Quantitative and Image-Based Analysis of Plant Structural Complexity for Genotype Classification

Prasenjit Betal

Quantitative and Image-Based Analysis of Plant Structural Complexity for Genotype Classification

Thesis submitted to
Indian Institute of Technology Kharagpur
for the award of the degree

of

Master of Science (By Research)

by

Prasenjit Betal

under the guidance of

prof. Pabitra Mitra
&
prof. Chittaranjan Mandal
(Department of Computer Science and Engineering)



CENTRE FOR COMPUTATIONAL AND DATA SCIENCES
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
June 2026

©2026 Prasenjit Betal. All rights reserved.

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Quantitative and Image-Based Analysis of Plant Structural Complexity for Genotype Classification"** submitted by **Prasenjit Betal** to the Indian Institute of Technology Kharagpur for the award of the degree of Master of Science (By Research), has been accepted by the external examiners and that the candidate has successfully defended the thesis in the viva voce examination held on 09-06-2026

Prof. Soumyajit Dey
(Chairman DAC)

Prof. Pabitra Mitra
(Supervisor)

Prof. Chittaranjan Mandal
(Co-Supervisor)

Prof. Mainack Mondal
(Member of the DAC)

Prof. K Sreenivas Rao
(Member of the DAC)

Prof. Sourav Mukhopadhyay
(Member of the DAC)



CENTRE FOR COMPUTATIONAL AND DATA SCIENCES

Indian Institute of Technology Kharagpur
Kharagpur, India-721302

Certificate

This is to certify that the thesis entitled **Quantitative and Image-Based Analysis of Plant Structural Complexity for Genotype Classification**, submitted by **Prasenjit Betal** to the Indian Institute of Technology Kharagpur, is a record of bona fide research work carried out by him under our supervision and guidance. In our opinion, the thesis is worthy of consideration for the award of the degree of **Master of Science (By Research)** of the Institute

Prof. Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India - 721302.

Prof. Chittaranjan Mandal
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India - 721302.

I.I.T. Kharagpur
June 2026

Declaration

I hereby declare that:

a. The work presented in this thesis is original and has been carried out by me under the supervision of Prof.Pabitra Mitra and Prof.Chittaranjan Mandal

b. This thesis has not been submitted, either in whole or in part, to any other Institute or University for the award of any degree, diploma, or other academic qualification.

c. I have followed the guidelines prescribed by the Institute in the preparation and presentation of this thesis.

d. I have adhered to the ethical standards and guidelines of the Institute in conducting the research and preparing this thesis.

e. Wherever materials, including data, figures, tables, models, images, or text, have been used from other sources, appropriate acknowledgement has been made through proper citation and referencing. Necessary permissions from copyright holders have been obtained wherever required.

Prasenjit Betal
Roll No: 23CD72P01

Acknowledgment

I would like to express my sincere gratitude to my supervisor, **Prof. Pabitra Mitra**, and my co-supervisor, **Prof. Chittaranjan Mandal**, for their invaluable guidance, continuous support, constructive suggestions, and encouragement throughout the course of this research. Their expertise and insightful feedback greatly contributed to the successful completion of this work.

I am also grateful to **ICAR, New Delhi (PUSA)** for providing the dataset used in this study. The availability of this dataset was essential for conducting the experiments and analyses presented in this dissertation.

Finally, I would like to thank all the faculty members, colleagues, friends, and family members who directly or indirectly supported me during the course of this work.

Prasenjit Betal
Roll No.: **23CD72P01**
Centre for Computational and Data Sciences
Indian Institute of Technology Kharagpur

Abstract

Quantitative characterization of plant architectural complexity is critical for understanding genotype-to-phenotype relationships and supporting modern crop improvement. Conventional phenotyping metrics, such as Leaf Area Index and Fractal Dimension, capture isolated geometric properties but fail to represent the integrated structural organization of plant architecture.

This thesis presents an image-based phenotyping framework for quantifying plant structural complexity and enabling genotype discrimination using interpretable architectural descriptors. A Structural Complexity Index (SCI) is first introduced by integrating complementary shape-based and skeleton-based features extracted from RGB images. Principal Component Analysis is employed to summarize high-dimensional morphological information into a single interpretable score, enabling scalable comparison across genotypes and growth stages.

To improve genotype discrimination, a multi-descriptor phenotyping pipeline is developed that combines geometric, skeletal, and structural traits with kernel-based non-linear representation learning and supervised classifiers. This framework captures conditional, combinatorial, and higher-order interactions among architectural features, enabling improved genotype separation beyond traditional linear metrics.

Finally, topological data analysis using Persistent Homology is incorporated to encode global structural organization, including branching connectivity and internal voids. The integration of topological descriptors with geometric features yields robust phenotypic representations that enhance genotype classification accuracy and resilience to noise and occlusion.

Overall, this work establishes a scalable and interpretable framework for topology-aware plant phenotyping and genotype-to-phenotype analysis.

Keywords: Plant structural complexity, Structural Complexity Index (SCI), Multi-descriptor phenotyping, Persistent Homology, Genotype classification, High-throughput phenotyping.

Contents

Abstract **xi**

1 Introduction **1**
1.1 Motivation 4
1.2 Research Objectives 5
1.3 Thesis Organization 6

2 Literature Survey **9**
2.1 Leaf Area Index as a Radiative Canopy Descriptor 9
2.2 Fractal Dimension as a Global Geometric Descriptor of Plant Architecture 11
2.3 RGB-Based Trait Extraction 13
2.4 RGB-Based Deep Learning for Phenotype Representation 13
2.5 Temporal RGB Modeling and Growth Dynamics 14
2.6 RGB-Based Genotype Discrimination and Its Limitations 14
2.7 Morphological and Skeleton-based Plant Phenotyping 15
2.8 Topological Data Analysis and Persistent Homology in Biological Imaging 16
2.9 Limitations of Traditional Phenotyping 17
2.10 Advancements in Non-Invasive, Image-Based Phenotyping 18
2.11 Summary of Identified Research Gaps 18

3 Background Methodology: Image-Based Morphological Feature Extraction **21**
3.1 Plant Shape and Skeleton Features 21
3.1.1 Segment Id 22
3.1.2 Tip and Branch point Algorithm 24

3.1.3 Segment Path Length Algorithm . . . 25
3.1.4 Segment Insertion Angle . . . 27
3.1.5 Segment Tangent Angle . . . 28
3.1.6 Segment Curvature Computation . . . 29
3.1.7 Euclidean Length . . . 31
3.1.8 Convex Hull & Solidity . . . 32
3.1.9 Contour Extraction . . . 33
3.1.10 Convex Hull Feature Computation . . . 34
3.1.11 Compactness Computation . . . 35
3.1.12 Eccentricity . . . 36
3.1.13 Different colour space . . . 37
3.1.14 Scatter Plot . . . 38
3.1.15 Data set . . . 39

4 Quantifying Plant Structural Complexity Using Structural Complexity Index 41
4.1 Introduction . . . 41
4.2 Proposed Work . . . 43
4.2.1 Structural Complexity Index (SCI) Framework: Step-by-Step Methodology . . . 43
4.2.2 Description of the SCI Computational Framework . . . 48
4.2.3 Structural Complexity Index (SCI) Computation Framework 48
4.3 Results . . . 50
4.3.1 Segment Curvature . . . 50
4.3.2 Segment Path length . . . 51
4.3.3 Different Growth Stages . . . 52
4.4 An Overview of Method . . . 55

5 Non-Linear Modeling of Multi-Descriptor Plant Architecture 57
5.1 Introduction . . . 57
5.1.1 U-Shaped and Non-Linear Trait Relationships in Plant Architecture . . . 59
5.1.2 Curvature and Path Length . . . 60
5.1.3 Branching Complexity and Allometry . . . 61
5.1.4 Multivariate Shape and Skeleton Traits . . . 61
5.1.5 Biological Justification for nSCI . . . 61
5.2 Proposed Work . . . 62

5.2.1 Non-Linear Structural Complexity Index (nSCI) 63
5.2.1.1 Phenotypic Feature Representation 64
5.2.1.2 Kernel-Induced RKHS Mapping 64
5.2.1.3 Kernel PCA for Structural Mode Extraction 64
5.2.1.4 Definition of nSCI Mode and Score 65
5.2.1.5 Biological Interpretation 65
5.2.1.6 Relation to Genotype Analysis 65
5.2.1.7 An Overview of Method 65
5.3 Kernel Fisher Discriminant for Genotype Separation 66
5.3.1 Definition 66
5.3.2 Dual Representation 67
5.3.3 Interpretation with Shape and Skeleton Features 67
5.3.4 Biological Interpretation 67
5.3.5 Key Points for Thesis Integration 68
5.4 SVM with RBF Kernel — Mathematical Formulation for Plant Traits 68
5.4.1 Decision Function 69
5.4.2 RBF (Gaussian) Kernel 69
5.4.3 Dual Optimization Problem 70
5.4.4 Biological Interpretation for Plant Traits 70
5.5 Deep Kernel Learning (DKL) for Plant Genotype Discrimination . . 72
5.5.1 Motivation 72
5.5.2 Mathematical Framework 72
5.5.2.1 Neural Network Feature Mapping 72
5.5.2.2 Kernel Construction in Latent Space 73
5.5.2.3 Supervised Genotype Classification 73
5.5.3 Local Trait Sensitivity and Interaction Structure 74
5.5.3.1 Trait Sensitivity 74
5.5.3.2 Trait Interaction Structure 74
5.5.4 Advantages over Standard Kernels 74
5.5.5 Empirical Example: U-Shaped Non-Monotonic Curvature Dependency 75
5.5.6 Deep Kernel Learning (DKL) — Supervised Non-Linear Genotype Discrimination 75
5.6 High-Dimensional Phenotyping and Genotype-to-Phenotype Mapping 76
5.6.1 An Overview of Method 76
5.6.2 Step 1: Image Acquisition, Preprocessing, and Segmentation 77
5.6.3 Step 2: Skeletonization and Structural Feature Extraction . 78

5.6.4 Step 3: Morphological Feature Extraction and Feature Matrix Formation . . . 78
5.6.5 Step 4: Dimensionality Reduction and Clustering . . . 78
5.6.6 Step 5: Genotype Classification and Feature Interpretation . 79
5.6.7 Output . . . 80
5.7 Results . . . 80
5.8 An Overview of Method . . . 83

6 Topology-Driven Genotype and Vigor Characterization Using Persistent Homology 85
6.1 Introduction . . . 85
6.2 Proposed Work . . . 89
6.2.1 Topological Features and Summary Statistics . . . 90
6.3 Theoretical Foundations of Persistent Homology–Based Learning . . 92
6.3.1 Image Domain and Function Space . . . 93
6.3.2 Perturbation Model . . . 93
6.3.3 Scalar Filtration Construction . . . 94
6.3.4 Persistent Homology and Diagrams . . . 94
6.3.5 Metric Structure on Persistence Diagrams . . . 95
6.3.6 Stability of Persistent Homology . . . 95
6.3.7 Topological Representation Operator . . . 95
6.4 Learning Model, Stability, and Transferability . . . 96
6.4.1 Learning Model via Topological Representations . . . 96
6.4.2 Lipschitz Continuity of the Regression Map . . . 96
6.4.3 Stability of the Composite Predictor . . . 97
6.4.4 Genotype–Phenotype Factorization . . . 97
6.4.5 Cross-Trait Transferability . . . 98
6.4.6 Implications for Generalization . . . 98
6.4.7 Euclidean Distance Transform (EDT) . . . 98
6.4.8 Compute Persistence Diagram (PD_i) . . . 100
6.4.9 Persistence Diagram . . . 100
6.4.10 Persistence Image Generation: . . . 101
6.4.11 Persistent Image . . . 103
6.5 Results . . . 103
6.5.1 Feature Concatenation . . . 103
6.5.2 UMAP-Based Dimensionality Reduction and Analysis . . . 104
6.6 Topological Structural Vigor Index . . . 105

6.6.1 Definition of Plant Vigor 106
6.6.2 Topological Representation Using Persistent Homology . . . 107
6.6.3 Geometric Vigor Index (GVI) 107
6.6.4 Data-Driven Weight Selection for Geometric Features 109
6.6.5 Data-Driven Selection and Interpretation of α 109
6.6.6 Hybrid Vigor Index (HVI) 110
6.6.7 Comparison with Traditional Systems 111
6.7 An Overview of Method 112
6.8 Summary of the Thesis 113

7 Conclusion **115**
References 117

List of Abbreviations

FD	Fractal Dimension
LAI	Leaf Area Index
SCI	Structural Complexity Index
PCA	Principal Component Analysis
nSCI	Non-linear Structural Complexity Index
RKHS	Reproducing Kernel Hilbert Space
KFD	Kernel Fisher Discriminant
SVM	Support Vector Machine
RBF	Radial Basis Function
DKL	Deep Kernel Learning
RF	Random Forest
XGBoost	Extreme Gradient Boosting
MLP	Multilayer Perceptron
UMAP	Uniform Manifold Approximation and Projection
TDA	Topological Data Analysis
PD	Persistence Diagram
PH	Persistent Homology
EDT	Euclidean Distance Transform
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
GVI	Geometric Vigor Index

List of Figures

2.1 Leaf Area Index. (Source: Compiled from Watson [1] and Mandelbrot [2].) . . . 10

3.1 Segment Id . . . 24
3.2 (a) Tip points . . . 25
3.3 (b) Branch points . . . 25
3.4 Segment Path Length . . . 26
3.5 Segment Insertion Angle . . . 28
3.6 Segment Tangent Angle . . . 29
3.7 Segment Curvature Angle . . . 30
3.8 Euclidean Length . . . 31
3.9 Convex Hull Area . . . 33
3.10 Canopy Geometry Visualization . . . 36
3.11 Colour Space Visualization . . . 37
3.12 Colour Scatter Plot Visualization . . . 38

4.1 SCI Pipeline . . . 43
4.2 Principle Component Analysis . . . 44
4.3 Segment Curvature Mean . . . 50
4.4 Path Length Across Growth Stages. . . . 51
4.5 Different Growth Stages . . . 52
4.6 Feature Correlation Heatmap. . . . 53
4.7 Scree Plot of Explained Variance by Principal Components. . . . 53
4.8 K-Means Cluster . . . 54

5.1 U-Shaped nSCI Manifold . . . 60
5.2 nSCI Visualization . . . 71
5.3 Deep Kernel Learning (DKL) for supervised genotype discrimination 76

5.4 Proposed Pipeline 77
5.5 UMAP-based clustering of *Brassica juncea* genotypes. 82

6.1 Euclidean Distance transform 99
6.2 Persistence Diagram 101
6.3 Persistence Image 102
6.4 UMAP-based clustering of Brassica juncea genotypes. 105

List of Tables

4.1 ANOVA Results for Top PCA Features 54

5.1 Classification Performance Comparison 81
5.2 Top 10 most important features ranked by Random Forest and XGBoost. 82

6.1 Comparison of classifier performance 104

CHAPTER 1

Introduction

Plant architecture [3, 4, 5] is a fundamental biological trait that reflects the integrated outcome of genetic regulation, developmental dynamics, and environmental interaction. Quantitative characterization of plant structure is central to understanding genotype–phenotype (G2P) relationships, accelerating crop improvement, and enabling predictive plant science. As global agriculture faces increasing pressure from climate variability, resource limitation, and population growth, there is a growing demand for phenotyping frameworks that are scalable, non-destructive, and capable of capturing biologically meaningful architectural information across species, genotypes, and developmental stages.

Despite major advances in high-throughput genotyping, phenotyping remains a critical bottleneck in the G2P pipeline due to its reliance on limited, labor-intensive, and often low-dimensional trait measurements [6, 7, 8]. Conventional plant phenotyping approaches typically emphasize isolated morphological or spectral attributes such as height, projected area, biomass, or color indices [9, 10, 11]. While these descriptors are useful for capturing specific aspects of growth, they fail to represent the multidimensional architectural organization of plant structure, including branching hierarchy, skeletal connectivity, spatial occupation, and internal void distribution [12, 13, 14].

Widely adopted summary metrics further illustrate these limitations. Biomass measurements provide aggregate estimates of growth but lack spatial and organizational detail. The *Leaf Area Index (LAI)*, defined as the ratio of total leaf area to ground area, $\mathrm{LAI} = A_{\mathrm{leaf}}/A_{\mathrm{ground}}$ [**?**], primarily reflects canopy density and radiative interception, without encoding branching structure or spatial arrangement. Similarly, Fractal Dimension (FD) characterizes global geometric complexity through image scaling behavior, $D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}$ [**?**], but remains a purely geometric abstraction that lacks topological awareness and biological specificity. Consequently, architecturally distinct plants may exhibit similar LAI or FD values despite profound differences in structural organization.

These limitations highlight a central challenge in modern phenotyping: plant architecture is an emergent phenotype arising from tightly coupled geometric, skeletal, and topological traits whose interactions are inherently non-linear, non-additive, and scale-dependent [15, 16, 17]. Curvature influences the functional relevance of path length, branching density modulates spatial occupation, and compactness governs how skeletal complexity translates into external form. Linear descriptors and Euclidean feature spaces are ill-suited to capture such dependencies, motivating the need for mathematically principled frameworks that explicitly model architectural interactions and global structure [18, 19, 20].

This thesis addresses these challenges through a progressive, multi-level phenotyping framework that integrates geometric, skeletal, non-linear, and topological representations of plant architecture from image-based data. The methodological development proceeds across three tightly connected stages.

First, a unified image-based baseline for architectural quantification is established through the *Structural Complexity Index (SCI)*. SCI combines complementary shape-based and skeleton-based descriptors extracted from RGB images using established computer vision pipelines such as PlantCV [21, 22]. By consolidating high-dimensional morphological information into a single interpretable measure, SCI enables scalable, non-destructive quantification of whole-plant structural variation across growth stages and genotypes. This provides a robust foundation for longitudinal phenotyping and comparative architectural analysis.

Second, recognizing that plant architecture resides on a high-dimensional, curved phenotypic manifold, the framework is extended to non-linear modeling

of architectural traits. Phenotypic measurements are embedded into reproducing kernel Hilbert spaces (RKHS), enabling the capture of conditional, combinatorial, and non-monotonic dependencies among geometric and skeletal features. Kernel Fisher discriminants, kernel-based Support Vector Machines, Deep Kernel Learning, and ensemble classifiers are employed to evaluate genotype-discriminative power and to identify architectural features that contribute most strongly to genotype separation [10, 23]. Non-linear dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) further reveals the intrinsic structure of the learned phenotypic space.

Third, to move beyond metric-dependent representations and explicitly capture global structural organization, the thesis incorporates concepts from *Topological Data Analysis* (TDA), with a focus on *Persistent Homology* (PH) [24, 25, 26]. Unlike conventional morphometric approaches, PH characterizes connectivity, loops, and internal voids across multiple spatial scales without reliance on fixed coordinate systems or rigid geometric assumptions. By transforming plant contours and skeleton graphs into suitable topological complexes, PH provides scale-invariant, noise-resilient descriptors of architectural complexity [27]. These topology-derived signatures capture branching patterns, internal gaps, and repetitive motifs that are not accessible through purely geometric metrics. Together, these components form a coherent phenotyping pipeline that advances from interpretable baseline descriptors to non-linear manifold modeling and topology-aware structural analysis. By unifying geometric, skeletal, and topological perspectives, this thesis provides a comprehensive framework for quantitative plant architecture analysis and genotype discrimination. The proposed methods support robust phenotypic comparison across genotypes, developmental stages, and environmental conditions, and establish a principled foundation for future research in plant phenotyping, breeding, and functional genomics, including applications in space-efficient cultivation systems, such as vertical and controlled-environment agriculture, and in the quantitative analysis of architectural resilience under abiotic stresses, where maintaining functional structural organization is critical for productivity and adaptation.

1.1 Motivation

Plant phenotyping is a cornerstone of modern crop science, enabling systematic investigation of genotype-to-phenotype relationships and supporting data-driven breeding strategies. Despite significant advances in genotyping technologies, phenotyping remains a major bottleneck due to its reliance on labor-intensive, subjective, and often low-throughput measurement protocols, which limit scalability, reproducibility, and cross-study comparability [11, 28]. These constraints have motivated the increasing adoption of high-throughput, non-destructive imaging technologies coupled with computational analysis [6, 29, 30].

Early image-based phenotyping approaches primarily focused on simple morphological and spectral descriptors such as projected leaf area, plant height, width, and color indices [31, 7, 32]. While effective for estimating growth and biomass, these descriptors capture isolated aspects of plant form and fail to encode the internal architectural organization of plants, including branching hierarchy, skeletal connectivity, and the spatial distribution of canopy voids [33, 34]. As a result, architecturally distinct genotypes may appear similar under conventional morphological summaries.

To facilitate standardized feature extraction, open-source platforms such as PlantCV and commercial phenotyping systems such as LemnaTec have become widely adopted [21, 35]. Although these platforms provide robust and repeatable pipelines for extracting shape- and color-based traits, they largely operate at fixed spatial scales and emphasize global geometric descriptors. Consequently, their ability to quantify structural complexity, architectural heterogeneity, and scale-dependent organization remains limited [8, 36].

Recent advances in machine learning and deep learning have significantly improved predictive performance in image-based plant phenotyping tasks [37, 38, 39]. However, many data-driven approaches rely on high-dimensional latent feature representations that function as black boxes, hindering biological interpretability and limiting insight into which structural attributes drive genotype-level differentiation [40, 41]. This trade-off between predictive accuracy and interpretability remains a critical challenge in phenotype-driven plant science.

Motivated by these gaps, this thesis aims to develop an interpretable, image-

based phenotyping framework that explicitly captures plant structural organization across multiple representation levels. By integrating complementary shape-based, skeleton-based, and topology-aware descriptors, the proposed approach seeks to quantify plant architectural complexity in a manner that is robust, scalable, and biologically meaningful. This unified framework is designed to move beyond isolated trait measurement and black-box modeling toward structured, explainable representations of plant architecture that support genotype discrimination and downstream biological interpretation.

1.2 Research Objectives

The primary objective of this thesis is to develop an interpretable, image-based computational framework for quantitative analysis of plant architecture and genotype discrimination using RGB imagery. The specific objectives are:

1. To extract robust **shape-based morphological descriptors** from RGB plant images that characterize global plant geometry and spatial extent [7, 12, 14].

2. To extract **skeleton-based structural descriptors** that capture internal connectivity, branching organization, and geometric orientation of plant architectures [22, 42, 43].

3. To formulate a **Structural Complexity Index (SCI)** that summarizes complementary morphological and skeletal features into a single interpretable measure of architectural complexity using dimensionality reduction techniques [44, 45, 46].

4. To evaluate the statistical significance and biological relevance of the contributing architectural features across genotypes and developmental stages using appropriate statistical analyses [47].

5. To develop a **multi-descriptor phenotyping framework** that integrates geometric, skeletal, and structural representations for robust genotype discrimination using supervised and kernel-based learning approaches [48, 49].

6. To incorporate **topological data analysis via Persistent Homology** for extracting higher-order structural descriptors that encode connectivity, canopy closure, and internal void structures from plant images [27, 24, 25, 26].

7. To design a **plant vigor assessment framework** that integrates topological, geometric, skeletal, and appearance-based descriptors into a data-driven Hybrid Vigor Index, enabling quantitative assessment of architectural integrity and growth organization under varying conditions [50, 51].

8. To demonstrate the applicability of the proposed framework for high-throughput phenotyping, genotype classification, and genotype-to-phenotype analysis using RGB image data [50, 51].

1.3 Thesis Organization

This thesis is organized into seven chapters, each addressing a specific aspect of image-based plant phenotyping and architectural analysis, progressing from foundational concepts to advanced modeling, experimental evaluation, and future directions.

Chapter 1 introduces the motivation and scope of the research. It outlines the challenges associated with traditional plant phenotyping approaches and highlights the need for scalable, interpretable, and structure-aware computational frameworks. The chapter reviews commonly used phenotyping metrics and motivates the development of advanced descriptors capable of capturing whole-plant architectural complexity.

Chapter 2 presents a comprehensive literature survey of existing plant phenotyping methodologies. It reviews traditional phenotyping indices, RGB-based image analysis techniques, deep learning approaches, and topological methods, identifying their strengths and limitations. This chapter establishes the conceptual and methodological gaps that motivate the proposed framework.

Chapter 3 presents the background and methodology for extracting morphological features from RGB images. This chapter describes the image preprocessing

pipeline, plant segmentation, and skeletonization procedures. It details the extraction of skeleton-based descriptors, including segment number (Fig. 3.1), segment path length (Fig. 3.4), segment curvature (Fig. 4.3), segment insertion angles (Fig. 3.5), and branching points, along with shape-based descriptors (Fig. 3.10) such as projected area, convex hull area, solidity, eccentricity, and compactness. These features collectively form the foundational phenotypic representation used in subsequent analyses.

Chapter 4 develops a mathematically grounded framework for modeling plant architecture in a non-linear phenotypic space. Kernel-based methods are introduced to capture conditional, non-additive, and non-monotonic relationships among shape- and skeleton-derived traits. The chapter formulates non-linear structural representations, including kernel-based indices and classification models, and evaluates their effectiveness for genotype discrimination using supervised learning techniques. Visualization of the learned phenotypic space is also presented to assess separability among genotypes.

Chapter 5 extends the architectural analysis through topological data analysis, with a focus on Persistent Homology. This chapter describes the construction of filtrations from plant image representations and the extraction of topological signatures such as connected components and loop structures. Persistence diagrams and vectorized topological descriptors are integrated with geometric and skeletal features to enhance robustness against noise, viewpoint variation, and partial occlusion. The combined framework is evaluated for genotype classification and phenotypic clustering. In addition, a unified plant vigor assessment framework is presented, integrating topological and geometric descriptors into a hybrid vigor index for genotype comparison and stress analysis.

Chapter 6 presents the experimental evaluation and analysis of the proposed framework. The chapter summarizes key empirical findings, discusses the performance of geometric, non-linear, and topology-aware representations, and analyzes their implications for genotype discrimination and architectural phenotyping across growth stages and environmental conditions.

Chapter 7 concludes the thesis by summarizing the principal contributions and key insights derived from this work. It discusses the broader implications of topology-aware plant phenotyping and outlines future research directions, in-

cluding multi-modal sensing, temporal topological analysis, genotype–phenotype association studies, and deployment in real-world phenotyping systems.

