted through traditional measurement methods from 1000 individuals of *A.halleri*, as well as plant color (chlorophyll content) features extracted from plant images using machine learning approaches. Secondly, To develop a multiple machine learning model to predict Zn concentration in *A.halleri* leaves based on the same plant features.

The logistic regression model demonstrated a remarkable training accuracy of 0.9423 when using the traditional features alone, indicating the model's exceptional ability to classify the training data into polluted and non-polluted conditions. Interestingly, when the plant color feature derived from images was incorporated, the training accuracy increased to 0.954, and the test accuracy improved from 0.905 to 0.94, confirming the importance of the image-based feature in enhancing the model's performance.

This study contributes to the growing body of evidence that emphasizes the significance of evaluating machine learning models not only on their training performance but also on their ability to generalize to new, unseen data. The developed model has the potential to assist in the prediction of Zn contamination in soil, leveraging both traditional plant features and image-based features, which can substantially reduce the cost and time requirements associated with traditional laboratory analyses

# Chapter I
# Introduction
# Theoretical Study

# 1 Introduction

## *1.1 Background on Zinc Contamination*

### 1.1.1 Importance of studying Zn contamination in the environment.

Zinc contamination in the environment has been the focus of numerous studies due to its widespread use in various industries, including galvanization, smelting, mining, alloy production, and agriculture (Fig.1). This extensive application increases the risk of contamination in soil and water, leading to significant environmental impacts. Elevated zinc levels can harm plant growth, soil health, and aquatic ecosystems, posing health risks to humans and wildlife alike (Kaur & Garg, 2021). The potential for bioaccumulation in the food chain further complicates the issue, necessitating ongoing research to understand its effects and develop effective remediation strategies (L. Liu et al., 2018). Regulatory compliance also drives these studies, as many countries have established limits for heavy metals, including zinc, to protect public health and the environment (European Commission, 2006). Additionally, the influence of climate change and urbanization on zinc distribution and mobility adds another layer of complexity, prompting interdisciplinary collaboration among ecologists, chemists, and toxicologist (Jarsjö et al., 2020). Ultimately, addressing zinc contamination is crucial for safeguarding both human health and the integrity of ecosystems.

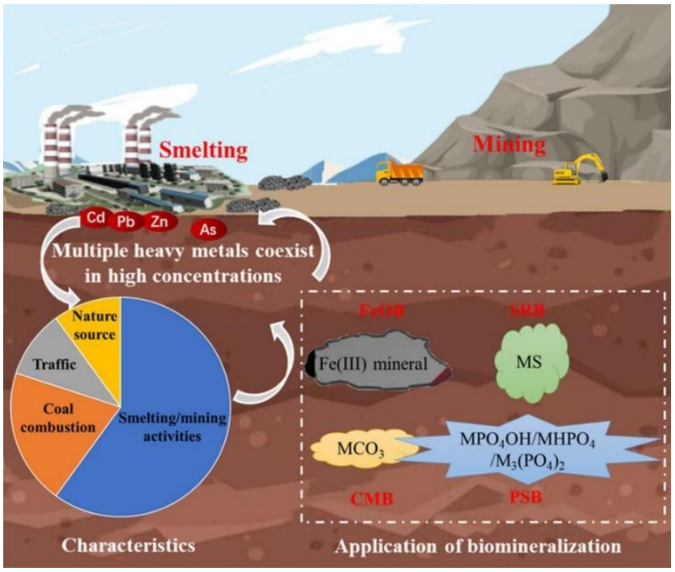



Figure 1: Soil heavy metal pollution from Pb/Zn smelting regions

### 1.1.2 Impact on ecosystems and human health.

Zinc contamination in ecosystems poses significant risks to both terrestrial and aquatic environments, impacting biodiversity and ecological health. Elevated levels of zinc can lead to soil degradation, affecting microbial communities and nutrient cycling (Okereafor et al., 2020). Soil microorganisms play a crucial role in maintaining soil fertility, and zinc toxicity can disrupt their populations, leading to reduced organic matter decomposition and impaired plant growth (Hussain et al., 2022) .

The impact of zinc on ecosystems extends beyond immediate toxicity; it also affects ecosystem services such as water purification, carbon storage, and habitat provision. As ecosystems become less resilient due to contamination, their ability to recover from disturbances diminishes, leading to long-term ecological consequences (P. Liu et al., 2021). Addressing zinc contamination is critical for maintaining the integrity of ecosystems and ensuring their capacity to support diverse life forms.

Zinc contamination in the environment can significantly impact human health, primarily through exposure via contaminated water, food, and air. While zinc is an essential trace element necessary for various biological functions, excessive exposure can lead to toxicity and adverse health effects. Acute zinc poisoning can result in symptoms such as nausea, vomiting, abdominal pain, and diarrhea. Chronic exposure, particularly in occupational settings or from consuming contaminated food and water, may lead to more severe health issues, including respiratory problems, immune dysfunction, and alterations in lipid metabolism (Plum et al., 2010) (Fig.2).

One of the primary concerns regarding zinc contamination is its potential to disrupt the endocrine system. Studies have shown that elevated zinc levels can interfere with hormone regulation, potentially leading to reproductive and developmental issues. For example, research indicates that high zinc concentrations can affect testosterone levels and impair fertility in both men and women (Pizent et al., 2012).

Furthermore, vulnerable populations, such as children and pregnant women, are at greater risk. Children exposed to high levels of zinc may experience developmental delays and cognitive impairments, while pregnant women may face complications that could affect fetal development.

Long-term exposure to zinc can also exacerbate existing health conditions, such as cardiovascular diseases and diabetes, due to its inflammatory properties.

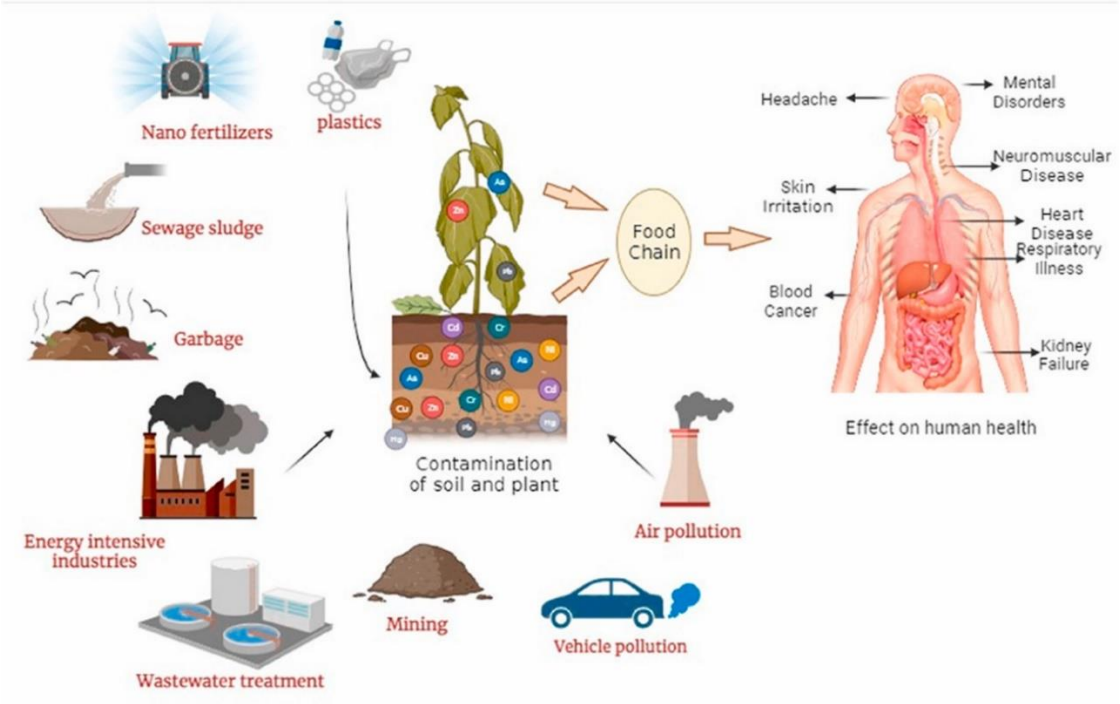



Figure2: Impact of Zn contamination on human health (Angon et al., 2024)

Given these potential health risks, monitoring and regulating zinc levels in the environment is crucial to protect public health. Efforts to reduce zinc contamination from industrial processes, agricultural runoff, and waste disposal are essential to minimize exposure and safeguard the well-being of communities.

### 1.1.3 Zinc contamination in soils and its effects on plant health

Zinc contamination in soils poses a dual threat to plant health and appearance, significantly influencing both physiological processes and visual characteristics. While zinc is vital for various enzymatic functions and growth, excessive levels can lead to toxicity, causing visible symptoms such as leaf chlorosis, necrosis, and stunted growth (Fig.3). Affected plants often exhibit a reduced leaf area and distorted growth patterns, making them less competitive in their environment. The stress induced by high zinc concentrations can also impair nutrient uptake, leading to deficiencies in essential elements like iron and magnesium, further exacerbating the decline in plant vigor. As a result, contaminated plants not only suffer from diminished health

but also display unsightly, unhealthy appearances, which can impact agricultural productivity and the aesthetic value of landscapes. Understanding these effects is crucial for developing strategies to mitigate soil contamination and promote healthier, more resilient plant communities (Yoon et al., 2006).

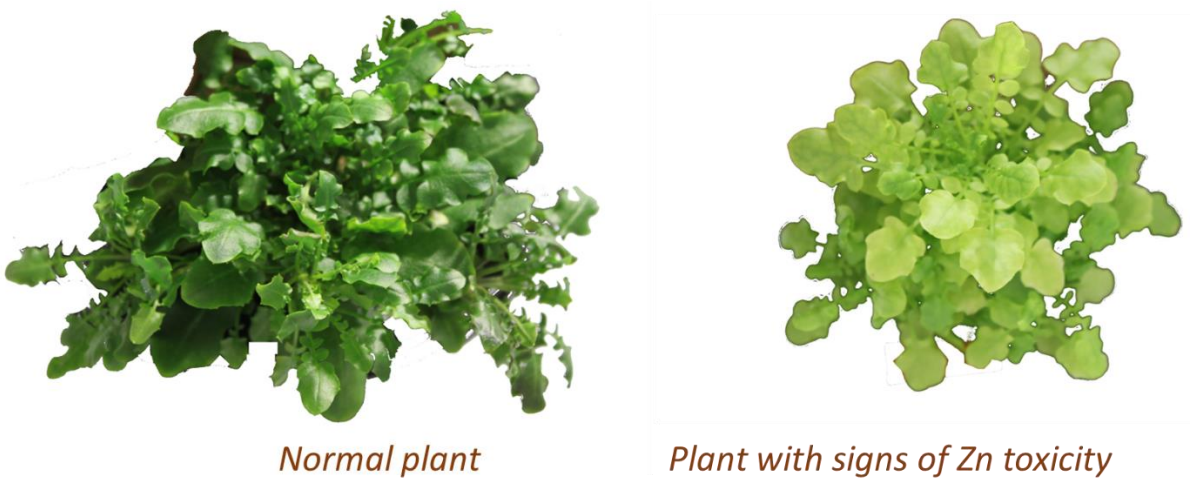



Figure 3: Normal plant vs plants with signs of Zn toxicity

Excessive zinc levels in plants can lead to a range of adverse effects that significantly impact their health and growth (Fig.4):

- ✓ **Physiological Changes**

1. **Nutrient Imbalance**: High zinc concentrations can interfere with the uptake and transport of essential nutrients, particularly iron and magnesium. This disruption can lead to deficiencies, causing further physiological stress.
2. **Oxidative Stress**: Elevated zinc levels can generate reactive oxygen species (ROS), leading to oxidative stress. This condition damages cellular structures, including membranes, proteins, and nucleic acids, ultimately impairing plant metabolism.
3. **Photosynthesis Inhibition**: Zinc toxicity can reduce chlorophyll synthesis, leading to chlorosis (yellowing of leaves). This reduction in chlorophyll affects the plant's ability to photosynthesize efficiently, resulting in decreased energy production.

✓ **Morphological Changes**

1. **Stunted Growth**: Excess zinc can inhibit root and shoot development, resulting in stunted growth. The overall size of the plant may be reduced, impacting its competitiveness and reproductive success.
2. **Leaf Morphology Alterations**: Plants exposed to high zinc levels often exhibit changes in leaf structure, such as reduced leaf area, leaf curling, and necrosis (death of tissue). These morphological changes can further hinder photosynthetic capacity and overall health.
3. **Root System Damage**: Zinc toxicity can lead to root damage, characterized by reduced root length and biomass. A compromised root system limits water and nutrient uptake, exacerbating the plant's stress response.