CHAPTER 2

Literature Survey

The evolution of plant phenotyping methodologies is driven by the necessity to efficiently and accurately link **genotype to phenotype** for crop improvement and enhanced productivity. The literature review highlights the transition from traditional, restrictive measurements to advanced, image-based quantification of plant structural complexity.

2.1 Leaf Area Index as a Radiative Canopy Descriptor

Leaf Area Index (LAI) is widely employed as a quantitative descriptor of vegetation canopies due to its strong linkage with *radiative transfer processes* and *biosphere–atmosphere exchanges*. Formally, LAI is defined as the ratio of one-sided leaf surface area to the corresponding ground area, LAI = A_L/A_G, where A_L denotes the total one-sided leaf area and A_G represents the ground area occupied by the canopy. By definition, LAI is a *dimensionless scalar* that characterizes canopy density rather than plant architectural organization.

The importance of LAI arises from the fact that key physiological and biophysical processes, including **photosynthesis**, **transpiration**, and **energy exchange**, occur predominantly at leaf surfaces. Consequently, LAI serves as a fundamental state variable in *ecological*, *hydrological*, and *climate models*, operating across spatial scales ranging from individual field plots to global Earth system simulations [52, 53].Figure 2.1 illustrates conceptual canopy models corresponding to two levels of structural complexity, where (a) represents a low-complexity canopy (complexity level 1) and (b) represents a higher-complexity canopy (complexity level 3).

The illustration is conceptually inspired by classical Leaf Area Index (LAI) canopy representations commonly employed in radiative transfer and plant canopy modeling literature [54].

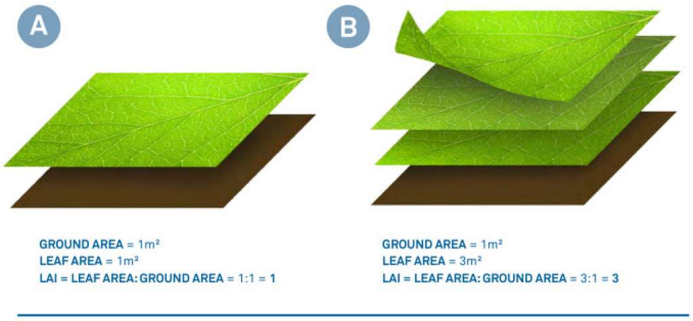


Figure 2.1: Leaf Area Index. (Source: Compiled from Watson [1] and Mandelbrot [2].)

In practice, LAI is rarely measured through direct destructive sampling. Instead, it is most commonly inferred using **indirect optical methods** that exploit canopy–radiation interactions. Within this framework, foliage is modeled as a light-attenuating medium, and the transmission of photosynthetically active radiation (PAR) through the canopy is described using a Beer–Lambert type formulation, $I(z)/I_0 = \exp(-k \cdot \mathrm{LAI})$, where I_0 is the incident radiation above the canopy, $I(z)$ is the transmitted radiation measured below the canopy, and k is the **canopy extinction coefficient**, which depends on *leaf angle distribution*, *solar*

zenith angle, and foliage optical properties [55, 56].

By inverting this relationship, LAI can be estimated from measurements of either **radiation transmittance** (e.g., PAR inversion techniques) or **radiation reflectance** (e.g., vegetation indices derived from red and near-infrared bands). These approaches form the theoretical basis of widely used field instruments and remote sensing products, enabling *non-destructive*, *repeatable*, and *spatially scalable* estimation of LAI [57, 58].

However, radiative measurements inherently integrate leaf area along the photon path, yielding an **effective LAI** rather than an explicit structural measure. As a result, canopies with substantially different *spatial arrangements*, *branching architectures*, or degrees of **foliage clumping** may produce similar LAI values despite marked differences in physical organization. In heterogeneous or woody canopies, non-photosynthetic elements such as stems and branches further contribute to radiation attenuation, leading to measurements that are more accurately interpreted as *plant area index (PAI)* rather than pure LAI [59].

Moreover, LAI exhibits **saturation behavior** in dense canopies, where additional leaf layers cause minimal changes in transmitted or reflected radiation, reducing sensitivity at high LAI values. While LAI effectively captures *temporal dynamics* such as canopy development, peak growth, and senescence, it does not encode information related to **branching topology**, **organ-level geometry**, or **spatial connectivity** within the canopy.

Thus, although LAI remains a powerful indicator of canopy functional capacity and vegetation productivity, its fundamentally radiative nature limits its ability to represent structural and topological complexity. This limitation motivates the development of complementary phenotyping descriptors capable of capturing plant architecture beyond leaf area alone.

2.2 Fractal Dimension as a Global Geometric Descriptor of Plant Architecture

Fractal Dimension (FD) has been widely adopted as a quantitative descriptor of plant morphology due to its ability to capture global geometric complexity

through scale-invariant image properties. In plant phenotyping, FD characterizes how effectively a plant structure fills space across multiple observation scales, thereby providing a compact scalar measure of structural irregularity.

Formally, fractal dimension is commonly estimated using the box-counting method and is defined as

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}, \tag{2.1}$$

where $N(\epsilon)$ denotes the number of covering elements (boxes) of side length ϵ required to cover the plant silhouette or skeleton at scale ϵ [**?**]. In digital plant images, this formulation quantifies the rate at which additional structural details emerge as the observation scale becomes finer.

The appeal of FD lies in its ability to summarize complex plant forms—including branching, serration, and overall irregularity—into a scale-independent metric. Consequently, FD has been applied to characterize plant growth patterns, canopy roughness, and morphological variability across environmental conditions and genotypes.

Despite these advantages, fractal dimension fundamentally operates as a *global geometric descriptor* and remains agnostic to the biological organization of plant structures. FD does not encode information about branching hierarchy, organ connectivity, insertion angles, or spatial arrangement of plant components. As a result, plant architectures with markedly different biological structures may exhibit similar FD values if their overall space-filling properties are comparable. This limitation becomes particularly pronounced across growth stages, where developmental changes modify topology and geometry in ways that cannot be faithfully represented by a single scaling exponent.

Furthermore, FD estimation is sensitive to image resolution, segmentation quality, and scale selection, which may further obscure biologically meaningful differences. Consequently, while fractal dimension captures *how much space a plant occupies across scales*, it does not describe *how that space is organized.*

Thus, similar to Leaf Area Index, fractal dimension provides a useful but incomplete abstraction of plant structure, emphasizing geometric complexity while neglecting architectural and topological specificity.

To address these limitations, the proposed Structural Complexity Index (SCI)

integrates ***complementary shape-based descriptors*** with ***skeleton-based descriptors*** into a unified framework. By jointly modeling global geometry and branching topology, SCI enables robust quantification of structural variation across ***growth stages*** and ***genotypes***, overcoming the inherent reductionism of purely scaling-based metrics such as fractal dimension [21, 22].

2.3 RGB-Based Trait Extraction

Early RGB-based phenotyping studies primarily focused on extracting low-level morphological and color-based traits such as projected area, biomass proxies, and growth rates from plant images. High-throughput RGB imaging platforms enabled automated measurement of these traits across large plant populations, significantly improving phenotyping throughput compared to manual approaches [32, 7, 21]. While effective for capturing overall growth trends, these methods rely mainly on global or pixel-level descriptors and do not explicitly model plant structural organization, such as branching topology or skeletal geometry.

2.4 RGB-Based Deep Learning for Phenotype Representation

With the emergence of deep learning, RGB-based phenotyping has shifted from hand-crafted feature extraction toward learned visual representations. Convolutional neural networks (CNNs) have been successfully applied to tasks such as plant segmentation, biomass estimation, and developmental stage prediction, learning complex visual patterns directly from RGB imagery [39, 40]. Although these approaches improve predictive performance, they often function as black-box models and provide limited interpretability regarding the specific architectural traits driving phenotypic variation.

2.5 Temporal RGB Modeling and Growth Dynamics

Several studies have extended RGB-based phenotyping by incorporating temporal image sequences to model plant growth dynamics. CNN–LSTM and related recurrent architectures combine spatial visual features with temporal dependencies, enabling genotype discrimination and treatment response analysis based on growth trajectories rather than single time-point observations [51, 7]. Despite their effectiveness in capturing developmental trends, these methods primarily encode appearance changes over time and do not explicitly quantify structural or topological adaptations in plant architecture.

2.6 RGB-Based Genotype Discrimination and Its Limitations

Recent studies have demonstrated that RGB imagery can be used directly for genotype classification. Deep learning models trained on RGB images have shown the ability to distinguish plant accessions or genotypes under controlled experimental conditions, indicating that genotype-specific phenotypic signatures are embedded in visual data [47, 60]. These approaches highlight the potential of image-based phenotyping for automated genotype discrimination.

However, genotype discrimination in RGB-based pipelines is typically driven by implicit feature learning rather than explicit architectural descriptors, limiting biological interpretability and robustness across environments. Most RGB-based phenotyping frameworks rely predominantly on appearance-driven features and linear or implicitly learned representations. As a result, structural traits such as branching complexity, curvature, insertion angles, and skeletal topology known to play a critical role in plant function, light interception, and environmental adaptation are rarely modeled explicitly [33, 22]. Consequently, these approaches may fail to capture non-linear and conditional relationships among architectural traits that are essential for robust and biologically meaningful genotype-to-phenotype analysis.

2.7 Morphological and Skeleton-based Plant Phenotyping

Image-based plant phenotyping has traditionally relied on geometric, morphometric, and skeleton-derived descriptors to quantify shoot architecture from RGB images. Widely used frameworks such as PlantCV [21], along with custom computer vision pipelines [39, 22], extract fundamental traits including projected area, convex hull area, aspect ratio, eccentricity, and compactness. These features provide quantitative summaries of plant size, shape, and overall growth status and have been extensively applied in phenotypic screening and stress assessment.

Complementing these geometric measures, skeleton-based descriptors capture structural properties related to plant topology at a local level. Features such as total skeleton length, number of branch points, leaf tip count, leaf insertion angle, and curvature metrics offer insight into branching complexity and canopy organization. Such descriptors are particularly useful for characterizing architectural traits associated with growth patterns, lodging resistance, and light interception.

Despite their widespread adoption, morphological and skeleton-based descriptors exhibit several inherent limitations. These features are often sensitive to segmentation quality, image resolution, occlusion, and camera viewpoint, leading to variability across imaging conditions and experimental setups. Metrics such as convex hull area or skeleton path length may fluctuate under minor pose changes or partial leaf overlap, reducing robustness and repeatability across genotypes. Furthermore, most conventional descriptors operate at fixed spatial scales and primarily capture local or part-wise geometry, limiting their ability to represent global structural organization, multi-scale connectivity, and repeated architectural motifs that are biologically meaningful. These limitations have motivated the exploration of alternative representations capable of capturing plant structure in a more holistic and scale-aware manner.

2.8 Topological Data Analysis and Persistent Homology in Biological Imaging

Topological Data Analysis (TDA) provides a complementary mathematical framework for characterizing complex structures by focusing on connectivity and shape across multiple spatial scales rather than precise geometric measurements. Unlike classical morphometric descriptors, TDA is inherently robust to small perturbations, noise, and sampling variability, making it well suited for biological imaging data [24, 25]. Persistent Homology (PH), the most widely used TDA technique, quantifies topological features such as connected components, loops, and voids that emerge and persist throughout a filtration process. These features are summarized using persistence diagrams, barcodes, or vectorized representations such as persistence images [26].

PH has been successfully applied across a range of biological domains, including protein structure analysis, neuronal morphology, cellular organization, and medical imaging, where it has demonstrated strong capability in capturing global structural patterns that are difficult to represent using conventional geometry alone. In plant science, applications of PH are comparatively recent but expanding. Studies have shown its effectiveness in characterizing root system architecture [61] and quantifying canopy structural complexity from imaging data [62], highlighting its ability to capture branching connectivity, canopy gaps, and global organization.

However, PH has not yet been systematically integrated with RGB-derived morphological and skeleton-based descriptors for comprehensive analysis of plant shoot architecture. Existing phenotyping pipelines typically treat geometric and topological information independently, limiting their capacity to jointly capture local shape traits and global structural organization. This separation constrains robust genotype discrimination and the detection of subtle architectural differences arising from growth dynamics or stress conditions.

The present work addresses this gap by integrating Persistent Homology with conventional morphological and skeleton-derived features. By extracting topological signatures from plant contours and skeleton representations and combining them with classical descriptors, the proposed framework enables multi-scale, topology-aware phenotypic characterization of plant architecture. This unified

representation improves robustness to noise and variability while preserving biologically meaningful structural information essential for genotype comparison and architectural analysis.

2.9 Limitations of Traditional Phenotyping

Despite their historical importance, conventional phenotyping approaches exhibit fundamental limitations that restrict their suitability for modern, high-throughput and data-driven breeding pipelines.

- **Limited Scalability and Repeatability:** Many traditional phenotyping methods rely on manual or destructive measurements, which hinder longitudinal analysis of individual plants across growth stages. Such approaches are labor-intensive and do not scale efficiently to the large populations required in contemporary genetic and breeding studies.
- **Insufficient Representation of Structural Complexity:** Widely used descriptors such as Leaf Area Index (LAI) and Fractal Dimension (FD) capture only partial aspects of plant structure. These measures summarize canopy density or global geometric complexity but do not encode branching organization, connectivity, or multi-scale architectural variation (see Figure 2.1). Consequently, structurally distinct plants may exhibit similar index values, limiting discrimination across genotypes or developmental stages.
- **Incompatibility with Image-Based High-Throughput Systems:** Traditional indices are often designed for specific measurement modalities or controlled conditions and cannot be seamlessly integrated into automated, image-based phenotyping pipelines. This limits their effectiveness in extracting biologically meaningful patterns from large-scale imaging datasets.

2.10 Advancements in Non-Invasive, Image-Based Phenotyping

To address the limitations of traditional phenotyping, recent research has increasingly adopted non-invasive, image-based approaches capable of capturing plant structure at scale.

- **Scalable Imaging Platforms:** High-throughput imaging modalities, including **RGB, hyperspectral, and 3D imaging**, enable rapid, non-destructive, and repeatable acquisition of plant data across large populations and multiple growth stages.
- **Challenges in Structural Quantification:** Despite advances in sensing technologies, robust quantification of plant architecture remains challenging due to self-occlusion, viewpoint dependency, and environmental variability. Furthermore, fusion of information from multiple views or modalities is often limited or ad hoc.
- **Fragmented Trait Analysis:** Existing image-based plant phenotyping studies predominantly focus on isolated or low-dimensional morphological traits (e.g., height, projected area, or leaf count). While such approaches enable automated analysis of specific traits, they fail to provide a unified, comprehensive, and scalable representation of whole-plant architectural complexity required for robust comparative and statistical analysis.

2.11 Summary of Identified Research Gaps

Based on the foregoing review of existing phenotyping methodologies, several key research gaps can be identified. Despite substantial progress in image-based plant phenotyping, important challenges remain unresolved:

- **Fragmented representation of plant architecture:** Most existing phenotyping pipelines rely on isolated morphological or shape-based traits, which fail to capture internal connectivity and branching organization [8, 34].

2.11 Summary of Identified Research Gaps

- **Absence of a unified complexity metric:** Although metrics such as Leaf Area Index and fractal dimension have been widely used, they do not provide a comprehensive or interpretable measure of plant structural complexity [31, 63].
- **Limited use of skeleton-based structural features:** Skeleton representations have been explored for specific traits, but are rarely integrated with shape descriptors to form a holistic architectural model [22, 42, 64].
- **Insufficient genotype discrimination using conventional traits:** Basic morphological and color features often lack sensitivity to subtle architectural differences among genotypes, particularly under noise, occlusion, and viewpoint variation [47, 65].
- **Underutilization of topological information:** Topological descriptors capable of encoding connectivity and void structures remain largely unexplored in integrated phenotyping pipelines [66, 67, 61].
- **Limited interpretability of learning-based models:** Many machine learning approaches prioritize predictive accuracy over interpretability, making it difficult to relate phenotypic differences to underlying biological structures [40, 60].

CHAPTER 3

Background Methodology: Image-Based Morphological Feature Extraction

3.1 Plant Shape and Skeleton Features

This chapter describes the feature extraction pipeline used to compute morphological descriptors of plant structures from RGB images. The framework follows standard image processing and plant phenotyping methodologies. A binary plant mask is first generated, followed by skeletonization to obtain a one-pixel-wide representation of the plant structure. From this skeleton, standard graph-based structural descriptors are extracted, including segment identification, segment length, tip and branch point detection, insertion angles, and curvature-based tangent angles. These skeleton-based features are computed using established image processing and graph-theoretic methods commonly used in plant phenotyping toolkits such as PlantCV [21] and related literature. In addition to skeleton descriptors, classical shape features such as convex hull area, contour area, solidity, eccentricity, and

compactness are extracted to characterize global plant morphology. These features are widely used in existing plant phenotyping frameworks (e.g., LemnaTec-based analysis pipelines). Together, these features provide a comprehensive morphological representation of plant architecture and serve as input for downstream analysis and the proposed SCI framework [35].

3.1.1 Segment Id

Segment ID refers to the unique numerical label assigned to each distinct structural component of a plant, such as individual branches, leaves, or skeleton segments derived from the binary mask. In image-based phenotyping, a plant is rarely treated as a single continuous object; instead, it is decomposed into smaller segments that represent biologically meaningful units. Assigning a unique Segment ID to each of these units enables the extraction of localized morphological information, including segment-specific length, curvature, orientation, insertion angle, and branching patterns. This segmentation-based representation is essential for understanding the hierarchical architecture of the plant, distinguishing between primary and secondary branches, and analyzing growth responses at a finer spatial resolution. By isolating and indexing each segment, the pipeline ensures accurate per-segment measurements and facilitates downstream tasks such as structural comparison, trait quantification, and phenotype classification.

Skeleton extraction and segment identification

Skeleton extraction. Given a binary plant mask M, the logical skeleton of the plant is first computed. The skeleton image is obtained through morphological thinning as ***skeleton image*** = ***skeletonize(M)***. Skeletonization produces a one-pixel-wide representation of the plant structure while preserving topology. This ensures that every path from a terminal pixel (tip) to a junction (branch point) corresponds to a distinct morphological segment.

Node identification. We classify skeleton pixels into three classes according to the degree in the 8-connected neighborhood.

- **Branch points:** $\texttt{BranchPts} = \{p \in \texttt{skeleton_image} : \texttt{countNeighbors}(p) \geq 3\}$ represent junctions where three or more branches meet.
- **Tip points:** $\texttt{TipPts} = \{p \in \texttt{skeleton_image} : \texttt{countNeighbors}(p) = 1\}$ represent terminal ends of branches or leaves.

Algorithm 1 Segment ID Assignment for Plant Skeleton

Require: Binary plant mask M
Ensure: Segment ID map for skeleton pixels
1: $\texttt{skeleton_image} \leftarrow \texttt{skeletonize}(M)$
2: Identify branch points:
$\texttt{BranchPts} \leftarrow \{p \in \texttt{skeleton_image} : \texttt{countNeighbors}(p) \geq 3\}$
3: Identify tip points:
$\texttt{TipPts} \leftarrow \{p \in \texttt{skeleton_image} : \texttt{countNeighbors}(p) = 1\}$
4: Initialize $\texttt{segment_id} \leftarrow 1$
5: **for all** $t \in \texttt{TipPts}$ **do**
6: $\texttt{path} \leftarrow \texttt{traceToNearestBranch}(t, \texttt{BranchPts})$
7: Assign $\texttt{segment_id}$ to all pixels in $\texttt{path}$
8: $\texttt{segment_id} \leftarrow \texttt{segment_id} + 1$
9: **end for**
10: Continue tracing until all skeleton pixels receive a segment ID
11: **return** Segment ID map

Segment tracing. For each tip $t \in \texttt{TipPts}$ we perform a local walk along the skeleton toward the plant interior. The trace proceeds pixel-by-pixel, selecting an adjacent skeleton pixel that has not yet been visited, and terminates when it reaches (a) a pixel in $\texttt{BranchPts}$ or (b) a pixel that already carries a segment identifier. All pixels traversed during this walk are assigned the same segment identifier. Because once labeled a pixel is never relabeled, traced segments are disjoint.

Completion step. After processing all tips, some skeleton pixels (for example central trunk regions or branch-to-branch paths that contain no tips) may remain unlabeled. We perform additional traversals starting from any unlabeled skeleton pixel and assign identifiers to these residual components until every skeleton pixel has been labeled. This guarantees full coverage of the skeleton graph.

Computational Complexity

Let N denote the number of skeleton pixels, T the number of tip points, and B the number of branch points. Skeletonization using standard morphological thinning operates in $O(N)$ time. Neighbor counting for tip and branch classification requires examining the 8-neighborhood of each skeleton pixel, which also results in $O(N)$ complexity. During tip-to-branch tracing, each skeleton pixel is visited at most once because previously labeled pixels are skipped, adding another $O(N)$ traversal. The remaining operations, including computing segment length, endpoints, and centroids, contribute an additional $O(N)$ cost.

Summing these stages, the overall time complexity of the algorithm is $O(N)$, indicating a linear relationship with the number of skeleton pixels. This is asymptotically optimal, since any algorithm must examine each skeleton pixel at least once. The space complexity is also $O(N)$, dominated by storage of the skeleton image and labeling structures.



Figure 3.1: Segment Id

3.1.2 Tip and Branch point Algorithm

The Algorithm 2 operates by scanning every pixel in the skeletonized image and counting the number of skeleton pixels in its 8-connected neighborhood. Pixels with exactly one neighbor are classified as tip points, while those with three or

Algorithm 2 Tip and Branch Point Detection

Require: Skeleton pixel set S (binary skeleton image)
Ensure: Tip point set T, Branch point set B

```
Initialize T ← ∅, B ← ∅
for all p ∈ S do
    n ← countNeighbors(p, S)        ▷ number of 8-connected neighbors in skeleton
    if n = 1 then
        T ← T ∪ {p}                 ▷ tip point
    else if n ≥ 3 then
        B ← B ∪ {p}                 ▷ branch point
    end if
end for
return T, B
```

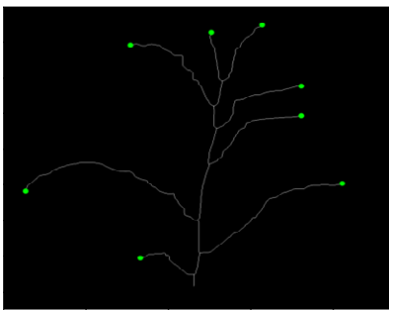


Figure 3.2: (a) Tip points

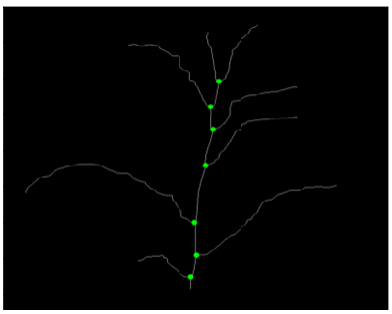


Figure 3.3: (b) Branch points

more neighbors are identified as branch points. Since each skeleton pixel is processed once and neighbor counting is a constant-time operation, the overall time complexity is $O(|S|)$, where $|S|$ is the number of skeleton pixels. These detected tip and branch points serve as the fundamental nodes for subsequent segment-ID algorithms, which trace and label the skeleton paths (segments) connecting these nodes to form a complete graph representation of the structure.

3.1.3 Segment Path Length Algorithm

The Algorithm 3 computes segment length by first extracting the ordered list of pixels along each skeleton segment. It then performs a single linear pass over this list, computing the Euclidean distance between each pair of consecutive pixels

and accumulating these values to obtain the final segment length. Since each pixel in a segment is visited once, the computation for a segment of N pixels is $O(N)$. Because segments are disjoint and every skeleton pixel belongs to exactly one segment, the total complexity over all segments is $O(|S|)$.

Algorithm 3 Segment Length Computation

Require: Binary plant mask M, skeleton image S
Ensure: Length value for each skeleton segment

1: Assign unique segment IDs to skeleton paths from each tip to its nearest branch point(see Algorithm 1)
2: **for all** segment s **do**
3: Extract ordered pixel list $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ along the skeleton path
4: `length` $\leftarrow 0$
5: **for** $i = 1$ to $N - 1$ **do**
6: $d \leftarrow \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$
7: `length` $\leftarrow$ `length` $+ d$
8: **end for**
9: Store `length` for segment s
10: **end for**
11: **return** All segment lengths

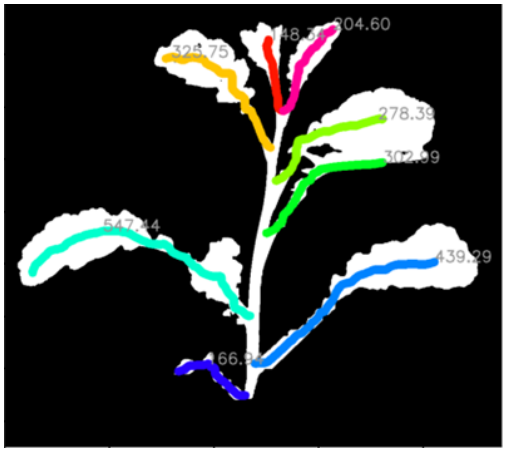


Figure 3.4: Segment Path Length

3.1.4 Segment Insertion Angle

The Algorithm 4 determines the insertion angle of a skeleton segment by first identifying the tip and branch point that define the segment. A direction vector is then formed by subtracting the branch coordinates from the tip coordinates. Using the vertical axis as the reference, the angle between this vector and the stem direction is computed via the `atan2` function, and the result is converted to degrees. When the stem is inclined, an optional rotation step aligns the stem with the vertical axis prior to angle computation. In this way, each branch segment is assigned a consistent insertion angle based solely on its geometric orientation.

Algorithm 4 Segment Insertion Angle Computation

Require: Skeleton segment (tip $\rightarrow$ branch point)
Ensure: Segment insertion angle θ (degrees)
Step 1: Identify key points
Detect tip point and branch point of the segment
Step 2: Compute vector
$V_x = x_{\text{tip}} - x_{\text{branch}}$
$V_y = y_{\text{tip}} - y_{\text{branch}}$
Step 3: Define reference axis
Reference axis = vertical stem direction (Y-axis)
Step 4: Calculate angle
$\theta = \text{atan2}(V_y, V_x)$
Convert θ to degrees if needed
Step 5: Adjust for inclined stems (optional)
If stem is inclined, rotate coordinate system to align stem vertically before computing angle
Step 6: Output
Return insertion angle θ

The algorithm determines the insertion angle of a skeleton segment by first identifying the tip and branch point that define the segment. A direction vector is then formed by subtracting the branch coordinates from the tip coordinates. Using the vertical axis as the reference, the angle between this vector and the stem direction is computed via the `atan2` function, and the result is converted to degrees. When the stem is inclined, an optional rotation step aligns the stem with the vertical axis prior to angle computation.

Since each segment consists of a small, fixed number of points (typically just the tip and branch), computing the vector and angle is a constant-time operation, resulting in an overall complexity of $O(1)$ per segment. In this way, each branch segment is assigned a consistent insertion angle based solely on its geometric orientation.

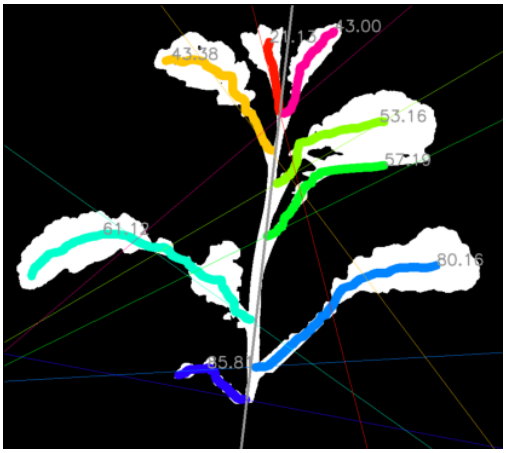


Figure 3.5: Segment Insertion Angle

3.1.5 Segment Tangent Angle

The Algorithm 5 computes the tangent angle at each pixel along an ordered skeleton segment to capture the local orientation of the branch. For each internal pixel i (excluding endpoints), the algorithm forms a vector between the previous and next pixels, (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}), effectively approximating the local tangent direction. The angle of this vector relative to the horizontal axis is then computed using the `atan2` function, giving the tangent angle θ_i at pixel i. Optionally, a sliding-window smoothing operation can be applied to θ_i to reduce noise and obtain a more stable orientation profile along the segment. The working logic ensures that each tangent reflects the direction of the curve at the pixel based on neighboring points. Since the algorithm performs a single pass through all N pixels of the segment with constant-time computations per pixel, the time complexity is $O(N)$ for a segment, and linear with respect to the total number of skeleton pixels when applied to all segments.

Algorithm 5 Segment Tangent Angle Computation

Require: Ordered pixels along a skeleton segment
Ensure: Tangent angle θ_i at each pixel (degrees)

Step 1: Loop through pixels
for $i = 2$ to $N - 1$ **do**
    Compute vector: $V_x = x_{i+1} - x_{i-1}$
    Compute vector: $V_y = y_{i+1} - y_{i-1}$
    Compute tangent angle: $\theta_i = \text{atan2}(V_y, V_x)$
end for
Step 2: Optional smoothing
Apply sliding window smoothing on θ_i to reduce noise
Step 3: Output
Return tangent angles θ_i

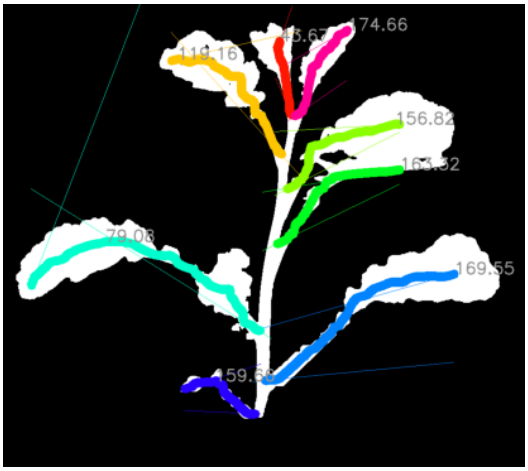


Figure 3.6: Segment Tangent Angle

3.1.6 Segment Curvature Computation

The Algorithm 6 computes the curvature angle at each pixel along an ordered skeleton segment to quantify the local bending of the branch. For each internal pixel i (excluding endpoints), two vectors are formed: V_1 from the previous pixel (x_{i-1}, y_{i-1}) to the current pixel (x_i, y_i), and V_2 from the current pixel to the next pixel (x_{i+1}, y_{i+1}). The curvature angle θ_i at pixel i is then calculated as the angle between these two vectors using the formula $\theta_i = \arccos\left(\frac{V_1 \cdot V_2}{|V_1||V_2|}\right)$. This step effectively captures the local deviation in direction along the skeleton path.

Optionally, a smoothing operation can be applied to the sequence of curvature angles θ_i to reduce noise and obtain a more stable curvature profile. Since the algorithm iterates once over all N pixels in the segment, performing constant-time vector and angle computations per pixel, the time complexity for a segment is $O(N)$, and linearly scales with the total number of skeleton pixels when applied to all segments.

Algorithm 6 Segment Curvature Angle Computation

Require: Ordered skeleton segment pixels
Ensure: Curvature angle along the segment (degrees)

Step 1: Loop through consecutive pixels
for $i = 2$ to $N - 1$ **do**
    Compute vectors: $V_1 = pixel_i - pixel_{i-1}$, $V_2 = pixel_{i+1} - pixel_i$
    Compute curvature angle: $\theta_i = \arccos\left(\frac{V_1 \cdot V_2}{|V_1||V_2|}\right)$
end for
Step 2: Optional smoothing
Apply smoothing to θ_i to reduce noise
Step 3: Output
Return curvature angles along the segment

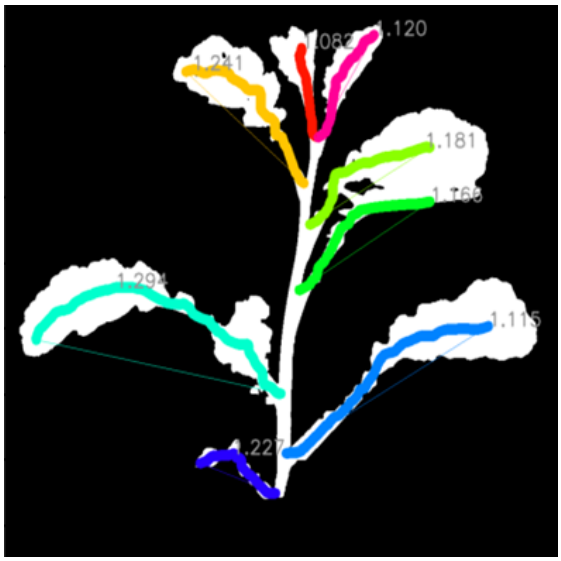


Figure 3.7: Segment Curvature Angle

3.1.7 Euclidean Length

The Euclidean length Algorithm 7 of a skeleton segment represents the straight-line distance between its two endpoints. It is computed using the standard distance formula in a two-dimensional plane, where the differences in the x- and y-coordinates determine how far apart the endpoints lie. This measure provides a simple geometric estimate of the segment's extent without considering intermediate curvature or pixel-level variations along the skeleton. Consequently, the Euclidean length is useful for quantifying the overall span of a branch and for comparing structural differences between segments.

Algorithm 7 Euclidean Length Computation for Skeleton Segments

Require: Ordered skeleton coordinates of a segment: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Ensure: Euclidean length of the segment
1: **Step 1: Initialize**
2: $Length = 0$
3: **Step 2: Compute pairwise distances**
4: **for** $i = 1$ to $n - 1$ **do**
5: $\quad d_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$
6: $\quad Length = Length + d_i$
7: **end for**
8: **Step 3: Output**
9: Return $Length$ as the Euclidean length of the skeleton segment

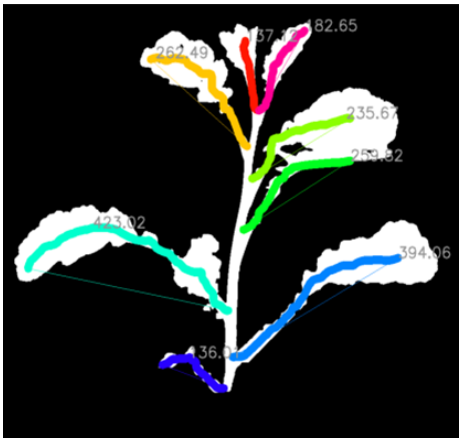


Figure 3.8: Euclidean Length

3.1.8 Convex Hull & Solidity

The Algorithm 8 computes the convex hull and solidity of a plant object from its binary mask. First, the contour of the plant is extracted using a contour detection method, producing an ordered set of boundary pixels representing the plant outline. Let C denote the number of contour points, i.e., the total number of pixels along the boundary. Next, the convex hull polygon that minimally encloses all contour points is computed, typically using algorithms such as Graham's scan or the monotone chain method. The areas of both the original object (from the contour) and the convex hull are then calculated. Solidity is defined as the ratio of the object area to the convex hull area, providing a measure of how compact or concave the plant shape is. The working logic ensures that the convex hull captures the maximal boundary envelope, while solidity quantifies deviations from convexity. Since the convex hull computation processes all C contour points, it runs in $O(C \log C)$ time, and the area calculations require $O(C)$ time. Therefore, the overall complexity is dominated by the convex hull step, giving a total time complexity of $O(C \log C)$, where C is the number of contour pixels.

Algorithm 8 Convex Hull and Solidity Computation

Require: Contour of the plant mask (binary image)
Ensure: Convex hull and solidity value
Step 1: Extract contour
Use contour detection to extract the outline of the plant from the binary mask
Step 2: Compute convex hull
Calculate the convex hull polygon that encloses all contour points
Step 3: Compute areas
$Hull_area$ = area(convex hull polygon)
$Object_area$ = area(contour)
Step 4: Compute solidity
$Solidity = \dfrac{Object_area}{Hull_area}$
Step 5: Output
Return convex hull and solidity value

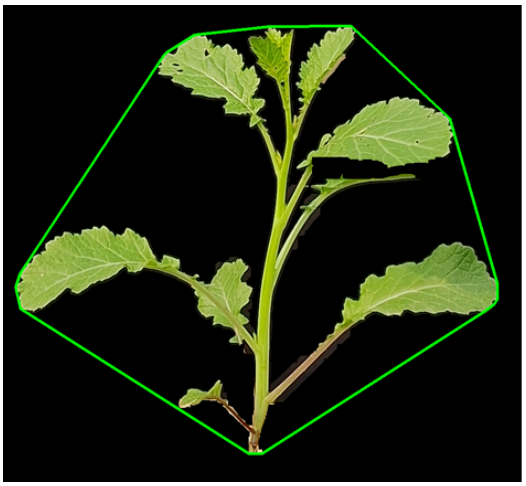


Figure 3.9: Convex Hull Area

3.1.9 Contour Extraction

A contour is a curve that joins all continuous points along the boundary of an object having the same intensity or color. In image analysis, contours are used to detect and represent the shape of objects in a binary or grayscale image. For plant phenotyping, contours provide a precise outline of plant regions, allowing the computation of geometric and morphological features such as area, perimeter, and convex hull. By representing the plant as a sequence of boundary points, contours enable shape-based analysis while ignoring internal texture or color variations.

Algorithm 9 Largest Contour Extraction

Require: Masked RGB image or binary plant mask
Ensure: Largest contour $C_{\max}$
1: Convert RGB image to grayscale (if needed)
2: Apply thresholding to obtain binary mask Bin
3: $Contours \leftarrow \text{findContours}(Bin)$
4: **if** $Contours$ is empty **then**
5: Return "No contour found"
6: **end if**
7: $C_{\max} \leftarrow$ contour with maximum area
8: Return $C_{\max}$

In this step, the objective is to isolate the primary plant region by extracting the largest contour from the binary plant mask. The process begins by converting

the masked RGB image into a grayscale representation to simplify the intensity information. A fixed threshold is then applied to generate a binary image Bin, where plant pixels are represented as white (foreground) and non-plant regions appear as black (background).

The contour detection algorithm then scans the binary image and traces the boundaries of all connected white regions. It identifies continuous pixel edges that form closed shapes, and each detected boundary is stored as an ordered sequence of coordinate points representing its outline. If no such boundary exists, the algorithm returns a "No contour found" message, indicating that the plant structure could not be segmented.

When multiple contours are detected, these may correspond to noise, small detached fragments, or overlapping plant parts. To ensure accurate segmentation, the contour with the maximum enclosed area is selected as the largest contour $C_{\max}$. This contour is taken to represent the main plant body, ensuring that subsequent geometric and morphological feature extraction is performed only on the biologically meaningful region of interest.

3.1.10 Convex Hull Feature Computation

The Convex Hull Feature Computation Algorithm 10 extracts geometric descriptors of the plant shape using the largest contour $C_{\max}$. First, the convex hull of $C_{\max}$ is computed, representing the smallest convex polygon that encloses the plant. The areas of the original contour and the convex hull are calculated, along with the perimeter of the hull. Using these values, solidity is computed as the ratio of contour area to hull area, indicating the compactness of the plant structure. The algorithm returns all features—hull area, contour area, hull perimeter, and solidity which collectively describe the overall shape and structural complexity of the plant.

Algorithm 10 Convex Hull Feature Computation

Require: Largest contour $C_{\max}$
Ensure: $Hull_Area$, $Contour_Area$, $Solidity$, $Hull_Perimeter$
1: $Hull \leftarrow \text{convexHull}(C_{\max})$
2: $Contour_Area \leftarrow \text{contourArea}(C_{\max})$
3: $Hull_Area \leftarrow \text{contourArea}(Hull)$
4: $Hull_Perimeter \leftarrow \text{arcLength}(Hull, \text{True})$
5: $Solidity \leftarrow \begin{cases} \frac{Contour_Area}{Hull_Area}, & \text{if } Hull_Area > 0 \\ 0, & \text{otherwise} \end{cases}$
6: Return all feature values

3.1.11 Compactness Computation

The Algorithm 11 computes the compactness of a plant object by quantifying how closely the object shape fills its convex hull. First, the contour of the plant is extracted from the binary mask, and the areas of both the object (from the contour) and its convex hull are calculated. The compactness is then defined as the ratio of the object area to the convex hull area, providing a measure of how convex or concave the plant shape is. This approach ensures that a perfectly convex object has a compactness of 1, while concave shapes yield values less than 1. If the contour contains C points, computing the convex hull requires $O(C \log C)$ time, while the area computations are $O(C)$. Therefore, the overall complexity is dominated by the convex hull calculation, resulting in a total time complexity of $O(C \log C)$, where C is the number of contour points.

Algorithm 11 Compactness Computation

Require: Contour of the plant mask (binary image)
Ensure: Compactness value
1: **Step 1: Compute areas**
2: $Object_area = \text{area(contour)}$
3: $Hull_area = \text{area(convex hull of contour)}$
4: **Step 2: Compute compactness**
5: $Compactness = \dfrac{Object_area}{Hull_area}$
6: **Step 3: Output**
7: Return compactness value

3.1.12 Eccentricity

Eccentricity is a shape descriptor that quantifies how elongated an object is relative to a perfect circle. It is calculated based on the best-fitting ellipse to the object contour. A value of 0 indicates a perfectly circular object, whereas values approaching 1 indicate a highly elongated shape. In plant phenotyping, eccentricity provides insight into leaf or branch elongation, helping to quantify variations in plant architecture and growth patterns.

where a and b are the lengths of the major and minor axes, respectively. A value of 0 indicates a circular shape, while values approaching 1 indicate high elongation.

Algorithm 12 Eccentricity Computation

Require: Contour of the plant segment
Ensure: Eccentricity value
1: Fit an ellipse to the contour → obtain major axis a and minor axis b
2: Compute eccentricity: $Eccentricity \leftarrow \sqrt{1 - (b^2/a^2)}$
3: Return $Eccentricity$

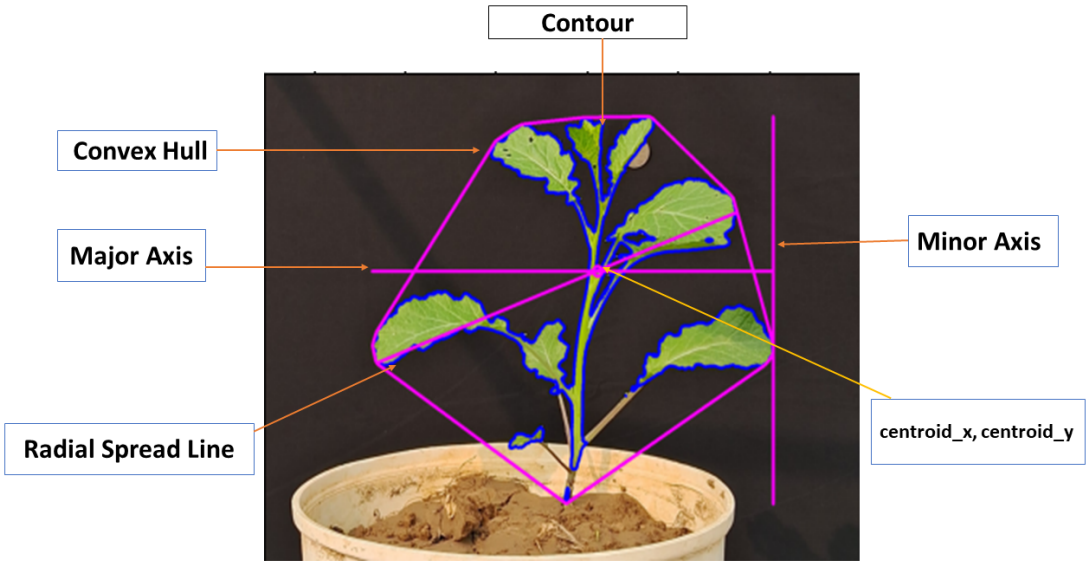


Figure 3.10: Canopy Geometry Visualization

3.1.13 Different colour space

The composite image illustrates how the plant is represented across multiple colour spaces, including HSV, Lab, and CMYK, with each individual channel (H, S, V & L, A, B & C, M, Y, K) displayed in grayscale to highlight the way different colour models encode chromatic and structural information. In the HSV space, the Hue channel captures broad colour differences but remains highly sensitive to illumination and background variation, while the Saturation and Value channels offer strong contrast yet often suppress fine leaf texture and small pigment gradients.

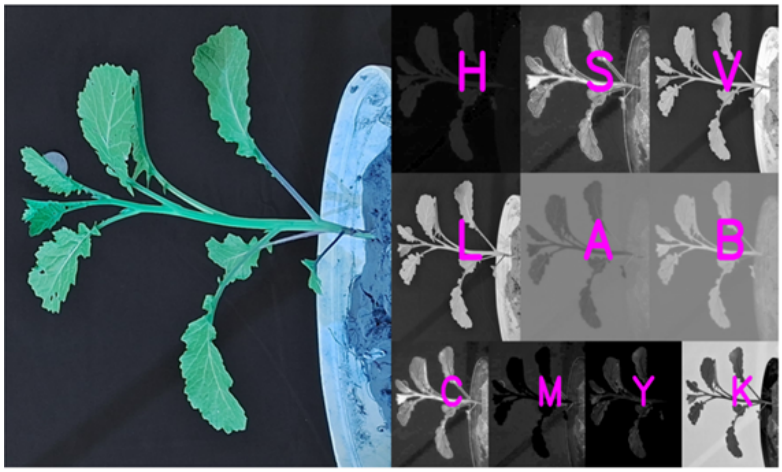


Figure 3.11: Colour Space Visualization

In contrast, the Lab colour space provides a more robust and biologically meaningful representation: the L channel offers smooth luminance separation useful for structural analysis, and the a and b channels distinctly isolate green–red and blue–yellow opponency, capturing subtle pigment differences and early stress signatures that are not well separated in RGB or HSV. The CMYK channels, although useful in printing, exhibit limited biological contrast and contribute little to physiological interpretation. Overall, the comparison clearly demonstrates that the Lab channels preserve both structural detail and chromatic variation with reduced noise, enabling more reliable clustering, segmentation, and colour-based trait extraction. For these reasons, the CIE–Lab colour space was selected in this study, as it provides perceptual uniformity, separates luminance from chromatic infor-

mation, and maps directly onto biologically relevant pigment transitions, making it particularly suitable for precise and reproducible plant phenotyping.

3.1.14 Scatter Plot

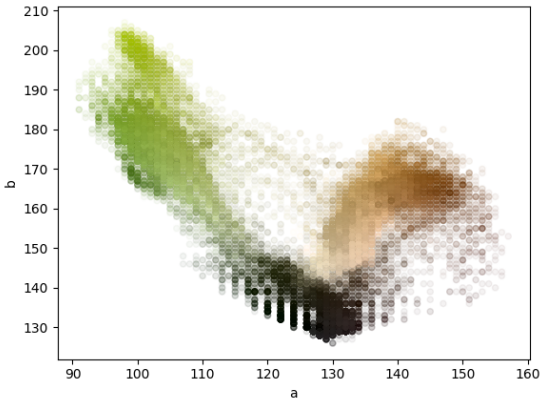


Figure 3.12: Colour Scatter Plot Visualization

Lab a–b

The first image presents a two-dimensional scatter plot of plant pixel values distributed across the a and b channels of the CIE–Lab colour space. Each point corresponds to a pixel from the plant image, positioned by its chromatic components: the a-axis represents the green–red opponency, while the b-axis represents the blue–yellow opponency. In the plot, the plant pixels form clear colour-dependent clusters. The cluster located toward lower a and higher b values corresponds to healthy green tissues, whereas the cluster with higher a values reflects browning or senescent regions. This structured yet overlapping distribution highlights how effectively the Lab colour space separates subtle chromatic differences in vegetation, enabling more robust identification of healthy, stressed, and senescent tissues. Consequently, Lab provides a strong foundation for downstream phenotyping applications such as stress detection, segmentation refinement, and pigment-related trait extraction.

3.1.15 Data set

The high-throughput phenotyping pipeline has been implemented on ten distinct genotypes of Brassica juncea (Indian mustard), namely: RE-2-2, RE-8, RE-13, RE-14, RE-35-4, CN-101849, CN-105306, CN-105308, CN-101846, and CN-112920. Each genotype represents a unique class, and the objective is to evaluate whether phenotypic traits extracted from RGB images such as morphological structure, leaf angle, curvature, and skeleton topology can be effectively used to classify plants according to their genotype. This genotype prediction task was framed as a multi-class supervised classification problem, wherein phenotypic descriptors serve as input features, and genotype labels are the target classes.

CHAPTER 4

Quantifying Plant Structural Complexity Using Structural Complexity Index

4.1 Introduction

Plant phenotyping is a cornerstone of modern agricultural research, enabling the systematic investigation of genotype–phenotype relationships that drive crop improvement. In the context of rising global food demand, climate variability, and population growth, enhancing plant productivity, resilience, and sustainability has become a critical priority for securing future food systems. Consequently, there is an increasing need for phenotyping frameworks that are scalable, repeatable, and capable of capturing biologically meaningful structural information across diverse species and growth conditions.

Traditional plant phenotyping approaches, however, face several fundamental limitations that restrict their applicability in modern high-throughput pipelines. These challenges primarily arise from reliance on destructive sampling, labor-intensive manual measurements, and limited ability to capture the full architectural complexity of plants across multiple developmental stages [6, 7, 8]. Such con-

straints hinder longitudinal analysis and reduce scalability, particularly in large breeding and genetic studies.

Widely used quantitative descriptors typically focus on isolated aspects of plant morphology. Biomass-based measurements provide aggregate estimates of growth but lack spatial or organizational detail. Similarly, the *Leaf Area Index (LAI)*, defined as the ratio of total leaf area to ground area, $\mathrm{LAI} = A_{\mathrm{leaf}}/A_{\mathrm{ground}}$ [1], primarily reflects canopy density rather than architectural organization. Because LAI integrates leaf area along radiative paths, it does not capture spatial arrangement, branching structure, or connectivity among plant organs.

Fractal Dimension (FD) has been employed as an alternative descriptor to characterize global geometric complexity based on image scaling behavior, $D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}$ [63], where $N(\epsilon)$ denotes the number of covering elements at scale ϵ. While FD effectively summarizes space-filling properties across scales, it remains a purely geometric abstraction and lacks biological specificity and topological awareness. As a result, structurally distinct plant architectures may exhibit similar FD values despite substantial differences in branching hierarchy, organ connectivity, or developmental organization.

Consequently, both LAI and FD provide incomplete representations of plant architecture across growth stages and genotypes. These limitations motivate the need for phenotyping descriptors that jointly encode external shape attributes and internal architectural organization in a unified and scalable manner.

To address this gap, this chapter introduces the *Structural Complexity Index (SCI)*, a non-destructive, image-based framework designed to quantify structural variation in whole-plant architecture. The SCI integrates ***complementary shape-based*** and ***skeleton-based descriptors*** extracted from plant images to derive a ***unified measure of structural variation*** across ***growth stages*** and ***genotypes*** [21, 22]. By consolidating high-dimensional morphological information into a single composite statistical index, SCI provides a robust and interpretable baseline for large-scale, longitudinal plant phenotyping. Importantly, it establishes a methodological foundation upon which the advanced non-linear modeling strategies and topological analyses developed in the subsequent chapters are built.

4.2 Proposed Work

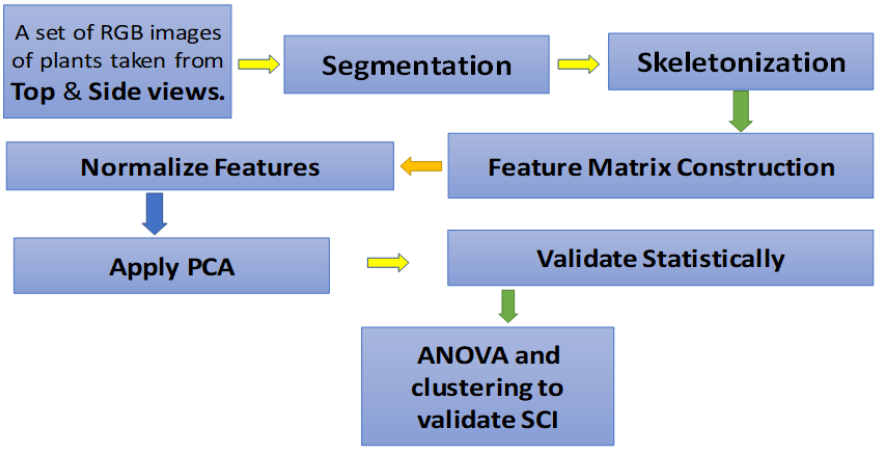


Figure 4.1: SCI Pipeline

4.2.1 Structural Complexity Index (SCI) Framework: Step-by-Step Methodology

The SCI framework is a comprehensive, non-destructive, and scalable methodology designed to characterize structural variation in plant architecture. It leverages image processing and statistical dimensionality reduction (**Principal Component Analysis - PCA**) to convert high-dimensional morphological data into a single, interpretable metric.