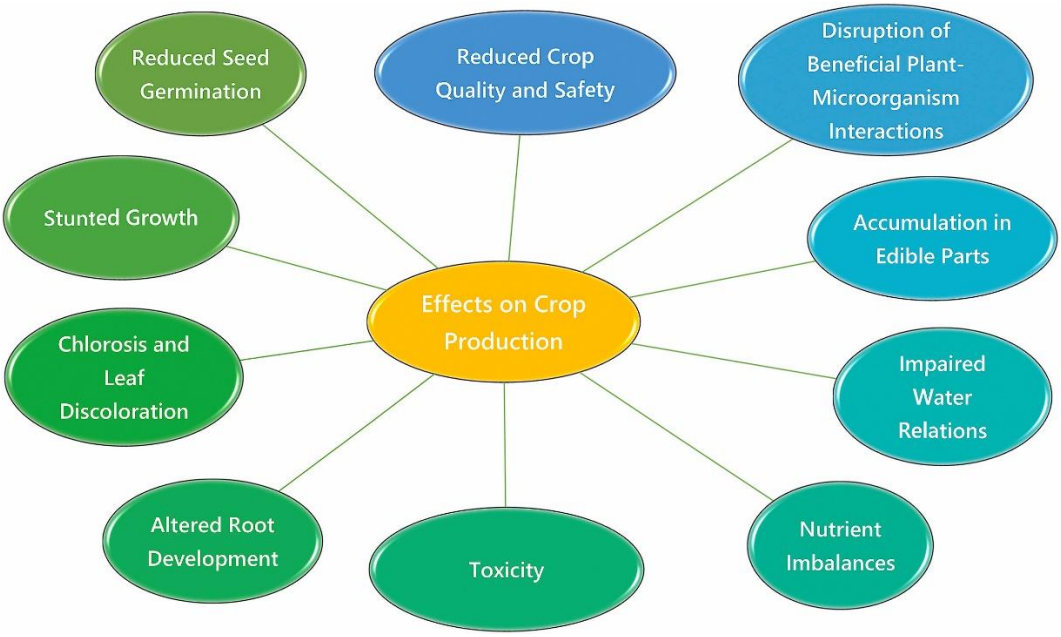



Figure 4: Effects of heavy metals HMs on plants and crop production (Angon et al., 2024)

## *1.2 Role of Arabidopsis helleri*

### 1.2.1 Characteristics of *Arabidopsis halleri* as a hyperaccumulator and tolerant plant.

*Arabidopsis halleri* is increasingly recognized for its remarkable ability to hyperaccumulate heavy metals, particularly zinc and cadmium, making it a valuable species for phytoremediation efforts. This small flowering plant, a relative of the model organism *Arabidopsis thaliana*, thrives in metal-rich soils, often found in regions contaminated by mining and industrial activities (Zhao et al., 2001) (Fig.5). Its unique physiological and biochemical adaptations allow it to tolerate and sequester high concentrations of metals, which would be toxic to most other plants.

Figure 5*: Arabidopsis helleri* in natural habitat

One of the key features that makes *A. halleri* an effective hyperaccumulator is its efficient metal uptake and transport mechanisms. The plant possesses specialized root structures and transport proteins that facilitate the absorption of heavy metals from the soil. Once taken up, these metals are stored in vacuoles and other cellular compartments, effectively reducing their bioavailability in the environment (Fig.6).

This capability not only helps in cleaning up contaminated sites but also contributes to the restoration of soil health and ecosystem balance (Bert et al., 2000; Cosio et al., 2004; Sarret et al., 2002; Schvartzman et al., n.d.; Zhao et al., 2001).

Research has shown that *A. halleri* can accumulate zinc concentrations exceeding 10,000 mg/kg in its tissues, significantly higher than typical plant thresholds (Zhao et al., 2006). This property has led to its use in bioremediation projects aimed at rehabilitating polluted lands. By cultivating *A. halleri* in contaminated areas, it is possible to extract heavy metals from the soil, thereby mitigating environmental risks and promoting biodiversity.

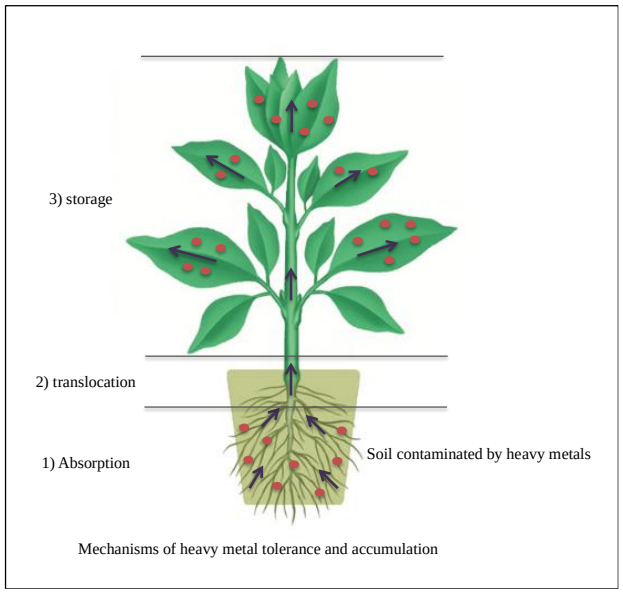



Figure 6: Mechanisms of heavy metal tolerance and accumulation

Moreover, the genetic and molecular studies of *A. halleri* have provided insights into the mechanisms of metal tolerance and accumulation (Fig.7). Understanding these processes can aid in the development of genetically engineered plants with enhanced capabilities for phytoremediation (Karam et al., 2019; Pant et al., 2023; Pauwels et al., 2008). As global concerns about heavy metal pollution grow, the role of *Arabidopsis halleri* as a hyperaccumulator will be crucial in developing sustainable strategies for environmental cleanup and management.

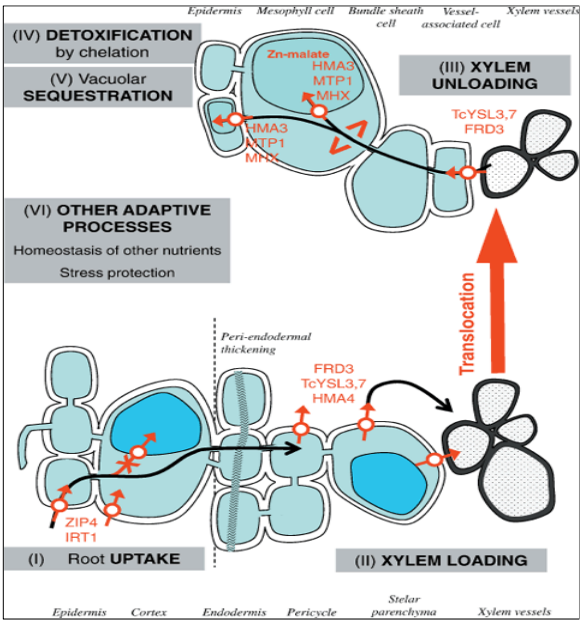



Figure 7: The main genes involved in the hyper-accumulation of Zn (Verbruggen et al., 2009)

### 1.2.2 *Arabidopsis halleri* as a bioindicator of Zn contamination.

*Arabidopsis halleri* has emerged as a prominent bioindicator for assessing zinc contamination in soils due to its unique ability to tolerate and accumulate heavy metals, particularly zinc. This species thrives in zinc-rich environments, making it an ideal model for studying metal uptake and phytoremediation processes. Research has shown that *A. halleri* can bioaccumulate significant amounts of zinc in its tissues without exhibiting severe phytotoxic effects, allowing scientists to monitor soil contamination levels effectively (Dietrich et al., 2021). The phenotypic responses of *A. halleri*, including changes in leaf morphology, chlorophyll content, roots length, Photosystem II yield $\Phi_{PSII}$ , nerves color….etc serve as reliable indicators of zinc stress, providing valuable insights into the extent of soil contamination (Karam et al., 2019). By leveraging the natural tolerance mechanisms of *A. halleri*, researchers can develop sustainable strategies for soil remediation and environmental monitoring, underscoring its importance in ecological studies and agricultural practices (Marschner et al., 2011).

## *1.3 AI in Environmental Science*

### 1.3.1 AI approaches that used in environmental science

The field of environmental science has seen a growing adoption of various AI approaches to tackle complex challenges (Konya & Nematzadeh, 2024). Here are some of the key AI techniques used in environmental science:

1. Machine Learning:
    - Supervised Learning Algorithms:
        - Logistic Regression
        - Decision Trees
        - Random Forests
        - Support Vector Machines
    - Unsupervised Learning Algorithms:
        - Clustering (e.g., K-means, DBSCAN)
        - Principal Component Analysis (PCA)

These algorithms are used for tasks like habitat modeling, species distribution prediction, land use/cover change analysis, and environmental risk assessment.

2. Deep Learning:
    - Convolutional Neural Networks (CNNs):

Used for image-based applications, such as satellite imagery analysis, remote sensing, and object detection in environmental monitoring.

- Recurrent Neural Networks (RNNs):

Employed for time-series data analysis, like forecasting environmental variables, predicting extreme events, and modeling ecosystem dynamics.

- Generative Adversarial Networks (GANs):

Utilized for synthetic data generation, which can be useful in data-scarce environmental domains.

3. Natural Language Processing (NLP):

Applied to analyze environmental reports, policies, and scientific literature to extract insights, identify emerging trends, and support decision-making.

4. Reinforcement Learning:

Used for optimizing environmental management strategies, such as resource allocation, energy management, and wildlife conservation.

5. Hybrid Approaches:

Combining multiple AI techniques, such as integrating machine learning models with physical or statistical models, to leverage the strengths of different approaches.

6. Explainable AI (XAI):

Developing AI models that can provide interpretable and transparent explanations for their predictions, which is crucial in environmental decision-making.

7. Federated Learning:

Enabling collaborative learning across distributed environmental datasets without the need to centralize the data, addressing privacy and data sovereignty concerns.

8. Edge Computing and IoT:

Deploying AI models on edge devices, such as sensors and drones, for real-time environmental monitoring and decision-making at the source.

These AI approaches are being actively employed in a wide range of environmental applications, including climate change modeling, biodiversity conservation, natural resource management, pollution control, and sustainable urban planning, among others. The integration of AI with environmental science is driving innovative solutions and transforming the way we understand, manage, and protect our natural world.

### 1.3.2 Overview of AI applications in environmental monitoring.

Artificial Intelligence (AI) is significantly enhancing environmental monitoring through various innovative applications. For instance, AI algorithms, particularly machine learning and deep learning, are revolutionizing remote sensing by analyzing satellite and aerial imagery to detect changes in land use, deforestation, and ecosystem health (Shi et al., 2020). Predictive modeling powered by AI enables the forecasting of environmental changes, such as climate impacts and pollution dispersion, by analyzing historical data to provide insights into future trends (Amiri et al., 2024). In air quality monitoring, AI systems process data from sensors to track real-time pollutant levels, helping communities respond effectively to health risks (Li et al., 2021). Similarly, AI applications in water quality assessment utilize sensor data to detect contaminants and monitor water health, allowing for rapid responses to pollution events (Yang et al., 2022). Additionally, AI plays a crucial role in disaster response by analyzing diverse data sources to predict natural disasters and optimize resource allocation during emergencies.

## *1.4 Logistic Regression*

### 1.4.1 Logistic Regression: The Art of Predicting Probabilities

Logistic regression is a powerful statistical technique used to model the relationship between a binary or categorical dependent variable and one or more independent variables. The key principles of logistic regression are:

1. Probability Modeling: Logistic regression estimates the probability of a binary outcome (e.g., success or failure, yes or no) based on the values of the independent variables. The output of a logistic regression model is a probability between 0 and 1, which can be interpreted as the likelihood of the event occurring.
2. Non-linear Relationship: Unlike linear regression, which models a linear relationship between the independent variables and the dependent variable, logistic regression can capture non-linear relationships by transforming the dependent variable using the logistic function (Fig.8).

3. Logit Transformation: Logistic regression uses the logit transformation, which converts the probability of an event occurring into the log of the odds ratio. This allows the model to work with a linear combination of the independent variables.

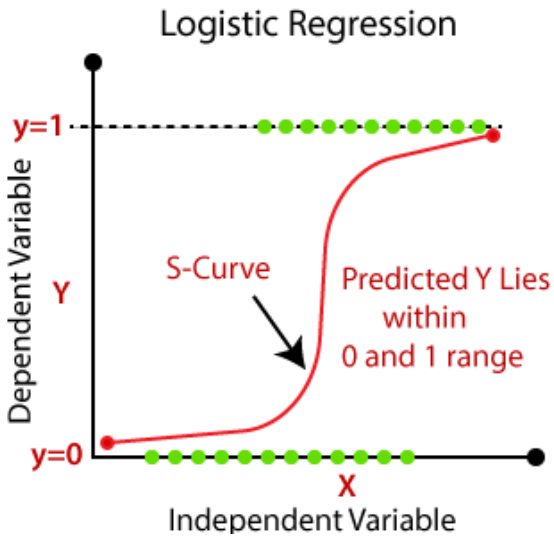



Figure 8: Non-linear Relationship of logistic regression

In terms of usage, logistic regression is widely applied in various fields, such as:

- Healthcare: Predicting the likelihood of a patient's disease, recovery, or response to treatment.
- Marketing: Analyzing factors that influence customer conversion, churn, or purchase decisions.
- Finance: Assessing the risk of loan defaults or credit card fraud.
- Social Sciences: Studying the factors that contribute to voting behavior, educational outcomes, or social phenomena.

The applications of logistic regression are diverse and versatile. Some common use cases include:

- Classification: Predicting whether an observation belongs to one of two or more categories (e.g., malignant vs. benign tumor).

- ❖ Risk Assessment: Estimating the probability of an event occurring, such as the likelihood of a customer defaulting on a loan or a patient developing a certain disease.
- ❖ Hypothesis Testing: Examining the significance of the relationship between independent variables and the binary dependent variable.
- ❖ Model Interpretation: Interpreting the coefficients of the logistic regression model to understand the relative importance and direction of the independent variables in predicting the outcome.

### 1.4.2 Logistic regression in the field of environmental science

Logistic regression has numerous applications in the field of environmental science. Such as:

1. Habitat Modeling and Species Distribution:
    - Logistic regression is commonly used to model the probability of a species' presence or absence based on environmental factors such as climate, land use, vegetation, and topography (Leach et al., 2016).
    - This helps ecologists and conservation biologists understand the habitat requirements of different species and predict their potential distribution (Gotelli & Stanton-Geddes, 2015).
2. Environmental Risk Assessment:
    - Logistic regression can be employed to assess the risk of environmental hazards, such as the likelihood of oil spills, chemical contamination, or natural disasters (e.g., floods, landslides) (Gotelli & Stanton-Geddes, 2015).
    - By modeling the relationship between environmental variables and the occurrence of these events, logistic regression aids in risk management and decision-making.
3. Land Use and Land Cover Change:
    - Logistic regression is used to model the probability of land use or land cover changes, such as deforestation, urbanization, or agricultural expansion.

- This helps researchers and policymakers understand the drivers of land use changes and develop strategies for sustainable land management.

4. Ecosystem Health Assessment:
    - Logistic regression can be used to assess the health of ecosystems by modeling the relationship between environmental indicators (e.g., water quality, soil properties, biodiversity) (Avila et al., 2018; Chaplin-Kramer et al., 2017; Joimel et al., 2017) and the presence or absence of healthy ecosystem functions.
    - This information is crucial for monitoring and managing the overall well-being of natural environments.
5. Environmental Impact Assessment:
    - Logistic regression is employed to evaluate the potential environmental impacts of human activities, such as the construction of roads, dams, or industrial facilities.
    - By modeling the probability of adverse environmental outcomes, logistic regression supports decision-making and helps mitigate the environmental consequences of development projects.
6. Climate Change Adaptation:
    - Logistic regression is used to model the likelihood of climate change-induced events, such as the occurrence of extreme weather, sea level rise, or species range shifts.
    - This information is essential for developing effective adaptation strategies and resilience measures in the face of a changing climate.

The versatility of logistic regression makes it a valuable tool in environmental science, allowing researchers and policymakers to better understand, predict, and manage the complex relationships between human activities, environmental factors, and ecological outcomes.

### *1.5 Machine Learning ML for predicting environmental metals pollutant*

Among the toxic substances released into the environment by human activities, metals are present in high quantity in the environment and thus require special attention because the short or long exposure to metallic pollution such as Zinc, cadmium and lead are known to cause harmful effects on living systems (terrestrial and aquatic ecosystem). Thus, predicting and treating this pollution is essential to save all living species.

Machine learning (ML) has emerged as a powerful tool in the detection and monitoring of metal contamination in the environment. Unlike traditional analytical methods, ML-based approaches can uncover the intricate and nonlinear relationships between metal pollutants and the complex factors influencing their dynamics. By leveraging large and diverse datasets, ML models can learn to identify patterns, predict concentrations, and assess ecological risks associated with metal contamination across various environmental media, including soil, water, and air.

Recent advancements in ML have enabled researchers to develop innovative models that can effectively source apportion metal pollutants, detect their presence with high accuracy, quantify their concentrations, and evaluate their potential ecological impacts. These models can incorporate a wide range of input variables, such as geochemical data, remote sensing imagery, and environmental monitoring measurements, to provide comprehensive and data-driven insights into metal pollution. The flexibility and adaptability of ML techniques have also allowed for the development of early warning systems and decision-support tools to aid in the prevention and management of metal contamination, ultimately contributing to the protection of human health and the environment.

As the field of environmental science continues to grapple with the persistent challenges posed by metal pollutants, the integration of ML-based approaches holds immense promise. By leveraging the power of these advanced analytical methods, researchers and policymakers can gain a deeper understanding of metal pollution dynamics, optimize monitoring and remediation strategies, and work towards a more sustainable and resilient future.

# Chapter II
# Literature study

# 2 Literature study

## *2.1 Studies using ML to detect metals contamination and accumulation*

- Recent study was performed to predict HMs (heavy metals) contents in crops based on the HMs contents in soil and other available information about the soil. Machine learning algorithm GBM, RF, and GLM were adopted to predict the BAFs of different HMs in soil-crop ecosystems based on 13 auxiliary variables, and the importance of the different variables in the models were quantified (Hu et al., 2020). This model could be used to assist the prediction of heavy metal contents in crops based on heavy metal contents in soil and other covariates, and can significantly reduce the cost and time requirements involved with laboratory analysis. It can also be used to quantify the importance of variables and identify potential control factors in heavy metal bioaccumulation in soil-crop ecosystems.

Note: The BAFs of the transfer of HMs from soil to crops were calculated using the following equation: BAFs = Ccrop / Csoil where Ccrop and Csoil are the HM contents in crops and soil, respectively.

- Other study was conducted to explore the performance of a back propagation neural network (BPNN) to predict plant uptake and accumulation of metallic Engineered Nanoparticles ENPs by terrestrial plants in both hydroponic and soil systems (Wang et al., 2021).

  this study aims also to identify key factors governing the uptake and accumulation of metallic ENPs by terrestrial plants;

  and to develop mathematical equations to estimate root concentration factor (RCF) and translocation factor (TF) values for different ENPs and plant species from selected quantifiable properties of ENPs.

  By imposing the capabilities of BPNN models, the researchers aimed to gain a deeper understanding of the complex interactions between metallic ENPs and plant systems, and to establish predictive tools that can assist in the assessment and management of ENP-related environmental risks. The findings from this study demonstrate that machine

learning could be an effective tool to predict the uptake and translocation of ENPs by plants.

Various machine learning (ML) algorithms have been extensively employed to model the adsorption of heavy metals in water systems. These algorithms include artificial neural networks (ANNs), random forests, and gradient boosted machines.

- For instance, (El Hanandeh et al., 2021) applied ML techniques to assess the availability and toxicity of heavy metals in soils amended with compost or biochar. Additionally, Li et al. (2021a) and Li et al. (2021b) developed ML-based models to estimate heavy metal accumulation in soil-crop systems.
- Cipullo et al., 2019) also contributed to this field by imposing ML algorithms to improve the understanding of heavy metal adsorption processes in water environments. The versatility and predictive power of these ML approaches have made them valuable tools for researchers and practitioners working on the assessment and management of heavy metal contamination in diverse environmental matrices.

The widespread adoption of ML techniques in modeling heavy metal adsorption highlights their potential to provide reliable and data-driven insights, which can inform decision-making and support the development of effective strategies for the remediation and prevention of heavy metal pollution in water systems.

## *2.2 Studies using plant images and machine learning ML*

### 2.2.1 Plant disease recognition

- (Maeda-Gutiérrez et al., 2020) aims to compare the performance of various convolutional neural network (CNN) architectures in classifying tomato plant diseases using plants leaves. The primary objective was to develop an automated system that can effectively identify and diagnose plant diseases, which would be highly beneficial for agricultural experts and technicians.
- The study of H.S. Abdullahi, *et al*. 2017 use the off-the-shelf convolutional neural network (ConvNet) representations to estimate plant health on a maize plantation.

The results showed an impressive average prediction accuracy of 99.58%, which is the best performance achieved compared to other techniques.

- H Al-Hiary; *et al*. 2011 presents the application of two machine learning techniques, K-means clustering and Neural Networks (NNs), for the clustering and classification of plant leaf diseases. Five specific diseases were tested: Early scorch, Cottony mold, ashen mold, late scorch, and tiny whiteness. The K-means clustering algorithm is used to group the plant leaf images based on their visual characteristics, while the Neural Network models are employed for the classification of the identified disease types.

  The experimental results demonstrate that the proposed approach is a valuable and effective method for the accurate detection of leaf diseases, with a minimal computational effort required.

### 2.2.2 Plant classification

- The study of (Lee et al., 2017) aim to investigate the use of deep learning techniques, specifically convolutional neural networks (CNNs), to harvest discriminatory features from leaf images and apply them as classifiers for plant identification.
  The main result of the study is that learning the features using CNNs can provide better feature representations of leaf images compared to using hand-crafted features.
- The main aim of (Anubha Pearline et al., 2019) paper was to investigate and compare the performance of traditional feature extraction methods and deep learning approaches for plant species recognition.
  Many Deep learning approaches were used including VGG 16 and VGG 19 Convolutional Neural Networks (CNNs).
  The main result of the study is that the deep learning-based approaches, particularly the VGG CNN models, outperformed the traditional feature extraction and classification methods in terms of plant species recognition accuracy across the tested datasets.

**Conclusion:**

No studies were conducted on *Arabidopsis halleri* to predict soil contamination using machine learning approaches. For this reason, we decide to conduct this work in order to use this plant as a bioindicator of soil contamination

## The main aims of this study

The main aims of this study were:

1. To develop a logistic regression model to predict Zn contamination in soil based on various plant features, extracted from 1000 individuals of *A.halleri* including:
    a. Morphological and physiological features measured using traditional methods.
    b. Plant color feature (chlorophyll content) estimated using plant images.
2. To evaluate the performance of the logistic regression model not only on the training data, but also on its ability to generalize to new, unseen data.
3. To assess the importance of the different features, including the plant color feature obtained from images, in the logistic regression model for predicting Zn contamination.
4. To develop a multiple machine learning model to predict Zn concentration in *A.halleri* leaves based on plant features, explained previously.

# Chapter III
# Materials and Methodology

# 3 Methodology

## *3.1 Data Collection*

### 3.1.1 Description of data and images collection methods for *Arabidopsis halleri*.

*Arabidopsis helleri* images were obtained from a conducted study by (Karam et al., 2019). The study was performed using two cuttings from each *Arabidopsis helleri* plant, resulting in a total of 1,062 plants. These cuttings were initially nurtured in a greenhouse on sand for eight weeks to encourage root development. Once established, the plants were transferred to pots and cultivated in a hydroponic solution under carefully controlled conditions, including temperature, humidity, light, and a moving table (Fig.9). The plants were then randomly allocated across the pots to ensure unbiased results. Then, the Zn tolerance test started by assigning one replicate per plant to either of two conditions of Zn concentrations in the hydroponic solution: 10 µM as a non-polluted condition and 2000 µM as a polluted condition (Meyer et al., 2010).

Figure 9: Hydroponic culture

The Zn tolerance test lasted six weeks. Zn tolerance was estimated by measuring five phenotypic features for all individuals and listed in the data file: biomass-related features that are root length (LonR2), leaf width (moyLarF), root dry biomass (massR) and shoot dry biomass (massF), and two physiological feature that are photosystem II yield (moyFluo2) and Zn content in the leaves

(Zn_acc) (Table 1). All these features were measured by the mean of classical methods such as, ruler, balance, fluorometer and chemical analyses.

An image was taken for each plant two times during the experiment, in this study we analyzed images taken at the end of the experiment. Digital camera was used to take these photos under the same condition for each plant (light, distance, camera parameters) (Fig.10).