Principal Component Analysis (PCA) Principal Component Analysis [44, 68] reduces high-dimensional data by finding new axes that summarize the main directions of variability. After the data are mean-centered, PCA analyzes the covariance among features and obtains orthogonal directions that express the strongest trends in the dataset. Using either the covariance eigen-decomposition or an SVD approach, PCA forms a smaller set of variables that preserves most of the information while removing redundancy.

SCI consolidates high-dimensional morphological data into a single, interpretable metric, enabling efficient comparisons across genotypes, growth stages, and en-

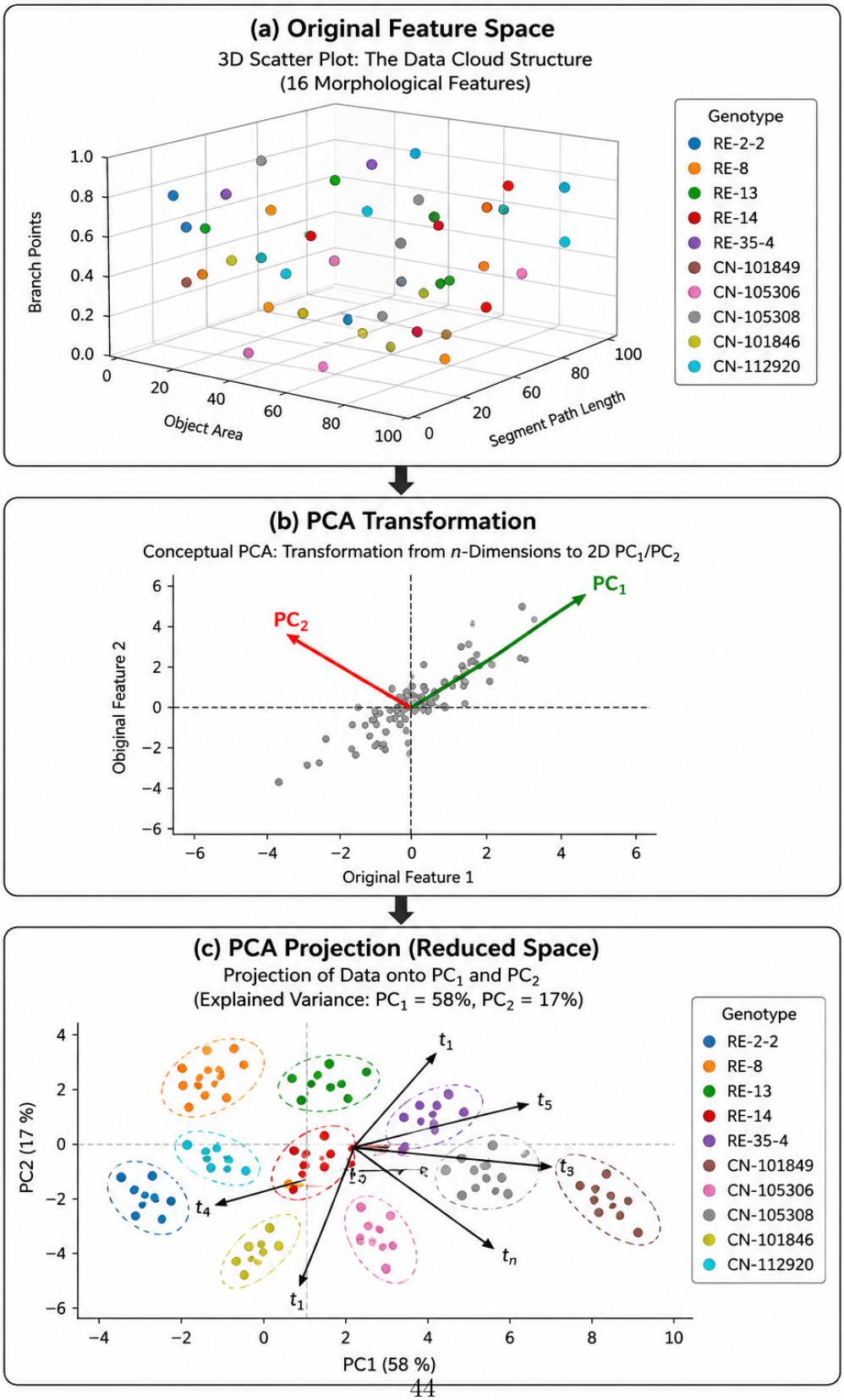


Figure 4.2: Principle Component Analysis

Algorithm 13 PCA Algorithm

Require: Data matrix $X \in \mathbb{R}^{n \times p}$
Ensure: Scores Z, Loadings V, Explained variance η
1: Compute column means: $\mu_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$
2: Center data: $\widetilde{X} = X - \mathbf{1}\mu^\top$
3: **Optional: Standardize columns:** $Y_{ij} = \widetilde{X}_{ij}/s_j$
4: Compute SVD: $Y = U\Sigma V^\top$
5: Compute eigenvalues: $\lambda_k = \sigma_k^2/(n-1)$
6: Compute explained variance ratio: $\eta_k = \lambda_k / \sum_j \lambda_j$
7: Compute scores: $Z = U\Sigma$
8: Select top r components
9: **return** $Z_{:,1:r}$, $V_{:,1:r}$, $\eta_{1:r}$

vironmental conditions [45]. Previous research has demonstrated the utility of advanced imaging and machine learning techniques in automating plant structure analysis [37], [69]. However, these studies have often focused on isolated aspects of plant morphology, unable to provide a comprehensive and scalable solution for capturing the full complexity of plant architecture. SCI fills this gap by combining multiple imaging features into a single index, which can be applied to a wide range of species and growth conditions without the need for destructive sampling or manual interventions [34],[43]. This approach not only simplifies the analysis but also ensures that phenotyping efforts are scalable, repeatable, and applicable in high-throughput breeding and genetic studies. In this study, the mustard plant has been used to validate the ability of the framework by quantifying structural variations across different growth stages. The outcome of the experimental results shows the potential of the SCI framework to overcome limitations of traditional phenotyping methods and provide a robust, non-destructive tool for plant breeders, geneticists, and precision agriculture. The SCI represents a significant advancement in plant phenotyping, offering a standardized metric for assessing plant complexity that can accelerate the development of crops with enhanced yield, resilience, and adaptability to changing environments

I. Image Processing and Feature Extraction

This initial phase captures the raw plant data and transforms it into quantifiable morphological features.

1. **Image Acquisition:** The process begins with collecting Input Data consisting of a set of **RGB images of plants taken from Top & Side views**. Capturing multi-view images (e.g., top and multiple side angles) is crucial for generating a robust 3D representation and accurately extracting architectural features, which helps overcome issues like **self-occlusion** that plague single-view methods.

2. **Segmentation:** The raw RGB images undergo **Segmentation** to isolate the plant (foreground) from the background (pot, soil, and environment). This is typically achieved using color space transformations (e.g., converting RGB to HSV or using ExG indices) and thresholding, or advanced machine learning techniques, to isolate the plant canopy for subsequent analysis.

3. **Skeletonization:** The segmented plant image (or resulting 3D point cloud) is subjected to **Skeletonization**. This process reduces the 3D structure of the plant down to a **one-pixel-wide topological skeleton** (a stick model). The skeleton is critical for extracting skeleton-based morphological features, which describe the internal connectivity and organization of the plant architecture.

4. **Feature Matrix Construction:** Data from both the segmented structure and the skeleton are compiled for **Feature Matrix Construction**. This is the critical step where multi-dimensional morphological features are quantified and compiled into a single data matrix.

    - **Skeleton-based features** include: branch count, branch order, path length, curvature, and branching angles.
    - **Shape-based features** include: projected leaf area, convex hull volume, solid volume, and measures of density/porosity.

    This comprehensive matrix captures both leaf area aspects and the underlying structure, unlike traditional methods like **Leaf Area Index (LAI)**.

II. SCI Computation and Validation

The second phase uses statistical techniques to condense the feature matrix into the final SCI score and verifies its effectiveness.

5. **Normalize Features:** The constructed feature matrix is scaled through **Normalize Features** (e.g., using Z-score or Min-Max scaling). This standardization ensures that all morphological features contribute equally to the final index, preventing features with larger numerical magnitudes (like total path length) from dominating the variance analysis.

6. **Apply PCA (Principal Component Analysis):** PCA is applied to the normalized feature matrix to obtain orthogonal components representing directions of maximum variance in the morphological feature space. The first principal component (PC_1) is retained for further analysis. The Structural Complexity Index (SCI) is defined as the projection score of each plant onto PC_1, providing a single scalar representation of structural variation across samples.

7. **Validate Statistically:** The resulting SCI values undergo preliminary **Statistical Validation**. This step ensures that the derived SCI can significantly differentiate structural variations across different populations (e.g., different genotypes or growth stages).

8. **ANOVA and Clustering to Validate SCI:** The final step involves formal testing:

    - **ANOVA (Analysis of Variance):** Used to formally test the hypothesis that the SCI values are statistically different across experimental groups (e.g., confirming a significant increase in SCI as the plant grows across different weeks).
    - **Clustering:** Techniques (e.g., K-means clustering) are applied to the SCI values to determine if the index can naturally group plants with similar structural characteristics, further confirming the SCI's ability to effectively discriminate between different architectural phenotypes.

Algorithm 14 Preprocessing and Feature Extraction

Input: Image set $P = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$
Output: Feature set $\{F_1, F_2, \ldots, F_n\}$
for each image $\alpha_i \in P$ **do**
    $(I_{\text{top}}, I_{\text{side}}) \leftarrow \texttt{extractTopSideViews}(\alpha_i)$
    $I_{\text{proc}} \leftarrow \texttt{preprocess}(I_{\text{top}}, I_{\text{side}})$
    $M_i \leftarrow \texttt{segmentPlant}(I_{\text{proc}})$
    $M_i \leftarrow \texttt{fillHolesAndDilate}(M_i)$
    $S_i \leftarrow \texttt{skeletonize}(M_i)$
    $S_i' \leftarrow \texttt{pruneSkeleton}(S_i, \text{size_threshold})$
    $(B_i, T_i) \leftarrow \texttt{extractSkeletonFeatures}(S_i')$
    $Sh_i \leftarrow \texttt{extractShapeFeatures}(M_i)$
    $F_i \leftarrow B_i \cup T_i \cup Sh_i$
end for

Algorithm 15 SCI Score Computation using PCA

$F \leftarrow \{F_1, \ldots, F_n\}$
Normalize features: F_{norm}
Perform PCA on F_{norm}
for $i = 1$ to n **do**
    $SCI_i \leftarrow$ Projection of F_i onto the first principal component (PC1)
end for
Return: SCI scores and statistical validation results

4.2.2 Description of the SCI Computational Framework

The Structural Complexity Index (SCI) framework is implemented via a two-stage computational process designed for high-throughput phenotyping: **Preprocessing and Feature Extraction** (Algorithm 14) and **SCI Score Computation** (Algorithm 15).

4.2.3 Structural Complexity Index (SCI) Computation Framework

The initial phase (Algorithm 14)of the SCI framework focuses on preparing the image data and extracting a comprehensive set of morphological features from each plant. Given an input image set $P = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, each plant sample is

processed through a standardized pipeline to ensure consistency.

For each image α_i, top-view (I_{top}) and side-view (I_{side}) representations are extracted and preprocessed using operations such as noise reduction and color normalization. Plant pixels are isolated from the background to generate a refined binary mask M_i, accurately delineating the plant structure.

The segmented mask is then skeletonized to produce a one-pixel-wide topological representation of plant architecture. Minor spurious branches are pruned for robustness. Morphological feature extraction combines **skeleton-based descriptors** (e.g., branch count, path length, branching angles) with **shape-based descriptors** (e.g., projected area, convex hull, volumetric measures). All extracted features are consolidated into a high-dimensional feature vector F_i, serving as input for SCI computation.

The second phase (Algorithm 15) computes a single SCI score from the extracted feature vectors. All individual feature vectors are compiled into a master matrix F and normalized to obtain F_{norm}, ensuring that features with larger numerical scales do not dominate the analysis. PCA is then applied to F_{norm} to extract the principal components that capture the dominant variance in the morphological feature space. The SCI score is computed as the projection of each plant onto the first principal component (PC1):

$$\text{SCI}_i = \sum_{j=1}^{p} w_j \, f_{ij}^{\text{norm}}$$

where:

- f_{ij}^{norm} represents the normalized value of the j-th feature for plant i,
- p denotes the total number of extracted morphological features, and
- w_j denotes the loading (weight) of the j-th feature on PC1.

Thus, although the SCI is derived from multiple morphological features, the resulting SCI value for each plant is a single scalar projection score along the dominant PCA direction. Features associated with larger absolute values of w_j contribute more strongly to the first principal component (PC1) and thus to the variance it captures.

Finally, statistical validation, including ANOVA and clustering analyses, is performed to evaluate the SCI's ability to discriminate structural variations across different plant growth stages and phenotypic groups.

4.3 Results

4.3.1 Segment Curvature

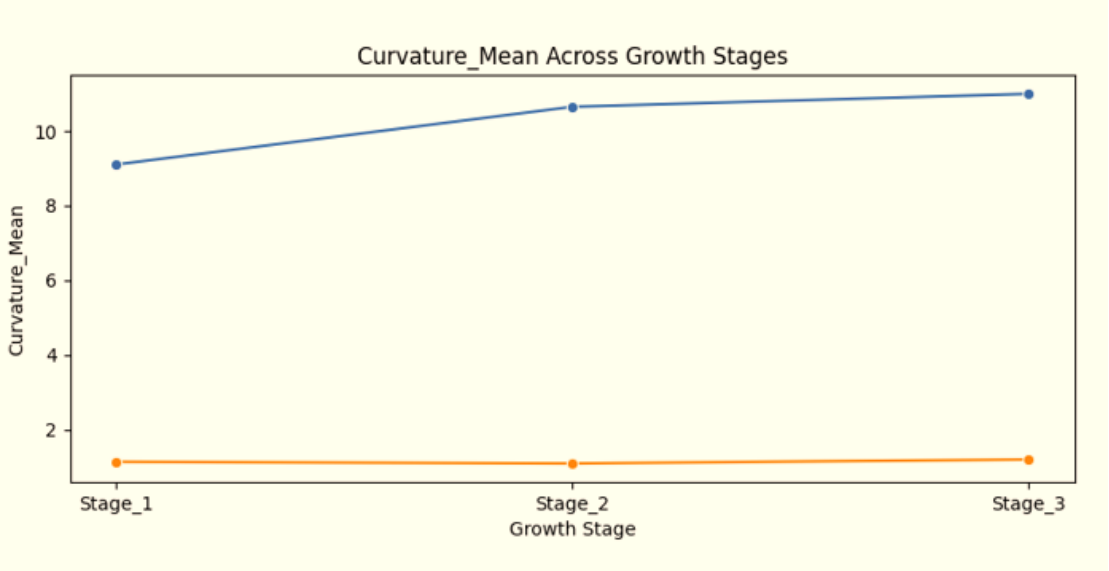


Figure 4.3: Segment Curvature Mean

Figure 4.3 presents the variation in mean curvature across the three growth stages. Curvature measures the local bending or deviation along the plant boundary or skeleton, making it a key geometric descriptor of structural complexity. The mean curvature (blue curve) increases progressively from Stage 1 to Stage 3, indicating that plant outlines become more intricate and irregular as development proceeds. In contrast, the second feature depicted (orange curve) shows minimal change across stages, representing a morphological attribute with limited growth-dependent variation. This contrast demonstrates that while some geometric descriptors are strongly responsive to developmental progression, others remain comparatively stable. Such variability reinforces the need for dimensionality-reduction

approaches like PCA, which integrate diverse feature dynamics into a single, interpretable index of structural complexity.

4.3.2 Segment Path length

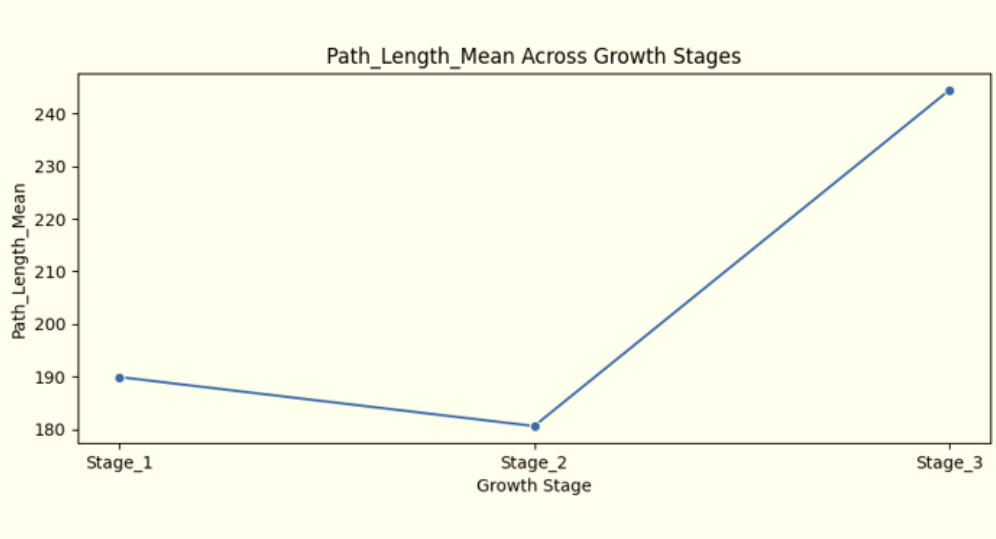


Figure 4.4: Path Length Across Growth Stages.

Figure 4.4 illustrates how the mean segment path length varies across the three growth stages. Skeleton path length represents the cumulative length of the plant's medial axis, making it a useful proxy for structural elongation and branching complexity. The results show a slight decrease from Stage 1 to Stage 2, which may reflect early developmental rearrangements or increased compactness during intermediate growth. In contrast, a sharp increase is observed at Stage 3, corresponding to substantial expansion, branching, and elongation characteristic of mature plant architecture. This nonlinear trend highlights the dynamic nature of structural development within the plant. While skeleton path length provides valuable information about architectural changes, it does not fully capture overall complexity. Therefore, combining this metric with additional morphological descriptors through PCA enables a more comprehensive quantification of structural variation, as reflected in the SCI score.

4.3.3 Different Growth Stages

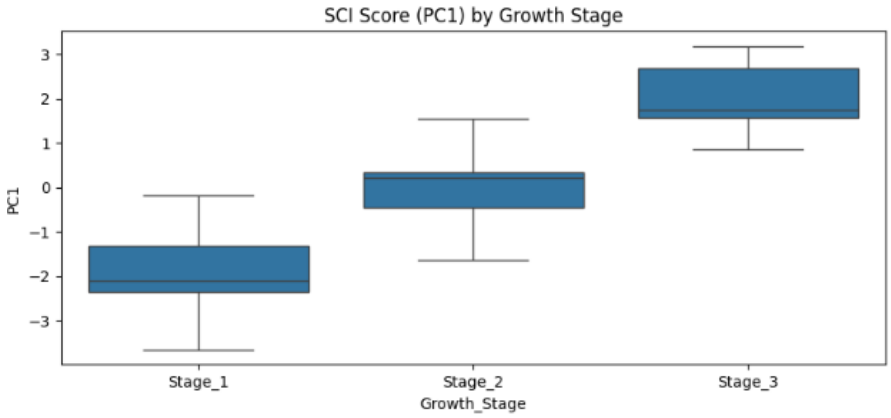


Figure 4.5: Different Growth Stages

Figure 4.5illustrates the distribution of Structural Complexity Index (SCI) values for the three growth stages. The SCI is defined as the first principal component (PC1) derived from the multi-feature dataset, and it provides a compact summary of the dominant shared variation among all morphological and topological descriptors. Plants in Stage 1 exhibit the lowest SCI values, indicating simple and minimally branched structures characteristic of early development. Stage 2 plants display intermediate SCI scores, corresponding to moderate architectural expansion and branching. Stage 3 plants show the highest SCI values, reflecting highly complex, well-developed,and mature plant structures. The clear separation among the three stages demonstrates that PCA successfully captures the developmental progression along a single axis. This validates the SCI as a robust and biologically meaningful measure of plant structural complexity, capable of differentiating growth stages based on their architectural traits

This heatmap Figure 4.6 shows the correlation heatmap of the extracted geometric and skeleton-based plant features. Size-related descriptors such as convex hull area, contour area, and perimeter exhibit strong positive correlations, while skeleton length measures (Euclidean and path lengths) are highly correlated among themselves. In contrast, angular and curvature-based features show weaker correlations with size descriptors, indicating that they capture complementary ar-

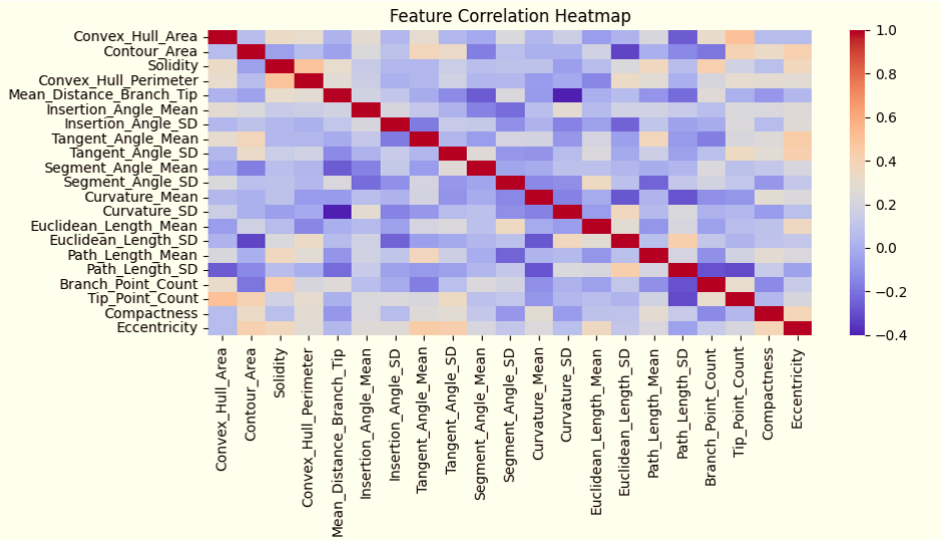


Figure 4.6: Feature Correlation Heatmap.

chitectural information. Overall, the heatmap suggests limited redundancy across major feature groups, supporting the use of a multi-descriptor representation for plant structural analysis.

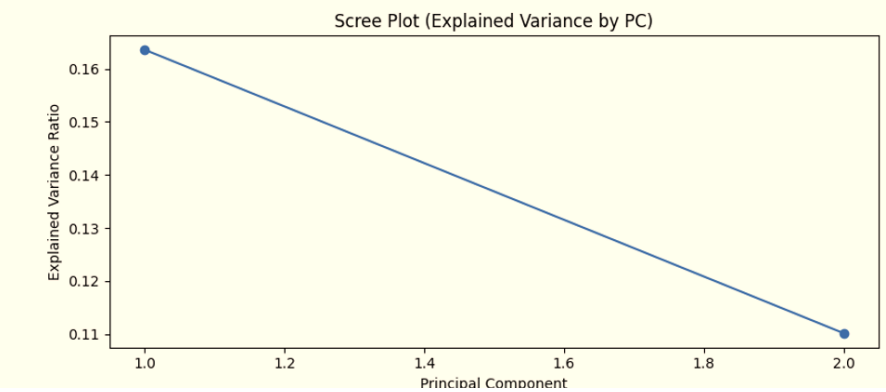


Figure 4.7: Scree Plot of Explained Variance by Principal Components.

The scree plot shown in Figure 4.7 summarizes the proportion of total variance explained by the first two principal components derived from the combined morphological and topological feature set. Principal Component 1 (PC1) captures the largest share of the variance, accounting for approximately 16%, and represents the most dominant structural trend across all measured traits. Principal Com-

ponent 2 (PC2) explains an additional portion of variability, around 11%, and reflects secondary patterns that are not represented in PC1.

Figure 4.8illustrates the result of applying K-means clustering with $k = 3$ to plant phenotypic features after Principal Component Analysis (PCA). The original high-dimensional feature set was first projected onto the first two principal components (PC1 and PC2), which capture the dominant sources of variance in the data. Each point represents an individual plant sample, and colors indicate cluster assignments obtained from K-means.

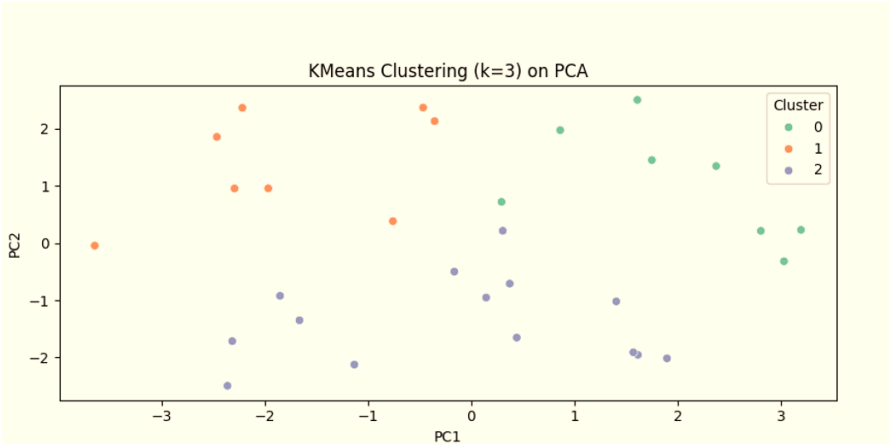


Figure 4.8: K-Means Cluster

Table 4.1: ANOVA Results for Top PCA Features

Feature	**F-Statistic**	**p-value**
Eccentricity	10.3668	0.0005
Tip Point Count	4.9192	0.0151
Solidity	7.2053	0.0031
Convex Hull Perimeter	3.9600	0.0310
Convex Hull Area	7.9699	0.0019

In this study, the ANOVA test was performed on the top five features identified through PCA to assess their statistical significance across different growth stages. The results (Table 4.1) revealed significant differences for several key features. **Eccentricity** ($F = 10.3668$, $p = 0.0005$) showed the strongest significance, indicating that the plant's shape varies significantly across growth stages. **Tip**

Point Count (F = 4.9192, p = 0.0151) and **Solidity** (F = 7.2053, p = 0.0031) also demonstrated significant differences, suggesting they are important for distinguishing plant structural complexity at various stages. Additionally, **Convex Hull Perimeter** (F = 3.9600, p = 0.0310) and **Convex Hull Area** (F = 7.9699, p = 0.0019) were found to vary significantly, highlighting their relevance in evaluating overall plant architecture during development.