Table 1: Description of *A.halleri* Morphological and Physiological Features

| **Feature Name** | **Feature Type** | **Explanation** |
|---|---|---|
| **LonR2** | Morphological Feature | Measures the root length of the plant, indicating growth and development under Zn tolerance conditions. |
| **moyLarF** | Morphological Feature | Represents the average leaf width, which can affect photosynthesis and overall plant health. |
| **massR** | Morphological Feature | Refers to the dry biomass of the roots, providing insights into root development and nutrient uptake capacity. |
| **massF** | Morphological Feature | Indicates the dry biomass of the shoot, reflecting the plant's overall growth and health. |
| **moyFluo2** | Physiological Feature | Represents the yield of photosystem II, an indicator of the plant's photosynthetic efficiency and health. |
| **Zn_acc** | Physiological Feature | Measures the zinc content in the leaves, which is crucial for assessing the plant's ability to tolerate zinc. |
| **Green** | Physiological Feature | Indicates the green color intensity of the plants extracted from RGB images, reflecting chlorophyll content. |

Plant in non-polluted condition

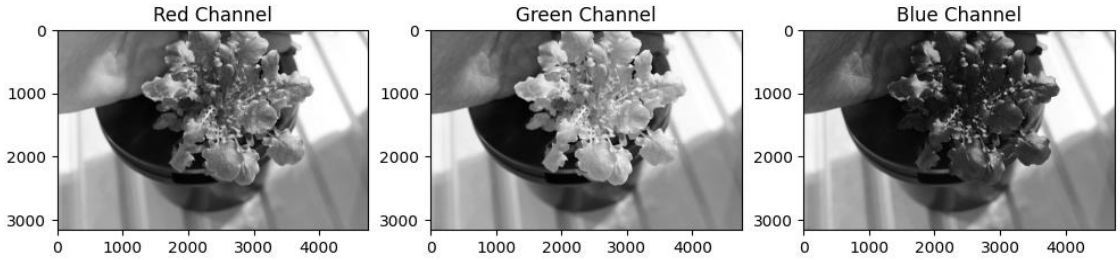
Plant in polluted condition

Figure 10: *Arabidopsis halleri* plants in polluted and non-polluted condition

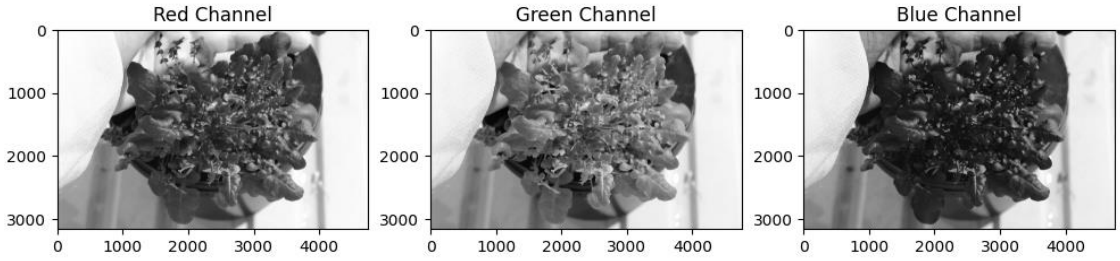
### *3.2 Image processing*

This step is designed to extract the green color content as a new feature from plants images. The green color intensity of the plants reflecting chlorophyll content. For each image, python's PIL, pandas and os libraries were used to converts it to the RGB color space (Fig.11), and extracts the green channel values. Then the average green value was calculated and store.

This method can be useful to analyze plant health or detect color-related issues in images, as the green color is often an important indicator of various biological and environmental factors. By automating the process of extracting and analyzing the green color data, this code can save time and effort for researchers, scientists, or anyone interested in studying the visual characteristics of a collection of images.

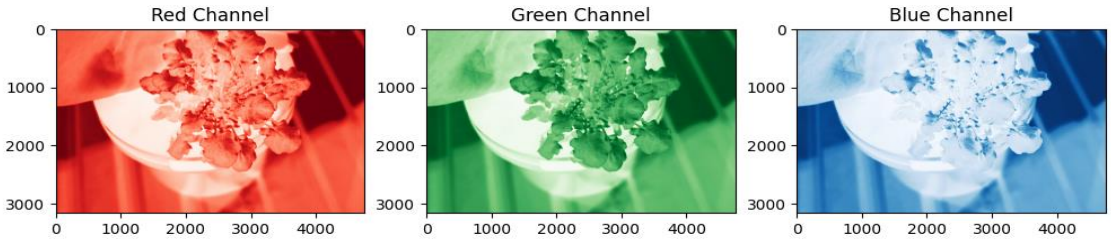



Plant with high content of Zn

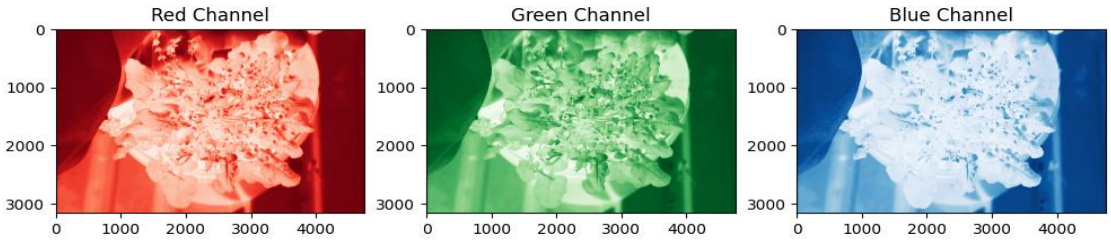



Plant with low content of Zn

Visualization of RGB Channels in Gray scale




Plant with high content of Zn




Plant with low content of Zn

Figure 11: Visualization of RGB Channels in Red, Green and Blue

## *3.3 Data Preprocessing*

### 3.3.1 Data cleaning

Python's pandas and numpy libraries were employed to perform data cleaning on the dataset. Initially, the dataset was loaded from a CSV file, and the first few rows were examined to understand its structure. To address missing values, we eliminate rows containing missing values. We also removed any duplicate entries to ensure the integrity of the dataset. Specific columns were converted to appropriate data types. Furthermore, we renamed certain columns to enhance clarity and reset the DataFrame index to maintain order after the removal of any rows. The cleaned dataset was subsequently saved to a new CSV file, thereby preparing it for further analysis and visualization. This systematic approach to data cleaning is essential for ensuring the reliability and validity of subsequent analyses.

After cleaning, 414 samples in polluted condition and 349 samples in non-polluted condition were chosen to conduct the next step (Fig.12).

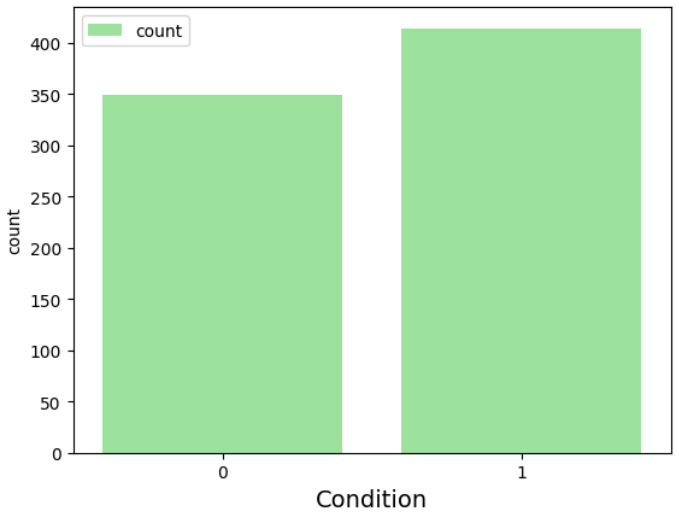



Figure 12: Samples count in polluted (1) and non-polluted (0) conditions

### 3.3.2 Data visualization:

✓ Morphological features

Two crucial libraries were used to build the boxplots:

- matplotlib.pyplot: which facilitates the creation of static, animated, and interactive visualizations in Python.
- Seaborn: a data visualization library built on Matplotlib that offers a high-level interface for producing visually appealing statistical graphics.

Then we construct a single figure featuring four side-by-side boxplots, each representing a morphological features of *Arabidopsis halleri* (leaves width, roots length, roots and leaves biomass) across two Zn concentrations hydroponic solution (polluted = 0 and non-polluted=1) (Fig.13). Each boxplot effectively summarizes the distribution of the respective feature values, enabling comparisons between different conditions. This visualization method is particularly valuable for identifying trends, outliers, and variations within the data, providing insights into how environmental factors influence plant morphology.

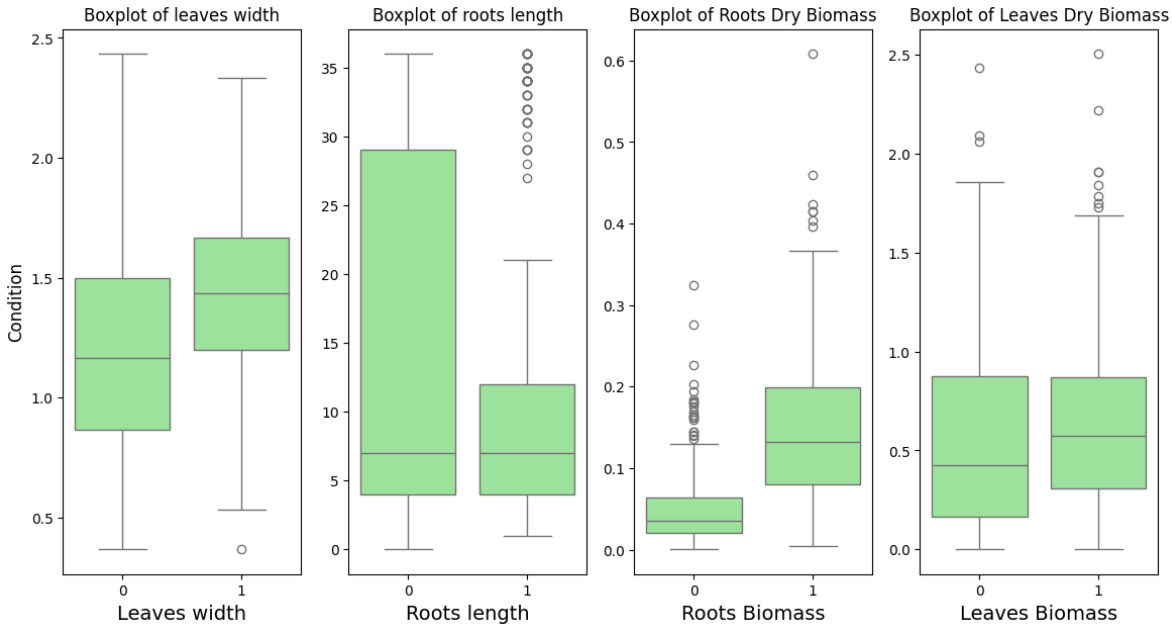



Figure 13: Morphological Features distributions in the non-polluted (0) and polluted (1) conditions for all plants

✓ Physiological features

The analysis is further enriched by examining three additional physiological features of plants: photosystem II yield, zinc (Zn) content in leaves and the green color intensity in plants reflecting chlorophyll content (Fig.14). Utilizing the same boxplot framework, the code generates visual representations for these features. These features are crucial for understanding the overall health and productivity of the plant, as higher yields indicate better photosynthetic performance. Meanwhile, the boxplot for zinc content in leaves provides insights into the nutritional status of the plants, highlighting variations in Zn levels across conditions.

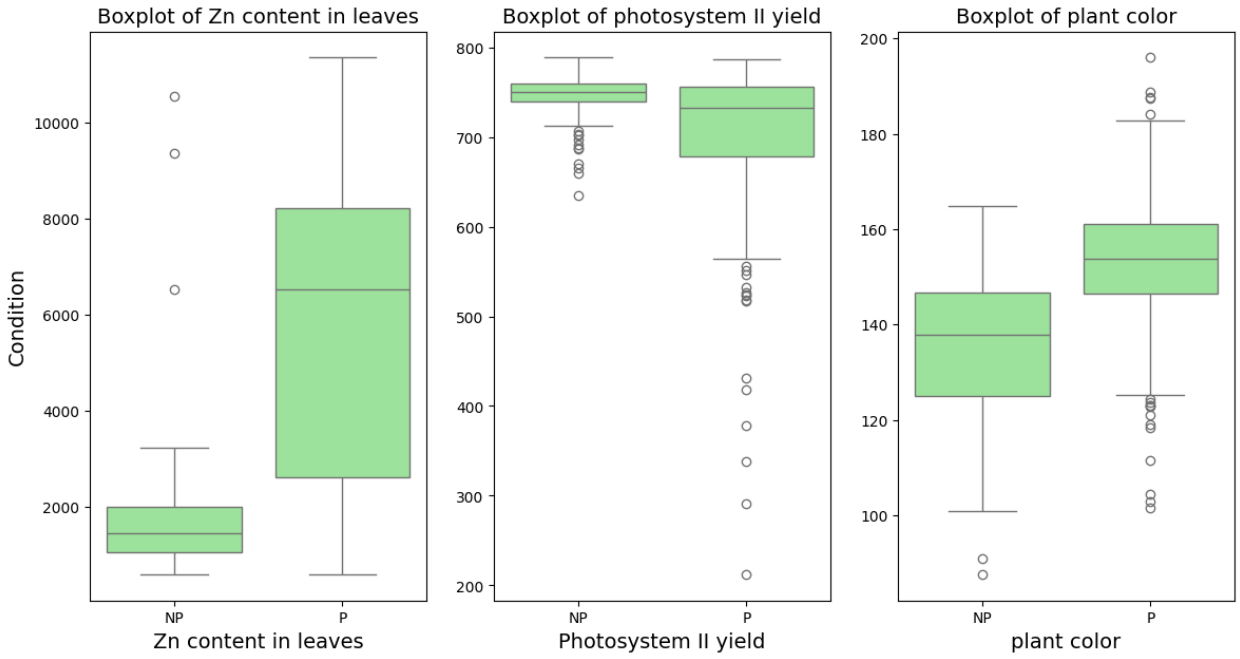



Figure 14: physiological Features distributions in the non-polluted (0) and polluted (1) conditions for all plants

A grid of scatter plots was performed, each box displaying samples distribution or relationship between variables in the dataset. Histograms and scatter plots illustrate all data in both polluted and non-polluted condition (Fig.15). The scatter plots cover a wide range of data types and patterns, making this a comprehensive visualization of the dataset.

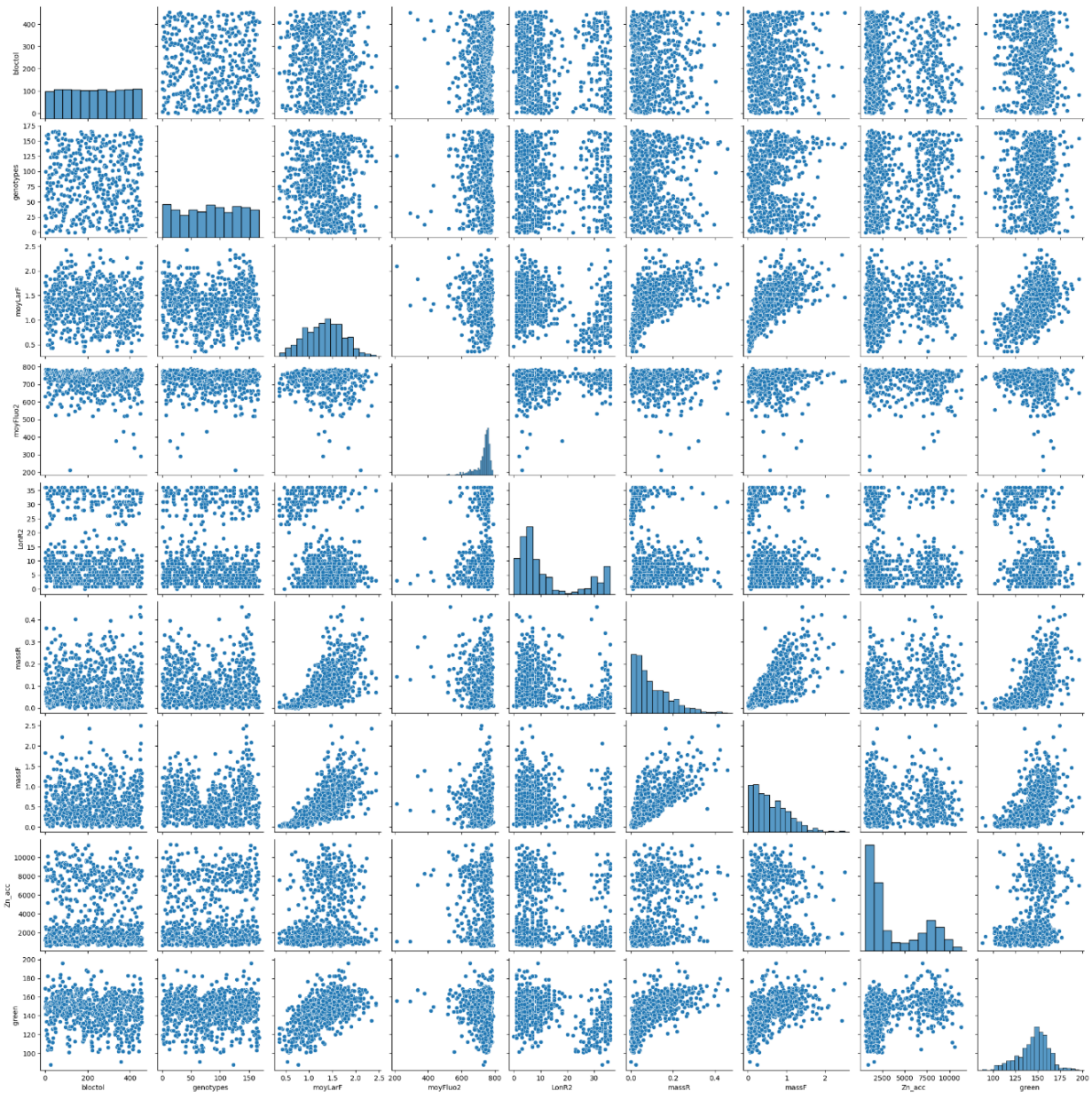



Figure 15: Visualization of relationship between variables using Scatter Plots, for all samples in polluted and non-polluted condition

Another grid of scatter plots was performed, each box displaying samples distribution and relationship between samples in polluted (in orange color) and non-polluted (in green color) condition. This figure show relationship between variables in the dataset. Histograms and scatter plots illustrate all data in both polluted and non-polluted condition (Fig.16). The scatter plots cover a wide range of data types and patterns, making this a comprehensive visualization of the dataset.

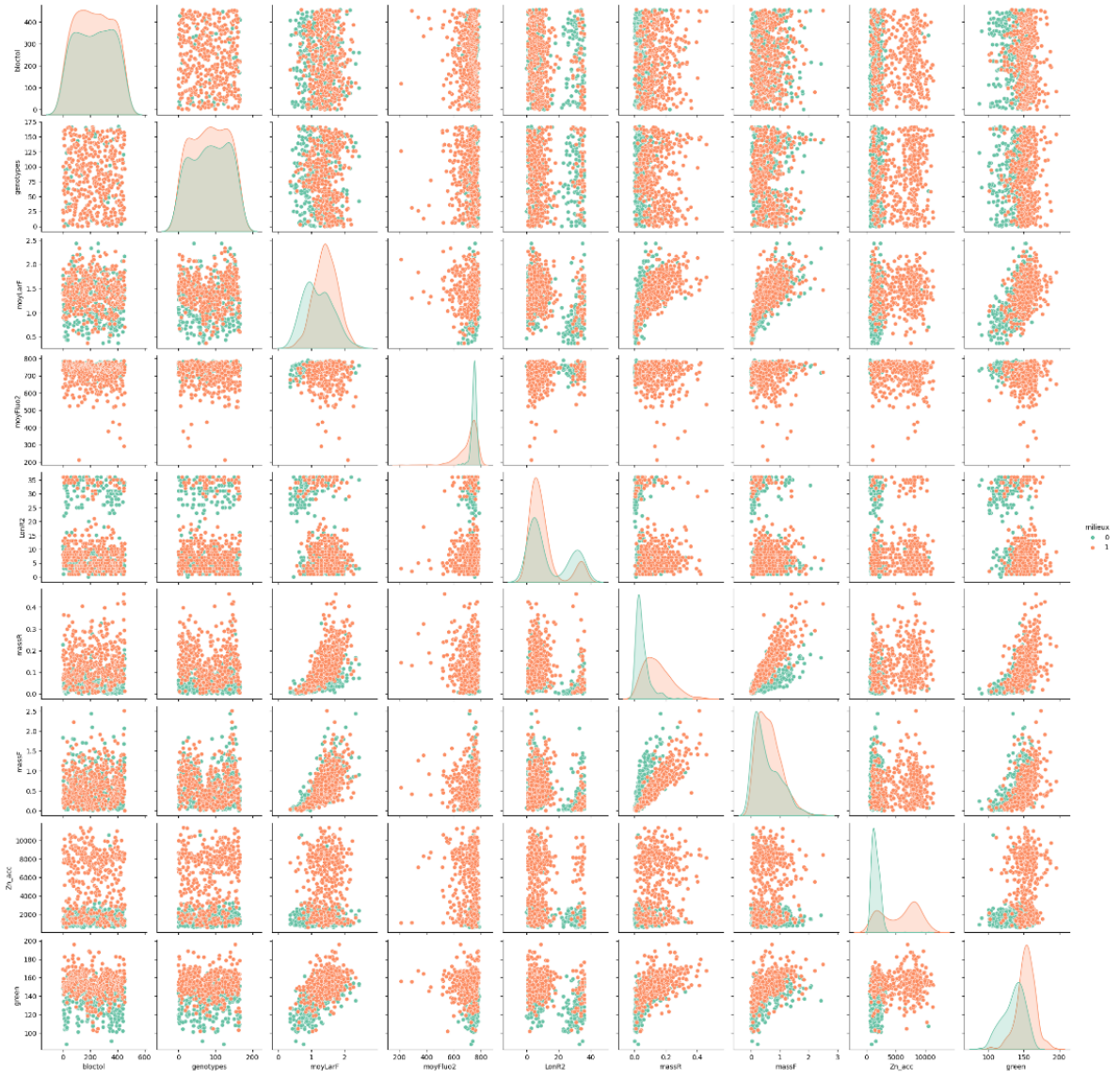



Figure 16: Visualization of relationship between variables using Scatter Plots. In green, samples in polluted condition, in orange, samples in non-polluted condition

### 3.3.3 Statistical analysis:

A heatmap tool was used to display the correlation between all morphological and physiological features. This tool helped to identify patterns and relationships among measured features.

The heatmap represents the correlation coefficients values ranging from -2% to 100% between each pair of features. A coefficient close to 100% indicates a strong positive correlation, meaning that as one feature increases, the other tends to increase as well. Conversely, a coefficient close to -2% signifies a strong negative correlation, where an increase in one feature corresponds to a decrease in the other. A coefficient around 0 suggests little to no correlation between the features (Fig.17).

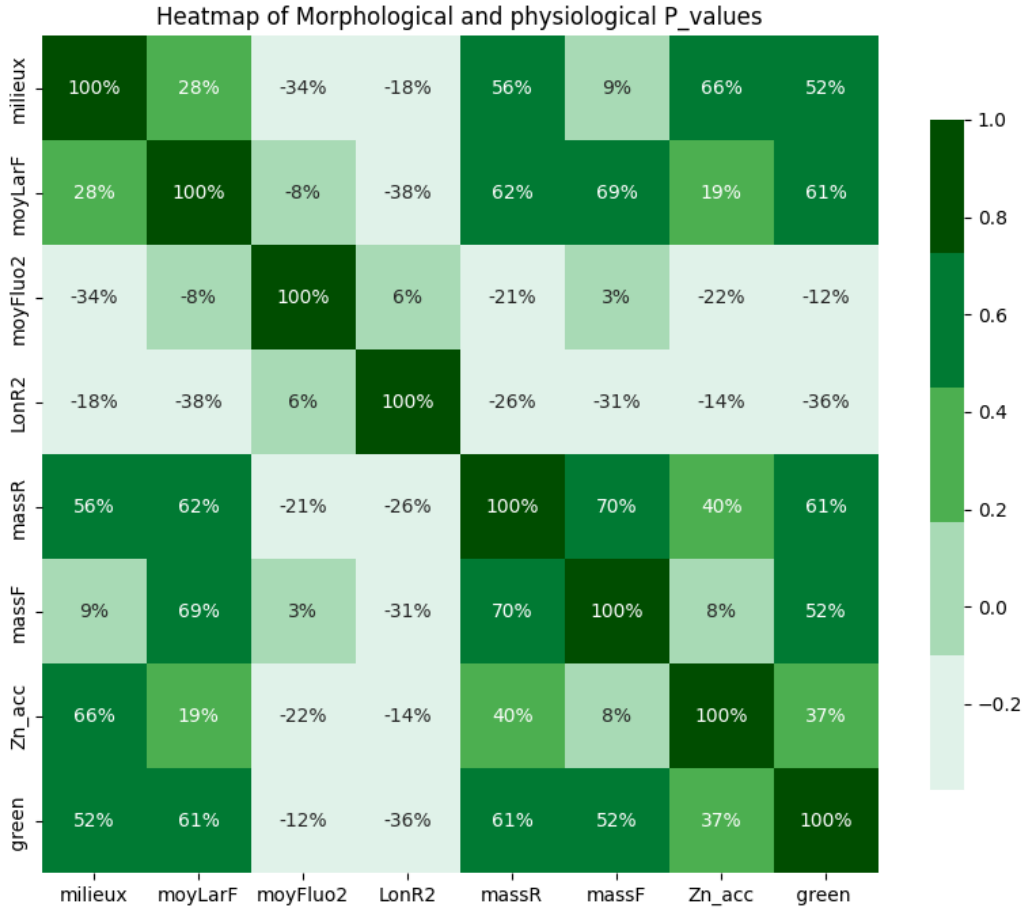



Figure 17: Correlation rates between morphological and physiological features

The color gradient in the heatmap visually represents these correlation values, distinct colors indicating the strength and direction of the relationships. This allows for quick identification of significant correlations, such as whether higher leaf width is associated with increased photosynthetic yield or if higher zinc content correlates with root biomass.