4.4 An Overview of Method

The SCI framework is based on a linear projection method (PCA), which may not fully capture the complex, nonlinear relationships among morphological features that characterize plant structures Although the selected feature set is biologically meaningful, it primarily reflects static snapshots of plant architecture and does not consider temporal growth dynamics or environmental factors that can influence phenotypic expression. Future enhancements to this framework could address these gaps. The use of nonlinear dimensionality reduction techniques, such as t-SNE or UMAP, may help reveal hidden patterns within high-dimensional feature spaces. Incorporating temporal data would enable the SCI to represent developmental trajectories over time, yielding more dynamic insights into plant growth. Finally, transitioning from 2D imaging to 3D plant reconstruction could substantially increase the accuracy and depth of structural complexity analysis.

CHAPTER 5

Non-Linear Modeling of Multi-Descriptor Plant Architecture

5.1 Introduction

Plant architecture is an emergent phenotype arising from tightly coupled geometric, skeletal, and topological traits that evolve jointly under genetic control and environmental modulation [15, 16, 17]. While genotyping technologies such as next-generation sequencing have progressed rapidly, phenotyping continues to pose a major bottleneck in the genotype-to-phenotype (G2P) pipeline due to its reliance on limited, labor-intensive, and often subjective trait measurements [9, 10, 11]. Traditional phenotyping approaches typically emphasize basic morphological or spectral attributes such as plant height, projected area, or color indices, which capture isolated aspects of growth but do not explicitly represent the multidimensional architectural organization of plant structure and development [12, 13, 14]. In real plant systems, the influence of individual traits is rarely independent: curvature modulates the functional role of path length, branching density alters the relevance of insertion angles, and compactness mediates how skeletal complexity

translates into spatial occupation. These dependencies are inherently non-linear, non-additive, and frequently non-monotonic, making linear summaries insufficient for capturing the true variability of plant architectural phenotypes [18, 19, 20].

This chapter develops a mathematically grounded framework for representing plant architecture as a high-dimensional, non-linear phenotypic space and for evaluating its utility in genotype discrimination. The proposed formulations—non-linear Structural Complexity Index (nSCI), kernel Fisher discriminants, kernel-based Support Vector Machines, and deep kernel learning—are designed to capture conditional, combinatorial, and non-monotonic dependencies among shape and skeleton-derived traits that cannot be adequately modeled using linear descriptors.

Rather than treating phenotypic traits as independent or additive variables, the methods presented in this chapter model plant architecture as a structured configuration arising from interactions among geometric properties, branching topology, and spatial organization. By embedding phenotypic measurements into reproducing kernel Hilbert spaces (RKHS), these approaches preserve curved manifolds, hierarchical feature interactions, and genotype-specific architectural signatures that are otherwise obscured in Euclidean feature spaces.

To operationalize these representations for genotype prediction, the resulting phenotypic feature matrix—comprising normalized shape, skeleton, and structural descriptors—is used as input to supervised machine learning classifiers for image-based genotype classification. This enables a quantitative assessment of how effectively the extracted architectural features encode genotype-discriminative information.

Random Forest (RF) classification is employed as a strong baseline due to its robustness to noisy and high-dimensional biological data, as well as its ability to estimate feature importance without relying on linear assumptions. To ensure comprehensive benchmarking, additional classifiers are evaluated, including Support Vector Machines (SVM) with radial basis function (RBF) kernels, XGBoost, and Multilayer Perceptrons (MLP).

SVMs provide a kernel-based margin maximization framework that aligns naturally with the RKHS formulations developed earlier in this chapter. XGBoost, as a gradient-boosted decision tree ensemble, enhances generalization through ad-

ditive model optimization and regularization, while implicitly capturing feature interactions. MLPs, as feedforward neural networks, model complex non-linear relationships among phenotypic traits through stacked hidden layers and back-propagation.

Together, these classifiers represent complementary learning paradigms ensemble based inference, margin based optimization, and neural representation learning allowing a rigorous evaluation of genotype–phenotype relationships from multiple methodological perspectives. Beyond predictive performance, this framework supports biological interpretation through feature importance analysis and the identification of architectural traits that contribute most strongly to genotype separation.

Finally, Uniform Manifold Approximation and Projection (UMAP) [18, 19, 20] is applied for non-linear dimensionality reduction to visualize the structure of the learned phenotypic space. By operating on the same multi-dimensional feature set, UMAP reveals compact and well-separated clusters corresponding to distinct Brassica juncea genotypes, providing geometric validation of the discriminative capacity of the extracted architectural descriptors.

Overall, this chapter establishes a coherent pipeline that links non-linear architectural representation to supervised genotype classification, demonstrating how mathematically principled feature modeling enables accurate prediction and interpretable analysis of genotype-specific plant structure. It provides valuable tools for advancing genotype-to-phenotype mapping and supports downstream applications in breeding, stress physiology, and functional genomics [10, 23].

5.1.1 U-Shaped and Non-Linear Trait Relationships in Plant Architecture

In plant morphometrics and architectural studies, many phenotypic traits interact in non-linear and conditional ways, rather than simply proportional or additive relationships. This provides a biological basis for modeling non-linear structural complexity (nSCI) in the RKHS framework.

5.1.2 Curvature and Path Length

The mean branch curvature (c) and path length (L) of plant skeletons are not linearly related across genotypes or growth forms. Empirical studies show that:

- Leaf and branch curvature varies with developmental stage, environmental constraints, and genetic background, creating non-monotonic responses to path length or overall size [70].
- Mechanical and hormonal treatments can induce curvature changes that depend on internode length, demonstrating conditional effects: curvature may increase with path length under one context and decrease under another [71].

These observations (Fig. 5.1) suggest that plants with very short or very long path lengths may exhibit relatively low curvature, whereas intermediate path lengths correspond to maximal curvature. This produces an effective U-shaped relationship [30, 51] between curvature and path length along the genotype spectrum.

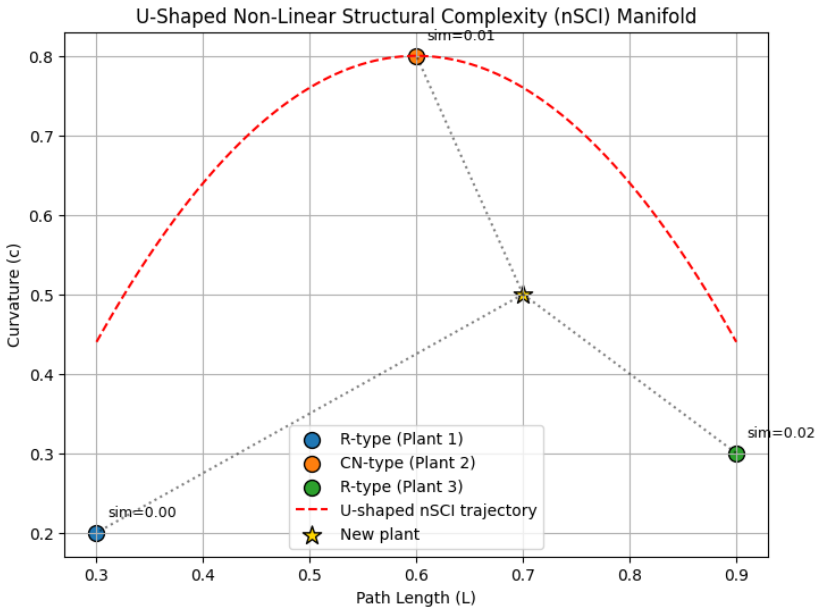


Figure 5.1: U-Shaped nSCI Manifold

5.1.3 Branching Complexity and Allometry

Branching traits, such as branch count and path length, follow fractal or power-law scaling, indicating non-linear associations:

- Branching complexity scales with overall plant size and path length in a non-additive manner, as shown in studies of tree architecture and metabolic scaling [72, 73].
- Combinatorial effects arise because curvature, branch angles, and tip-point distributions interact differently depending on overall branch density, making linear correlations insufficient to capture true architectural variation.

5.1.4 Multivariate Shape and Skeleton Traits

Geometric morphometrics studies further confirm that shape descriptors interact non-linearly:

- Traits such as projected area, solidity, eccentricity, and compactness are jointly influenced by branching structure and curvature [74].
- Multivariate analyses show conditional dependencies where a trait's effect on phenotype depends on the value of other traits, supporting the need for a kernel-based, non-linear approach to capture structural complexity.

5.1.5 Biological Justification for nSCI

These empirical findings justify the use of non-linear structural complexity indices:

- The observed U-shaped or non-monotonic relationships[30, 51] between curvature and path length illustrate why linear indices (like PCA-based SCI) would fail to capture the dominant modes of architectural variation.
- Conditional and combinatorial dependencies among shape and skeleton parameters (e.g., curvature only matters for intermediate path lengths, or insertion angles are relevant only when branch count is high) motivate the RKHS framework and kernel covariance operator used in nSCI.

- The nSCI modes can capture these non-linear trajectories, quantify genotype differences, and provide a biologically meaningful scalar score reflecting overall structural complexity.

5.2 Proposed Work

Although multiple kernel-based techniques are employed in this thesis, they do not represent independent or disconnected algorithms. Instead, they are formulated as complementary components of a single, unified Reproducing Kernel Hilbert Space (RKHS)–based phenotyping framework, designed to characterize plant architectural complexity and genotype-dependent variation in a mathematically coherent and biologically interpretable manner.

Within this unified framework, the Non-Linear Structural Complexity Index (nSCI) provides an unsupervised representation of intrinsic plant architecture. By operating in a kernel-induced feature space constructed from joint shape- and skeleton-based descriptors, nSCI captures the dominant non-linear modes of structural variation that arise from complex, conditional, and combinatorial interactions among architectural traits. Importantly, these modes summarize how plant form and structure co-vary across samples without incorporating genotype labels, thereby yielding a biologically grounded and data-driven measure of structural complexity.

Building upon the same kernel-induced feature space, the Kernel Fisher Discriminant (KFD) is employed as a supervised analytical tool to identify genotype-discriminative structural axes. By explicitly maximizing between-genotype variance relative to within-genotype variability in RKHS, KFD reveals interpretable non-linear directions along which architectural differences between genotypes are most pronounced. In this sense, KFD serves to validate and interpret the structural patterns captured by nSCI, rather than acting as an independent predictive model.

Finally, Support Vector Machines (SVMs) with a Radial Basis Function (RBF) kernel are used to construct a margin-optimal supervised decision function for genotype prediction. Leveraging the same kernel representation, the SVM learns

non-linear decision boundaries that optimally separate genotypes while ensuring robust generalization through margin maximization and regularization.

Crucially, nSCI, KFD, and SVM all operate on an identical kernel-induced feature space derived from shape and skeleton descriptors. This shared mathematical foundation ensures methodological consistency, facilitates biological interpretation, and maintains conceptual coherence across unsupervised structural characterization, discriminative analysis, and predictive modeling. Rather than introducing multiple unrelated models, the proposed framework integrates these components into a single, principled phenotyping pipeline for comprehensive analysis of plant architectural complexity and genotype–phenotype relationships.

5.2.1 Non-Linear Structural Complexity Index (nSCI)

The *non-linear Structural Complexity Index* (nSCI) is an unsupervised measure of plant architectural complexity that captures non-linear, conditional, and combinatorial interactions among multiple morphological traits. Unlike linear indices such as PCA-based SCI, nSCI operates in a kernel-induced feature space, enabling the extraction of intrinsic structural patterns that are not additive in the original phenotypic domain.nSCI does not use genotype labels and is not optimized for classification. Instead, it characterizes the dominant modes of architectural variation present in the population.

Algorithm 16 nSCI Mode Extraction (Unsupervised Structural Complexity)

Require: Phenotype matrix $X = \{x_i \in \mathbb{R}^{21}\}_{i=1}^{N}$, Kernel width $\sigma > 0$
Ensure: nSCI structural modes $\psi_\ell(x)$
1: **Step 1: Kernel Construction** Compute Gaussian kernel matrix: $K_{ij} = \exp\Big(-\|x_i - x_j\|^2/(2\sigma^2)\Big)$
2: **Step 2: Kernel Centering** $\tilde{K} \leftarrow K - 1_N K - K 1_N + 1_N K 1_N$
3: **Step 3: Eigen-Decomposition** Solve eigenvalue problem: $\tilde{K}\beta_\ell = N\lambda_\ell \beta_\ell$
4: **Step 4: nSCI Mode Definition** Compute nSCI mode for plant x: $\psi_\ell(x) = \sum_{i=1}^{N} \beta_{\ell i} k(x, x_i)$
5: **return** $\{\psi_\ell(x)\}$

5.2.1.1 Phenotypic Feature Representation

Each plant is represented by a 21-dimensional phenotypic vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i21}) \in \mathbb{R}^{21}, \tag{5.1}$$

comprising shape descriptors (e.g., projected area, solidity, compactness) and skeleton-derived traits (e.g., path length, branching statistics, curvature, and insertion angles).

5.2.1.2 Kernel-Induced RKHS Mapping

To capture non-linear interactions among traits, phenotypic vectors are implicitly mapped into a Reproducing Kernel Hilbert Space (RKHS) using a Gaussian kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \tag{5.2}$$

where σ controls the scale of non-linearity.

This mapping enables similarity comparisons based on complex trait interactions rather than individual feature differences.

5.2.1.3 Kernel PCA for Structural Mode Extraction

nSCI is computed using **kernel principal component analysis (kPCA)** in RKHS. The centered kernel matrix $\tilde{K}$ is constructed as

$$\tilde{K} = K - 1_N K - K 1_N + 1_N K 1_N, \tag{5.3}$$

and eigen-decomposition is performed:

$$\tilde{K}\boldsymbol{\beta}_\ell = N\lambda_\ell \boldsymbol{\beta}_\ell. \tag{5.4}$$

Each eigenvector defines a non-linear structural mode in RKHS.

5.2.1.4 Definition of nSCI Mode and Score

The ℓ-th nSCI mode evaluated at plant $\mathbf{x}$ is

$$\psi_\ell(\mathbf{x}) = \sum_{j=1}^{N} \beta_{\ell j} k(\mathbf{x}, \mathbf{x}_j). \tag{5.5}$$

The **nSCI score** for plant i is defined as the projection onto the first mode:

$$\text{nSCI}_i = \psi_1(\mathbf{x}_i) = \sum_{j=1}^{N} \beta_{1j} k(\mathbf{x}_i, \mathbf{x}_j), \tag{5.6}$$

which captures the dominant non-linear axis of structural complexity.

5.2.1.5 Biological Interpretation

The first nSCI mode encodes intrinsic architectural variation, often corresponding to curved or U-shaped manifolds in trait space. For example, curvature may contribute strongly to structural complexity only at intermediate path lengths, reflecting conditional trait effects that are invisible to linear methods.

Combinatorial interactions among branching density, curvature, and insertion angles are jointly represented within the RKHS, allowing nSCI to summarize complex plant architectures using a single scalar score.

5.2.1.6 Relation to Genotype Analysis

Although nSCI is unsupervised and independent of genotype labels, its modes can be used as inputs to supervised classifiers. Plants located near non-linear transition regions in the nSCI space often correspond to ambiguous or intermediate architectural forms, which are informative for downstream genotype discrimination.

5.2.1.7 An Overview of Method

The nSCI framework:

- Uses kernel PCA in RKHS to extract non-linear structural modes from plant phenotypes.
- Captures intrinsic architectural complexity without using genotype labels.
- Models conditional and combinatorial trait interactions.
- Provides a scalar structural complexity score for each plant.

5.3 Kernel Fisher Discriminant for Genotype Separation

5.3.1 Definition

Let $\mathcal{H}$ be the Reproducing Kernel Hilbert Space (RKHS) induced by an RBF kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\Big),$$

and let plant phenotypes $\mathbf{x}_i \in \mathbb{R}^{21}$ have genotypes $y_i \in \{+1, -1\}$.

The Kernel Fisher Discriminant (KFD) finds a direction $w^* \in \mathcal{H}$ that maximizes the ratio of between-class to within-class variance:

$$w^* = \arg\max_{w \in \mathcal{H}} \frac{w^\top S_B w}{w^\top S_W w},$$

where the RKHS covariance operators are

$$\begin{aligned} S_B &= (\mu_+ - \mu_-)(\mu_+ - \mu_-)^\top, \\ S_W &= \sum_{i:y_i=+1} (\phi(\mathbf{x}_i) - \mu_+)(\phi(\mathbf{x}_i) - \mu_+)^\top \\ &\quad + \sum_{i:y_i=-1} (\phi(\mathbf{x}_i) - \mu_-)(\phi(\mathbf{x}_i) - \mu_-)^\top \end{aligned}$$

and class means:

$$\mu_\pm = \frac{1}{N_\pm} \sum_{i:y_i=\pm 1} \phi(\mathbf{x}_i).$$

5.3.2 Dual Representation

The optimal discriminant lies in the span of training points:

$$w^* = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i),$$

allowing all computations to be expressed using kernel functions:

$$f(\mathbf{x}) = {w^*}^\top \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}).$$

5.3.3 Interpretation with Shape and Skeleton Features

Let $\mathbf{x} = (\mathbf{x}_{\text{shape}}, \mathbf{x}_{\text{skel}}) \in \mathbb{R}^{21}$, where

- $\mathbf{x}_{\text{shape}} = (\text{area, solidity, eccentricity, compactness, } \dots)$
- $\mathbf{x}_{\text{skel}} = (\text{path length, branch count, curvature, insertion angles, } \dots)$

For genotypes exhibiting a U-shaped curvature-path length pattern, a linear discriminant fails to separate them. KFD in RKHS projects plants onto a *non-linear axis* that separates genotypes while preserving architectural similarity among others.

The predicted genotype is then

$$\hat{y} = \begin{cases} +1 & f(\mathbf{x}) > 0 \\ -1 & f(\mathbf{x}) < 0 \end{cases}.$$

5.3.4 Biological Interpretation

- **Conditional effects:** Curvature may influence the discriminant only for intermediate path lengths.
- **Combinatorial effects:** Branch count, insertion angles, and tip distribution collectively define structural differences.
- **Non-linear projection:** Each plant is represented by a single scalar along the discriminative axis, forming a 1D non-linear manifold:

- Compact upright genotypes → negative end
- Intermediate branching genotypes → middle
- Sprawling genotypes → positive end

5.3.5 Key Points for Thesis Integration

- KFD is **supervised**, directly optimized for genotype separation using labels.
- Operates in the same RKHS as nSCI, allowing comparison and potential integration.
- Captures non-linear, conditional, and combinatorial trait effects explicitly along a single discriminative axis.

5.4 SVM with RBF Kernel — Mathematical Formulation for Plant Traits

Algorithm 17 Supervised SVM with RBF Kernel for Genotype Prediction

Require: Phenotype matrix $X = \{x_i \in \mathbb{R}^{21}\}_{i=1}^{N}$, Genotype labels $y_i \in \{+1, -1\}$, Kernel width $\sigma > 0$, Regularization parameter $C > 0$

Ensure: SVM decision function $g(x)$, Predicted genotype $\hat{y}$

1: **Step 1: Kernel Construction** Compute Gaussian kernel matrix: $K_{ij} = \exp\Big(-\|x_i - x_j\|^2/(2\sigma^2)\Big)$

2: **Step 2: SVM Dual Optimization** Solve for α_i:

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K_{ij}$$

subject to $0 \leq \alpha_i \leq C$, $\sum_i \alpha_i y_i = 0$

3: **Step 3: Construct Decision Function** $g(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i)$

4: **Step 4: Genotype Prediction** $\hat{y} = \text{sign}(g(x))$

5: **return** $g(x), \hat{y}$

5.4.1 Decision Function

For a dataset of plant phenotypes:

$$x_i = (x_{\text{shape}}, x_{\text{skel}}) \in \mathbb{R}^{21}, \quad y_i \in \{-1, +1\}, \quad i = 1, \ldots, N,$$

the SVM decision function with a kernel is defined as:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b \right),$$

where:

- α_i are Lagrange multipliers obtained from the dual optimization problem.
- b is the bias term.
- $k(x_i, x_j)$ is the kernel function mapping input vectors into a high-dimensional feature space.

5.4.2 RBF (Gaussian) Kernel

The Radial Basis Function (RBF) kernel is:

$$k(x_i, x_j) = \exp\left(- \frac{\|x_i - x_j\|^2}{2\sigma^2} \right),$$

where σ controls the smoothness of the decision boundary:

- Large $\sigma \rightarrow$ smoother boundary, risk of underfitting.
- Small $\sigma \rightarrow$ highly flexible boundary, risk of overfitting.

The RBF kernel automatically captures **non-linear, conditional, and combinatorial interactions** among plant traits.

Example in our context:

- Curvature c contributes primarily when path length L is intermediate (conditional effect).

- Branch angles θ and tip distribution t interact non-additively (combinatorial effect).

5.4.3 Dual Optimization Problem

The SVM dual problem with the RBF kernel is:

$$\begin{aligned} \text{maximize}_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \end{aligned}$$

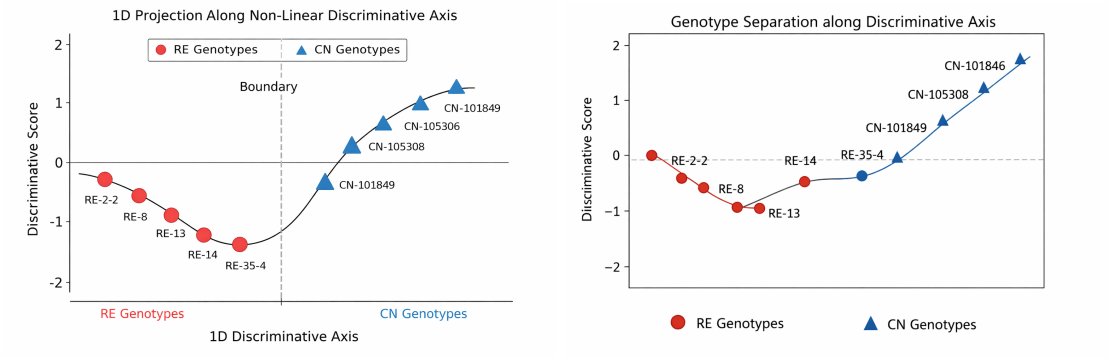
where C is the regularization parameter controlling the trade-off between margin width and misclassification tolerance.

Only points with $\alpha_i > 0$ are **support vectors**, which define the decision boundary in the high-dimensional feature space.

5.4.4 Biological Interpretation for Plant Traits

- Each plant is projected into the high-dimensional RKHS via the RBF kernel.
- SVM identifies a non-linear decision boundary separating genotypes based on shape (x_{shape}) and skeleton (x_{skel}) traits.
- Conditional and combinatorial interactions are automatically captured:
  - Curvature c matters most at intermediate path lengths L.
  - Branch density, insertion angles, and tip distribution combine non-additively.
- **Output:** Each plant receives a single scalar SVM score $f(x)$ for classification.
- Supports interpretation of genotype separation in the non-linear RKHS, providing a supervised counterpart to unsupervised nSCI modes.

5.4 SVM with RBF Kernel — Mathematical Formulation for Plant Traits



(a) nSCI — Unsupervised Structural Complexity

(b) SVM with RBF Kernel and KFD — Supervised Genotype Separation

Figure 5.2: nSCI Visualization

Figure 5.2 Visualization of non-linear structural complexity (nSCI) and supervised genotype separation (SVM/KFD) in Indian mustard. The figure illustrates the distribution of Indian mustard genotypes along non-linear axes derived from both unsupervised and supervised analyses.

(a) nSCI: The first nSCI mode captures dominant non-linear architectural variation. RE genotypes (RE-2-2, RE-8, RE-13, RE-14, RE-35-4) cluster on the negative end, whereas CN genotypes (CN-101849, CN-105306, CN-105308, CN-101846) cluster on the positive end. Each plant's position represents its scalar projection along the first nSCI mode, reflecting latent structural complexity.

(b) SVM with RBF Kernel and KFD: Using the same RKHS, the SVM learns a non-linear decision boundary. Plants are mapped to a scalar SVM score (y-axis) for classification. RE genotypes are represented as red circles, CN genotypes as blue triangles. Conditional effects (e.g., curvature at intermediate path lengths) and combinatorial interactions (branch number, insertion angles, tip distribution) are automatically captured. KFD projects plants onto a supervised non-linear axis optimized to maximize between-class separation, explicitly revealing trait contributions along the discriminative axis.

5.5 Deep Kernel Learning (DKL) for Plant Genotype Discrimination

5.5.1 Motivation

Standard RBF kernels applied directly to raw plant features may fail to capture complex, hierarchical, and conditional dependencies among traits such as curvature, path length, branch count, and insertion angles.

Deep Kernel Learning (DKL) combines neural networks with kernel methods to learn a *data-driven kernel* that adapts to the underlying architecture of phenotypes. This enables **non-linear, conditional, and combinatorial feature interactions** to be captured automatically.

5.5.2 Mathematical Framework

Let $x \in \mathbb{R}^{21}$ denote a plant phenotype vector:

$$x = (x_{\text{shape}}, x_{\text{skel}})$$

where

- x_{shape} = (area, solidity, eccentricity, compactness ...)
- x_{skel} = (path length, branch count, curvature, insertion angles ...)

5.5.2.1 Neural Network Feature Mapping

A neural network $\phi_\theta : \mathbb{R}^{21} \to \mathbb{R}^d$ maps raw features to a latent representation:

$$z_i = \phi_\theta(x_i), \quad i = 1, \ldots, N$$

where $d \ll 21$ is the latent dimension and θ denotes the network parameters (weights and biases).

For a single-hidden-layer network:

$$\phi_\theta(x) = \sigma(Wx + b)$$

with

$$z_k = \sigma\left(\sum_{j=1}^{21} W_{kj} x_j + b_k\right), \quad k = 1, \ldots, d$$

where $W \in \mathbb{R}^{d \times 21}$, $b \in \mathbb{R}^d$, and $\sigma(\cdot)$ is a non-linear activation function (ReLU, tanh, etc.)