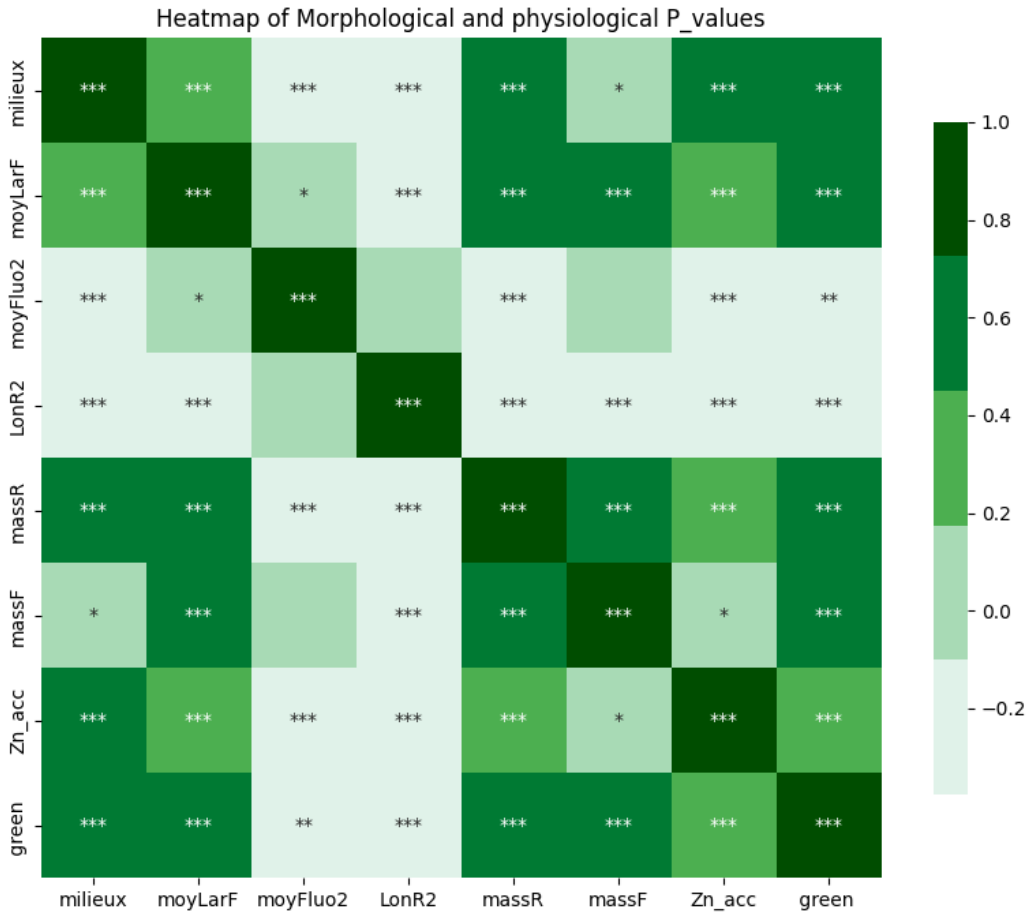



Figure 18: Correlation between morphological and physiological features, stars indicate the significance of this correlation as following: $p < 0.001$ = ***, $p < 0.01$ = **, $p < 0.05$ = *

In analyzing our dataset, the p-values derived from the correlation heatmap provide critical insights into the relationships between the variables. In our analysis, p-values less than 0.001 suggest statistically significant correlations, meaning we can reject the null hypothesis with confidence (Fig.18).

## *3.4 AI Model Development*

### 3.4.1 Data preparation:

Dataset was prepared for training a machine learning model, specifically a Logistic Regression model. The process involves several key steps:

**Splitting the Dataset**:

- We separate the dataset into independent variables X = (features including morphological and physiological features) and the dependent variable Y = (condition). In this case, X consists of selected columns from the dataset that are believed to influence the outcome Y (condition; P and NP, polluted vs non-polluted) the target variable we aim to predict.

  We performed this step twice, once without the feature that obtained from *A.halleri* images named green (chlorophyll content) and second with the values obtained from image processing. We did this in order to compare the results.

- We then split the data into training 75% and testing 25% of the dataset. This is crucial for evaluating the model's performance on unseen data, ensuring that our model generalizes well and does not overfit to the training data.

- We apply standardization to the feature set, transforming the data such that it has a mean of 0 and a standard deviation of 1. This step is important for many machine learning algorithms, including Logistic Regression, as it ensures that all features contribute equally to the distance calculations and gradient descent optimization, improving the model's convergence and performance.

### 3.4.2 Model Training for Logistic Regression

We initialize the Logistic Regression model and fit it to the training data. This process involves learning the relationship between the independent variables X and the dependent variable Y, allowing the model to make predictions based on new data.

Finally, we print the training accuracy, which provides insight into how well the model has learned from the training data. A high training accuracy indicates that the model has effectively captured the underlying patterns in the data.

Overall, this step is essential for setting up the machine learning workflow, allowing us to train and evaluate the Logistic Regression model, ultimately helping us understand and predict the target variable (condition which represent soil contamination) based on the independent features (morphological and physiological).

### 3.4.3 Metrics for evaluating model performance

Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC) were provided using the standard metrics to assess the performance of the logistic regression model (Fig.20). By the mean of these metrics, we could understand the model's ability to correctly classify instances into the two classes (polluted and non-polluted).

- ✓ Accuracy measures the overall proportion of correctly classified instances
- ✓ Precision and Recall focus on the model's performance for the positive class.
- ✓ Precision indicates the proportion of true positives among all instances classified as positive
- ✓ Recall measures the proportion of true positives that the model correctly identified.
- ✓ The F1-Score is the harmonic mean of Precision and Recall, providing a balanced metric that considers both.

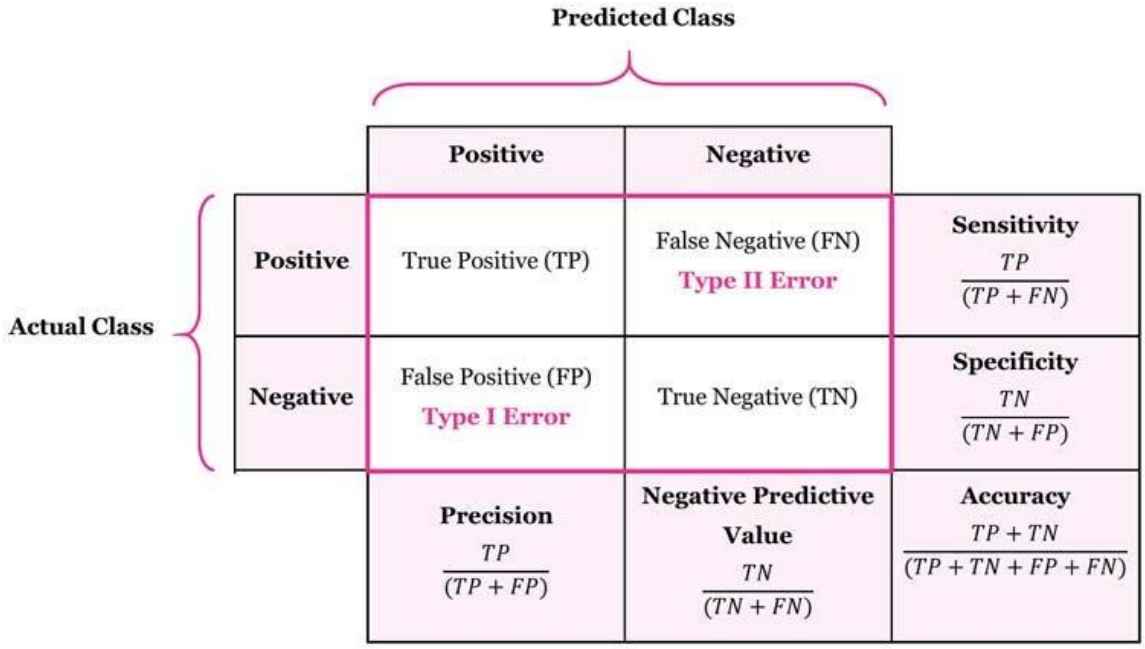



Figure 20: Confusion matrix with advanced classification metrics

### *3.5 Zn content prediction in A.halleri leaves*

A machine learning model was used to predict a target variable that is Zn concentration in *A.halleri* leaves depending on morphological and physiological features including images information. The model combines different regression models (Random Forest and Gradient Boosting) using a Voting Regressor, which aggregates their predictions to improve overall performance.

### 3.5.1 Data preparation

- The first step was to import Libraries like: pandas for data manipulation, sklearn for machine learning models and evaluation metrics, and StandardScaler for data preprocessing.
- We separate the dataset into independent variables X = (including morphological and physiological features) and the dependent variable Y = (Zn concentration in leaves). In this case, X consists of selected columns from the dataset that are believed to influence the outcome Y the target variable we aim to predict.
- We then split the data into training 75% and testing 25% of the dataset.
- Standardization is performed to scale the features such that they have a mean of 0 and a standard deviation of 1.

### 3.5.2 Model development

- Two regression models, Random Forest and Gradient Boosting, were implemented.
- The hyperparameters for the Random Forest model were selected to optimize performance and generalization. The number of trees (n_estimators) was set to 100 and 200, as these values typically balance accuracy and computational efficiency. The maximum depth (max_depth) was chosen as None, 10, and 20 to allow flexibility in capturing complexity while preventing overfitting. Lastly, the minimum samples required to split an internal node (min_samples_split) was set to 2, 5, and 10 to ensure that splits are made with sufficient data, promoting more generalized trees.
- A Voting Regressor is created, which combines the predictions of the best Random Forest and Gradient Boosting models.
- The Voting Regressor is trained on the training dataset. Then, predictions are made on the test dataset.

### 3.5.3 Evaluate model performance

The model's performance is evaluated using three metrics:

- Mean Absolute Error (MAE): Average absolute difference between predicted and actual values.
- Mean Squared Error (MSE): Average of the squares of the errors.
- R-squared Score: Proportion of variance in the dependent variable that can be explained by the independent variables.

# Chapter IV

# Main Results

# 4 Results

## *4.1 Statistical results for all features*

We plotted the features distribution between polluted and non-polluted conditions. Our results show a significant increase in morphological values for all biomass traits in polluted condition P. Conversely, a significant decrease in the physiological values of the photosystem II yield feature and plant color feature were observed in response to Zn toxicity (Fig 13,14). Remarkably, a positive correlation between the photosystem II yield trait, plant color and morphological features appeared, namely for root length, shoot biomass and root biomass (Spearman coefficient) (Fig. 17,18).

The boxplot for photosystem II yield and the plant color illustrates the differences of photosynthesis efficiency between polluted and non-polluted condition (Fig.14), revealing how different factors may influence the plant's ability to harness light energy effectively.

## *4.2 Model Performance for Zn contamination prediction*

### 4.2.1 Without images feature

- **Accuracy**

The training accuracy of the logistic regression model for all features except plant color was a remarkable 0.9423, indicating that the algorithm had an exceptional ability to classify the features in the training dataset into two categories polluted and non-polluted conditions. This level of performance on the familiar training data suggests that the model had effectively learned the underlying patterns.

The model's performance on the test dataset was evaluated, revealing an accuracy of 0.905. This result is particularly notable, as it demonstrates the model's capacity to adapt and thrive in the face of the unpredictable challenges that characterize real-world applications.

- **Confusion matrix**

The provided confusion matrix shows the performance of logistic regression classification model, performed on all features except plant color. The key metrics are:

True Positive (TP): 84, correctly classified as plants exist on polluted condition

False Positive (FP): 12, incorrectly classified as plants exist on non-polluted condition

True Negative (TN): 89, correctly classified as plants exist on non-polluted condition

False Negative (FN): 6, incorrectly classified as plant exist on polluted condition

The overall accuracy of the model is 0.91, which indicates that the model correctly classified 91% of the instances (Fig.21).

This confusion matrix provides a detailed breakdown of the model's performance, highlighting the number of true positive, false positive, true negative, and false negative predictions. This information can be used to further analyze the model's strengths, weaknesses, and potential areas for improvement.

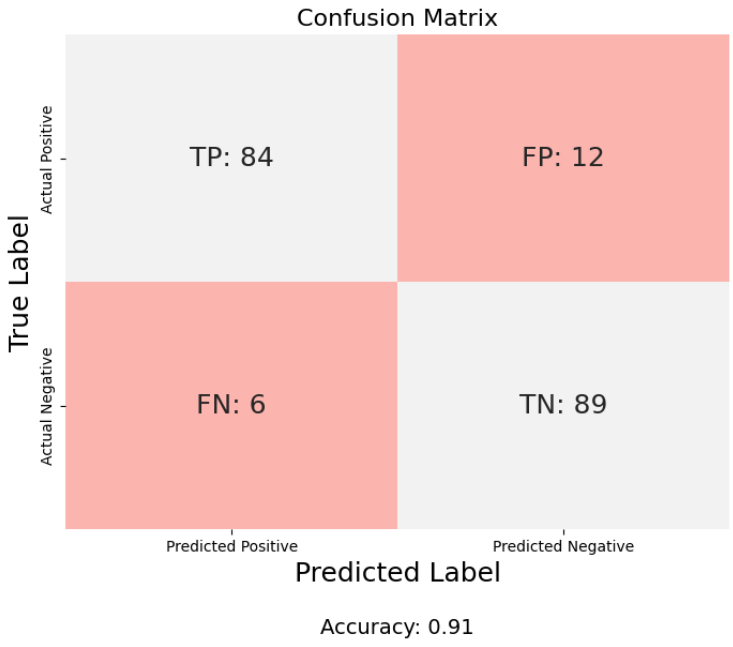



Figure 21: confusion matrix of the model's performance without images features

- **Accuracy, precision, recall, and F1 score**.

The precision is 0.88, indicating that 88% of the positive predictions made by the model were correct. The recall is 0.88, meaning the model correctly identified 88% of the positive instances. The F1-score is 0.91, which is a strong result. The overall accuracy of the model is 0.91, meaning it correctly classified 91% of the instances. Macro Average is the average of the precision, recall, and F1-score across all classes, giving equal weight to each class. The macro average values are all 0.91, indicating consistent performance across classes.

Table 2: Accuracy, precision, recall, and F1 score for the first model, features without images

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.93 | 0.90 | 90 |
| 1 | 0.94 | 0.88 | 0.91 | 101 |
| accuracy | | | 0.91 | 191 |
| macro avg | 0.91 | 0.91 | 0.91 | 191 |
| weighted avg | 0.91 | 0.91 | 0.91 | 191 |

0.9057591623036649

- **Another accuracy test**

Another code was used for accuracy test (pred = log.predict(X_test)), The purpose of this code is to compare the predicted values (pred) with the true values (Y_test) of the test data (Fig.22). This is a common step in the model evaluation process, where the model's predictions are compared to the ground truth to assess the model's performance.

```
#Another accuracy test
pred = log.predict(X_test)
print(pred)
print()
print(Y_test)
```

```
[0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0 1 0 0 1 0 0 1 1 0 1 1 1
 1 0 0 0 1 1 1 1 0 0 1 0 1 0 1 1 0 1 1 0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 1 0
 0 0 0 1 0 0 1 1 0 1 1 0 0 0 1 0 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0
 0 0 1 1 0 1 0 0 0 1 1 1 0 1 1 0 1 0 0 0 1 1 1 0 0 0 1 0 1 1 1 0 0 1 0 1 1
 1 1 1 0 1 1 0 1 1 1 0 0 0 0 1 1 1 0 1 0 1 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 0
 1 1 1 0 1 1]

[1 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0 1 1 0 1 0 0 1 0 0 1 1 1
 1 0 0 0 1 1 1 1 0 0 1 0 1 0 1 1 0 1 1 0 0 1 0 1 1 0 1 0 1 0 1 0 1 1 0 1 0
 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0
 0 0 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 1 0 1 1
 1 1 1 0 1 1 0 1 1 1 0 0 0 1 1 1 1 0 1 0 1 0 0 1 1 0 1 0 1 1 0 1 1 0 0 0 0
 1 1 1 0 1 1]
```

Figure 22: Accuuracy test that compare predict values with the true values

By printing both the predicted values (pred) and the true values (Y_test), the developer can visually inspect the model's predictions and identify any discrepancies or patterns. This information can be further used to analyze the model's strengths, weaknesses, and potential areas for improvement.

### 4.2.2 With images data

- **Accuracy**

The training accuracy of the logistic regression model for all features including plant color was a remarkable 0.954, indicating that the algorithm had an exceptional ability to classify the features in the training dataset into two categories polluted and non-polluted conditions.

The model's performance on the test dataset was evaluated, revealing an accuracy of 0.94. This result is particularly notable, as it demonstrates the model's capacity to adapt and thrive in the face of the unpredictable challenges that characterize real-world applications.

- **Confusion matrix**

The provided confusion matrix shows the performance of logistic regression classification model, performed on all features including plant color (Fig.23).

True Positive (TP): 77, correctly classified as plants exist on polluted condition

False Positive (FP): 10, incorrectly classified as plants exist on non-polluted condition

True Negative (TN): 99, correctly classified as plants exist on non-polluted condition

False Negative (FN): 1, incorrectly classified as plant exist on polluted condition

The overall accuracy of the model is 0.94, which indicates that the model correctly classified 94% of the instances.

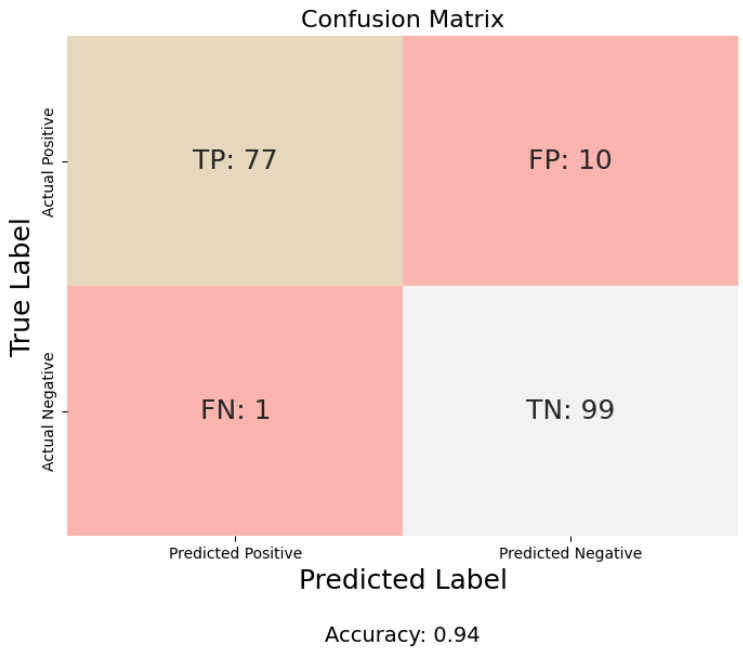



Figure 23: confusion matrix of the model's performance with images features

- **Accuracy, precision, recall, and F1 score**.

The precision is 0.89, indicating that 89% of the positive predictions made by the model were

correct. The recall is 0.91, meaning the model correctly identified 91% of the positive instances. The F1-score is 0.95, which is a strong result. The overall accuracy of the model is 0.95, meaning it correctly classified 95% of the instances. Macro Average is the average of the precision, recall, and F1-score across all classes, giving equal weight to each class. The macro average values are all 0.94, indicating consistent performance across classes (Table 3).

Table 3: Accuracy, precision, recall, and F1 score for the first model, with images feature

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.99 | 0.93 | 78 |
| 1 | 0.99 | 0.91 | 0.95 | 110 |
| accuracy | | | 0.94 | 188 |
| macro avg | 0.94 | 0.95 | 0.94 | 188 |
| weighted avg | 0.95 | 0.94 | 0.94 | 188 |

- **Another accuracy test**

Another code was used for accuracy test (pred = log.predict(X_test)), The purpose of this code is to compare the predicted values (pred) with the true values (Y_test) of the test data (Fig.24). This is a common step in the model evaluation process, where the model's predictions are compared to the ground truth to assess the model's performance.

```
0.9414893617021277
[0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0
 0 0 0 1 1 1 0 0 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 1 0 1 1 1 0 0 1 1 1
 1 1 1 1 0 0 0 1 0 0 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1
 0 1 1 0 0 1 1 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 0
 1 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 0 1 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0
 1 0 1]

[1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 0 0 1 1 0 0 1 0
 0 1 0 1 1 1 0 0 1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1
 1 1 1 1 0 0 0 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1
 0 1 1 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 1 1 1 0
 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 1 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 1 1 0
 1 0 1]
```

Figure 24: Accuracy test that compare predict values with the true values

### *4.1 Model Performance for Zn content prediction*

The evaluation metrics for the combination of Random Forest and Gradient Boosting models indicate this results:

1. Mean Absolute Error (MAE): is 2277, this means that, on average, the model's predictions are off by approximately 2277 units from the actual values.
2. Mean Squared Error (MSE): is 7560384, this indicates the average of the squared differences between predicted and actual values which is large. A lower MSE is preferable.
3. R-squared Score: is 0.2383, this suggests that about 23.83% of the variance in the target variable is explained by the model. This indicates that the model is not capturing a significant portion of the variance.

# Chapter V
# Discussion and perspective

# 5 Discussion

## *5.1 Interpretation of Statistical Results*

Zn exposure caused a general decrease in the photosynthetic yield of the F2 plants while, unexpectedly, biomass was increased with Zn exposure. These findings differ from what was observed by (Meyer et al., 2010) in the same conditions as our experiment, where the biomass-related traits (shoot and root biomass, root length and leaf width) had lower values in the polluted condition. This contradictory behavior could be related to different genetic backgrounds among the compared plants. The increase in the biomass-related features (morphological) we observed could be mainly due to a biomass that is not affected by the Zn pollution in some plants (as observed in (Zhao et al., 2000).

The boxplot for photosystem II yield and the plant color (chlorophyll content) illustrates the differences of photosynthesis efficiency between polluted and non-polluted condition, revealing how different factors such as Zn contamination may influence the plant's ability to harness light energy effectively.

When we looked at the correlations between the features we measured, Zn exposure seemed to break positive correlations existing between biomass, photosynthetic yield and plant color. This suggests that Zn may differently affect the biomass-related features, the photosynthetic yield and plant color.

## *5.1 Interpretation of logistic regression Results*

A model of logistic regression was performed to predict Zn contamination in soil. This model was build using several features in an accumulator and tolerant plant *A.halleri*. All features were measured using traditional methods, while plant color which reveals chlorophyll content in plant leaves was estimated using plant images.

The training accuracy of the logistic regression model for all features except plant color was a remarkable 0.9423, indicating that the algorithm had an exceptional ability to classify the features in the training dataset into two categories polluted and non-polluted conditions. This level of performance on the familiar training data suggests that the model had effectively learned

the underlying patterns. The model's performance on the test dataset was evaluated, revealing an accuracy of 0.905.

When we build the model again with images feature (plant color), the training accuracy of the logistic regression increased from 0.9423 to 0.954, and the model's performance on the test dataset increased from 0.905 to 0.94. This increase in logistic regression accuracy improves that images features. This result is particularly notable, as it demonstrates the model's capacity to adapt and thrive in the face of the unpredictable challenges that characterize real-world applications.

The findings of this study contribute to the growing body of evidence that emphasizes the importance of evaluating machine learning models not only on their training performance but also on their ability to generalize to new, unseen data.

Accuracy, precision, recall, and F1 score for each model appear to have strong and balanced performance across the various evaluation metrics, which is a positive sign of its effectiveness.

### *5.2 Interpretation of multiple regression Results for Zn content prediction*

A machine learning model combines different regression models (Random Forest and Gradient Boosting) using a Voting Regressor, was build to predict Zn content in *A.halleri* leaves. While the model has shown some improvement, further refinements can lead to better performance. By employing other strategies and techniques. Make another combination of models or test other models are good method to get better results. We have to look for the best way to improve our model performance.

### *5.3 Advantages of Using ML*

This model could be used to assist the prediction of Zn contamination in soil, based on plants morphological and physiological features, the majority of these features could be obtained using computer and moderate knowledge of programming. ML approaches can significantly reduce the cost and time requirements involved with laboratory analysis. It can also be used to quantify the importance of variables and identify potential control factors in heavy metal bioaccumulation in soil-crop ecosystems.

### *5.4 Limitations of the Study*

This study was conducted on plant images and other variables obtained from controlled experiment which may bias the results. Another dataset with different culture condition is available and may help to improve the model. Thus, in the future we could use images from different experiments, as *A.halleri* is widely used in this field of study.