Biological Interpretation: Each latent coordinate z_k represents a *learned architectural motif*, capturing higher-order structural patterns. For example:

$$z_1 = \sigma(0.7 \cdot \text{curvature} + 0.6 \cdot \text{path length} - 0.8 \cdot \text{solidity} + b)$$

activating strongly for plants with long, highly curved branches and low solidity—characteristic of a specific genotype.

5.5.2.2 Kernel Construction in Latent Space

Once mapped to the latent space, similarity is computed using a Gaussian kernel:

$$k_\theta(x_i, x_j) = \exp\left(-\frac{\|\phi_\theta(x_i) - \phi_\theta(x_j)\|^2}{2\sigma^2}\right)$$

Two plants are considered similar if they are close in the *learned architectural space*, even if their raw feature values differ substantially.

5.5.2.3 Supervised Genotype Classification

The kernel-based decision function in the learned latent space is:

$$g_\theta(x) = \sum_{i=1}^{N} \alpha_i y_i k_\theta(x, x_i) + b$$

where $y_i \in \{+1, -1\}$ denotes genotype labels, α_i are learned coefficients, and b is a bias term.

The joint optimization over α and θ minimizes the regularized empirical risk:

$$\min_{\alpha,\theta} \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i g_\theta(x_i)) + \frac{\lambda}{2}\|\alpha\|^2$$

5.5.3 Local Trait Sensitivity and Interaction Structure

5.5.3.1 Trait Sensitivity

The gradient of the decision function with respect to input features measures the *local influence* of each trait:

$$S_\theta(x) = \nabla_x g_\theta(x) = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \alpha_i y_i k_\theta(x, x_i)(\nabla_x \phi_\theta(x))^\top (\phi_\theta(x) - \phi_\theta(x_i))$$

Traits influence predictions primarily within specific structural neighborhoods defined by the support vectors.

5.5.3.2 Trait Interaction Structure

Non-additive interactions are captured by the Hessian:

$$\begin{aligned} H_\theta(x) &= \nabla_x^2 g_\theta(x) \\ &= \frac{1}{\sigma^4} \sum_{i=1}^{N} \alpha_i y_i k_\theta(x, x_i) J_\theta(x) \big(\phi_\theta(x) - \phi_\theta(x_i)\big) \big(\phi_\theta(x) - \phi_\theta(x_i)\big)^\top J_\theta(x)^\top \\ &\quad - \frac{1}{\sigma^2} \sum_{i=1}^{N} \alpha_i y_i k_\theta(x, x_i) \nabla_x^2 \phi_\theta(x) \end{aligned} \tag{5.7}$$

where $J_\theta(x) = \nabla_x \phi_\theta(x)$ is the Jacobian of the neural network mapping.

Interpretation: Non-zero mixed partial derivatives indicate combinatorial effects. For example, *branch count* $\times$ *solidity* interacts strongly only in dense architectures.

5.5.4 Advantages over Standard Kernels

- Learns **hierarchical feature interactions** automatically.
- Captures **conditional and non-monotonic dependencies**, including U-shaped trait trends.

- Maps biologically similar but non-identical plants closer in latent space, improving genotype separation.

5.5.5 Empirical Example: U-Shaped Non-Monotonic Curvature Dependency

Consider three genotypes characterized by curvature c and path length L:

$$(c, L) : \text{Genotype } 1 \to (0.2, 0.3), \quad \text{Genotype } 2 \to (0.8, 0.6), \quad \text{Genotype } 3 \to (0.3, 0.9)$$

Linear methods or raw RBF kernels fail to capture the U-shaped dependency, misclassifying Genotype 3. DKL maps these points into a latent RKHS where **conditional dependencies and combinatorial interactions** are represented non-linearly, enabling correct genotype discrimination.

5.5.6 Deep Kernel Learning (DKL) — Supervised Non-Linear Genotype Discrimination

Figure 5.3 illustrates the behavior of Deep Kernel Learning (DKL) for genotype discrimination in Indian mustard. Unlike nSCI, which characterizes intrinsic architectural complexity in an unsupervised manner, and unlike SVM with a fixed RBF kernel or Kernel Fisher Discriminant (KFD), which operate in a predefined Reproducing Kernel Hilbert Space (RKHS), DKL learns the kernel itself through a neural network–based feature mapping.

Raw phenotypic descriptors are first transformed into a latent architectural space, where similarity is computed using a Gaussian kernel. This learned embedding reshapes the RKHS such that plants sharing biologically meaningful structural patterns are mapped closer together, even when their raw phenotypic traits differ substantially.

The figure demonstrates that DKL resolves complex non-linear genotype structure that cannot be fully separated using fixed kernels. In particular, plants exhibiting U-shaped or conditional dependencies between curvature and path length

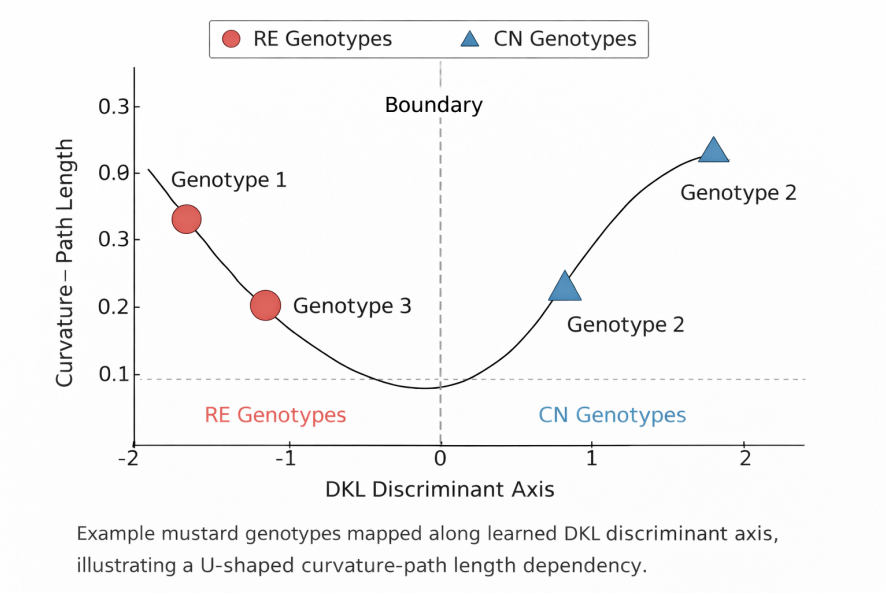


Figure 5.3: Deep Kernel Learning (DKL) for supervised genotype discrimination

are correctly organized along a discriminative axis in the learned RKHS. Hierarchical and combinatorial trait interactions such as curvature acting jointly with branching density and insertion angles are implicitly encoded within the latent representation. As a result, genotype separation becomes more compact and robust, with reduced class overlap compared to SVM with an RBF kernel and KFD.

Overall, the visualization confirms that DKL provides the most expressive supervised model among the evaluated approaches, achieving improved genotype discrimination by jointly learning feature representations and kernel similarity within a unified RKHS framework.

5.6 High-Dimensional Phenotyping and Genotype-to-Phenotype Mapping

5.6.1 An Overview of Method

A modular image-based phenotyping pipeline has been proposed to extract structural features from RGB images of plants and map phenotype representations to genotype labels using supervised machine learning. The pipeline includes fea-

ture engineering, nonlinear dimensionality reduction, clustering, and classification using both Random Forest and XGBoost. It outputs a reduced-dimensional phenotype space, phenotype clusters, genotype prediction models, and important phenotype traits.

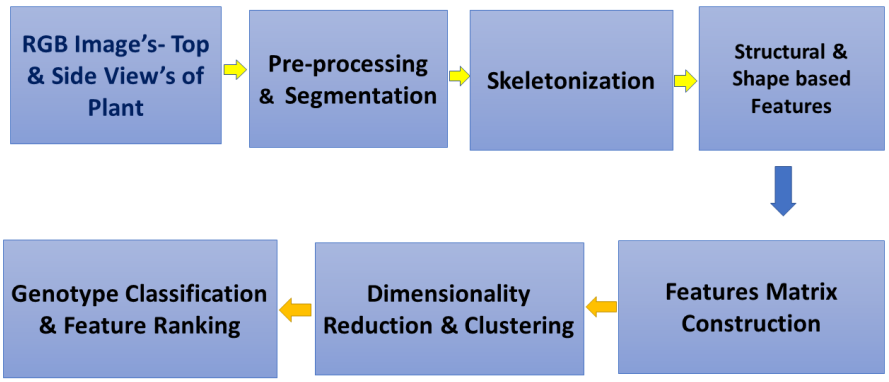


Figure 5.4: Proposed Pipeline

5.6.2 Step 1: Image Acquisition, Preprocessing, and Segmentation

For each genotype, a set of RGB images

$$P = \{\alpha_1, \alpha_2, \ldots, \alpha_n\},$$

where each α_i is an RGB image containing both top and side views of the plant, preprocessing is performed to obtain clean binary masks for subsequent analysis. For each image $\alpha_i \in P$, the top and side views $(I_{\text{top}}, I_{\text{side}})$ are first extracted. Color correction, including white balancing and color calibration using reference cards, is applied to I_{top}. The corrected image is then spatially normalized through cropping and rotation to ensure consistent orientation across samples.

Plant regions are segmented from the background using thresholding and region-of-interest filtering in the Lab color space, resulting in a binary mask M_i. Morphological operations, including hole filling and dilation, are subsequently applied to remove segmentation artifacts and improve mask continuity.

5.6.3 Step 2: Skeletonization and Structural Feature Extraction

From each cleaned binary mask M_i, a medial axis skeleton S_i is computed using skeletonization techniques. To eliminate spurious branches introduced by noise or minor irregularities, the skeleton is pruned using a predefined length threshold τ, yielding a refined skeleton S_i'. Structural features are then extracted from S_i', including the number and spatial distribution of branch points (B_i), terminal tips (T_i), branch lengths, angular relationships, and curvature measures, which collectively characterize the plant's structural topology.

5.6.4 Step 3: Morphological Feature Extraction and Feature Matrix Formation

In parallel, global morphological features are computed directly from the binary mask M_i. These include shape descriptors such as area, convex hull area, solidity, aspect ratio, and compactness, forming a morphological feature set H_i. The skeleton-based features and shape descriptors are then combined to form a unified feature vector

$$F_i = B_i \cup T_i \cup H_i.$$

The complete feature matrix is constructed as

$$F = \{F_1, F_2, \ldots, F_n\},$$

which is subsequently normalized to obtain F_{norm}. This normalized feature matrix serves as the input for downstream analysis and classification.

5.6.5 Step 4: Dimensionality Reduction and Clustering

The extracted feature vectors are aggregated into a unified feature matrix $F = \{F_1, F_2, \ldots, F_n\}$, where each F_i represents the morphological descriptor of an individual plant. Prior to analysis, feature normalization is performed using either Z-score normalization or Min–Max scaling to ensure comparability across heterogeneous descriptors. To capture non-linear relationships and reduce dimensionality,

Algorithm 18 Dimensionality Reduction and Clustering of Plant Features

Require: Normalized feature matrix F_{norm}
Ensure: Reduced feature matrix F_{DR}, clusters
Dimensionality Reduction:
Choose DR method: `DR_method` $\in$ {`KernelPCA`, `UMAP`, `t-SNE`}
$DR_model \leftarrow$ `fitDR`$(F_{\text{norm}}$, `DR_method`$)$
$F_{\text{DR}} \leftarrow$ `transform`$(F_{\text{norm}}, DR_model)$
Clustering:
$Clusters \leftarrow$ `KMeans`(F_{DR}, k)
return $F_{\text{DR}}, Clusters$

kernel-based or manifold learning methods such as Kernel PCA, UMAP, or t-SNE are applied, yielding a reduced representation F_{DR}. Finally, K-Means clustering is performed in the F_{DR} space to identify intrinsic phenotype clusters based on structural similarity.

Algorithm 19 Genotype Classification and Feature Interpretation

Require: Reduced feature matrix F_{DR}, image set P
Ensure: Classification accuracies $RF_Accuracy$, $XGB_Accuracy$ and feature importances $RF_Features$, $XGB_Features$
$Labels \leftarrow$ `genotypeLabels`(P)
Training Classifiers:
$RF \leftarrow$ `trainRandomForest`$(F_{\text{DR}}, Labels)$
$XGB \leftarrow$ `trainXGBoost`$(F_{\text{DR}}, Labels)$
Cross-validation:
$RF_Accuracy \leftarrow$ `crossValidate`(RF)
$XGB_Accuracy \leftarrow$ `crossValidate`(XGB)
Feature Importance:
$RF_Features \leftarrow$ `featureImportance`(RF)
$XGB_Features \leftarrow$ `featureImportance`(XGB)
return $RF_Accuracy, XGB_Accuracy, RF_Features, XGB_Features$

5.6.6 Step 5: Genotype Classification and Feature Interpretation

Genotype classification is performed by training Random Forest (RF) and XGBoost (XGB) models on the reduced feature representation F_{DR} using known

genotype labels. Model performance is evaluated through cross-validation to ensure robustness. Feature importance scores derived from both classifiers are subsequently analyzed to identify and rank the most influential traits contributing to genotype discrimination.

5.6.7 Output

The pipeline outputs Table 5.1:

- A reduced-dimensional phenotype space F_{DR}.
- Phenotypic clusters representing structural variation.
- Trained RF and XGB classifiers with cross-validated accuracies.
- Ranked phenotypic features important for genotype prediction.

5.7 Results

In this study, the SCI framework is extended by integrating it into a comprehensive phenotypic feature set that combines both skeleton-based traits (e.g., path length, curvature, insertion angles) and shape-based traits (e.g., area, solidity, eccentricity). The resulting feature matrix is normalized and used as input to a Random Forest (RF) classifier to enable supervised genotype prediction.

To benchmark classification performance, we evaluate additional classifiers including Support Vector Machines (SVM) [75], XGBoost [76], and Multilayer Perceptrons (MLP) [77]. RF [78] is employed as a strong baseline due to its robustness against noisy and high-dimensional biological data, and its ability to estimate feature importance without relying on linear assumptions. SVM, with its kernel-based approach, effectively separates complex phenotypic classes in high-dimensional spaces. XGBoost, a gradient-boosted decision tree ensemble, improves generalization performance through additive model optimization and built-in regularization. MLP, a type of feedforward artificial neural network, captures nonlinear interactions among features via multiple hidden layers and backpropagation. Together, these classifiers represent a diverse set of learning

paradigms—ensemble methods, margin-based optimization, and deep neural architectures—allowing us to explore the relationship between phenotype and genotype with high fidelity. This multivariate, nonlinear modeling pipeline not only enables accurate genotype classification, but also facilitates the identification of key phenotypic traits through feature importance analysis.

Table 5.1: Classification Performance Comparison

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**
Random Forest	92.4%	0.91	0.92	0.91	0.95
SVM (RBF kernel)	89.6%	0.88	0.89	0.88	0.92
XGBoost	94.1%	0.93	0.94	0.93	0.96
MLP	91.2%	0.90	0.91	0.90	0.93

As shown in Table 5.1, **XGBoost** achieved the highest overall performance across all metrics, with 94.1% accuracy and balanced precision, recall, and F1-score values of 0.93, 0.94, and 0.93, respectively. **Random Forest** also performed strongly, reaching 92.4% accuracy and consistent metric values, demonstrating its robustness in handling complex phenotypic data. **MLP** offered competitive performance, though slightly lower than ensemble methods, while **SVM** with an RBF kernel produced reasonable results but showed a marginal drop in accuracy and recall. These results suggest that ensemble-based methods, particularly XGBoost, are well-suited for phenotype-based genotype prediction, offering both accuracy and reliability across diverse genotypic classes.

Feature importance analysis identified structural traits such as total path length, curvature, and insertion angle variability as among the most influential predictors for genotype classification by both Random Forest and XGBoost models. Additionally, shape descriptors including solidity, eccentricity, and convex hull area contributed substantially to classifier performance. As shown in Table 5.2, these features received the highest importance scores from both methods, underscoring their role in capturing complex phenotypic variations that distinguish *Brassica juncea* genotypes. Features reflecting branching complexity—such as branch point count and number of skeleton tips—were also ranked highly, emphasizing the relevance of detailed architectural traits in genotype discrimination. The strong concordance in rankings across two ensemble classifiers demonstrates

Table 5.2: Top 10 most important features ranked by Random Forest and XGBoost.

Feature	RF Importance	XGB Importance
Total Path Length	0.127	0.144
Mean Curvature	0.115	0.130
Solidity	0.098	0.101
Eccentricity	0.093	0.089
Insertion Angle	0.087	0.085
Convex Hull Area	0.075	0.078
Skewness (Shape)	0.069	0.072
Branch Point Count	0.064	0.070
Aspect Ratio	0.060	0.066
Number of Skeleton Tips	0.055	0.062

the robustness and biological relevance of these phenotypic markers.

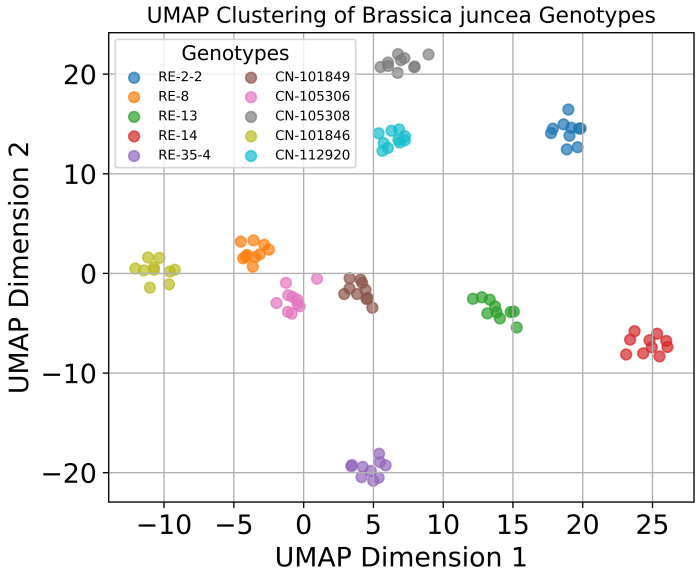


Figure 5.5: UMAP-based clustering of *Brassica juncea* genotypes.

Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction, leveraging a rich set of phenotypic descriptors derived from shape, skeleton, and structural features. UMAP provided superior cluster compactness and better preservation of global structure compared to other nonlinear techniques. As shown in Figure 5.5, the UMAP visualization reveals distinct,

well-separated clusters corresponding to the ten *Brassica juncea* genotypes.

5.8 An Overview of Method

This chapter investigated non-linear modeling strategies for quantifying plant architectural complexity and predicting genotype from image-derived phenotypic traits. A comprehensive phenotypic feature set combining skeleton-based descriptors (e.g., path length, curvature, branching structure, insertion angles) and shape-based descriptors (e.g., area, solidity, eccentricity) was constructed to represent plant architecture.

An unsupervised non-linear Structural Complexity Index (nSCI) was introduced to capture intrinsic architectural variation using a kernel-based representation in RKHS, without relying on genotype labels. The resulting nSCI score provides a biologically interpretable scalar measure of plant structural complexity.

To evaluate genotype discrimination performance, multiple supervised learning models were applied, including SVM with an RBF kernel, Kernel Fisher Discriminant (KFD), Deep Kernel Learning (DKL), Random Forest, XGBoost, and Multilayer Perceptrons. Among these, ensemble-based methods—particularly XGBoost—achieved the highest classification accuracy, while kernel-based models effectively captured non-linear and conditional trait interactions.

Feature importance analysis confirmed that both skeletal and shape traits contribute significantly to genotype separation, highlighting the biological relevance of architectural complexity. Overall, the results demonstrate that combining unsupervised structural complexity modeling with supervised non-linear classifiers provides a robust and scalable framework for phenotype-based genotype prediction

CHAPTER 6

Topology-Driven Genotype and Vigor Characterization Using Persistent Homology

6.1 Introduction

Understanding plant architecture requires analytical frameworks capable of capturing not only local geometric traits but also the global structural organization that emerges from complex growth processes. While conventional morphological descriptors quantify individual traits such as length, area, or curvature, they often fail to represent higher-order structural relationships such as branching connectivity, internal voids, or repeated architectural motifs that characterize plant form across genotypes and developmental stages.

Recent advances in image-based phenotyping have enabled the extraction of rich geometric and skeleton-derived features from RGB images using tools such as PlantCV [21] and related computer vision pipelines [39, 22]. Although these features provide valuable quantitative descriptors of plant shape, they are typically sensitive to noise, image resolution, viewpoint variation, and partial occlusion. Moreover, such descriptors operate at fixed spatial scales and rely on

predefined geometric assumptions, limiting their ability to robustly characterize scale-dependent and global structural properties of plant architecture [62, 79, 61].

To overcome these limitations, this study incorporates concepts from *Topological Data Analysis* (TDA), with a particular focus on *Persistent Homology* (PH), to characterize plant structural complexity in a scale-invariant and noise-resilient manner. Unlike traditional morphometric approaches, PH analyzes the connectivity and spatial organization of structures across multiple resolutions without relying on explicit coordinate systems or rigid shape models [24, 25, 26]. This makes PH particularly suitable for biological structures such as plants, where variability and deformation are intrinsic properties rather than sources of error.

Persistent Homology captures topological features that emerge and persist as a function of a filtration parameter. These features include connected components (0-dimensional homology), loops or holes (1-dimensional homology), and higher-dimensional voids when applicable. As the filtration progresses, features are born, persist across scales, and eventually disappear, allowing meaningful structural patterns to be distinguished from noise. The resulting topological summaries are represented using persistence diagrams or barcodes, which compactly encode the lifespan of each feature [27, 24].

In this work, PH is applied to RGB-derived plant representations by transforming shape contours and skeleton graphs into suitable simplicial or cubical complexes. Homology in dimensions 0 and 1 is computed to identify connected regions and loop-like structures corresponding to branching patterns, internal gaps, and repetitive architectural motifs within the plant canopy. These topological signatures are subsequently vectorized using persistence images or statistical summaries and integrated with machine learning models to enrich phenotypic representations beyond conventional geometric descriptors.

Compared to classical morphological metrics, PH-derived features are inherently invariant to scale, rotation, and minor perturbations, while capturing global structural relationships that are critical for distinguishing subtle architectural differences among genotypes. This enables robust classification, comparative analysis, and clustering of plant architectures across genetic backgrounds or growth conditions.

Persistent Homology can be applied to both static and time-series plant imag-

ing data. For static images, PH provides compact descriptors of overall canopy complexity and internal structural organization. When applied to longitudinal image sequences, PH enables the tracking of topological feature evolution over time, revealing growth dynamics, developmental transitions, and structural responses to environmental stress. Consequently, PH is particularly powerful when used with multi-day phenotyping data, where temporal changes in plant architecture can be systematically quantified.

Beyond methodological contributions, topology-based descriptors have increasing relevance in applied plant science domains. In automated and mechanical harvesting systems, quantifying global canopy structure and branching uniformity is critical for identifying ideotypes compatible with robotics [33, 29]. In controlled-environment and vertical farming systems, topology-aware measures of canopy density and internal void distribution support optimization of light interception and space utilization [80, 81]. Furthermore, deviations in topological signatures can reflect stress-induced structural alterations, enabling robust detection of drought, heat, or nutrient stress responses [82, 47, 65]. Thus, Persistent Homology provides a principled and extensible framework for advancing topology-aware plant phenotyping and genotype-to-phenotype analysis.

Random Forest (RF): Random Forest is an ensemble-based supervised learning method that constructs a collection of decision trees using randomized feature subsets and bootstrapped training samples, and aggregates their predictions through majority voting [49]. This ensemble strategy reduces variance and mitigates overfitting, making Random Forest particularly robust when dealing with high-dimensional and potentially noisy phenotypic descriptors.

In the context of image-based plant phenotyping, the extracted feature space—comprising shape-based metrics, skeleton-derived descriptors, and Persistent Homology (PH) representations—exhibits heterogeneity in scale, distribution, and statistical properties. Random Forest is well suited to such conditions, as it does not assume linear separability or specific feature distributions and can naturally model complex, non-linear interactions among features.

Within this study, Random Forest is employed as a non-parametric reference model to evaluate the discriminative capacity of the combined geometric and topo-

logical feature set. Its inherent ability to capture feature importance and hierarchical decision structures provides additional insight into the relative contribution of morphological versus topological descriptors in plant architecture classification.

Multilayer Perceptron (MLP): The Multilayer Perceptron (MLP) is a feed-forward neural network architecture designed to model complex, non-linear relationships between input features and output classes [83]. In the context of plant phenotyping, the relationships between image-derived descriptors and phenotypic classes are rarely linear, as architectural traits emerge from the interaction of multiple geometric, structural, and topological factors.

Unlike linear classifiers, MLPs learn hierarchical feature representations by progressively transforming the input space through multiple hidden layers. This property makes them particularly suitable for integrating heterogeneous feature sets, such as skeleton-based descriptors and Persistent Homology (PH)–derived topological signatures, which encode complementary aspects of plant structure. Through backpropagation, the network adapts its connection weights to capture non-linear dependencies between these features and the underlying phenotypic categories.

In this study, the MLP serves as a unifying non-linear model that maps the fused feature space—comprising skeleton geometry and PH-based representations—onto phenotypic class labels. Rectified Linear Unit (ReLU) activation functions were employed to facilitate efficient optimization and stable gradient propagation, enabling the network to learn discriminative patterns arising from complex plant architectural organization. By leveraging both geometric and topological information, the MLP effectively captures higher-order structural variations that are not adequately modeled by linear decision boundaries.

Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised learning framework widely used for classification problems involving high-dimensional and structured feature spaces [48]. In image-based plant phenotyping, the extracted descriptors—particularly skeleton-based metrics and Persistent Homology (PH) representations—often reside in a high-dimensional space where class separation is not trivially linear.

SVM addresses this challenge by learning an optimal separating hyperplane

that maximizes the margin between phenotypic classes, thereby enhancing generalization performance. To accommodate the non-linear relationships inherent in plant architectural features, the radial basis function (RBF) kernel was employed. This kernel implicitly maps the fused geometric and topological feature space into a higher-dimensional representation, enabling the separation of complex and overlapping phenotypic patterns.

Within the proposed framework, SVM serves as a strong non-linear baseline for evaluating the discriminative power of the combined skeleton and PH-derived features. Its margin-based formulation provides robustness to noise and high-dimensionality, making it well-suited for assessing whether the extracted structural and topological descriptors yield separable phenotypic classes.