## 6 Conclusion

### *6.1 Summary of Findings*

The main goal of this study was to predict Zn contamination in the soil using machine learning model. A logistic regression model was developed to predict Zn contamination using features of *A.halleri* and extracted feature from images. The model showed a remarkable training accuracy of 0.9423 using the traditional features, indicating the model's exceptional ability to classify the features in the training dataset into polluted and non-polluted conditions. When the plant color feature (chlorophyll content) was added, the training accuracy increased from 0.9423 to 0.954, and the test accuracy increased from 0.905 to 0.94, confirming the importance of the plant color feature.

Moreover, the study showed a slight capability to predict Zn concentration in *A.halleri* leaves using Random Forest and Gradient Boosting models, this result need an improvement but, it isn’t impssible.

This study contributes to the growing evidence emphasizing the importance of evaluating machine learning models not only on their training performance but also on their ability to generalize to new, unseen data. The model can be used to assist in the prediction of Zn contamination in soil using plant morphological and physiological features, which can significantly reduce the cost and time requirements involved with laboratory analysis.

### *6.2 Future Directions*

Many perspectives are possible in this field of study such as:

- Add new dataset to test this model, dataset from different experiment condition and collected plants in situ.
- Make deep images processing to extract many other features that may help to increase the performance of the model.
- Use other machine learning approaches to predict Zn contamination and make a comparison among them
- The main perspective is to process images to predict Zn content in plants, photosynthesis efficiency. This perspective can reduce time and cost of laboratory work.
- *A.halleri* is widely used in QTL mapping studies, ML approaches can help to make an accurate phenotyping, saving time and efforts.

### *Acknowledgment*

This research is supported by L'Oreal UNESCO for women in science. Financial support was given by this organization to conduct this master.

# 7 References

Amiri, Z., Heidari, A., & Navimipour, N. J. (2024). Comprehensive survey of artificial intelligence techniques and strategies for climate change mitigation. *Energy*, *308*, 132827. https://doi.org/10.1016/j.energy.2024.132827

Angon, P. B., Islam, Md. S., Kc, S., Das, A., Anjum, N., Poudel, A., & Suchi, S. A. (2024). Sources, effects and present perspectives of heavy metals contamination: Soil, plants and human food chain. *Heliyon*, *10*(7), e28357. https://doi.org/10.1016/j.heliyon.2024.e28357

Anubha Pearline, S., Sathiesh Kumar, V., & Harini, S. (2019). A study on plant recognition using conventional image processing and deep learning approaches. *Journal of Intelligent & Fuzzy Systems*, *36*(3), 1997–2004. https://doi.org/10.3233/JIFS-169911

Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018). Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management*, *206*, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049

Bert, V., Macnair, M. R., Laguerie, P. D., Saumitou-Laprade, P., & Petit, D. (2000). Zinc tolerance and accumulation in metallicolous and nonmetallicolous populations of Arabidopsis halleri (Brassicaceae). *The New Phytologist*, *146*(2), 225–233. https://doi.org/10.1046/j.1469-8137.2000.00634.x

Chaplin-Kramer, R., Sim, S., Hamel, P., Bryant, B., Noe, R., Mueller, C., Rigarlsford, G., Kulak, M., Kowal, V., Sharp, R., Clavreul, J., Price, E., Polasky, S., Ruckelshaus, M., & Daily, G. (2017). Life cycle assessment needs predictive spatial modelling for biodiversity and ecosystem services. *Nature Communications*, *8*(1), 15065. https://doi.org/10.1038/ncomms15065

Cipullo, S., Snapir, B., Prpich, G., Campo, P., & Coulon, F. (2019). Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. *Chemosphere*, *215*, 388–395. https://doi.org/10.1016/j.chemosphere.2018.10.056

Cosio, C., Martinoia, E., & Keller, C. (2004). Hyperaccumulation of Cadmium and Zinc in Thlaspi caerulescens and Arabidopsis halleri at the Leaf Cellular Level. *Plant Physiology*, *134*(2), 716–725. https://doi.org/10.1104/pp.103.031948

Dietrich, C. C., Tandy, S., Murawska-Wlodarczyk, K., Banaś, A., Korzeniak, U., Seget, B., & Babst-Kostecka, A. (2021). Phytoextraction efficiency of *Arabidopsis halleri* is driven by the plant and not by soil metal concentration. *Chemosphere*, *285*, 131437. https://doi.org/10.1016/j.chemosphere.2021.131437

El Hanandeh, A., Mahdi, Z., & Imtiaz, M. S. (2021). Modelling of the adsorption of Pb, Cu and Ni ions from single and multi-component aqueous solutions by date seed derived biochar: Comparison of six machine learning approaches. *Environmental Research*, *192*, 110338. https://doi.org/10.1016/j.envres.2020.110338

Gotelli, N. J., & Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, *42*(9), 1577–1585. https://doi.org/10.1111/jbi.12562

Hu, B., Xue, J., Zhou, Y., Shao, S., Fu, Z., Li, Y., Chen, S., Qi, L., & Shi, Z. (2020). Modelling bioaccumulation of heavy metals in soil-crop ecosystems and identifying its controlling factors using machine learning. *Environmental Pollution*, *262*, 114308. https://doi.org/10.1016/j.envpol.2020.114308

Hussain, S., Khan, M., Sheikh, T. M. M., Mumtaz, M. Z., Chohan, T. A., Shamim, S., & Liu, Y. (2022). Zinc Essentiality, Toxicity, and Its Bacterial Bioremediation: A Comprehensive Insight. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.900740

Jarsjö, J., Andersson-Sköld, Y., Fröberg, M., Pietroń, J., Borgström, R., Löv, Å., & Kleja, D. B. (2020). Projecting impacts of climate change on metal mobilization at contaminated sites: Controls by the groundwater level. *Science of The Total Environment*, *712*, 135560. https://doi.org/10.1016/j.scitotenv.2019.135560

Joimel, S., Schwartz, C., Hedde, M., Kiyota, S., Krogh, P. H., Nahmani, J., Pérès, G., Vergnes, A., & Cortet, J. (2017). Urban and industrial land uses have a higher soil biological quality than expected from physicochemical quality. *Science of The Total Environment*, *584–585*, 614–621. https://doi.org/10.1016/j.scitotenv.2017.01.086

Karam, M.-J., Souleman, D., Schvartzman, M. S., Gallina, S., Spielmann, J., Poncet, C., Bouchez, O., Pauwels, M., Hanikenne, M., & Frérot, H. (2019). Genetic architecture of a plant adaptive trait: QTL mapping of intraspecific variation for tolerance to metal pollution in Arabidopsis halleri. *Heredity*, *122*(6), 877–892. https://doi.org/10.1038/s41437-019-0184-4

Kaur, H., & Garg, N. (2021). Zinc toxicity in plants: A review. *Planta*, *253*(6), 129. https://doi.org/10.1007/s00425-021-03642-z

Konya, A., & Nematzadeh, P. (2024). Recent applications of AI to environmental disciplines: A review. *Science of The Total Environment*, *906*, 167705. https://doi.org/10.1016/j.scitotenv.2023.167705

Leach, K., Montgomery, W. I., & Reid, N. (2016). Modelling the influence of biotic factors on species distribution patterns. *Ecological Modelling*, *337*, 96–106. https://doi.org/10.1016/j.ecolmodel.2016.06.008

Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, *71*, 1–13. https://doi.org/10.1016/j.patcog.2017.05.015

Li, V. O. K., Lam, J. C. K., Han, Y., & Chow, K. (2021). A Big Data and Artificial Intelligence Framework for Smart and Personalized Air Pollution Monitoring and Health Management in Hong Kong. *Environmental Science & Policy*, *124*, 441–450. https://doi.org/10.1016/j.envsci.2021.06.011

Liu, L., Li, W., Song, W., & Guo, M. (2018). Remediation techniques for heavy metal-contaminated soils: Principles and applicability. *Science of The Total Environment*, *633*, 206–219. https://doi.org/10.1016/j.scitotenv.2018.03.161

Liu, P., Zhang, Y., Tang, Q., & Shi, S. (2021). Bioremediation of metal-contaminated soils by microbially-induced carbonate precipitation and its effects on ecotoxicity and long-term

stability. *Biochemical Engineering Journal*, *166*, 107856. https://doi.org/10.1016/j.bej.2020.107856

Maeda-Gutiérrez, V., Galván-Tejada, C. E., Zanella-Calzada, L. A., Celaya-Padilla, J. M., Galván-Tejada, J. I., Gamboa-Rosales, H., Luna-García, H., Magallanes-Quintanar, R., Guerrero Méndez, C. A., & Olvera-Olvera, C. A. (2020). Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases. *Applied Sciences*, *10*(4), Article 4. https://doi.org/10.3390/app10041245

Meyer, C.-L., Kostecka, A. A., Saumitou-Laprade, P., Créach, A., Castric, V., Pauwels, M., & Frérot, H. (2010). Variability of zinc tolerance among and within populations of the pseudometallophyte species Arabidopsis halleri and possible role of directional selection. *New Phytologist*, *185*(1), 130–142. https://doi.org/10.1111/j.1469-8137.2009.03062.x

Okereafor, U., Makhatha, M., Mekuto, L., Uche-Okereafor, N., Sebola, T., & Mavumengwana, V. (2020). Toxic Metal Implications on Agricultural Soils, Plants, Animals, Aquatic life and Human Health. *International Journal of Environmental Research and Public Health*, *17*(7), Article 7. https://doi.org/10.3390/ijerph17072204

Pant, M., Dolma, S., Gahlot, M., Sharma, A., & Mundepi, S. (2023). Phytoremediation of Heavy Metals. In R. P. Singh, P. Singh, & A. Srivastava (Eds.), *Heavy Metal Toxicity: Environmental Concerns, Remediation and Opportunities* (pp. 313–340). Springer Nature. https://doi.org/10.1007/978-981-99-0397-9_15

Pauwels, M., Willems, G., Roosens, N., Frérot, H., & Saumitou-Laprade, P. (2008). Merging methods in molecular and ecological genetics to study the adaptation of plants to anthropogenic metal-polluted sites: Implications for phytoremediation. *Molecular Ecology*, *17*(1), 108–119. https://doi.org/10.1111/j.1365-294X.2007.03486.x

Pizent, A., Tariba, B., & Živković, T. (2012). Reproductive Toxicity of Metals in Men. *Arhiv Za Higijenu Rada i Toksikologiju*, *63*(Supplement 1), 35–45. https://doi.org/10.2478/10004-1254-63-2012-2151

Plum, L. M., Rink, L., & Haase, H. (2010). The Essential Toxin: Impact of Zinc on Human Health. *International Journal of Environmental Research and Public Health*, *7*(4), Article 4. https://doi.org/10.3390/ijerph7041342

Sarret, G., Saumitou-Laprade, P., Bert, V., Proux, O., Hazemann, J.-L., Traverse, A., Marcus, M. A., & Manceau, A. (2002). Forms of Zinc Accumulated in the HyperaccumulatorArabidopsis halleri. *Plant Physiology*, *130*(4), 1815–1826. https://doi.org/10.1104/pp.007799

Schvartzman, M. S., Corso, M., Fataftah, N., Scheepers, M., Nouet, C., Bosman, B., Carnol, M., Motte, P., Verbruggen, N., & Hanikenne, M. (n.d.). *Adaptation to high zinc depends on distinct mechanisms in metallicolous populations of Arabidopsis halleri*. https://doi.org/10.1111/nph.14949

Shi, W., Zhang, M., Zhang, R., Chen, S., & Zhan, Z. (2020). Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sensing*, *12*(10), Article 10. https://doi.org/10.3390/rs12101688

Verbruggen, N., Hermans, C., & Schat, H. (2009). Molecular mechanisms of metal hyperaccumulation in plants. *New Phytologist*, *181*(4), 759–776. https://doi.org/10.1111/j.1469-8137.2008.02748.x

Wang, X., Liu, L., Zhang, W., & Ma, X. (2021). Prediction of Plant Uptake and Translocation of Engineered Metallic Nanoparticles by Machine Learning. *Environmental Science & Technology*, *55*(11), 7491–7500. https://doi.org/10.1021/acs.est.1c01603

Yang, L., Driscol, J., Sarigai, S., Wu, Q., Lippitt, C. D., & Morgan, M. (2022). Towards Synoptic Water Monitoring Systems: A Review of AI Methods for Automating Water Body Detection and Water Quality Monitoring Using Remote Sensing. *Sensors*, *22*(6), Article 6. https://doi.org/10.3390/s22062416

Yoon, J., Cao, X., Zhou, Q., & Ma, L. Q. (2006). Accumulation of Pb, Cu, and Zn in native plants growing on a contaminated Florida site. *Science of The Total Environment*, *368*(2), 456–464. https://doi.org/10.1016/j.scitotenv.2006.01.016

Zhao, F. J., Lombi, E., Breedon, T., & M, S. P. (2001). *Zinc hyperaccumulation and cellular distribution in Arabidopsis halleri*. https://onlinelibrary.wiley.com/doi/10.1046/j.1365-3040.2000.00569.x