6.2 Proposed Work

Skeleton-Based Feature Extraction. Given an image set $P = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, each plant image α_i is first preprocessed to reduce noise and normalize imaging conditions. The plant region is then segmented to obtain a binary mask $M_i = \text{segmentPlant}(\alpha_i)$. A skeletal representation $S_i = \text{skeletonize}(M_i)$ is computed to capture the medial axis and branching structure of the plant. To suppress spurious branches caused by segmentation artifacts, the skeleton is pruned using a predefined threshold, yielding $S_i' = \text{pruneSkeleton}(S_i)$. Skeleton-based descriptors are then extracted from the pruned skeleton,

$$F_{\text{skel}}^{(i)} = \text{extractSkeletonFeatures}(S_i'),$$

and aggregated over all samples to form the skeleton feature set F_{skel}.

Persistent Homology Feature Fusion and Genotype Classification. For each segmented plant mask M_i, the Euclidean Distance Transform $D_i = \text{computeEDT}(M_i)$ is computed to generate a scalar field encoding local thickness and spatial organization. Persistent homology is applied to the sublevel-set filtration induced by

D_i, producing a persistence diagram

$$PD_i = \text{computePersistenceDiagram}(D_i),$$

which summarizes multi-scale topological features of the plant structure. From each persistence diagram, a set of statistical topological descriptors

$$F_{\text{PHstat}}^{(i)} = \text{extractPHStatistics}(PD_i)$$

is extracted, and a persistence image representation

$$PI_i = \text{generatePersistenceImage}(PD_i)$$

is constructed to obtain a fixed-dimensional encoding of the topological information.

The final fused feature vector for each image is defined as

$$F_i = \text{concatenate}\big(F_{\text{skel}}^{(i)},\, F_{\text{PHstat}}^{(i)},\, PI_i\big).$$

The collection of fused features $F = \{F_i\}_{i=1}^{n}$, together with the corresponding genotype labels $L = \{g_i\}_{i=1}^{n}$, is used to train a supervised classification model. The trained model predicts genotype classes based on the combined skeletal and topological representations.

6.2.1 Topological Features and Summary Statistics

Topological features were extracted from the persistence diagrams corresponding to connected components (H_0) and loops (H_1), capturing complementary aspects of plant structural organization.

Connected Components (H_0). Persistence diagrams in dimension H_0 characterize the emergence and merging of connected regions within the plant structure. From $PD_i^{H_0}$, the following summary statistics were computed:

1. **Total number of connected components:**

$$|PD_i^{H_0}|$$

representing the number of distinct foreground components appearing during the filtration.

2. **Maximum persistence:**

$$\max_j(d_j - b_j)$$

corresponding to the most dominant connected component.

3. **Mean persistence:**

$$\frac{1}{|PD_i^{H_0}|}\sum_j(d_j - b_j)$$

capturing the average lifespan of connected components across scales.

4. **Total persistence:**

$$\sum_j(d_j - b_j)$$

which quantifies the overall stability of connectivity within the plant structure.

5. **Connectivity persistence ratio:**

$$\frac{\text{H0_MaxPersistence}}{\text{H0_MeanPersistence} + \epsilon}$$

measuring the dominance of the principal connected component relative to smaller components, where ϵ is a small constant to ensure numerical stability.

Loops and Holes (H_1). Persistence diagrams in dimension H_1 capture cyclic structures and internal voids that arise from branching patterns, canopy gaps, and repeated motifs. From $PD_i^{H_1}$ the following features were extracted:

1. **Total number of loops:**

$$|PD_i^{H_1}|$$

2. **Maximum persistence:**
$$\max_j(d_j - b_j)$$

3. **Mean persistence:**
$$\frac{1}{|PD_i^{H_1}|}\sum_j(d_j - b_j)$$

4. **Total persistence:**
$$\sum_j(d_j - b_j)$$

5. **Loop density:** number of loops per unit plant area A,
$$\frac{|PD_i^{H_1}|}{A}$$

6. **Loop area (weighted persistence):**
$$\sum_j(d_j - b_j)^2$$

7. **Loop persistence ratio:** dominance of the most persistent loop
$$\frac{\text{H1_MaxPersistence}}{\text{H1_MeanPersistence} + \epsilon}$$
where ϵ avoids division by zero.

6.3 Theoretical Foundations of Persistent Homology–Based Learning

This section establishes the theoretical foundation underlying the proposed image-based genotype–phenotype learning framework. Rather than presenting experimental results, it formalizes how plant images are represented mathematically, how persistent homology acts as a stable structural representation, and how phenotype prediction is built upon this representation. The objective of this section is to provide mathematical justification for robustness, stability, and transferability

of the proposed approach, thereby supporting the methodological choices made in subsequent chapters.

6.3.1 Image Domain and Function Space

Let $\Omega \subset \mathbb{R}^2$ be a compact spatial domain corresponding to the image plane. An RGB plant image is modeled as a bounded function

$$I : \Omega \to \mathbb{R}^3$$

where $I(x) = (I_R(x), I_G(x), I_B(x))$ denotes the color intensity at pixel location $x \in \Omega$.

We define the image space

$$\mathcal{I} := L^\infty(\Omega, \mathbb{R}^3)$$

equipped with the supremum norm

$$\|I\|_\infty := \sup_{x \in \Omega} \|I(x)\|_2$$

This norm provides a natural measure of worst-case pixel perturbations induced by noise, illumination variation, or sensor artifacts.

6.3.2 Perturbation Model

Environmental and imaging variability are modeled as additive perturbations $\delta I \in \mathcal{I}$. Given two observed images $I_1, I_2 \in \mathcal{I}$, we quantify their difference by

$$\|I_1 - I_2\|_\infty \leq \varepsilon$$

for some $\varepsilon > 0$. Stability under this perturbation model is a central requirement for robust genotype–phenotype learning.

6.3.3 Scalar Filtration Construction

To construct persistent homology, we associate each RGB image with a scalar filtration function. Let

$$\phi : \mathbb{R}^3 \to \mathbb{R}$$

be a continuous mapping from color space to intensity, such as grayscale projection, vegetation index, or learned scalarization.

Define the scalar function

$$f_I(x) := \phi(I(x)), \quad x \in \Omega.$$

For each threshold $t \in \mathbb{R}$, we define the sublevel set

$$X_t := \{x \in \Omega \mid f_I(x) \leq t\}$$

The collection $\{X_t\}_{t\in\mathbb{R}}$ forms a filtration of topological spaces satisfying $X_s \subseteq X_t$ for $s \leq t$

6.3.4 Persistent Homology and Diagrams

For each homological dimension $k \geq 0$, we compute the persistent homology of the filtration $\{X_t\}$. This yields a multiset of birth–death pairs

$$D_k(I) := \{(b_i, d_i)\}_{i=1}^{N_k}$$

where each pair corresponds to the appearance and disappearance of a k-dimensional topological feature.

The collection

$$D(I) := \{D_k(I)\}_{k=0}^{K}$$

is referred to as the persistence diagram of the image I up to dimension K. These diagrams summarize the multi-scale topological structure of plant morphology.

6.3.5 Metric Structure on Persistence Diagrams

Let $\mathcal{D}$ denote the space of persistence diagrams. To compare diagrams, we employ the bottleneck distance

$$d_B(D_1, D_2) := \inf_{\gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

where γ ranges over bijections between diagrams allowing matching to the diagonal.

The bottleneck distance provides a natural metric on $\mathcal{D}$ and plays a central role in stability analysis.

6.3.6 Stability of Persistent Homology

The following theorem formalizes the robustness of persistent homology under image perturbations.

Theorem 6.1 (Stability of Sublevel-Set Persistent Homology) *Let $I_1, I_2 \in \mathcal{I}$ be two images, and let $D(I_1), D(I_2)$ denote their corresponding persistence diagrams. Then*

$$d_B(D(I_1), D(I_2)) \leq \|f_{I_1} - f_{I_2}\|_\infty$$

If ϕ is Lipschitz continuous with constant L_ϕ, then

$$d_B(D(I_1), D(I_2)) \leq L_\phi \|I_1 - I_2\|_\infty$$

This result implies that infinitesimal pixel-level perturbations cannot induce large topological changes, establishing persistent homology as a stable representation for image-based learning.

6.3.7 Topological Representation Operator

We define the persistent homology operator

$$\mathcal{P} : \mathcal{I} \to \mathcal{D}, \quad \mathcal{P}(I) := D(I)$$

This operator maps images to a structured, multi-scale topological representation that is invariant to small perturbations and robust to noise.

The separation between representation learning (via $\mathcal{P}$) and prediction is a key aspect of our framework and forms the basis for the stability and transferability results presented in the next section.

6.4 Learning Model, Stability, and Transferability

In this section, we define the learning model built upon persistent homology representations and establish theoretical guarantees for stability and cross-trait transferability.

6.4.1 Learning Model via Topological Representations

Let $\mathcal{P} : \mathcal{I} \to \mathcal{D}$ denote the persistent homology operator defined in Section 3. We consider predictors of the form

$$f = h \circ \mathcal{P}$$

where $h : \mathcal{D} \to \mathbb{R}$ is a regression function mapping persistence diagrams to scalar phenotypes.

This formulation explicitly separates representation learning from prediction. All image variability is handled by the topological operator $\mathcal{P}$, while phenotype-specific information is encoded in h

6.4.2 Lipschitz Continuity of the Regression Map

We assume that the regression function h is Lipschitz continuous with respect to the bottleneck distance. That is, there exists a constant $L_h > 0$ such that for all persistence diagrams $D_1, D_2 \in \mathcal{D}$,

$$|h(D_1) - h(D_2)| \leq L_h \, d_B(D_1, D_2)$$

This assumption is mild and holds for a wide class of learning models, includ-

ing linear regression on vectorized persistence diagrams, kernel methods defined on $\mathcal{D}$, and neural networks with bounded weights operating on stable diagram embeddings.

6.4.3 Stability of the Composite Predictor

We now establish stability of the full genotype–phenotype predictor under image perturbations.

Theorem 6.2 (Stability of the Composite Predictor) *Let $I_1, I_2 \in \mathcal{I}$ be two images such that $\|I_1 - I_2\|_\infty \leq \varepsilon$. Then*

$$|f(I_1) - f(I_2)| \leq L_h L_\phi \, \varepsilon$$

By definition,

$$|f(I_1) - f(I_2)| = |h(\mathcal{P}(I_1)) - h(\mathcal{P}(I_2))|.$$

Using Lipschitz continuity of h, we obtain

$$|f(I_1) - f(I_2)| \leq L_h \, d_B(\mathcal{P}(I_1), \mathcal{P}(I_2))$$

Applying the stability theorem of persistent homology from (Theorem 6.1) yields

$$d_B(\mathcal{P}(I_1), \mathcal{P}(I_2)) \leq L_\phi \|I_1 - I_2\|_\infty$$

Combining the inequalities completes the proof

This result provides an explicit robustness guarantee: small pixel-level perturbations induce proportionally small changes in phenotype predictions.

6.4.4 Genotype–Phenotype Factorization

We now relate the image-based learning model to underlying biological structure. Let G denote genotype, and let $M(G)$ denote the resulting plant morphology, modeled as an abstract geometric object. The observed image I is a projection of $M(G)$ under environmental and imaging conditions.

We assume that phenotypic traits depend smoothly on morphology, i.e.,

$$y = \psi(M(G))$$

for some smooth function ψ. Since persistent homology captures intrinsic topological properties of morphology, we posit the factorization

$$y \approx h(\mathcal{P}(I))$$

6.4.5 Cross-Trait Transferability

Consider multiple phenotypic traits $\{y^{(j)}\}_{j=1}^{J}$ derived from the same underlying morphology. Each trait admits a predictor

$$f^{(j)} = h^{(j)} \circ \mathcal{P}$$

sharing the same topological representation $\mathcal{P}$ but differing in the regression map $h^{(j)}$

This shared representation explains empirical observations of cross-trait transferability: once $\mathcal{P}$ captures the structural essence of plant morphology, multiple phenotypes can be learned efficiently from the same topological features.

6.4.6 Implications for Generalization

The stability and factorization results jointly imply that persistent homology acts as a structural bottleneck that filters out nuisance variability while preserving biologically meaningful information. This explains both improved robustness and improved generalization across environments, genotypes, and phenotypic traits.

6.4.7 Euclidean Distance Transform (EDT)

Given a binary segmentation mask M_i, where $M_i(x, y) = 1$ for foreground (plant) pixels and 0 otherwise, we compute the Euclidean Distance Transform (EDT) to obtain a grayscale image D_i. Each pixel in D_i is assigned the shortest distance to the nearest background pixel, formally computed as:

$$D_i(x, y) = \min_{(x',y') \in \text{background}} \sqrt{(x - x')^2 + (y - y')^2}.$$

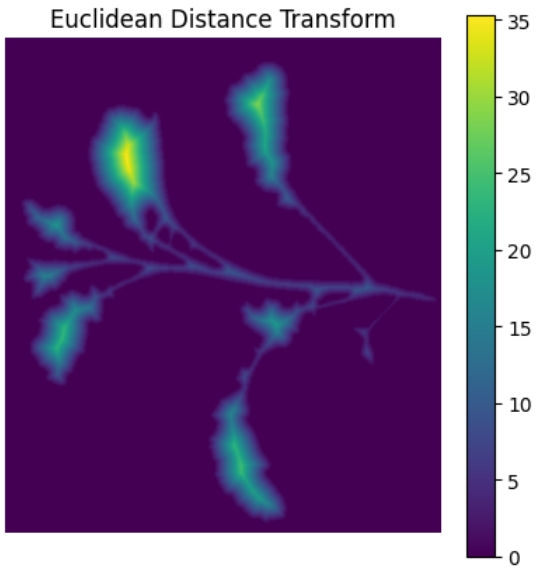


Figure 6.1: Euclidean Distance transform

This transformation provides a "landscape" that highlights the thickness of plant structures—pixels farther from the background receive higher values, effectively encoding the internal density of the plant.

To extract topological features from the plant's shape, we begin by computing a Euclidean Distance Transform (EDT) of the plant mask M_i, producing a grayscale image D_i. This distance map encodes how deep each pixel is within the plant structure interior pixels have higher values, while boundary pixels are closer to zero. From D_i, we construct a topological filtration using a sequence of superlevel sets:

$$F_t = \{(x, y) \mid D_i(x, y) \geq t\}$$

where t varies from the maximum to the minimum pixel value in D_i. As t decreases, more of the plant structure is included in the superlevel set F_t, gradually revealing the entire shape.

We treat the image domain as a 2D cubical complex, where each pixel is a 2-cell (square), and the filtration creates a nested sequence of topological spaces.

6.4.8 Compute Persistence Diagram (PD_i)

Once we have the Euclidean Distance Transform D_i, the next step is to extract a **Persistence Diagram** PD_i using topological data analysis (TDA). This diagram captures the appearance and disappearance of topological features—such as connected components and loops—during a filtration process based on pixel distance values.

- **Filtration:** A *superlevel set filtration* is applied to D_i, where pixels with higher values (typically interior plant regions) are included first, and lower values (near edges) later.
- A **cubical complex** representation—well-suited for grid-based image data—is used to track topological changes throughout this filtration.
- As the threshold decreases:
  - New **connected components** (0-dimensional features, H_0) are born.
  - Existing components may merge, leading to the **death** of one.
  - **Loops or holes** (1-dimensional features, H_1) can form and eventually get filled.

6.4.9 Persistence Diagram

The above figure Figure 6.2 presents a Persistence Diagram for one-dimensional homology (H_1), illustrating the birth and death of topological features detected in the plant structure. Each point represents a loop-like feature whose "birth" marks the scale at which it first appears, and whose "death" indicates the scale at which it disappears during the filtration process. Points positioned far from the diagonal correspond to more persistent and biologically meaningful structural patterns, whereas points close to the diagonal indicate noise or minor geometric fluctuations. The distribution of points across the diagram reflects the complexity of the plant's branching geometry, capturing subtle topological variations that traditional RGB-based or skeletonization methods often miss. This demonstrates the robustness of persistent homology in identifying hidden structural relationships, even under conditions where branch overlap or occlusion occurs.

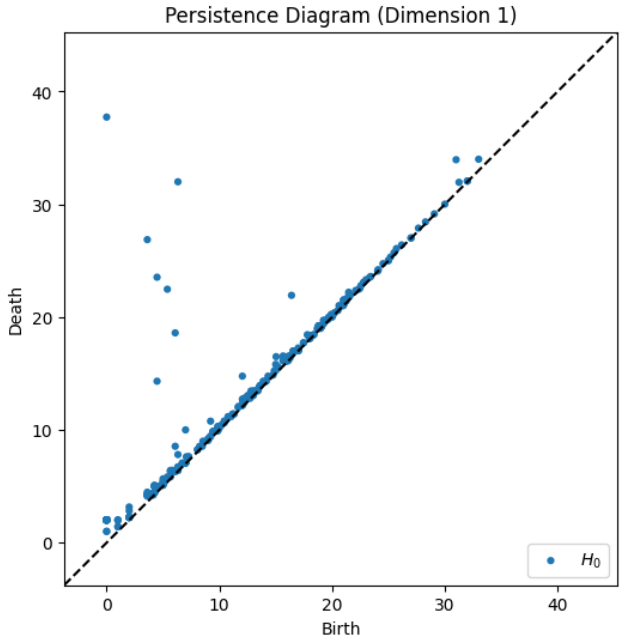


Figure 6.2: Persistence Diagram

6.4.10 Persistence Image Generation:

To provide a fixed-dimensional, differentiable topological descriptor suitable for machine learning, we convert each persistence diagram into a persistence image (PI). First, all (b, d) pairs are transformed into (b, p), where $p = d - b$, emphasizing persistence. This new coordinate space is discretized into a uniform grid (typically 30×30 bins). For each transformed point, a Gaussian kernel is centered at (b, p) and spread across the grid. The pixel value at location (x, y) is computed by summing:

$$I(x, y) += \sum_{(b,p)\in PD} \exp\left(-\frac{(x-b)^2 + (y-p)^2}{2\sigma^2}\right),$$

where σ controls the spread of each point's contribution. The resulting 30×30 matrix is flattened into a 900-dimensional vector $PI_i \in \mathbb{R}^{900}$. This persistence image preserves spatial information of topological features in a format suitable for standard classifiers.

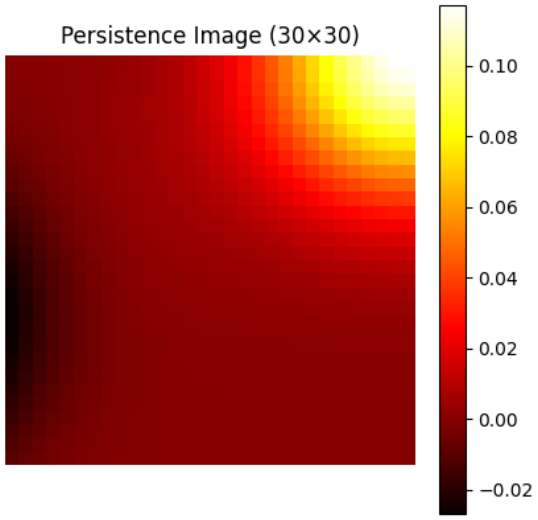


Figure 6.3: Persistence Image

Output: Persistence Diagram PD_i: The result is a multiset of **birth-death pairs** (b_j, d_j), where:

- b_j: filtration value when a topological feature appears,
- d_j: filtration value when it disappears.

Each point in the persistence diagram corresponds to a single topological feature. Features with a large persistence $d_j - b_j$ (i.e., far from the diagonal) are typically significant. In contrast, short-lived features near the diagonal may represent noise.

- Captures **global structural traits** like symmetry, repeated motifs, and internal voids.
- Robust to **noise, deformation, and scale**, offering complementary insights to conventional shape descriptors.
- Enables **quantification of complex morphology** in a principled, scale-invariant way.

6.4.11 Persistent Image

The persistence image (30×30) shown in Figure 6.3 provides a compact, grid-based representation of the topological features encoded in the persistence diagram. Each pixel corresponds to the aggregated contribution of nearby birth–death pairs after applying a Gaussian weighting function. Brighter regions (yellow to white) indicate zones with a high density or strong influence of persistent topological features, while darker regions (black to deep red) represent areas with fewer or less persistent features. In this image, the strong intensity concentrated toward the upper-right portion of the grid suggests the presence of features with relatively higher persistence, which typically correspond to stable structural patterns within the plant morphology. The smooth gradient across the image demonstrates how persistence information is diffused into a continuous feature space, enabling more robust comparisons across samples. This transformation is particularly advantageous for machine-learning workflows, as the persistence image converts complex topological signatures into a fixed-size numerical representation suitable for clustering, classification, or regression tasks.

The second component, persistent homology summary statistics (F_{PHstat}), captures topological information derived from persistence diagrams. Finally, the third component is the persistence image (PI_i), a 900-dimensional vector obtained by flattening a 2D representation of the persistence diagram. This image encodes the spatial distribution and significance of topological features in a form suitable for machine learning. The combined vector $F_i \in \mathbb{R}^{921}$ thus provides a comprehensive morphological and topological characterization of each plant image.

6.5 Results

6.5.1 Feature Concatenation

Once the individual feature components have been extracted for each plant image α_i, they are combined into a single high-dimensional feature vector F_i. This vector integrates three types of features.

All feature vectors are compiled into a dataset:

$$F = \{F_1, F_2, \ldots, F_n\} \in \mathbb{R}^{n \times 921}$$

where n is the number of plant images.

Each vector F_i is paired with a genotype label g_i, forming the supervised label set:

$$G = \{g_1, g_2, \ldots, g_n\}$$

Together, the dataset (F, G) serves as the input for downstream statistical and machine learning models. This integrated feature representation enables robust and interpretable classification of plant genotypes, especially where morphological variations are subtle or distributed across spatial scales.

6.5.2 UMAP-Based Dimensionality Reduction and Analysis

Figure 6.4 was employed for dimensionality reduction, utilizing a comprehensive set of phenotypic descriptors derived from shape, skeleton, and structural features. Compared to other nonlinear methods, UMAP demonstrated superior cluster compactness and better preservation of the global data structure. The UMAP visualization distinctly separates the ten Brassica juncea genotypes into well-defined clusters.

Table 6.1: Comparison of classifier performance

Classifier	**Feature Set**	**Accuracy (%)**	**F1-score (%)**
SVM	Skeleton only	87.4	86.6
SVM	Skeleton + PH	93.7	93.1
Random Forest	Skeleton only	88.5	87.9
Random Forest	Skeleton + PH	94.2	93.6
MLP	Skeleton + PH	92.5	91.1

The experimental results demonstrate that the integration of persistent homology (PH) features with conventional skeleton-based descriptors significantly enhances genotype classification performance across all classifiers. Skeleton-only

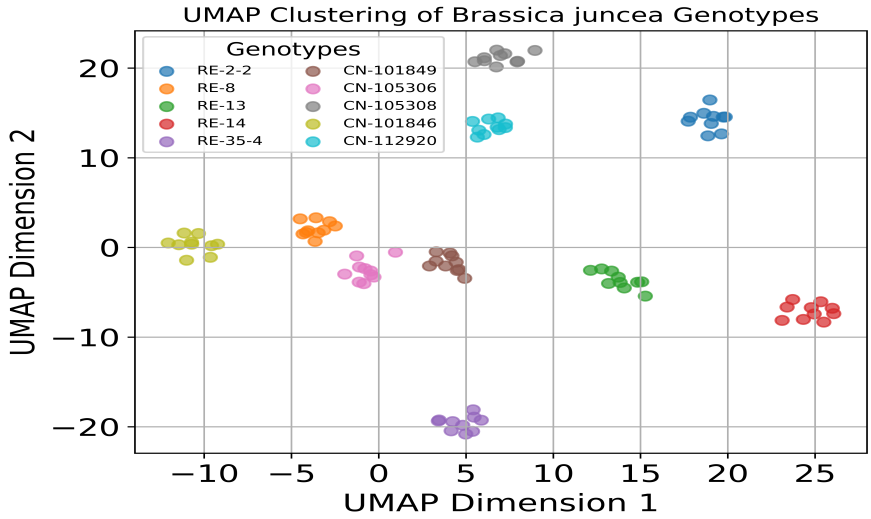


Figure 6.4: UMAP-based clustering of Brassica juncea genotypes.

features, while informative of plant structure, tend to capture local branching patterns and length distributions. However, they often miss global spatial arrangements and voids that are critical for distinguishing complex canopy architectures. PH features, derived from topological data analysis, provide a robust representation of these global patterns by quantifying loops, connectivity, and voids across multiple spatial scales. When combined, the two feature sets offer complementary perspectives on plant morphology—skeleton features encode fine structural traits, while PH captures scale-invariant topological complexity. This fusion results in a richer, more discriminative feature space, which explains the marked improvement in accuracy and F1-score observed in all classifiers, particularly Random Forest and SVM.

6.6 Topological Structural Vigor Index

Most existing vigor studies focus on seed or early-stage vigor using spectral signals and machine learning. In contrast, the present work targets plant-level vigor by analyzing structural and morphological traits of the growing plant. By leveraging topological data analysis, the proposed framework captures architectural complexity and growth patterns that are intrinsically linked to plant vigor but are

not explicitly represented in seed-based or appearance-driven approaches.

This section presents the conceptual formulation, mathematical definition, and biological interpretation of the proposed vigor indices, establishing a theoretically grounded basis for subsequent experimental validation.

6.6.1 Definition of Plant Vigor

Plant vigor is a holistic biological trait that reflects a plant's overall growth potential, structural integrity, and architectural organization under given environmental conditions. In image-based phenotyping, vigor is not solely a function of visual appearance or canopy greenness, but rather emerges from the coordinated interaction between physiological health and structural development.

In particular, plant vigor is characterized by the following structural and spatial properties:

- **Stem and branch connectivity**, which reflects mechanical strength, vascular continuity, and efficient transport of water and nutrients.
- **Canopy closure and loop formation**, indicating leaf overlap, self-supporting architecture, and the organization of photosynthetically active surfaces.
- **Spatial compactness**, capturing how efficiently plant biomass is distributed within the occupied canopy region.
- **Geometric balance**, describing proportional growth along vertical and lateral axes and overall architectural symmetry.

Healthy (control) plants typically exhibit high vigor, characterized by strong skeletal connectivity, dense and closed canopies, compact spatial organization, and balanced geometric growth. In contrast, plants subjected to abiotic or biotic stress show reduced vigor, manifested through stem thinning, fragmented connectivity, loss of canopy closure, and increased spatial sparsity. Importantly, such structural degradation may occur even when color-based indicators remain relatively stable, underscoring the need for vigor descriptors that extend beyond appearance alone.

6.6.2 Topological Representation Using Persistent Homology

To capture structural and canopy-level properties from RGB plant images, the Euclidean Distance Transform (EDT) is computed from the binary plant mask. Persistent homology is then applied on the EDT-based filtration to extract topological features.

Two homology dimensions are considered:

- H_0: Represents connected components and captures global structural connectivity such as stem thickness and branching continuity.
- H_1: Represents loops and captures canopy closure, leaf overlap, and architectural complexity.

Persistent homology descriptors P_{H_0} and P_{H_1} are obtained by summarizing the corresponding persistence diagrams.

6.6.3 Geometric Vigor Index (GVI)

While topological features capture global structural organization, plant vigor is also reflected in measurable geometric, skeletal, and appearance-based properties. To complement the Topological Vigor Index, a Geometric Vigor Index (GVI) is constructed using traits extracted from PlantCV and skeleton-based analysis.

Geometric Shape Descriptors The following geometric traits are extracted from the segmented plant mask Figure 3.10:

- **Compactness**: Measures canopy density and spatial efficiency, reflecting how tightly plant biomass is distributed.
- **Convex Hull Area**: Represents overall plant spread and lateral expansion.
- **Major Axis Length**: Indicates plant height and dominant growth direction.
- **Minor Axis Length**: Indicates lateral growth and canopy width.

Skeleton-Based Geometric Descriptors To capture internal growth organization beyond global shape, skeleton-derived geometric properties are also included:

- **Total Skeleton Length**: Measures cumulative stem and branch extent.
- **Mean Segment Length**: Reflects internode elongation and growth uniformity. Figure 3.4
- **Branch Angle Statistics**: Quantifies angular distribution of branches, related to canopy openness and architecture. Figure 3.5
- **Euclidean Distance Measures**: Distances between key skeletal endpoints and the plant centroid, capturing growth balance and symmetry.Figure 3.8

Color-Based Appearance Descriptors In addition to structural geometry, physiological vigor influences plant color distribution. Therefore, color statistics derived from pixel scatter in color space are incorporated Figure 6.3:

- **Mean and variance of Lab a–b scatter**: Captures chlorophyll dominance and stress-induced color shifts.
- **Color compactness**: Measures dispersion of plant pixels in chromatic space.

Geometric Vigor Index Definition After normalization of all features, the Geometric Vigor Index is defined as a weighted linear combination:

$$\mathrm{GVI} = \sum_{i=1}^{p} w_i \, G_i \tag{6.1}$$

where G_i denotes the i-th geometric, skeleton-based, or color-based feature, and w_i is its corresponding weight. This formulation allows flexible integration of multiple non-topological descriptors within a single vigor measure.

6.6.4 Data-Driven Weight Selection for Geometric Features

The weights $\{w_i\}$ are learned using a data-driven optimization strategy based on Fisher's separation criterion. Each feature is standardized prior to weighting to ensure scale invariance. Optimal weights are selected to maximize separation between control and stressed plants, ensuring that features most sensitive to vigor degradation contribute more strongly to the final index.

This data-adaptive formulation allows the Geometric Vigor Index to capture changes in plant size, growth direction, skeletal organization, and color distribution that are not explicitly encoded by topological descriptors.

6.6.5 Data-Driven Selection and Interpretation of α

The weighting parameter $\alpha \in [0, 1]$ plays a central role in the definition of the Topological Vigor Index, as it governs the relative contribution of structural connectivity (H_0) and canopy complexity (H_1). Rather than fixing α heuristically, we determine its value using a data-driven optimization strategy to ensure objective and biologically meaningful weighting.

For a given value of α, the Topological Vigor Index is computed for all plants as

$$\mathrm{TVI}_{\mathrm{topo}}(\alpha) = \alpha\, P_{H_0} + (1 - \alpha)\, P_{H_1} \tag{6.2}$$

To quantify how well a particular α separates healthy (control) plants from stressed plants, we employ the Fisher separation score:

$$F(\alpha) = \frac{(\mu_{\mathrm{control}} - \mu_{\mathrm{stress}})^2}{\sigma^2_{\mathrm{control}} + \sigma^2_{\mathrm{stress}}} \tag{6.3}$$

where μ_{control} and $\sigma^2_{\mathrm{control}}$ denote the mean and variance of $\mathrm{TVI}_{\mathrm{topo}}$ for control plants, and μ_{stress} and $\sigma^2_{\mathrm{stress}}$ denote the corresponding statistics for stressed plants.

This criterion favors values of α that simultaneously maximize the difference between class means while minimizing within-class variability, thereby promoting both discriminability and robustness.

The optimal weighting parameter is selected as

$$\alpha^* = \arg\max_{\alpha \in [0,1]} F(\alpha) \tag{6.4}$$

Interpretation of α

The learned value α^* admits a clear biological interpretation:

- $\alpha^* \approx 1$ indicates that plant vigor is primarily driven by structural connectivity, such as stem thickness, branch continuity, and global skeletal integrity.
- $\alpha^* \approx 0$ indicates that canopy complexity and loop formation dominate vigor differences, reflecting leaf overlap, canopy closure, and architectural density.
- Intermediate values of α^* suggest that both connectivity and canopy structure jointly contribute to vigor, consistent with balanced plant growth.

Practical Implications

Once determined, α^* is fixed and used consistently across all experiments, genotypes, and stress conditions. This prevents overfitting and ensures that vigor comparisons remain stable and interpretable. Importantly, this data-driven calibration allows the Topological Vigor Index to adapt to different species, developmental stages, or stress regimes while preserving a single, unified mathematical formulation.

6.6.6 Hybrid Vigor Index (HVI)

Finally, a Hybrid Vigor Index is defined to integrate both topological and geometric information:

$$\mathrm{HVI} = \gamma \cdot \mathrm{TVI}_{\mathrm{topo}} + (1 - \gamma) \cdot \mathrm{GVI} \tag{6.5}$$

where $\gamma \in [0, 1]$ controls the balance between topological and geometric contributions and is selected using data-driven optimization.

Biological Interpretation of γ

The weighting parameter γ has direct biological significance. Topological traits, such as connectivity and branching patterns, are typically more informative during early growth stages, as they guide developmental potential. In contrast, geometric traits, including canopy size and spatial spread, dominate during later stages of growth. Under stress conditions, structural degradation often precedes visible reductions in plant size; consequently, topological descriptors serve as early indicators of vigor loss, whereas geometric traits capture more advanced growth impairment.

Data-Driven Optimization of γ

Rather than assigning γ heuristically, its value is determined using a data-driven optimization strategy that maximizes separability between control and stressed plants. This ensures that the relative weighting of topological and geometric contributions reflects observed phenotypic differences, rather than subjective assumptions. Once the optimal value γ^* is identified, it is applied consistently across all genotypes and experimental conditions, ensuring reproducibility and comparability of the Hybrid Vigor Index.

6.6.7 Comparison with Traditional Systems

Existing plant phenotyping platforms predominantly rely on color- and geometry-based descriptors to assess plant status. Commercial systems such as LemnaTec emphasize spectral indices and global geometric traits, which are effective for monitoring biomass accumulation and stress-induced color changes but do not explicitly encode internal architectural organization or connectivity. Open-source frameworks such as PlantCV provide a rich set of modular tools for extracting morphological and skeletal traits; however, these features are typically analyzed independently and are not integrated into a unified vigor index. As a result, structural degradation affecting branching continuity or canopy closure may remain underrepresented, particularly when color-based indicators remain stable.

In contrast, the proposed Hybrid Vigor Index (HVI) integrates geometric,

skeletal, and topological descriptors within a principled mathematical framework. By incorporating persistent homology, the framework captures higher-order architectural properties such as structural connectivity and canopy closure, while data-driven weighting ensures robust and interpretable integration of complementary trait domains. This enables a more comprehensive assessment of plant vigor that reflects both growth magnitude and architectural integrity.

6.7 An Overview of Method

This chapter demonstrated the effectiveness of Persistent Homology (PH) for enhancing genotype discrimination by capturing global architectural characteristics that are not represented by conventional geometric or skeleton-based descriptors. UMAP-based visualization revealed that the inclusion of PH features led to more compact and well-separated genotype clusters, indicating improved preservation of intrinsic structural relationships among Brassica juncea genotypes.

The classification results further confirmed that integrating PH with skeleton-based features consistently improved performance across all evaluated classifiers. While skeleton descriptors effectively encode local branching patterns and length distributions, they are limited in representing global spatial arrangements, canopy closure, and internal void structures. PH complements these descriptors by quantifying topological properties such as connectivity, loops, and void persistence across multiple spatial scales. This complementary fusion results in a richer and more discriminative feature space, explaining the substantial gains in accuracy and F1-score observed for SVM, Random Forest, and MLP classifiers.

Beyond genotype discrimination, the topology-aware feature representation provides meaningful insight into plant vigor and hybrid performance. Hybrid vigor manifests not only as increased size but also as improved architectural coherence, connectivity, and structural integrity. PH-derived descriptors capture these higher-order organizational traits by reflecting enhanced persistence of connected components and loops in vigorous plants. When combined with skeleton-based features, PH enables a quantitative and interpretable assessment of architectural superiority in hybrids, establishing a robust framework for topology-driven vigor analysis and genotype-to-phenotype interpretation.

6.8 Summary of the Thesis

This work presents an image-based phenotyping framework for quantitative analysis of plant vigor, architecture, and genotype differentiation, where visual plant structures are systematically converted into numerical, measurable descriptors rather than relying on qualitative or subjective assessment. High-dimensional geometric and topological features were extracted from segmented plant images, capturing both structural connectivity and spatial form. Dimensionality reduction techniques were employed to obtain compact representations that preserve dominant phenotypic variation, enabling effective clustering and visualization of growth patterns. Unsupervised clustering revealed distinct phenotypic groups, while supervised learning models were used to classify genotypes and interpret key discriminative traits. The proposed framework provides a robust, scalable, and data-driven approach for plant phenotyping, supporting objective assessment of growth dynamics and genotype-specific architectural characteristics.

CHAPTER 7

Conclusion

The major conclusions drawn from this study are summarized as follows:

- The proposed *Structural Complexity Index (SCI)* provides a compact and interpretable scalar representation of plant architectural complexity by integrating complementary shape and skeleton-based descriptors.
- The SCI framework successfully captures biologically meaningful structural variation across developmental stages and genotypes.
- The proposed *non-linear Structural Complexity Index (nSCI)* extends the linear SCI formulation through RKHS-based embedding, enabling modeling of non-linear and conditional trait interactions.
- Kernel-based modeling improves genotype separability by preserving complex architectural dependencies beyond linear representations.
- The integration of *Persistent Homology (PH)* introduces topology-aware structural characterization for detecting higher-order connectivity patterns.

- EDT-based persistent homology descriptors demonstrate robustness to noise, segmentation artifacts, and structural perturbations.
- Persistence image transformation enables fixed-dimensional machine-learning-compatible topological encoding.
- The combined SCI–nSCI–PH framework significantly enhances genotype discrimination by jointly capturing geometric, structural, and topological properties.
- Ensemble and kernel-based classifiers effectively exploit the proposed feature space for accurate genotype prediction.
- The proposed framework establishes a scalable computational pipeline for high-throughput genotype-to-phenotype analysis.

The results collectively demonstrate that plant architectural complexity can be effectively quantified through a unified multi-level representation framework. By integrating linear structural descriptors, non-linear kernel representations, and topology-aware analysis, this work provides a mathematically grounded and biologically interpretable approach for robust plant phenotyping. The proposed framework not only improves genotype discrimination performance but also establishes a foundation for scalable architectural analysis across developmental stages and varying environmental conditions.

References

[1] D. J. Watson, "Comparative physiological studies on the growth of field crops: I. variation in net assimilation rate and leaf area between species and varieties, and within and between years," *Annals of Botany*, vol. 11, pp. 41–76, 1947.

[2] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W. H. Freeman and Company, 1983.

[3] C. Fournier and B. Andrieu, "A 3d architectural model of maize development," *Annals of Botany*, vol. 81, no. 1, pp. 107–119, 1998.

[4] R. T. Furbank and M. Tester, "Phenomics–technologies to relieve the phenotyping bottleneck," *Trends in Plant Science*, vol. 16, no. 12, pp. 635–644, 2011.

[5] F. Hallé, R. A. A. Oldeman, and P. B. Tomlinson, *Tropical Trees and Forests: An Architectural Analysis*. Berlin, Heidelberg: Springer, 1978.

[6] J. Cohen, V. Alchanatis *et al.*, "A comprehensive review of non-destructive phenotyping techniques in plant breeding," *Journal of Experimental Botany*, vol. 67, no. 12, pp. 345–358, 2016.

[7] N. Fahlgren, M. A. Gehan *et al.*, "A comprehensive analysis of plant phenotyping using imaging technologies," *Nature Methods*, vol. 12, no. 1, pp. 21–29, 2015.

[8] P. Sadeghi-Tehran, K. Sabermanesh *et al.*, "Automated phenotyping of plant architecture using rgb images: A review," *Frontiers in Plant Science*, vol. 8, pp. 1–17, 2017.

[9] S. Meraj *et al.*, "Automated plant phenotyping: a comprehensive survey," unpublished.

[10] B. Singhal *et al.*, "High-throughput image-based crop phenotyping in india: challenges and prospects," *Frontiers in Plant Science*, vol. 14, 2023.

[11] P. Singh *et al.*, "Labor intensity in traditional crop trait scoring," *Field Crops Research*, vol. 256, p. 108014, 2020.

[12] A. Ghosal *et al.*, "Leaf area and biomass modeling via image-based features," *Plant Methods*, vol. 15, pp. 37–45, 2019.

[13] S. Rao *et al.*, "Color index-based heat stress detection," *Frontiers in Plant Science*, vol. 13, pp. 841–850, 2024.

[14] J. Singh *et al.*, "Shape metrics for drought stress classification," *Plant Methods*, vol. 19, pp. 42–52, 2023.

[15] J. Yang *et al.*, "Plant phenomics research trends," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 526–534.

[16] D. Cherry and M. Picker, "Skeleton-based phenotyping methods," unpublished.

[17] L. Kumar *et al.*, "Curvature and branch-point analysis in plant imagery," *Plant Methods*, vol. 14, pp. 68–80, 2023.

[18] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint*, 2018.

[21] M. A. Gehan, N. Fahlgren *et al.*, "Plantcv v2: Image analysis software for high-throughput plant phenotyping," *PeerJ*, vol. 5, p. e4088, 2017.

[22] Z. Liang, H. Zhou, and X. Wang, "Skeleton-based plant phenotyping for leaf shape and growth analysis," *Computers and Electronics in Agriculture*, vol. 178, 2020.

[23] E. Ubbens and I. Stavness, "Global trends in plant phenotyping," *Frontiers in Plant Science*, vol. 11, pp. 700–710, 2020.

[24] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[25] F. Chazal and B. Michel, "An introduction to topological data analysis: fundamental and practical aspects for data scientists," *arXiv preprint arXiv:1710.04019*, 2017.

[26] L. Wasserman, "Topological data analysis," *Annual Review of Statistics and Its Application*, vol. 5, pp. 501–532, 2018.

[27] H. Edelsbrunner and J. Harer, "Persistent homology—a survey," *Contemporary Mathematics*, vol. 453, pp. 257–282, 2008.

[28] K. Das *et al.*, "Observer bias in manual plant measurements," *Plant Methods*, vol. 14, pp. 5:1–5:9, 2018.

[29] L. Li, Q. Zhang, and D. Huang, "A review of imaging techniques for plant phenotyping," *Sensors*, vol. 14, no. 11, pp. 20 078–20 111, 2014.

[30] R. T. Furbank and M. Tester, "Phenomics – technologies to relieve the phenotyping bottleneck," *Trends in Plant Science*, vol. 25, no. 3, pp. 184–195, 2020.

[31] J. M. Chen *et al.*, "Leaf area index: Its definition, measurement, and applications," *Remote Sensing of Environment*, vol. 69, no. 2, pp. 71–84, 1997.

[32] M. R. Golzarian, R. Frick, and K. Rajendran, "Accurate inference of shoot biomass from high-throughput images of cereal plants," *Plant Methods*, vol. 7, no. 1, p. 2, 2011.

[33] G. Hammer, E. van Oosterom, R. McLean, and et al., "Plant architecture and implications for crop design and harvestability," *Journal of Experimental Botany*, vol. 70, no. 18, pp. 4481–4494, 2019.

[34] L. Gonsalves *et al.*, "The role of imaging technologies in understanding plant architecture," *Plant Physiology*, vol. 182, no. 1, pp. 1–12, 2020.

[35] LemnaTec GmbH, "Lemnatec high-content screening: Analysis parameters," https://www.lemnatec.com/phenoaixpert/, accessed 2024.

[36] R. Kolhar and P. Jagtap, "Rgb image-based plant phenotyping: a review," *Sensors*, vol. 23, no. 4, pp. 1147–1166, 2023.

[37] M. Osman, S. Sankaran *et al.*, "Deep learning for image-based phenotyping in plants," *Agricultural Systems*, vol. 179, pp. 102–113, 2020.

[38] D. Jha, A. Singh *et al.*, "Advances in machine learning for plant phenotyping," *Plant Science*, vol. 310, p. 110828, 2021.

[39] M. Minervini, A. Abdelsamea, and S. Tsaftaris, "Image-based plant phenotyping with deep learning," *Trends in Plant Science*, vol. 26, no. 11, pp. 1092–1104, 2021.

[40] Y. Wang, X. Zhang, Z. Zhao, and H. Xu, "Explainable deep learning for plant phenotyping: A review," *Computers and Electronics in Agriculture*, vol. 195, 2022.

[41] B. Xu, J. Tang, and H. Gao, "A survey on explainable ai: Techniques and applications in plant science," *Artificial Intelligence Review*, vol. 53, pp. 4973–4993, 2020.

[42] H. Tang, J. Zhou, and Y. Wu, "A new framework for plant architecture analysis combining skeleton extraction and machine learning," *Plant Methods*, vol. 17, no. 1, 2021.

[43] S. Kumar, A. Shukla *et al.*, "Integrating advanced imaging with phenotyping techniques for agriculture," *Agricultural Systems*, vol. 172, pp. 88–98, 2019.

[44] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[45] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[46] Y. Li, X. Chen *et al.*, "Application of pca-based methods for plant structure analysis," *Agronomy*, vol. 12, no. 4, p. 803, 2022.

[47] A. Singh, K. Ganapathysubramanian, S. K. Singh, and A. Sarkar, "Machine learning for high-throughput stress phenotyping in plants," *Trends in Plant Science*, vol. 26, no. 1, pp. 88–100, 2021.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[49] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[50] F. Tardieu *et al.*, "Phenotyping and plant functional genomics: Reaching the next level," *Trends in Plant Science*, vol. 22, no. 4, pp. 314–324, 2017.

[51] J. R. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness, "Latent space phenotyping: automatic image-based phenotyping for treatment studies," *Plant Methods*, vol. 16, no. 1, pp. 1–16, 2020.

[52] S. W. Running and J. C. Coughlan, "A general model of forest ecosystem processes for regional applications. i. hydrologic balance, canopy gas exchange and primary production processes," *Ecological Modelling*, vol. 42, no. 2, pp. 125–154, 1988.

[53] P. J. Sellers, R. E. Dickinson, D. A. Randall, A. K. Betts, F. G. Hall, J. A. Berry, G. J. Collatz, A. S. Denning, H. A. Mooney, C. A. Nobre, N. Sato, C. B. Field, and A. Henderson-Sellers, "Biosphere–atmosphere exchanges of carbon, water, and energy," *Science*, vol. 275, no. 5299, pp. 502–509, 1997.

[54] G. S. Campbell and J. M. Norman, *An Introduction to Environmental Biophysics*. Springer, 1998.

[55] M. Monsi and T. Saeki, "Über den lichtfaktor in den pflanzengesellschaften," *Japanese Journal of Botany*, vol. 14, pp. 22–52, 1953.

[56] J. M. Norman and P. G. Jarvis, "Photosynthesis in sitka spruce," *Journal of Applied Ecology*, vol. 12, pp. 839–878, 1975.

[57] M. Weiss, F. Baret, R. B. Myneni, A. Pragnère, and Y. Knyazikhin, "Investigation of a model inversion technique to estimate canopy biophysical variables from spectral and directional reflectance data," *Agricultural and Forest Meteorology*, vol. 127, no. 1–2, pp. 1–16, 2004.

[58] S. e. a. Garrigues, "Validation and intercomparison of global leaf area index products," *Journal of Geophysical Research*, vol. 113, 2008.

[59] J. M. Chen and T. A. Black, "Leaf area index of boreal forests: Theory, techniques, and measurements," *Journal of Geophysical Research*, vol. 96, pp. 2865–2877, 1991.

[60] L. Yang, M. Li, and S. Wu, "Interpretable ai models for plant phenotyping using multimodal imaging data," *Frontiers in Plant Science*, vol. 14, 2023.

[61] M. G. et al., "Topological data analysis expands the genotype to phenotype map for 3d maize root system architecture," *Frontiers in Plant Science*, 2023. [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1260005/full

[62] H. Oda, M. Kida, Y. Nakata, and H. Kurihara, "Novel definition and quantitative analysis of branch structure with topological data analysis," *arXiv preprint arXiv:2402.07436*, 2024. [Online]. Available: https://arxiv.org/abs/2402.07436

[63] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W. H. Freeman and Co., 1982.

[64] X. Zhang, Y. Huang, and W. Zhang, "Morphological and skeleton-based leaf phenotyping using deep learning," *Computers and Electronics in Agriculture*, vol. 187, 2021.

[65] A. Raza, M. M. Siraj, and K. Lee, "Machine learning for plant stress phenotyping," *Plant Science*, vol. 307, 2021.

[66] H. Li, Z. Huang, J. Zhao, and X. Li, "Applying persistent homology to analyze leaf shapes for plant species classification," *Frontiers in Plant Science*, vol. 12, 2021.

[67] L. T. Berner, T. H. White, J. P. Anderson, and J. T. Aber, "Topological data analysis reveals plant root architecture variation," *New Phytologist*, vol. 234, no. 5, pp. 1796–1808, 2022.

[68] H. Hotelling, "Analysis of a complex of statistical variables into principal components (part ii)," *Journal of Educational Psychology*, vol. 24, no. 7, pp. 498–520, 1933.

[69] Y. Liu, Z. Wu *et al.*, "Deep learning for plant phenotyping: A case study of plant architecture analysis using 3d images," *Plant Methods*, vol. 17, p. 45, 2021.

[70] D. H. Chitwood, L. R. Headland, A. Ranjan, C. C. Martinez, S. A. Braybrook, D. Koenig, S. B. Kang, S. Chacko, J. N. Maloof, and N. R. Sinha, "Leaf shape and curvature are regulated by multivariate genetic interactions in tomato," *Journal of Experimental Botany*, vol. 61, no. 10, pp. 2757–2768, 2013.

[71] X. Zhang, Y. Li, J. Chen, and H. Wang, "Hormonal and mechanical regulation of branch curvature in plants," *Scientific Reports*, vol. 15, p. 17474, 2025.

[72] G. B. West, J. H. Brown, and B. J. Enquist, "A general model for the origin of allometric scaling laws in biology," *Science*, vol. 276, no. 5309, pp. 122–126, 1997.

[73] K. J. Niklas, *Plant Allometry: The Scaling of Form and Process*. Chicago, IL: University of Chicago Press, 2016.

[74] D. L. Royer, P. Wilf, and B. K. Janes, "Geometric morphometrics of plant leaves: Capturing multivariate shape variation in ecology and evolution," *Plants*, vol. 14, no. 5, p. 808, 2022.

[75] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[76] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of ACM SIGKDD*, 2016, pp. 785–794.

[77] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*. MIT Press, 1986, vol. 1, pp. 318–362.

[78] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[79] L. Meyer, A. Gilson, O. Scholz, and M. Stamminger, "Cherrypicker: Semantic skeletonization and topological reconstruction of cherry trees," *arXiv preprint arXiv:2304.04708*, 2023. [Online]. Available: https://arxiv.org/abs/2304.04708

[80] T. Kozai, G. Niu, and M. Takagaki, *Plant Factory: An Indoor Vertical Farming System for Efficient Quality Food Production*. Academic Press, 2019.

[81] K. Benke and B. Tomkins, "Future food-production systems: vertical farming and controlled-environment agriculture," *Sustainability: Science, Practice and Policy*, vol. 13, no. 1, pp. 13–26, 2017.

[82] R. Pieruschka and H. Poorter, "Phenotyping plant responses to environmental change," *Functional Plant Biology*, vol. 39, pp. 839–850, 2012.

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

Publications from this Thesis

Conferences

A part of this research has been peer-reviewed and published in an international conference proceedings indexed by Springer:

- **Betal, P.,** Mitra, P., Mandal, C., Dasgupta, A., Kumar Deb, C., Das, M. (2026). ***Quantifying Plant Structural Complexity Using an Image-Based Structural Complexity Index (SCI)***. In: Subudhi, B.N., Rout, R.K., Sengar, B.S. (eds) *Computer Vision, Pattern Recognition, Image Processing, and Graphics*. NCVPRIPG 2025. Communications in Computer and Information Science, vol. 2684. Springer, Cham. `https://doi.org/10.1007/978-3-032-08508-5_19`

- **Betal, P.**, Mitra, P., Mandal, C., Dasgupta, A. (2025). ***A Multi-Descriptor Framework for Image-Based Genotype Classification in Plants***. In: *Proceedings of the 2025 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, pp. 1–6. IEEE. `https://doi.org/10.1109/CVMI66673.2025.11337833`

- **Betal, P.,** Mitra, P., Dasgupta, A., Mandal, C., Gupta, P. (2025). ***Topological Analysis of Plant Structure Using Persistent Homology for Genotype Classification***. In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2025)*, Indian Institute of Technology Mandi, Mandi, India. To appear in *ACM*. ACM ISBN: 979-8-4007-1930-1. `https://doi.org/10.1145/3774521.3774623`

Author's Biography

Prasenjit Betal is an MS candidate from **CENTRE FOR COMPUTATIONAL & DATA SCIENCES , Indian Institute of Technology Kharagpur**.

Contact Information

Permanent Address: Satpati, Paschim Medinipur
West Bengal, PIN–721516, India.
Mobile: +91-9547260760
Email: prasenjitbetal2@gmail.com

Research Interests

Computer Vision, Deep Learning, Artificial Intelligence.